Lineamientos conceptuales para la visualización estadística

Camila Acosta Ramirez

2021-05-27

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).

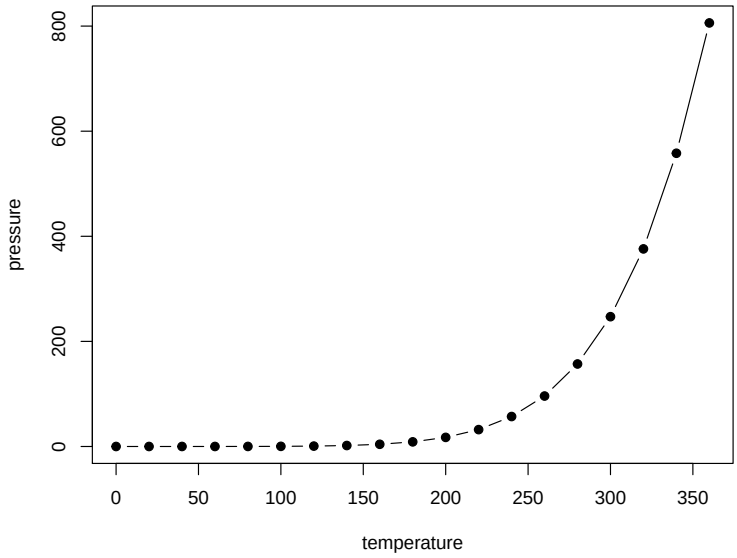

Figure 1.1: Here is a nice figure!

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 garantizara 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, titulo y leyenda.

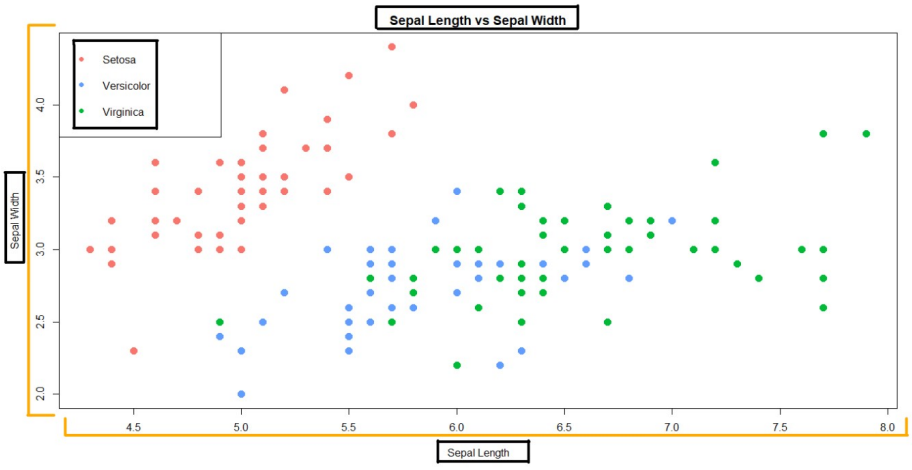

Figure 2.1: Principales partes de un gráfico

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 determina 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.

