eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Facultad de Economía y Empresa
Sección de Gipuzkoa
Grado en Administración y dirección de Empresas

Trabajo de fin de grado

Diagramas de sectores en R

Jorge Ramos Arcas

Índice

1 Introducción

El ámbito de estudio de este trabajo es el método de análisis gráfico conocido como diagrama sectorial, así como su elaboración y sus posibles alternativas dentro del programa estadístico R.

Es evidente que el uso de dichos gráficos se ha popularizado a la hora de realizar una presentación de datos al público, pese a estar muy estigmatizado en los entornos académicos. Analizaremos sus debilidades tanto como sus fortalezas con el fin de obtener una respuesta clara a esta cuestión.

Una de las razones principales por las que me he decantado por este tema, ha sido por el máster de *big data* y analítica empresarial que comencé a cursar online, ya que los conocimientos adquiridos en él sobre el software y el lenguaje de R me han sido de gran utilidad.

Respecto al marco teórico, haré hincapié en los tipos de variables estadísticas, en el origen y en una crítica hacia el diagrama de sectores para posteriormente entrar a explicar R en profundidad.

De cara a el apartado práctico del trabajo, he decidido estructurar los supuestos progresivamente en base a su complejidad comenzando con los *pie chart* básicos de R, llegando hasta los más sofisticados.

2 Tipos de variables estadísticas

Antes de entrar a analizar los diagramas de sectores, es necesario hacer una diferenciación entre los tipos de variables y sus correspondientes gráficas existentes dentro de la estadística descriptiva.

Una variable es cualquier condición susceptible de modificarse o de variar en cuanto a cantidad y calidad que sea medible. Debido a las estimaciones de carácter subjetivo, la percepción del investigador juega un papel muy importante en el análisis de los resultados obtenidos.

Dentro del estudio de la estadística descriptiva, existen dos grupos de variables cuyos comportamientos tienden a analizarse, las variables cualitativas y cuantitativas, ya sea a través de las mediciones numéricas o abstracciones no medibles numéricamente.

Las variables cualitativas hacen referencia a los atributos y propiedades que no permiten una medición numérica; las variables cuantitativas sin embargo, sí permiten ser medidas por medio de valores numéricos.

2.1 Variables cuantitativas

Las variables cuantitativas son aquellas variables estadísticas que nos proporcionan un resultado mediante un valor numérico, facilitando una sencilla comprensión del valor de estudio. Algunos ejemplos de variables cuantitativas podrían ser la temperatura (20ºC, 24ºC, etc.), la altura (1.75 cm, 1.82 cm, etc.) o el número de alumnos en una aula (25, 30, 35…).

Entre las características más relevantes que poseen estas variables distinguiremos las siguientes: uso de los números como indicadores, empleo conjunto de los valores con diagramas de barra, diagramas integrales y también con diagramas diferenciales para indicar la frecuencia relativa de dichas variables.

2.1.1 Discretas

Las cifras de estas variables se encuentran separadas unas de otras mediante escalas, por tanto, no hay otros valores entre los valores o cifras en cuestión (valores decimales), en vez de ello, nos hallamos con valores exactos.

He aquí algunos ejemplos de variables que únicamente poseen un valor de un conjunto de números exactos. El número de los hijos en una unidad familiar (1, 2, 3, etc.) o la evolución del número de habitantes en un municipio (10.000, 20.000, etc.) siempre estarán compuestos por números exactos y nunca por números decimales (2,25 hijos o 15.384,5 habitantes por ejemplo).

2.1.2 Continuas

A diferencia de las variables discretas, estas variables se encuentran dentro de una escala continuada, de tal forma que cada uno de los posibles individuos u objetos puedan tener su

propia puntuación. Entre las variables más reconocibles encontraríamos las calificaciones en los exámenes, la altura, el peso o la presión atmosférica. Los intervalos entre las variables pueden ser muy pequeños y pueden llegar a agruparse en función de las necesidades del estudio.

2.1.3 Métodos gráficos para representar muestras de datos cuantitativos

Existen dos tipos de diagramas estadísticos para las variables cuantitativas en relación de si se usan las frecuencias tanto absolutas como relativas o las frecuencias acumuladas: los diagramas diferenciales y los integrales.

Diagramas diferenciales son gráficos que representan frecuencias absolutas o relativas. En ellos se representa el número o porcentaje de elementos que presenta una modalidad dada.

Los diagramas integrales sin embargo, son aquellos en los que se representan el número de elementos que muestra una modalidad inferior o igual a una indicada. Se elaboran a partir de las frecuencias acumuladas, por tanto, los gráficos son ascendentes, y es por ello que no tienen ninguna aplicación práctica con las variables cualitativas.

He aquí algunas de las representaciones gráficas más eficientes y más empleadas tanto con variables cuantitativas continuas como discretas.

Histogramas

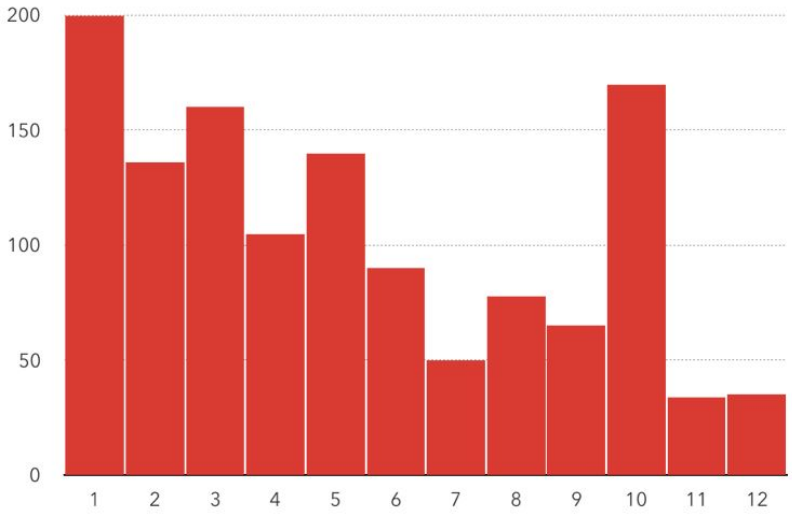


Figura 1. Ejemplo de histograma

Un histograma es un tipo de gráfico que muestra la distribución de frecuencias de una variable cuantitativa y continua en forma de barras, en el cual la superficie de cada barra es proporcional a la frecuencia de los valores ilustrados. Generalmente, cuando las variables a mostrar en el histograma son discretas, los valores a trazar pasan a ser específicos y en consecuencia se añade una separación entre columnas que representa la ausencia de un espectro continuo de valores.

Polígonos de frecuencia

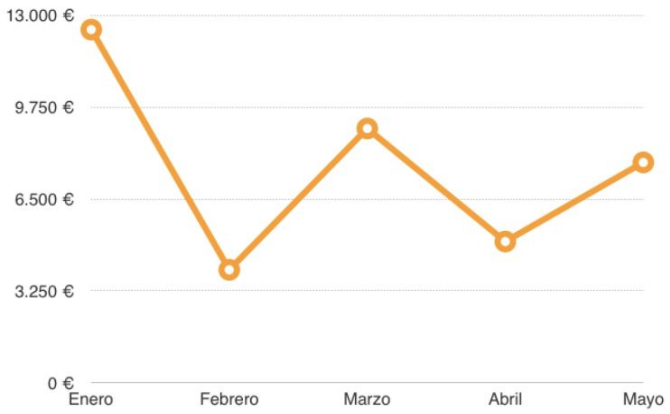


Figura 2. Ejemplo de polígono de frecuencias

Llamamos polígonos de frecuencia al gráfico que se obtiene de la unión mediante segmentos de los puntos de máxima altura dentro de las columnas de un histograma. Aunque su uso se puede limitar para representar variables cuantitativas principalmente, también admite clasificar una variable continua cualitativa con otra cuantitativa.

Entre algunas de las características de este gráfico podemos destacar que no nos proporciona ningún tipo de frecuencia acumulada, que el punto con mayor altura es el que indica la mayor frecuencia y que la zona por debajo de los segmentos representa toda la muestra.

Diagrama de barras

Se emplea para representar datos cuantitativos de tipo discreto (aunque también puede emplearse con datos cualitativos como veremos más adelante). La información se presenta mediante unas barras cuya altura es proporcional a la frecuencia. En su eje de ordenadas encontramos las frecuencias absolutas o relativas acumuladas y en su eje de abscisas situamos los distintos valores de la variable.

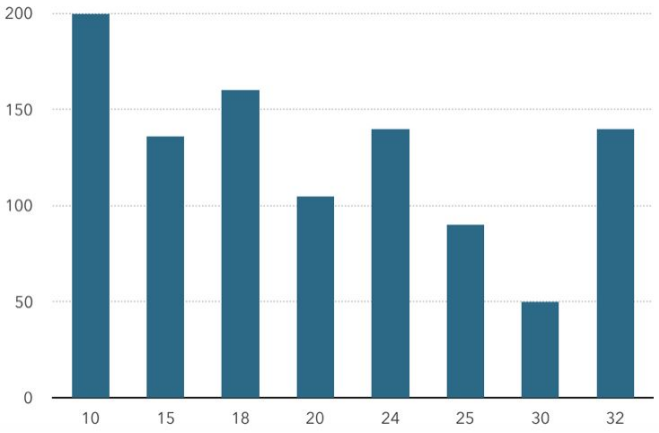


Figura 3. Ejemplo de diagrama de barras

2.2 Variables cualitativas

Denominamos a aquellas variables que toman un conjunto de valores que son cualidades no numéricas como pueden ser categorías o niveles. La marca de un automóvil (Ford, Peugeot, Renault, Skoda…), el color de pelo (negro, castaño, rubio, rojo) entre otros, serían ejemplos de variables cualitativas.

Las variables cualitativas se centran en expresar los atributos, características, categorías, circunstancias o cualidades de algún individuo o de algún objeto en vez de proporcionarnos información numérica, por tanto, se tratan de variables que no tienen sentido natural para su orden.

Para una posible aplicación y análisis de dichas variables en modelos tanto económicos, matemáticos como financieros, es posible la codificación de cada valor asignándole un

número específico (sin necesidad de que los números tengan sentido entre sí). Tomemos por ejemplo el estado civil como variable cualitativa, tomando como opciones soltero, casado, divorciado y viudo. En un análisis matemático esta variable se incluiría por ejemplo dando a todos los individuos solteros el valor 0, a los casado el 1, a los divorciados el 2 y el número 3 al grupo restante.

Para su correcta implementación en un análisis estadístico, se requiere que las variables cualitativas sean exhaustivas (es decir, que sirvan para clasificar a todos los objetos o individuos dentro del estudio) y que sean mutuamente excluyentes (que cada sujeto pueda pertenecer solamente a una categoría).

Podemos realizar dos clasificaciones dentro de las variables cualitativas. La primera sería en referencia al número de categorías y la segunda a la escala de medida de las mismas.

2.2.1 Según el número de categorías

Dentro de este grupo podemos diferenciar entre las variables dicotómicas y politómicas.

En las dicotómicas solo dispondremos de dos posibilidades, es decir, la respuesta de cada variable es binaria. Un ejemplo de esto sería el sexo (hombre, mujer), pertenecer a una familia numerosa (sí, no) o el resultado de la parte teórica del carné de conducir (aprobar o suspender).

En el caso de las politómicas, se dan cuando las variables admiten más de dos categorías, como podría ser el grupo sanguíneo, el lugar de residencia o las marcas de automóviles.

2.2.2 Según la escala de medida de las categorías

Es posible realizar una distinción entre las variables cualitativas según su escala en las distintas categorías que la conforman. Existen dos posibles clasificaciones de las variables según dicho criterio.

La primera clase estaría compuesta por las variables nominales, las cuales no dispondrían de un valor numérico asignado y sería imposible establecer un orden natural o criterio específico entre sus categorías como sería el caso de la nacionalidad, el color de pelo o la religión.

En el segundo grupo nos encontraríamos con las variables ordinales que serían aquellas cuyos valores fueran susceptibles de un orden o jerarquía y por tanto, que permitieran definir relaciones de orden de tipo mayor, menor, igual o preferencia entre los sujetos. Podemos emplear como ejemplo el nivel de estudios (educación secundaria obligatoria, bachillerato, grado superior, grado universitario, máster, etc.), el rango militar (soldado, cabo, sargento, teniente, capitán, etc.) o los niveles de eficiencia energética (A, B, C, D, E, F y G). A pesar de estas clasificaciones nos facilitan un orden determinado, no se puede analizar la distancia absoluta que se halla entre las distintas categorías.

También es posible agrupar algunas variables cuantitativas en intervalos, y tratar a dicho intervalo como una variable ordinal, de este modo sería posible calcular la distancia

numérica entre los distintos niveles de la escala ordinal como sucedería con la edad, el nivel salarial o la renta per cápita.

2.2.3 Métodos gráficos para representar muestras de datos cualitativos

Entre los distintos métodos gráficos disponibles para presentar las variables cualitativas nos centraremos en los dos diagramas más empleados: el diagrama de barras y el de sectores.

Diagrama de barras

Ya hemos hablado sobre el diagrama de barras previamente por su uso con datos cuantitativos. En este caso analizaremos su empleo con las variables cualitativas. Su principal punto fuerte radica en la posibilidad de comparar dos o más valores. En el eje de abscisas *x*, se representan los datos o modalidades y en el eje de ordenadas *y*, se representan las frecuencias de cada dato o modalidad.

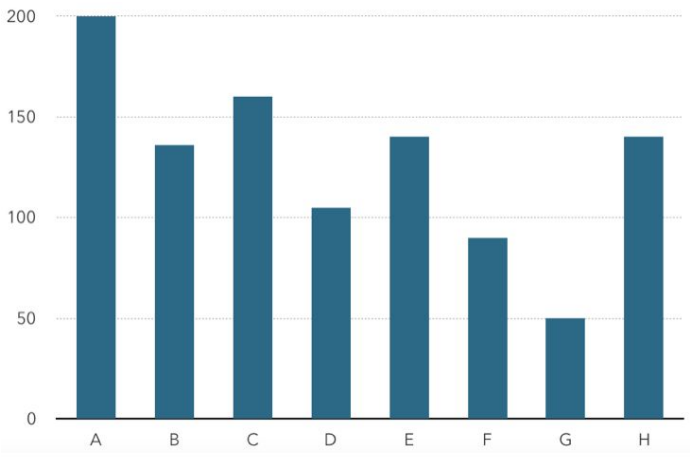


Figura 4. Ejemplo de diagrama de barras

Diagramas de sectores

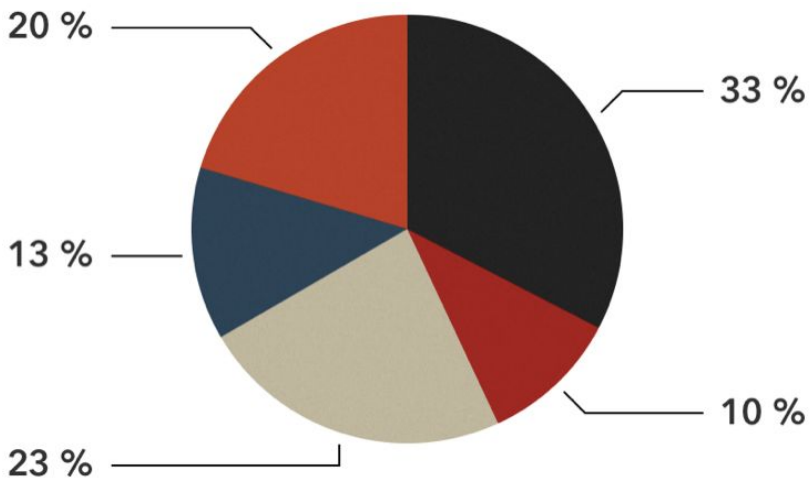


Figura 5. Ejemplo de diagrama de sectores

Consiste en una gráfica circular subdividida en áreas donde cada una de las cuales es proporcional a la frecuencia de la modalidad que representa. La circunferencia entera representa el 100% de un todo mientra las partes forman proporciones de la cantidad total. Su uso está completamente arraigado en la población y es por esto que resulta muy habitual verlo en todos los niveles (salvo a nivel académico, y más adelante veremos porque), desde los negocios hasta las encuestas.

3 Orígenes del diagrama de sectores

3.1 The Statistical Breviary

La primera aparición del diagrama de sectores de la cual se tiene constancia se dio en 1801 en una publicación de William Playfair, titulada "The Statistical Breviary". Previamente en 1786, ideó y popularizó el uso de los gráficos de barras y líneas para representar las series temporales en la estadística.

En "The Statistical Breviary", Playfair emplea todo tipo de gráficas (principalmente los gráficos de tarta) para analizar las áreas geográficas, las poblaciones y los ingresos de los estados europeos de finales del siglo XVIII hasta comienzos del XIX. En el documento, William optó por el uso de representaciones gráficas de información cuantitativa dado que él creía que haciendo visibles las magnitudes y proporciones involucradas era el método más eficiente para transmitir ideas diferentes.

