Week 2: Exploring Data

EMSE 4197 | John Paul Helveston | January 22, 2020

Thanks for the hero gifs :)

24,901

24,901 miles

Earth's circumference at the equator: 24,901 miles

Types of Data

Categorical

Subdivide things into useful groups

- What type?
- Which category?

Variable type:

- Nominal
- Ordinal

Numerical

Measure things with numbers

- How many?
- How much?

Scale type:

- Interval
- Ratio

Categorical (discrete) variables

Nominal

- Order doesn't matter
- Differ in "name" (nominal) only

Example: `country` in *TB cases*

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

Ordinal

- Order matters
- Distance between units not equal

Example: `Placement` *2017 Boston marathon*

```
## # A tibble: 6 x 3
##   Placement `Official Time` Name
##       <dbl> <drtn>          <chr>
## 1         1 02:09           Kirui, Geoffrey
## 2         2 02:09           Rupp, Galen
## 3         3 02:10           Osako, Suguru
## 4         4 02:12           Biwott, Shadrac
## 5         5 02:12           Chebet, Wilson
## 6         6 02:12           Abdirahman, Abd
```

Numerical data

Interval

- Numerical scale with arbitrary starting point
- No "0" point
- Can't say "x" is double "y"

Example: `day`, `time`, & `temp` in *Beaver temperature*

```
##   day time  temp activ
## 1 346 2140 36.91     0
## 2 346 1810 37.00     0
## 3 346 1730 37.07     1
## 4 346  940 36.71     0
## 5 346 2110 36.87     0
## 6 346 1050 36.89     0
```

Ratio

- Has a "0" point
- Can be described as percentages
- Can say "x" is double "y"

Example: `height` & `speed` in wildlife impacts

```
## # A tibble: 6 x 3
##   incident_date       height speed
##   <dttm>               <dbl> <dbl>
## 1 2018-12-31 00:00:00    700   200
## 2 2018-12-27 00:00:00    600   145
## 3 2018-12-23 00:00:00      0   130
## 4 2018-12-22 00:00:00    500   160
## 5 2018-12-21 00:00:00    100   150
## 6 2018-12-18 00:00:00   4500   250
```

Be careful of how variables are encoded

- When numbers are categories
  - "Dummy coding": "Has Graduated" = `1`, "Has not Graduated" = `0`)
  - "North", "South", "East", "West" = `1`, `2`, `3`, `4`
- When ratio data are discrete (i.e. counts)
  - Number of eggs in a carton, heart beats per minute, etc.
  - Continuous variables measured discretely (e.g. age)
- Time:
  - As *ordinal* categories: "Jan.", "Feb.", "Mar.", etc.
  - As *interval* scale: "Jan.", "Feb.", "Mar.", etc.
  - As *ratio* scale: "Day 1", "Day 2", "Day 3", etc.

Practice with data types

1) Read in the following data sets:

- `milk_production.csv`
- `lotr_words.csv`

2) For each variable in each data set, note the data type:

Categorical	Numerical
Nominal	Interval
Ordinal	Ratio

3) Share your results with your neighbor

Summary measures:

1. Centrality
2. Variability

Centrality ("Average")

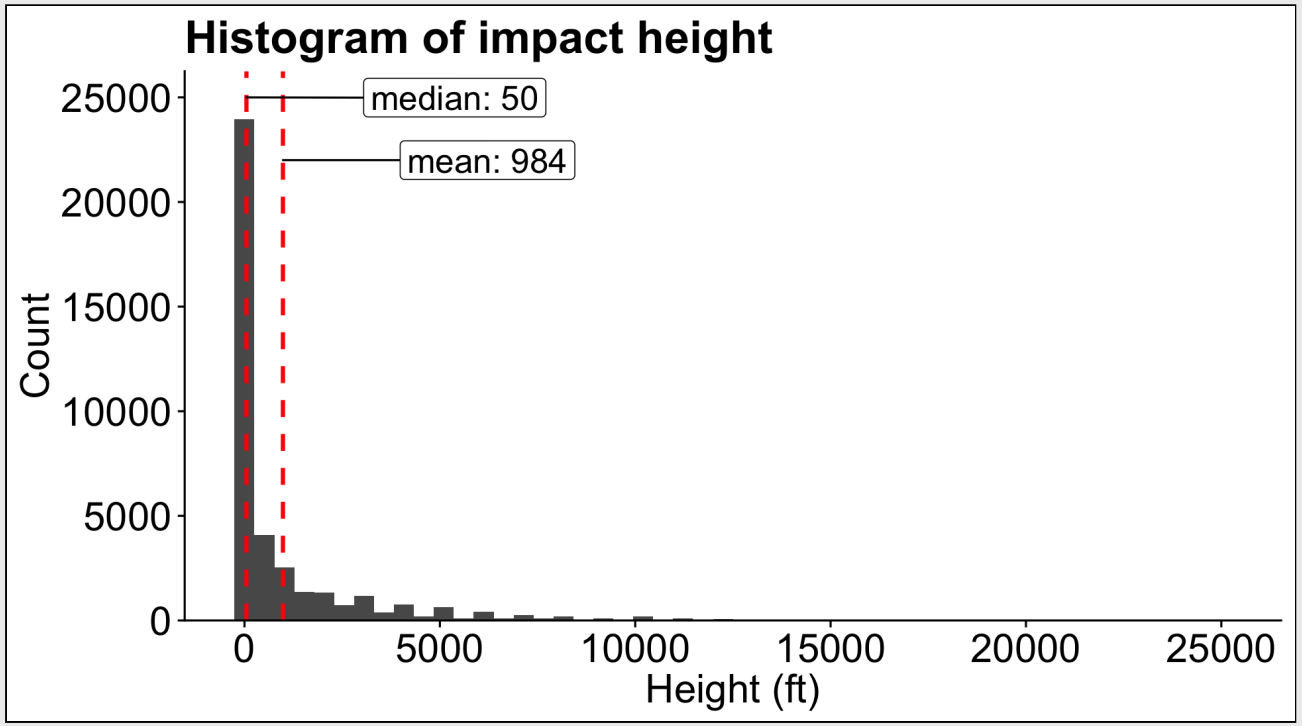
A single number representing the *middle* of a set of numbers

