Revised: 3/21/2021

Descriptive Statistics

Michael C. Hackett

Assistant Professor, Computer Science

Community College of Philadelphia

ggplot2

- ggplot2 is a data visualization package in R's tidyverse

- Allows for declaratively creating graphics
  - Based on the text The Grammar of Graphics

- *"You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details."*
  - Project homepage: https://ggplot2.tidyverse.org/

ggplot2

- ggplot2 is installed along with the tidyverse:

```
install.packages("tidyverse")
```

- Can be installed as a stand-alone package:

```
install.packages("ggplot2")
```

- Extensions:

  https://exts.ggplot2.tidyverse.org/gallery/

ggplot2

- ggplot2 is loaded along with the rest of the tidyverse:
  ```
  library(tidyverse)
  ```

- Can be loaded by itself:
  ```
  library(ggplot2)
  ```

- ggplot2 has a sample data frame for demonstration purposes
  - The **mpg** dataset contains observations collected by the US Environmental Protection Agency on 38 models of cars

ggplot2

- If tidyverse was loaded:

```
library(tidyverse)
ggplot2::mpg
```

- If ggplot2 was loaded by itself:

```
library(ggplot2)
mpg
```

- We’ll assume ggplot2 was loaded by itself for the remainder of the lecture

ggplot2

- We begin creating a plot with the **`ggplot()`** function
  - This creates a coordinate system that layers can be added on to

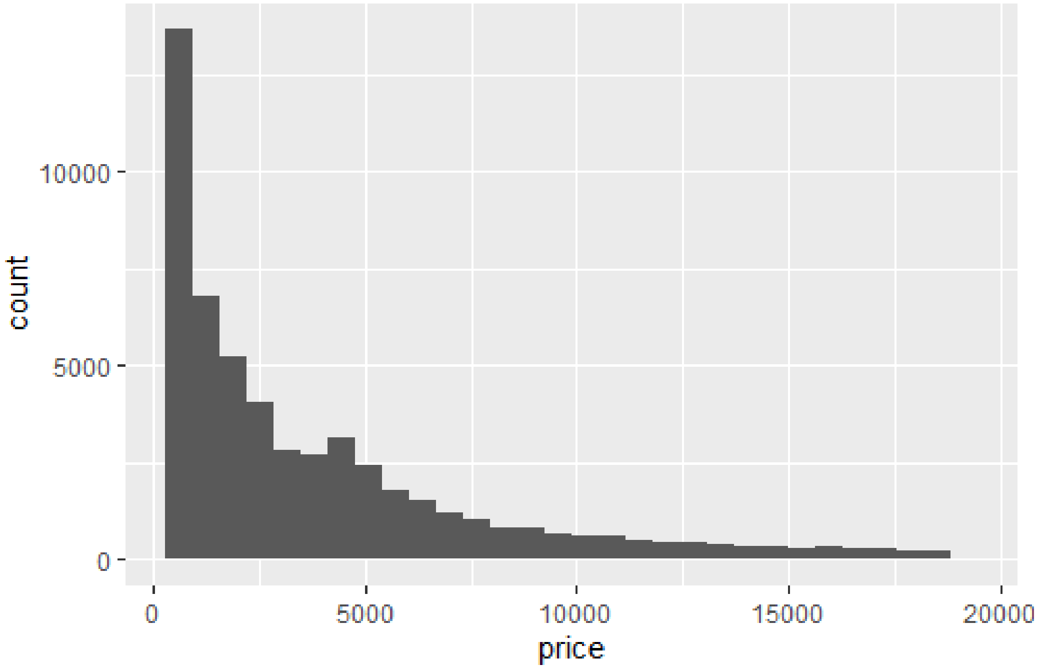
- The first argument to the ggplot function is the data we wish to plot

```
ggplot(data = mpg)
```

- Now that the plot has its data, layers are added that specify how to data is to be displayed.
  - Layers of data are referred to as *geometries* or ***geoms***

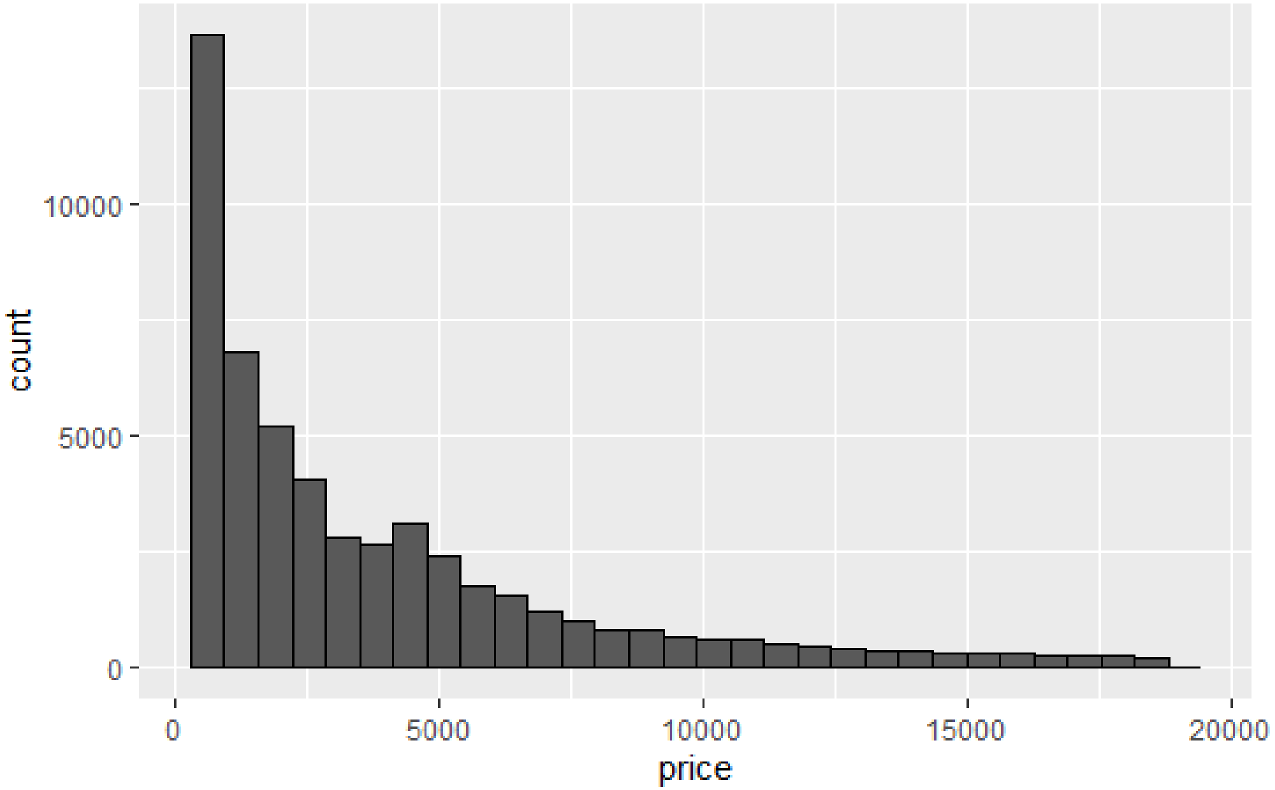
Histograms

- **Histograms** are used to visualize the distribution of a numerical variable by grouping data into “bins”
  - Histograms show **data density**; Higher bars = fuller bins