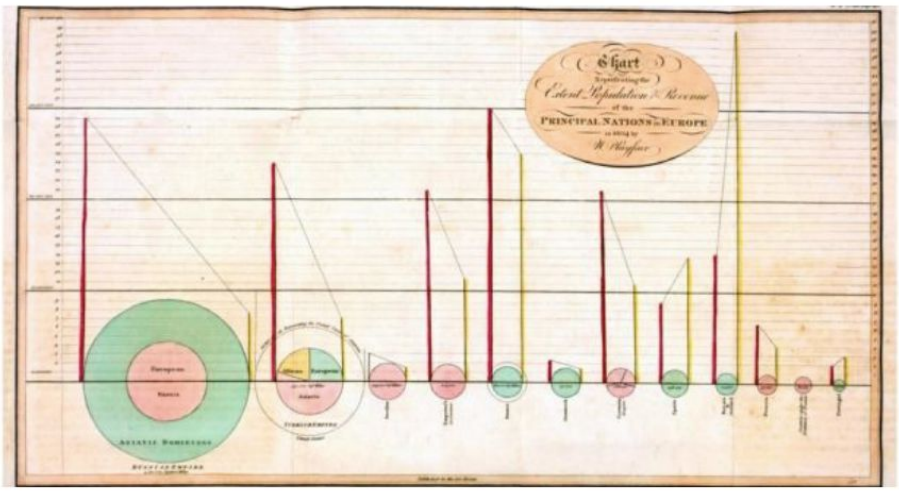


Figura 6. Gráficos empleados por Playfair en "The Statistical Breviary"

El primer gráfico circular del que se tiene constancia representaba la situación previa de los países europeos antes de la Revolución Francesa de 1789, y el segundo gráfico mostraba la evolución de dichos países en 1801. El área del círculo que formaba cada gráfica era proporcional al área real de cada país. Además de la información mencionada, los gráficos también mostraban si el país en cuestión era una potencia marítima o terrestre, mostrando las áreas de cada gráfico en verde o en rojo respectivamente. Los ingresos por impuestos venían representados por una línea vertical amarilla situada a la derecha de cada círculo mientras que la población venía representada por otra línea vertical roja situada a la izquierda. Para mostrar las diversas subdivisiones políticas, Playfair dividió el imperio ruso en la región europea y en la región asiática. El anillo circundante que representaba a la parte oriental, lo indicó en verde dado a su poder marítimo y su valor estratégico, la zona interior que representaba a la zona europea sin embargo, la marcó en rojo como potencia terrestre.

En el caso del imperio turco, tuvo que dividir el dominio en tres áreas, la zona asiática, la europea y la africana. Como dividir el área en tres círculos concéntricos habría dificultado la comparativa visual (como ya había hecho con el imperio ruso) y como el objetivo de Playfair era facilitar la comparación y hacer la información más entendible, optó por descartar este método. En su lugar, dividió el círculo en tres sectores proporcionales a las dimensiones de la región asiática, europea y africana. Indicó la zona asiática de verde indicando su potencia marítima, la europea en rojo como potencia terrestre y empleó el amarillo para la africana. Todavía a día de hoy se siguen sin saber si William empleó el amarillo por decisión unilateral o porque era consciente de que mezclando las luces verdes y rojas se obtenía el amarillo, señalando con esto que la zona africana era una una zona marítima y terrestre equilibrada.

Figura 7. Diagrama sectorial del Imperio Turco empleado en "The Statistical Breviary"

En el segundo gráfico del documento (p.49; figura 1) centrado en el año 1801 (el año del tratado de paz de Luneville entre Francia y Austria), es parecido al primero pero con unos pocos países menos. Todo esto ocasionado por la guerra europea y los avances de Napoleón, que obligó a hacer varios realineamientos y cambios a nivel político. En dicha tabla, Playfair dividió el imperio alemán en otro diagrama de sectores para representar el área de los territorios pertenecientes a Austria, Prusia y el del imperio alemán. En el caso del imperio turco, aparte de la gráfica circular, también empleó un diagrama de tres círculos superpuestos (diagrama de Venn) para indicar la propiedad conjunta del estado. El círculo izquierdo indicaba los intereses austríacos, en el de la derecha los terrenos de Prusia y en el círculo (central el cual tenía la misma área que el gráfico tarta) los del imperio alemán.

Como hemos podido observar, Playfair introdujo tres nuevas formas de diagramas estadísticos en dos gráficas del mismo documento: el diagrama circular, el de sectores y el diagrama de Venn. En el caso de los diagramas de sectores, empleó el ángulo de los segmentos para indicar la proporción, también utilizó el coloreado y el etiquetado para referenciar a cada uno de los sectores.

Los diseños iniciales de los diagramas de sectores de Playfair se han ido optimizando a lo largo de los años, al igual que ocurrió con los gráficos de barras y líneas que él mismo introdujo 15 antes del "The Statistical Breviary". En el caso de los diagramas de Venn, dado que su uso no suele ser muy práctico a la hora de analizar datos estadísticos cuantitativos, no suele ser muy habitual su aplicación.

3.2 Autores que influenciaron a Playfair

Los orígenes de las otras invenciones gráficas de Playfair como son el gráfico de barras y el de líneas no son desconocidas. En el caso del primero, se basó en el diagrama cronológico de Priestley (publicado en 1765). Playfair era un asiduo de las gráficas de líneas temporales, las cuales empleaba para mostrar la esperanza de los individuos respecto a una determinada escala temporal. En el caso de los diagramas de líneas, William le cedió todo el

crédito a su hermano mayor dado que de joven le obligaba a realizar registros diarios de la temperatura y a trazar los datos de manera gráfica. A diferencia de los dos métodos gráficos mencionados en los cuales el autor se centró en abordar sus fuentes de inspiración en multitud de ocasiones, no sucede los mismo con los diagramas de sectores. Cabe la posibilidad de que Playfair no diera ninguna información al respecto de las posibles fuentes en las que se basó por pensar que se trataban de diagramas muy evidentes, los cuales podrían haber sido descubiertos por cualquiera.

Para intentar abordar la causa de llevó a Playfair a pensar por qué las explicaciones sobre estas fuentes no eran necesarias, nos tendríamos que remontar a su juventud. Con la muerte de su padre a los al tener cuando él tenía 12 años, su hermano John se quedó a cargo del cuidado y educación de sus hermanos. Posteriormente, John se convirtió en uno de los matemáticos más prestigiosos de Escocia como profesor de matemáticas en la Universidad de Edinburgo, y también en renombrado geólogo y físico. Las conversaciones sobre matemáticas y la resolución de puzzles debían de ser habituales en la familia Playfair, es por ello que tal vez John influenciara a su hermano incluso inconscientemente. William a lo largo de su trayectoria como autor tiene varios ejemplos sobre adaptaciones de trabajos ajenos. Como resultado a sus adaptaciones, Playfair derivó en las gráficas de barras y líneas, pero es muy probable que este no fuera en caso con las gráficas circulares.

En referencia a sus diagramas de intersecciones circulares, en 1880, Venn los empleó en su trabajo sobre el sistema lógico de Boole. Pero no fue el primero en emplearlos dado que Euler los usó con la misma intención en 1768 y Leibniz los introdujo en su trabajo de proposiciones lógicas en 1666, influenciados por Ramón Llull y Giordano Bruno. Se tiene constancia de que John Playfair estaba estrechamente relacionado con los trabajos de Leibniz y Euler, además, John fue ministro ordenado en la iglesia de Escocia y por consiguiente, era altamente improbable que no hubiera tenido constancia de los escritos de los pensadores religiosos como Ramón Llull y Giordano Bruno.

Remontándonos a su orígenes, Ramón Llull introdujo las bases de los diagramas lógicos en su obra *Ars Magna* (1305). Él representó los conceptos mediante círculos y sugirió semejanza entre ellos superponiendo los círculos entre sí, a pesar de que no llegó a desarrollar el concepto moderno de intersección de Baron (1969). Mecanizó parte de su esquema ordenando las categorías en discos con diámetros diversos que podían ser girados con el propósito de crear el mayor número de combinaciones posibles de conceptos. Sin embargo, dichas ideas no fueron desarrolladas hasta la llegada de Giordano Bruno en el siglo XV. Bruno, empleó los diagramas circulares donde el todo estaba representado por el círculo y las relaciones por los segmentos, también incrementó el número de estos últimos en sus diagramas. Cabe destacar que ninguno de los dos utilizó estos diagramas para exponer datos empíricos.

Leibniz infundido por las ideas de estos dos pensadores medievales, modificó sus ideas y estableció el inicio de la lógica objetiva. Además, fue el primer matemático en poner el foco sobre las proposiciones lógicas mediante el empleo de estos diagramas a través de su trabajo *Dissertatio de Arte Combinatoria* de 1666. En él explica los diversos métodos geométricos empleados para mostrar los silogismos aristotélicos que empleó en su estudio de las combinaciones, las cuales fueron elaboraciones hechas a partir de los trabajos de Llull como el mismo Leibniz indicó en más de una ocasión es sus escritos. Entre los diagramas

que usó, se encontrarían los diagramas de Venn y unas versiones lineales de estos ideadas por el propio autor, aunque no se popularizaron los diagramas circulares hasta 1768 gracias a Euler.

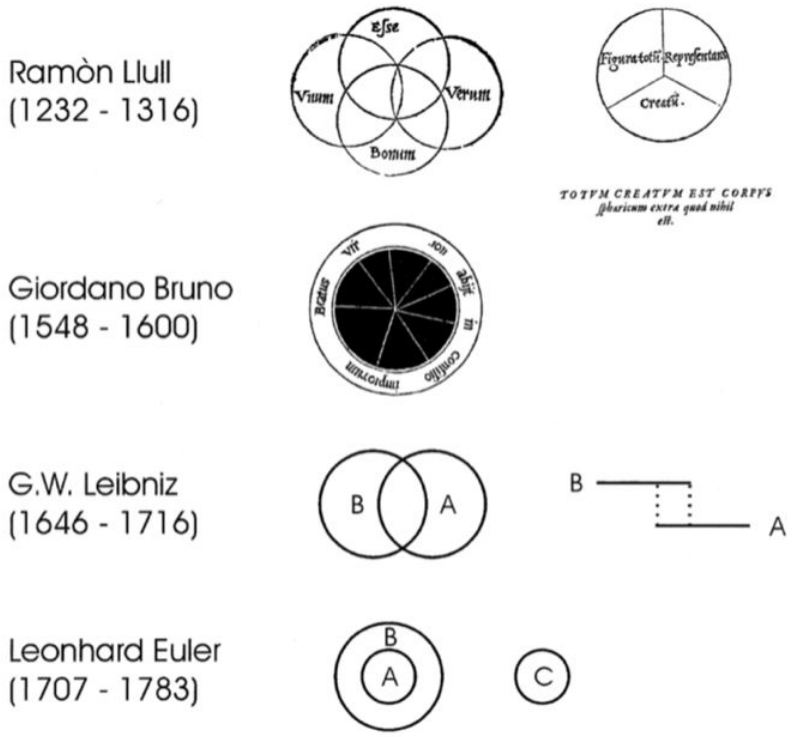


Figura 8. Antecedentes al diagrama sectorial propuesto por Playfair

3.3 La adopción de los gráficos circulares

A pesar de los gráficos innovadores que llegó a elaborar, el trabajo de Playfair fue completamente desestimado en Gran Bretaña, principalmente por la reputación que él mismo había obtenido. Estuvo involucrado un una multitud de negocios comerciales a comienzos de 1780 y acarreó una gran cantidad de deudas por ello. También fue acusado de apropiación intelectual en alguna de sus patentes. Posteriormente, estuvo envuelto en numerosos escándalos financieros en la París revolucionaria y más adelante con el Banco de Inglaterra, los cuales le llevaron a ser perseguido por la justicia en varias ocasiones. Toda esta reputación que Playfair cosechó, dificultó la propagación y aceptación de sus ideas en los entornos académicos.

Sin embargo, los escritos de Playfair calaron mejor en Francia y Alemania donde personajes ilustres como Alexander von Humboldt, y Charles-Joseph Minard (años más tarde a mediados de siglo) los implementaron en sus trabajos, especialmente este último, el cual fue un acérrimo defensor de los diagramas de sectores, como se puede observar en su conocido trabajo *Cartes Figuratives* (1858). Después de Mainard, otra personalidades como Palsky o Bertillon (Atlas de la ciudad de París, 1891) se hicieron eco de ellos.

3.4 Análisis empírico de los diagramas de sectores durante sus primeras aplicaciones

Los gráficos estadísticos no fueron empleados en el Reino Unido hasta la llegada de William Stanley Jevons, quien los adaptó para su atlas económico en 1860. Con el paso del tiempo, Jevons influenció a Karl Pearson y este fue uno de los principales responsables en que se propagara el uso de los gráficos entre los estadistas británicos.

Por otra parte, con el paso de los años el gráfico circular fue objeto de crítica y fue desaprobado su uso en gráficos estadísticos dado que tachaban a los diagramas de sectores como una forma errónea de presentación cuantitativa (Brinton, 1914). Todo este pensamiento crítico sobre su uso derivó en los primeros experimentos psicológicos sobre los gráficos.

Uno de los pioneros en esta clase de experimentos fue Walter Crosby Eells. En su trabajo "The Relative Merits of Circles and Bars for Representing Components parts", Eells sometió a dos métodos estadísticos gráficos (el diagrama circular y el gráfico de barras apiladas) a diversas pruebas con tal de intentar averiguar la eficiencia de cada uno a la hora de exponer información cuantitativa. Además, cayó en la cuenta de que la gran preferencia existente por parte de los autores de la época hacia el diagrama de barras carecía de datos empíricos que justificaran su eficacia.

En su estudio, Eells contó con 100 estudiantes de psicología de Whitman College, a los cuales sometió a un experimento comparativo entre los gráficos de barras apiladas y circulares y concluyó que los diagramas circulares eran visualmente mejores que los gráficos de barras apiladas.

9 meses después de las publicaciones de Eells, von Huhn refutó el experimento. Según von Huhn, los datos obtenidos por Eells fueron limitados por un mal planteamiento del experimento, en el cual no podían compararse el gráfico circular y el de barras. La principal carencia que von Huhn criticó, fue la ausencia de una escala con las magnitudes de diversos segmentos junto a las barras, la cual habría aumentado la precisión de los sujetos, y la falta de un etiquetado para cada segmento del gráfico circular.

Posteriormente, Croxton el cual también fue muy crítico con Eells, realizó el mismo experimento que él. Para ello empleó dos gráficos circulares y dos gráficos de barras apiladas entre los cuales los 287 participantes tenían que indicar el ratio de los segmentos. Las muestras obtenidas favorecieron en este caso a los gráficos de barras apiladas.

Los resultados de Croxton mostraron serias ineficiencias al igual que los de Eells, por tanto, prosiguió con sus experimentos pero sin llegar a comparar nunca el gráfico de barras respecto al diagrama de sectores (Croxton & Stein, 1932; Croxton & Striker, 1927).

Los resultados de estos estudios permanecieron inconclusos, por ello, las ideas de Eells, von Huhn, Croxton, Stein y Striker no llegaron a cuajar en la comunidad estadística.

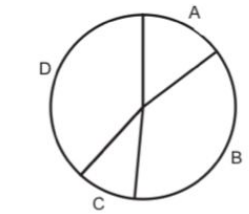
Los primeros en obtener resultados concluyentes fruto de sus estudios fueron Cleveland y McGill. El primer experimento que realizaron comparaba los gráficos de barras apiladas

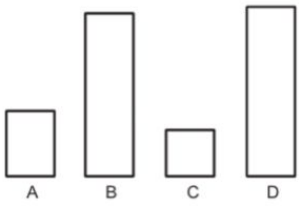
verticales frente a las barras verticales. En su segundo estudio, compararon los gráficos de barras verticales respecto a los circulares de manera aleatoria.
El análisis estadístico obtenido de ambos experimentos mostró que la precisión variaba con cada método gráfico. Cleveland y McGill llegaron a la conclusión de que el gráfico con menor margen de error fue el de barras verticales seguido del circular y por último el de barras apiladas verticales.

Cleveland y McGill fueron pioneros en el desarrollo de una teoría sobre la percepción gráfica, basada en la observación de que en ciertos análisis gráficos se emplearon elementos que ayudaban a mejorar la precisión y los resultados obtenidos. Dedujeron que los juicios sobre los ángulos en el caso de los diagramas de sectores no eran tan precisos como los juicios de longitud de los gráficos de barras y por tanto, que los gráficos circulares no eran tan eficientes como los de barras a la hora de estimar o comparar proporciones.

Posteriormente en 1987, Simkin y Hastie indicaron que en los ejercicios usados por Cleveland y McGill, no hubo una justa comparativa entre ambos métodos. En el caso de los segmentos alineados, los sujetos tenían que decidir qué proporción tenía una línea respecto a otra y lo mismo sucedía con la estimación de los ángulos. También concluyeron que el procedimiento utilizado era coherente con la forma en la que la gente calculaba la proporción en un gráfico de barras, aunque no sucedía lo mismo en el caso de los diagramas circulares porque no reflejaban como se estimaba la proporción, ya que se comparaba el ángulo del segmento frente a los 360 grados del círculo.

Figura 9. Gráficas empleadas en los estudios de Spence y Lewandowsky





A	B	C	D
10	25	7	26

En 1990, Spence, en diversas pruebas de experimentos psicofísicos demostró que las proporciones juzgadas a través de los diagramas de sectores eran más acertadas que las de otras formas básicas empleadas. Sus resultados les condujeron a equiparar el desempeño de las gráficas de barras y circulares al realizar estimaciones simples de proporciones. Sin embargo, un año después Spence y Lewandowsky demostraron que el diagrama sectorial era superior a las tablas en lo que a comparaciones rápidas se refería y que además, superaba al de barras en precisión cuanto más complejas fueran las estimaciones a realizar. En adicción, también demostraron que el desempeño de las gráficas circulares no era inferior al de las barras en las comparaciones de proporciones. De hecho, si se trataba de una comparación compleja entre proporciones compuestas (A+B y C+D), la eficiencia del diagrama circular superaba al de barras.

El estudio fue titulado "Mostrando Proporciones y Porcentajes" y fue publicado en la revista *Applied Cognitive Psychology* (volumen 5, páginas 61-77). La pregunta que los sujetos tenían que responder era cuál de las dos secciones (A+B o C+D) era mayor (en base a los siguientes gráficos de la derecha).

Spence, en sus posteriores trabajos con Holland (1992, 1998 y 2001), probó que el gráfico circular funcionaba igual de bien que otros gráficos de uso común a la hora de mostrar el tamaño relativo de un pequeño número de proporciones.

