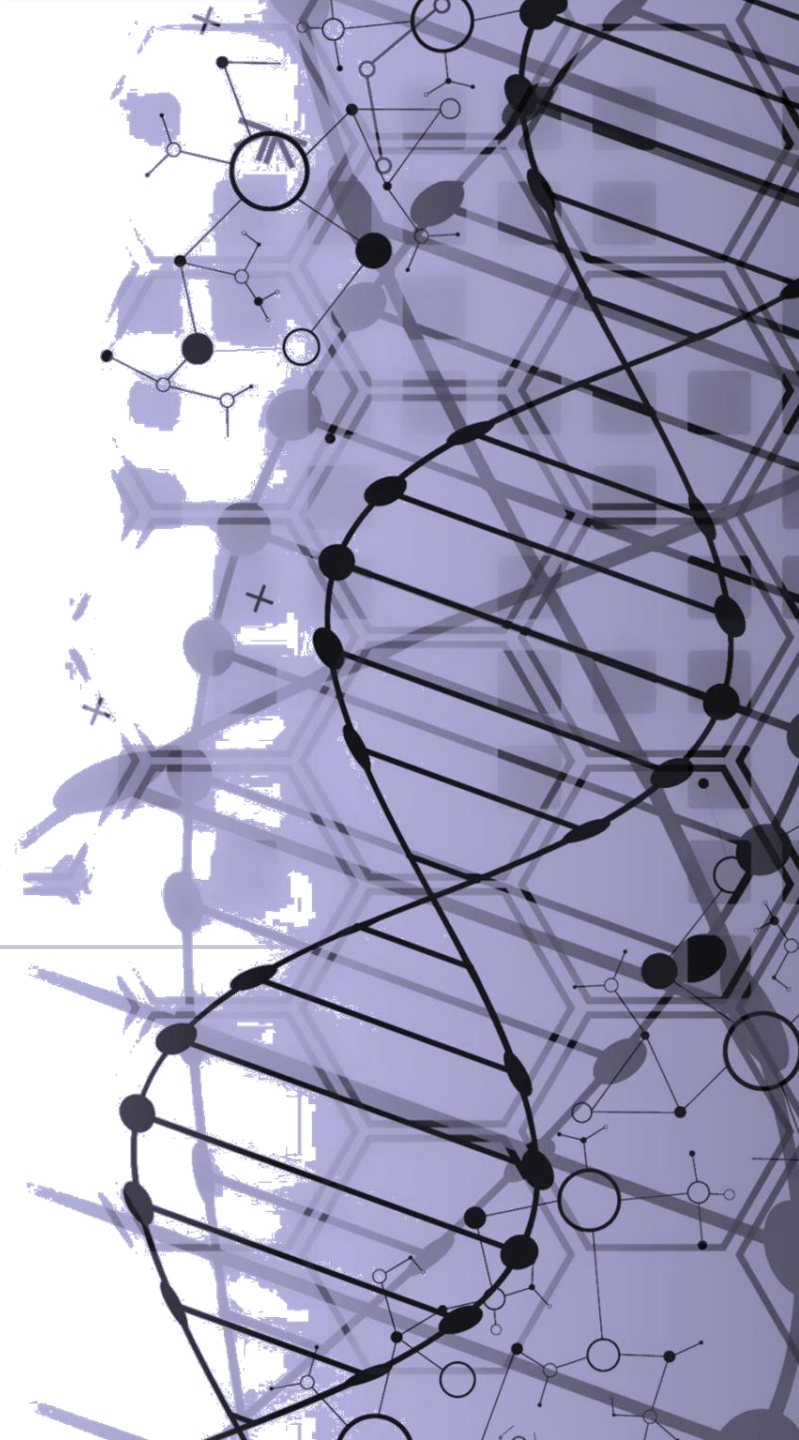





Análisis exploratorio de datos

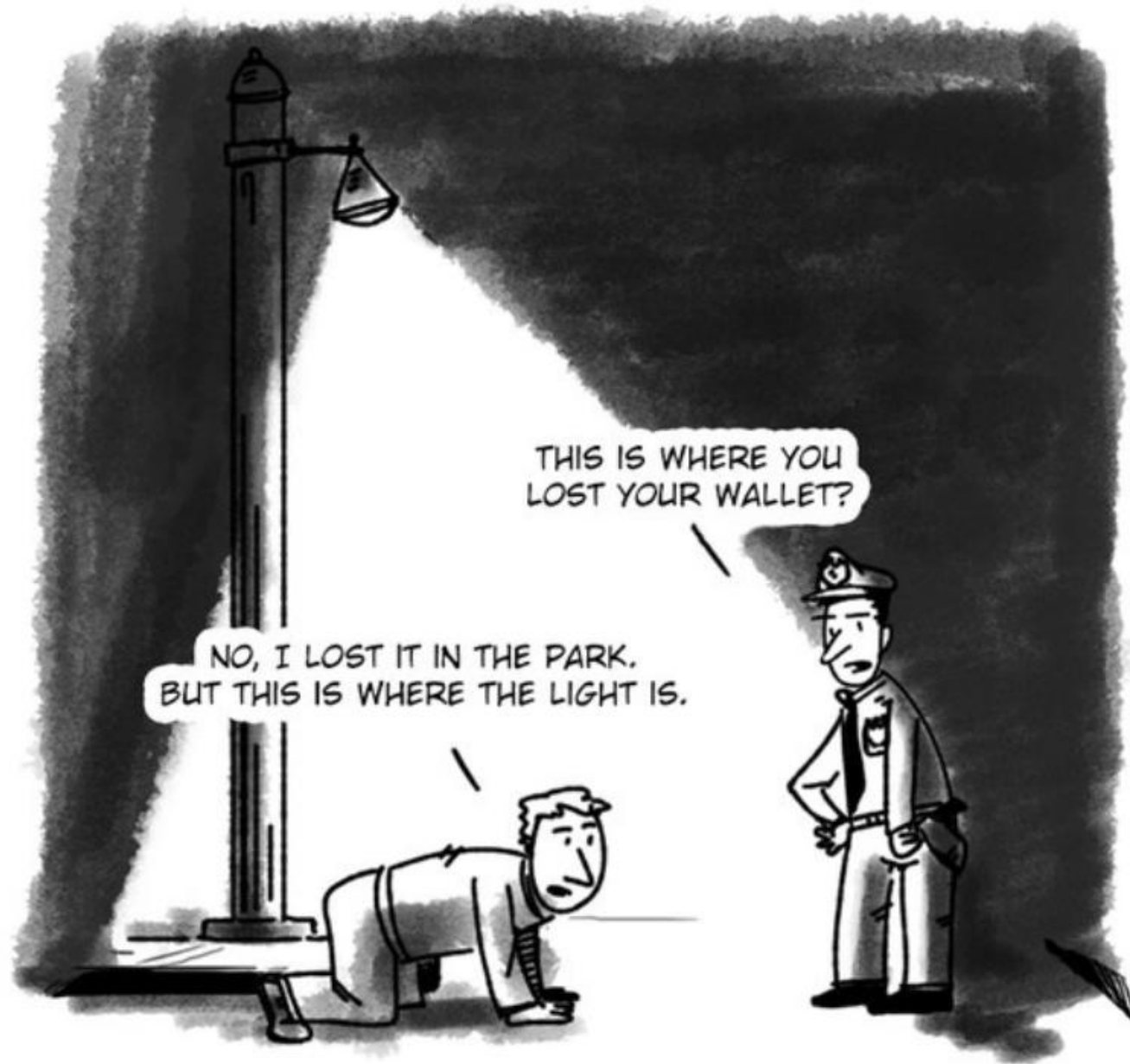
Basado en *R for Data Science*, de Wickman y Grolemond

Ignacio Díaz Oreiro
CI0131. Diseño de Experimentos
Universidad de Costa Rica





Agradecimiento a la profesora Dra. Kryscia Ramírez Benavides, por facilitar material usado en esta presentación



Trevor Klee, 2023

“Es mucho mejor una respuesta aproximada a la pregunta correcta, que a menudo es vaga, que una respuesta exacta a la pregunta equivocada, que siempre se puede precisar”.

– John Tukey

Análisis Exploratorio de Datos (EDA)

- El análisis exploratorio de datos se refiere al proceso crítico de realizar investigaciones iniciales sobre los datos para descubrir patrones, detectar anomalías, y verificar suposiciones con la ayuda de estadísticas resumidas y representaciones gráficas.

Análisis Exploratorio de Datos (EDA)

EDA es un ciclo iterativo en el que:

- Se generan preguntas sobre los datos.
- Se buscan respuestas visualizando, transformando y modelando los datos.
- Se utiliza lo que se aprende para refinar las preguntas y/o generar nuevas preguntas.

Análisis Exploratorio de Datos (EDA)

- No es un proceso formal con un conjunto estricto de reglas, sino más bien un *estado de ánimo o mental*.
- Durante las fases iniciales de EDA, el investigador debe sentirse libre de investigar cada idea que se le ocurra.
- Algunas de estas ideas funcionarán y otras serán callejones sin salida.
- A medida que continúe la exploración, se centrará en áreas particularmente productivas que eventualmente escribirá y comunicará a otros.

Análisis Exploratorio de Datos (EDA)

- EDA es una parte importante de cualquier análisis de datos porque siempre se necesita investigar la calidad de los datos, y así saber si los datos cumplen con las expectativas o no.
- También sirve para realizar la limpieza de datos, utilizando herramientas y técnicas como visualización, transformación y modelado.

Análisis Exploratorio de Datos (EDA)

- El objetivo es desarrollar una comprensión de los datos.
- La forma más fácil de hacer esto es usar preguntas como herramientas para guiar la investigación.
- Cuando se formula una pregunta, esta pregunta enfoca su atención en una parte específica del conjunto de datos y ayuda a decidir qué gráficos, modelos o transformaciones hacer.

Análisis Exploratorio de Datos (EDA)

- Es fundamentalmente un proceso creativo, donde la clave para hacer preguntas de calidad es generar una gran cantidad de preguntas.
- Es difícil hacer preguntas reveladoras al comienzo del análisis porque no se sabe qué información contiene el conjunto de datos.
- Por otro lado, cada nueva pregunta expondrá un nuevo aspecto de los datos y aumentará las posibilidades de hacer un descubrimiento.

Análisis Exploratorio de Datos (EDA)

- Permite profundizar rápidamente en las partes más interesantes de los datos y desarrollar un conjunto de preguntas que invitan a la reflexión, si se hace un seguimiento de cada pregunta con una nueva pregunta basada en lo que encuentre.

Análisis Exploratorio de Datos (EDA)

- Aunque no hay una regla sobre qué preguntas debe hacer para guiar la investigación, dos tipos de preguntas siempre serán útiles para hacer descubrimientos dentro de los datos:
- ¿Qué tipo de variación ocurre dentro de las variables?
- ¿Qué tipo de covariación ocurre entre las variables?

Variación

- Es la tendencia de los valores de una variable a cambiar de una medición a otra.
- Si se mide cualquier variable numérica dos veces, se obtendrán dos resultados diferentes, incluso si se miden cantidades que son constantes, como la velocidad de la luz.
- Cada una de las mediciones incluirá una pequeña cantidad de error que varía de una medición a otra.

Variación

- Pueden ocurrir variaciones tanto en variables categóricas, como en discretas y continuas.
- Una variable categórica contiene un número finito de categorías o grupos y pueden no tener un orden lógico.
- En R, las variables categóricas generalmente se guardan como factores o vectores de caracteres.