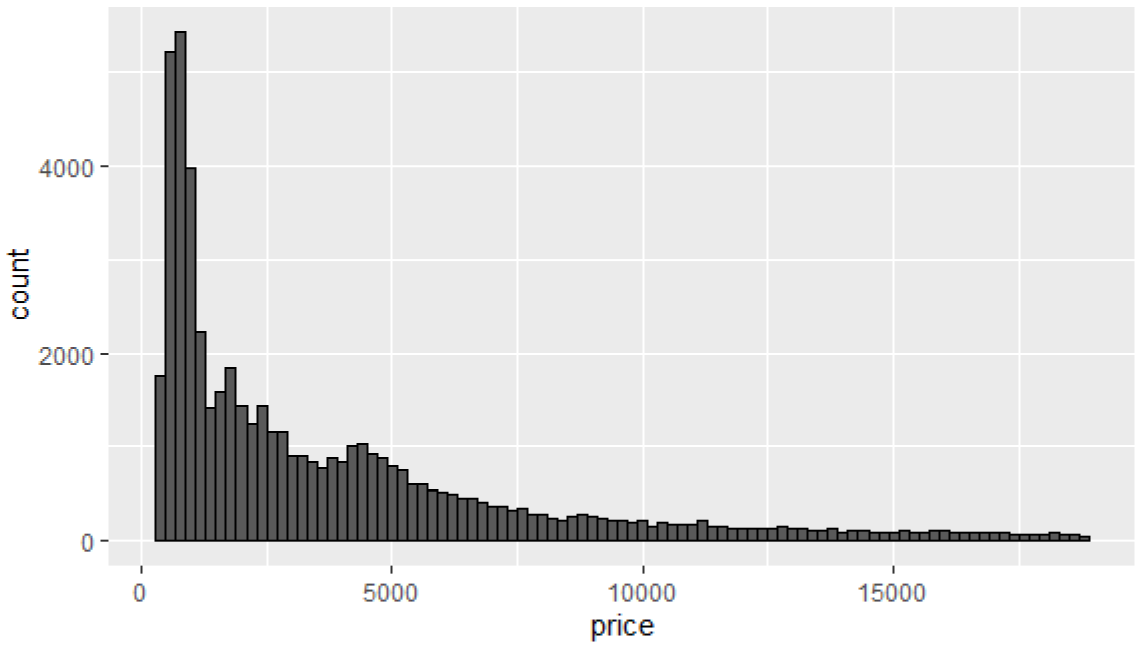
Histograms

- Add borders for better visibility



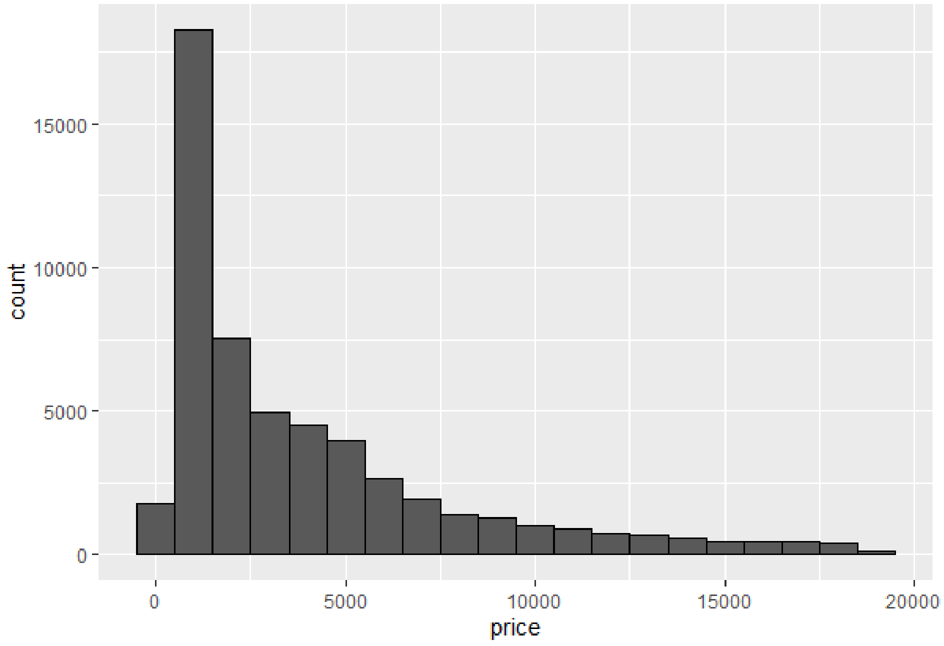
Histograms

- This histogram has binwidths of 200
  - Provides greater detail about how data is distributed, but sometimes, less is more



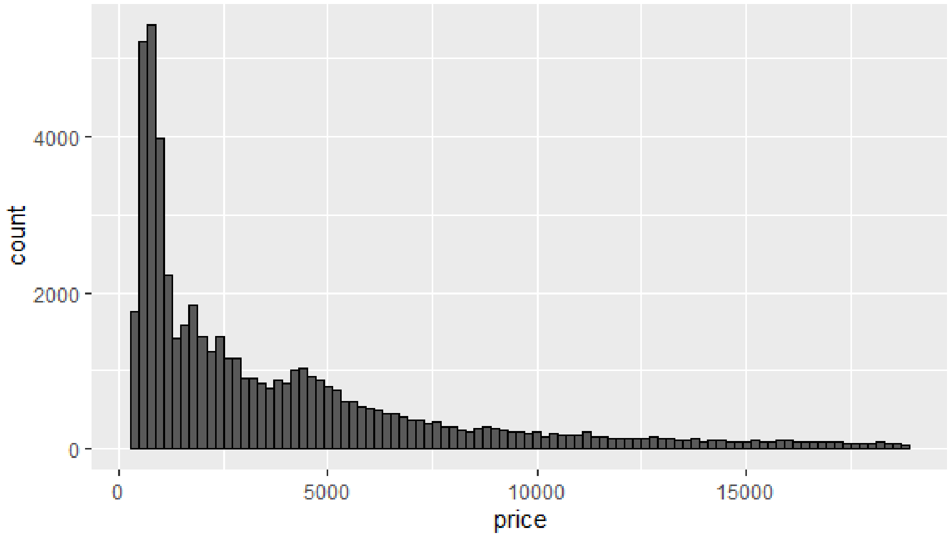
Histograms

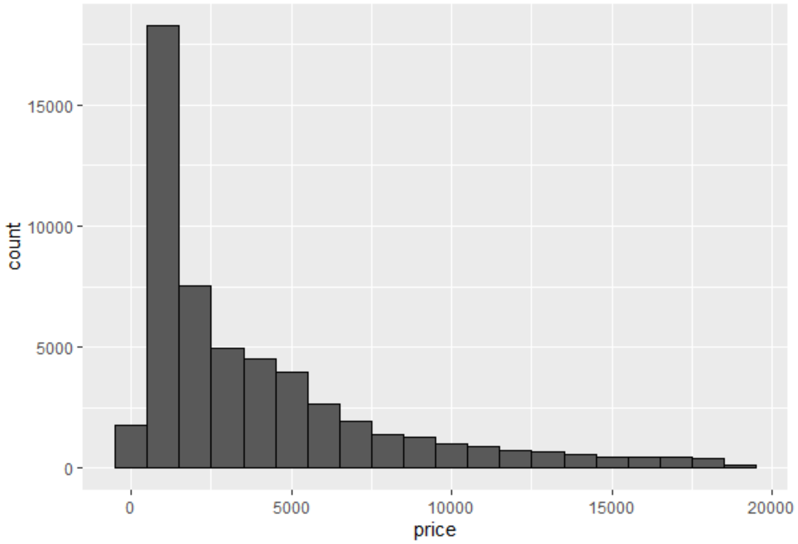
- This histogram has binwidths of 1000



Histograms

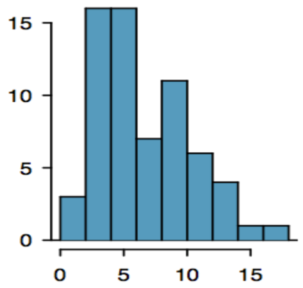
- Both are essentially telling the same story, but one histogram tells it with greater detail than the other.
  - Sometimes, less detail makes it easier to "digest" what the visualization is saying.