4 Crítica al diagrama de sectores

4.1 Objeciones ante el uso de los diagramas de sectores en las publicaciones académicas.

Uno de los puntos fuertes de esta gráfica es que el mensaje es dividido en partes que conforman el total de la muestra a analizar. Sin embargo, aunque a priori dichas partes puedan visualmente resultar obvias, a la hora de decodificar su significado en términos cuantitativos podríamos tener algún inconveniente si no se trataran de ratios (un cuarto, un tercio, tres cuartos, etc.).

Para facilitar la descodificación del mensaje en términos cuantitativos, es preferible el diagrama de barras ya que este nos ofrece la posibilidad de ver las diferencias (gracias a la escala cuantitativa) entre las partes que conforman el total de una forma mucho más sencilla y rápida que un diagrama de sectores, salvo cuando en estos últimos las magnitudes de las partes son cercanas al 0%, 25%, 50%, 75% y 100%. En cualquier otro caso, el diagrama de sectores nos dificultaría la comprensión del tamaño de los porcentajes de cada una de las partes. Incluso aunque haya secciones que equivalgan al 25% o 75%, si dichas secciones no están orientadas respecto a los ejes tanto de las ordenadas como de las abscisas y se encuentran inclinadas, será mucho más complicado identificar incluso los valores visualmente más evidentes.

He aquí algunos ejemplos mostrando lo mencionado anteriormente:

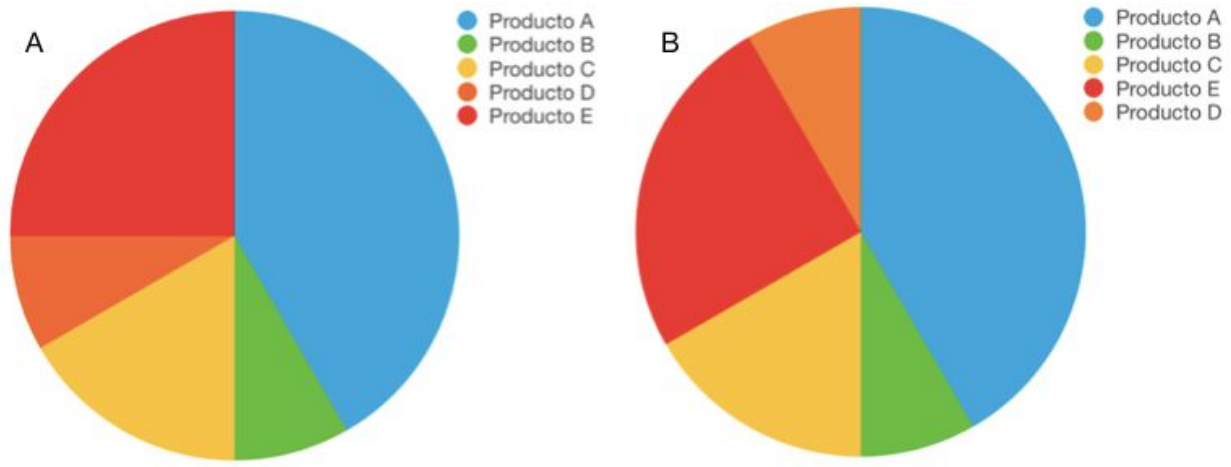


Figura 10. Comparativa entre dos gráficas sectoriales con la misma distribución

Tomando como referencia el producto E que equivale al 25%, la orientación de su sección en la gráfica A nos permite identificar su tamaño rápidamente. No obstante, no sucede lo mismo en la gráfica B, en la cual resulta complicado visualizar el cuarto del producto E con exactitud. Por tanto, podemos afirmar que los gráficos circulares no permiten transmitir cantidades exactas de manera visual (al menos sin ningún etiquetado).

Una posible forma de solucionar esto podría ser mediante la adicción de los porcentajes a cada sección e incluso podemos añadirles sus respectivos nombres para no tener que estar mirando la leyenda lateral constantemente para identificar cada producto. El problema que esto conlleva es que estamos aportando la misma información que la que nos podría proporcionar una simple tabla, pero de manera caótica, desordenada y poco eficiente dado que tendremos que ir fijando nuestra vista en cada sección del diagrama y después proceder a identificarlo en la leyenda.

Productos	Porcentajes
Producto A	42 %
Producto B	8 %
Producto C	17 %
Producto D	8 %
Producto E	25 %

Tabla 1. Distribución de los porcentajes de cada producto.

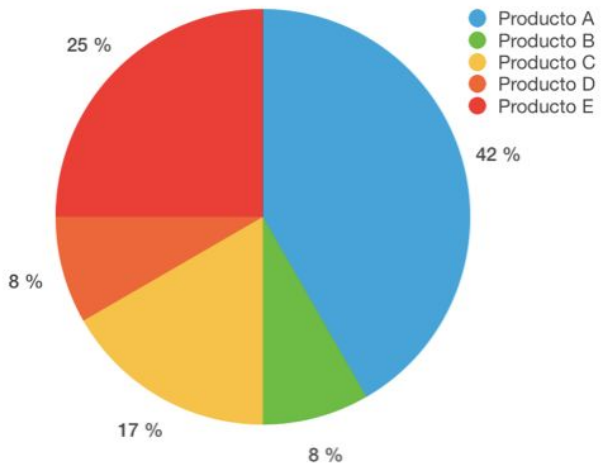
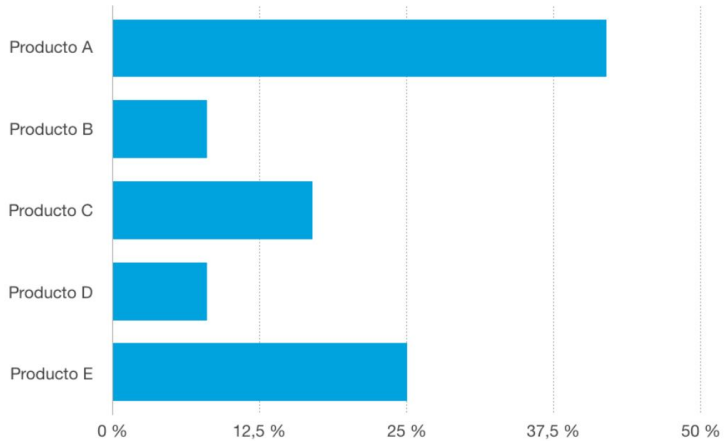


Figura 11. Representación gráfica de la tabla mediante un diagrama circular con los porcentajes de cada sección.

Como acabamos de ver, es mucho más eficiente emplear las tablas en grupos pequeños de datos en vez de los diagramas de sectores habituales (Tufte, p. 178). En los entornos académicos se opina que no se deberían emplear dichos diagramas a la hora de exponer información numérica (Cleveland, *The Elements of Graphing Data*, 1994). Según este criterio, estos gráficos sólo deberían emplearse para ayudar a discernir y comprender las relaciones tales como patrones, tendencias y excepciones que a priori no serían evidentes si la información estuviera expuesta en una tabla.

Figura 12. Gráfico de barras de los datos expuestos anteriormente



Es posible también, exponer toda la información de la tabla anterior de tal manera que podamos comparar y visualizar los valores entre sí mediante un gráfico de barras evitando las complicaciones de un gráfico de pastel.

A diferencia del diagrama de sectores, el diagrama de barras nos ofrece toda la información de manera directa y propensa a ser comparada. Probablemente no sea tan eficaz como la tabla pero a diferencia de esta, permite al lector visualizar las diferencias gráficamente, hecho que conlleva a una mejor comprensión de los datos (especialmente si se trata de datos cuantitativos).

Los diagramas de sectores codifican los valores cuantitativos según dos factores: el área bidimensional de cada sección y los ángulos formados por estos que nacen del centro de la gráfica. Es evidente (según los diversos estudios realizados como los de Kosara y Skau entre otros) que ninguno de esos dos atributos resulta fácil de comparar. No ocurre lo mismo cuando comparamos las áreas bidimensionales de los gráficos de barras en las cuales las únicas diferencias radican en la longitud de la línea de cada una.

El ojo humano no puede calcular con exactitud la diferencia de volumen que se halla entre una circunferencia con un radio de 10 centímetros a otra de 25 centímetros, como tampoco puede calcular con precisión las áreas de los sectores que conforman el gráfico circular, algo esencial a la hora de analizar las evoluciones de los valores o realizar comparativas entre ellos, especialmente cuando las secciones varían sutilmente.

Y algo parecido ocurre cuando se equiparan secciones del diagrama con ángulos similares, ya que si dependemos únicamente de la vista, nos resultará complicado decidirnos por el sector más grande.

En ambos casos (tanto con el área o el ángulo), cuanto más pequeños sean mayor dificultad tendrá el observador a la hora de analizarlos. Por ello, cuantas más secciones posea un diagrama circular, más ineficiente y complejo se volverá, siendo descartado este método por otros tipos de análisis gráficos (diagramas de barra o puntos).

Según Naomi Robbins:

"*Para nuestros fines, un gráfico es considerado más efectivo que otro si su información cuantitativa puede ser decodificada con mayor rapidez o con mayor facilidad por la mayoría de los observadores*" (Naomi Robbins*, Communicating Data Clearly*, Strata Conference, 2011, p. 3).

En este mismo trabajo, Naomi es muy crítica respecto al diagrama de sectores. Entre los mayores defectos de la gráfica destaca la dificultad a la hora de ordenar por tamaños secciones con ángulos similares y defiende el uso de gráficas que nos faciliten ver estas diferencias (como las gráficas de puntos o de barras).

Un gran detractor de los diagramas de sectores (como ya he mencionado con anterioridad en la comparativa entre el gráfico circular y la tabla) sería Edward Tufte, el cual estableció que prácticamente en la gran mayoría de los casos, debería sustituirse el uso de las gráficas de sectores por tablas (especialmente en grupos de datos reducidos). También declaró que el único peor diseño que el diagrama de sectores sería una combinación de varios de ellos, por tener que exigir al observador equiparar cuantías dentro y entre cada sector, provocando así una distorsión de las proporciones.

Cleveland y Becker (1996, p. 50) a través de sus múltiples experimentos de percepción gráfica, comentaron que los gráficos de pasteles tenían serios problemas perceptuales en comparación con los diagramas de puntos y por ello pecaban de poco fiables. Sin embargo, concluyeron que si el verdadero objetivo del emisor era mostrar la información, sin darle mucha importancia a que los lectores la comprendieran, el uso de estos gráficos sería el más adecuado.

4.2 Errores de juicio en las variaciones de los diagramas sectoriales

El artículo de Robert Kosara y Drew Skau publicado en 2016, analizaba cuatro de las más comunes variaciones del gráfico de sectores y calculaba el índice estadístico que conllevaba a una lectura errónea. Para ello analizaron varias versiones de los diagramas de sectores entre las cuales se encontraba el gráfico circular con separación entre los sectores (exploded pie chart), el gráfico con un sector ampliado, un gráfico elíptico y otro cuadrado.

Para llegar a la obtención de las tasas de errores logarítmicas, Kosara y Skau emplearon la siguiente fórmula que idearon en uno de sus trabajos previos, la cual servía para calcular el error logarítmico absoluto:

Figura 13. Gráficas sectoriales empleadas por Kosara y Skau

$$log_2(|\text{judgedvalue} - \text{truevalue}| + \tfrac{1}{8})$$

En la siguiente tabla se muestran las tasas de errores logarítmicas obtenidas en el estudio. El error varía en función del ángulo mostrado por cada diagrama de sector. Como podemos observar, los dos gráficos con mayor tasa de error son aquellos con la forma distorsionada (el elíptico y el cuadrado).

Chart Variation	Mean Log Error
Baseline Pie	1.151
Exploded	1.236
Larger Slice	1.338
Square	1.487
Ellipse	1.570

Tabla 2. Tasas de error logarítmico por gráfica

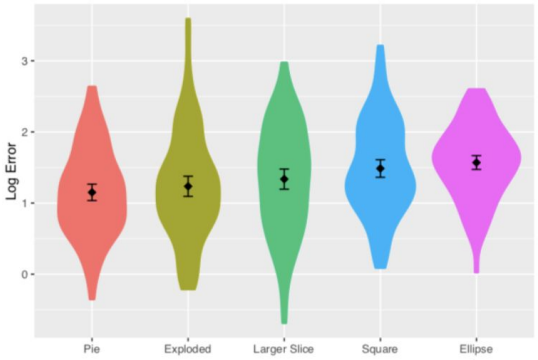


Figura 14. Distribuciones del error entre las distintas gráficas

Los errores difieren por el ángulo mostrado derivando esto en una sobrestimación de los valores pequeños y una subestimación de los más grandes. En general, todos los diagramas sectoriales que contaban con un sector ampliado, implicaban una subestimación de los datos. Los gráficos con los sectores separados mostraban una tasa de error mayor debido a la distracción que suponía el hueco entre ambos pedazos. Por último, como bien he mencionado en el párrafo anterior, el gráfico elíptico y el cuadrado indicaban las mayores tasas de errores de toda la muestra, siendo el primero el que mas error acumulaba a diferencia de las predicciones realizadas.

Entre las principales conclusiones obtenidas por el estudio destacaba el hecho de poner en duda una posible lectura del gráfico circular por el ángulo central, dado que si este fuera el caso, el error obtenido no tendría que haber diferido tanto entre los diagramas normales, los de un sector ampliado y los que contaban con una separación entre las secciones. Además, aclararon que los cambios en la forma y contorno de los gráficos inducían a una considerable distorsión, aconsejando sustituirlos por el diagrama de sectores estándar o el gráfico de anillos. A pesar de los datos obtenidos, Kosara y Skau opinaron que al existir tantas formas diversas de generar un error en la interpretación de estas gráficas (como desplazando un trozo del centro, entre otra muchas alternativas), era necesario un estudio mucho más minucioso abordando estas cuestiones.

5 Software estadístico R

5.1 Introducción

R es un entorno de programación libre que permite realizar análisis estadísticos de datos y representaciones de los mismos empleando el lenguaje S de GNU (sistema operativo de tipo Unix). Al tratarse de un software gratuito y de código abierto, R permite ampliar sus funcionalidades mediante extensiones, librerías o modificaciones propias hechas por el usuario.

El desarrollo de R fue llevado a cabo por Robert Gentleman y Ross Ihaka en 1993, que formaban parte del Departamento de Estadística de la Universidad de Auckland. En la actualidad el software R está a cargo del R Development Core Team.

R posee también su propio lenguaje de programación orientado a objetos y su enorme y flexible conjunto de módulos lo convierten en uno de los mejores programas de estadística computacional.

5.2 Instalación

Al tratarse de un programa de código libre, se puede acceder a su descarga desde https://cran.r-project.org. Una vez en que estemos en la página de Cran, seleccionaremos cualquier repositorio y elegiremos nuestra versión dependiendo de nuestro sistema operativo.

6 Variantes de diagramas de sectores en R

6.1 Diagrama de sectores, sectores con separación y de tres dimensiones

Diagrama circular en el cual se muestran las áreas de las secciones de manera que compongan una representación proporcional de un todo. El segundo tipo de diagrama sería exactamente igual al primero pero añadiendo cierto margen entre las distintas secciones. El tercero haría referencia a un gráfico circular con altura, anchura y profundidad.

6.2 Diagrama de anillos

Es un diagrama de sectores con un agujero circular en el núcleo, el cual puede contener desde un texto hasta una imagen (que conforma el todo al que hacen referencia las secciones), en función de lo que el autor crea conveniente. Además, puede contener varias estadísticas a la vez y proporciona una mejor relación de intensidad respecto a los datos.

6.3 Diagrama de rayos de sol o de sectores multinivel

Este gráfico multinivel está conformado por un conjunto de círculos concéntricos empleados para mostrar las relaciones jerárquicas. El tamaño de cada nivel representa su contribución a la categoría interior que le precede. El gráfico comienza desde un único valor denominado en nodo raíz y va creando una relación jerárquica con los segmento del círculo externo. El diagrama de rayos de sol puede representarse también sin el nodo raíz con un agujero en el núcleo, formando un diagrama de anillos multinivel.

6.4 Diagrama de área polar o rosa de Nightingale

Aunque se trata de un diagrama de circular, sus sectores no están distribuidos de manera proporcional como lo estarían en un diagrama de sectores. En vez de ello, cada sector dispone del mismo ángulo, pero difiere en la distancia en la que se aleja del centro. El número de secciones dependerá de la muestra a analizar (en caso de ser una muestra de un año dividida mensualmente sería 12), y estas estarán subdivididas en diferentes áreas, las cuales serán proporcionales entre sí. El radio de cada sector será proporcional a la raíz cuadrada de la variable a analizar (en el conocido ejemplo de Florence Nightingale la variable que empleó fue la tasa de muertes mensuales por la Guerra de Crimea).

Aunque se relaciona a Nightingale como la pionera de esta gráfica, lo cierto es que el primer uso de este diagrama fue atribuido a André Michael Guerry en 1829. Leon Lalannelater lo empleó en 1843 para mostrar la frecuencia de las direcciones del viento y pero no fue hasta 1858 cuando ganó renombre a través de la publicación de Nightingale de su diagrama de rosa.

6.5 Spie chart

Se trata de una variante de diagrama de área polar (ideada por Dror Feitelson) que nos ofrece dos variables distintas en un mismo gráfico (una cualitativa y la otra cuantitativa). A diferencia de un gráfico circular habitual, este tiene superpuesto otro gráfico en forma de radar, es por ello que el área de los sectores no es proporcional a las variables cualitativas y

el radio de cada sector varía en función de la variable cuantitativa del radar, mientras que el ángulo del diagrama permanece invariable. Esta modificación de las proporciones puede llegar a dificultar la correcta comprensión de los datos por parte del lector.