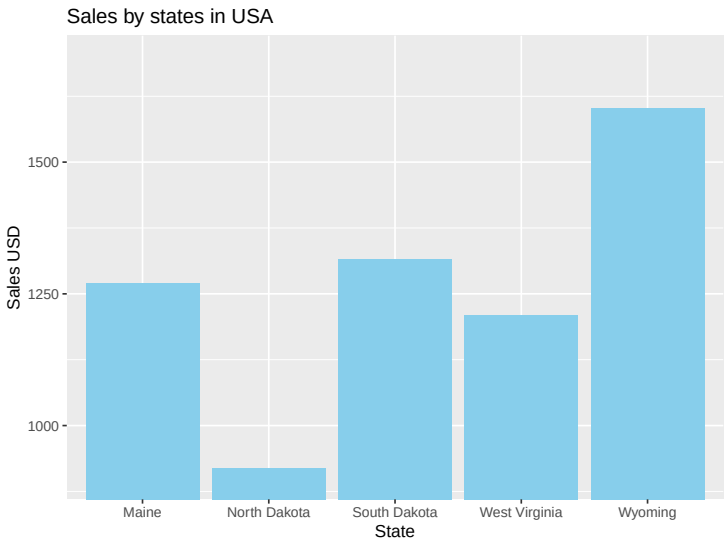

Figure 2.2: Ventas por estados de EE.UU, visualiazción engañosa

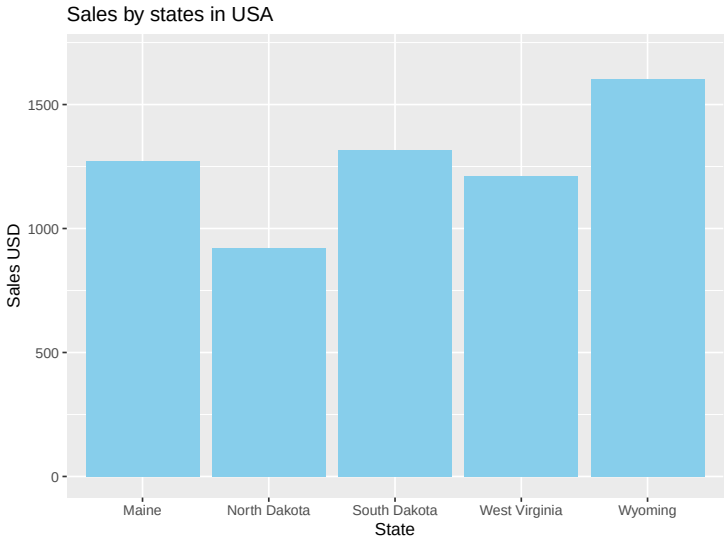

Figure 2.3: Ventas por estados de EE.UU, uso correcto de la escala lineal

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 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 continuadas 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 2.4 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, Visicolor 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 2.5.

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 2.6 ilustra el uso de una variable discreta para asignar diferentes tamaños a las observaciones.

La figura 2.7 muestra el uso de una variable continua para determinar el tamaño de cada observación, note que a demás de la geometría relacionada al tamaño

