Lineamientos conceptuales para la visualización estadística

Camila Acosta Ramirez

2021-07-02

Contents

Portada

Espacio para la portada ....

Chapter 1

Introduction

What is Lorem Ipsum Lorem Ipsum is simply dummy text of the printing and typesetting industry Lorem Ipsum has been the industry's standard dummy text ever since the 1500s when an unknown printer took a galley of type and scrambled it to make a type specimen book it has?

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 1. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 1.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 1.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2021) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

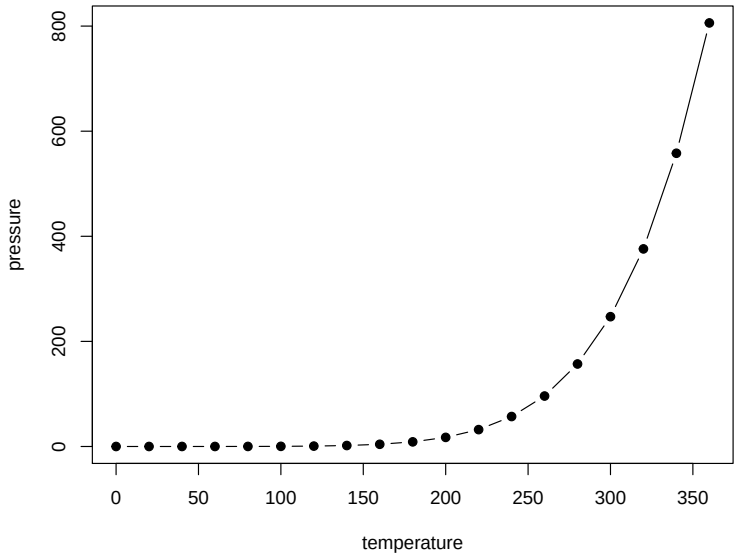

Figure 1.1: Here is a nice figure!

Table 1.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 2

Partes principales de un gráfico

A la hora de crear un gráfico es necesario tener presente cada uno de los elementos que lo conforman y determinar cual es la mejor manera de representar cada uno de estos para lograr el impacto deseado en la visualización. El diseño correcto de estos elementos garantizará el éxito de su gráfico, al comunicar de manera acertada la información que pretende presentar. Dentro de la gama de gráficos estadísticos básicos se identifican tres elementos importantes los cuales son ejes, geometría en la cual se incluyen la forma, tamaño y color, tipos de líneas y texto el cual incluye las etiquetas de los ejes, título y leyenda.

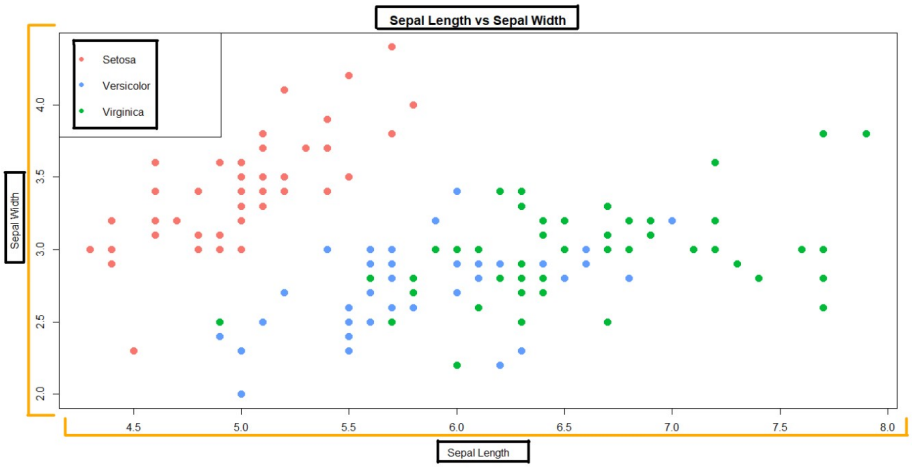

Figure 2.1: Principales partes de un gráfico

La figura 2.1 presenta estos elementos importantes. En los recuadros negros se encierra todo lo relacionado con texto, etiquetas de los ejes, título y leyenda; las líneas naranjadas representan los ejes del gráfico de manera horizontal corre el eje X y de manera vertical el eje Y. En la parte central de la visualización se ubican las observaciones a las cuales se les aplica la geometría, dependiendo del tipo de gráfico es posible cambiar el tamaño, forma y color de cado dato.

A continuación, se presentan las características que se consideraron más importantes para tener en cuenta a la hora de dar formato y personalizar cada uno de los elementos mencionados.

2.1 Ejes

Los ejes son de los elementos de mayor relevancia dentro de cada gráfico ya que determinan la posición donde se ubica cada dato. Cuando se trata de gráficos en dos dimensiones, los más comunes, las posiciones son descritas a través de dos valores que especifican un punto de forma única, y por lo tanto se necesitan escalas de posición, estas escalas son generalmente los ejes X y Y. Por convención general el eje X corre horizontalmente y el eje Y lo hace de manera vertical, aunque esto no siempre debe ser así, hay gráficos en los cuales los ejes son radiales. El objetivo principal de las visualizaciones que se crean es comparar los datos, es decir, identificar el comportamiento de cada observación en relación con las demás que posee el conjunto de datos. Para realizar estas comparaciones es importante definir la escala de los ejes de manera adecuada, una mala elección de estas escalas lo puede conducir a interpretar la información de manera errada; es recomendable iniciar los ejes en 0, aunque no siempre es necesario si es importante considerar que los datos sean comparables.

Para ilustrar la importancia de la correcta elección del inicio del eje Y consideremos la visualización de cantidades a lo largo de una escala lineal. La figura 2.2 muestra las ventas en cinco estados de EE.UU; una vista rápida a esta visualización indica que las ventas en North Dakota son extremadamente bajas en comparación con los demás estados, sin embargo, este gráfico es engañoso ya que las ventas inician en $900 USD, por lo tanto, mientras que el punto final de cada barra indica de manera correcta el total de ventas, la altura de la barra representa la medida en que las ventas superan los $900 dólares; la percepción humana entenderá la altura de cada barra como las ventas por estado lo que conlleva a una interpretación errónea.

La forma correcta de visualizar estos datos se presenta en la figura 2.3, es claro que existen diferencias entre las ventas por estados, pero no son tan distantes como lo muestra la figura 2.2, las ventas en los cinco estados presentados son comparables. En este caso es particular se debe seguir la regla de iniciar los ejes en cero.

Como se mencionó anteriormente no siempre es necesario iniciar los ejes en

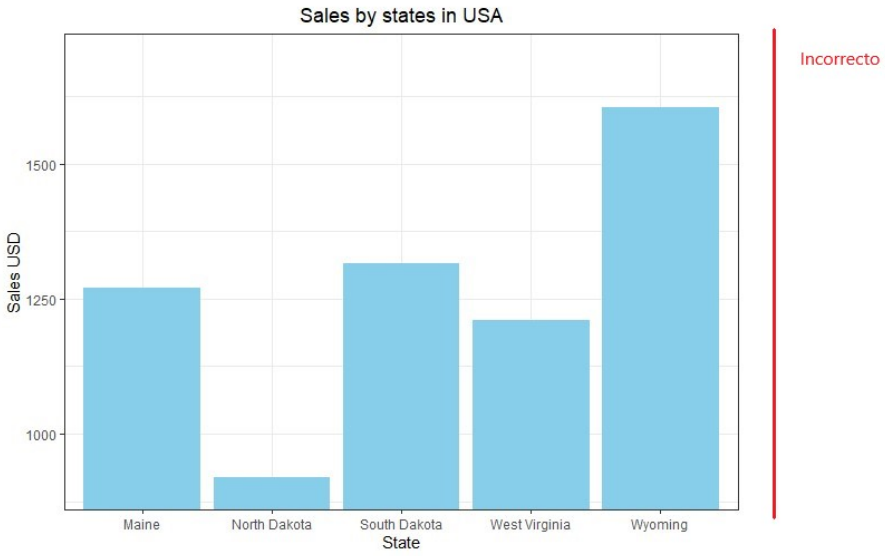

Figure 2.2: Ventas por estados de EE.UU, visualiazción engañosa

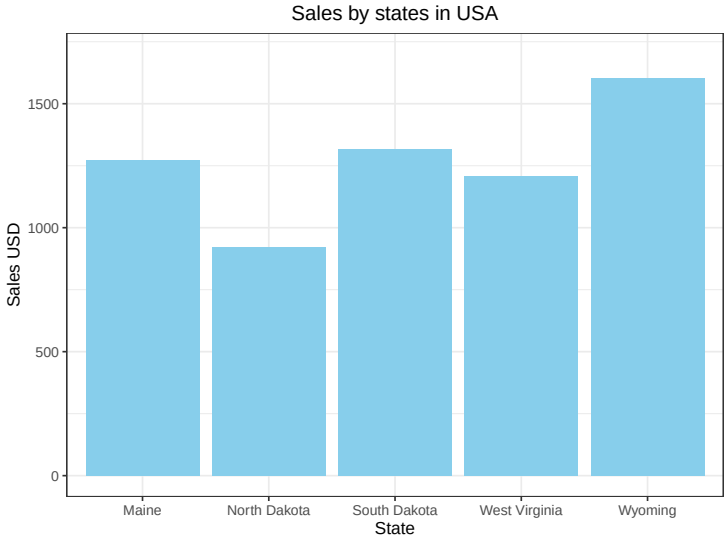

Figure 2.3: Ventas por estados de EE.UU, uso correcto de la escala lineal

cero, existen ocasiones en las cuales los datos se encuentran agrupados en un intervalo específico y comenzar los ejes en cero hará que las discrepancias entre los datos no se observen con claridad, la figura 2.4 presenta la relación X-Y de datos ficticios, fue etiquetada como incorrecta ya que iniciar el eje Y en cero no permite visualizar con claridad las ubicaciones de los puntos a lo largo del eje vertical.

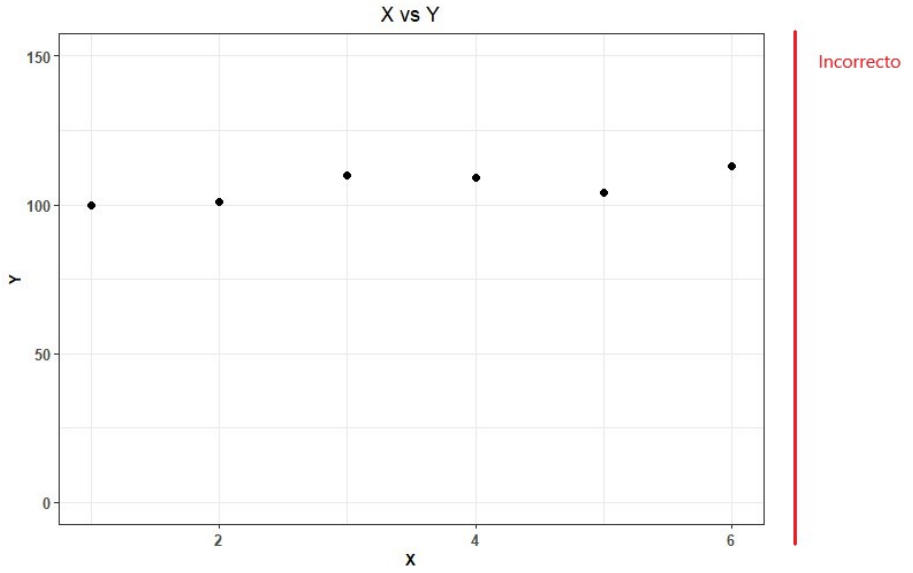

Figure 2.4: Visualización engañosa por iniciar eje en cero

En este caso se debe restringir el dominio del eje Y al intervalo 100 a 115 aproximadamente ya que allí es donde se concentran las observaciones, note que la figura 2.5 es más informativa ya que permite identificar con claridad la ubicación de los puntos dando como resultado una visualización clara y explicativa.

2.2 Geometría

La geometría es una parte primordial y que hará las visualizaciones mucho mas claras y entendibles. Dentro de las geometrías principales podemos considerar la forma, tamaño, tipo de línea y color.

2.2.1 Forma y tipo de línea

Estas dos son estéticas o atributos que generalmente se usan para representar datos categóricos, dentro de visualizaciones discretas o continuas. Cuando se

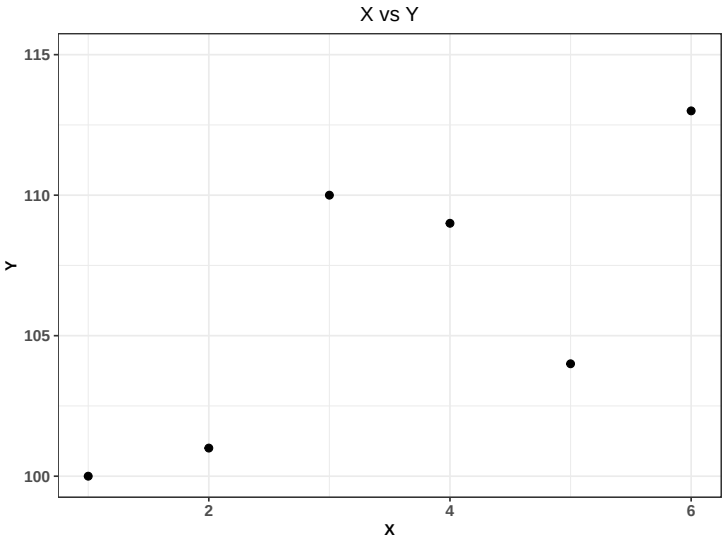

Figure 2.5: Restringir dominio del eje Y para una visualización más clara

trata de gráficos de dispersión se opta por usar diferentes formas a partir de una variable categórica con el fin de comparar los comportamientos de cada uno de los valores que toma la variable cualitativa. En el caso de los gráficos de líneas se opta por usar diferentes estilos o tipos de líneas nuevamente con el fin de diferenciar la categoría de los datos, por lo general los estilos usados son líneas continuas y punteadas. Ambos elementos pueden ser usados para distinguir o resaltar, en el caso de ser usados para distinguir se asigna una forma o tipo de línea a cada uno de los niveles de la variable categórica y en el caso de resaltado se usa la misma forma o tipo de línea para todos los datos excepto para aquellos elementos que queremos resaltar.

La figura 2.6 ilustra el uso de distintas formas para distinguir las especies de flores registradas en la base de datos Iris. Como ya se menciono se usan tantas formas como niveles tenga la variable categórica, en este caso se usan círculos, triángulos y cuadrados.

Si quisiéramos resaltar una de las especies por ejemplo, Versicolor debemos asignar la misma forma a las especies Setosa y Virginica y una distinta a la especie a resaltar, por ejemplo usar círculos y triángulos, como se presenta en la figura 2.7.