6.6 Diagrama cuadrado o gráfico de gofre

El gráfico de gofre se emplea (al igual que un diagrama de sectores normal) para representar las partes que componen un todo de cantidades categóricas. Para emular los porcentajes del gráfico circular, se suele emplear una red cuadriculada de 10x10 cuadrados representando cada uno un 1% del total, aunque no es una norma escrita y por tanto es común ver redes cuadriculadas de forma rectangular. Al igual que sucede con los gráficos de tarta, en estos es preferible mantener un número bajo de categorías.

7 Paquetes y extensiones de R que emplearemos

7.1 Graphics

Paquete fundamental de R que viene instalado por defecto y que contiene las funciones de las gráficas más básicas.

7.2 Datasets

Se trata de otro paquete que también viene instalado de serie en R y que contiene grupos de datos. Más adelante explicaré cómo trabajar con un marco de datos empleando un grupo de datos que trae el software como muestra.

7.3 GRdevices

Paquete inicial de R que contiene funciones de apoyo tanto para los gráficos base como para las gráficas cuadriculadas.

7.4 Plotrix

Extensión que nos provee de un gran número de gráficas en dos y tres dimensiones, varios tipos de etiquetados, varios tipos de ejes y funciones de escalado de colores.

7.5 Vcd

Las siglas de la extensión hacen referencia a "visualización de datos categóricos". Contiene técnicas de visualización, grupos de datos, resúmenes e inferencias de procedimientos enfocados principalmente en datos categóricos como su nombre indica. No viene instalada en R así que tendremos que añadirla nosotros.

7.6 Ggplot2

En 2005, Hadley Wickham creó el paquete ggplot2 para R, con la intención de mejorar la graficación por defecto y la extensión "Lattice" de R, para así poder establecer un modelo más eficiente y efectivo de graficación.

En él, elaboró un potente lenguaje gráfico para diseñar complejos y sutiles diagramas. Con el auge de la data Science a lo largo de los años, esta extensión ha ido ganando popularidad entre la comunidad de usuarios de R, llegando al punto de convertirse en uno de los paquetes más usados siendo indispensable a la hora de representar gráficos estadísticos.

Wickham se inspiró en las ideas de "Gramática de Gráficos" de Leland Wilkinson a la hora de crear ggplot2, es por ello que al contrario que la mayoría de paquetes de R, ggplot2 posee un lenguaje de programación propio (basado en dicho trabajo de Wilkinson). Este paquete nos permite crear gráficos que representan datos numéricos y categóricos tanto univariados como multivariados de manera directa. De esta forma, ggplot2 puede crear diagramas muy flexibles y específicos para cualquier caso.

Ggplot2 nos permite comenzar con una capa de datos base a la cual podremos ir añadiendo más capas en función de nuestras necesidades estadísticas. El lenguaje de ggplot2 nos permite elaborar gráficos mucho más exigentes que los que nos ofrece R en un principio, pudiendo modificar y combinar elementos más complejos en otros grupos de datos o gráficas. Procederemos a instalar el paquete ggplot2 dado que no viene instalado por defecto en R (lo explicaré más adelante).

7.7 Rgl

Este paquete ofrece a R funciones para gráficos en tres dimensiones interjectivos, entre las cuales se incluirían las funciones modeladas a partir de gráficos básicos como funciones para construir representaciones geométricas de objetos. El resultado gráfico será mostrado en pantalla a través de OpenGL, que se trata de una interfaz de programación de aplicaciones comúnmente usada para la computación gráfica en dos y tres dimensiones, que también viene incluida en la extensión y es compatible con los formatos estándares de los archivos.

7.8 Dplyr

Paquete que nos proporciona una herramienta imprescindible para trabajar con marcos de datos como objetos, tanto en la memoria como fuera de ella.

7.9 Gplots

Provee a R de varias herramientas para representar los datos gráficamente que pueden variar desde funciones para manipular colores, diagramas de Venn, espaciado de puntos, gráfico de texto, gráfico de globo, etc.

7.10 RColorBrewer

Extensión que permite a R el empleo de multitud de paletas de colores.

8 Ejemplos de diagramas de sectores básicos en R

En vez de dedicarle una sección al lenguaje de programación de R, he optado por ir viendo progresivamente el código a través de los ejemplos desde lo más sencillo a lo más complejo.

Comenzaremos abriendo y creando un nuevo proyecto en la aplicación. R viene instalado con una gran variedad de paquetes de análisis gráfico por defecto, entre los cuales se incluye la función *pie()*. Antes de empezar a explicar el proceso a realizar empleando esta función, me centraré en la instalación de paquetes en primer lugar.

Para realizar la instalación de los paquetes de R que vayamos a necesitar y no nos hayan venido instalados, iremos a "Packages & Data" de la barra de opciones superior y en el menú desplegable seleccionaremos "Package Installer".

Figura 15. Listado de los repositorios de R

Se nos abrirá el instalador de paquetes de R, y en la barra desplegable del recuadro "Packages Repository" elegiremos la opción "Other Repository". La primera vez que lo hagamos, R nos mostrará una gran cantidad de repositorios distribuidos por todo el mundo. En mi caso, elegiré el repositorio de Bristol (Reino Unido) dado su gran repertorio y utilidad, además de ser uno de los repositorios más recomendados en los foros estadísticos de R. El software nos preguntará si queremos que el repositorio seleccionado se quede guardado de cara a futuras sesiones así que pulsaremos el botón afirmativo. A continuación, marcaremos la casilla de "Binary Format Packages" y en el recuadro de ubicación de instalación, la opción "At System Level (In R Framework)". Finalmente tendremos que hacer click en "Get List".

El instalador nos mostrará todos los paquetes del repositorio con sus correspondientes versiones disponibles y también nos indicará que versiones de cada paquete tenemos instaladas actualmente en R (siempre y cuando las tengamos descargadas). En nuestro caso seleccionamos el paquete "graphics" que viene a ser el paquete de gráficas estándar que nos proporciona R y pulsamos "Install Selected" para descargarlo. En caso de necesitar una gráfica o una función más compleja o poco habitual de R, tendremos que buscar el paquete en cuestión en el repositorio de nuestra preferencia y seguir los mismos pasos que he mostrado en este apartado.

También, debo aclarar que esta no es la única forma de instalación de paquetes en R. Otra forma bastante común de hacerlo sería ejecutando el siguiente script (lo veremos en los ejemplos posteriores):

```
> install.packages("ggplot2") #instalar ggplot2
> library(ggplot2) #cargar el paquete en la memoria
```

8.1 Creación del pie chart básico mediante vectores

Comenzaremos con el diagrama de sectores básico. En este primer ejemplo crearé dos objetos (concretamente un vector) los cuales serán mostrados por el pie chart. En este ejemplo, emplearé información ficticia sobre los ingresos anuales del 2018 de las cinco empresas tecnológicas más grandes de EEUU (Google, Amazon, Facebook, Microsoft, Apple) para crear el vector.

Antes que nada, para definir un vector es necesario crear una clase con la función *c()*, en la cual introducimos los datos para posteriormente crear el vector. Cuando elaboramos un vector hay que introducir los datos de una clase en él, y para ello se emplea el signo "<-", que básicamente crearía una copia de la clase en el nuevo objeto creado. En este caso, he definido como "gafam_incomes" al vector que contiene la información sobre la clase que posee los ingresos de estas empresas en miles de millones de euros (los datos son ficticios).

```
> gafam_incomes <- c(143,194,108,119,137)
```

A continuación, crearé otro vector con los nombres de las empresas llamado "gafam_enterprises", el cual emplearé posteriormente para identificar cada sección del gráfico. El procedimiento en este caso es idéntico al anterior salvo que en vez de emplear valores numéricos en la clase, usaré los denominados "strings" o cadenas de caracteres.

```
> gafam_enterprises <- c("Google", "Amazon", "Facebook", "Apple", "Microsoft")
```

Una vez creados los dos vectores, emplearemos la función *pie()* para crear el diagrama de sectores que muestre toda la información que hemos introducido. Para crear la gráfica tendremos que especificar una serie de atributos dentro de la función. Aunque la función *pie()* tiene una gran variedad de atributos (los cuales abordaré en detalle más adelante), en este supuesto solo usaremos tres: el vector con datos numéricos que nos proporciona la información, el otro vector con los nombres de las empresas que nos ayudará a etiquetar los segmentos y por último el título de la gráfica.

```
> pie(gafam_incomes, label=gafam_enterprises, main="Ingresos de las 5 empresas tecnológicas estadounidenses más importantes")
```

Como podemos observar, para la fuente de datos he empleado el vector "gafam_incomes", para identificar cada sector he empleado el otro vector "gafam_enterprises" y por último, he titulado el diagrama como "Ingresos de las 5 empresas tecnológicas estadounidenses más importantes".

En la siguiente imagen se muestra el gráfico de sectores que acabamos de crear. Este sería el diseño más sencillo de diagramas de sectores que ofrece R.

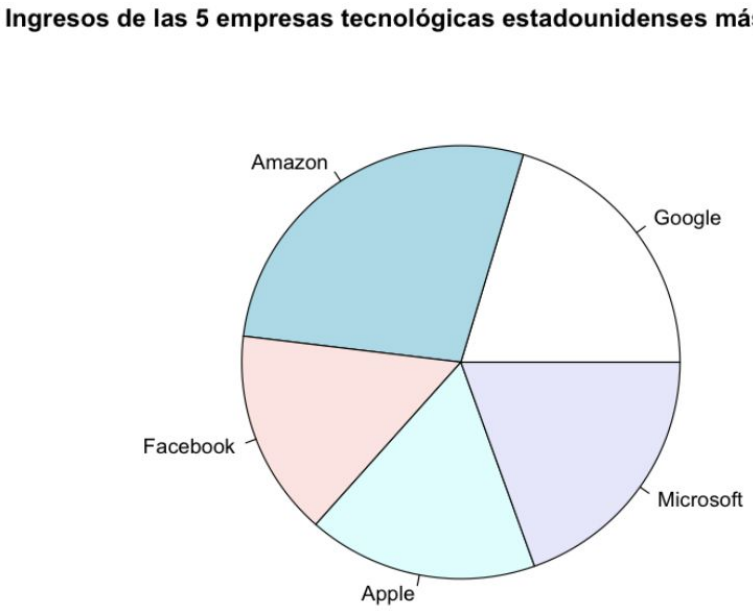


Figura 16. Pie chart básico en R

8.2 Creación del pie chart básico mediante (data frames) tablas enlazadas

En el anterior ejemplo, hemos introducido los datos manualmente mediante el uso de los vectores. Sin embargo, ahora enlazaré una tabla de datos externa al gráfico de pasteles, sin que haya necesidad de introducir datos. También, le añadiré un filtro a la tabla para que podamos operar con mayor facilidad. Estos dos métodos, suelen ser procedimientos muy habituales al trabajar con grandes bases de datos ya que nos ahorraremos tener que introducir los datos de uno en uno o tener que andar buscando entre un gran cúmulo de datos.

> CO2

	Plant	Type	Treatment	conc	uptake
1	Qn1	Quebec	nonchilled	95	16.0
2	Qn1	Quebec	nonchilled	175	30.4
3	Qn1	Quebec	nonchilled	250	34.8
4	Qn1	Quebec	nonchilled	350	37.2
5	Qn1	Quebec	nonchilled	500	35.3
6	Qn1	Quebec	nonchilled	675	39.2
7	Qn1	Quebec	nonchilled	1000	39.7
8	Qn2	Quebec	nonchilled	95	13.6
9	Qn2	Quebec	nonchilled	175	27.3
10	Qn2	Quebec	nonchilled	250	37.1
11	Qn2	Quebec	nonchilled	350	41.8
12	Qn2	Quebec	nonchilled	500	40.6
13	Qn2	Quebec	nonchilled	675	41.4
14	Qn2	Quebec	nonchilled	1000	44.3
15	Qn3	Quebec	nonchilled	95	16.2
16	Qn3	Quebec	nonchilled	175	32.4
17	Qn3	Quebec	nonchilled	250	40.3
18	Qn3	Quebec	nonchilled	350	42.1
19	Qn3	Quebec	nonchilled	500	42.9
20	Qn3	Quebec	nonchilled	675	43.9
21	Qn3	Quebec	nonchilled	1000	45.5
22	Qc1	Quebec	chilled	95	14.2
23	Qc1	Quebec	chilled	175	24.1
24	Qc1	Quebec	chilled	250	30.3
25	Qc1	Quebec	chilled	350	34.6
26	Qc1	Quebec	chilled	500	32.5
27	Qc1	Quebec	chilled	675	35.4
28	Qc1	Quebec	chilled	1000	38.7
29	Qc2	Quebec	chilled	95	9.3
30	Qc2	Quebec	chilled	175	27.3
31	Qc2	Quebec	chilled	250	35.0
32	Qc2	Quebec	chilled	350	38.8
33	Qc2	Quebec	chilled	500	38.6
34	Qc2	Quebec	chilled	675	37.5
35	Qc2	Quebec	chilled	1000	42.4
36	Qc3	Quebec	chilled	95	15.1
37	Qc3	Quebec	chilled	175	21.0
38	Qc3	Quebec	chilled	250	38.1
39	Qc3	Quebec	chilled	350	34.0
40	Qc3	Quebec	chilled	500	38.9
41	Qc3	Quebec	chilled	675	39.6
42	Qc3	Quebec	chilled	1000	41.4
43	Mn1	Mississippi	nonchilled	95	10.6
44	Mn1	Mississippi	nonchilled	175	19.2
45	Mn1	Mississippi	nonchilled	250	26.2
46	Mn1	Mississippi	nonchilled	350	30.0
47	Mn1	Mississippi	nonchilled	500	30.9
48	Mn1	Mississippi	nonchilled	675	32.4
49	Mn1	Mississippi	nonchilled	1000	35.5

Tabla 3. Data frame CO2

Para este ejemplo usaré una tabla predeterminada de R llamada "CO2". Dicha tabla hace referencia a un experimento sobre la tolerancia del frío de la especie de hierba *echinochloa crus-galli* realizado en dos zonas, Mississippi y Quebec. Lo que realizaremos en este caso será un diagrama de sectores que nos filtre por ambas zonas.

En este caso en concreto, dado que la única información que necesitamos es la de la zona, filtraremos la tabla "CO2" por la columna "Type". Para ello, escribiremos el nombre del archivo de datos (CO2) seguido de unos corchetes, dentro de los cuales el primer espacio lo dejaríamos en blanco dado que hace referencia a las filas, y en el segundo emplearemos la función c() para que nos filtre únicamente los valores del vector (columna) "Type". Hay varias formas de hacer esto pero las dos más comunes serían indicar el número de la

columna o indicar su nombre entre comillas, ya que en R al igual que en muchos lenguajes de programación, las cadenas de caracteres se indican mediante comillas.

```
CO2[, c("Type")]
```

Realizado este paso, convertiremos nuestros nuevos datos en una nueva tabla a la cual le asignaremos un nuevo nombre, para tener que evitar hacer el filtrado constantemente y así tener los datos más simplificados. En nuestro caso, he llamado a la nueva variable como "MyCO2".

```
MyCO2 <- table(CO2[, c("Type")])
```

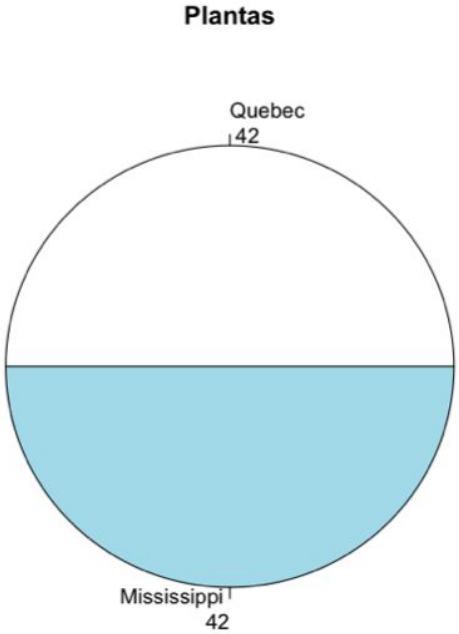
Después, para poder introducir en el atributo de "label" del diagrama de sectores los nombres de las zonas de la columna "Type" de la tabla CO2, tenemos que crear un nuevo objeto que almacene dichos valores. Para lograr esto, empleamos la función *names* (la cual devuelve los valores de las cadenas de caracteres) dentro de la función *paste*, que nos servirá para pegar dichos strings en nuestro nuevo objeto al que llamaremos "labels_myCO2".

```
labels_myCO2 <- paste(names(MyCO2), "\n", MyCO2, sep=" ")
```

Finalmente, solo queda insertar los argumentos correspondientes el la función pie. En primer lugar el vector "MyCO2", seguido de las etiquetas de "labels_myCO2" y el título del gráfico (Plantas).

```
pie(MyCO2, label=labels_myCO2, main="Plantas")
```

Figura 17. Gráfico sectorial de la tabla *CO2*



8.3 Análisis de los argumentos de la función *pie()*

En los apartados anteriores, sólo hemos hecho hincapié en los tipos de fuentes posibles (vectores y marcos de datos) de los cuales podemos extraer la información para posteriormente plasmarla en las gráficas. Ahora profundizaré en diversos aspectos del pie chart y los argumentos que conforman la función *pie()*, esenciales para crear diagramas más complejos. Aquí tenemos una muestra de algunos de los argumentos más utilizados de la función *pie()*:

pie(x, label=names(x), shadow=FALSE, edges=200, radius=0.8, col=NULL, main=NULL, ...)

Analicemos paso por paso las estructura básica de *pie()*. En primer lugar analizaremos los tres atributos que ya hemos empleado con antelación. El primero de ellos lo compondría un vector (al cual denominaremos "x") de cantidades positivas, representadas como los diversos sectores que conforman el gráfico. El siguiente argumento sería "label", el vector con cadenas de caracteres (strings) que sirven para nombrar a cada área. El tercero sería "main", con el cual establecemos un título para el gráfico.