Variación

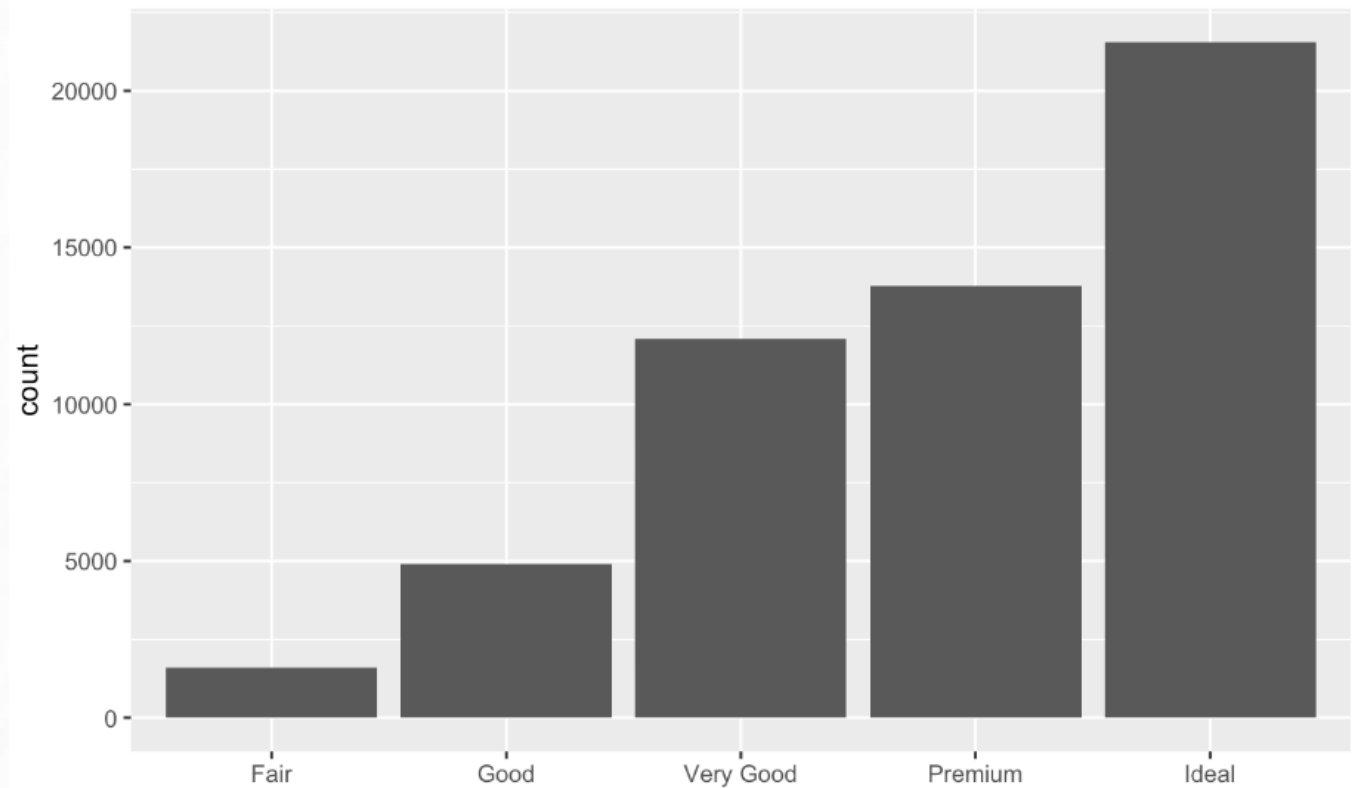
- Las variables discretas son numéricas y contienen un número contable de valores.
- Una variable es continua si puede tomar cualquiera de un conjunto infinito de valores ordenados.

Variación

- Medir variables categóricas también pueden reportar variaciones, por ejemplo, el color de ojos de una persona puede identificarse en más de una categoría.
- Cada variable tiene su propio patrón de variación, que puede revelar información interesante.
- La mejor manera de entender ese patrón es visualizar la distribución de los valores de la variable.

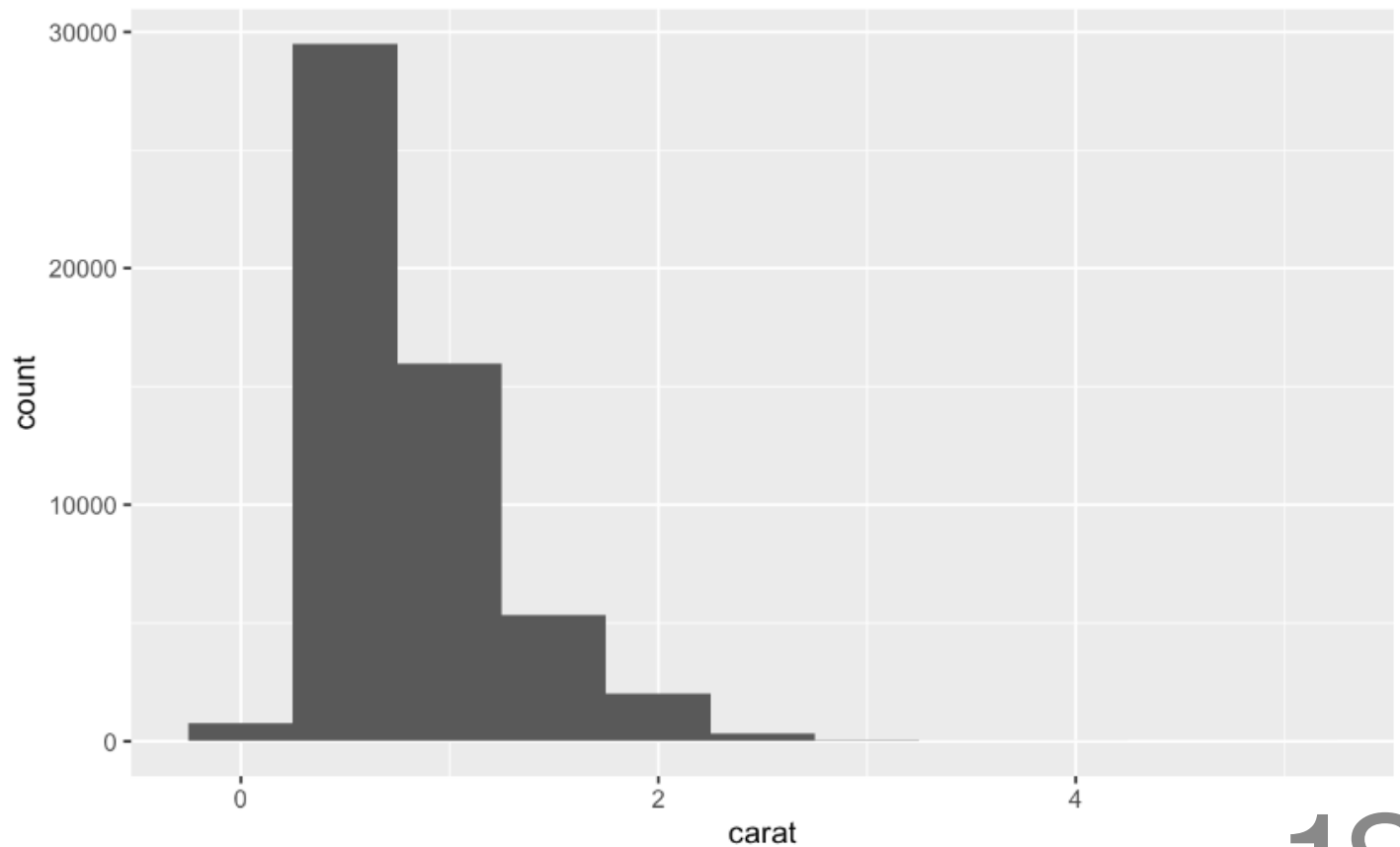
Variación – Visualizar distribuciones

Para examinar la distribución de una variable categórica, se puede utilizar un gráfico de barras:



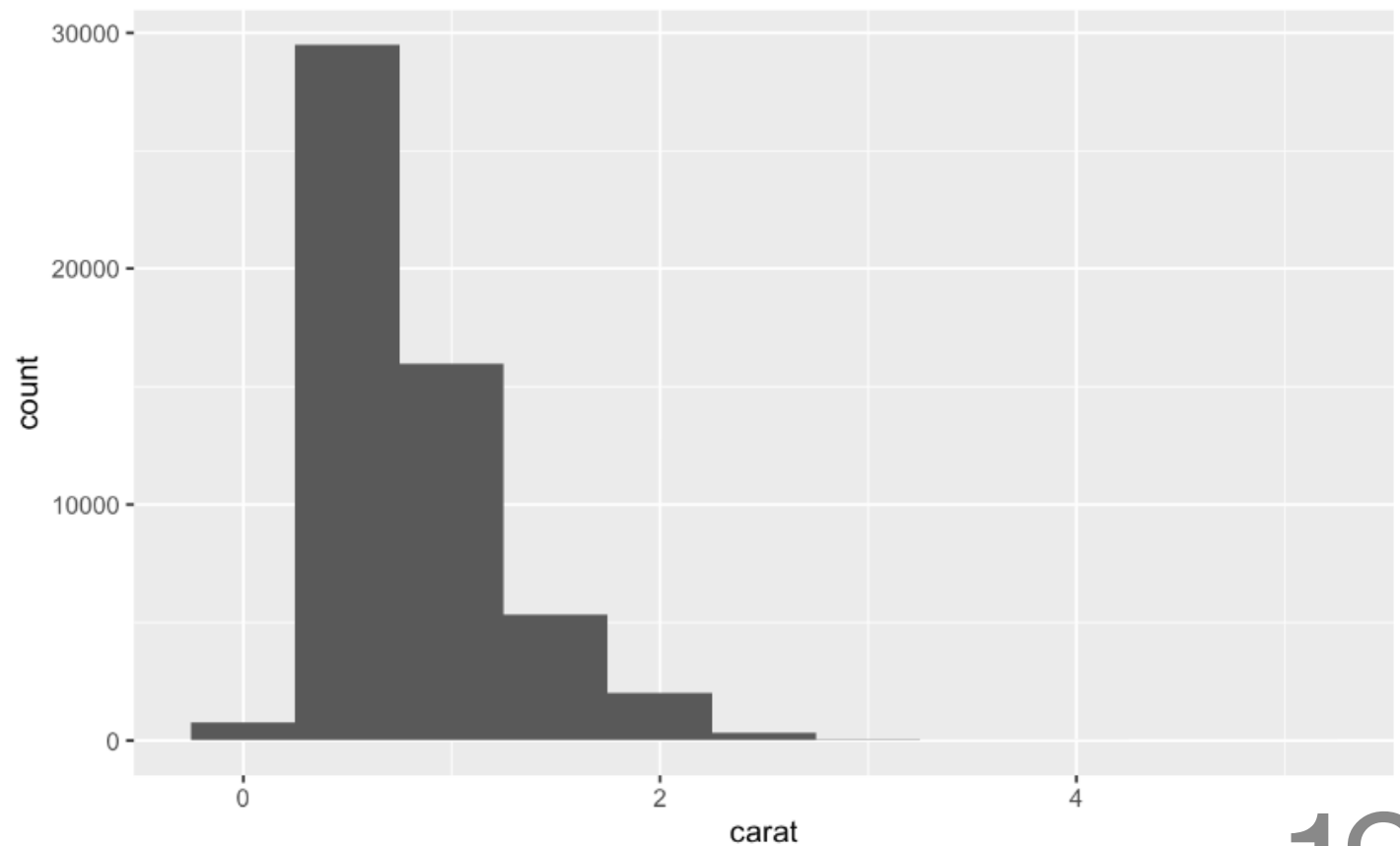
Variación – Visualizar distribuciones

Para examinar la distribución de una variable numérica, se puede utilizar un histograma que divide el eje x en contenedores igualmente espaciados y luego usa la altura de una barra para mostrar el número de observaciones que caen en cada contenedor.



Variación – Visualizar distribuciones

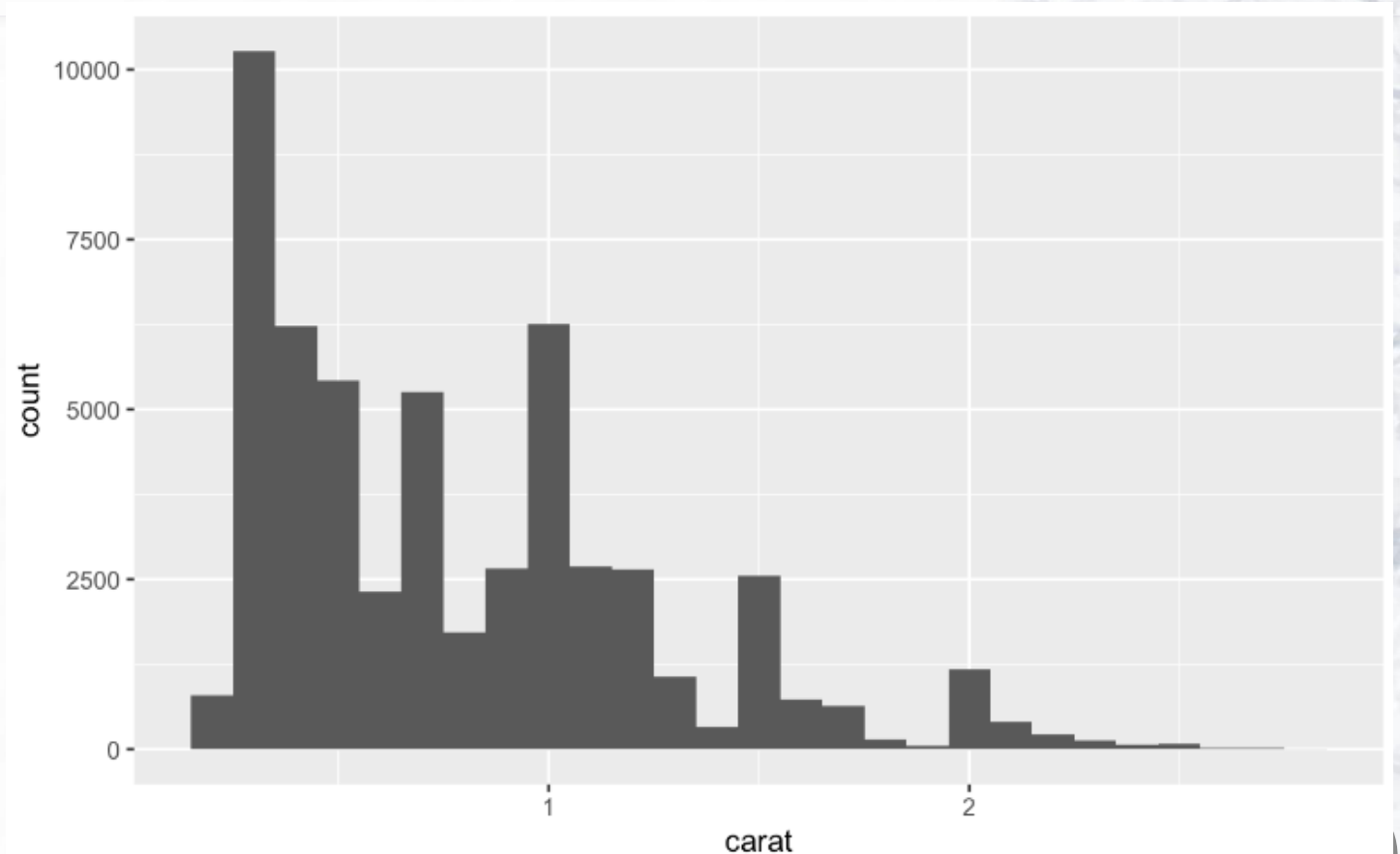
En este caso se trata de la distribución del peso (carat) de una muestra de alrededor de 54.000 diamantes.