Cuando se trata de gráficos de líneas que comúnmente son utilizados para representar series de tiempo también es posible hacer uso del tipo de línea tanto para distinguir como para resaltar, al igual que en los gráficos de dispersión debemos usar tantas formas como categorías tenga la variable discreta si la intención es distinguir. Se usara un conjunto de datos de R llamado babynames que contiene

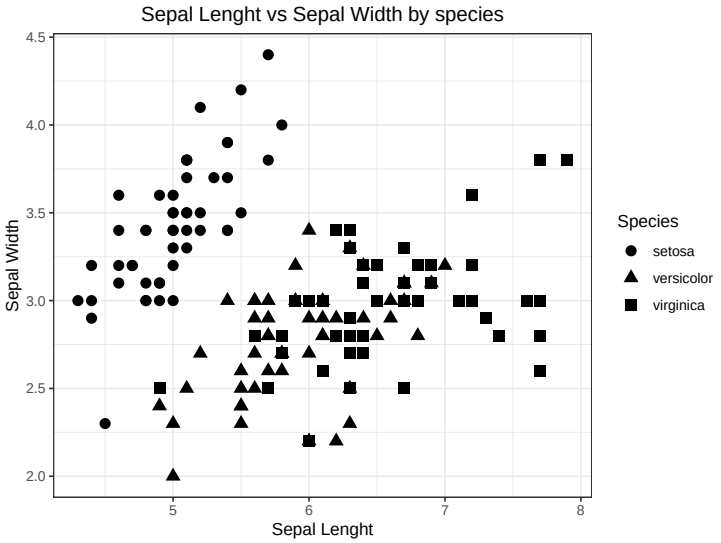

Figure 2.6: Uso de las formas para distinguir grupos de datos

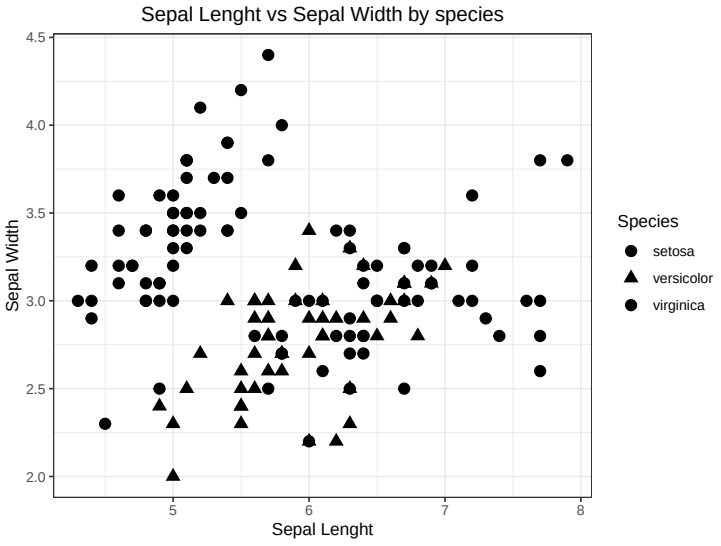

Figure 2.7: Uso de las formas para resaltar un grupo de observaciones

información de la cantidad de bebes nacidos desde el año 1880 hasta 2017 con los respectivos nombres, se realiza un gráfico de líneas con los tres nombres de niñas más populares y se cambia el tipo de línea para distinguir entre nombres como se muestra en la figura 2.8.

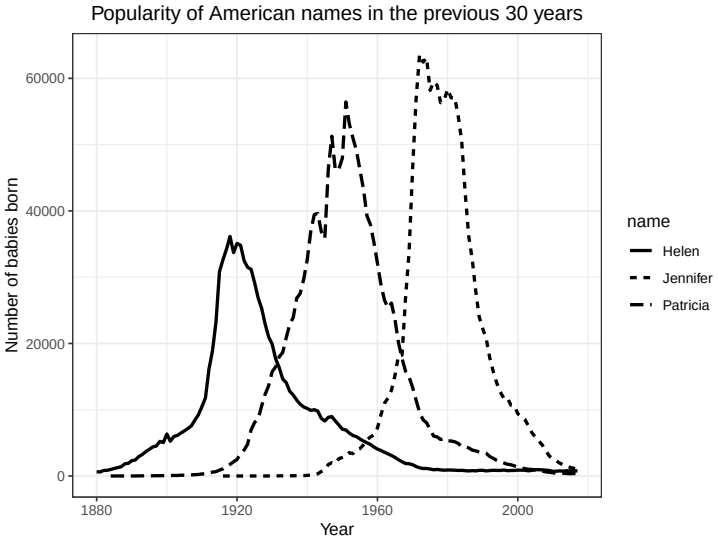

Figure 2.8: Uso del tipo de líneas para distinguir grupos de observaciones

En el caso de querer resaltar algún nombre en particular debemos usar un tipo de línea diferente para el nombre a resaltar, por ejemplo, en la figura 2.9 se usa una línea punteada para representar la popularidad a través del tiempo del nombre Jennifer y los otros dos se dejaron como líneas continuas.

Adicional a las formas y tipos de líneas se puede cambiar con el color para generar visualizaciones más atractivas e informativas, ya que al color se le pueden dar distintos usos, esto se muestra en la subsección 2.2.3.

2.2.2 Tamaño

El tamaño generalmente es una estética usada en gráficos de dispersión, se incluye una nueva variable continua o discreta que determina el tamaño de cada observación representada en el gráfico. Este atributo es de gran utilidad, pero se debe tener mucho cuidado al usarlo, ya que en el caso de datos desproporcionados un solo punto ocupará un tamaño exagerado que será poco comparable con los demás datos.

La figura 2.10 ilustra el uso de una variable discreta para asignar diferentes tamaños a las observaciones.

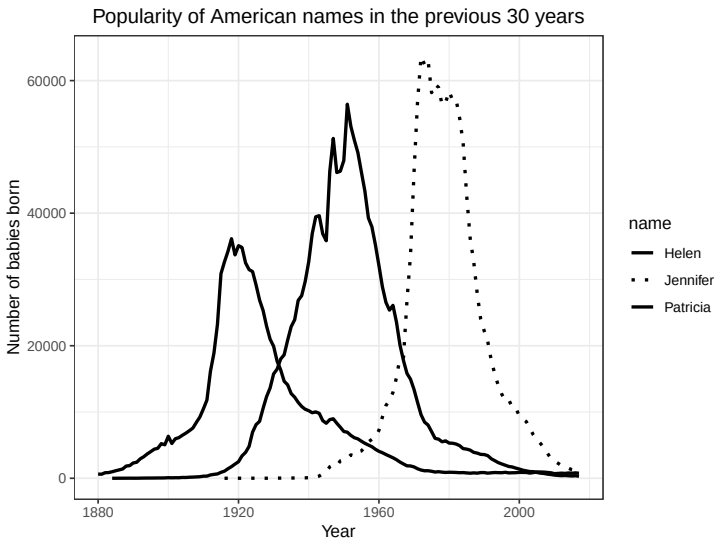

Figure 2.9: Uso del tipo de líneas para resaltar un grupo de observaciones

Figure 2.10: Asignación de tamaños mediante una variable discreta

La figura 2.11 muestra el uso de una variable continua para determinar el tamaño de cada observación, note que a demás de la geometría relacionada al tamaño se debe usar la transparencia para evitar que los países con mayor población oculten o se sobrepongan a aquellos a países de menor población.

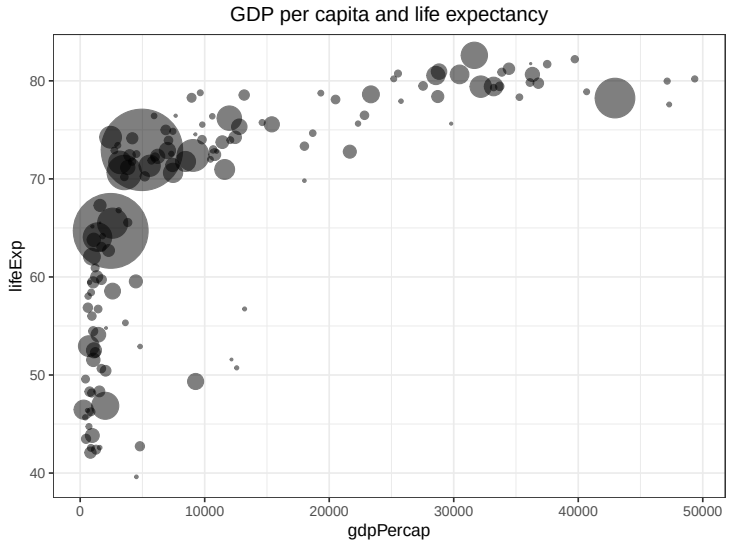

Figure 2.11: Asignación de tamaños mediante una variable continua

2.2.3 Color

El color es una de las estéticas más importantes y que pueden marcar una gran diferencia en la interpretación de sus datos. Existen algunos colores que destacan más que otros por lo que darán un peso innecesario a los datos, es decir, que atraen la atención de los usuarios a esos puntos y que pueden no necesariamente ser los de interés central, también es recomendable no superar los seis colores por gráfico. El color dentro de una visualización puede ser usado principalmente para tres casos: para distinguir grupos de datos entre sí, uso del color para representar valores de datos y finalmente puede ser usado para resaltar.

2.2.3.1 Distinguir grupos de datos

Emplear el color como una herramienta para distinguir es uno de los usos más comunes que se le da al color cuando se trata de gráficos que incluyen variables categóricas y que no tienen un orden específico como diferentes niveles de formación dentro de una universidad o departamentos dentro de un mapa. En

este caso, se utiliza una escala de colores cualitativa la cual contiene un conjunto finito de colores específicos que se eligen para versen claramente distintos entre sí y que al mismo tiempo deben ser equivalentes entre sí. Es decir que lo colores seleccionados se deben poder diferenciar de manera clara y precisa, pero uno no debe resaltar más que otro. También es importante que el conjunto de colores seleccionados no presente un orden ya que esto creara un orden en la visualización que por definición de los datos no se tiene. Como recomendación general, las escalas de color cualitativas funcionan mejor cuando hay de tres a cinco categorías diferentes; tener ocho o diez categorías hará que la tarea de hacer coincidir los colores sea tediosa, a demás la leyenda será demasiado extensa y el usuario tendrá que hacer un fuerte trabajo de búsqueda para identificar el color correspondiente a cada categoría; en el caso de muchas categorías se recomienda usar etiquetas directas sobre la observación para así facilitar la comprensión del gráfico aunque con esto también se debe tener cuidado ya que muchas etiquetas hará que la visualización se sature y la información no sea transmitida de la manera correcta.

La figura 2.12 muestra el uso correcto de los colores como herramienta para distinguir, se seleccionaron colores que contrastan entre sí pero no compiten por la atención, este gráfico en particular posee ocho categorías distintas pero aún así se logra identificar claramente cada una de las sedes de admisión.

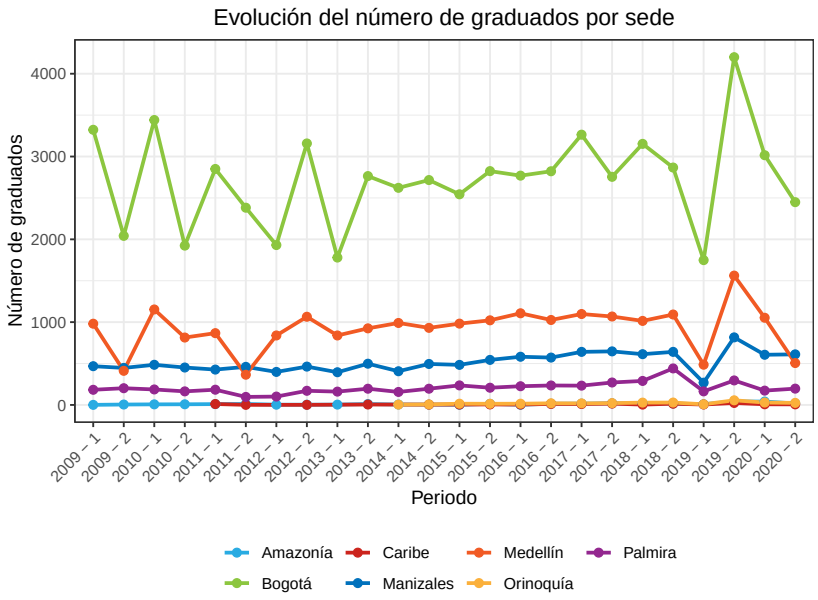

Figure 2.12: Uso del color como herramienta para distinguir grupos de datos

2.2.3.2 Representar valores

El color también puede ser usado como herramienta para representar variables cuantitativas como ingresos, temperatura, entre otros. En este caso se usa una escala de color secuencial, la cual indica claramente que valores son más grandes o pequeños y que tan distantes se encuentran dos valores específicos entre sí. Estas escalas secuenciales de color pueden basarse en un solo tono por ejemplo de azul oscuro a azul claro, o en múltiples matices por ejemplo de rojo oscuro a amarillo claro.

En la figura 2.13 presenta un uso adecuado de la escala de colores secuencial, se usa una paleta de un solo tono que inicia en azul claro y termina en un azul un poco más fuerte. Esta visualización presenta la cantidad de estudiantes graduados en el periodo 2020-II por departamento de nacimiento y la escala de color fue usada para colorear el conteo por cada uno de estos departamentos.

Figure 2.13: Uso de la escala de color secuencial para representar valores

Existen algunas ocasiones en las cuales es necesario visualizar la desviación de los valores de los datos en una de dos direcciones en relación con un punto medio neutral. Un ejemplo sencillo y muy básico de estos casos es cuando se tienen números positivos y negativos, que se representan por dos colores, puede ser verde para los números positivos y rojo para los negativos, a partir de la intensidad de estos colores se indica la lejanía con el cero. La escala de color usada en estos casos se denomina divergente y puede pensarse como dos escalas de colores secuenciales que se unen en un punto medio; estas escalas deben ser equilibradas, de modo que la progresión de los colores claros en el centro a los colores oscuros del exterior debe ser la aproximadamente igual para ambas direcciones.

La figura 2.14 presenta una de los tantos usos que se le puede dar a las escalas de

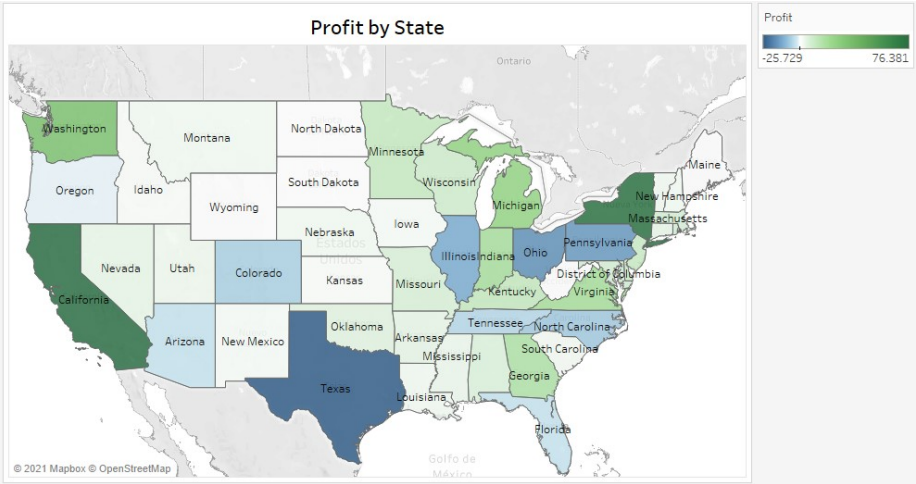

Figure 2.14: Escala divergente para representar valores

colores divergentes, se presenta los beneficios por estados. Es importante notar que la escala de colores no se encuentra equilibrada y esto se debe a la desproporción de los datos, a pesar de esto la visualización sigue siendo informativa y se logra distinción entre los colores.

2.2.3.3 Resaltar observaciones

Emplear el color para resaltar observaciones es de gran utilidad cuando el conjunto de datos contiene información clave sobre la historia que se quiere contar a través del gráfico y enfatizar en estos elementos conlleva a una mejor comprensión de la información que se desea comunicar. Para lograr este énfasis podemos colorear estos elementos de la figura con colores o tonalidades que se destaquen vívidamente contra el resto de la figura; generalmente se usan escalas de color de acento, las cuales contienen tanto un conjunto de colores tenues como un conjunto coincidente de colores más fuertes.

Cuando se trabaja con una escala de color de acento, es fundamental que los colores básicos no compitan por la atención. Estos colores de base deben ser monótonos pero que apoye bien el color de acento, es muy común cometer el error de usar colores de línea de base que son demasiado coloridos, de modo que terminan compitiendo por la atención del lector. Una alternativa fácil es usar un color neutro en todos los elementos de la figura, excepto para la categoría de puntos que se quiere resaltar.