Aparte de estos tres, podemos diferenciar otros argumentos básicos como "shadow", un vector booleano (verdadero o falso) aplica un efecto de sombra al gráfico; "edges", que establece el número de lados del polígono que conforma la línea exterior circular (su valor por defecto es de 200); "col", argumento del diagrama de sectores que sirve para colorear los mismo mediante un vector y finalmente "radius", que indica el radio del círculo del diagrama y toma valores entre 1 y -1.

8.4 Creación de diagramas de sectores con porcentajes y colores

Comenzamos creando el vector que contendrá las cantidades que conformarán los sectores del gráfico al cual llamaré "gdp" (en este caso haré una comparativa ficticia del PIB entre algunos países europeos). Al otro vector con las strings de cada sector lo nombraré "countries" y le estableceré nombres.

```
> gdp <- c(1200,2000,1800,1300)
> countries <- c("España","Alemania","Francia","Italia")
```

Después, para calcular el porcentaje procederemos a dividir el PIB de un país entre el sumatorio del PIB y lo multiplicaremos por cien y enviaremos el resultado a un nueva variable que denominaremos "pct". Habiendo obtenido este dato, añadimos el reciente objeto "pct" al vector "countries" mediante el comando *paste()*. Repetimos el mismo proceso para añadirle el símbolo del porcentaje a cada string de "countries", pero en este caso, señalaremos un espaciado en el argumento "sep" dentro de *paste().*

```
> pct <- round(gdp/sum(gdp)*100)
> countries <- paste(countries,pct)
> countries <- paste(countries,"%",sep=" ")
```

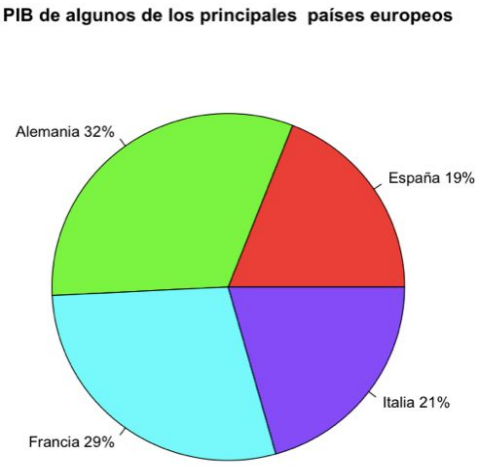


Figura 18. Gráfico circular con porcentajes y tonalidad "rainbow"

En la función *pie()*, rellenaremos los argumentos del vector "X" (gdp) y "labels" como en los apartados anteriores. En este caso añadiremos uno nuevo, el argumento de color (col) del gráfico, en el cual los sectores quedarán diferenciados por multicolores al definirlo como "rainbow", y para que cada área disponga de su propio color, usaremos el comando *length()* dentro de rainbow y así lograremos un color diferente por cada string (país) de "countries".

```
> pie(gdp, labels = countries,
col=rainbow(length(countries)), main="PIB de
algunos de los principales países europeos")
```

8.5 Creación de diagramas de sectores en tres dimensiones con argumentos modificados de la función *pie()*

En primer lugar, necesitaremos el paquete denominado "plotrix", la extensión correspondiente a las gráficas en tres dimensiones. Por ello, procederemos a ir a "Packages & Data" en la barra de herramientas superior de R, y en el menú desplegable seleccionaremos "Package Installer". Una vez en el instalador, buscaremos en el repositorio la extensión "plotrix" y procederemos a instalarla. Una vez realizada la instalación, la primera línea de código indicará la librería a la que llamará R, en este caso "plotrix".

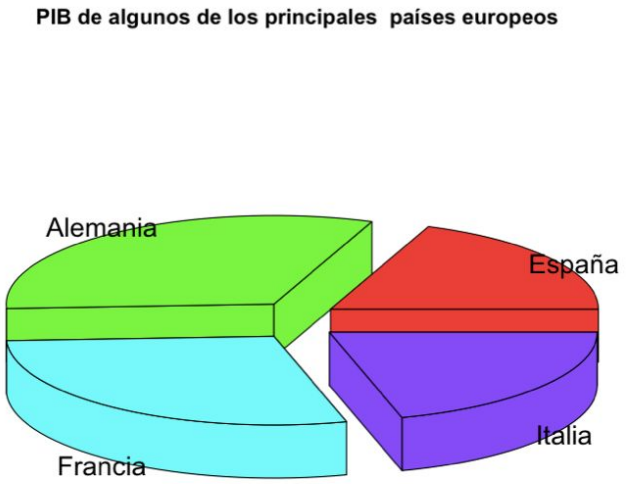


Figura 19. Gráfico circular en 3D con separaciones entre las secciones

```
> library(plotrix)
```

Continuamos con los mismos pasos de los ejercicios previos, por tanto, para este gráfico emplearé los mismos datos que el supuesto anterior. Para representar el diagrama de sectores en tres dimensiones le daremos uso a la función pie3D() e introduciremos en los argumentos "x" y "labels" los vectores correspondientes (*gdp* en el primero con la información numérica y *countries* con las strings). También usaremos el argumento "explode" para separar las secciones los gráficos entre sí.

```
> pie3D(gdp, labels=countries, explode=0.15, main="PIB de algunos de los principales
países europeos",shade=0)
```

9 Creación de diagramas de sectores complejos con la extensión GGPLOT2

9.1 Instalación de la extensión GGPLOT2

En este apartado propondré algunos supuestos y emplearé los distintos diagramas circulares que me ofrece la extensión más completa de R para las representaciones y etiquetado de gráficas, *ggplot2*, que se trata de unas de las herramientas más usadas tanto a nivel profesional como académico.

Para ello procederé a descargar el paquete y lo cargaré en la biblioteca de R junto con la extensión *dplyr* (que dispone de la gramática necesaria para la manipulación de datos) mediante los siguientes scripts:

```
> install.packages("ggplot2", "dplyr")
> library("ggplot2")
> library("dplyr")
```

9.2 Elaboración de un diagrama de sectores a través de un marco de datos y un diagrama de barras

Para este supuesto, he cogido como referencia los distintos tipos de mascotas que poseen los niños de una determinada aula compuesta de 26 alumnos. Nuestro objetivo será crear un diagrama de sectores que refleje el porcentaje de cada especie respecto al total de la cantidad de mascotas que hay en la clase. Una vez recopilada la información sobre cada uno de los alumnos, hemos obtenido la siguiente tabla:

Mascotas	Cantidad
Perros	8
Gatos	6
Pajaros	2
Peces	3
Hamsters	1
Ninguna	10

Tabla 4. Tabla mascotas

Como podemos visualizar, hay algunos alumnos con varias mascotas, es por ello que el sumatorio de los resultados no concuerda con el número de alumnos total.

Lo primero que tendremos que hacer es elaborar un marco de datos dentro de R que represente la información adecuadamente, el cual usaremos posteriormente como puente para elaborar nuestros diagramas.

Para ello crearemos dos vectores, el vector *mascotas* que incluiría las cadenas de caracteres y el vector *cantidad* que estaría compuesto de datos numéricos enteros (el proceso es el mismo que en los ejemplos anteriores).

```
> mascotas = c("Perros", "Gatos", "Pajaros", "Peces", "Hamsters", "Ninguna")
> cantidad = c(8, 6, 2, 3, 1, 10)
```

Acto seguido crearemos un marco de datos con la función data.frame(), en la cual incluiremos tanto el vector *mascotas* como el vector *cantidad*. Le asignaremos a este nuevo data frame el nombre "df1".

```
> df1 <- data.frame(mascotas = c("Perros", "Gatos", "Pajaros", "Peces", "Hamsters",
"Ninguna"), cantidad = c(8, 6, 2, 3, 1, 10)
```

Podemos visualizar el marco de datos *df1* en un formato similar al de una tabla con la función *head()*, que sirve para devolvernos las primeras parte de un vector, una tabla, una matriz, un data frame o una función.

```
> head(df1)
  mascotas cantidad
1   Perros        8
2    Gatos        6
3  Pajaros        2
4    Peces        3
5 Hamsters        1
6  Ninguna       10
```

Tabla 5. Data frame *df1*

Puesto que lo que queramos hacer se trataría una comparativa entre los tipos de mascotas respecto a la cantidad total de ellas, tendremos que redefinir el marco de datos excluyendo de él la información relacionada con aquellos alumnos que no posean mascotas.

```
> head(df2)
  mascotas2 cantidad
1    Perros        8
2     Gatos        6
3   Pajaros        2
4     Peces        3
5  Hamsters        1
```

Tabla 6. Data frame *df2*

En este caso, he optado por crear un nuevo data frame llamado *df2*, reduciendo en uno la longitud de los dos vectores que formaban *df1*. Para mantener la información base de dichos vectores, en vez de sobreescribir las variables, las he redefinido y les he añadido una terminación en "2" para diferenciarlas. Después he ejecutado la función *head()* nuevamente par ver como ha quedado *df2*.

```
> mascotas2 = c("Perros", "Gatos", "Pajaros", "Peces", "Hamsters")
> cantidad2 = c(8, 6, 2, 3, 1)
> df2 <- data.frame(mascotas2 = c("Perros", "Gatos", "Pajaros", "Peces", "Hamsters"),
      cantidad = c(8, 6, 2, 3, 1))
```

Habiendo establecido el marco de datos correctamente, procederemos a crear el diagrama de barras que posteriormente convertiremos en un diagrama circular (ver apéndice 1).

```
> bp1 <- ggplot(df2, aes(x = "", y = cantidad2, fill = mascotas2)) +
      geom_bar(width = 1, stat = "identity")
> bp1
```

Para crear el objeto del diagrama de barras al que llamaremos *bp1*, utilizaremos la función *ggplot(),* que sirve para crea un marco de datos de entrada para un gráfico y especifica el conjunto estético que sea común en las capas posteriores. Dentro de la estructura de la función("*ggplot(data=NULL, mapping = aes(), … , environment = parent.frame())*"), primero introduciremos el conjunto de datos df2, en *aes(),* asignaremos las propiedades estéticas de la gráfica, dado que nuestra intención es crear un diagrama de barras dejaremos la variable "x" en blanco. A la variable "y" le estableceremos los datos del vector *cantidad2*, donde las categorías serán representadas en el eje de ordenadas de manera acumulativa. Enlazaremos en *fill* el vector con los strings de las categorías.

Acto seguido, a *ggplot()* le sumaremos la función *geom_bar()*, que hace que la altura de la barra sea proporcional al número de registros de cada grupo. Rellenamos la anchura con el valor 1, y declaramos "identity" en el atributo *stat*, para indicarle a R que las alturas de las barras representan valores distintos. Ejecutamos el script y volvemos a introducir el objeto *bp1* únicamente para que el programa nos muestre la gráfica.

Ahora procederemos a convertir el diagrama de barras en uno de sectores. Para esto, haremos que en vez de que la disposición de la barra sea vertical, los valores roten respecto a un eje mediante coordenadas polares. Por tanto, definiremos un objeto (*pie_mascotas*) que sea la suma de conjunto de datos *bp1* y la función *coord_polar()*, en la que indicaremos como centro al eje de ordenadas (y) con un punto de inicio igual a 0, y después introduciremos dicho objeto para que R nos muestre el diagrama de sectores.

```
> pie_mascotas <- bp1 + coord_polar("y", start=0)
> pie_mascotas
```

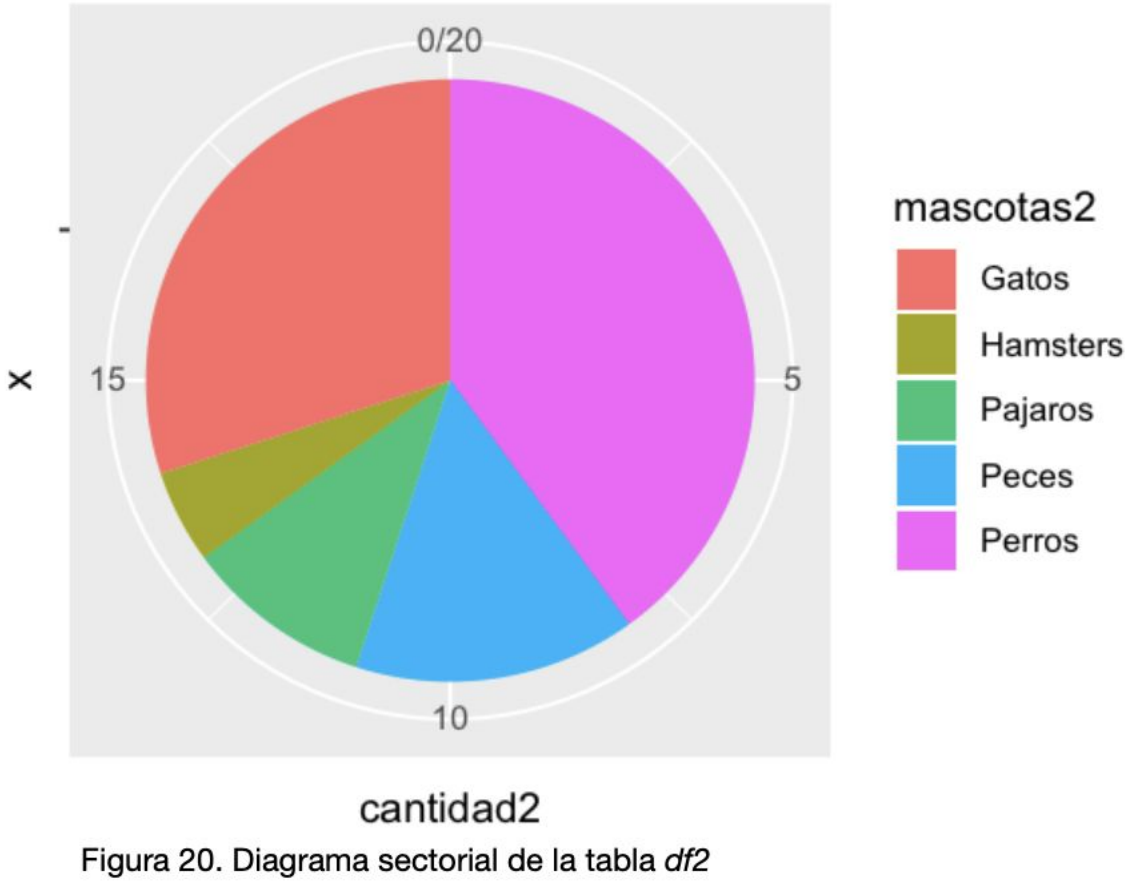


Figura 20. Diagrama sectorial de la tabla *df2*

A pesar de que el etiquetado de ggplot2 nos muestra un eje cuantitativo alrededor del gráfico, nosotros le añadiremos los porcentajes correspondiente a cada porción para no tener que hacer aproximaciones a simple vista.

Para añadir los porcentajes al gráfico la función a utilizar sería *geom_text()*. Dentro de esta, emplearemos la función aestethic (aes) y dentro de ella en el argumento label, usaremos *paste0()* para concatenar las strings de los porcentajes que vamos a generar añadiendo el símbolo "%" sin ningún espaciado. La función round nos redondea el valor en cuestión, que en nuestro caso será el porcentaje de cada sección y para ello introduciremos el valor de la sección entre el valor del sumatorio del vector *cantidad2* y lo multiplicaremos por 100. Por último, en el argumento de posición, la función position_stack() nos servirá para alinear nuestra string verticalmente (vjust) y establecerla en el centro (en este caso establecerá los valores en la mitad del radio del círculo).

```
> pie_mascotas <- pie_mascotas + geom_text(aes(label =
paste0(round((cantidad2/sum(cantidad2))*100, "%")), position = position_stack(vjust =
0.5))
```

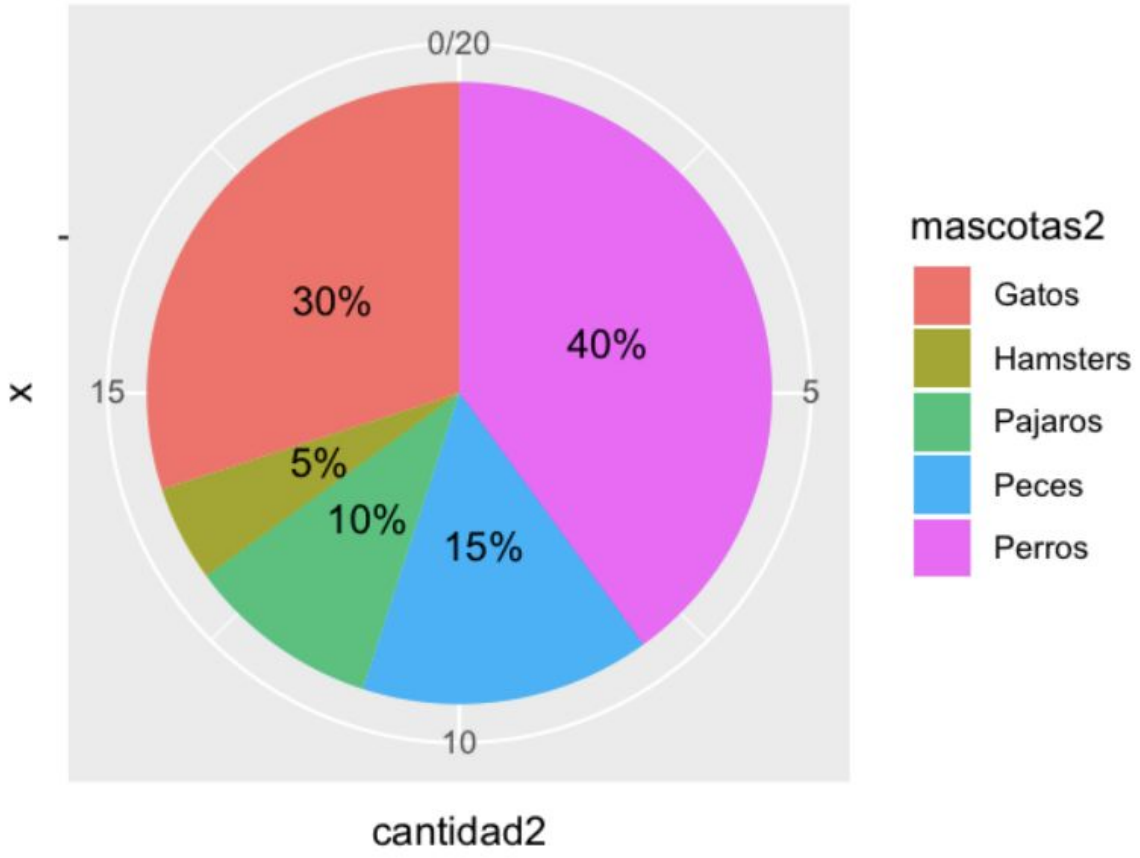


Figura 21. Diagrama sectorial de la tabla *df2* con porcentajes

Esta es una forma de crear un diagrama de sectores de un marco de datos con *ggplot2*, usando como apoyo un gráfico de barras con secciones proporcionales y la función de coordenadas polares. A continuación veremos algunos ejemplos para personalizar el gráfico a nuestro antojo.