Mean: $\frac{\text{Sum of values}}{\#\text{ of values}}$

Median: Middle value (50% of data above & below)

Mode: Most frequent value (rarely use)

"Mean" isn't always the best choice




```
wildlife_impacts %>%
    filter(! is.na(height)) %>%
    summarise(
      mean = mean(height),
      median = median(height))
```

```
## # A tibble: 1 x 2
##    mean median
##   <dbl>  <dbl>
## 1  984.     50
```

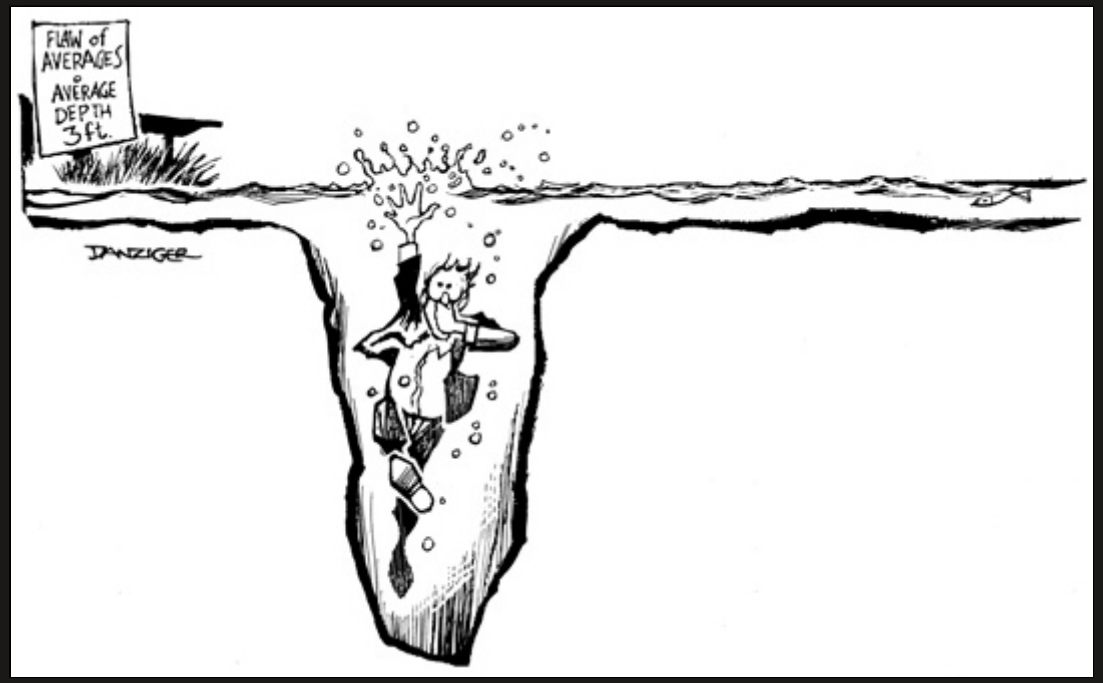
Percent of data below mean:

```
## [1] "73.9%"
```

Beware the "flaw of averages"

What happened to the statistician that crossed a river with an average depth of 3 feet?

...he drowned



Variability ("Spread")

Range: max - min

Standard deviation: distribution of values relative to the mean

Interquartile range (IQR): $Q_3 - Q_1$ (middle 50% of data)

Example: Days to ship

Complaints are coming in about orders shipped from warehouse B, so you collect some data:

```
##    order warehouseA warehouseB
## 1      1          3          1
## 2      2          3          1
## 3      3          3          1
## 4      4          4          3
## 5      5          4          3
## 6      6          4          4
## 7      7          5          5
## 8      8          5          5
## 9      9          5          5
## 10    10          5          6
## 11    11          5          7
## 12    12          5         10
```

Here, **averages** are misleading:

```
daysToShip %>%
    gather(warehouse, days, warehouseA:warehouseB) %>%
    group_by(warehouse) %>%
    summarise(
        mean   = mean(days),
        median = median(days))
```

```
## # A tibble: 2 x 3
##   warehouse   mean median
##   <chr>      <dbl>  <dbl>
## 1 warehouseA  4.25    4.5
## 2 warehouseB  4.25    4.5
```

Example: Days to ship

Complaints are coming in about orders shipped from warehouse B, so you collect some data:

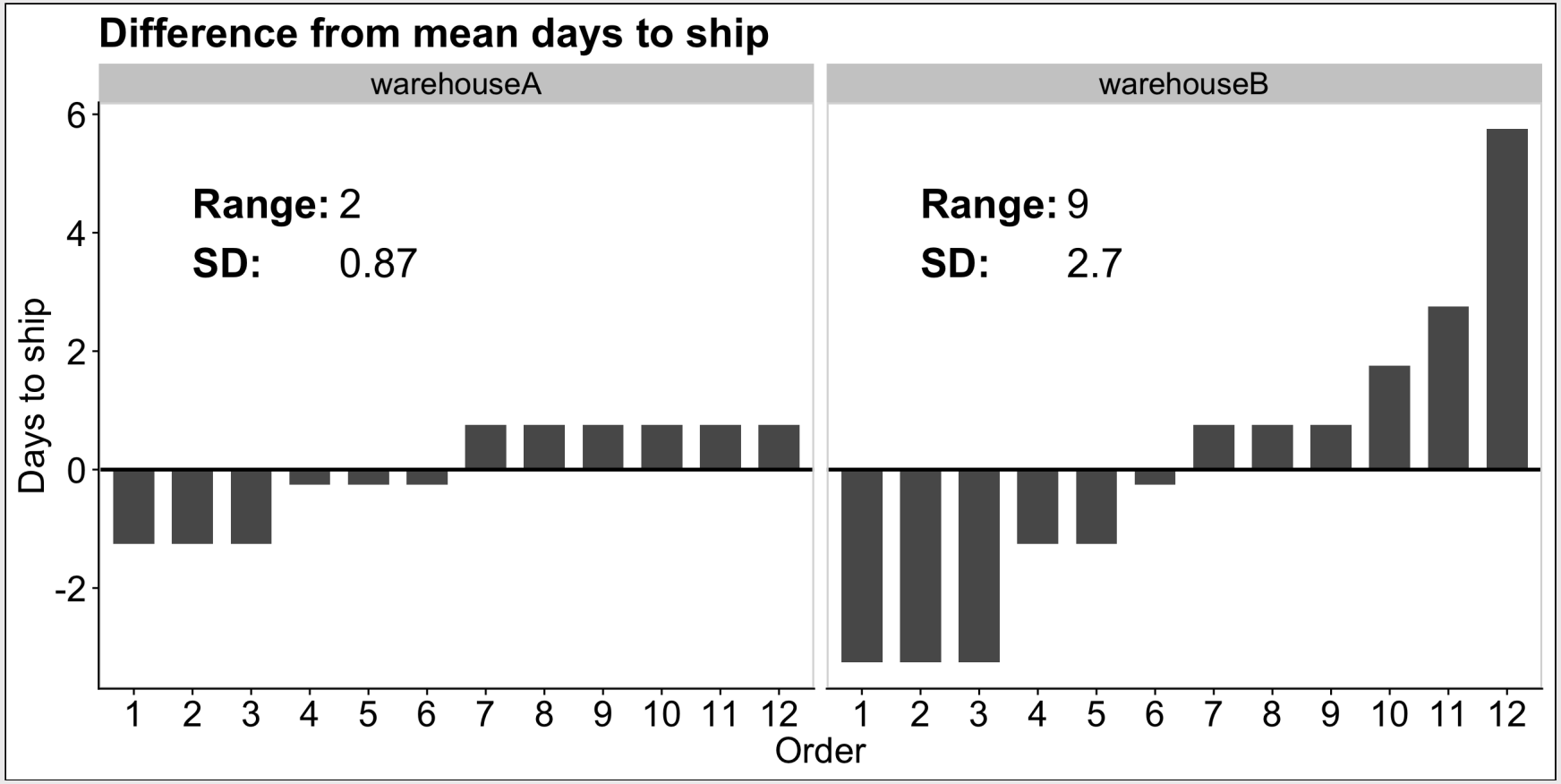
```
##    order warehouseA warehouseB
## 1      1          3          1
## 2      2          3          1
## 3      3          3          1
## 4      4          4          3
## 5      5          4          3
## 6      6          4          4
## 7      7          5          5
## 8      8          5          5
## 9      9          5          5
## 10    10          5          6
## 11    11          5          7
## 12    12          5         10
```

Variability reveals difference in days to ship:

```
daysToShip %>%
    gather(warehouse, days, warehouseA:warehouseB) %>%
    group_by(warehouse) %>%
    summarise(
        mean   = mean(days),
        median = median(days),
        range = max(days) - min(days),
        sd    = sd(days))
```

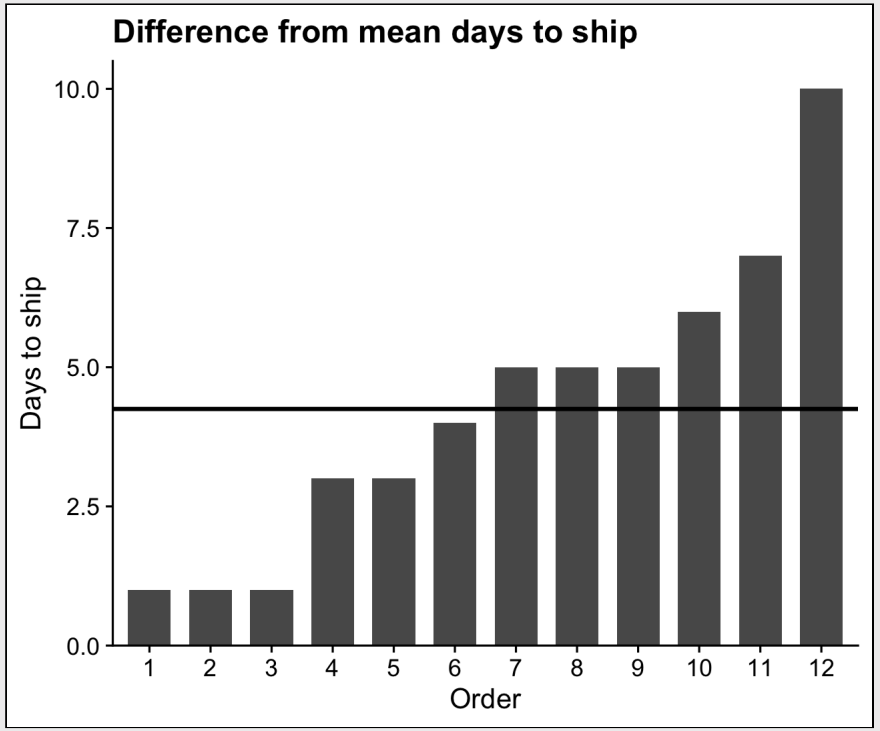
```
## # A tibble: 2 x 5
##   warehouse   mean median range    sd
##   <chr>      <dbl>  <dbl> <dbl> <dbl>
## 1 warehouseA  4.25    4.5     2 0.866
## 2 warehouseB  4.25    4.5     9 2.70
```