La figura 2.15 muestra el total de estudiantes admitidos por departamento de nacimiento, se hace uso de una escala de color de acento para resaltar los departamentos pertenecientes a la región andina. Observe que se usa un color

neutro para los departamentos que no son de interés y un color azul llamativo para atraer la atención del usuario a la región andina. Esta figura presenta uno de los tantos usos que se le puede dar a las escalas de colores de acento para resaltar datos, es posible aplicarla a gráficos de líneas, diagramas de dispersión, entre otros.

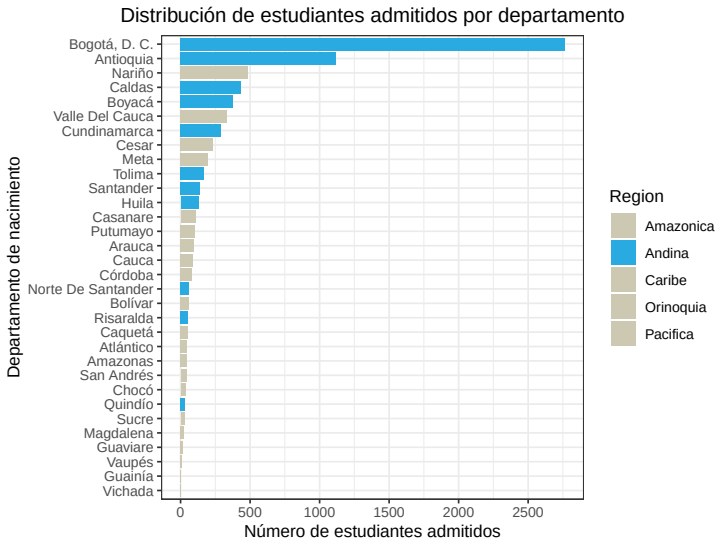

Figure 2.15: Escala de color de acento para resaltar observaciones

2.3 Texto

Al momento de realizar una visualización el texto es uno de los elementos a los cuales se les presta poca atención pero que podrían hacer el gráfico aún mas informativo, se debe manejar una misma fuente pero proporcionar un balance entre los tamaños con el fin de destacar los elementos importantes, por ejemplo, el título de la visualización debe atraer más la atención del usuario que las etiquetas del eje por esta razón no es correcto utilizar un tamaño mayor para las etiquetas del eje que para el título. El éxito de la estética de este elemento se basa en establecer correctamente la jerarquía que existe entre los textos que involucra el gráfico.

El objetivo principal de una visualización es informar y transmitir información de manera clara y concisa, por esta razón se deben colocar los datos en contexto proporcionando títulos, subtítulos y otras anotaciones que los acompañen. A continuación, analizaremos algunas recomendaciones utiles que nos ayudarán a contextualizar los datos de manera correcta.

Uno de los elementos principales dentro de un gráfico es su título, este debe ser claro e informativo ya que su función es transmitir con precisión al lector de qué se trata la figura, también existen ocasiones en las que es necesario usar subtítulos para contextualizar por completo al usuario acerca de la información que se presenta. Otro elemento importante y que logra que las visualizaciones se expliquen por sí mismas son los títulos de los ejes, estos deben indicar de manera clara lo que representan y la unidad en que se miden, observe la figura 2.3 en la cual el eje Y esta titulado de manera correcta ya que indica que representa las ventas y esta medido en dólares; los títulos de las leyendas también deben ser claros e indicar lo que representan como se muestra en la figura 2.15, donde el título de la leyenda hace referencia a la región geográfica del país, en algunas ocasiones es posible omitir el título de los ejes o de las leyendas, es decir, cuando las etiquetas son completamnete explícitas.

La figura 2.16 ilustra lo que no se debe hacer en una visualización ya que se le da enfoque princical a las etiquetas del eje y leyenda y no al título de la figura, el cual informa acerca de la visualización.

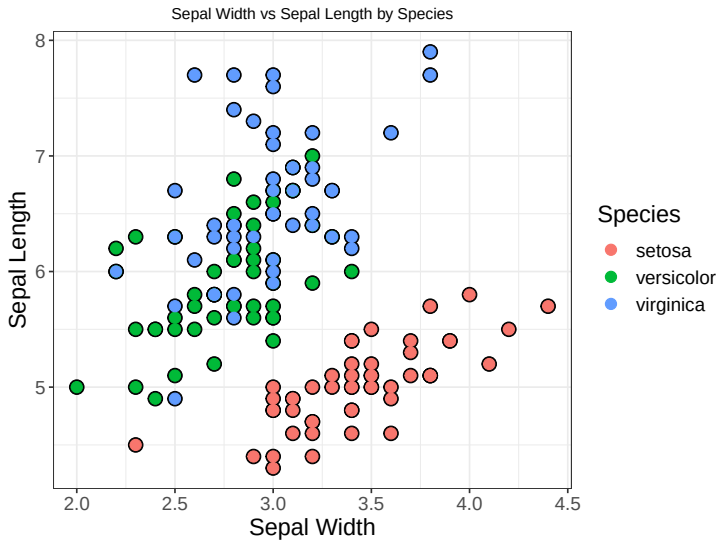

Figure 2.16: Uso incorrecto del texto

La figura 2.17 presenta la clara jerarquía existente entre los textos del gráfico, observe que se trabaja con la misma fuente pero se juega con mayúsculas y minúsculas para dar la importancia que cada elemento requiere, esta visualización es un ejemplo en el cual se puede omitir el título de la leyenda, ya que las etiquetas son tan claras que usar un título conlleva a un gráfico saturado.

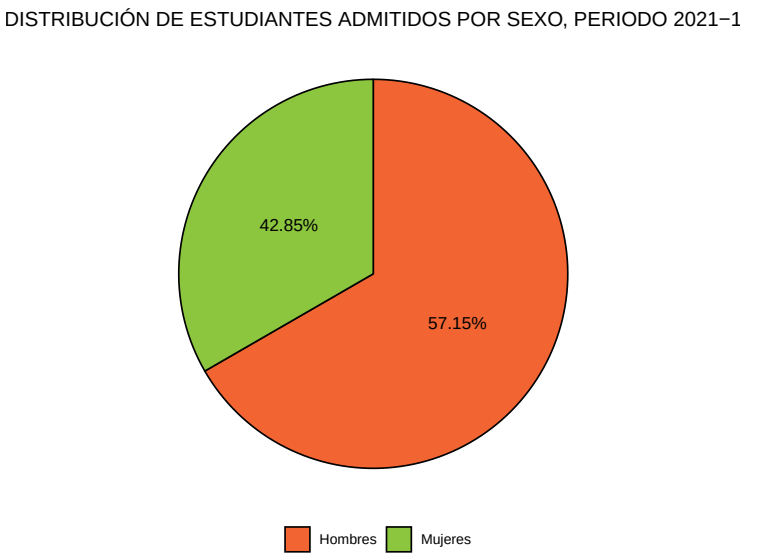

Figure 2.17: Uso correcto del texto para contextualizar

Chapter 3

¿Cómo visualizar los datos?

En la actualidad hay muchos gráficos disponibles para visualizar nuestros datos, pero no todos los gráficos pueden ser usados para lo que se quiere representar, por esta razón es importante conocer cuáles son los gráficos apropiados para los datos que se quieren mostrar. Dentro de los datos más comunes a visualizar se tienen las cantidades, proporciones, distribuciones, series de tiempo, datos geoespaciales, relaciones X-Y e indicadores hacia una meta en particular.

En este capítulo se presenta la forma correcta de representar estos datos con los gráficos que se tienen disponibles, teniendo el contraste entre lo correcto e incorrecto con el fin de informar al usuario acerca de lo que se debe o no hacer para crear las visualizaciones.

3.1 Visualización de cantidades

En muchas ocasiones estamos interesados en visualizar la magnitud de algún conjunto de números, por ejemplo, visualizar el volumen de ventas por estados, total de estudiantes admitidos por modalidad de formación, estudiantes graduados por grupos de edad o departamento, entre muchos otros ejemplos. Observe que en todos estos casos se tiene un conjunto de categorías y un valor cuantitativo para cada categoría. La visualización recomendada y más usada en este escenario es el gráfico de barras en el cual se incluyen distintas variaciones tales como las barras simples, agrupadas y apiladas tanto verticales como horizontales. Las alternativas al diagrama de barras son los diagramas de puntos y mapas de calor.

3.1.1 Gráfico de barras

Suponga que queremos visualizar la cantidad de estudiantes admitidos por nivel de formación para el periodo 2021-1, este tipo de datos se visualiza comúnmente con barras verticales, para cada nivel de formación se dibuja una barra que inicia en cero y se extiende hasta la cantidad de estudiantes admitidos. La figura 3.1 muestra el uso del gráfico de barras para visualizar estas cantidades.

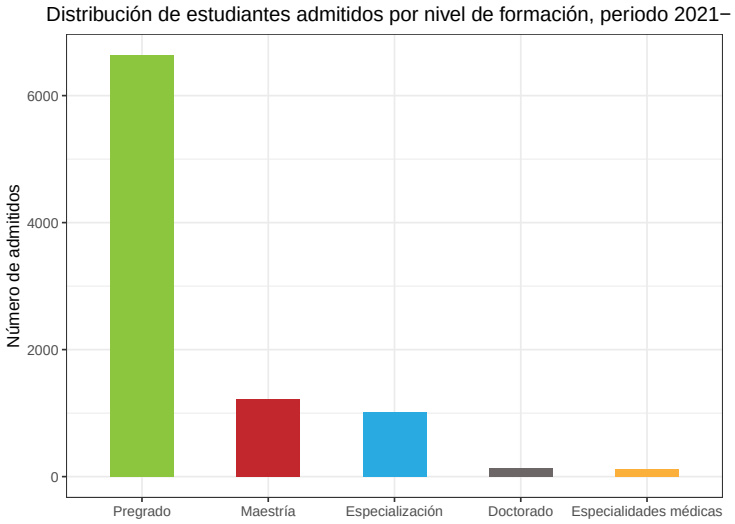

Figure 3.1: Uso del gráfico de barras para representar cantidades

Uno de los problemas más comunes con los gráficos de barras verticales es que las etiquetas que identifican cada barra ocupan mucho espacio horizontal. Por esta razón en la figura 3.1 fue necesario aumentar la separación entre las barras para poder ubicar las etiquetas en la parte inferior de cada barra y que estas no se traslaparan. Una solución a este problema es girar las etiquetas de cada barra como se muestra en la figura 3.2, pero esto no es estéticamente correcto, ya que es incómodo para el usuario.

La mejor solución para etiquetas largas es cambiar a un gráfico de barras horizontales, de esta manera no será necesario aumentar el espaciado entre barras ni girar las etiquetas y se obtiene una visualización compacta en la cual todos los elementos visuales están ubicados de manera horizontal y hace que el gráfico sea más fácil de leer y comprender.

Sin importar la posición de las barras es decir si son horizontales o verticales se debe prestar mucha atención al orden en el cual se ubica cada barra, en muchas ocasiones las barras están dispuestas de forma arbitraria o por algún criterio que no es significativo en el contexto de la figura, algunos programas simplemente ubican las etiquetas por orden alfabético o algún otro criterio.

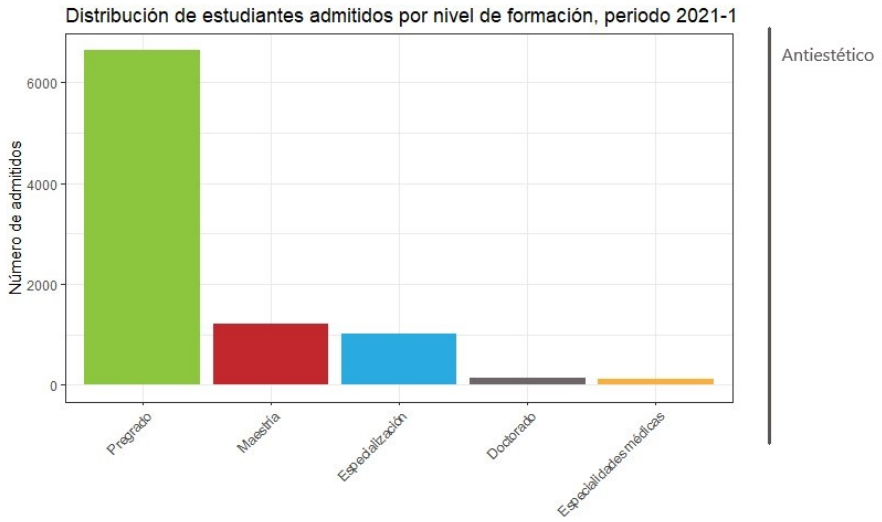

Figure 3.2: Uso del gráfico de barras para representar cantidades, girando etiquetas

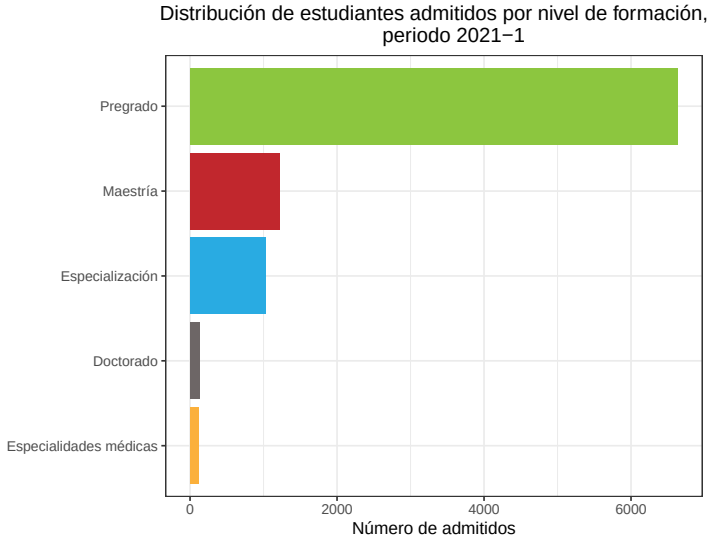

Figure 3.3: Uso del gráfico de barras horizontales para representar cantidades

Sin embargo, las etiquetas solo pueden ser reordenadas cuando las categorías que representan no tienen un orden natural establecido. Siempre que exista un orden natural en los datos es necesario mantener este orden para representar los datos de la manera correcta. Suponga que se desea visualizar la cantidad de estudiantes graduados en el periodo 2020-2 por grupos de edad. En este caso las barras deben ordenarse de manera creciente según el grupo de edad como se ilustra en la figura 3.4. En este caso no tiene sentido ordenar por la altura de la barra, es decir de forma ascendente o descendente ya que las etiquetas perderán el orden natural que poseen.