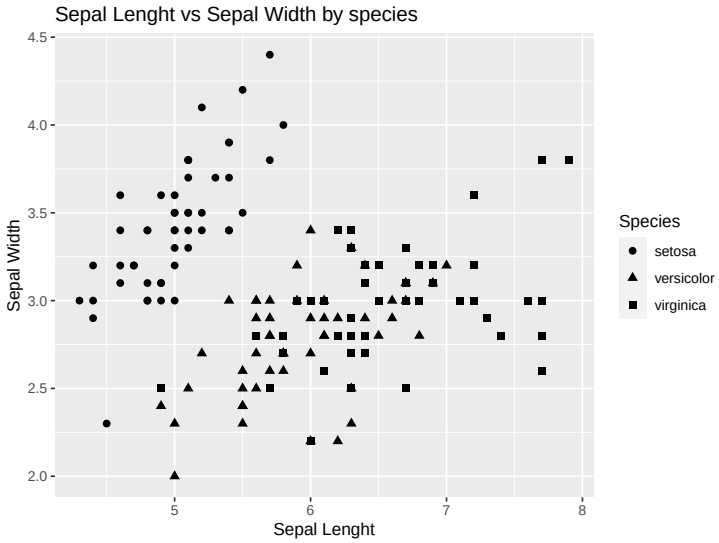

Figure 2.4: Uso de las formas para distinguir grupos de datos

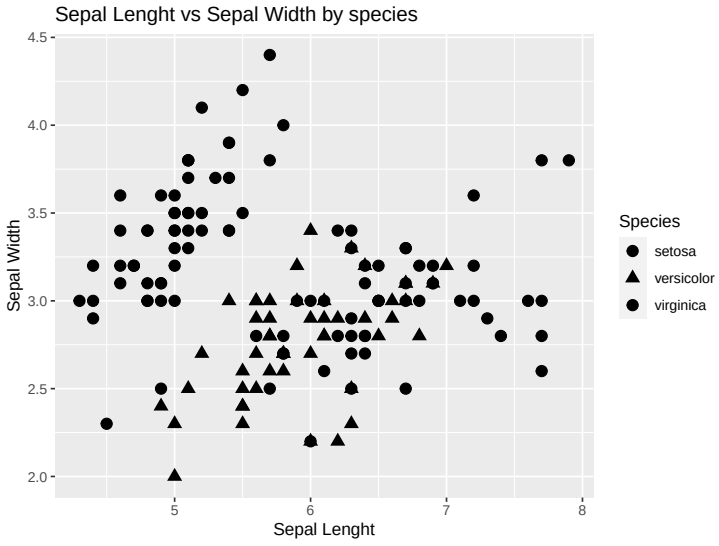

Figure 2.5: Uso de las formas para resaltar un grupo de observaciones

Figure 2.6: Asignación de tamaños mediante una variable discreta

se debe usar la transparencia para evitar que los países con mayor población oculten o se sobrepongan a aquellos a países de menor población.

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

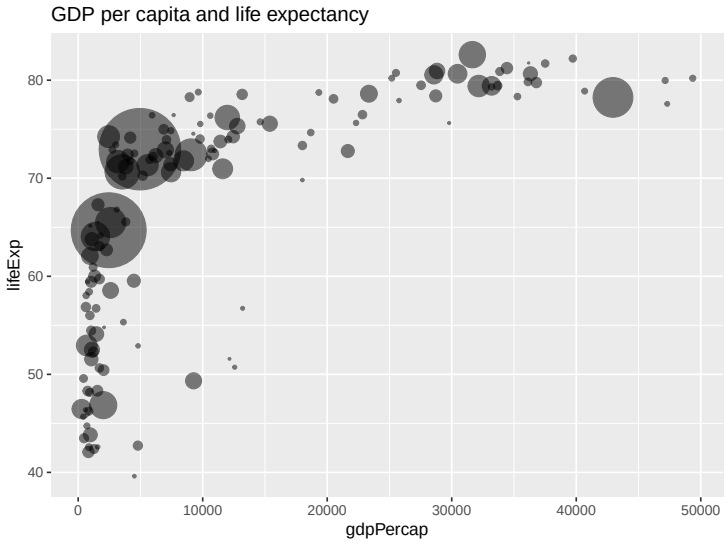

Figure 2.7: Asignación de tamaños mediante una variable continua

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 **??** 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.

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 mas comunes a visualizar se tienen las cantidades, proporciones, distribuciones, series de tiempo, datos geoespaciales y las relaciones X-Y. En este capitulo 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. Observe que en todos estos casos de 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 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 de

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.