Análisis Exploratorio de Datos (EDA)

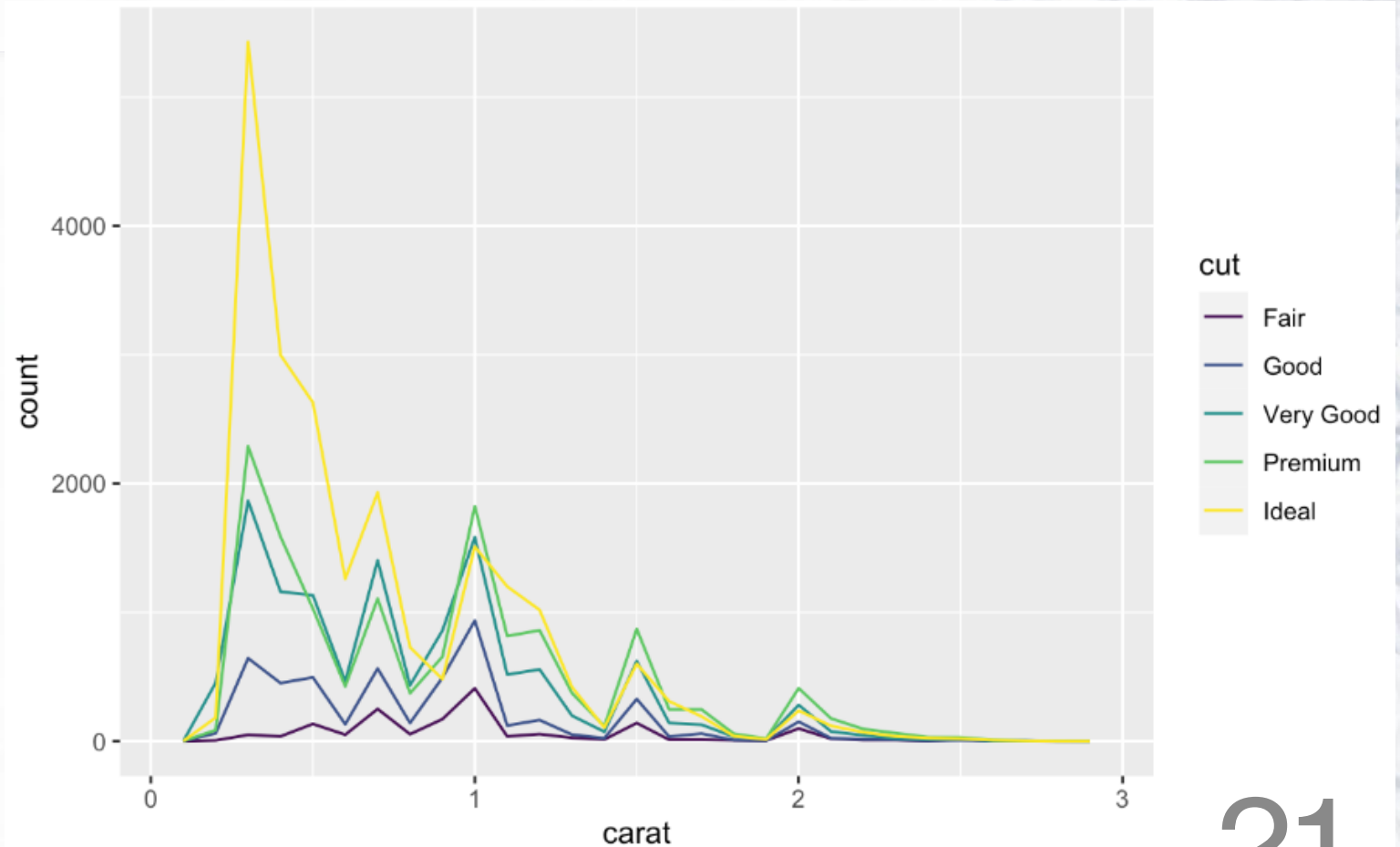
Variación – Visualizar distribuciones

Si se modifica el ancho del contenedor se pueden observar diferentes parámetros, por lo que es conveniente explorar diferentes anchos.



Variación – Visualizar distribuciones

Si se desea traslapar la información de varios histogramas, se puede utilizar un polígono de frecuencia, dado que es más sencillo entender líneas superpuestas que barras.

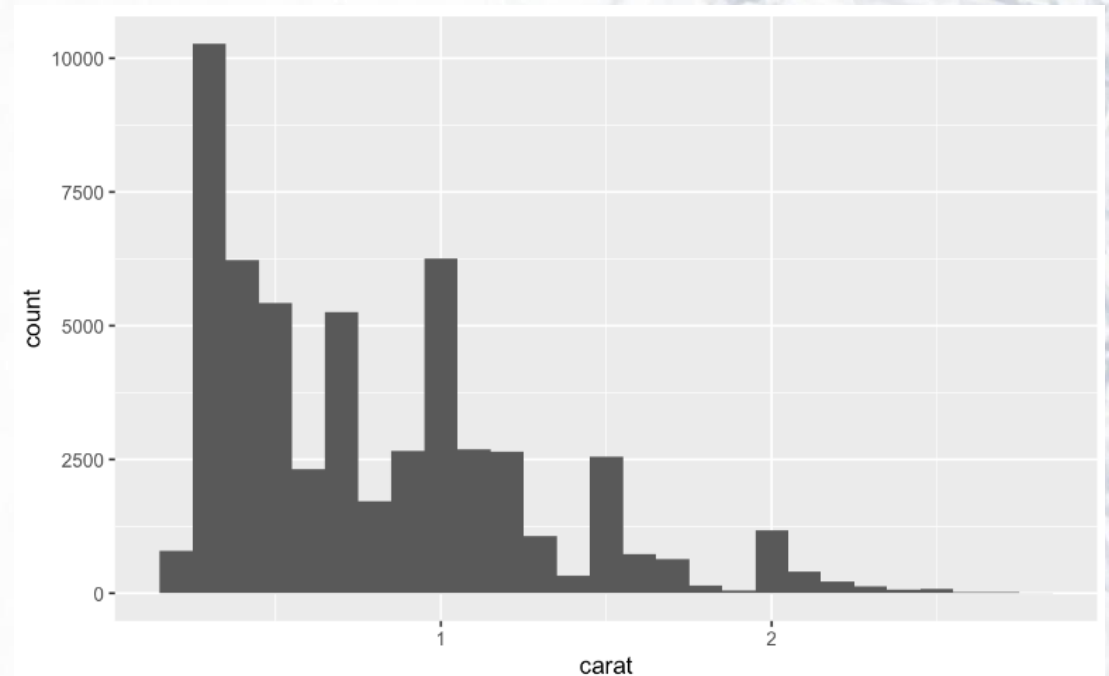


Variación

- Al poder visualizar la variación, ¿qué se debe buscar en los gráficos? ¿qué tipo de preguntas de seguimiento se deben hacer?
- La clave para hacer buenas preguntas de seguimiento será confiar en la curiosidad (¿sobre qué se quiere aprender más?) así como en el escepticismo (¿cómo podría ser engañoso?).

Variación – Valores típicos

- Tanto en gráficos de barras como en histogramas, las barras altas muestran los valores comunes de una variable y las barras más cortas muestran los valores menos comunes.
- Los lugares que no tienen barras revelan valores que no se vieron en los datos.



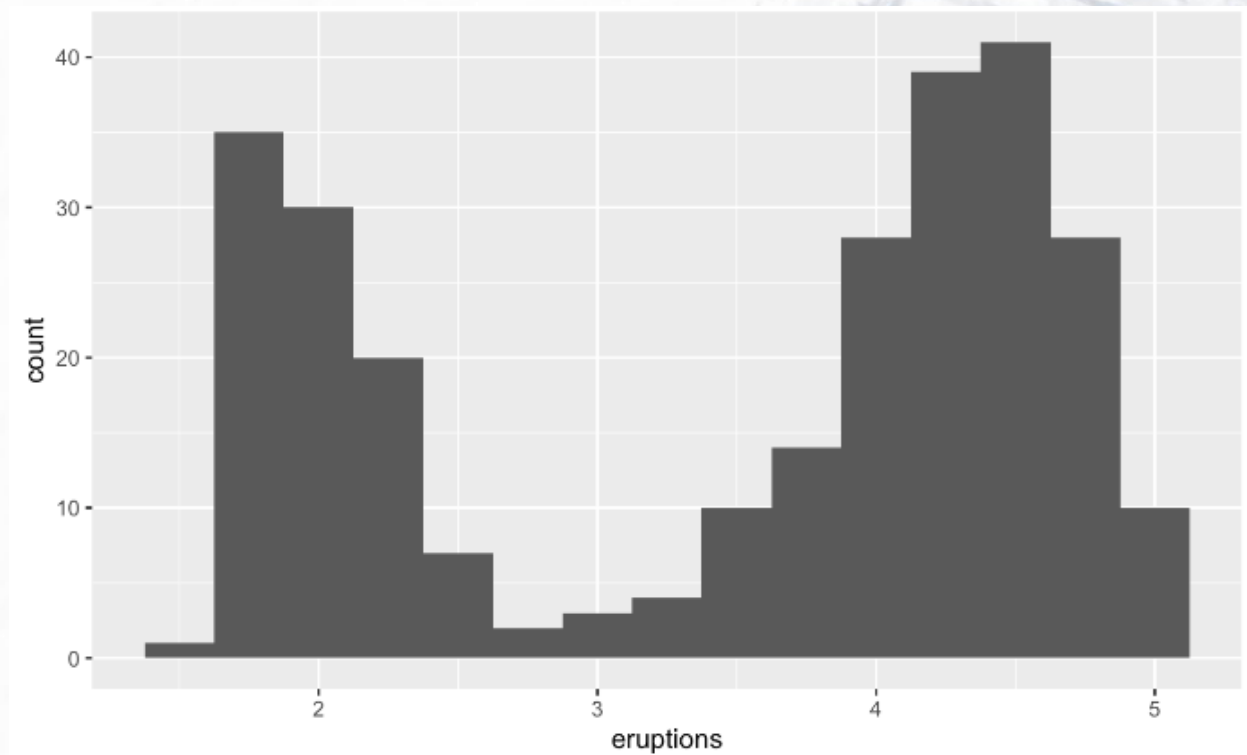
Variación – Valores típicos

- Para convertir esta información en preguntas útiles:
- ¿Qué valores son los más comunes? ¿Por qué?
- ¿Qué valores son raros? ¿Por qué? ¿Eso coincide con las expectativas?
- ¿Se puede ver algún patrón inusual? ¿Qué podría explicarlo?

Variación – Valores típicos

El siguiente histograma muestra la duración (en minutos) de 272 erupciones del géiser Old Faithful en el Parque Nacional de Yellowstone.

Los tiempos de erupción parecen estar agrupados en dos: erupciones cortas (de alrededor de 2 minutos) y erupciones largas (4-5 minutos).