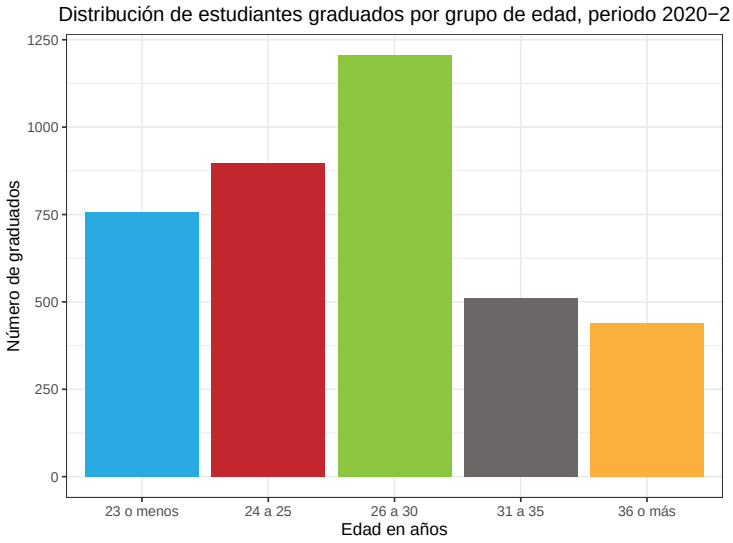

Figure 3.4: Etiquetas con un orden natural

3.1.2 Barras agrupadas y apiladas

Las figuras mostradas en la subsección anterior representan variables cualitativas en relación con una variable categórica. Sin embargo, a menudo es de interés visualizar como estos valores varían según dos variables categóricas; en un diagrama de barras apiladas, se dibuja un grupo de barras en cada posición del eje X, determinado por una variable categórica, y luego se dibujan barras dentro de cada grupo con la otra variable categórica de interés, por ejemplo, es posible representar el número de funcionarios adminsitrativos por años de servicio prestado y sexo para el periodo 2020-2 tal y como se ilustra en la figura 3.6.

Con los gráficos de barras agrupadas se debe tener cuidado ya que las variables categóricas elegidas pueden tener muchos niveles y harán que le gráfico se sature y el usuario no comprenda la información de manera correcta, la figura 3.7

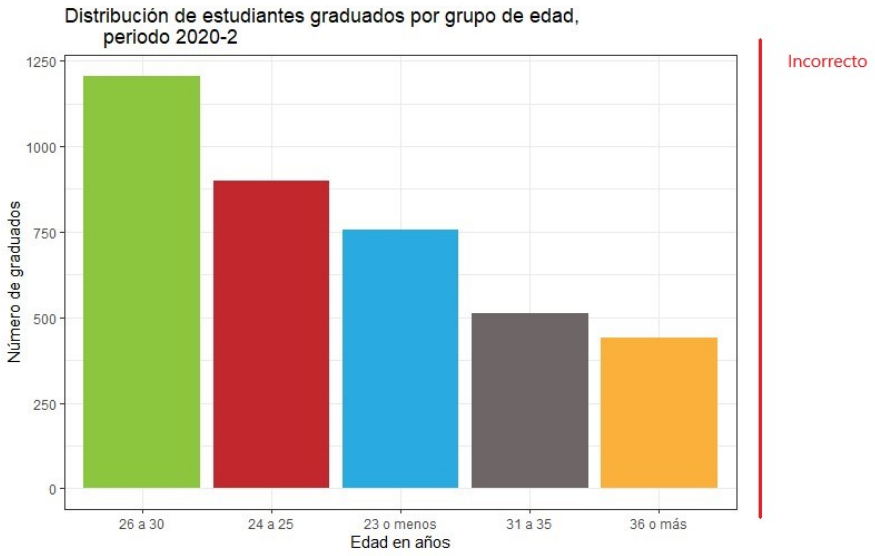

Figure 3.5: Reordenar etiquetas cuando se tiene un orden natural

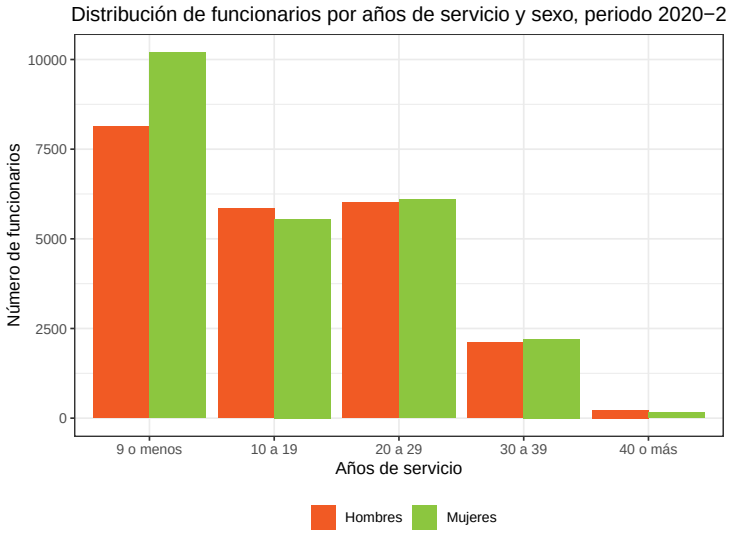

Figure 3.6: Uso de barras agrupadas para representar valores usando dos categorías

muestra la distribución del número de funcionarios por años de servicio y sede para el periodo 2020-2, aunque la figura es correcta resulta difícil de interpretar por la cantidad de sedes existentes dentro de la Universidad Nacional de Colombia, observe que para cada grupo de años de servicio se crean diez barras.

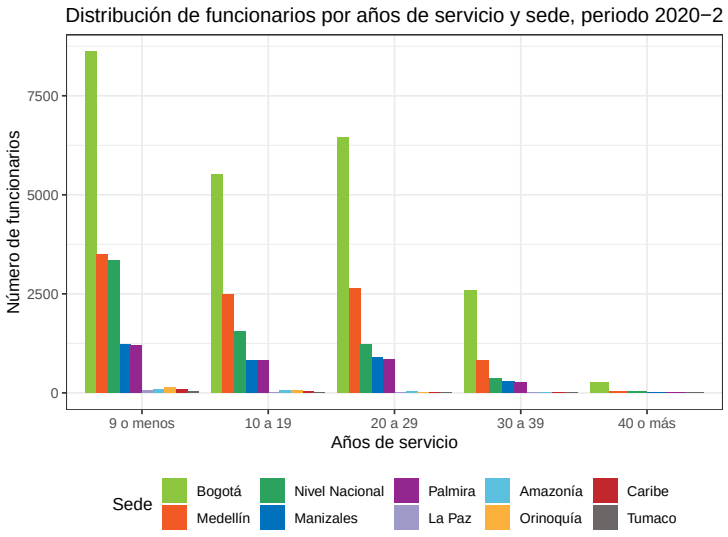

Figure 3.7: Cantidad de funcionarios por años de servicio y sede

Una alternativa a esta figura es visualizar como categoría principal la sede a la que pertenece cada funcionario y que los grupos de barras para cada sede se creen a partir de los años de servicio prestados, usando una escala de color secuencial para representar cada uno de los años de servicio prestado por los funcionarios administrativos, con esto solo se tendrán 5 barras por cada sede y el gráfico será más sencillo y comprensible, como se muestra en la figura 3.8.

Como ya vimos los gráficos de barras agrupadas consisten en dibujar una barra al lado de la otra, pero hay ocasiones en las cuales se prefiere apiliar las barras, es decir, ubicar una encima de la otra, esto generalmente se realiza cuando la cantidad representada por las barras dispuestas en esta posición es significativa, también es necesario tener cuidado con este tipo de gráficos ya que utilizar muchas categorías para apilar las barras resultara en una visualización saturada y poco informativa. Un uso común de este tipo de gráficos es cuando las barras individuales representan recuentos, por ejemplo, el conjunto de datos llamado Administrativos posee el recuento de funcionarios por sexo, para este caso si apilamos una barra que representa el recuento de mujeres encima de una barra que representa el recuento de hombres, entonces la altura de la barra combinada mostrara el total de funcionarios independiente del género. La figura 3.9 presenta el uso de barras apiladas como alternativa del uso de barras agrupadas ilustrado en la figura 3.6.

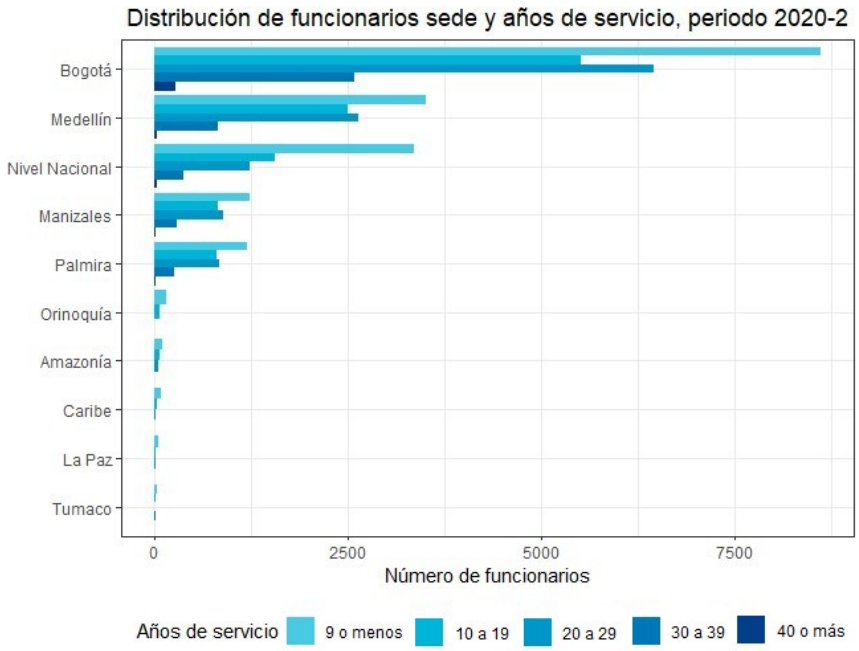

Figure 3.8: Edad promedio de estudiantes graduados

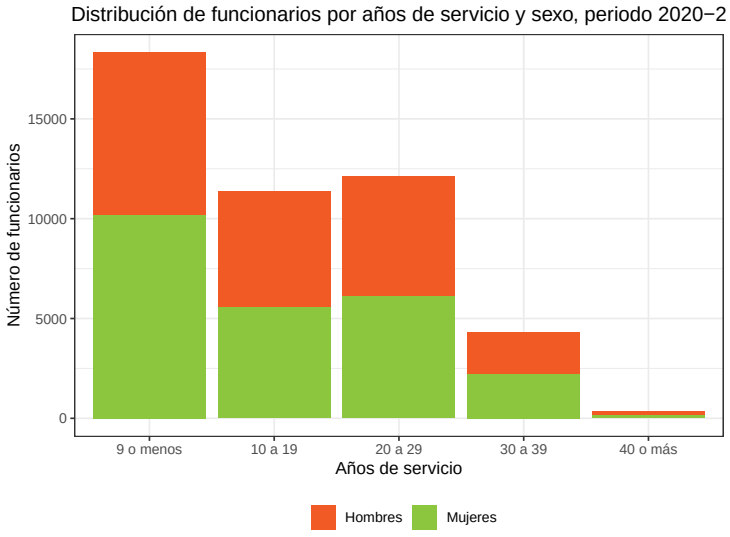

Figure 3.9: Barras apiladas como una alternativa a las barras agrupadas

3.1.3 Gráficos de puntos y mapas de calor

Una de las mayores limitaciones al usar los gráficos de barras ya sean simples o alguna de sus variaciones es que los ejes deben iniciar en cero para lograr que la altura de la barra sea proporcional a la cantidad que representa, existen muchas ocasiones en las cuales es poco práctico iniciar siempre los ejes en cero y una alternativa es usar puntos ubicados en lugares apropiados a lo largo del eje X o Y.

Suponga que se quiere visualizar la edad promedio de los estudiantes graduados por nivel de formación, observe que en este caso no tiene mucho sentido iniciar el eje Y en cero ya que la edad promedio estará por encima de los 20 años aproximadamente y las discrepancias entre los promedios de edades no son muy grandes, por esta razón es conveniente restringir el dominio del eje x al intervalo de 20 a 40 años, para que las diferencias sean notorias y la información sea interpretada con mayor facilidad, como se ilustra en la figura 3.10.

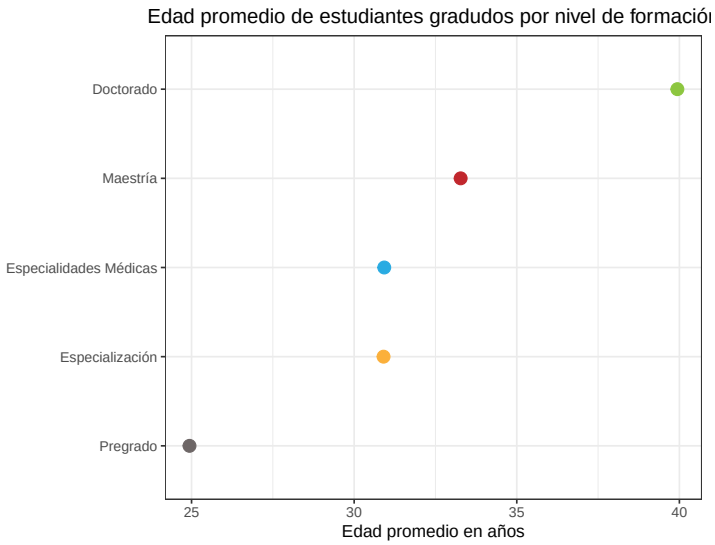

Figure 3.10: Diagrama de puntos como alternativa a los gráficos de barras

La figura 3.11 muestra la edad promedio de los estudiantes graduados por nivel de formación usando un gráfico de barras, esta figura fue etiquetada como incorrecta ya que al usar barras el eje debe iniciar en cero y las discrepancias existentes entre las edades promedios de los niveles especialidades médicas y especialización no se logra apreciar con claridad.

Para este tipo de visualización también aplica reordenar las categorías por la variable numérica si las categorías no tienen un orden natural, este grafico es recomendable cuando no se tiene un orden natural en los datos ya que pueden ser reordenados logrando una visualización atractiva y entendible.

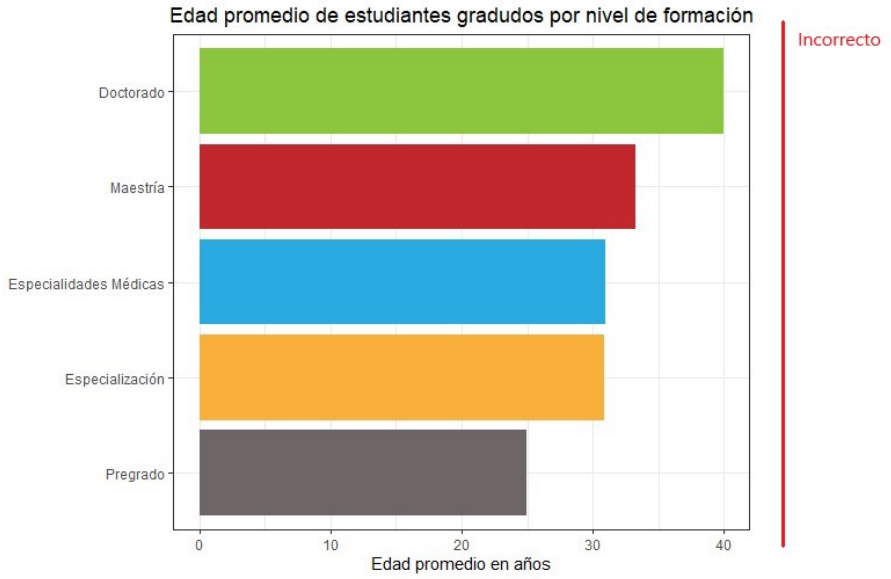

Figure 3.11: Edad promedio de estudiantes graduados

Las visualizaciones presentadas hasta el momento representan variables numéricas usando categorías a través de las posiciones en uno de los ejes, ya sea con un punto final en el valor que representa o el tamaño de la barra. Sin embargo, estos gráficos no son adecuados cuando el conjunto de datos muy grandes, ya que la cantidad de barras será excesiva y proporcionará una figura saturada. Ya se había observado en la figura 3.7 que grupos de 8 barras resultan en una visualización compleja y no tan fácil de interpretar, imagine que en lugar de 8 barras por grupo se tuvieran 20 o más, resultaría aun más complejo y probablemente muy confusa.