Para ello contamos con la función *scale_fill_manual()*, con la cual intercambiaremos el color del gráfico *pie_mascotas* introduciendo el vector "values" y asignando un color para cada categoría de la gráfica mediante strings. Crearemos otro objeto a partir de *pie_mascotas* usando estas modificaciones al que llamaremos *pie_mascotas2*.

```
> pie_mascotas2 <- pie_mascotas + scale_fill_manual(values=c("red", "blue",
"lightblue", "grey", "orange"))
```

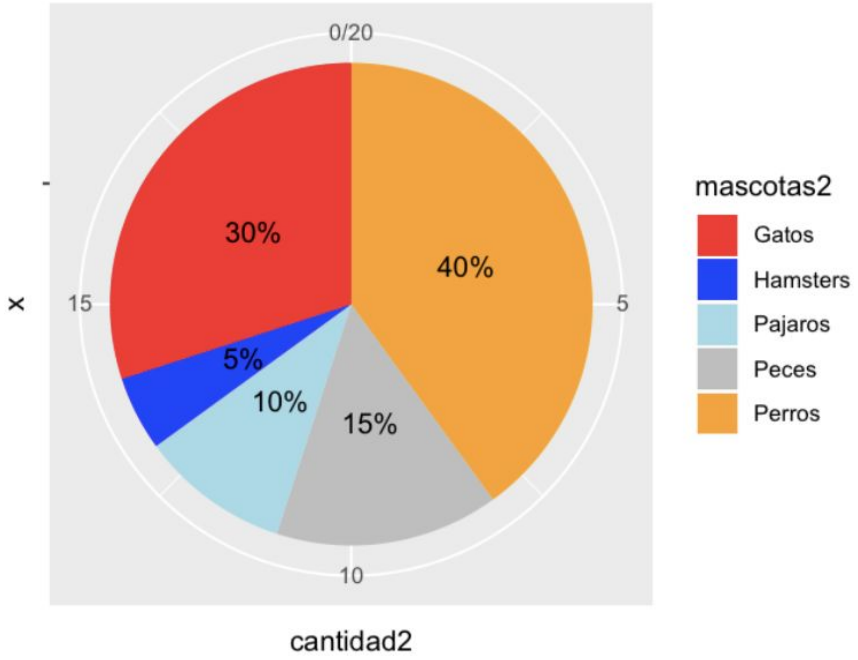


Figura 22. Diagrama sectorial de la tabla *df2* con la función *scale_fill_manual()*

Entre las diversas formas de cambiar los colores de las secciones estarían las funciones *scale_fill_grey()* con la cual diferenciaríamos los sectores por una escala de grises y *scale_fill_brewer()* que nos permitiría seleccionar una de las muchas paletas de colores de las que nos abastece R.

9.3 Comparativa entre dos diagramas de sectores en ggplot2

Mascotas	Cantidad
Perros	10
Gatos	4
Pajaros	4
Peces	2
Hamsters	3
Ninguna	7

Tabla 7. Tabla *mascotas* de la clase B

En este supuesto, equipararemos los resultados de la primera clase (A) con otra sometida al mismo estudio (a la que denominaremos clase B). La tabla de la izquierda recoge la información correspondiente a la clase B.

Nuestro objetivo es que R nos muestre ambos gráficos de sectores a la par para poder realizar una comparación rápida. La única forma de lograr esto sería mediante la creación de un marco de datos conjunto entre ambas clases.

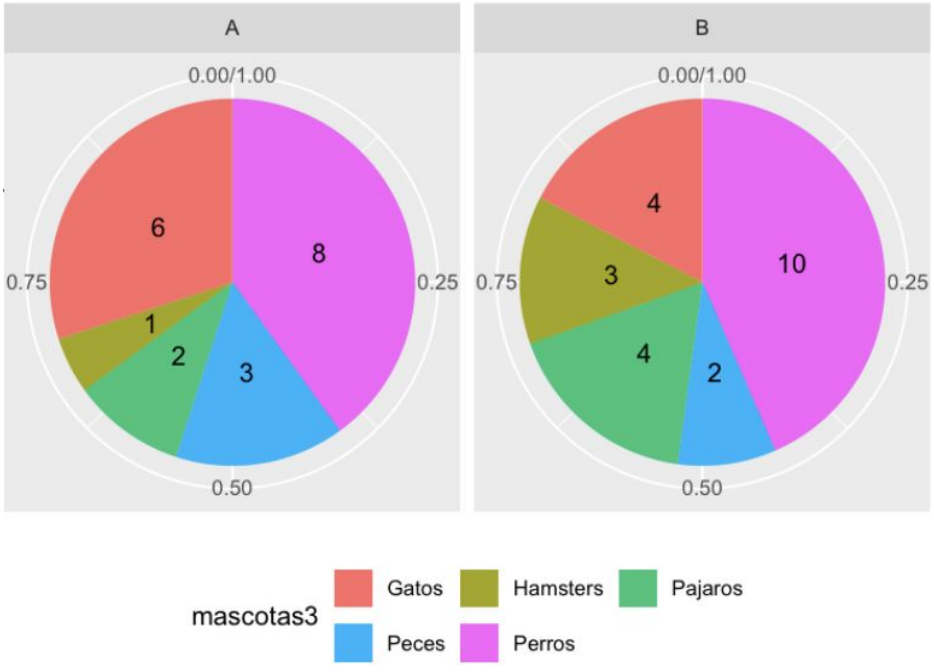
De este modo, crearemos el data frame *df_10*, que dispondrá de tres vectores. El vector "mascotas3" con la información categórica (nótese que se han creado dos grupos idénticos de nombres de las mascotas, uno para cada clase). Los datos cuantitativos de "cantidades3" también están compuestos por otros dos bloques (los 5 primeros datos de la A y los restantes 5 de la B) y se repite el mismo patrón con el vector "clase".

```
> df_10 <- data.frame(mascotas3=c("Perros", "Gatos", "Pajaros", "Peces", "Hamsters",
      "Perros", "Gatos", "Pajaros", "Peces", "Hamsters"),
      cantidades3=c(8,6,2,3,1,10,4,4,2,3), clase=c("A", "A", "A", "A", "A", "B",
      "B", "B", "B", "B"))
```

Continuaremos con la función *ggplot()* dentro del cual introduciremos nuestro reciente marco de datos, estableceremos como variable categórica a "mascotas3" y en el eje Y al vector "cantidades3". A *geom_bar()* le añadiremos la función *position.fill()*, la cual nos apila las barras y las estandariza, manteniendo así una altura constante. La etiquetas de geom_text() irán vinculadas al "cantidades3". Emplearemos *facet_wrap()* con el vector "clase" para envolver un espacio monodimensional en uno bidimesional (agrupa los datos). Por supuesto, la función *coord_polar()* no puede faltar dado estamos representando gráficos de sectores. Finalmente, eliminaremos las etiquetas de los ejes mediante *element_blank()* y situaremos la leyenda del diagrama debajo de las gráficas en dos filas.

```
> ggplot(data = df_10, aes(x = "", y = cantidades3, fill = mascotas3)) +
      geom_bar(stat = "identity", position = position_fill()) +
      geom_text(aes(label = cantidades3), position = position_fill(vjust = 0.5)) +
      coord_polar(theta = "y") + facet_wrap(~ clase) +
      theme(axis.title.x = elemente_blank(), axis.title.y = element_blank()) +
      theme(legend.position ="bottom") + guides(fill=guide_legend(nrow=2,
      byrow=TRUE))
```

Figura 23. Comparativa entre los diagramas sectoriales de las clases A y B



9.4 Elaboración de un diagrama de anillos con ggplot2

Un profesor de una universidad para uno de sus estudios realiza una encuesta a sus estudiantes con el objetivo de analizar el medio de transporte que emplean mayoritariamente para llegar al centro. A la derecha aparecen mostrados los resultados en una simple tabla:

Transporte	
Autobús	15
Motocicleta	13
Coche	9
Metro	3
Bicicleta	8
A pie	6

Tabla 8. Tabla *transporte*

Vamos a proceder a crear un gráfico de sectores que nos muestre las proporciones entre los distintos medios de transporte dado que el proceso para crear el diagrama de anillo en R es muy similar al del diagrama de sectores. La principal diferencia radicaría que en el caso de donut, tendremos que añadir una función para generar un agujero en el centro.

En este supuesto, procederemos a clasificar la tabla en un marco de datos para su posterior aplicación en un gráfico de circular. Elaboraremos un data frame (*df_transporte*) con dos vectores a los que denominaremos "transporte" y "personas" (el primero sería el de las cadenas de caracteres y el segundo el de los datos numéricos).

Continuamos con el proceso de creación habitual de un diagrama de sectores. Con la función *ggplot()* declararemos la función de datos de entrada para gráfica, con *geom_bar()* crearemos el diagrama de barras proporcionales, con *coord_polar()* haremos que la barra pivote sobre el centro del círculo, la función *geom_text()* la emplearemos para mostrar el texto con los porcentajes de cada sección. Cabe destacar el cambio del color de la fuente a blanco mediante el atributo "color", dado que de no especificar, el color por defecto sería el negro y al establecer una sección que también comparta ese mismo color (como en este caso) el texto sería ilegible. Seguimos con la función *scale_fill_manual(),* con la que indicaremos los colores de cada categoría y por último, con *theme_void()* eliminaremos el fondo gris y los ejes laterales que acompañan a la gráfica en *ggplot2*.

```
> df_transporte <- data.frame(transporte=c("Autobus", "Motocicleta", "Coche",
"Metro", "Bicicleta", "A pie"), personas=c(15, 13, 9, 3, 8, 6))

> ggplot(df_transporte, aes(x = "", y = personas, fill = transporte)) +
      geom_bar(width = 1, stat = "identity", color = "white") +
      coord_polar("y", start = 0) +
      geom_text(aes(label = paste0(round(personas/sum(personas))*100, "%")),
      position = position_stack(vjust = 0.5), color = "white") +
      scale_fill_manual(values=c("grey", "blue", "red", "black", "green",
      "lightblue")) + theme_void()
```

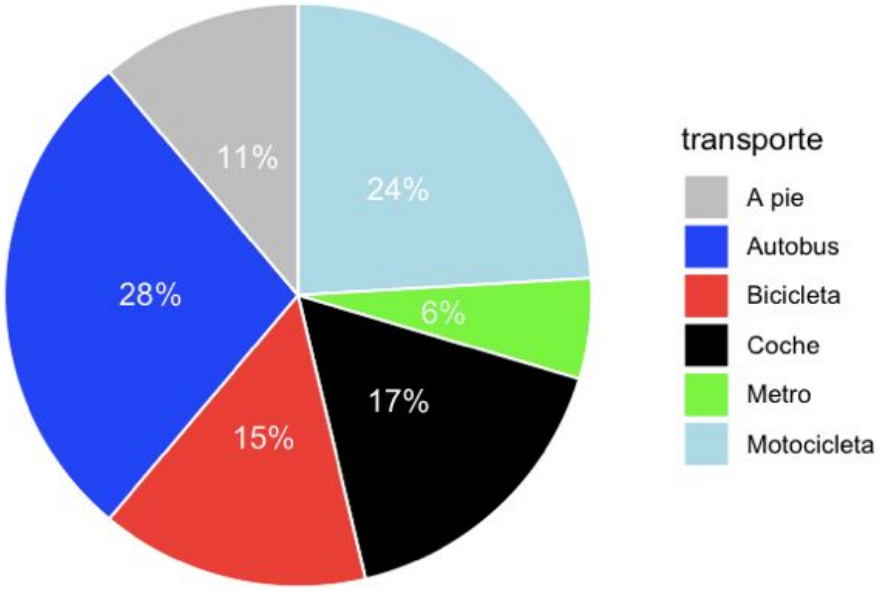
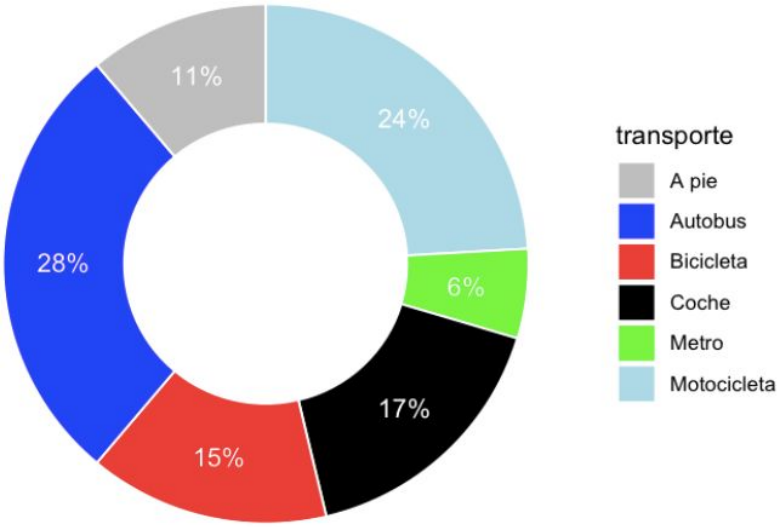


Figura 24. Diagrama sectorial de la tabla *transporte*

Ahora nos centraremos en crear el agujero interior de la gráfica. Esto se consigue usando la función *xlim()*, en la cual establecemos límites en los ejes X e Y (0.5 y 2.5 respectivamente). De esta forma, las observaciones que no se encuentren en este rango serán desechadas y no se mostrarán en ninguna capa. Además, suprimiremos el atributo "width" de la función *geom_bar()* y modificaremos el argumento "x" de la función *aes()* (aesthetic) a 2.

```
> ggplot(df_transporte, aes(x = 2, y = personas, fill = transporte)) +
      geom_bar(stat = "identity", color = "white") + coord_polar("y", start = 0) +
      geom_text(aes(label = paste0(round((personas/sum(personas))*100), "%")),
      position = position_stack(vjust = 0.5), color = "white") +
      scale_fill_manual(values=c("grey", "blue", "red", "black", "green",
      "lightblue")) + theme_void() + xlim(0.5, 2.5)
```

Figura 25. Diagrama de anillo de la tabla *transporte*



9.5 Elaboración de un diagrama de área polar con ggplot2

La rosa de Nightingale combina grupos de datos cuantitativos y cualitativos en un determinado periodo de tiempo es por ello que para desarrollar un diagrama de área polar he optado por crear un marco de datos (df_viajes_bus) en R que recopile información sobre los viajes en autobús de un estudiante universitario. En él, he plasmado cuantos viajes y en qué fase del día los realiza (mañana, tarde y noche) durante 12 días.

```
> df_viajes_bus
   dias veces fase
1     1     2     M
2     1     1     T
3     1     0     N
4     2     1     M
5     2     2     T
6     2     1     N
7     3     3     M
8     3     0     T
9     3     1     N
10    4     2     M
11    4     2     T
12    4     0     N
13    5     1     M
14    5     1     T
15    5     0     N
16    6     2     M
17    6     1     T
18    6     1     N
19    7     3     M
20    7     0     T
21    7     1     N
22    8     2     M
23    8     0     T
24    8     0     N
25    9     1     M
26    9     2     T
27    9     1     N
28   10     3     M
29   10     2     T
30   10     1     N
31   11     2     M
32   11     2     T
33   11     1     N
34   12     1     M
35   12     0     T
36   12     0     N
```

Tabla 9. Data frame *df_viajes_bus*

En primer lugar emplearemos la función ggplot() para indicar la entrada de los datos estadísticos. Usaremos el data frame de *df_viajes_bus* como fuente de información de la gráfica, donde estableceremos los vectores "dias" (sin tilde) y "veces" en los ejes X e Y respectivamente y mediante el argumento fill, introduciremos el vector "fase" que contendrá las tres distintas categorías.

Acto seguido, estableceremos un gráfico de barras usando *geom_bar()*. Tengamos en cuenta que para desarrollar el diagrama de área polar, al igual que sucede con el resto de gráficas circulares en ggplot2, primero tendremos que plasmar los datos en un diagrama de barras que luego convertiremos en un diagrama sectorial mediante la función *coord_polar()*.

Por motivos meramente estéticos, cambiaremos los colores mediante *scale_fill_manual()*, quitaremos la etiqueta (del vector "fase") de las categorías y eliminaremos todos los ejes y fondos de apoyo que emplea ggplot en sus gráficas, principalmente porque en este gráfico en concreto, entorpecierían al lector.

```
> ggplot(data = df_viajes_bus, aes(x = dias, y = veces, fill = fase)) +
      geom_bar(stat = "identity", width = 1, color = "black", size = 0.1) +
      coord_polar() + labs(fill="") +
      scale_fill_manual(values=c("lightblue", "grey", "orange")) + theme_void()
```

El diagrama de Nightingale que se muestra a continuación refleja exactamente la información del marco de datos *df_viajes_bus* de manera mucho más eficaz que una tabla, mediante el cual podemos diferenciar claramente el número de viajes realizados y el periodo en el que fueron realizados (en azul los viajes matutinos, en naranja los vespertinos y en gris los nocturnos). Esto sucede porque el gráfico no dispone que una elevada tasa de viajes diarios, lo cual hace que las subdivisiones de cada sector sean más amplias puediéndose así diferenciar las cantidades de los viajes con mayor facilidad.