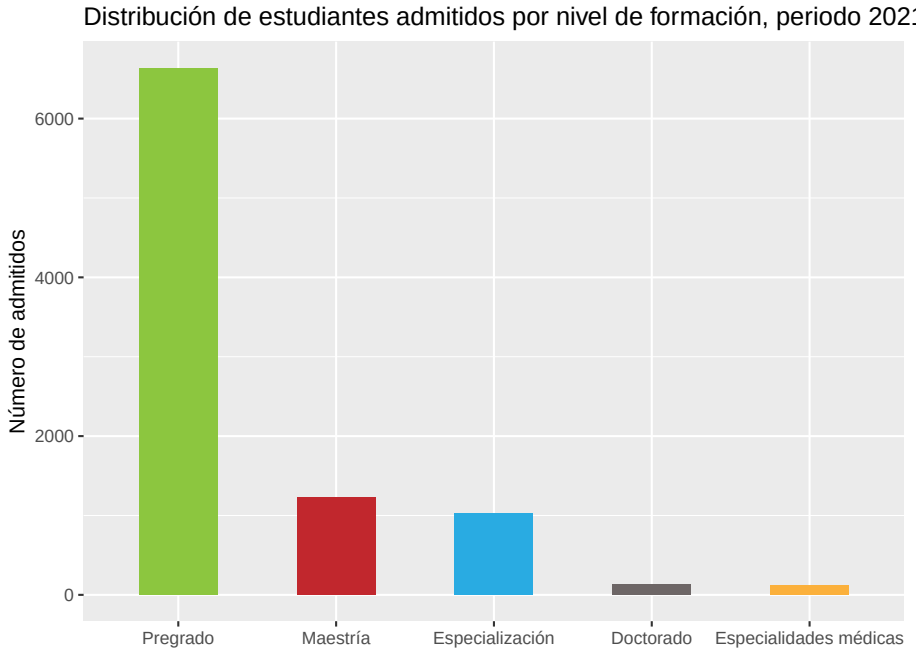

Figure 3.1: Uso del gráfico de barras para representar cantidades

Uno de los problemas mas 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 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.

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

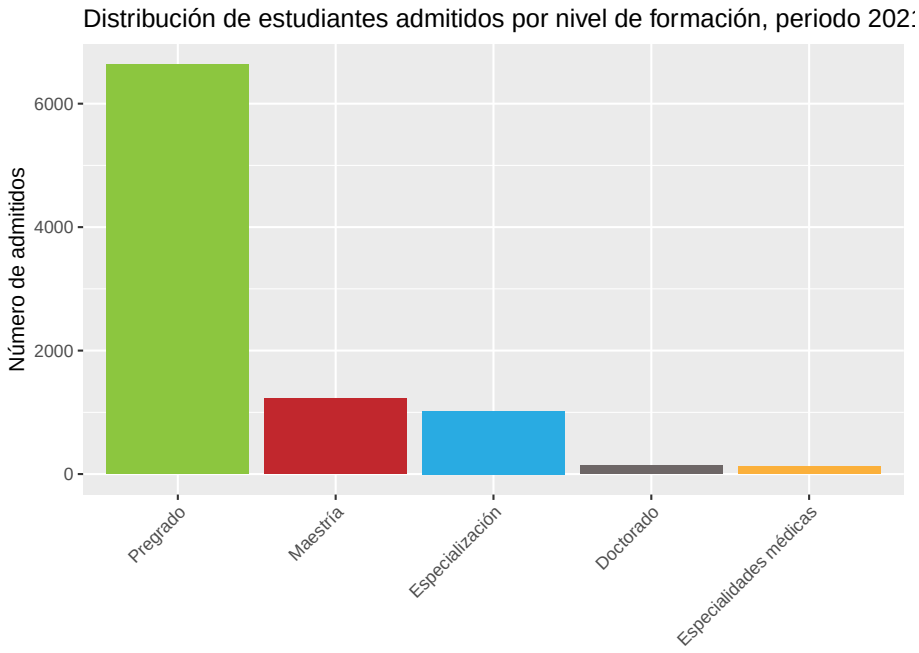

Figure 3.2: Uso del gráfico de barras para representar cantidades, girando etiquetas