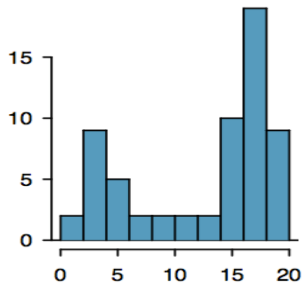


Histograms

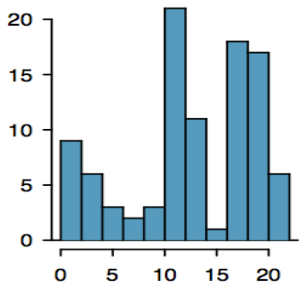
- The **modality** of a distribution is one way to describe its shape
  - *Unimodal*: One prominent peak
  - *Bimodal*: Two prominent peaks
  - *Multimodal*: More than two prominent peaks
  - *Uniform*: No prominent peaks



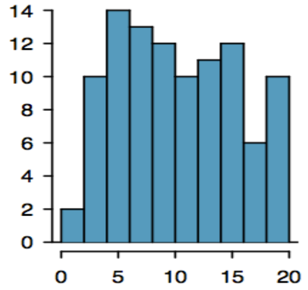
Unimodal



Bimodal



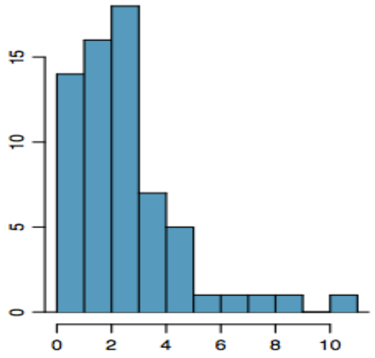
Multimodal



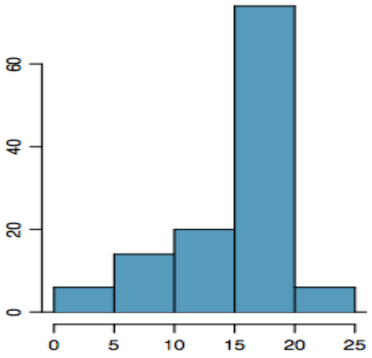
Uniform

Histograms

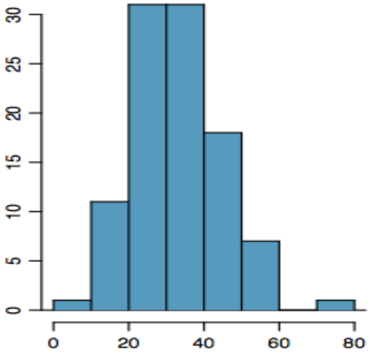
- The **skew** of a distribution is another way to describe its shape
  - *Right Skewed*: The data trails off to the right
  - *Left Skewed*: The data trails off to the left
  - *Symmetric*: The data trails off in both directions (roughly) equally



Right skewed



Left skewed



Symmetric

Mean

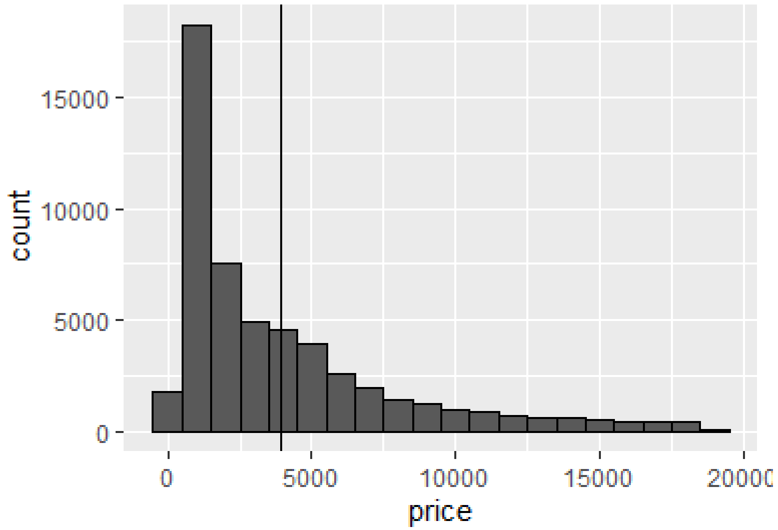
- The **mean** (or average) is one method to find the center of a distribution.
  - The sum of the observed values divided by the total number of observed values.

- The mean is denoted by $\bar{x}$
  - More specifically, the *sample mean* is denoted by $\bar{x}$
  - The *population mean* is denoted by μ

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Mean

- R's [mean function]
  mean(diamonds$price)
  ~3932.80
- A histogram with a vertical line geometry at the x-intercept of the mean:



Variance and Standard Deviation

- The distance of an observation from the mean is called **deviation**.

$$deviation = x_n - \bar{x}$$

- The average of the squared deviations from the mean is called the **variance (s^2)**.
  - Variance describes how spread out the data in a distribution is around the mean

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1}$$

Variance and Standard Deviation

- R's (variance) var function

  `var(diamonds$price)`

  ~15915629.42 (A very high variance)

- The square root of the variance is called the **standard deviation (s)**.
  - The standard deviation is the typical deviation of any data from the mean.

$$s = \sqrt{s^2}$$

Variance and Standard Deviation

- R's (standard deviation) sd function

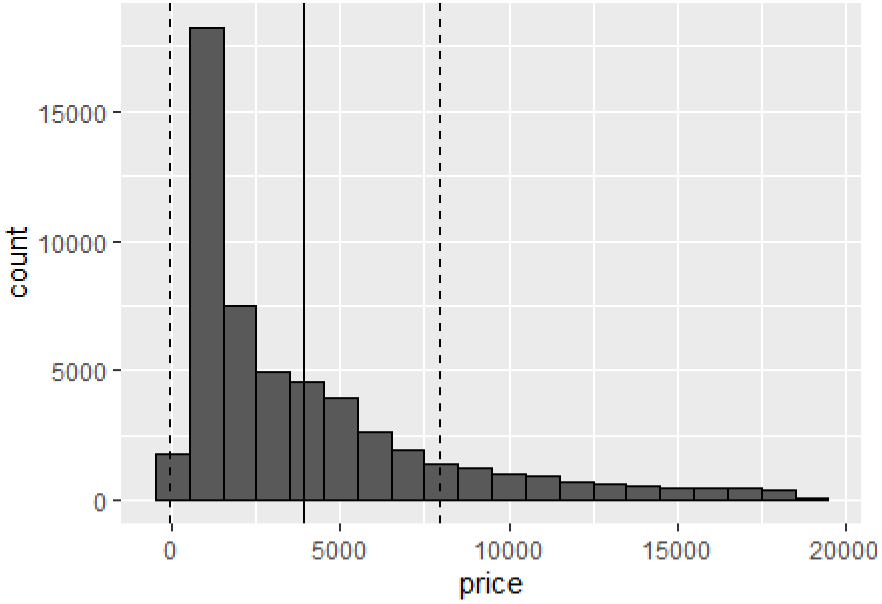
  `sd(diamonds$price)`

  ~3989.43 (A very high variance)

- A general rule of thumb is that 70% of the data in a distribution will be within one standard deviation of the mean; 95% will be within two standard deviations.
  - We'll revisit this in more detail when we get into probability

Variance and Standard Deviation

- ~70% of the data is between the dashed lines (one standard deviation away from the mean)