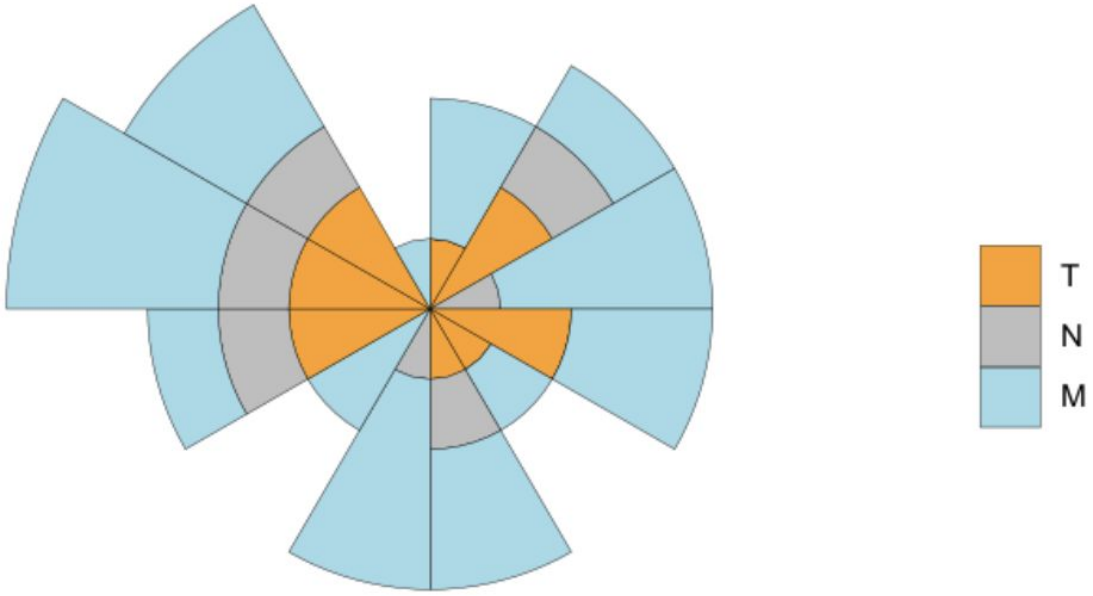


Figura 26. Diagrama de área polar de la tabla *df_viajes_bus*

9.6 Elaboración de un diagrama multinivel

Supongamos que queremos analizar y representar gráficamente a los alumnos de una determinada clase en función de sus géneros y sus notas, realizando subdivisiones entre sí. Para ello disponemos de la siguiente tabla:

Alumnos	Género	
Excelentes	Hombres	3
	Mujeres	3
Buenos	Hombres	5
	Mujeres	5
Normales	Hombres	7
	Mujeres	7

Tabla 10. Tabla *alumnos*

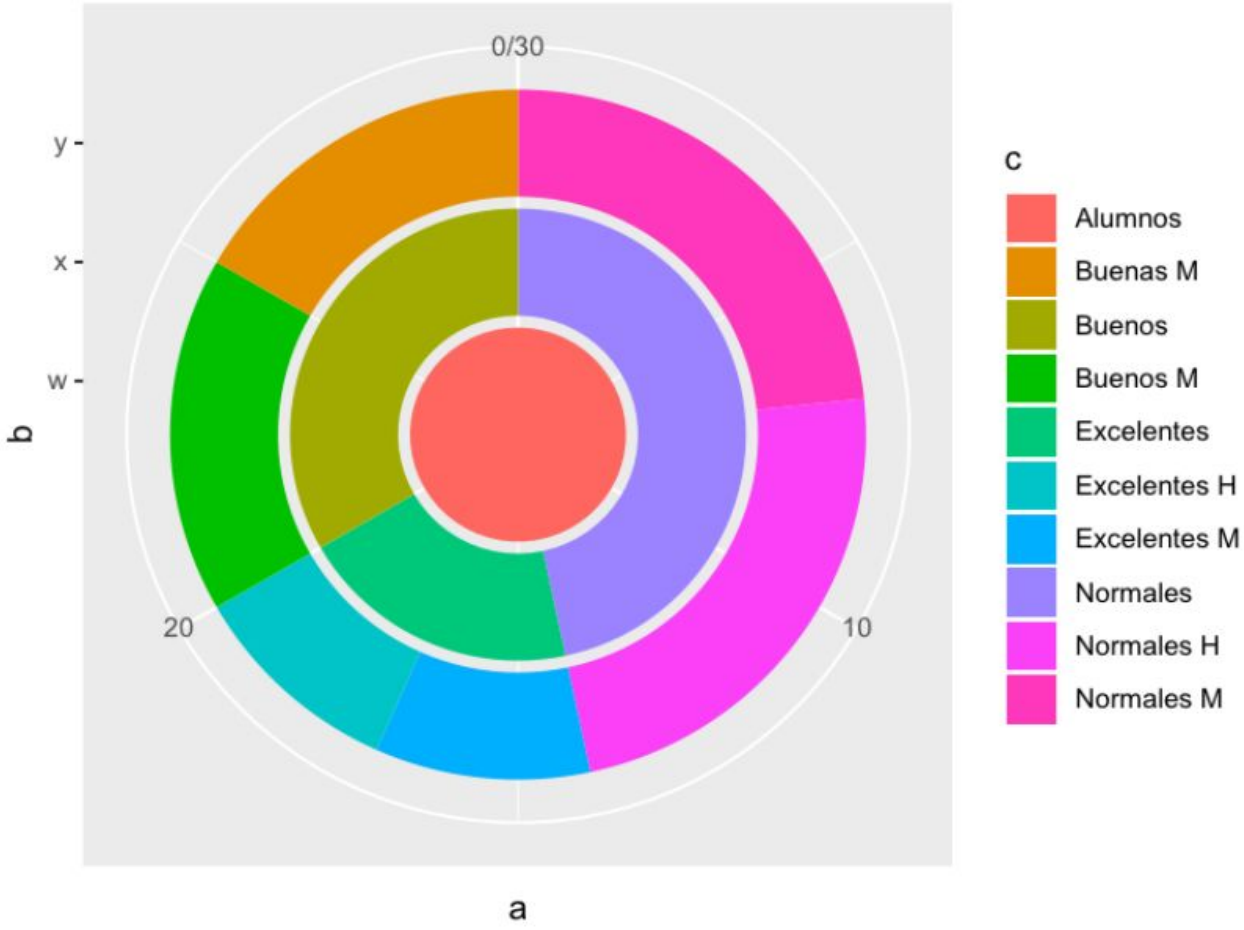
La mayor dificultad que nos presenta este tipo de gráfico reside en el marco de datos, por tanto, elaborar un data frame adecuado para su posterior manipulación es esencial. Para ello, crearemos uno (que llamaremos df_multi), al que asignaremos tres vectores: uno con la información numérica, otro con un carácter que repetiremos entre los datos de cada nivel (ya que servirá para que agrupemos los datos) y el último con todas las categorías que conformarán la gráfica.

```
> df_multi <- data.frame(a = c(6, 10, 14, 3, 3, 5, 5, 7, 7, 30),
      b = c("x", "x", "x", "y", "y", "y", "y", "y", "y", "w"),
      c = c("Excelentes", "Buenos", "Normales", "Excelentes H", "Excelentes M",
      "Buenos H", "Buenas M", "Normales H", "Normales M", "Alumnos"))
```

Después haber definido el conjunto de datos *df_multi*, la introduciremos en la función *ggplot()* adjuntando el vector "b" al eje de abscisas para agrupar los datos y el vector "a" en el de ordenadas. Procedemos a crear el gráfico de barras con *geom_bar()* y le añadimos la función *coord_polar()* para convertirlo en una gráfica circular.

```
> ggplot(df_multi, aes(x = b, y = a, fill = c)) +
       geom_bar(stat = "identity") +
       coord_polar(theta= "y")
```

Figura 27. Diagrama multinivel de la tabla *alumnos*



9.7 Elaboración de un spie chart

Para este supuesto, usaremos una media ficticia de los pasajeros anuales (por día de la semana) de una de las líneas de una compañía de autobuses.

En la siguiente tabla podemos observar las medias de pasajeros y el número total de cada día de la semana en el año X de la línea 29.

Pasajeros de la línea 29		
Días de la semana	Media de pasajeros por día	Número de días en el año X
Lunes	1010	53
Martes	1000	52
Miércoles	970	52
Jueves	990	52
Viernes	940	52
Sábado	760	52
Domingo	660	52

Tabla 11. Tabla de la línea 29

Para realizar el spie chart el procedimiento sería el siguiente. Primero realizaremos un marco de datos reagrupando la información dada por la tabla de la línea 29, después, elaboraremos un gráfico de barras (apéndice 2) con dichos datos y por último, convertiremos el gráfico en un spie chart haciéndolo girar sobre su eje de abscisas.

```
> df_pasajeros <-
             data.frame(dia=c("L","M","X","J","V","S","D"),
             media=c(1010,1000,970,990,940,760,660),
             num_dias=c(53,52,52,52,52,52,52))
> df_pasajeros$porcentaje_dias <- df_pasajeros$num_dias / sum(df_pasajeros$num_dias)
> df_pasajeros$cum_porcent <- cumsum(df_pasajeros$num_dias)
```

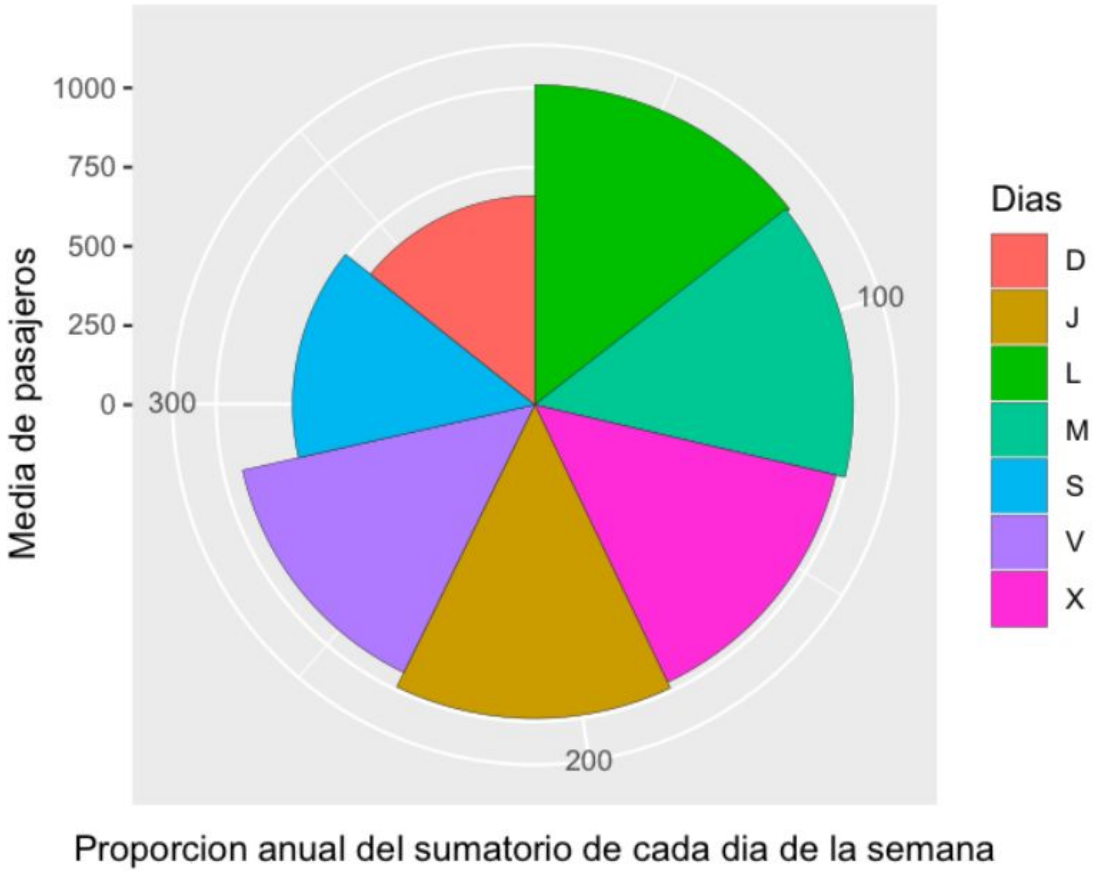
Como podemos observar, hemos creado el marco de datos "*df_pasajeros*" introduciendo los datos de la tabla superior empleando los objetos "dia", "media" y "num_dias". En este paso, crearíamos también nuevos objetos a los que denominaremos "porcentaje_dias" y "cum_porcent" para facilitarnos el cálculo de las proporciones de "*df_pasajeros*", que serán los datos que tendremos como referencia a la hora de crear el spie chart.

```
> ggplot(df_pasajeros, aes(x = cum_porcent, y= media, fill = dia)) +
              geom_bar(aes(width = num_dias, y = media),
              color = "grey10", size = 0.1, stat = "identity") +
             coord_polar(theta = "x") +
             labs(x = "Proporcion anual del sumatorio de cada dia de la semana",
             y = "Media de pasajeros", fill = "Dias")
```

Finalmente, empleamos la función *ggplot()* con el data frame de "df_pasajeros", estableciendo en X el vector "cum_porcent", en Y el vector "media" y en fill el vector "dias". Usaremos *geom_bar()* aplicando en el ancho de la función aesthetics el objeto "num_dias" y el vector "media" en el eje de ordenadas. En el atributo *color* pondremos "grey10", lo que hará que los bordes del gráfico se vuelvan grises y en *size*, estableceremos un valor de 0,1. Como de costumbre, definiremos a *stat* como "identity" para que la gráfica nos muestre debidamente sus valores numéricos.

En esta ocasión, la función *coord_polar()* hará que la gráfica rote en torno al eje X (a diferencia de los supuestos anteriores en los cuales la rotación se hacía sobre el eje Y) y con *labs()*, sobreescribiremos el etiquetado de la gráfica.

Figura 28. Spie chart de la tabla de la línea 29



9.8 Elaboración de un diagrama cuadrado mediante ggplot2

Para crear un "waffle chart", emplearemos las estimaciones de ventas por cada 100 automóviles vendidos por la empresa de automoción "Bercedes Menz" para el año 2022.

Estimación de las ventas de automóviles para el año 2022 por cada 100 coches vendidos	
Gasoil	37
Diesel	20
Eléctrico	28
Híbrido	15

Tabla 12. Estimaciones de Bercedes Menz para el 2022

En primer lugar crearemos un data frame empleando la función *structure()*. En ella, crearemos una lista con la función *list()* en la cual introduciremos los datos en los vectores "motor" y "ncoches", donde "L" es un sufijo que emplearemos con los datos numéricos para indicarle a R que se tratan de números enteros. Mediante *.Names* y *row.names* estableceremos los nombres de ambas columnas y con *class*, le indicaremos a *structure()* que se trata de un marco de datos.

```
> tb <- structure(list(motor = c("Gasoil", "Diesel", "Electricos", "Hibridos"),
       ncoches = c(37L, 20L, 28L, 15L)), .Names = c("motor", "ncoches"),
       row.names = c(NA, -4L), class = "data.frame")
```

El siguiente paso será emplear *ggplot()* introduciendo los datos del nuevo data frame especificando a R que en *aesthetics()*, nos factorice por el primer nivel en el eje X. Después, le añadiremos *geom_bar()* para crear un gráfico de barras (ver apéndice 3), que convertiremos en un gráfico sectorial mediante *coord_polar().* También, usaremos *labs()* para dejar las etiquetas de los ejes vacías.

```
> ggplot(tb, aes(x = factor(1), weight = ncoches, fill = motor)) +
       geom_bar() + scale_y_continuous() + coord_polar(theta = "y") +
       labs(x = "", y = "")
```

El próximo paso consistirá en transformar nuestro diagrama sectorial en uno cuadrado. Por este motivo crearemos el objeto *"tb_waffle"* mediante la función *expand.grid()*. Con ella, crearemos otro data frame que resultaría en el producto cartesiano de sus argumentos. Para establecer las dimensiones del diagrama, el eje de ordenadas irá de 1 a 5 (el valor que tomará *ndeep*). En el de abscisas, crearemos con *seq_len()* una secuencia hasta el valor numérico que hayamos designado (20 en este caso) y *ceiling()* nos redondeará el valor numérico hacia arriba.

```
> tb_waffle <- expand.grid(y = 1:ndeep, x = seq_len(ceiling(sum(tb$ncoches)/ndeep)))
```

Acto seguido crearemos el objeto "vector" mediante la función *rep()*, que replicaría los valores de los vectores "motor" y "ncoches" en X.

```
> vector <- rep(tb$motor, tb$ncoches)
```