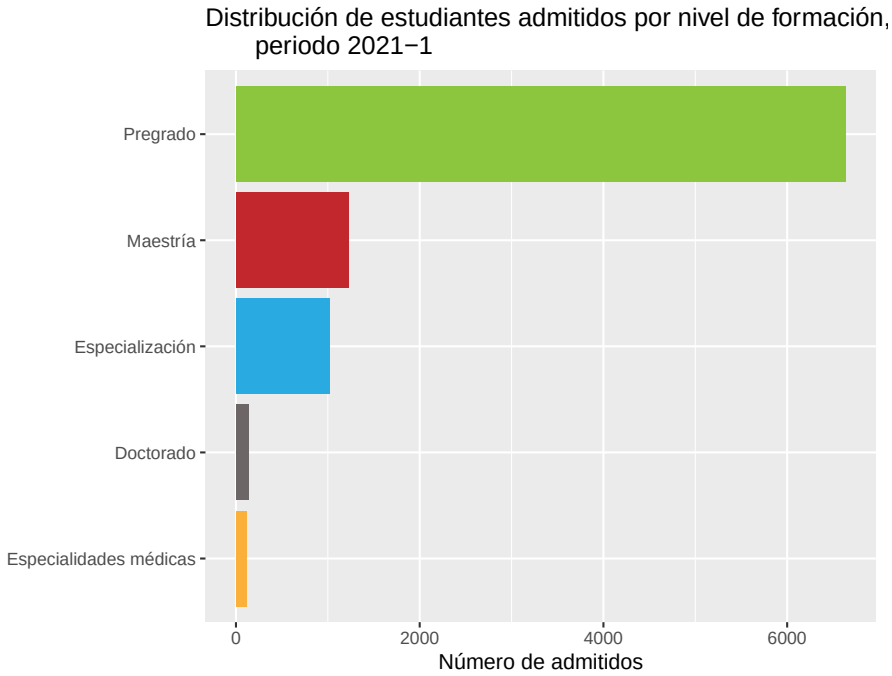

Figure 3.3: Uso del gráfico de barras horizontales para representar cantidades

ubican las etiquetas por orden alfabético o algún otro criterio.

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.

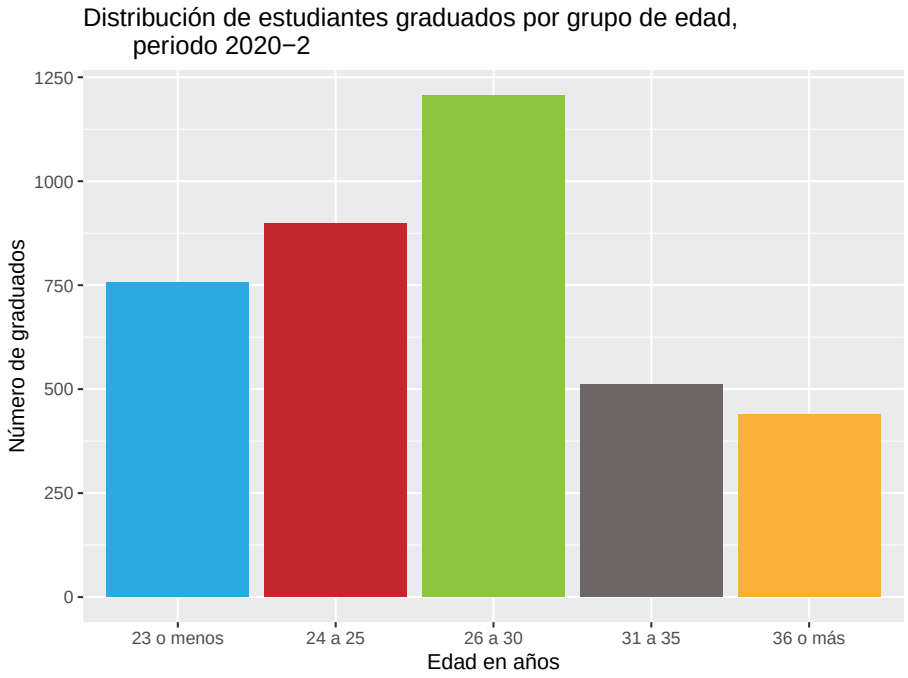

Figure 3.4: Etiquetas con un orden natural

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 3.5.