Una alternativa a las barras y puntos son los mapas de calor, los cuales están formados por dos variables categóricas, una en cada eje, y los valores son representados a través de una escala de color secuencial. La figura 3.12 utiliza este enfoque para mostrar el número de accidentes ocurridos por día y mes en los años 2014 a 2016, es una figura útil para detectar tendencias más amplias que para visualizar exactamente cada uno de los valores que representa, se identifica con claridad que los primeros 15 días del mes de enero se presenta menor accidentalidad comparado con los días restantes de este mismo mes.

La figura 3.12 es informativa pero incorrecta ya que no se respeta el orden natural que poseen los datos, la figura correcta para visualizar los datos usando un mapa de calor se presenta a continuación.

Como ya se ha mencionado en múltiples ocasiones se debe prestar mucha atención al orden de las variables categóricas, en este caso dichas variables presentan

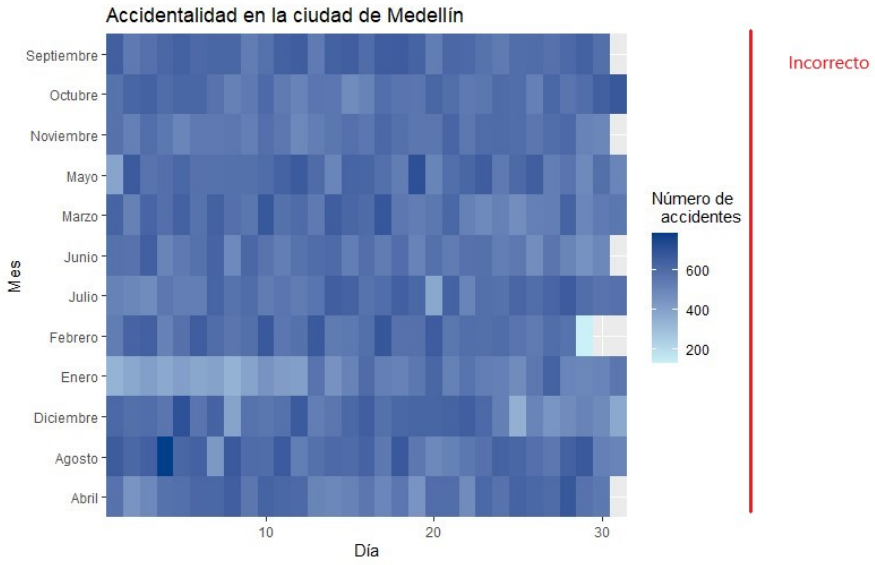

Figure 3.12: Mapa de calor para representar la accidentalidad en la ciudad de Medelllín

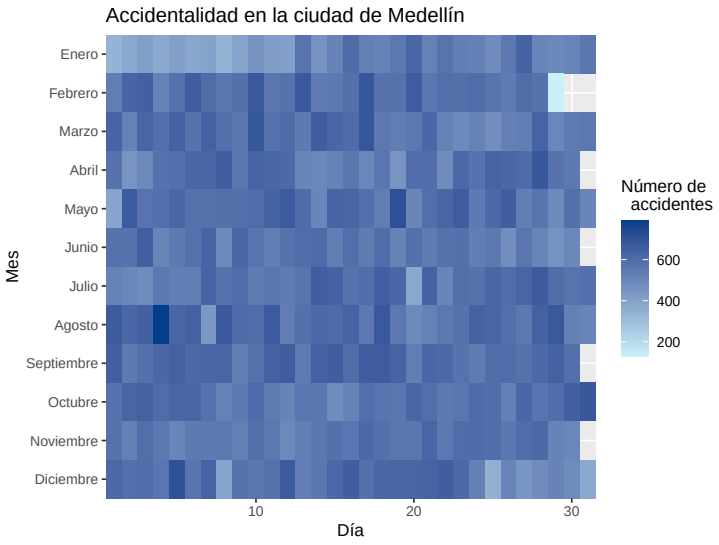

Figure 3.13: Mapa de calor para representar la accidentalidad en la ciudad de Medelllín

un orden natural ya que son meses y días, en el caso de no presentarse un orden natural podría reordenarse las categorías usando la variable numérica para lograr una visualización mas clara, atractiva y estéticamente llamativa.

3.2 Visualización de proporciones

Cuando se están analizando datos se presentan ocasiones en las cuales se quiere representar como algún grupo se divide en partes individuales que representan un porcentaje o proporción de un todo, entre los ejemplos más comunes se incluyen las proporciones de hombres y mujeres dentro de un grupo de personas, el porcentaje de estudiantes graduados por modalidad de formación, el porcentaje de estudiantes admitidos por facultad en cada una de las sedes de la Universidad Nacional de Colombia, entre muchos otros ejemplos.

La visualización más utilizada en este caso es el gráfico circular, gráfico odiado y amado por muchos, el éxito de este gráfico depende principalmente de la correcta elección de grupos y colores. Cuando el todo se divide en muchas partes diferentes o cuando queremos ver los cambios en las proporciones a través de una ventana de tiempo mostrar las proporciones se convierte en un verdadero reto y no hay una visualización única que funcione en todos los casos.

Generalmente para mostrar proporciones se usan gráficos circulares, barras apiladas y barras una al lado de la otra, dependiendo del propósito de la visualización se elige cada una de las opciones mencionadas.

3.2.1 Gráficos circulares

Un gráfico circular está compuesto de una variable discreta que determina la cantidad de divisiones del circulo y adicionalmente una variable numérica con la cual se generan los fragmentos y se debe lograr que cada uno sea proporcional a la fracción del total que representa.

Dentro de los datos almacenados para los estudiantes admitidos a la Universidad Nacional se encuentra el estrato socioeconómico el cual está dividido o categorizado en tres grupos, en este caso es posible usar un gráfico circular como se ilustra en la figura 3.14, se observa que del total de estudiantes admitidos el 50.91% pertenece al grupo de estrado 2 o menos, el 32.88% al estrato 3 y el restante a la categoría de estrato 4 o más.

La figura anterior se considera como un uso correcto de los gráficos circulares ya que estos son recomendados para visualizar los datos como proporciones de un todo y enfatizar visualmente en fracciones simples como la mitad, un tercio o un cuarto, es decir, funciona correctamente cuando la cantidad de divisiones no es mayor a cuatro o cinco y es posible utilizarlo en conjuntos de datos pequeños.

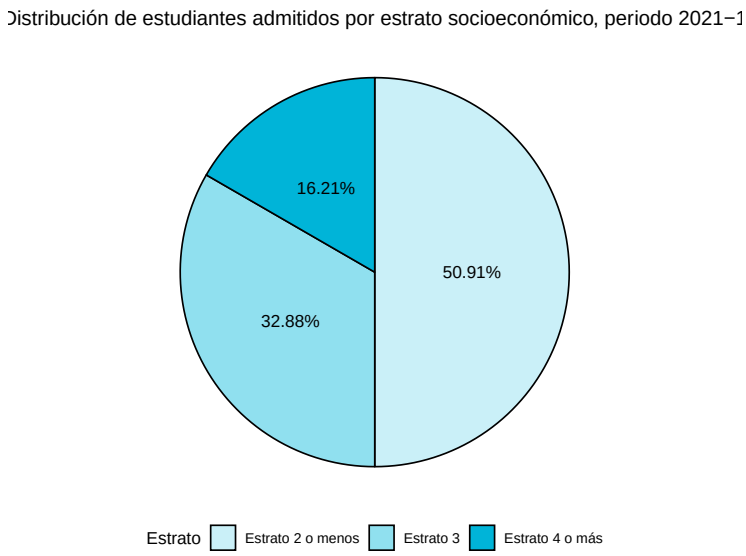

Figure 3.14: Gráfico circular para representar 3 categorías

Cuando se tienen variables con muchas categorías no se recomienda utilizar los gráficos circulares ya que no se logra enfatizar en las proporciones y las etiquetas de cada categoría no se observan con claridad, la figura 3.15 se considera un uso incorrecto de los gráficos circulares ya que la variable discreta utilizada tiene nueve categorías, para esta ocasión se prefiere un gráfico de barras, aunque se pierdan algunas características como la relación con el total.

Existe un caso especial en el cual los gráficos circulares fallan, suponga que se quiere mostrar el cambio en porcentaje para categorías de productos en cierta empresa de 2017 a 2019, los productos son clasificados en tres grupos Furniture, Technology y Office Supplies para representar los porcentajes de ventas de estas categorías se usa un gráfico circular como se ilustra en la figura 3.16, cuando se utiliza este tipo de visualización es difícil observar con claridad lo que está sucediendo exactamente. Parece que las ventas en la categoría Office Supplies se incrementaron para el año 2018, mientras que las ventas de Tecnología fueron menores en ese mismo año, pero más allá de estas observaciones no podemos describir con claridad lo que esta pasando. En particular, no está claro como se comparan exactamente las ventas de las diferentes categorías dentro de cada año, adicionalmente los cambios en la participación de los grupos de productos en las ventas a lo largo de años son difíciles de observar.

3.2.2 Barras apiladas

Las barras apiladas se recomiendan cuando la intención del grafico es visualizar claramente los datos como proporciones de un todo, para visualizar muchos

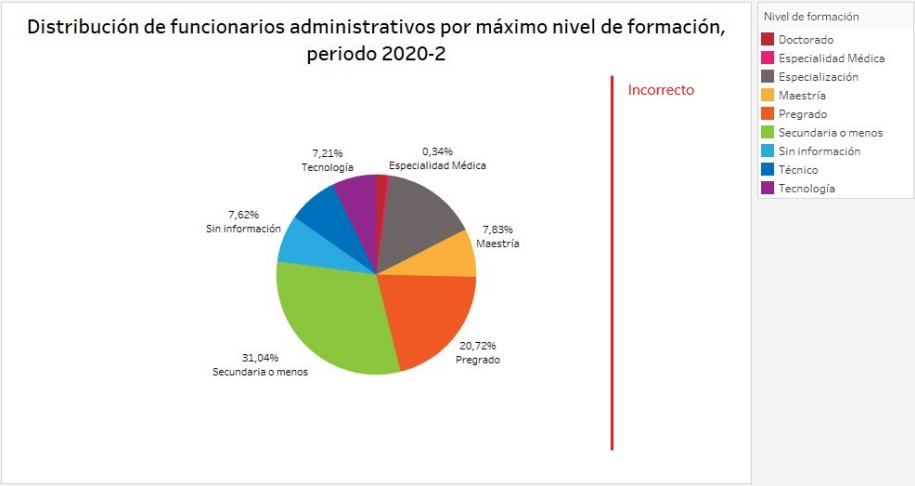

Figure 3.15: Uso incorrecto de los gráficos circulares

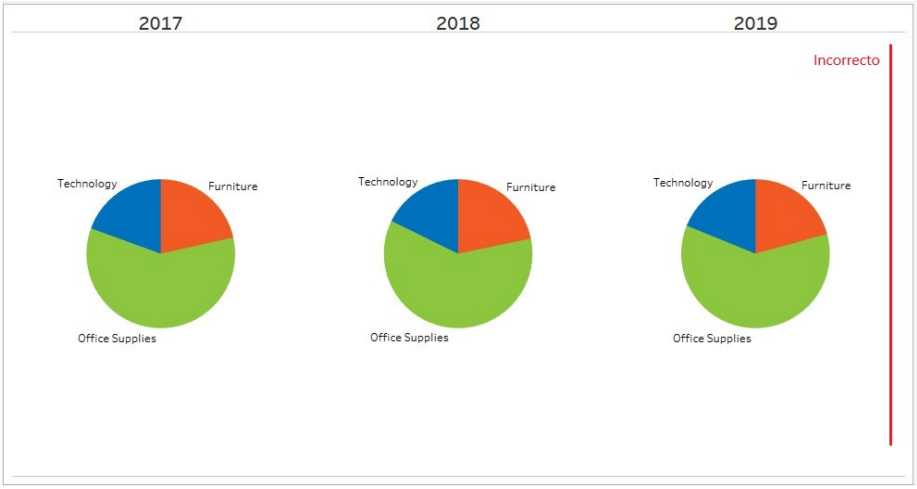

Figure 3.16: Participación de productos sobre ventas generales para tres años

conjuntos de proporciones o representar series temporales de proporciones.

Con las barras apiladas sucede algo similar a los gráficos de torta ya que tener muchas categorías saturan la visualización y la hacen poco informativa, nuevamente es recomendable para representar no más de 4 categorías y cuidar siempre el orden de cada grupo dentro de cada barra cuando se representan varios conjuntos de proporciones o se tiene una serie temporal de proporciones.

La figura 3.17 utiliza el enfoque de barras apiladas verticalmente para visualizar la distribución de estudiantes admitidos por estrato socioeconómico para el periodo 2021-1, a pesar de ser una figura informativa se etiqueta como antiestética ya que tener una única barra no es visualmente atractivo, para representar esta información en particular se prefiere un gráfico circular.

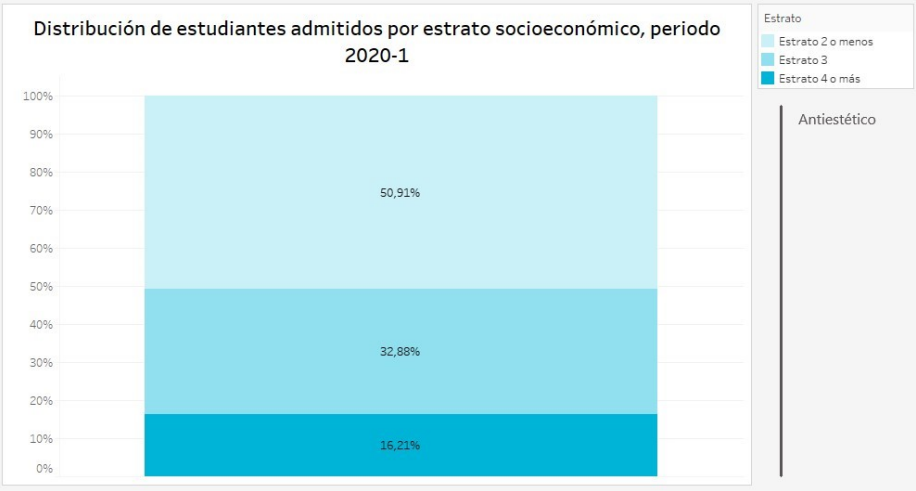

Figure 3.17: Barra apilada verticalmente para representar estratos socioeconómicos

Anteriormente se menciono que las barras apiladas son una de las visualizaciones sugeridas cuando la intención es representar series temporales de proporciones, a pesar de esto hay un caso en el cual este tipo de gráfico falla y es cuando se tienen demasiadas categorías y están desporporcionadas. Suponga que queremos visualizar la cantidad de estudiantes graduados por nivel de formación en los últimos tres años usando el enfoque de barras apiladas como se ilustra en la figura 3.18.

La figura anterior permite visualizar los datos como partes de un todo, pero al tener datos desproporcionados no es clara la contribución que tiene cada nivel de formación sobre el total de estudiantes graduados, un poco más del 50% de estudiantes graduados pertenecen a pregrado, cerca del 25% a maestría, alrededor del 20% a especialización y los aportes de especialidades médicas y doctorado no son claras ya que son muy pequeñas, por esta razón dicha figura

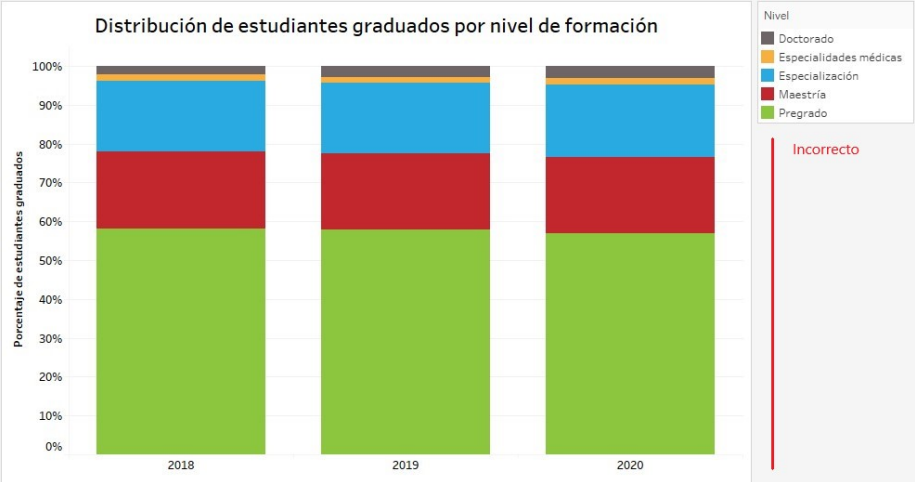

Figure 3.18: Barras apiladas para nivel de formación

fue etiquetada como incorrecta.