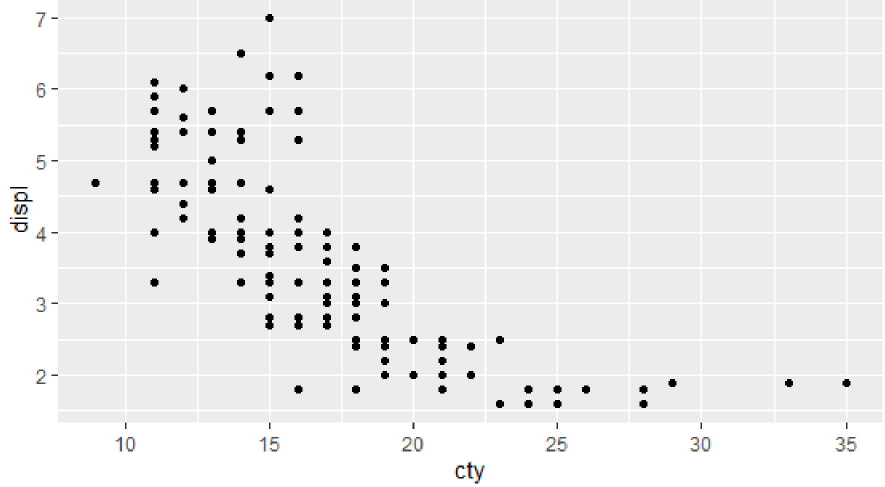


Variance and Standard Deviation

- Symbols:
  - Sample variance: s^2
  - Sample standard deviation: s
  - Population variance: σ^2
  - Population standard deviation: σ

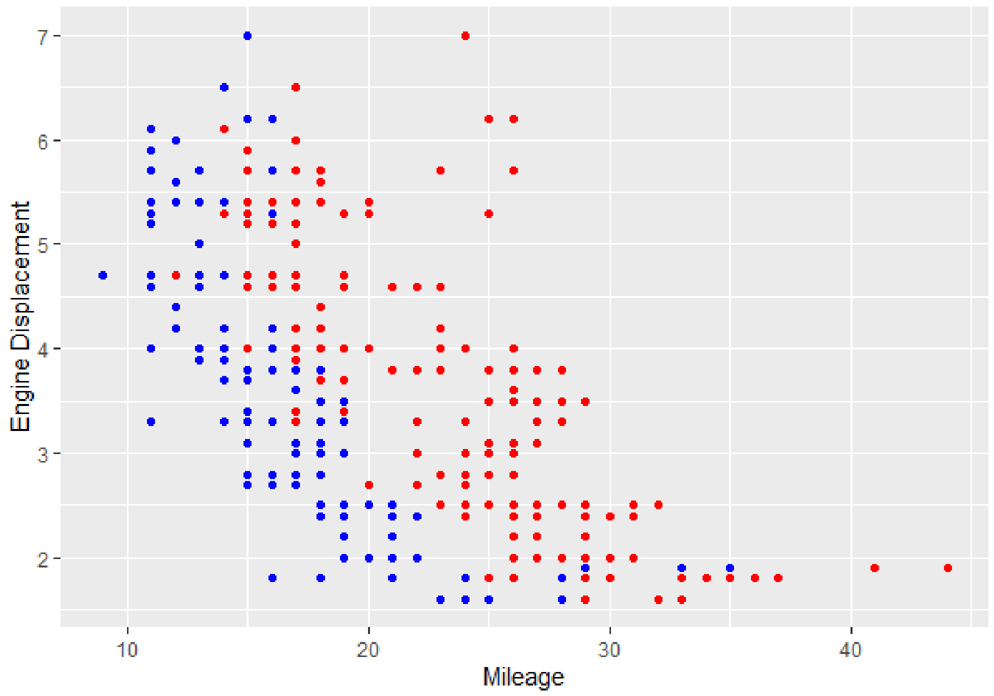
Scatterplots

- Scatterplots visualize the association between two numerical variables.
  - This scatterplot shows the relationship between engine displacement and city milage
  - City milage improves in cars with smaller engine displacement



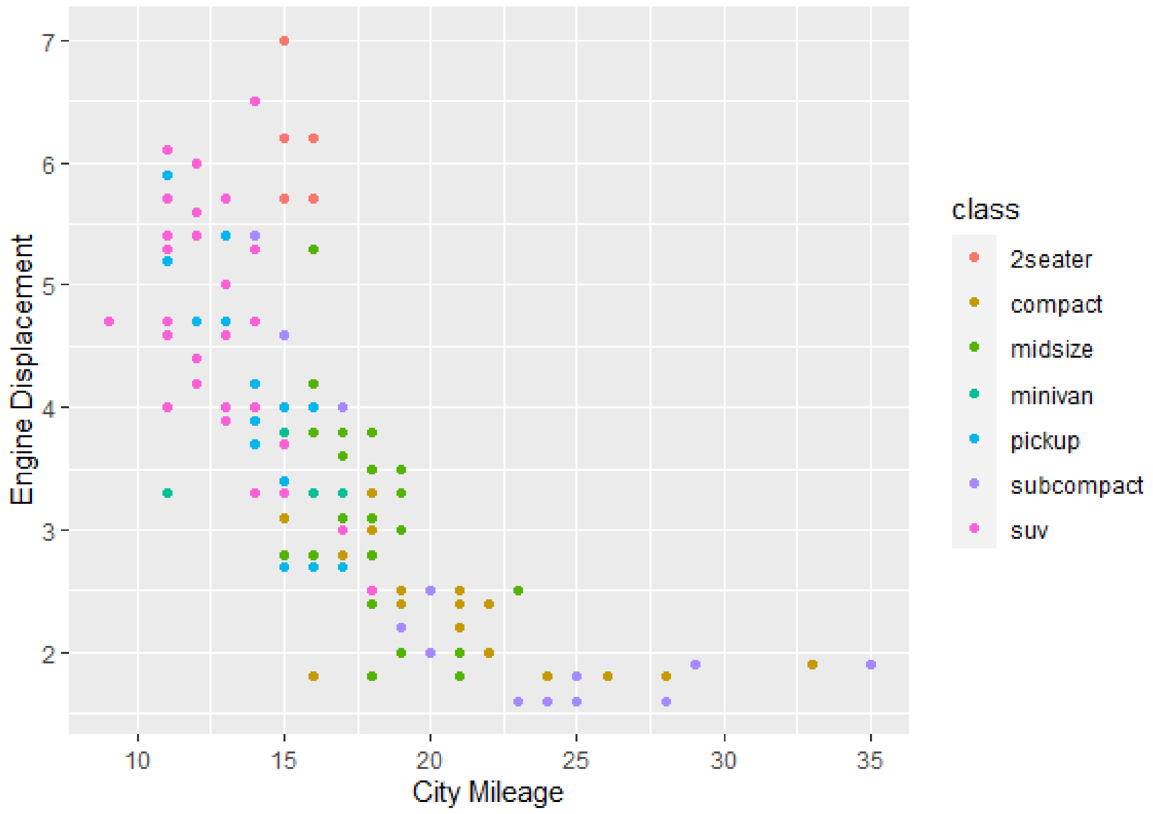
Scatterplots

- Plotting both city (blue) and highway (red) milage
  - What associations does this scatterplot show?



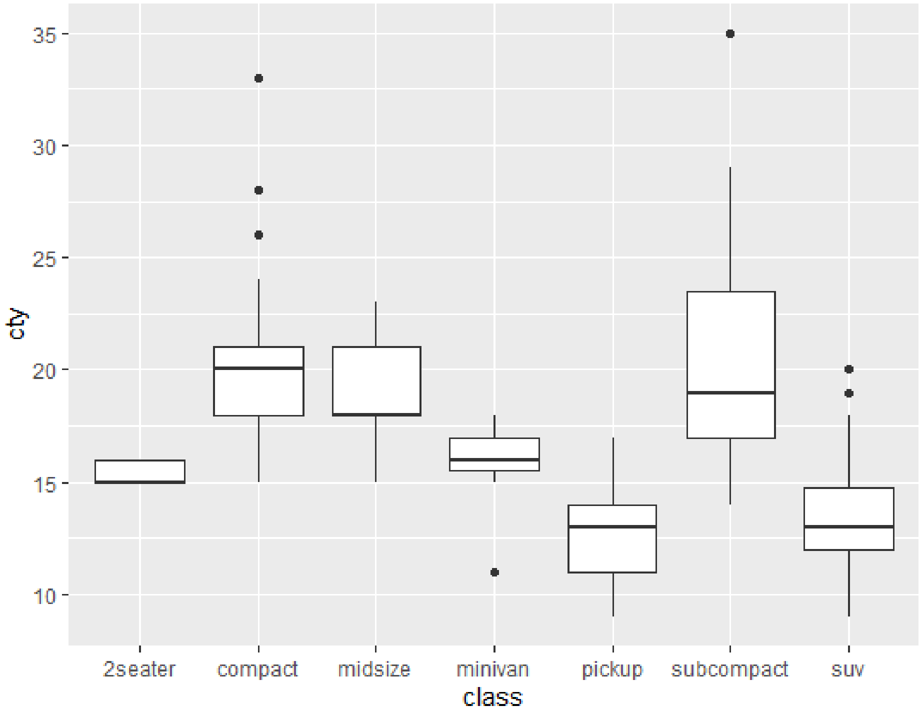
Scatterplots

- What associations does this scatterplot show?