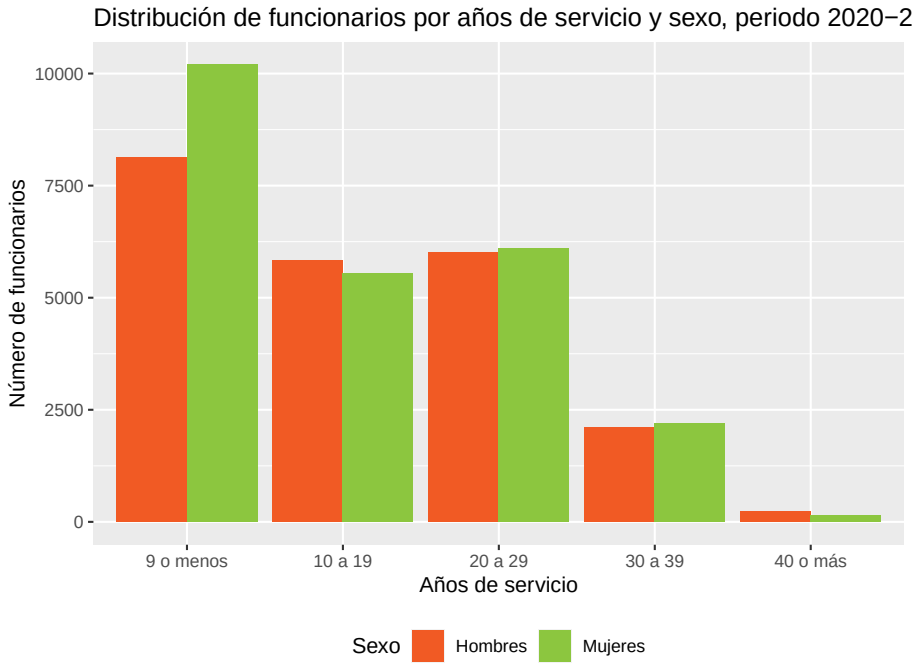

Figure 3.5: Uso de barras agrupadas para representar valores usando dos categorías

Con los gráficos de barras agrupadas se debe tener cuidado ya que las variables categóricas elegidas pueden tener muchos niveles y harán que le grafico se sature y el usuario no comprenda la información de manera correcta, la figura 3.6 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 ocho barras

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 3.7.