Un caso útil y en el cual realmente se explota el potencial de los gráficos de barras apiladas es cuando se representan datos con solo dos categorías ya que los cambios a través del tiempo serán mucho más notorios y claros o incluso si se tratara de conjuntos de barras apiladas y no series temporales, la figura 3.19 permite visualizar la distribución de estudiantes admitidos por sexo desde el año 2008 hasta 2020, esta visualización permite identificar claramente las contribuciones de cada uno de los sexos en el total de estudiantes admitidos y los cambios en el tiempo también son claros a pesar de que las proporciones se mantienen estables.

Como se mencionó anteriormente los gráficos de barras apiladas también son útiles para representar conjuntos de proporciones en el caso de usar este enfoque es necesario tener dos variables categóricas una que marcara las divisiones de cada barra y la otra será usada en uno de los ejes para generar tantas barras como categorías se tengan. Suponga que queremos visualizar la distribución de estudiantes admitidos por modalidad formación y sexo, para realizar utilizaremos el enfoque de barras apiladas horizontalmente como se muestra en la figura 3.20.

La visualización 3.20 permite identificar con claridad como se comporta la proporción de hombres y mujeres admitidos en los niveles de formación ofrecidos por la Universidad Nacional de Colombia, se identifica que en pregrado cerca del 30% de las personas admitidas son mujeres mientras que para los niveles de especialización la participación de las mujeres aumenta, llegado alrededor de 50%.

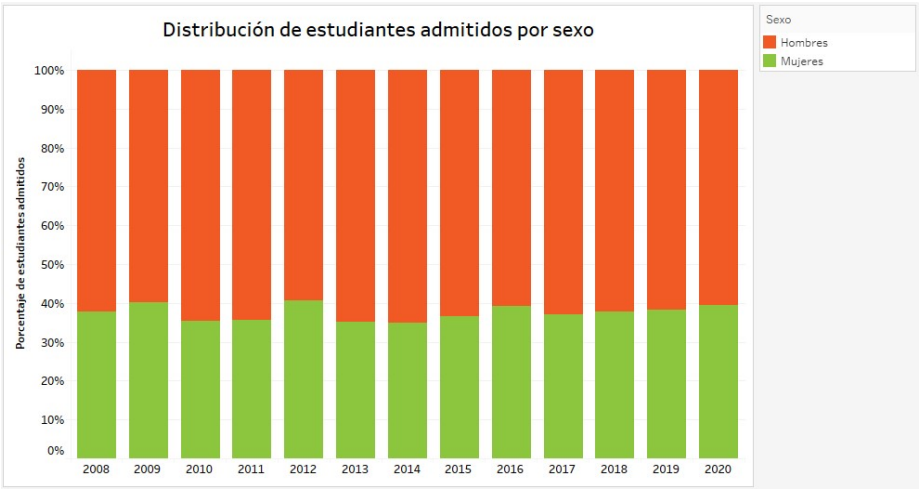

Figure 3.19: Barras apiladas para nivel de formación

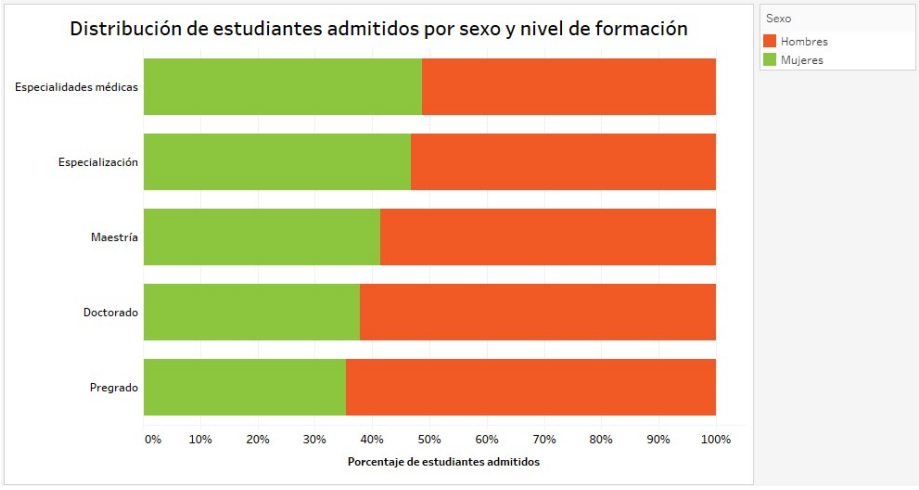

Figure 3.20: Barras apiladas para nivel de formación

3.2.3 Barras agrupadas

Las barras dispuestas una al lado de la otra permite una fácil comparación visual de las proporciones, es informativo incluso cuando el conjunto de datos es pequeño y si los datos a mostrar se dividen en muchas partes este es el grafico indicado para visualizar las proporciones, también es uno de los más recomendados y utilizados cuando se quiere ilustrar series temporales de proporciones.

Considere nuevamente la figura 3.18, en la cual se intento ilustrar la distribución de estudiantes graduados por nivel de formación desde el año 2018 hasta 2020 y esa figura fue etiquetada como incorrecta por que no se lograba observar con claridad las contribuciones de los niveles doctorado y especialidades médicas. Ahora representaremos la misma información, pero con el enfoque de barras agrupadas, es decir tendremos tres grupos de barras donde cada grupo tiene cinco barras que representan cada uno de los niveles de formación ofrecidos en la Universidad.

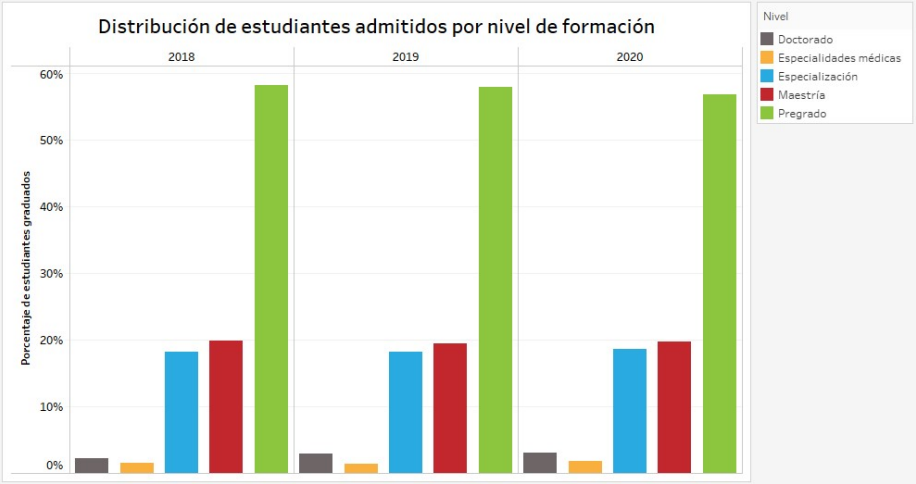

Figure 3.21: Barras apiladas para nivel de formación

La figura 3.21 permite visualizar con claridad las contribuciones de cada uno de los niveles de formación sobre el total de estudiantes graduados, los niveles de especialización y maestría generan alrededor del 20% cada una de estudiantes graduados por año, mientras que las contribuciones de doctorado y especialidades médicas están alrededor del 2% para cada uno de estos niveles. Adicional a poder comparar con claridad estos porcentajes también se puede observar la evolución anual de la proporción de estudiantes graduados por nivel de formación, para los años 2019 y 2020 la cantidad de estudiantes graduados de doctorado fue mayor que para el año 2018, también se observa que la cantidad de graduados de pregrado es similar a lo largo de los tres años representados.

En la sección 3.1 se hacen algunas recomendaciones acerca de la ubicación de

etiquetas, ejes y otras que también se aplican para esta sección con el fin de lograr visualizaciones claras, informativas y atractivas.

3.3 Series de tiempo

La visualización de series de tiempo puede verse como un caso especial de los diagramas de dispersión, en el cual una de las dos variables puede considerarse como tiempo ya que el tiempo impone una estructura adicional a los datos. En estos casos los puntos tienen un orden inherente y es posible ordenar las observaciones de manera creciente en el tiempo y definir un antecesor y sucesor para cada punto. Generalmente la visualización de las series de tiempo se realiza usando gráficos de líneas, sin embargo, este tipo de gráficos no se limitan únicamente a esta información, son útiles siempre que una variable impone un orden a los datos.

3.3.1 Series de tiempo individuales

Como primer ejemplo de una serie de tiempo, consideremos el patrón de estudiantes graduados en la Universidad Nacional a nivel de cifras generales, estas admisiones se presentan dos veces al año, por tanto, el eje temporal de nuestra visualización se conforma del año y el semestre al que fue admitido el estudiante. Podemos visualizar el comportamiento en la cantidad de estudiantes admitidos haciendo una forma de diagrama de dispersión donde dibujamos los puntos que representan el número de admitidos en cada semestre, como se presenta en la figura 3.24.

Sin embargo, existe una importante diferencia con los diagramas de dispersión que se analizan en la sección *referencia*. En la figura 3.24 los puntos están espaciados uniformemente a lo largo del eje X y existe un orden específico definido por la línea temporal entre ellos. Cada punto tiene exactamente un vecino a la izquierda y uno a la derecha, excepto los puntos de inicio y final que solo tienen un vecino. Una manera de enfatizar visualmente en este orden presentado por los datos es conectando las observaciones con líneas como se ilustra en la figura 3.23 y este tipo de grafico se denomina gráfico de líneas.

A pesar de que muchos científicos de datos indican que esta práctica es incorrecta ya que las líneas no representan datos observados y de alguna manera corresponden a datos inventados, usar líneas para conectar los puntos ayuda con la percepción cuando los puntos están muy separados o están ubicados de manera desigual.

Sin embargo, el uso de líneas para representar series de tiempo es un practica generalmente aceptada y, con frecuencia, los puntos se omiten por completo. Sin puntos la figura logra centrar la atención del usuario en la tendencia general de los datos y menos en las observaciones individuales adicionalmente una figura sin

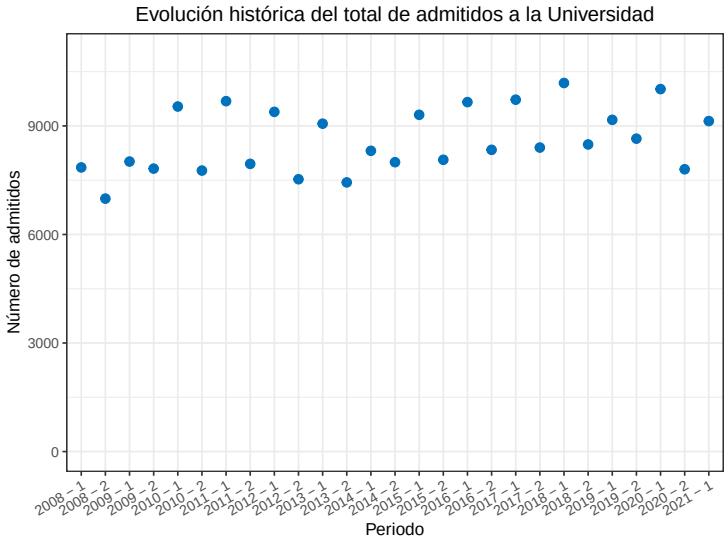

Figure 3.22: Gráfico de dispersión para la evolución de admitidos

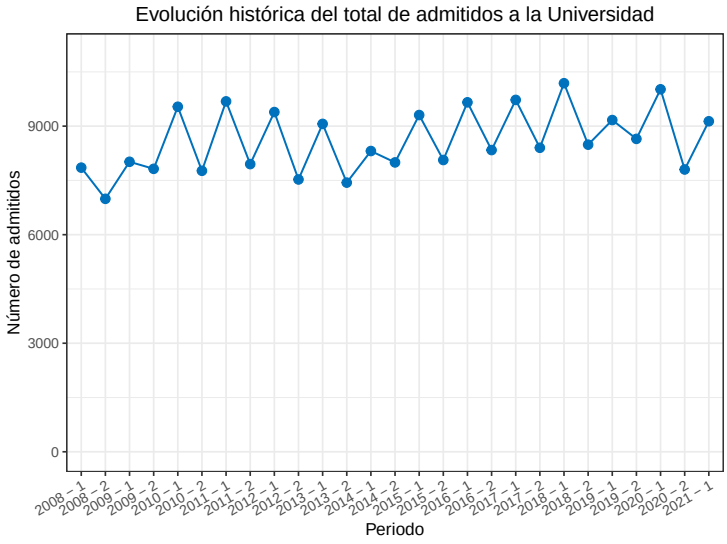

Figure 3.23: Gráfico de dispersión para la evolución de admitidos

puntos esta menos ocupada visualmente lo que indica claridad para el usuario. En general, cuanto mas densa es la serie de tiempo, es decir mas puntos tiene, se hace menos importante enfatizar en observaciones individuales usando puntos. Para el conjunto de datos de admisiones es posible omitir los puntos y lograr un gráfico más limpio. La siguiente visualización fue realizada usando Flourish, con el fin de incluir herramientas interactivas.

Otra opción bastante utilizada es rellenar el área bajo la curva con un color sólido, como se ilustra en la figura 3.24. Esta elección enfatiza aún más en la tendencia general de los datos ya que separa visualmente el área por encima de la curva del área que esta por debajo. Sin embargo, se debe tener cuidado con el inicio del eje Y, ya que si seleccionamos un valor de inicio diferente de cero el área bajo la curva representa la medida en que el valor de la observación supera el inicio del eje, mientras que el final de la sombra representa de manera correcta la cantidad de estudiantes admitidos.

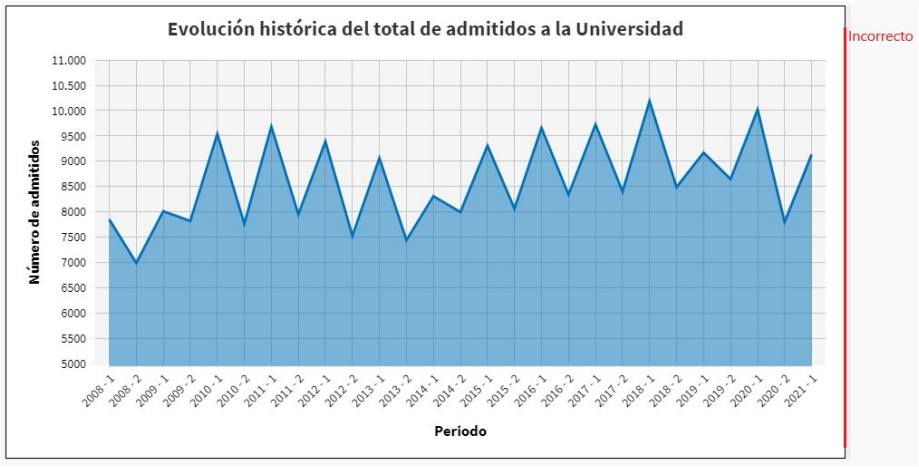

Figure 3.24: Gráfico de áreas para representar series de tiempo

Lo correcto para esta visualización es iniciar el eje Y siempre en cero, para lograr que el área sombreada y la línea representen exactamente la misma cantidad, que para este caso en particular corresponde a la cantidad de estudiantes admitidos por semestre.