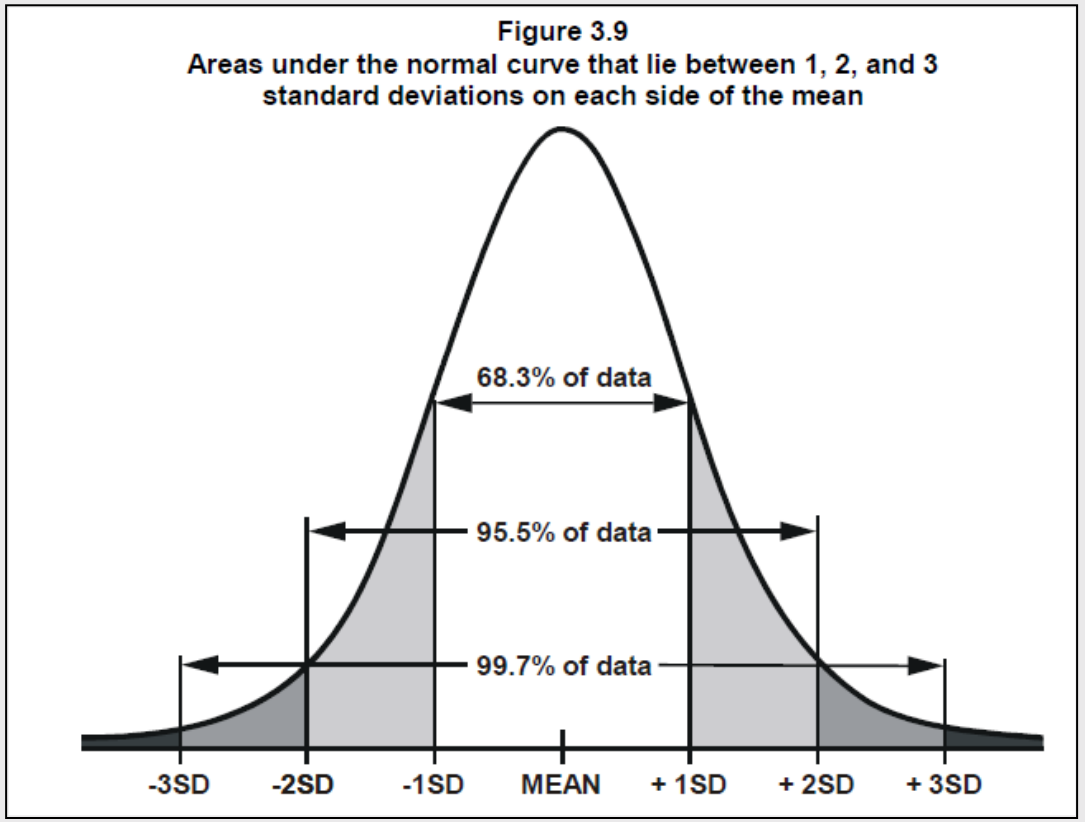
Example: Days to ship



Interpreting the standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$





Practice with summary measurements

1) Read in the following data sets:

- `milk_production.csv`
- `lotr_words.csv`

2) For each variable in each data set, if possible, summarize its

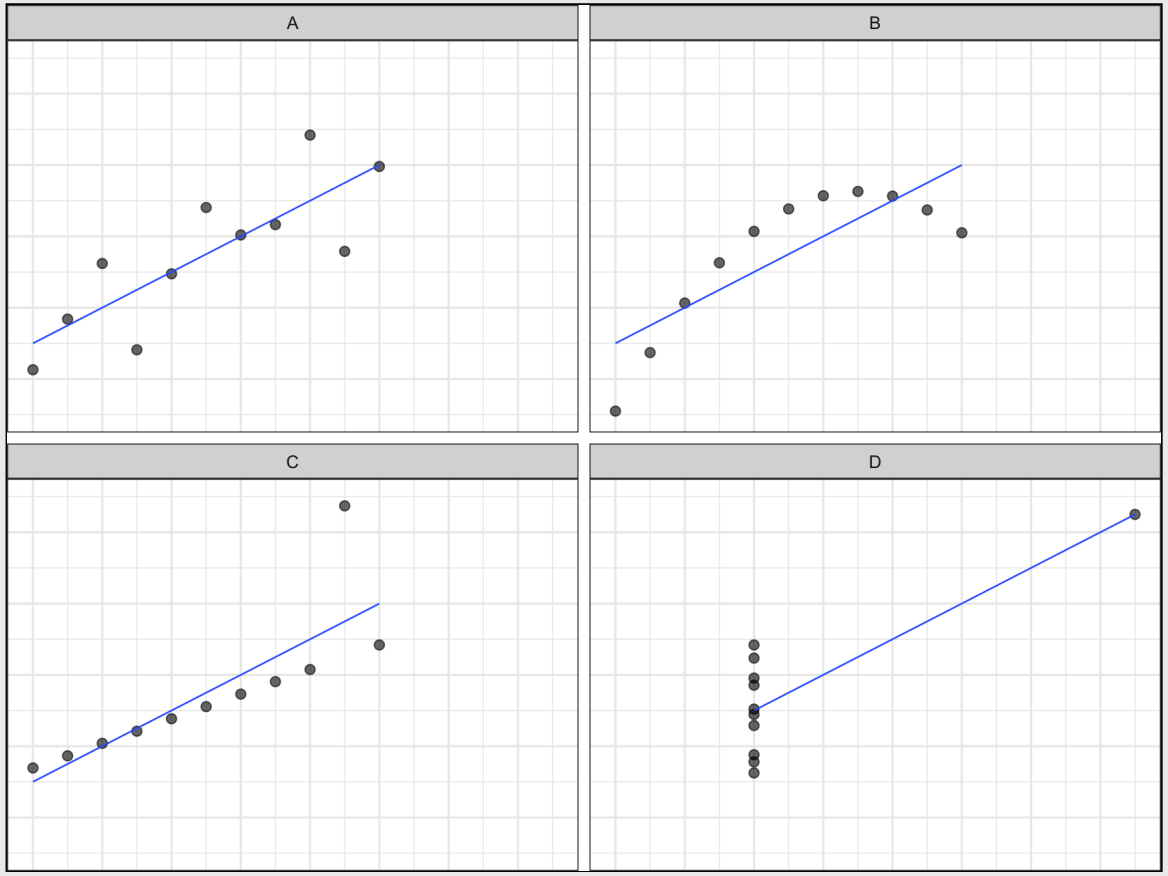
- *Centrality*
- *Variability*

3) Share your results with your neighbor

"Visualizing data helps us think"[1]