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,

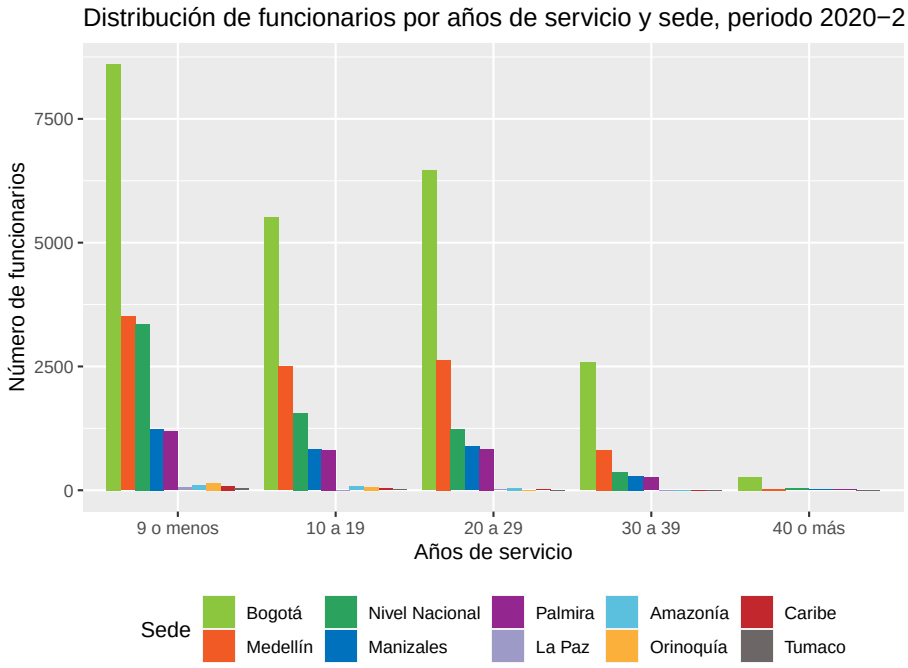

Figure 3.6: Cantidad de funcionarios por años de servicio y sede

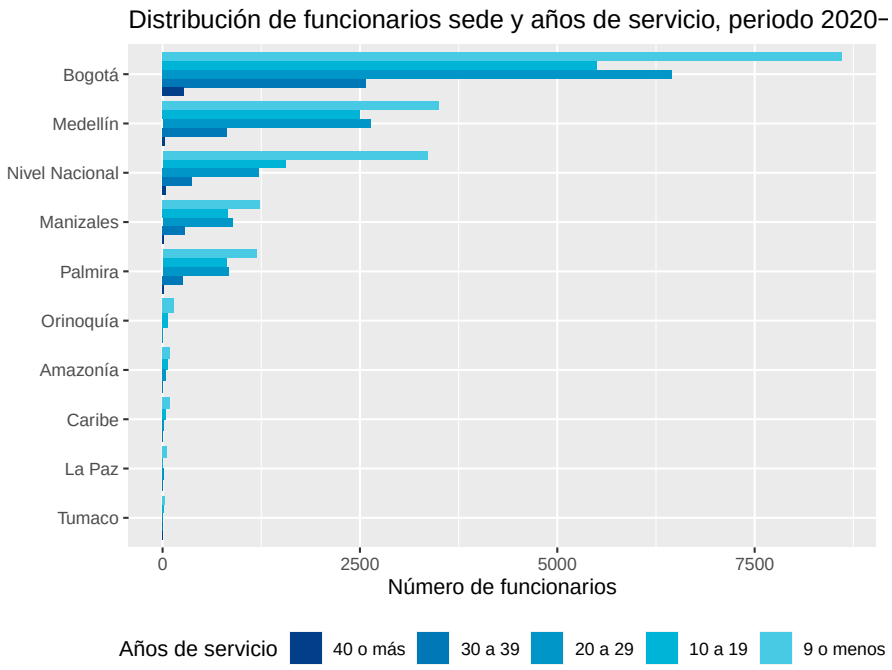

Figure 3.7: Cantidad de funcionarios por sede y años de servicio

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 casi 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 3.8 presenta el uso de barras apiladas como alternativa del uso de barras agrupadas ilustrado en la figura 3.5.

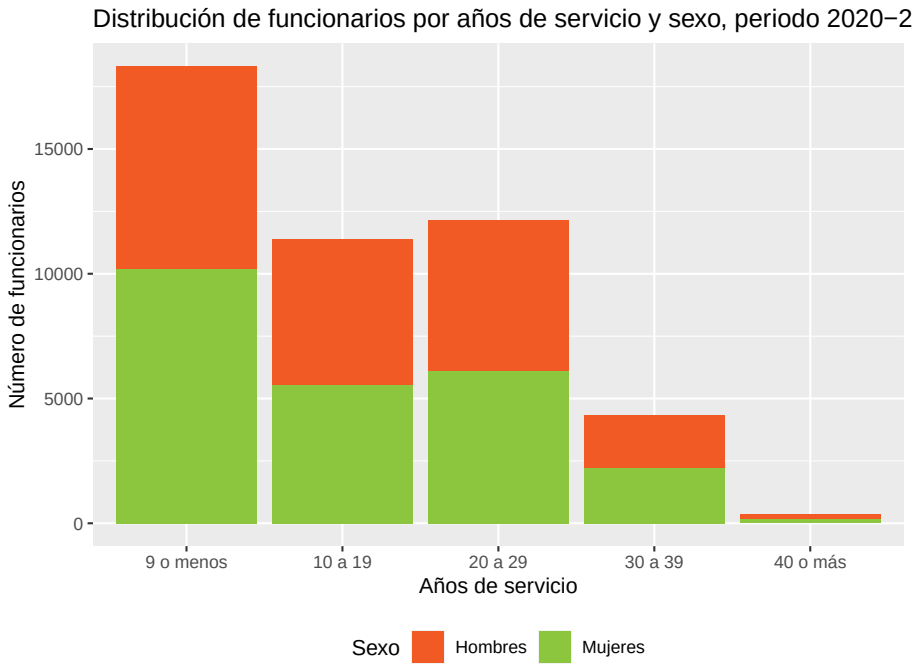

Figure 3.8: Barras apiladas como una alternativa a las barras agrupadas

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 practico 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 de 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 3.9.