3.3.2 Varias series de tiempo

Es muy común querer visualizar varios datos a través del tiempo, por lo general usando una variable categórica para diferenciar las observaciones. En estos casos se debe tener especial cuidado en como graficamos los datos ya que la figura puede tornarse confusa o difícil de interpretar. Suponga que queremos visualizar de manera semestral la cantidad de estudiantes graduados por nivel

de formación, un gráfico de dispersión usando colores para distinguir no es una buena idea porque los cursos del tiempo se encuentran entre sí.

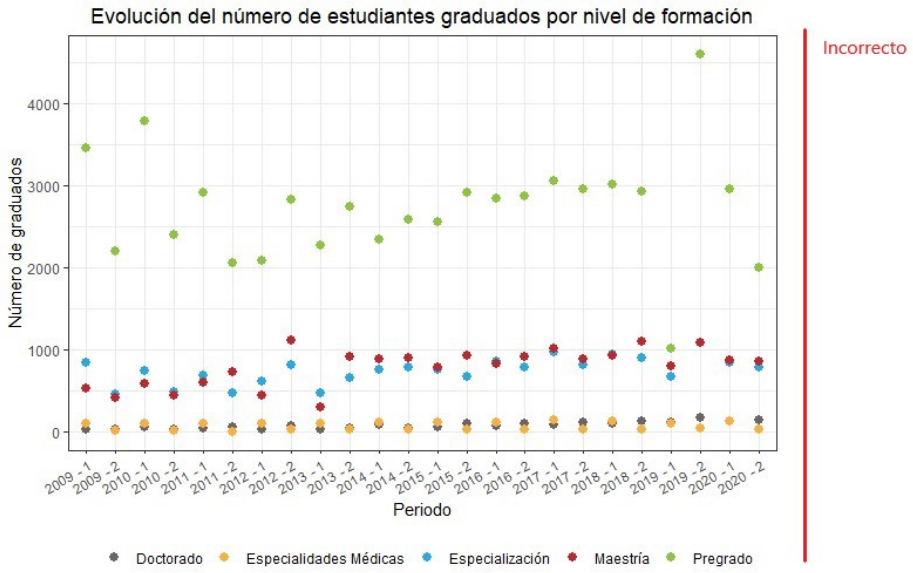

Figure 3.25: Uso incorrecto de los puntos para mostrar series de tiempo

La figura 3.25 es un ejemplo claro en el cual conectar los puntos mediante líneas es de vital importancia para generar visualizaciones claras y concisas que permiten una fácil interpretación por parte del usuario.

Los gráficos de líneas no se limitan únicamente a series de tiempo, son apropiados siempre que los puntos de datos tengan un orden natural que pueda ser reflejado en la variable que se ubica a lo largo del eje X, de modo que los puntos vecinos se puedan conectar con una línea. Este enfoque de gráfico de líneas puede ser usado para visualizar la cantidad de admitidos por rango de edad y nivel de formación.

La visualización anterior permite identificar claramente la cantidad de estudiantes admitidos por rango de edad para cada nivel de formación ofrecido por la Universidad, este enfoque se considera correcto ya que los rangos de edad a pesar de ser una variable discreta presentan un orden natural que se refleja claramente sobre el eje X.

3.4 Visualización de distribuciones

Cuando se realiza análisis de datos con frecuencia se presentan situaciones en las que nos interesa comprender como se distribuye una variable en particular

en un conjunto de datos, para estos casos el tipo de gráficos utilizado corresponde a histogramas y densidades. Es común tener tanto una como múltiples distribuciones, razón por la cual esta sección será dividida en visualización de una y varias distribuciones.

3.4.1 Visualización de una única distribución

Para esta sección usaremos los datos relacionados con la edad de estudiantes y administrativos ya que son variables numéricas y continuas para las cuales es posible visualizar la distribución. Como se mencionó anteriormente existen principalmente dos posibilidades gráficas las cuales son histogramas y densidades.

Los histogramas consisten en rectángulos llenos cuya altura corresponde a los conteos y los anchos corresponden al ancho de los contenedores, es importante resaltar que todos los contenedores deberán tener el mismo ancho para que la visualización represente un histograma válido. La figura 3.26 presenta el histograma correspondiente a la edad de los estudiantes graduados de la Universidad Nacional sin diferenciar por ningún tipo de característica como sexo o nivel de formación.

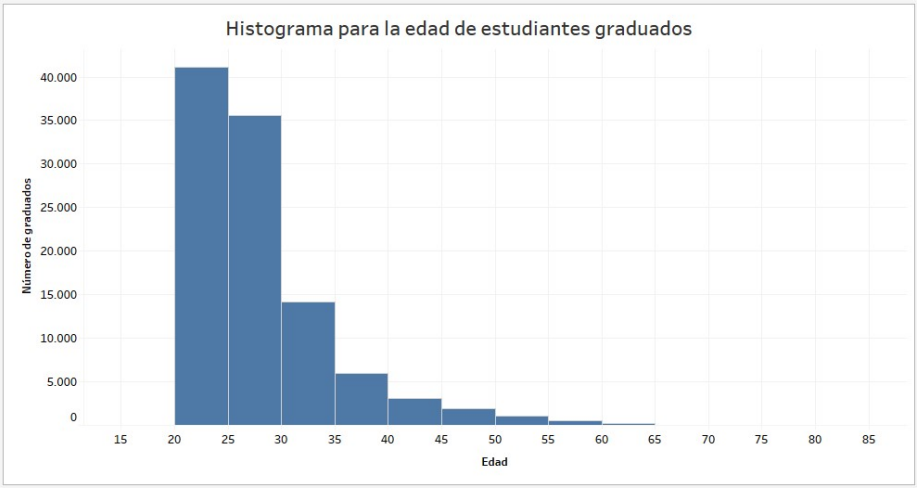

Figure 3.26: Histograma para la edad de estudiantes graduados

Los histogramas son generados a partir de una agrupación de datos por lo que su apariencia visual dependerá de la elección del ancho de los intervalos o contenedores, generalmente los programas de visualización que permiten graficar histogramas eligen el ancho del intervalo de manera predeterminada, pero es posible que este ancho no sea el más apropiado para la información que se quiere presentar. Por esta razón es fundamental siempre probar con diferentes

anchos de intervalo para verificar que el histograma resultante refleje los datos subyacentes con precisión. En general, si el ancho del contenedor es demasiado pequeño el histograma se vuelve demasiado puntual, visualmente ocupado y las principales tendencias en los datos se ocultan. Por otro lado, si el ancho del contenedor es demasiado grande, las características más pequeñas en la distribución de los datos tienden a desaparecer.

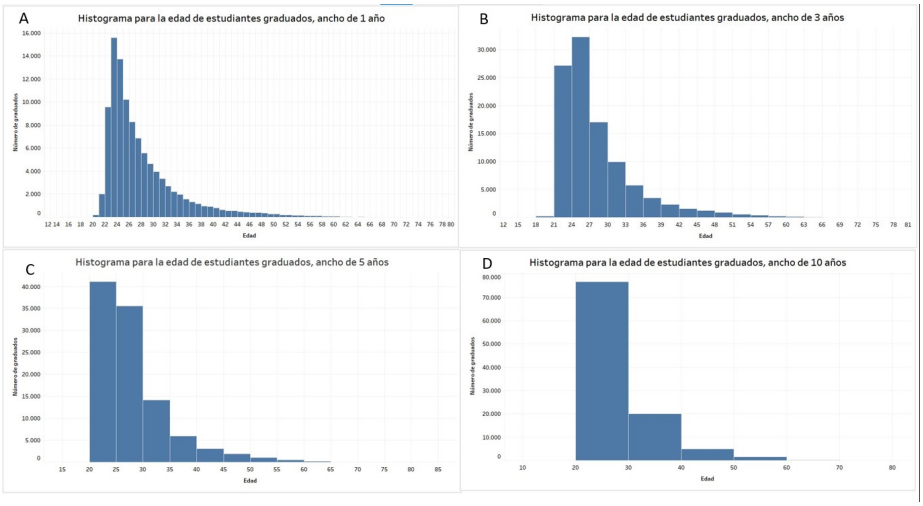

Figure 3.27: Histograma para la edad de estudiantes graduados, diferentes anchos de intervalo

Para la distribución de edades de los estudiantes graduados, se observa de la figura 3.27 que un ancho de intervalo de 1 año es demasiado pequeño y un ancho de intervalo de 10 años es demasiado grande, mientras que los anchos de contenedor entre 3 y 5 años funcionan bien. Es de resaltar que no existe una regla general que indique de manera exacta el ancho de intervalo correcto, es más una elección visual que permita ver las tendencias generales sin llegar a mostrar los datos de manera puntual pero tampoco agruparlos en intervalos muy grandes que oculten las características de la distribución.

A pesar de la popularidad de los histogramas como opción para la visualización de distribuciones por su facilidad de generación de manera manual, a medida que le tiempo avanza se desarrollan tecnologías que desarrollan los gráficos de densidad y han logrado opacar la popularidad de los histogramas. En una gráfica de densidad, lo que se intenta es visualizar la distribución de probabilidad subyacente de los datos dibujando una curva continúa apropiada. La figura 3.28 presenta el gráfico de densidad para la edad de los administrativos de la Universidad Nacional en el periodo 2020-2.

Esta curva de densidad debe estimarse a partir de los datos, y el método más utilizado para este procedimiento de estimación se denomina estimación de densidad de kernel, con este método de estimación se dibuja una curva continua (el

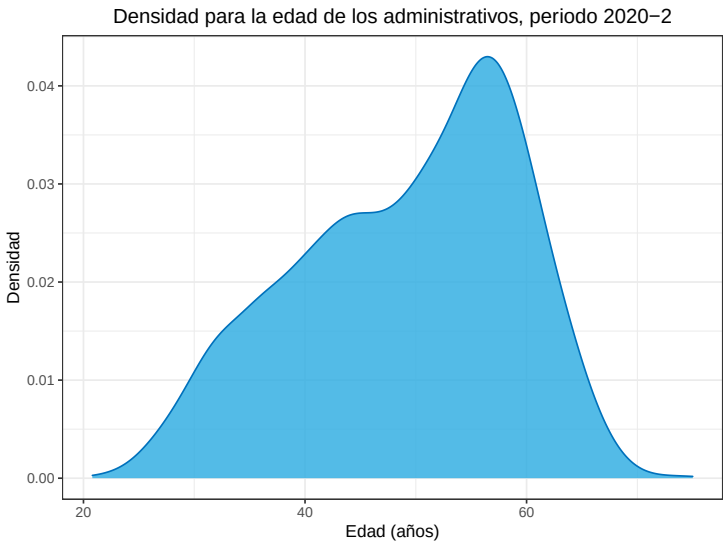

Figure 3.28: Densidad para la edad del personal administrativo, periodo 2020-2

kernel) con un ancho pequeño en la ubicación de cada punto de datos, y luego se suman todas estas curvas para obtener la estimación de la densidad final. El kernel más utilizado es el kernel Gaussiano pero hay muchas otras opciones.

Al igual que en el caso de los histogramas, la apariencia visual exacta de una gráfica de densidad depende de las opciones del kernel. Por ejemplo, un kernel gaussiano tendrá tendencia a producir estimaciones de densidad que parece gaussianas con características y colas suaves, por el contrario, un núcleo o kernel rectangular puede generar escalones en la curva de densidad. La figura *referencia* muestra la distribución de la edad de los administrativos que laboran en la Universidad usando un kernel rectangular, observe que al usar este tipo de kernel se pierde la suavidad en la curva.

En general, la elección del kernel no afecta de manera drástica el conjunto de datos cuando este tiene una cantidad grande de información, pero puede resultar en visualizaciones engañosas cuando el conjunto de datos posee poca información.

Como en todo no hay una visualización completamente correcta la elección entre histogramas y gráficos de densidad dependerá de lo que se quiera comunicar y como se quieran ver reflejadas las características especificas de los datos. También existe la posibilidad de no utilizar estos dos gráficos y optar por realizar funciones empíricas de densidad acumulativa o qqplot; lo que si es cierto es que las densidades tienen una ventaja inherente sobre los histogramas cuando se trata de visualizar mas de una distribución a la vez.

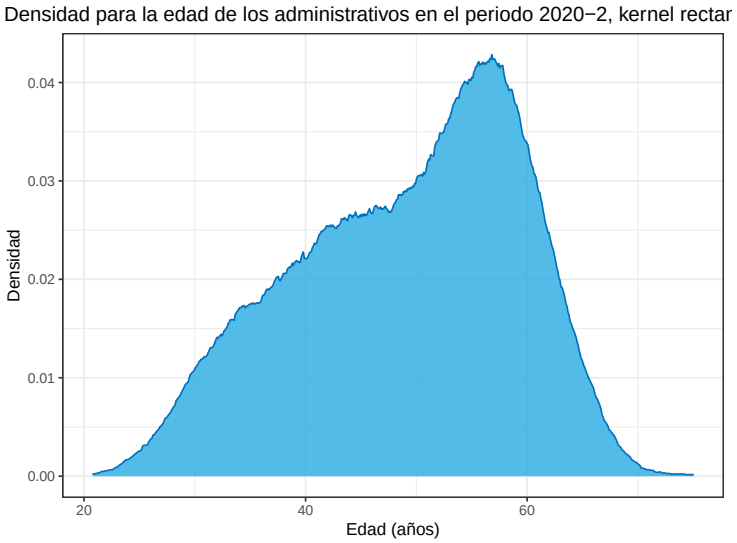

Figure 3.29: Densidad para la edad del personal administrativo en el periodo 2020-2, kernel rectangular

3.4.2 Visualización de múltiples distribuciones

En múltiples escenarios del análisis de datos existen ocasiones en las cuales es necesario la visualización de múltiples distribuciones de manera simultánea. Por ejemplo, suponga que estamos interesados en visualizar como se distribuyen las edades de los estudiantes graduados por sexo, la idea es identificar las diferencias existentes entre las edades de hombres y mujeres, ¿los hombres y mujeres son generalmente de la misma edad o existe una diferencia de edad entre los sexos?; una estrategia de visualización comúnmente utilizada en estas situaciones es un histograma apilado, donde se dibujan las barras del histograma para hombres en la parte superior de las barras para las mujeres, en un color diferente como se ilustra en la figura 3.30.

La figura anterior fue etiquetada como incorrecta ya que este tipo de visualización no es recomendable ya que es posible identificar dos problemas claves. El primero de ellos tiene que ver con solo mirar la figura, no esta completamente claro donde comienzan exactamente las barras, podríamos preguntarnos si inician donde cambia de color o están destinadas a iniciar desde cero, es decir, para el rango de 21 a 14 años hay alrededor de 15.000 hombres o cerca de 27.000. Em segundo lugar, las alturas de las barras para los conteos de hombres no se pueden comparar directamente entre sí, por que todas las barras comienzan en una altura diferente.