Box Plots

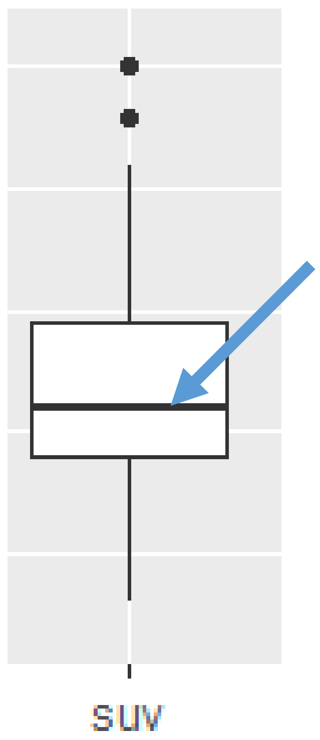
- The box plot summarizes a data set with five statistics.



Box Plots

1. The **median** is the observation in the middle of all observations

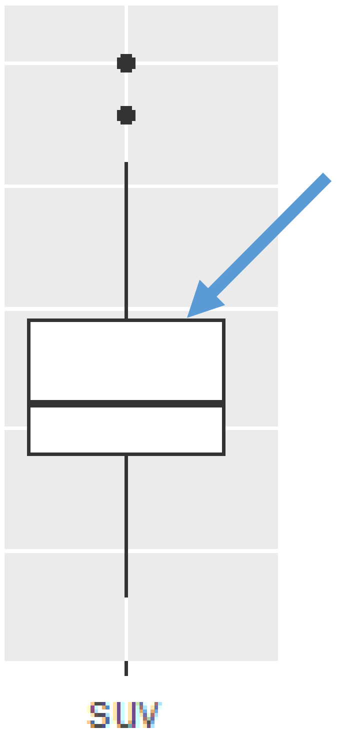
- If there are an even number of observations, the average of the two middle observations is used.
- 50% of data fall above the median; the other 50% falls below it



Box Plots

2. The **third quartile** (Q_3 or "75th percentile") indicates where 75% of values in the data set fall under

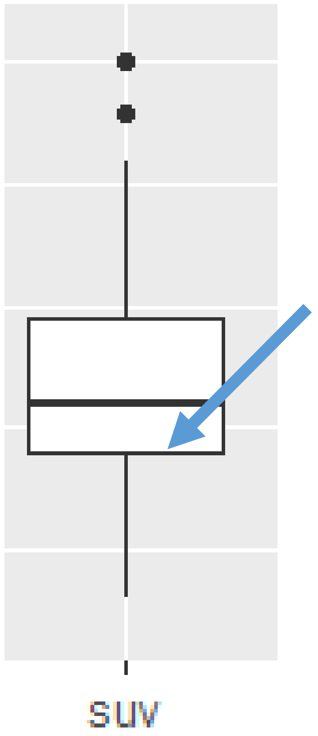
- 75% of observations fall below that line



Box Plots

3. The **first quartile** (Q_1 or "25th percentile") indicates where 25% of values in the data set fall under

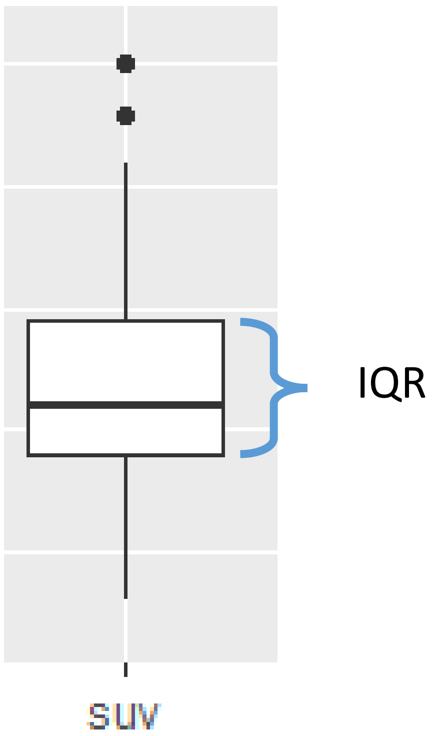
- 25% of observations fall below that line



Box Plots

- Together, they mark the boundaries of the **interquartile range** or **IQR**.
  - 75% of observations fall below to top line
  - 25% of observations fall below the bottom line
  - Thus, 50% of all observations will fall between them (in the box)

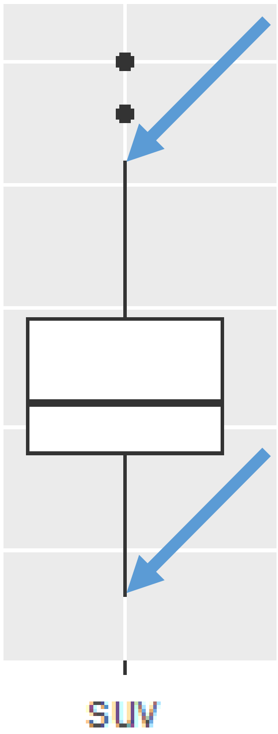
$$IQR = Q_3 - Q_1$$



Box Plots

4 and 5. The **whiskers** try to capture the data outside of the IQR

- At most, they can extend $1.5 \times IQR$
- Max upper whisker = $Q_3 + 1.5 \times IQR$
- Max lower whisker = $Q_1 - 1.5 \times IQR$



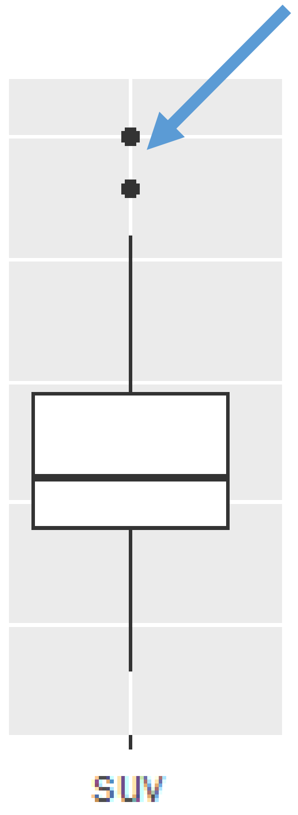
Box Plots

- The upper whisker extends as far as it can go

  $$(Q_3 + 1.5 \times IQR)$$

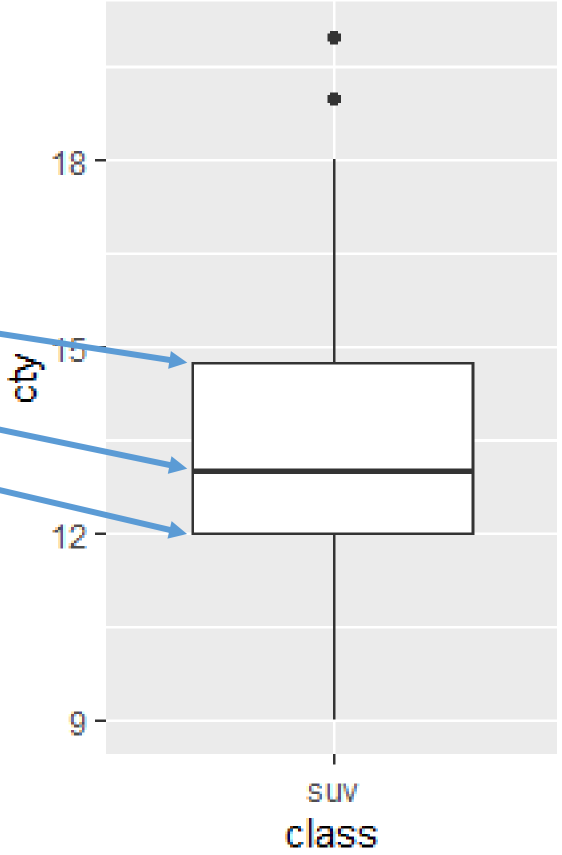
- There are data points still outside of its reach.
  - These two data points (distant from the rest of the data) are called **outliers**