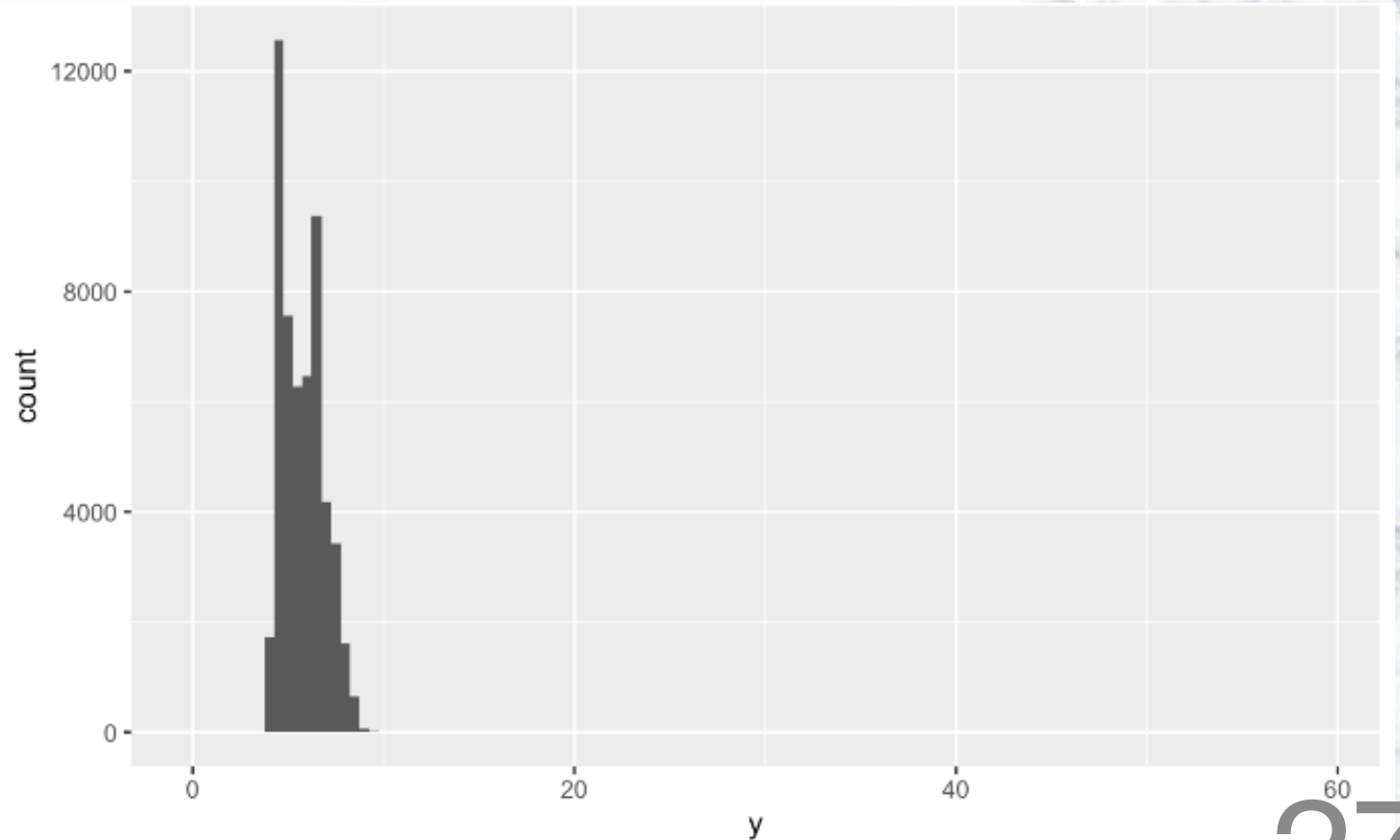
Variación – Valores inusuales

- Los valores atípicos (*outliers*) son observaciones que son inusuales; puntos de datos que no parecen encajar en el patrón.
- A veces, los valores atípicos son errores de entrada de datos, a veces son simplemente valores en los extremos que se observaron en la recolección de datos y otras veces sugieren nuevos descubrimientos importantes.

Variación – Valores inusuales

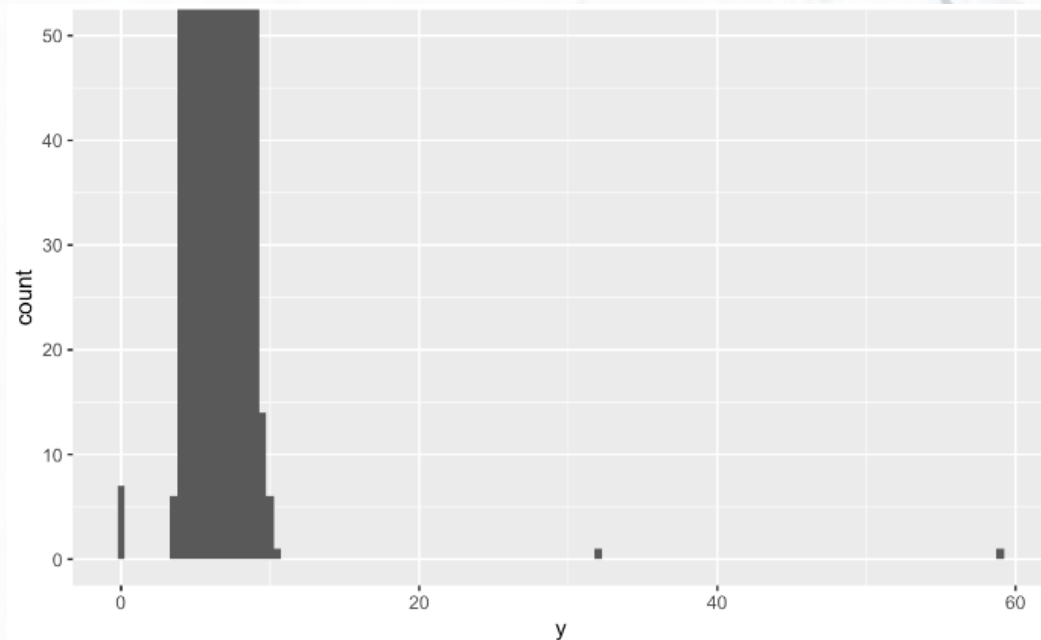
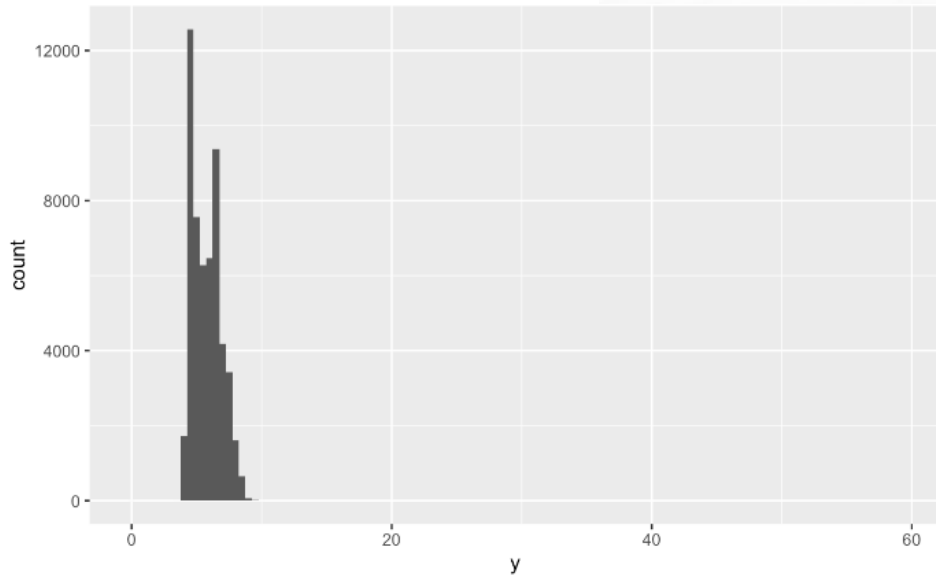
Cuando hay muchos datos, los valores atípicos a veces son difíciles de ver en un histograma.

Por ejemplo, en el siguiente histograma los límites en el eje x son inusualmente amplios



Variación – Valores inusuales

- Si se amplía el gráfico a valores más pequeños del eje y pueden observarse valores atípicos en 0, ~30, y ~60 que deben analizarse.



Valores faltantes

Si se encontraron valores inusuales en el conjunto de datos y simplemente se desea continuar con el resto de su análisis, hay dos opciones.

1) Elimine toda la fila con los valores extraños.

- No es recomendable porque el hecho de que una medida no sea válida no significa que todas las medidas lo sean.
- Además, si se tienen datos de baja calidad, aplicar este enfoque a cada variable, ¡es posible que no queden datos!

Valores faltantes

2) Declarar el valor como un valor faltante “NA” (*missing value*).

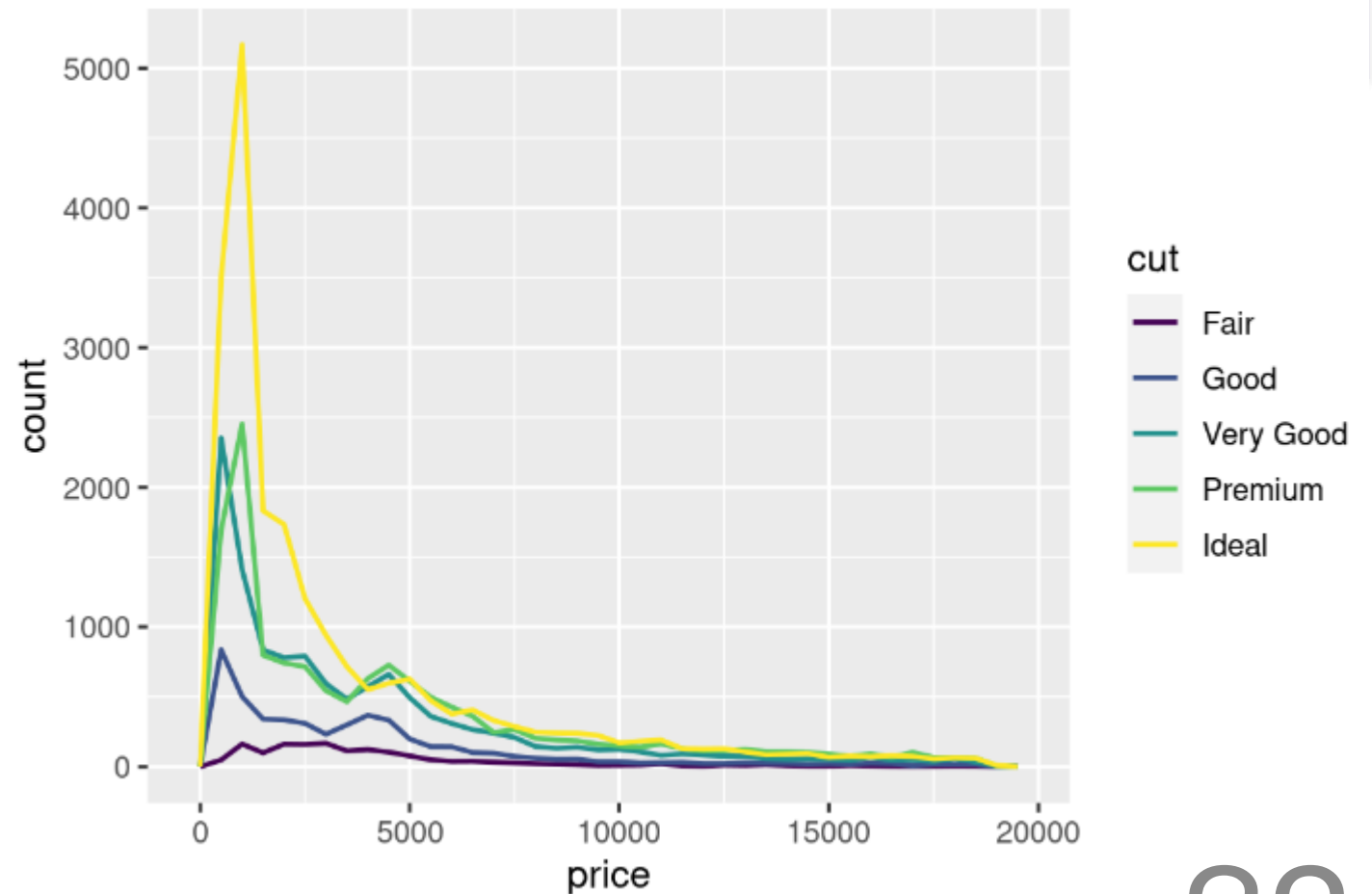
- Los programas de análisis y graficación de datos no los incluirán en los gráficos, pero sí advertirán que han sido eliminados.
- Otras veces, se desea comprender qué hace que las observaciones con valores faltantes sean diferentes a las observaciones con valores registrados.

Covariación

- Si la variación describe el comportamiento dentro de una variable, la covariación describe el comportamiento entre variables.
- La covariación es la tendencia de los valores de dos o más variables a variar juntos de manera relacionada.
- Una manera de detectar la covariación es visualizar la relación entre dos o más variables.