	A		B		C		D	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99	82.51	99	82.51	99	82.5	99	82.51
Mean:	9	7.5	9	7.5	9	7.5	9	7.5
St. Dev:	3.3	2	3.3	2	3.3	2	3.3	2

Anscombe's Quartet



Stephen Few (2009, pg. 6)

The data type determines how to summarize it

Nominal (Categorical)

Measures:

- Frequency counts
- Proportions

Charts:

- Bars

Ordinal (Categorical)

Measures:

- Frequency counts
- Proportions
- Centrality: Median, Mode
- Variability: IQR

Charts:

- Bars

Numerical (Continuous)

Measures:

- Centrality: Mean, median
- Variability: Range, standard deviation, IQR

Charts:

- Histogram
- Boxplot

Summarizing **Nominal** data

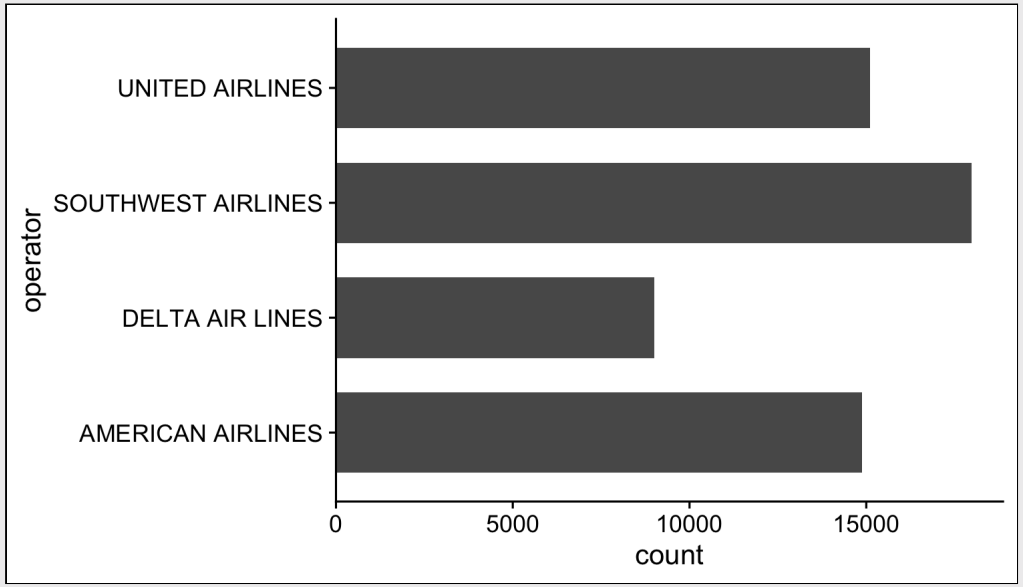
Summarize with counts / percentages

```
wildlife_impacts %>%
    count(operator) %>%
    mutate(
        p = n / sum(n),
        percent = round(100*p, 2))
```

```
## # A tibble: 4 x 4
##   operator              n     p percent
##   <chr>             <int> <dbl>   <dbl>
## 1 AMERICAN AIRLINES  14887 0.261    26.1
## 2 DELTA AIR LINES     9005 0.158    15.8
## 3 SOUTHWEST AIRLINES 17970 0.315    31.5
## 4 UNITED AIRLINES    15116 0.265    26.5
```

Visualize with bars

```
wildlife_impacts %>%
    ggplot() +
    geom_bar(aes(x = operator), width = 0.7) +
    coord_flip() +
    theme_half_open()
```



Summarizing **Ordinal** data

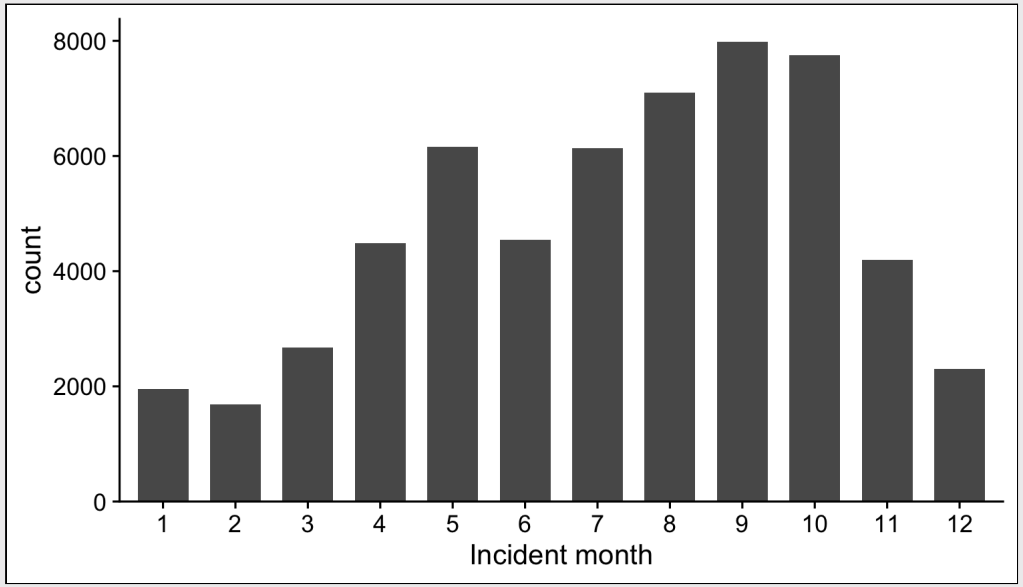
Summarize with counts / percentages

```
wildlife_impacts %>%
    count(incident_month) %>%
    mutate(
        p = n / sum(n),
        percent = round(100*p, 2))
```

```
## # A tibble: 12 x 4
##    incident_month     n      p percent
##             <dbl> <int>  <dbl>   <dbl>
##  1              1  1951 0.0342    3.42
##  2              2  1692 0.0297    2.97
##  3              3  2678 0.0470    4.7
##  4              4  4490 0.0788    7.88
##  5              5  6161 0.108    10.8
##  6              6  4541 0.0797    7.97
##  7              7  6133 0.108    10.8
##  8              8  7104 0.125    12.5
##  9              9  7980 0.140    14.0
## 10             10  7754 0.136    13.6
## 11             11  4191 0.0736    7.36
## 12             12  2303 0.0404    4.04
```