- Looking for outliers is useful for:
  - Identifying strong skew
  - Identifying data collection or data entry errors
  - Offering insight into interesting properties of the data



Box Plots

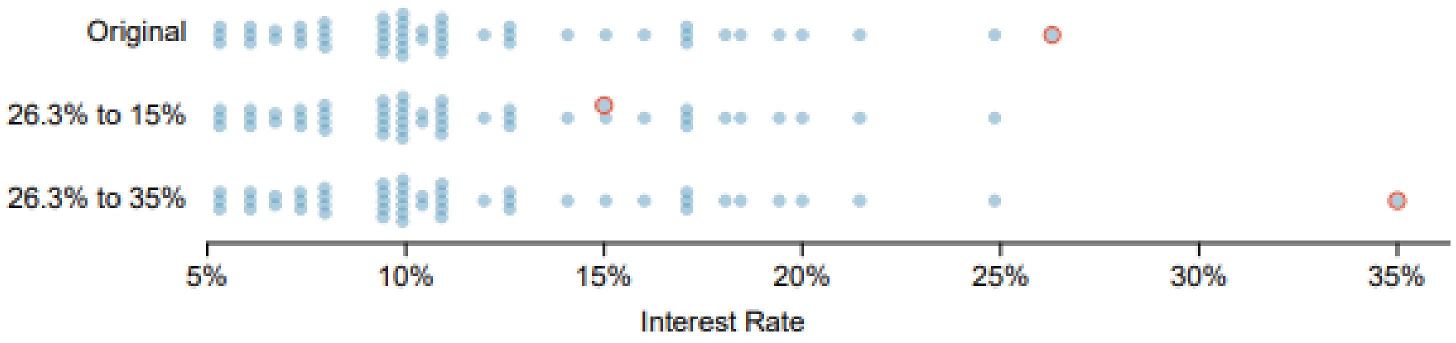
```
> summary(mpg$cty)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   9.00   14.00   17.00   16.86   19.00   35.00 
> summary(subset(mpg, class=="suv")$cty)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   9.00   12.00   13.00   13.50   14.75   20.00 
```



- R's [summary function]

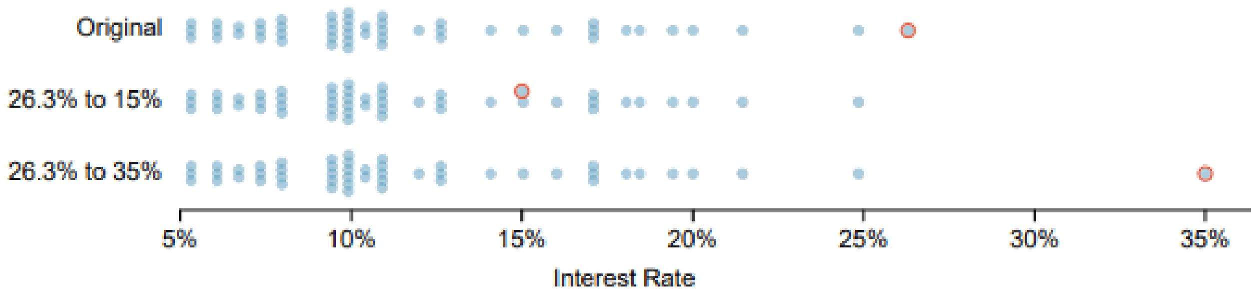
Robust Statistics

- Median and IQR are **robust statistics** in that extreme outliers have little effect on their values.
- This example shows an observation being changed three times.
  - What effect will this have on the sample statistics?



Robust Statistics

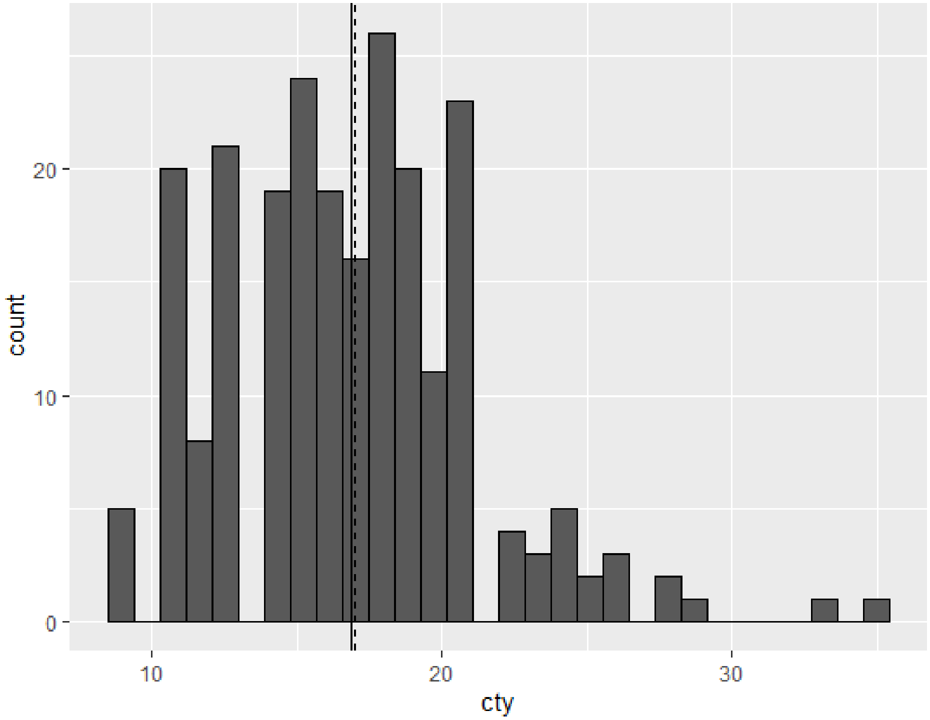
- No impact on median and IQR measurements
- *Did* impact the mean and standard deviation measurements



	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original `interest_rate` data	9.93%	5.76%	11.57%	5.05%
move 26.3% → 15%	9.93%	5.76%	11.34%	4.61%
move 26.3% → 35%	9.93%	5.76%	11.74%	5.68%