Covariación

- Por ejemplo, se puede graficar la relación entre el precio (variable *price*) de los diamantes en los diferentes cortes (variable *cut*).

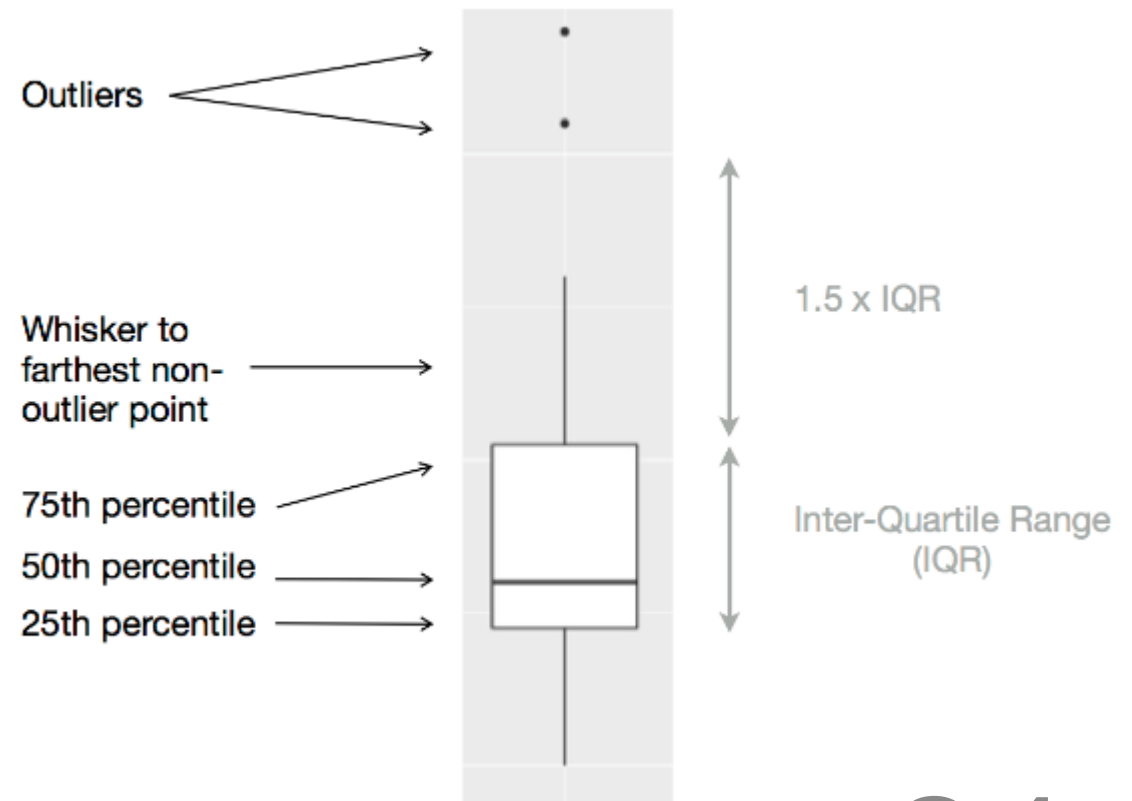


Covariación

- Para mostrar la distribución de una variable numérica desglosada por una variable categórica se puede utilizar el diagrama de caja (boxplot).
- Un diagrama de caja es un tipo de abreviatura visual para una distribución de valores que es popular entre los estadísticos.
- Cada boxplot contiene:

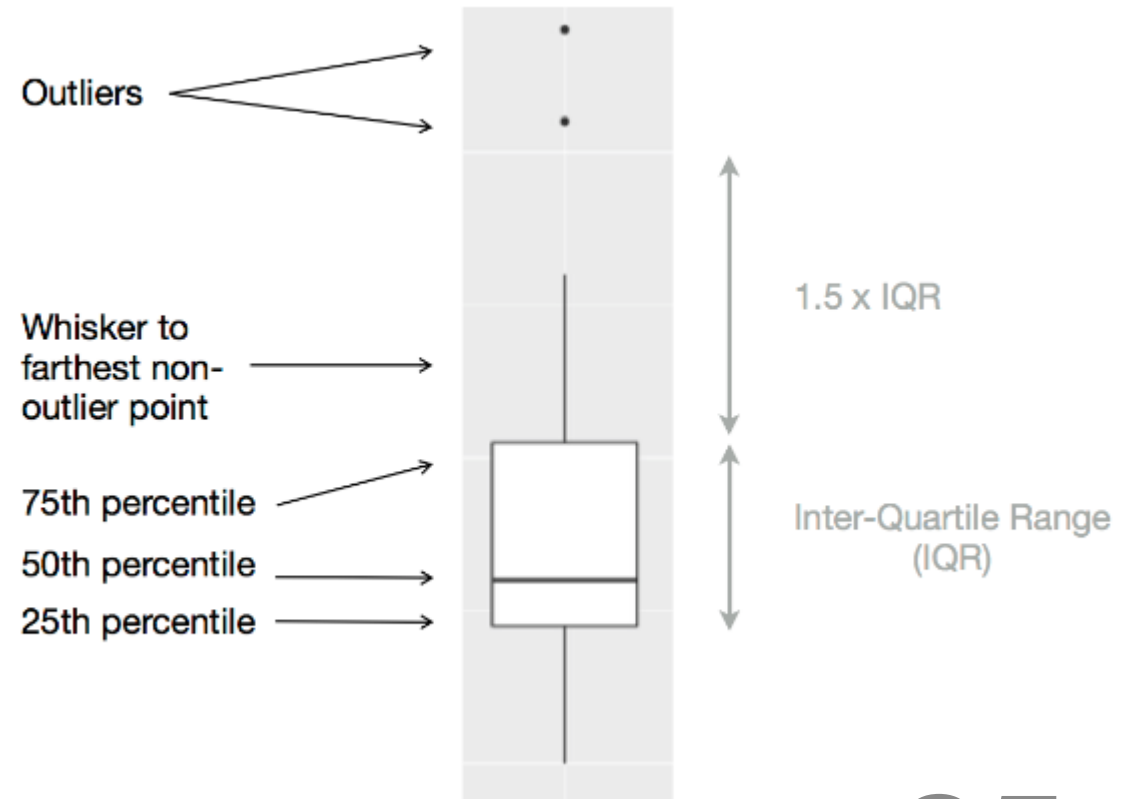
Covariación – Cada boxplot contiene:

- Una caja que va desde el percentil 25 de la distribución hasta el percentil 75, distancia conocida como rango intercuartílico (IQR).
- En el medio del cuadro hay una línea que muestra la mediana, es decir, el percentil 50.
- Estas tres líneas dan una idea de la dispersión de la distribución y si la distribución es o no simétrica con respecto a la mediana o sesgada hacia un lado.



Covariación – Cada boxplot contiene:

- Pueden aparecer marcas visuales que muestran observaciones que caen más de 1,5 veces el IQR desde cualquier borde del cuadro.
- Estos puntos atípicos (outliers) se trazan individualmente.
- Una línea (o bigote) que se extiende desde cada extremo de la caja y va hasta el punto no atípico más alejado de la distribución.

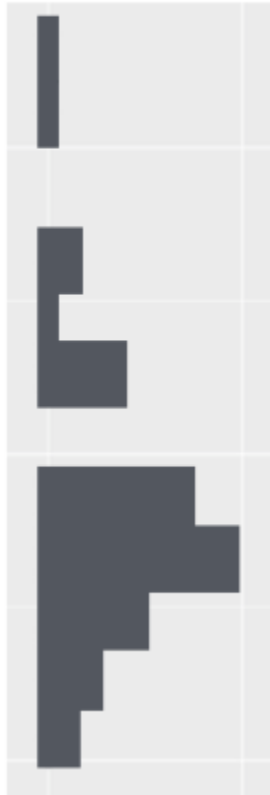


Covariación – Cada boxplot contiene:

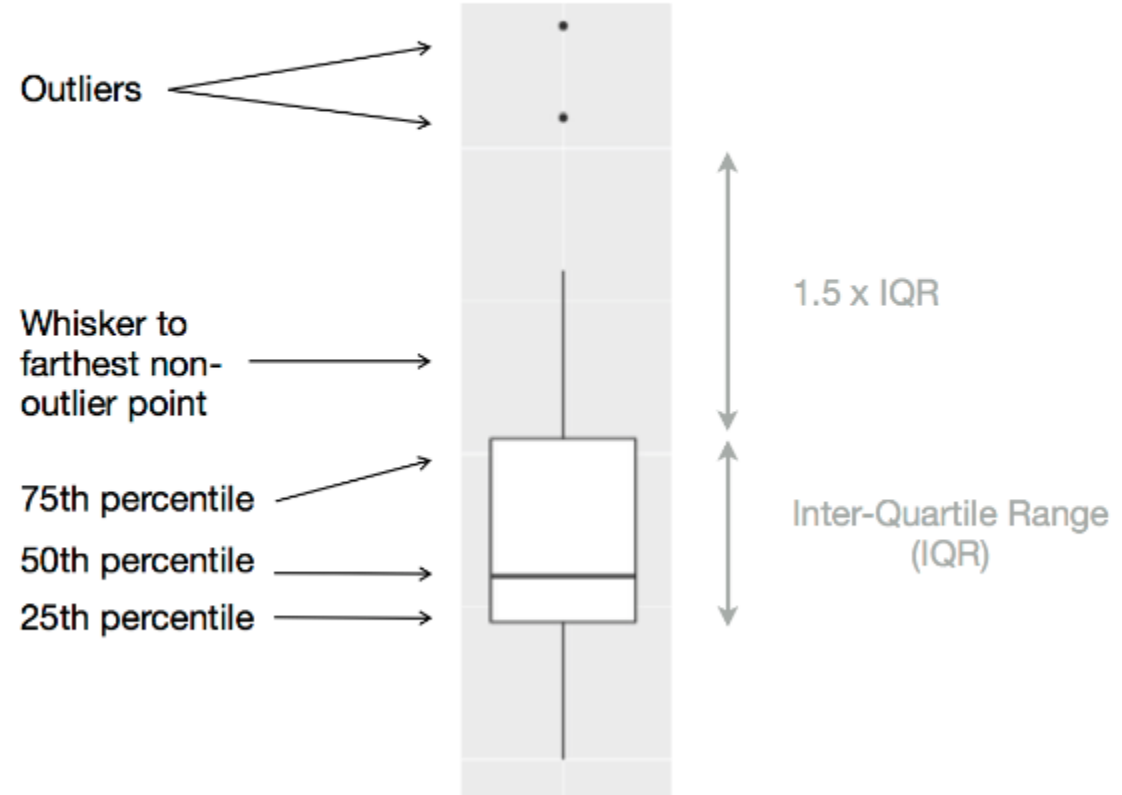
The actual values in a distribution



How a histogram would display the values (rotated)