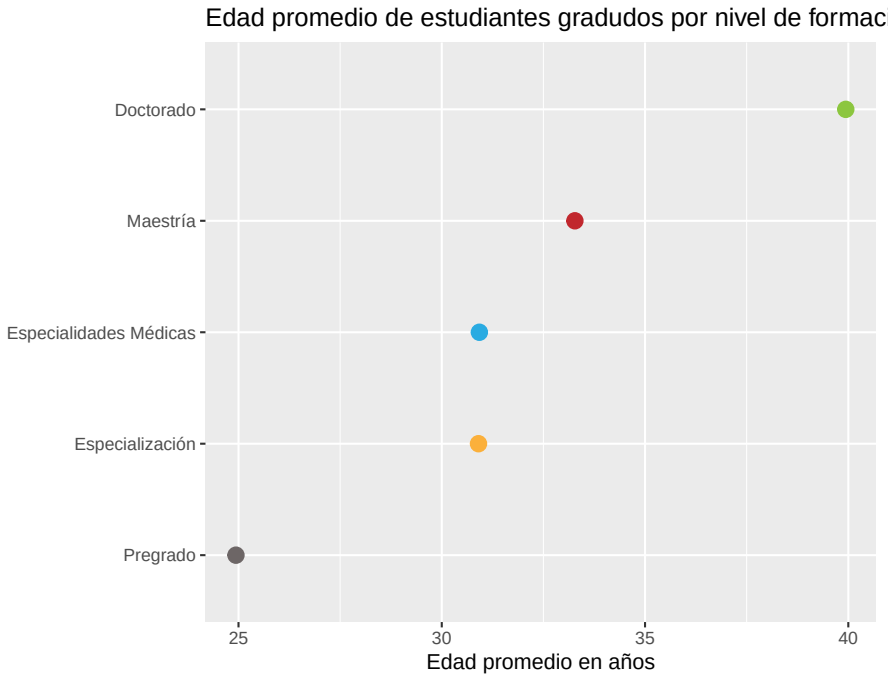

Figure 3.9: Diagrama de puntos como alternativa a los gráficos de barras

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.

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 3.6 que grupos de 8 barras resultan en una visu-

alización 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 @ref(fig: mapadecalor-fig) 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.

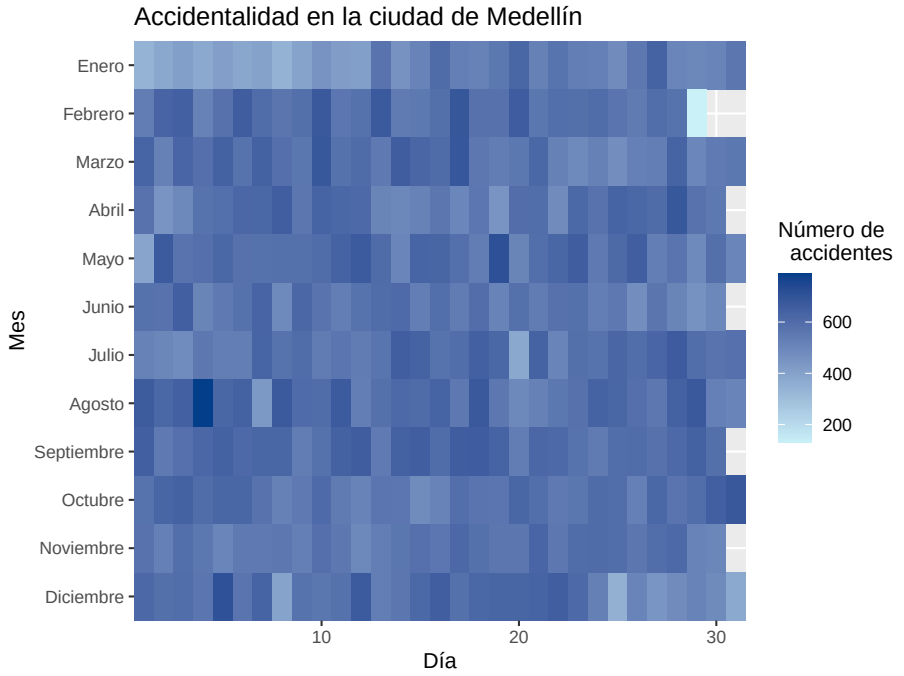

Figure 3.10: Mapa de calor para representar la accidentalidad en la ciudad de Medelllín

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

Chapter 4

Applications

Some *significant* applications are demonstrated in this chapter.

4.1 Example one

4.2 Example two

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.