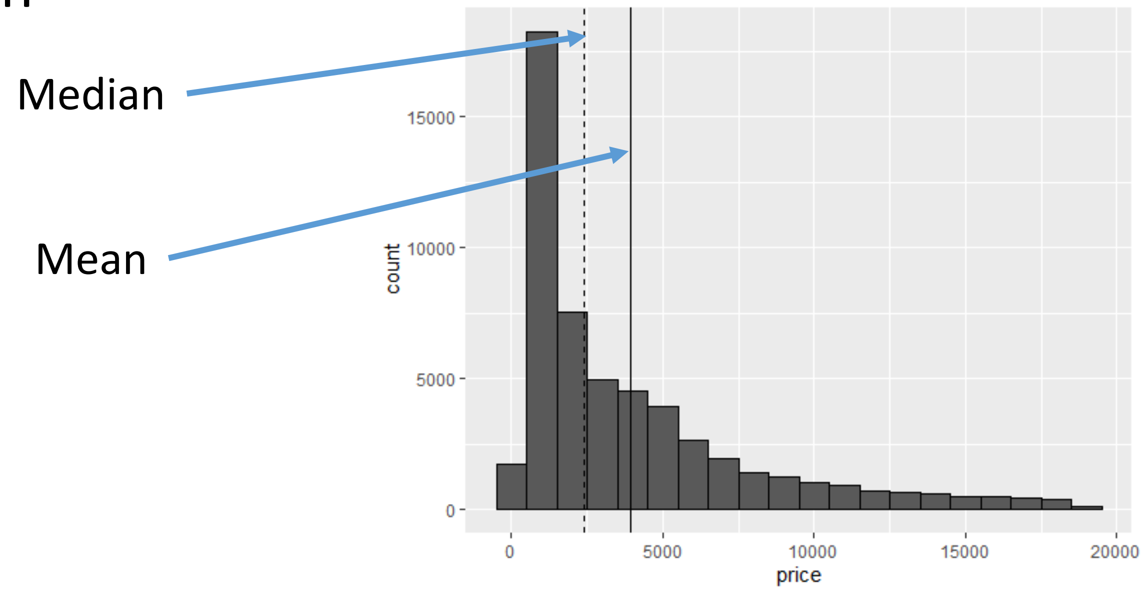
Robust Statistics

- In symmetric distributions, the mean is typically used to describe the center
    - mean ~ median



Robust Statistics

- In skewed distributions or where extreme outliers are present, the median is typically used to describe the center
  - Right skewed: mean > median
  - Left skewed: mean < median



Robust Statistics

- For symmetric distributions, use $\bar{x}$ and s to describe the center and spread
- For skewed distributions: use median and IQR to describe the center and spread

Contingency Tables

- A **contingency table** is a table that summarizes data for two categorical variables.

- Contingency tables show the frequency of combinations between the two variables.
  - For example, the table below shows there were 3496 observations in the data set that had an application type of “individual” and homeownership type of “rent”
  - Another example, there were 183 observations in the data set that had an application type of “joint” and homeownership type of “own”

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Contingency Tables

```
> library(readr)
> loans <- read_csv("loans.csv")
Parsed with column specification:
cols(
  .default = col_double(),
  emp_title = col_character(),
  state = col_character(),
  homeownership = col_character(),
  verified_income = col_character(),
  verification_income_joint = col_character(),
  loan_purpose = col_character(),
  application_type = col_character(),
  grade = col_character(),
  sub_grade = col_character(),
  issue_month = col_character(),
  loan_status = col_character(),
  initial_listing_status = col_character(),
  disbursement_method = col_character()
)
See spec(...) for full column specifications.
> table(loans$application_type, loans$homeownership)

             MORTGAGE  OWN RENT
  individual     3839 1170 3496
  joint           950  183  362
> addmargins(table(loans$application_type, loans$homeownership))

             MORTGAGE   OWN  RENT   Sum
  individual     3839  1170  3496  8505
  joint           950   183   362  1495
  Sum            4789  1353  3858 10000
> |
```

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

[table function]

[addmargins function]

Contingency Tables

- A contingency table can also be used to summarize one categorical variable.

```
> table(loans$homeownership)

MORTGAGE      OWN     RENT
    4789     1353     3858
> addmargins(table(loans$homeownership))

MORTGAGE      OWN     RENT      Sum
    4789     1353     3858    10000
```

homeownership	Count
rent	3858
mortgage	4789
own	1353
Total	10000

Contingency Tables

- Sometimes, it is useful for contingency tables display proportions instead of frequencies.

```
> t<-table(loans$application_type, loans$homeownership)
> t

             MORTGAGE  OWN RENT
  individual     3839 1170 3496
  joint           950  183  362
> prop.table(t)

             MORTGAGE    OWN   RENT
  individual   0.3839 0.1170 0.3496
  joint        0.0950 0.0183 0.0362
```

Frequencies

Proportions

prop.table function

Contingency Tables

- Row Proportions:
  - 63.5% of observations with an application type of “joint” have a homeownership type of “mortgage”.
  - 12.2% of observations with an application type of “joint” have a homeownership type of “own”.
  - 24.2% of observations with an application type of “joint” have a homeownership type of “rent”.

```
> t<-table(loans$application_type, loans$homeownership)
> t

             MORTGAGE  OWN RENT
  individual     3839 1170 3496
  joint           950  183  362
> prop.table(t, margin=1)

              MORTGAGE       OWN      RENT
  individual 0.4513815 0.1375661 0.4110523
  joint      0.6354515 0.1224080 0.2421405
```

Contingency Tables

- Column Proportions:
  - 86.5% of observations with a homeownership type of “own” have an application type of “individual”.
  - 13.5% of observations with a homeownership type of “own” have an application type of “joint”.

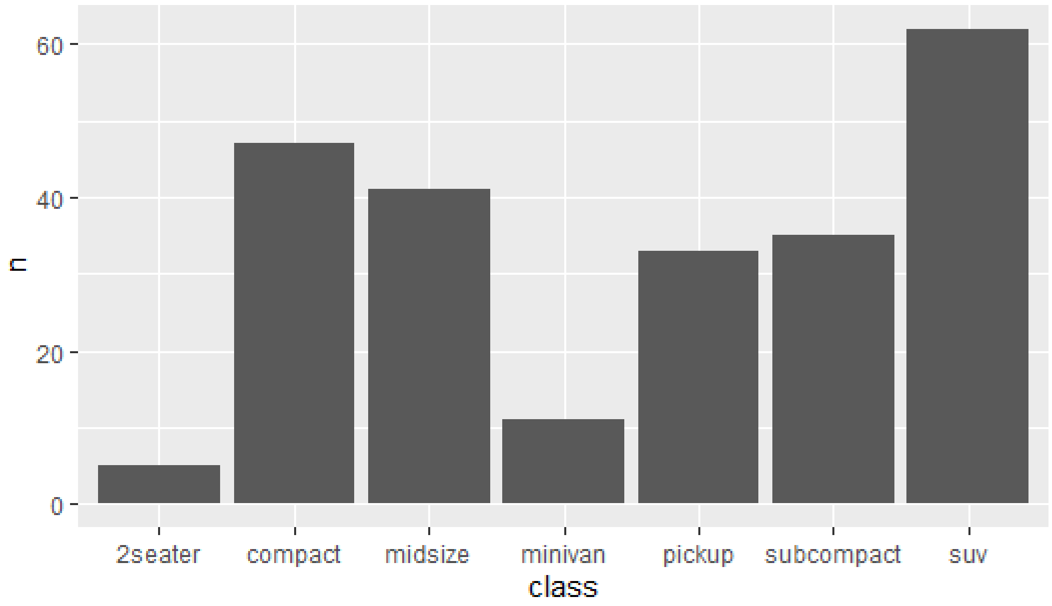
```
> t<-table(loans$application_type, loans$homeownership)
> t

             MORTGAGE  OWN RENT
  individual     3839 1170 3496
  joint           950  183  362
> prop.table(t, margin=2)

              MORTGAGE       OWN      RENT
  individual 0.8016287 0.8647450 0.9061690
  joint      0.1983713 0.1352550 0.0938310
```