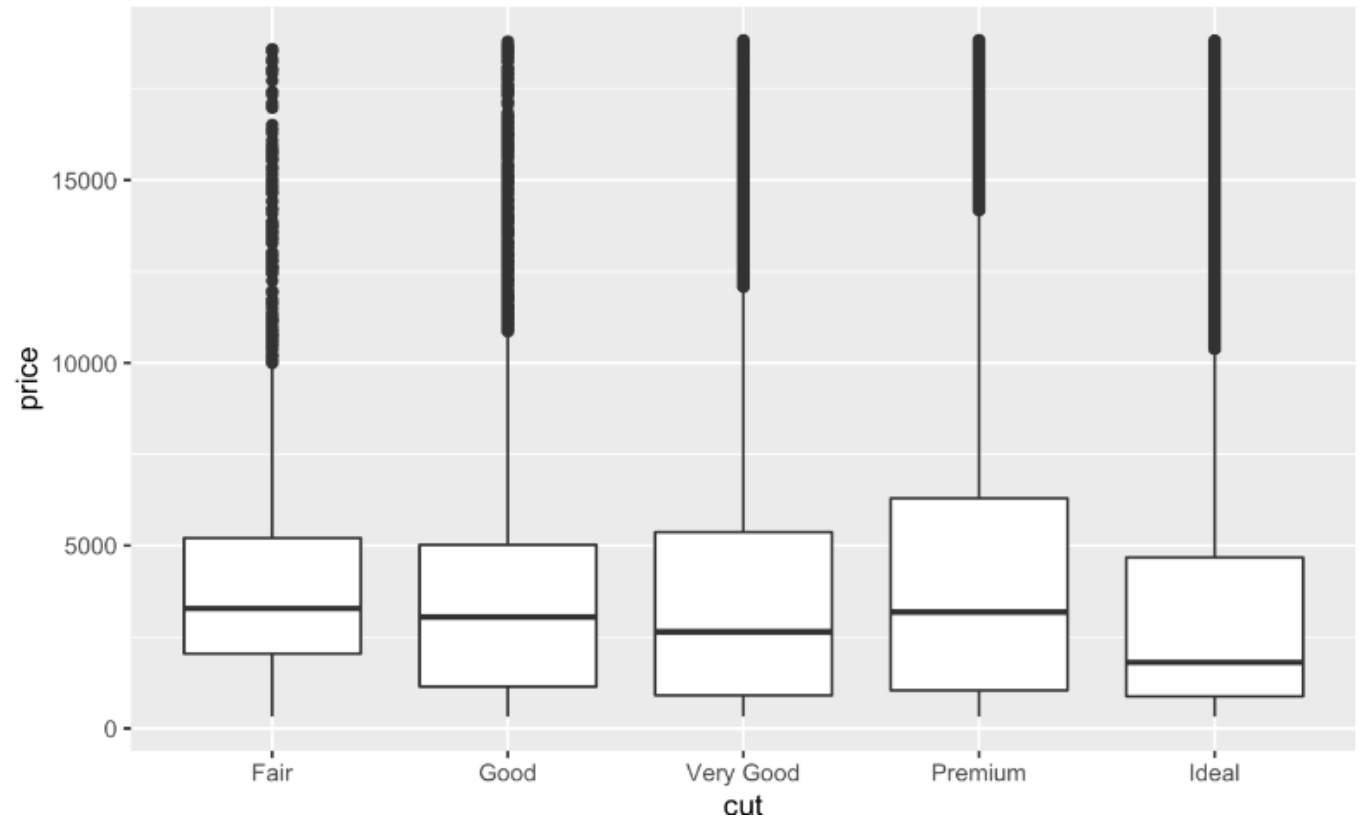


How a boxplot would display the values



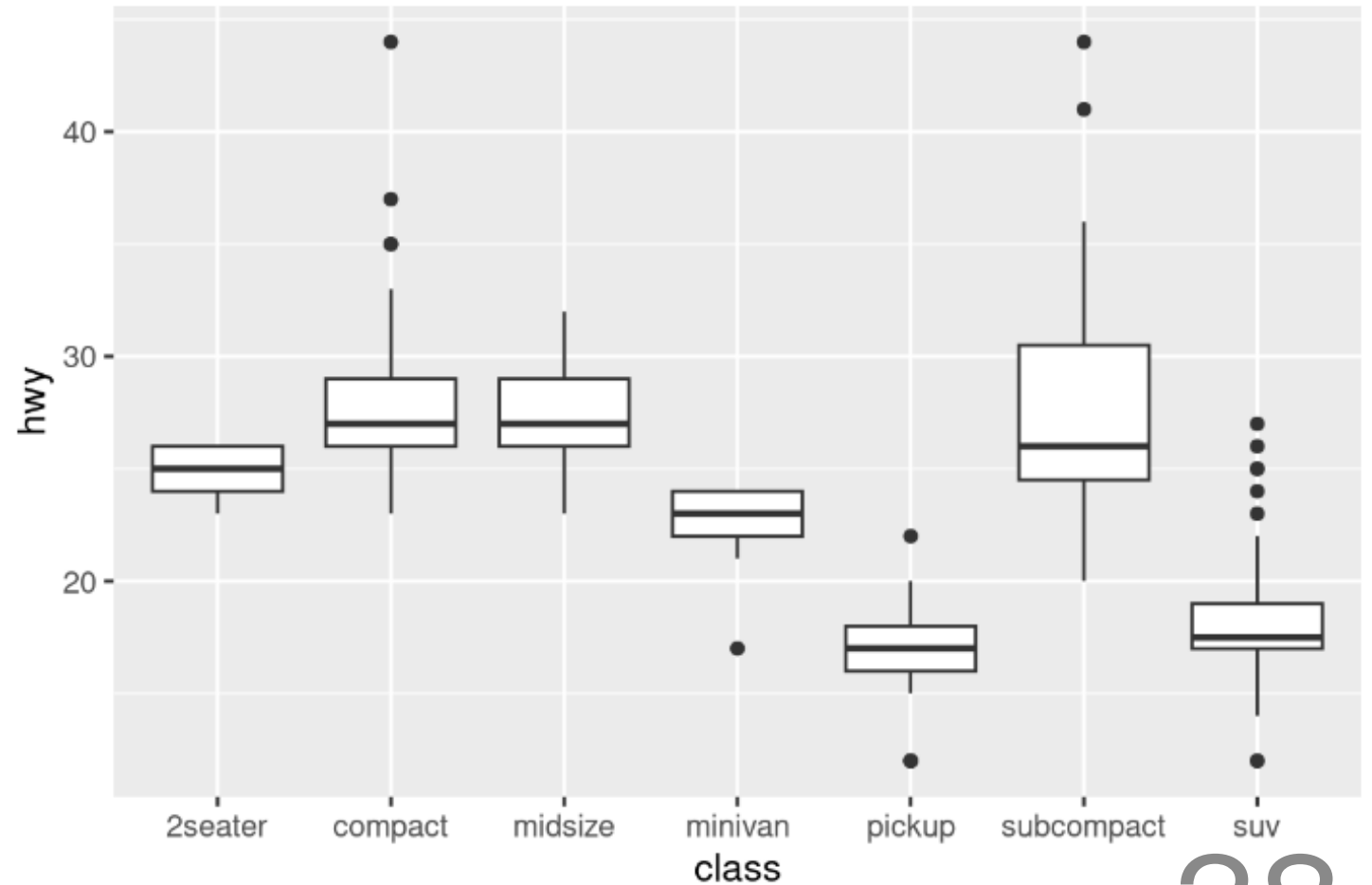
Covariación

- La distribución de la variable numérica *price* desglosada por la variable categórica *cut*:
- Se ve menos información de la distribución de cada categoría que en un polígono de frecuencia, pero es más sencillo de compararlas.



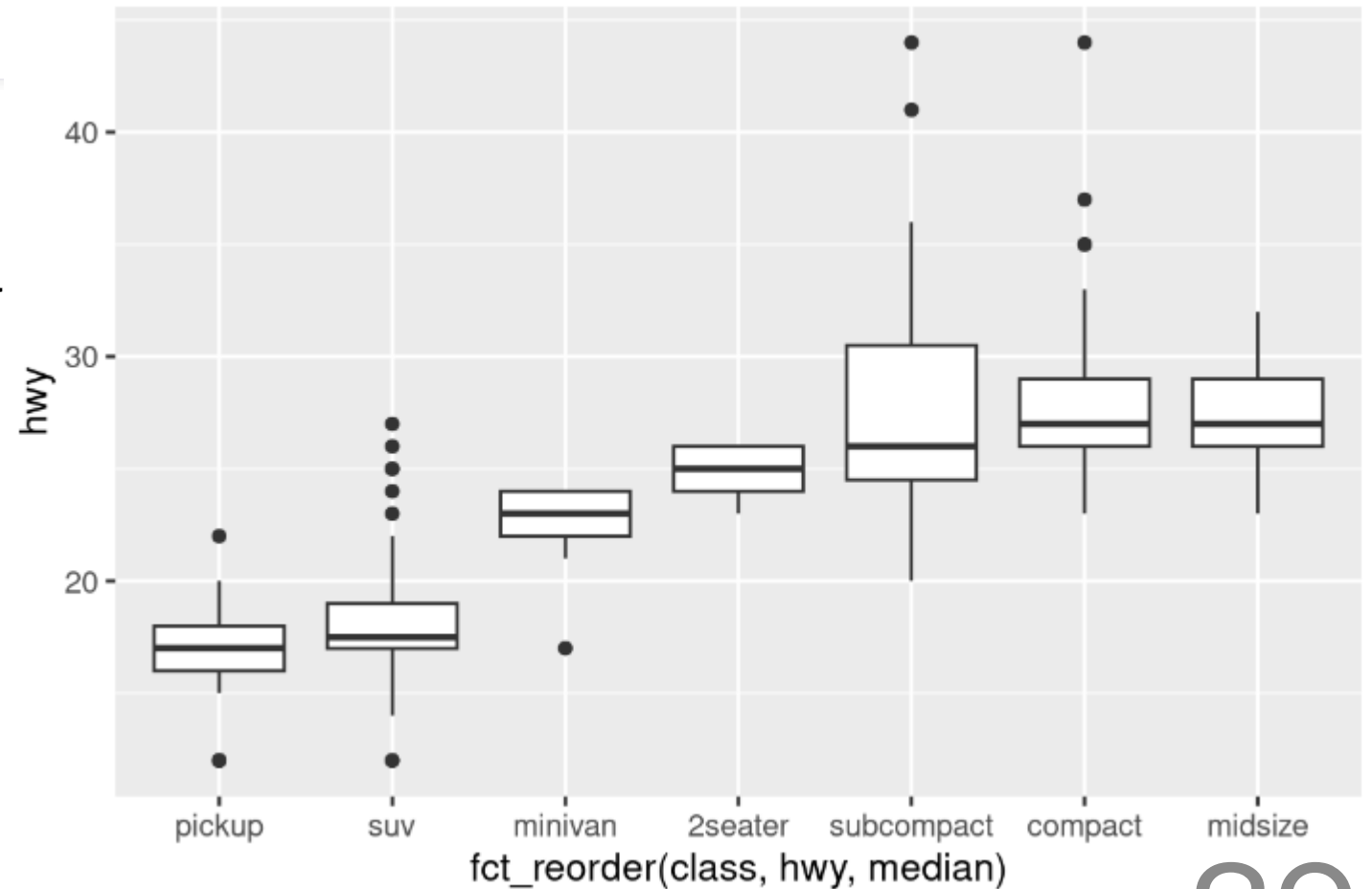
Covariación

- Algunas variables categóricas no tienen orden intrínseco.
- Por ejemplo, el tipo de automóvil.



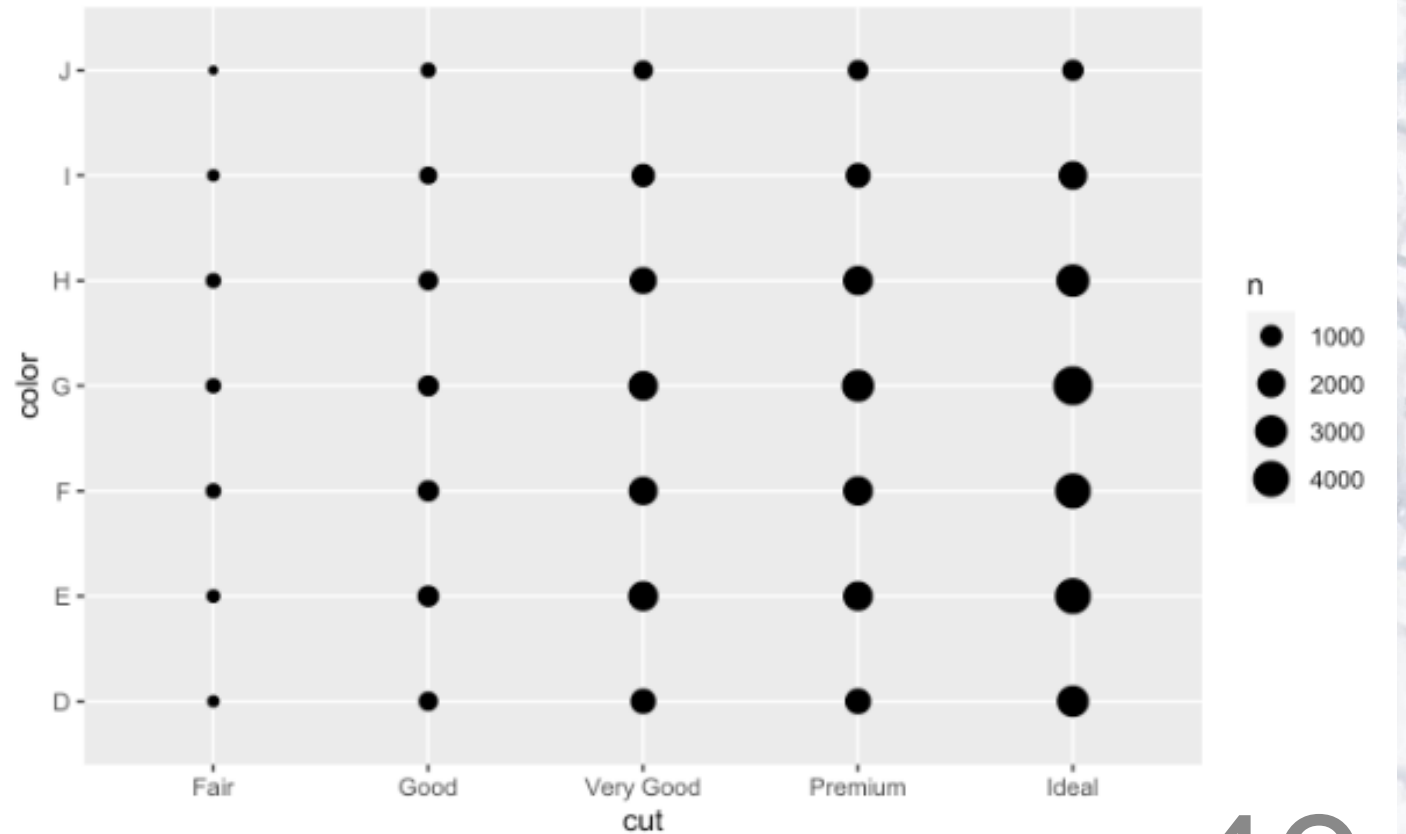
Covariación

- Entonces, puede ser útil reordenar las categorías para que el despliegue sea más informativo.