Visualize with bars

```
wildlife_impacts %>%
    ggplot() +
    geom_bar(aes(x = as.factor(incident_month)),
             width = 0.7) +
    theme_half_open() +
    labs(x = 'Incident month')
```



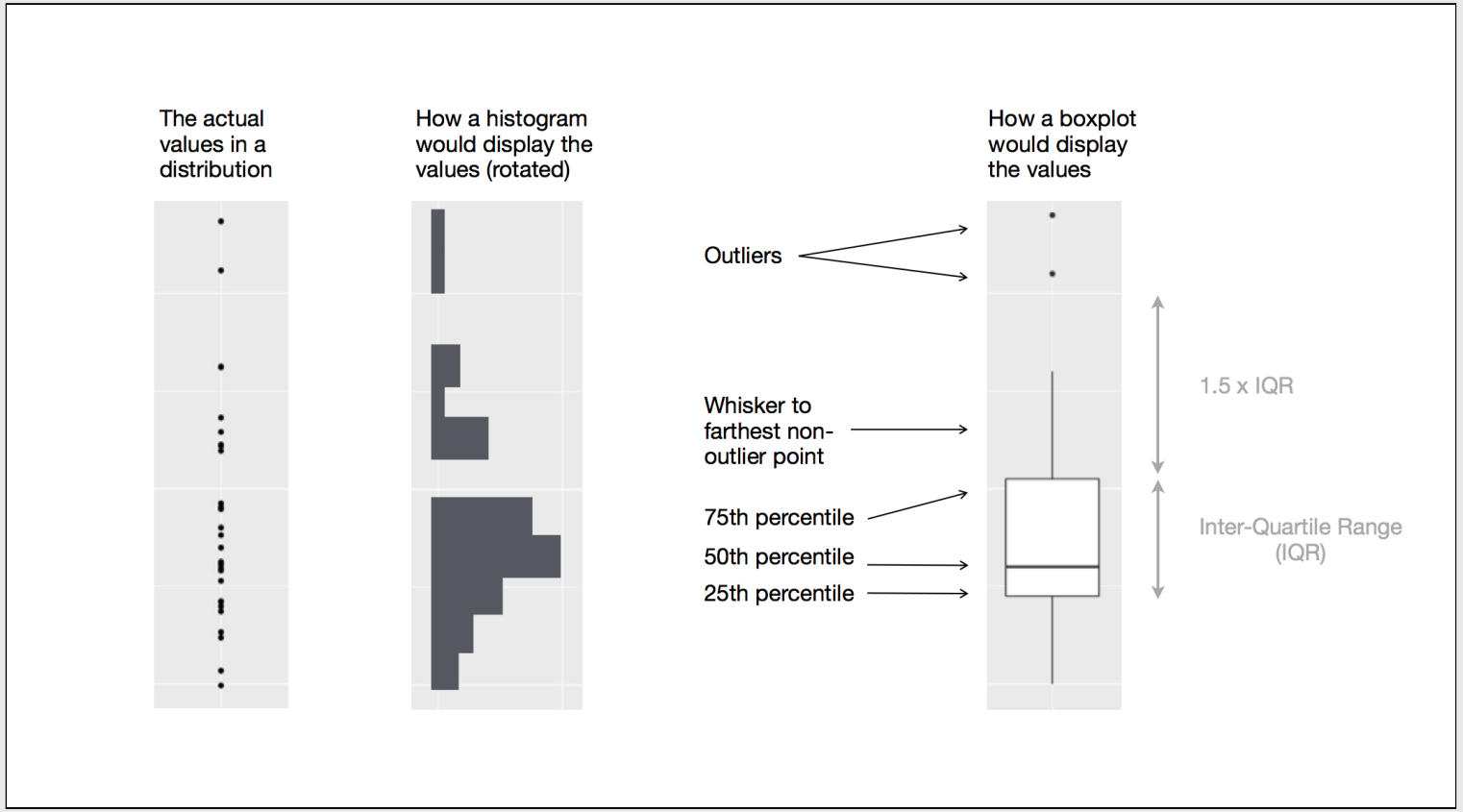
Summarizing **continuous** variables

Histograms:

- Identifying skewness
- Identifying # of modes

Boxplots:

- Identifying outliers
- Comparing distributions across groups



Continuous variables: **histogram**

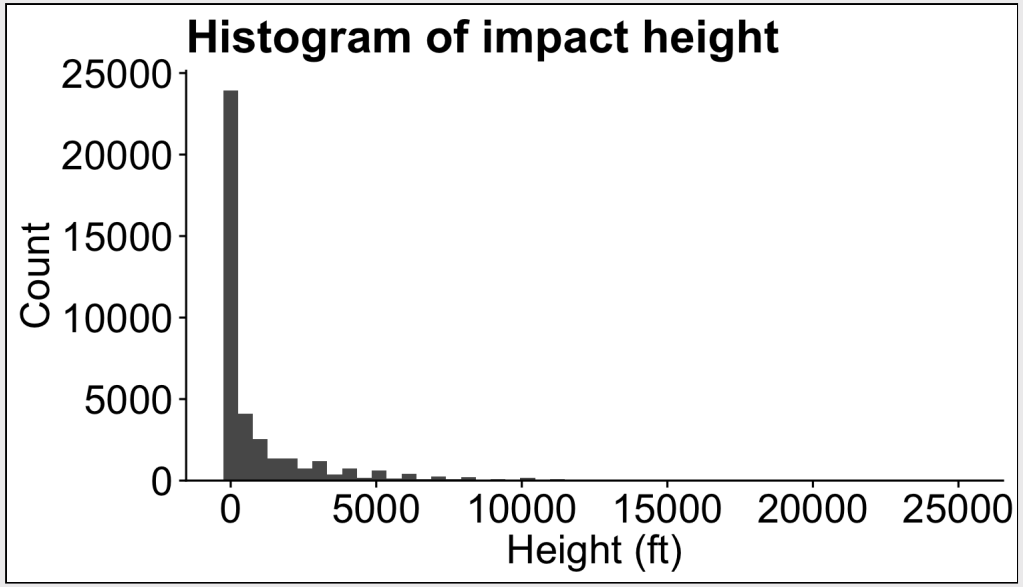
Summarise with mean, median, sd, range, & IQR:

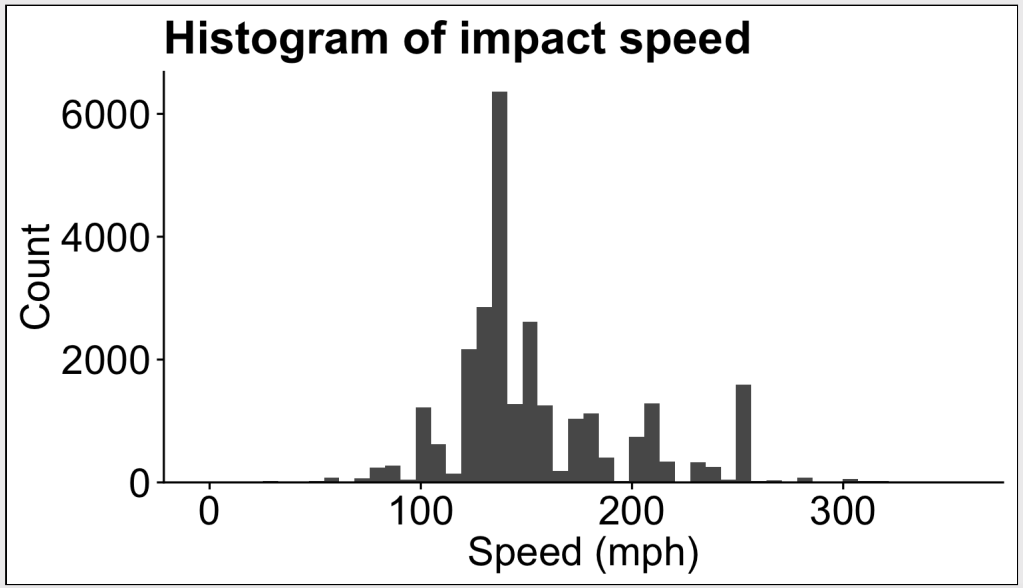
```
## # A tibble: 2 x 6
##   var     mean median     sd range   IQR
##   <chr>  <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1 height  1212    200 2157.  25000  1500
## 2 speed    154    140   42.3   354    40
```

Visualize with **histogram** to:

- Identify skewness
- Identify # of modes

```
wildlife_impacts %>%
    ggplot() +
    geom_histogram(aes(x = height), bins=50)
    theme_half_open()
```





Continuous variables: **boxplot**

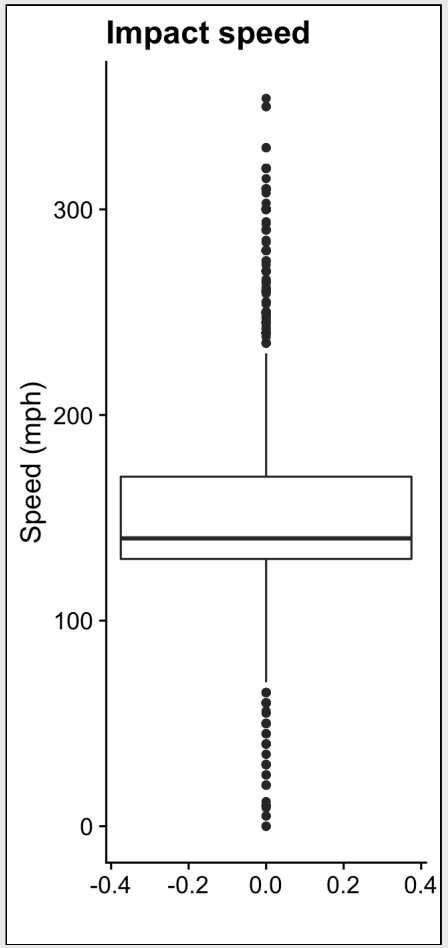
Summarise with mean, median, sd, range, & IQR:

```
## # A tibble: 2 x 6
##   var     mean median     sd range   IQR
##   <chr>  <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1 height  1212    200 2157.  25000  1500
## 2 speed    154    140   42.3   354    40
```

Visualize with **boxplot** to:

- Identify outliers

```
wildlife_impacts %>%
    ggplot() +
    geom_boxplot(aes(y = speed)) +
    theme_half_open()
```



Practice with visual summaries

1) Read in the following data sets:

- `faithful.csv`
- `marathon.csv`

2) Summarize the following variables using an appropriate chart (bar chart, histogram, and / or boxplot):

- faithful: `eruptions`
- faithful: `waiting`
- marathon: `Age`
- marathon: `State`
- marathon: `Country`
- marathon: `` `Official Time` ``

3) Share what you learned about each variable with your neighbor.