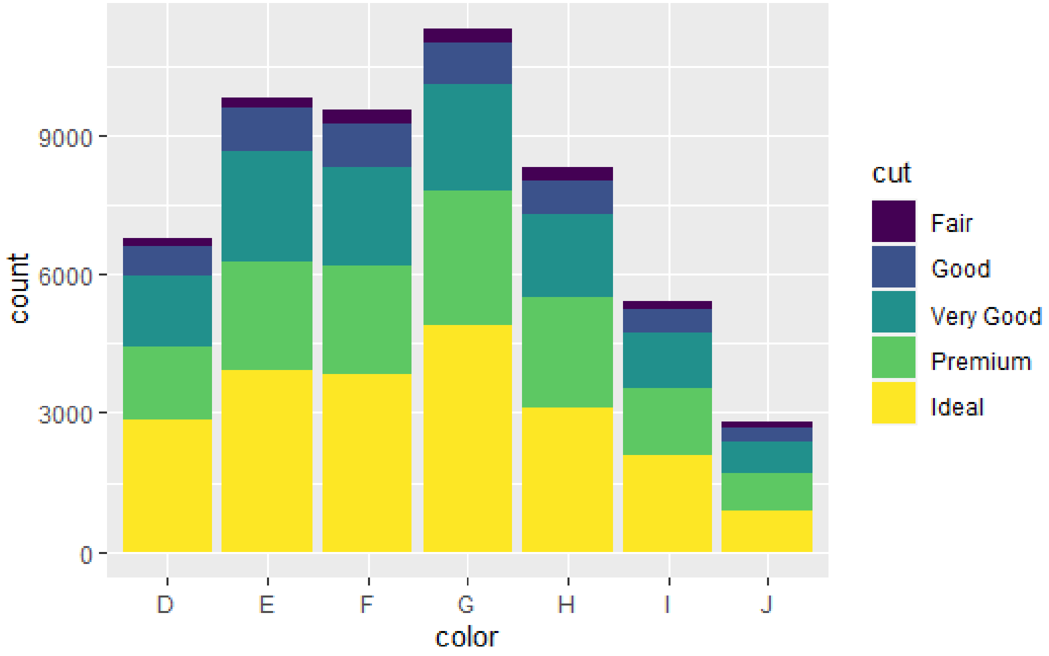
Bar plots

- Bar plots are used to visualize the quantities of categorical variables.



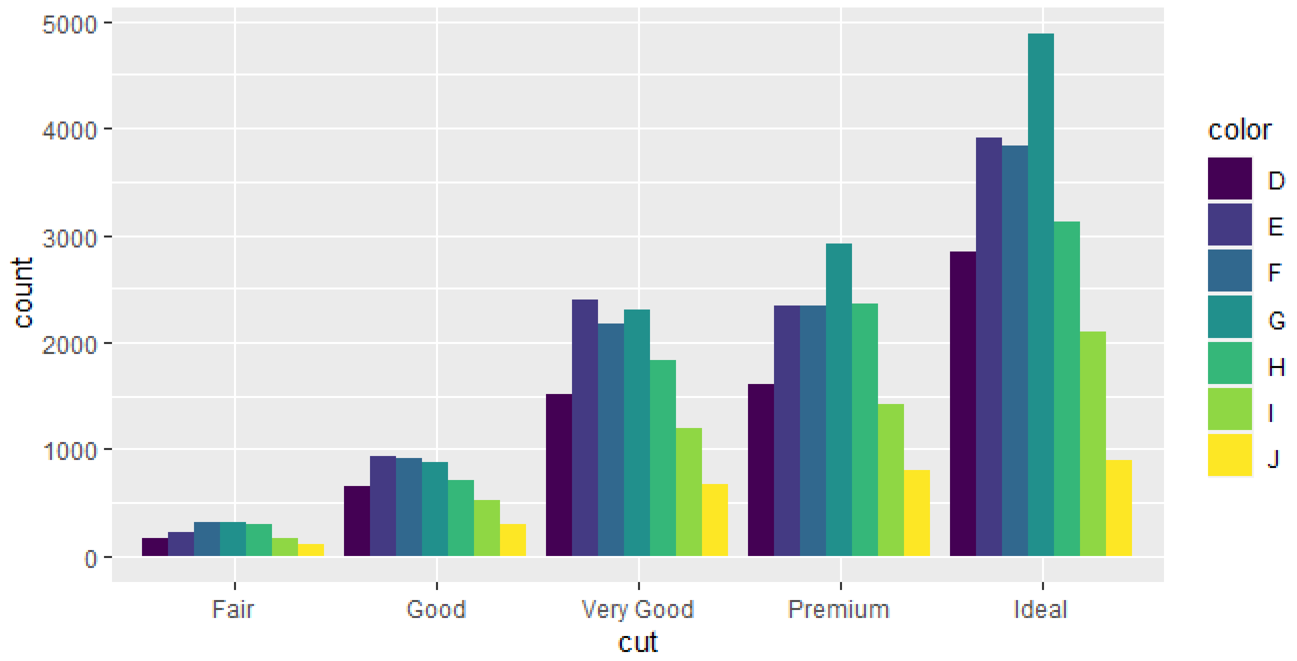
Bar plots

- A *stacked bar plot* allows for plotting two categorial variables at once
  - Still emphasizes the totals of the x-axis variables



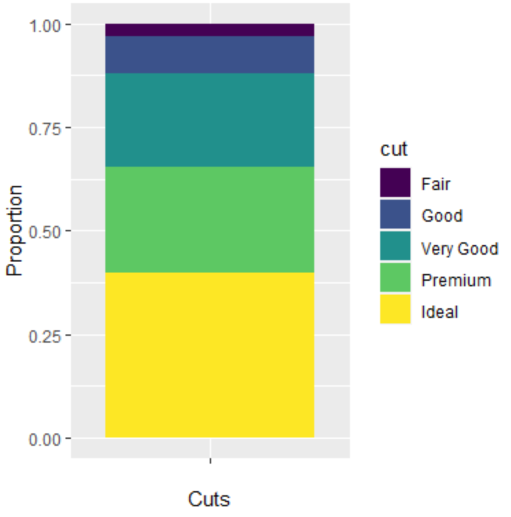
Bar plots

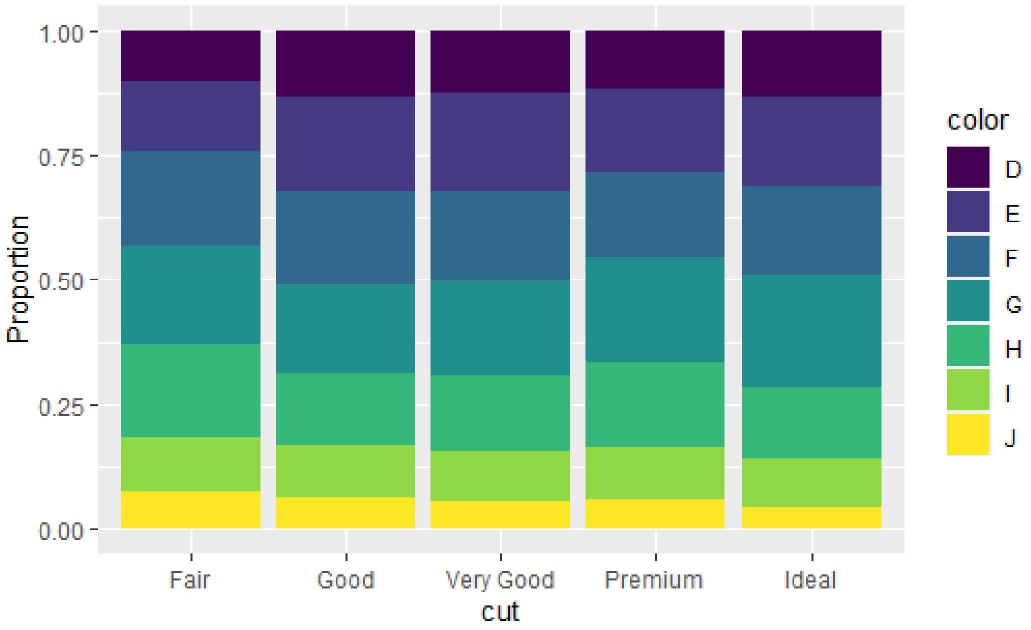
- A *grouped bar plot* also allows for plotting two categorial variables at once.



Bar plots

- Stacked bar plots can also be used to visualize proportions





Bar plots

- Seen previously, grouped bar plots also show proportionality
  - In this case, the right plot is better for comparing the proportions of colors in each cut type