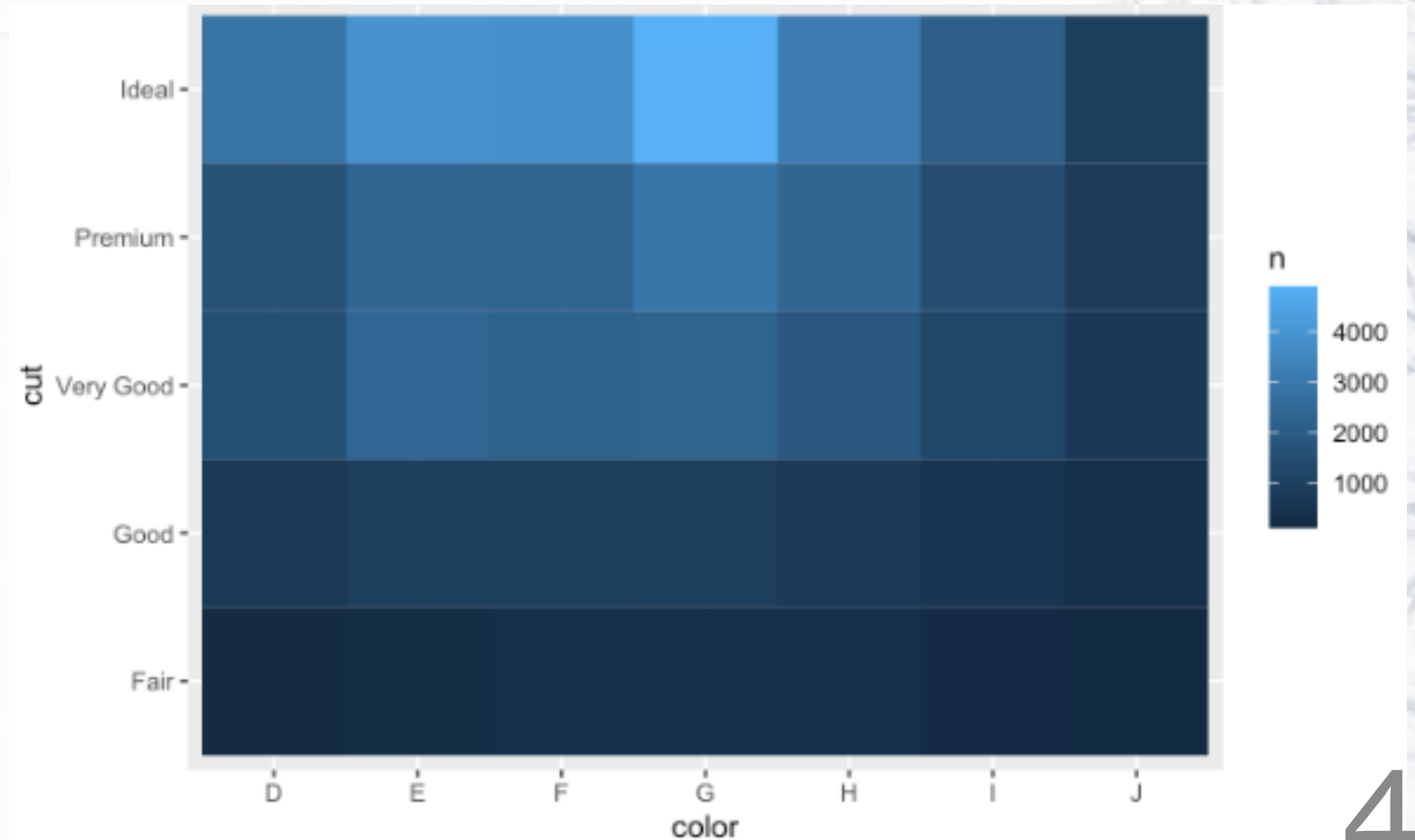
Covariación

- Para visualizar la covariación entre dos variables categóricas se debe contar el número de observaciones de cada combinación.
- La covariación se representará por el tamaño de los círculos.



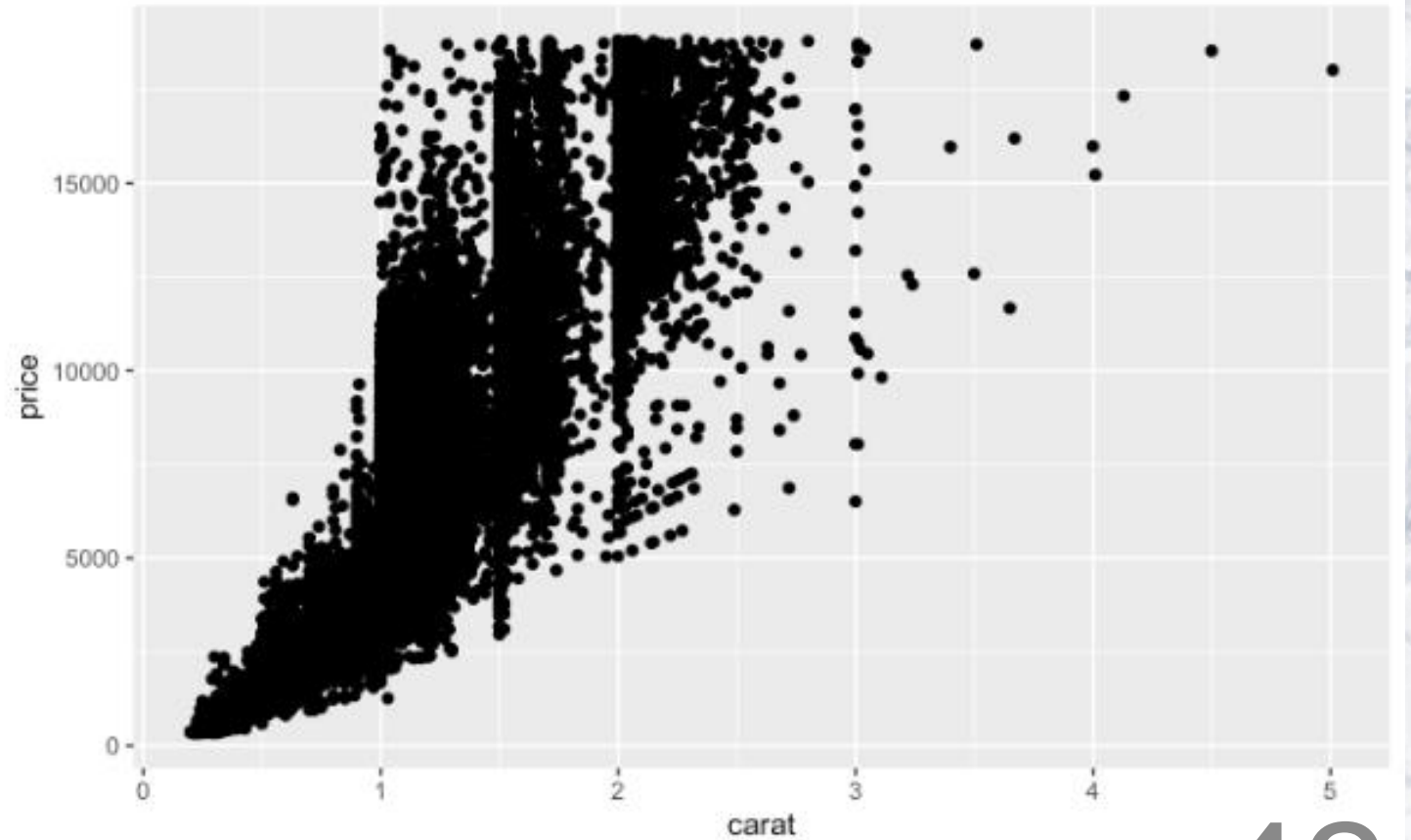
Covariación

- Otra variante es representar los contadores con colores de diferente intensidad:



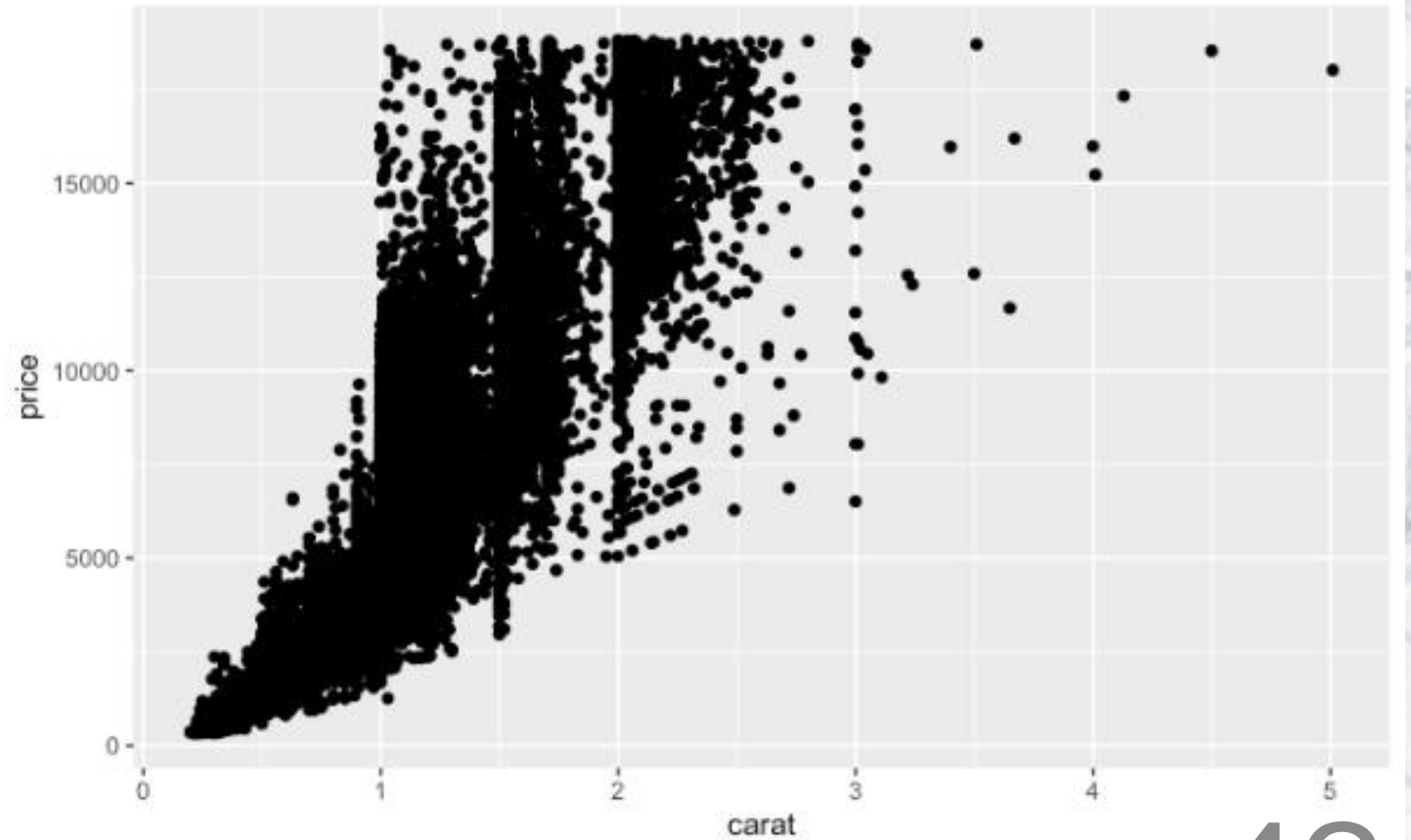
Covariación

- Para visualizar la covariación entre dos variables numéricas se puede usar un gráfico de dispersión (*scatterplot*), que se presenta como patrones de puntos