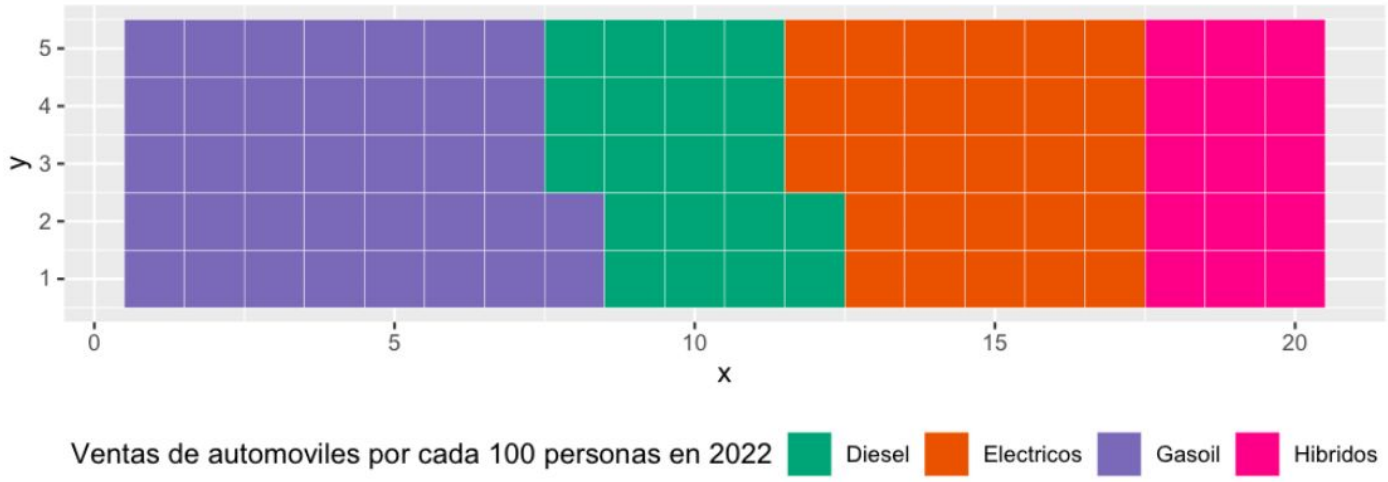
Ahora modificaremos el atributo "motor" del recién creado data frame "*tb_waffle*". Para ello combinaremos mediante *c()* el objeto "*vector*" y duplicaremos el valor de las filas de "*tb_waffle*" con *nrow()* menos la longitud de "*vector*", todo ello usando la función *rep()*.

```
> tb_waffle$motor <- c(vector, rep(NA, nrow(tb_waffle) - length(vector)))
```

Finalmente, usaremos *ggplot()* enlazándolo con "*tb_waffle*". El resto del código se enfocará en cuestiones estéticas, como *geom_tile()* que lo emplearemos para poner en blanco las líneas de separación y *scale_fill_manual()*, que lo usaremos para añadir un título y una paleta de colores. Por último, la función *legend.position()* dentro de *theme()*, posicionará la leyenda debajo del gráfico (en vez de en un lateral como vendría por defecto).

```
> ggplot(tb_waffle, aes(x = x, y = y, fill = motor)) + geom_tile(color = "white") +
      scale_fill_manual("Ventas de automoviles por cada 100 personas en 2022",
      values = RColorBrewer::brewer.pal(4, "Dark2")) +
      theme(legend.position="bottom")
```

Figura 29. Diagrama cuadrado de la tabla de Bercedez Menz



10 Conclusiones finales

Es evidente que los puntos fuertes de los gráficos circulares residen en su carácter estético y opciones de personalización, no obstante, pese a la gran variedad de alternativas de diagramas de sectores que hemos podido abordar a lo largo del trabajo, no considero que sea la forma más adecuada de exposición de datos cuantitativos.

Como ya he mencionado anteriormente, los diagramas de barras y puntos ofrecen una mayor claridad y una mayor facilidad para su decodificación que los de sectores, además de ofrecernos una comparativa entre las distintas muestras, cosa que no sucedería con los diagramas de sectores dado que su codificación según el ángulo y el área de cada sección dificultarían esta tarea.

Además, cuantas más muestras haya que analizar o cuanto mayor sea la diferencia de tamaño entre ellas, más ineficiente se volverá el diagrama sectorial. Estos factores son los que nos harían decantarnos por cualquiera de las otras dos opciones mencionadas, dada su mayor eficacia a la hora de visualizar los datos.

En definitiva, si lo que el emisor busca es resaltar el método de análisis gráfico en vez de la información en sí, emplear un diagrama de sectores no sería una mala idea dada su polivalencia y atractivo estético. Sin embargo, si lo que buscara se tratara principalmente en transmitir la información de manera clara, ordenada y rápida, sugeriría cualquier otro método gráfico como los diagramas de barras o puntos.

11 Glosario

Big data: término que describe cualquier cantidad numerosa de datos estructurados, semiestructurados y no estructurados propensos a ser extraídos para obtener información.

Data frame: estructura de datos bidimensional compuesta por filas y columnas.

Echinochloa crus-galli: especie de planta del género echinochloa perteneciente a la familia de las poáceas.

Pie chart: gráfico circular o de tarta en inglés.

Silogismo: razonamiento deductivo o inductivo que está formado por dos premisas y una conclusión que es el resultado lógico que se deduce de las dos premisas.

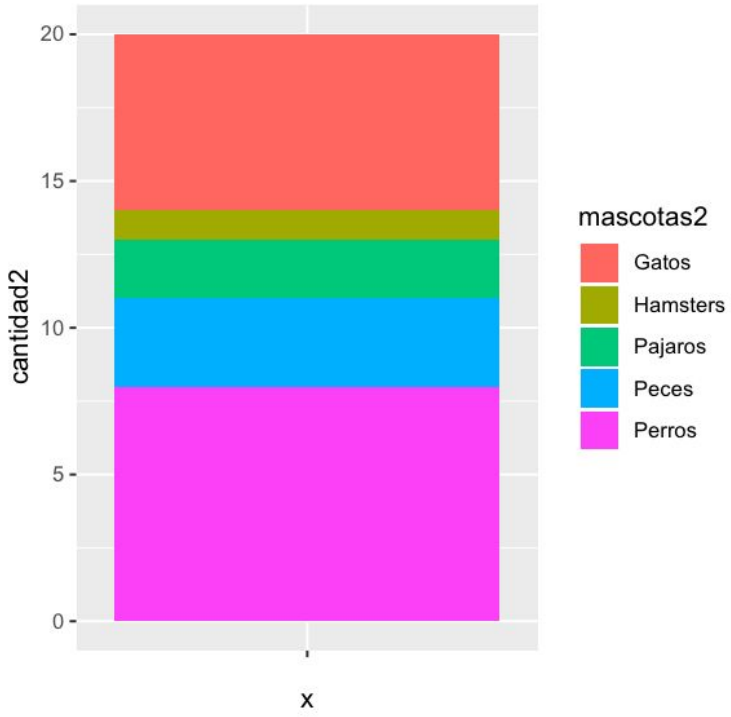
Spie chart: gráfico circular que posee la característica de comparar variables cuantitativas mediante la longitud del rádio de sus secciones.

String: cadena de caracteres.

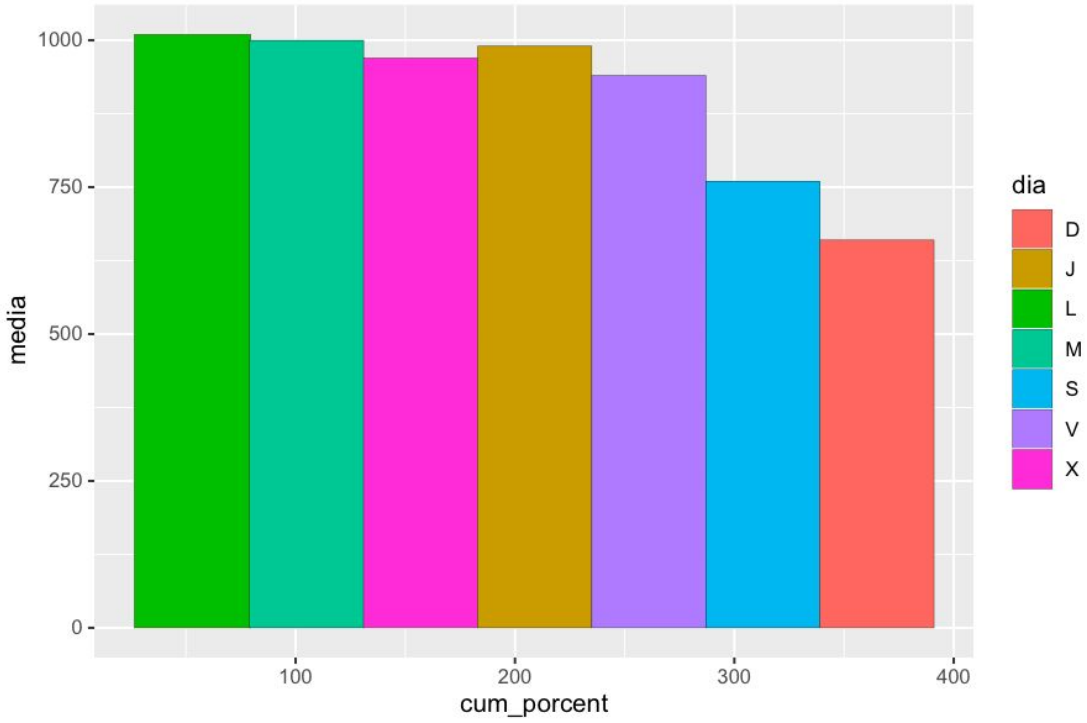
Waffle: gofre en inglés.

12 Apéndices

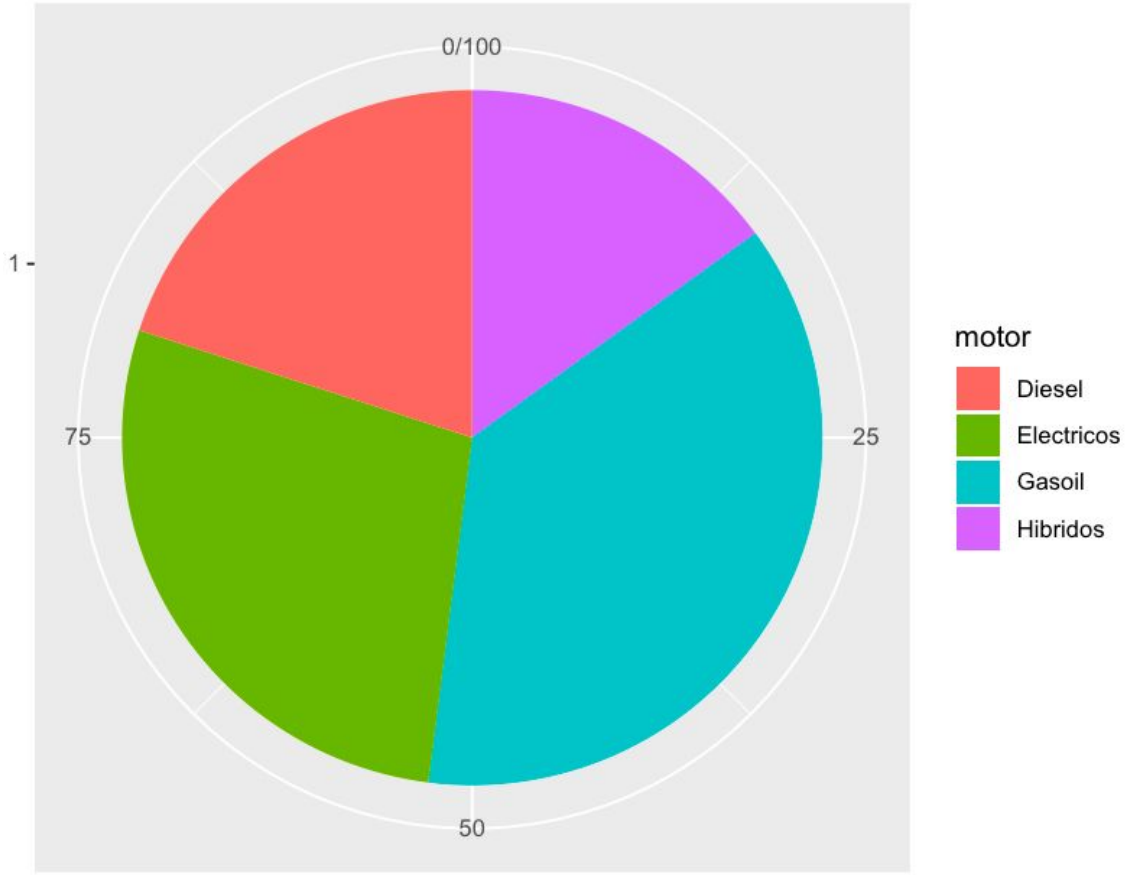
12.1 Diagrama de barras complementario del primer ejemplo con la extensión GGPLOT2



12.2 Diagrama de barras complementario del ejemplo del spie chart



12.3 Diagrama de sectores complementario del ejemplo del diagrama cuadrado



13 Bibliografía

Arteaga, Blanca. (s.f.). Descripción de una variable: Tablas de frecuencias. *Universidad Carlos III de Madrid*. Recuperado de: http://halweb.uc3m.es/esp/Personal/personas/imolina/MiDocencia/Estadistica_INFDOC/Tema2DescripUnaVar_TablasFrec.pdf

Cleveland, W. S. (1993). Visualizing Data. *Hobart Press*, Summit, NJ.

Cleveland, W. S. (1994). The Elements of Graphing Data (Revised Edition), *Hobart Press*, Summit, NJ.

De La Fuente Fernández, S. (2011). Tablas contingencia. *Universidad Autónoma de Madrid*. Recuperado de: http://www.estadistica.net/ECONOMETRIA/CUALITATIVAS/CONTINGENCIA/tablas-contingencia.pdf

Few, S. (2006). Beautiful Evidence: A Journey through the Mind of Edward Tufte. *B Eye Network*. Recuperado de: http://www.b-eye-network.com/view/3226

Few, S. (Agosto, 2007). Save the pies for dessert. *Perceptual Edge Visual Business Intelligence Newsletter.* Recuperado de: http://www.perceptualedge.com/articles/08-21-07.pdf

Field, A., Field, Zoe & Miles, J. (2012). Discovering statistics using R. *Thousand Oaks*, *Sage*. London.

Gomila J. G. (s.f.). Representación gráfica en R. *Juan Gabriel Gomila*. Recuperado de: http://juangabrielgomila.com/graphics-in-r/

Gráficos para variables cuantitativas. *Universidad de Málaga*. Recuperado el 4 de octubre de 2019 de: https://virtual.uptc.edu.co/ova/estadistica/docs/libros/ftp.bioestadistica.uma.es/libro/node10.htm

Jacobs, A. (October, 2014). Visualizing Health Expenditure using Spie Charts (and R). *Unconstant conjunction*. Recuperado de: https://unconj.ca/blog/visualizing-health-expenditure-using-spie-charts-and-r.html

Kosara R. & Skau D. (2016). Judgment error in pie chart variations. *Short Paper Proceedings of the Eurographics/IEEE VGTC Symposium on Visualization (EuroVis)*. Recuperado de: https://kosara.net/papers/2016/Kosara-EuroVis-2016.pdf

Maron L. & Espinoza R. (s.f.). Gráficos elegantes con ggplot2. Recuperado de: https://luisxsuper.github.io/m_ggplot2.html

Mezo, J. (s.f.). Curso de estadística aplicada a las ciencias sociales. *Universidad de Castilla la Mancha*. Recuperado de:
https://previa.uclm.es/profesorado/jmezo/estadistica/t2.pdf

Nightingale, F. (1858). Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army. Recuperado de:
http://www.scottlan.edu/lriddle/women/nightpiechart.htm.

Pacheco, J. (Julio, 2019). Polígono De Frecuencia (Definición, Objetivo, Características). *Web y Empresas*. Recuperado de:
https://www.webyempresas.com/poligono-de-frecuencia/

Playfair, W. (1801). The Statistical Breviary. *T. Bensley*, pp. 4-13.

Polígonos de frecuencia: variables discretas. *Ditutor*. Recuperado el 4 de octubre de 2019 de: https://www.ditutor.com/estadistica/poligonos_frecuencias.html

Robbins, N. B. (2005). Creating More Effective Graphs. *Wiley*, Hoboken, NJ.

Robbins, N. B. (2011). Communicating Data Clearly, *Strata Conference*. New York.

Simkin, D. & Hastie, R. (1987). An information-processing analysis of graph perception. J. Am. 82(398), pp. 454-465.

Skau, D. & Kosara, R. (2016). Arcs, angles, or areas: individual data encodings in pie and donut charts. Comput graph forum 35, pp. 121-130.

Spence, I. & Lewandowsky S.(Enero, 1991). Displaying Proportions and Percentages. *Applied Cognitive Psychology*, Vol. 5, pp. 61-77.

Spence, I. (2005). No Humble Pie: The Origins and Usage of a Statistical Chart. *Journal of Educational and Behavioral Statistics*, Vol. 30, No. 4, pp. 353-368.

The Comprehensive R Archive Network. *R Cran*. Enlace de descarga de R: https://cran.r-project.org

Trick, A. (Julio, 2018). Nightingale Rose in R. *Andrew Trick*. Recuperado de: http://andrewtrick.com/coffee_rose.html

Tufte, E. (1983). The Visual Display of Quantitative Information. *Graphics Press*, Cheshire, CT.

Variable cualitativa. *Enciclopedia financiera*. Recuperado el 3 de octubre de 2019 de: https://www.enciclopediafinanciera.com/definicion-variable-cualitativa.html

Variable cuantitativa. *Enciclopedia económica*. Recuperado el 3 de octubre de 2019 de: https://enciclopediaeconomica.com/variable-cuantitativa/

Von Huhn, R. (1927). Further studies in the graphic use of circles and bars. *Taylor & Francis Online*. Recuperado de: http://www.tandfonline.com/doi/abs/10.1080/01621459.1927.10502938?journalCode=uasa20

Wickham, H. (2010). Ggplot2: elegant graphics for data analysis. *Springer*. New York.

Wiper, M. (s.f.). Gráficos para datos cuantitativos. *Universidad Carlos III de Madrid*. Recuperado de: http://halweb.uc3m.es/esp/Personal/personas/mwiper/docencia/Spanish/GC/guardia%20civil/lecture%20notes/class4.pdf