Lo correcto en estas ocasiones es usar gráficos de densidad superpuestos, ya que estos ayudan al ojo a mantener las distribuciones separadas. Sin embargo, para

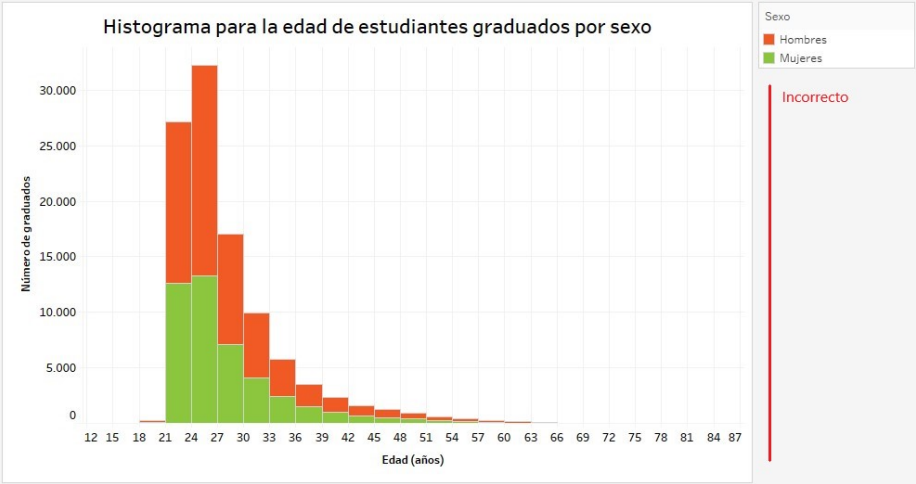

Figure 3.30: Histograma para la edad de estudiantes graduados discriminado por sexo

el conjunto de datos de administrativos, la distribución de edad los empleados hombres y mujeres son casi idénticas hasta alrededor de los 25 años y luego divergen, esto se observa en la figura 3.31, por lo que la visualización resultante es válida pero no la ideal.

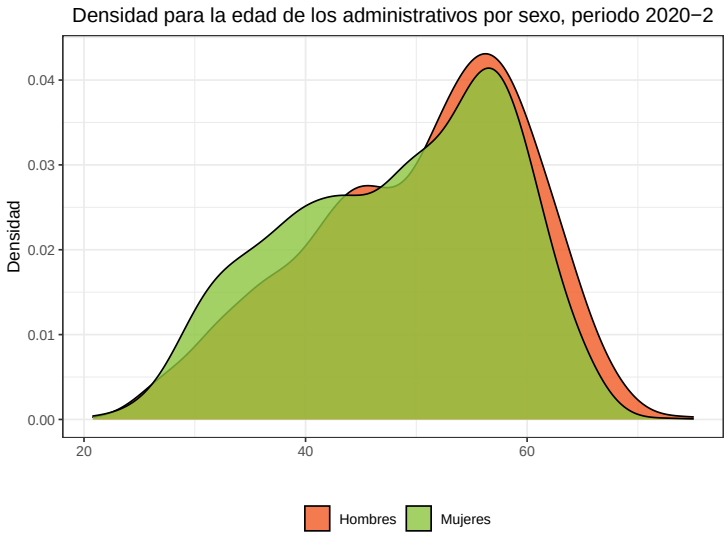

Figure 3.31: Densidad para la edad del personal administrativo en el periodo 2020-2 discriminado por sexo

Chapter 4

Errores en la trama

El principal y más grande error dentro de la visualización de datos es no respetar la integridad gráfica, por integridad gráfica nos referimos a los factores de engaño o adornos que se añaden, pensando que harán las figuras más atractivas y llamativas, pero en realidad saturan y opacan la verdadera intención de los gráficos que es informar. Un ejemplo claro de esto son las figuras 3D, las rejillas y los rellenos con patrones.

El factor de engaño corresponde a las distorsiones generadas por la mala elección de los elementos gráficos, también se incluyen elementos gráficos que no corresponden a variaciones de datos o que hacen más compleja la interpretación de estos.

Un diseño muy popular que introduce una alta distorsión en los datos son los gráficos en 3D, muchos softwares de visualización permiten arreglar los gráficos convirtiéndolos en objetos tridimensionales que generalmente son rotados hasta conseguir una proyección bidimensional, comúnmente este efecto es aplicado a gráficos de torta, barras y dispersión. El principal problema con los diseños 3D es que la proyección a dos dimensiones para poder visualizarlos en un monitor distorsiona los datos, para ilustrar esto observe la figura 4.1 que muestra la distribución de aspirantes a pregrado del programa PAES para el periodo 2021-1, note que la rebana correspondiente a la población afrocolombiana se ve más grande que la rebanada de comunidades indígenas, aunque claramente este no es el caso, ya que de los aspirantes a pregrado el 26% pertenece a comunidades indígenas y el 17% a la población afroamericana. Este es un claro ejemplo de los factores de engaño que se pueden introducir al gráfico a través de elementos estéticos como el diseño tridimensional.

Para eliminar el factor de engaño es necesario eliminar de la visualización el efecto 3D, para lograr que cada rebanada sea proporcional a cantidad que representa, como se muestra en la figura 4.2, a pesar de que esta visualización es mucho más clara que la anterior no es del todo correcta, ya que la variable

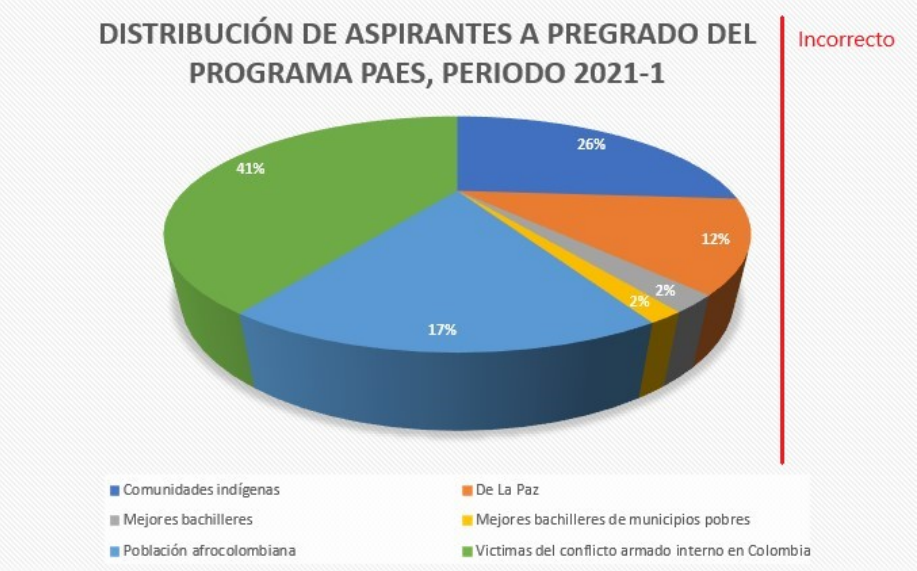

Figure 4.1: Uso incorrecto de las figuras en 3D

discreta presenta seis categorías y este tipo de gráficos es recomendable cuando no se superan las cuatro categorías.

Cuando se presentan más de cuatro categorías es recomendable utilizar un gráfico de barras y ordenar las clases de manera decreciente para lograr una mejor comprensión por parte del usuario.

En muchas ocasiones cuando se realizan gráficos de dispersión se opta por añadir rejillas con la intención de que el usuario identifique con facilidad las coordenadas X y Y de las observaciones, pero es común usar colores muy fuertes para estas rejillas opacando por completo las observaciones, como se ilustra en la figura 4.3.

La figura 4.3 representa la relación existente entre el peso en miles de libras y la distancia recorrida en millas por galón para 32 autos, esta figura es etiquetada como incorrecta ya que la rejilla se sobrepone a las observaciones creando una visualización saturada, poco comprensible y degrada la percepción de patrones en los datos. Lo recomendado en estas ocasiones es utilizar puntos rellenos y colores suaves para la rejilla por ejemplo tonalidades grises como se presenta en la figura 4.4.

Otro adorno comúnmente utilizado que es desagradable y quita la atención de los datos es usar patrones para rellenar las figuras, por ejemplo, utilizar líneas, puntos, estrellas u otras figuras para rellenar las barras, la figura 4.5 muestra la distribución de los graduados por grupo de edad para el periodo 2020-2, se etiqueto como incorrecta ya que añadir patrones para rellenar las barras se

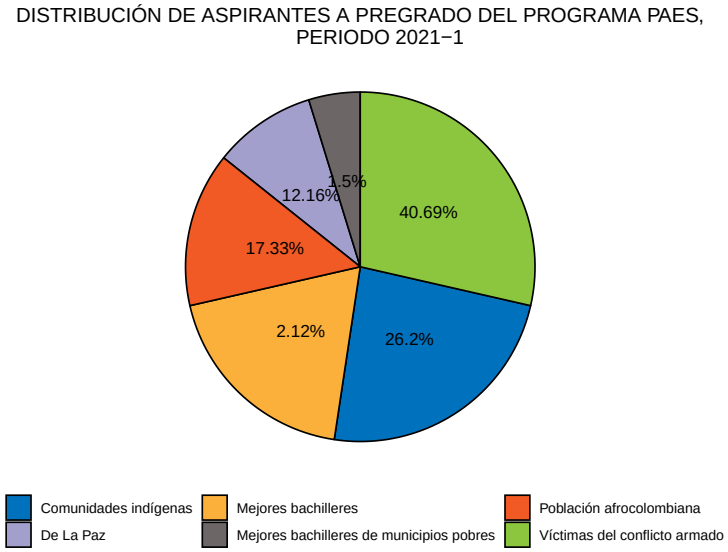

Figure 4.2: Eliminar efecto 3D

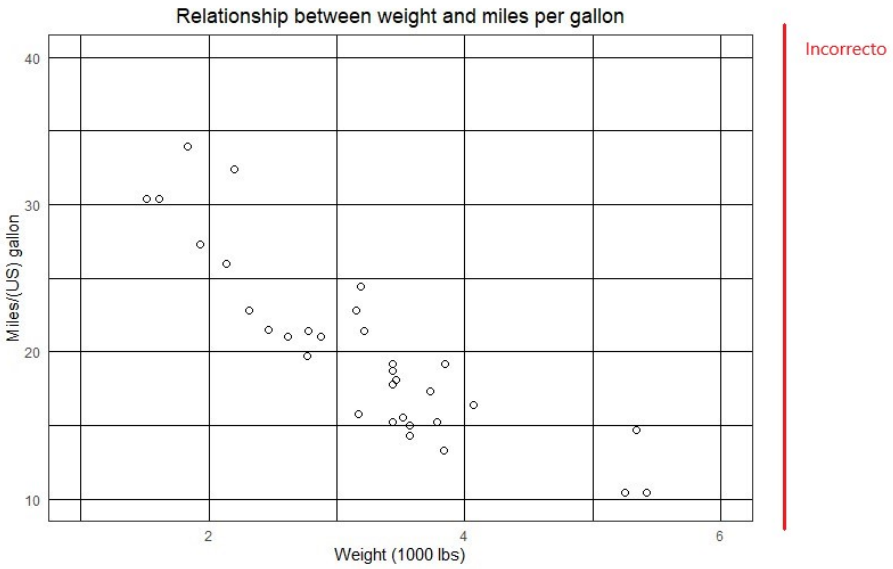

Figure 4.3: Uso incorrecto de rejillas

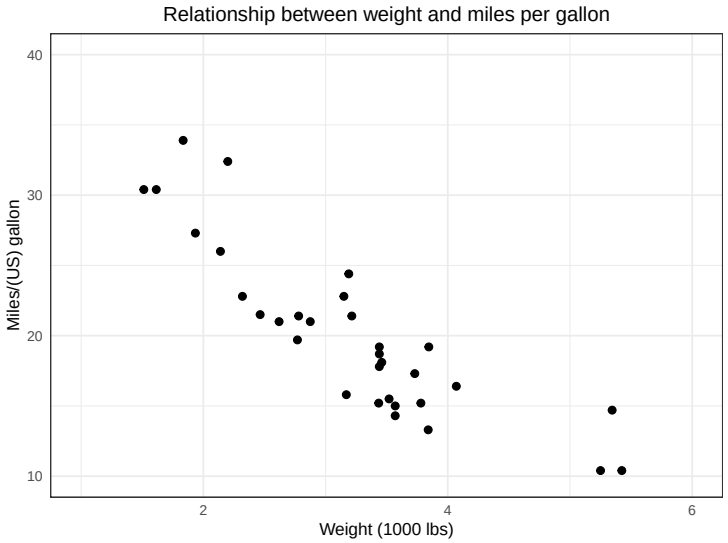

Figure 4.4: Uso de rejillas en tonalidades suaves

considera poco atractivo y distrae la atención del usuario.

Cuando se quiere añadir algún tipo de relleno a las barras se recomienda hacerlo con colores, usando escalas de colores cualitativas si la intención es distinguir datos o una escala secuencial si se trata de representar cantidades, la forma correcta o recomendada para realizar esta gráfica se muestra en la figura 4.6.

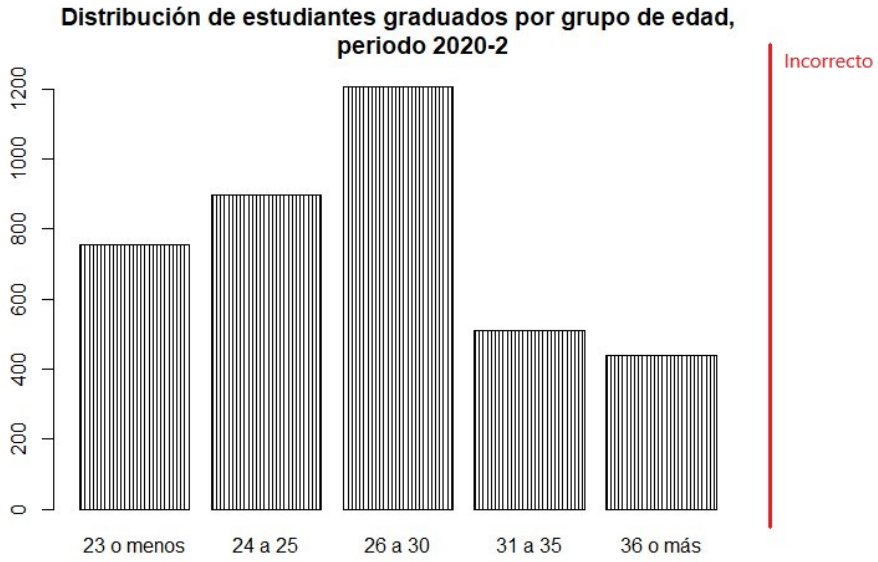

Figure 4.5: Uso incorrecto de los patrones para rellenar

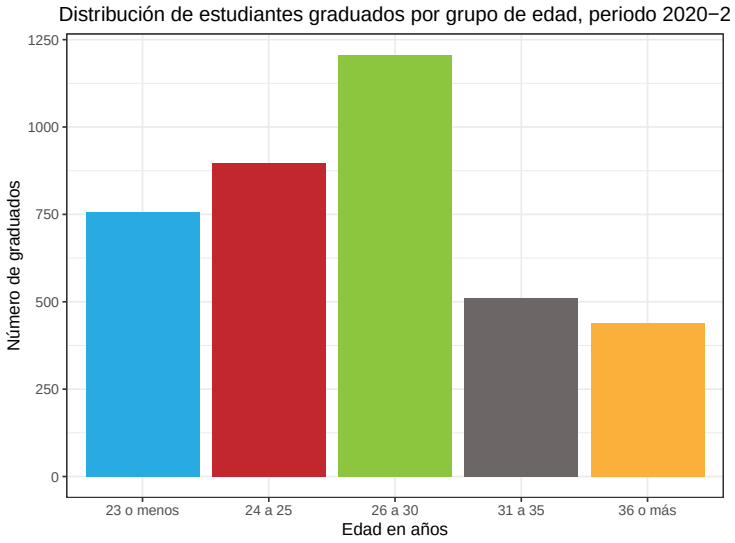

Figure 4.6: Uso de los rellenos de color

Chapter 5

Final Words

We have finished a nice book.

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2021). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.22.