Covariación

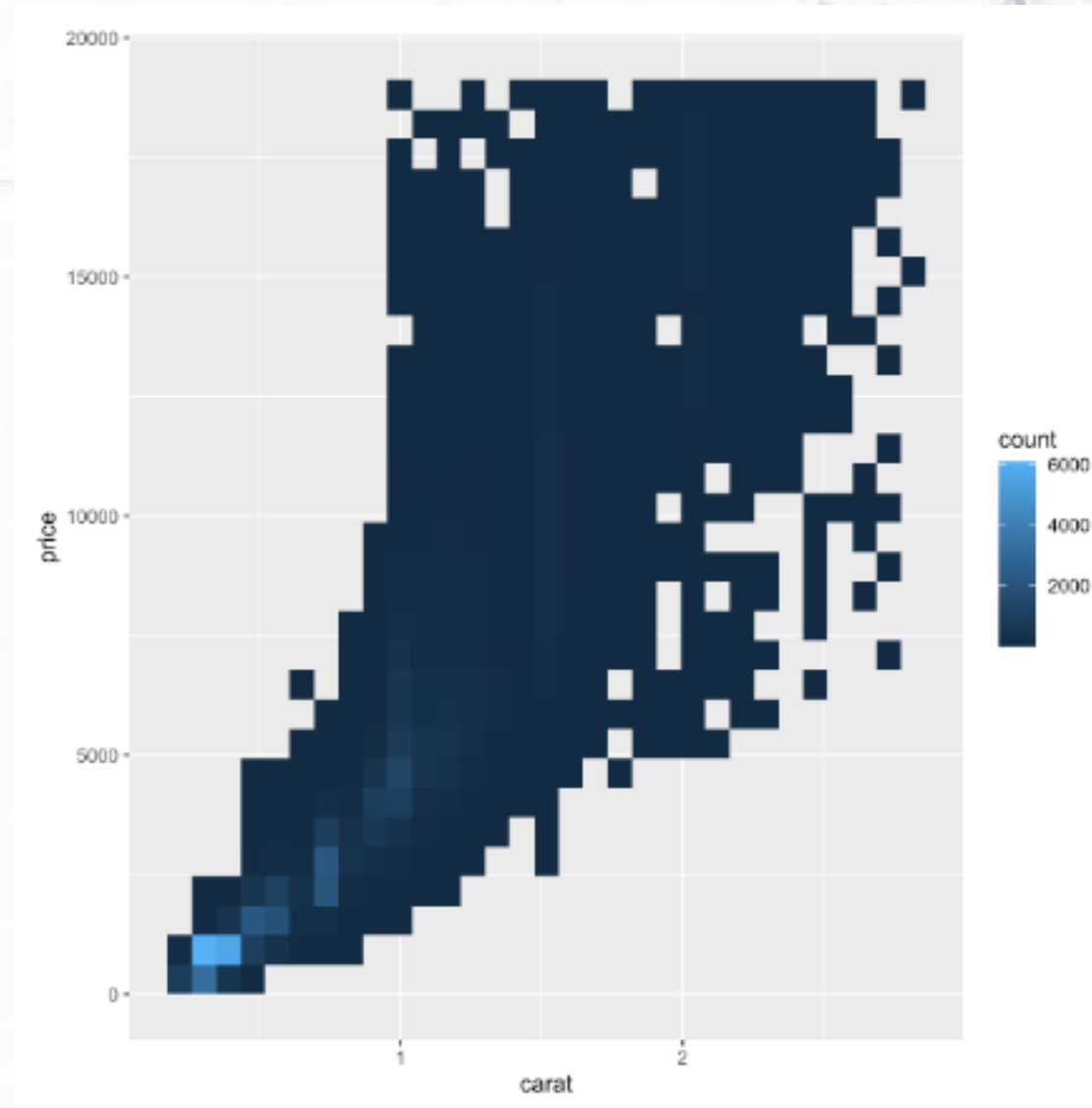
- Los diagramas de dispersión se vuelven menos útiles a medida que crece el tamaño de su conjunto de datos, porque los puntos comienzan a superponerse y se acumulan en áreas negras.



Análisis Exploratorio de Datos (EDA)

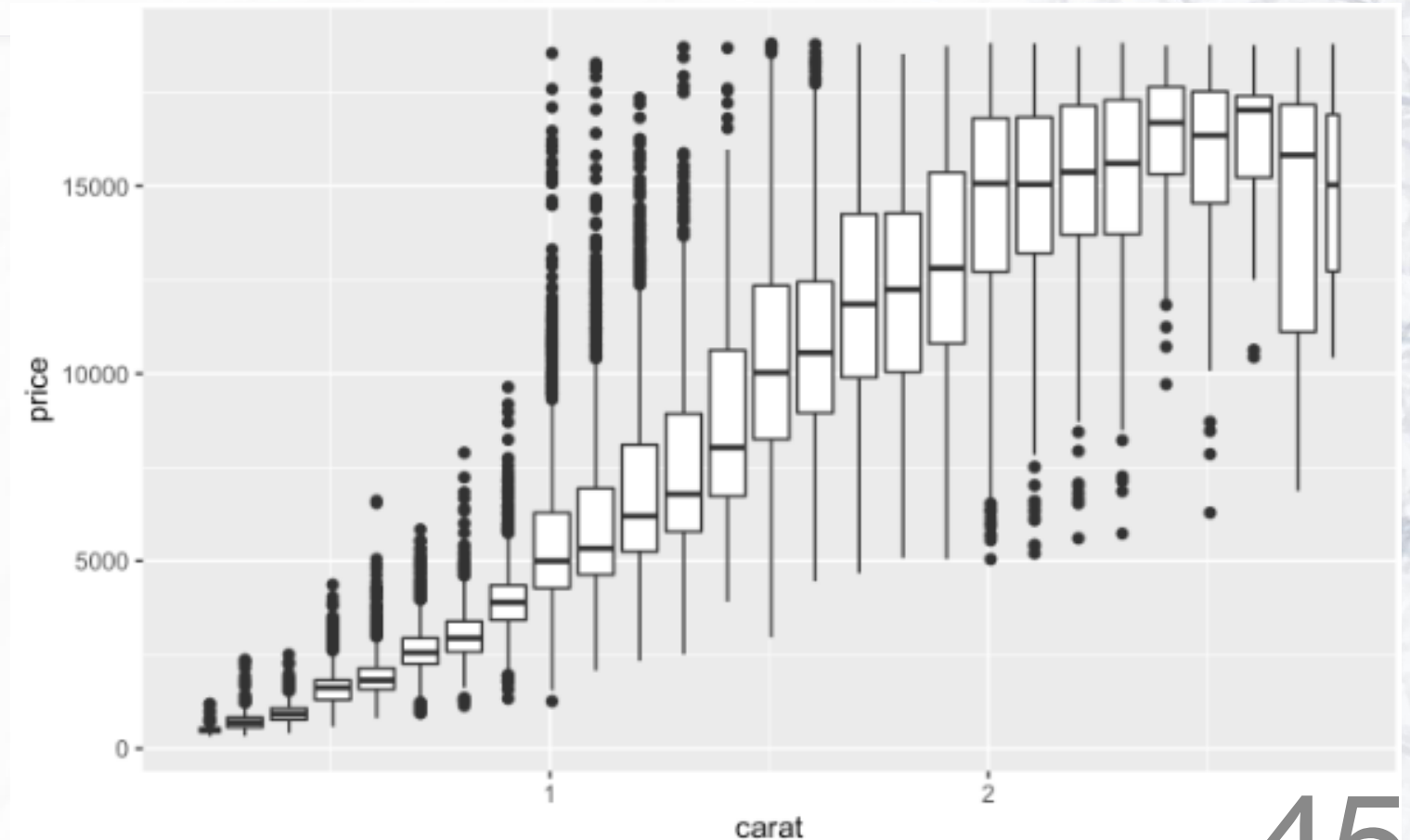
Covariación

Una opción es usar
color para identificar
la cantidad de puntos :



Covariación

Otra opción es agrupar una variable numérica para que actúe como una variable categórica.



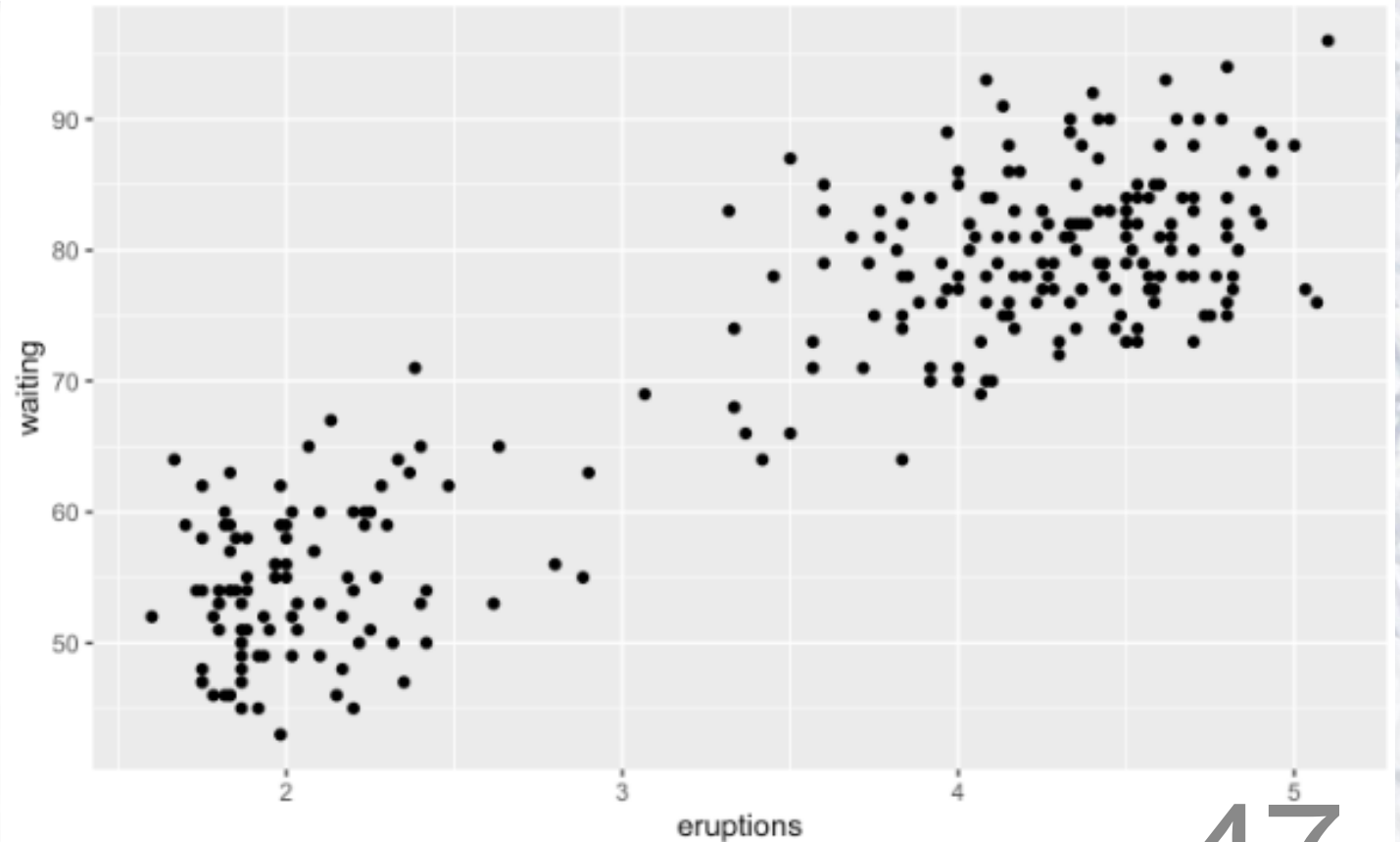
Patrones

Los patrones en los datos brindan pistas sobre las relaciones. Si existe una relación sistemática entre dos variables, aparecerá como un patrón. Entonces:

- ¿Podría este patrón deberse a una coincidencia (es decir, al azar)?
- ¿Cómo se puede describir la relación implícita en el patrón?
- ¿Qué tan fuerte es la relación implícita en el patrón?
- ¿Qué otras variables podrían afectar la relación?
- ¿Cambia la relación si se observan subgrupos individuales de datos?

Patrones

Un gráfico de dispersión de la duración de las erupciones de Old Faithful frente al tiempo de espera entre erupciones muestra un patrón: los tiempos de espera más largos se asocian con erupciones más largas.



Patrones

- Los patrones proporcionan una de las herramientas más útiles para los científicos de datos porque revelan la covariación.
- Si se piensa en la variación como un fenómeno que crea incertidumbre, la covariación es un fenómeno que la reduce.
- Si dos variables covarían, se puede usar los valores de una variable para hacer mejores predicciones sobre los valores de la segunda.
- Si la covariación se debe a una relación causal (un caso especial), se puede usar el valor de una variable para controlar el valor de la segunda.

Referencias

- Wickman, H., Grolemond, G. (2017). *R for Data Science. Visualize, model, transform, tidy, and import data.* O'Reilly.
Disponible en: <https://r4ds.hadley.nz/>
- Lawson, J. (2014). *Design and Analysis of Experiments with R* (Vol. 115). CRC press.