5 minute break!

Stand up

Move around

Stretch!

Relationship between two variables

Two categorical variables

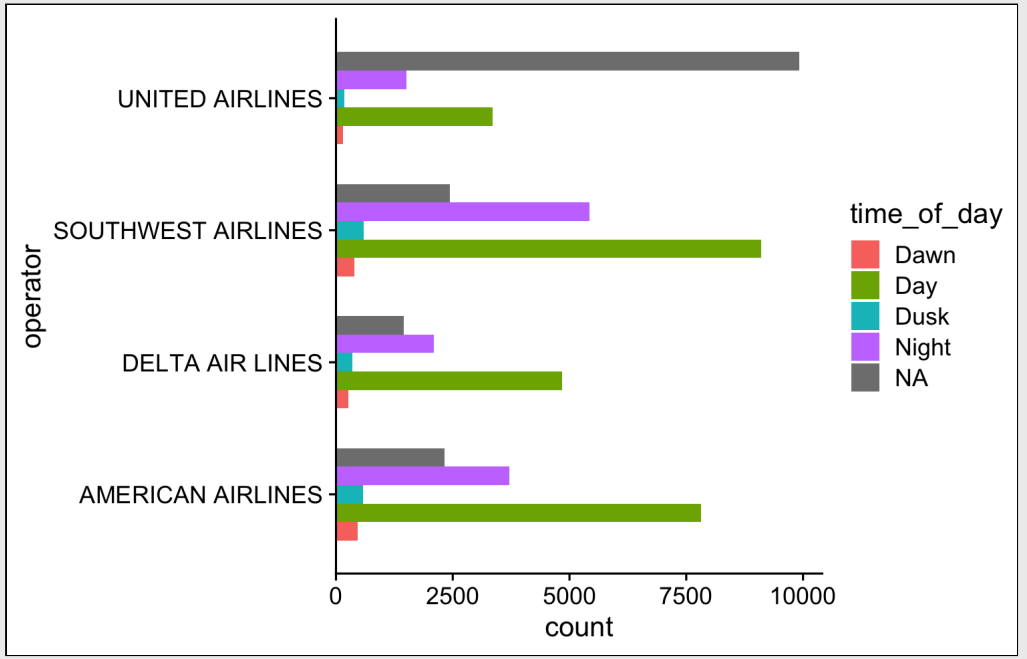
Summarize with a table of counts

```
wildlife_impacts %>%
    count(operator, time_of_day) %>%
    spread(time_of_day, n)
```

```
## # A tibble: 4 x 6
##   operator            Dawn   Day  Dusk Night `<NA>`
##   <chr>              <int> <int> <int> <int>  <int>
## 1 AMERICAN AIRLINES    458  7809   584  3710   2326
## 2 DELTA AIR LINES      267  4846   353  2090   1449
## 3 SOUTHWEST AIRLINES   394  9109   599  5425   2443
## 4 UNITED AIRLINES      151  3359   181  1510   9915
```

Map **color** aesthetic to denote 2nd categorical var

```
wildlife_impacts %>%
    ggplot() +
    geom_bar(aes(x = operator, fill = time_of_day),
             width = 0.7, position = 'dodge') +
    coord_flip()
```



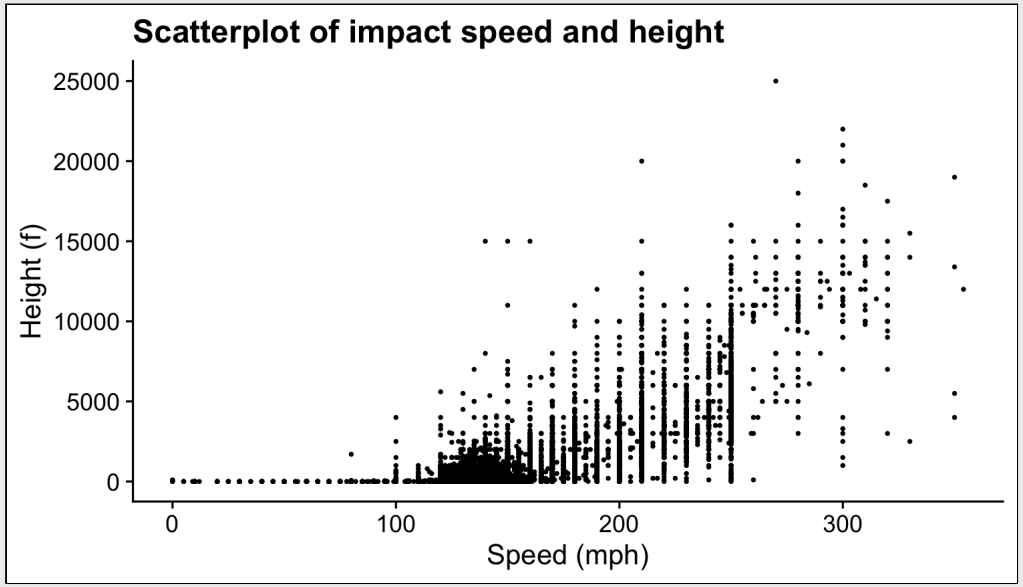
Two continuous variables

Summarise with mean, median, sd, range, & IQR:

```
## # A tibble: 2 x 6
##   var     mean median     sd range   IQR
##   <chr>  <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1 height  1212    200 2157.  25000  1500
## 2 speed    154    140   42.3   354    40
```

Visualize with **scatterplot**

```
wildlife_impacts %>%
    ggplot() +
    geom_point(aes(x = speed, y = height),
               size = 0.5)
```



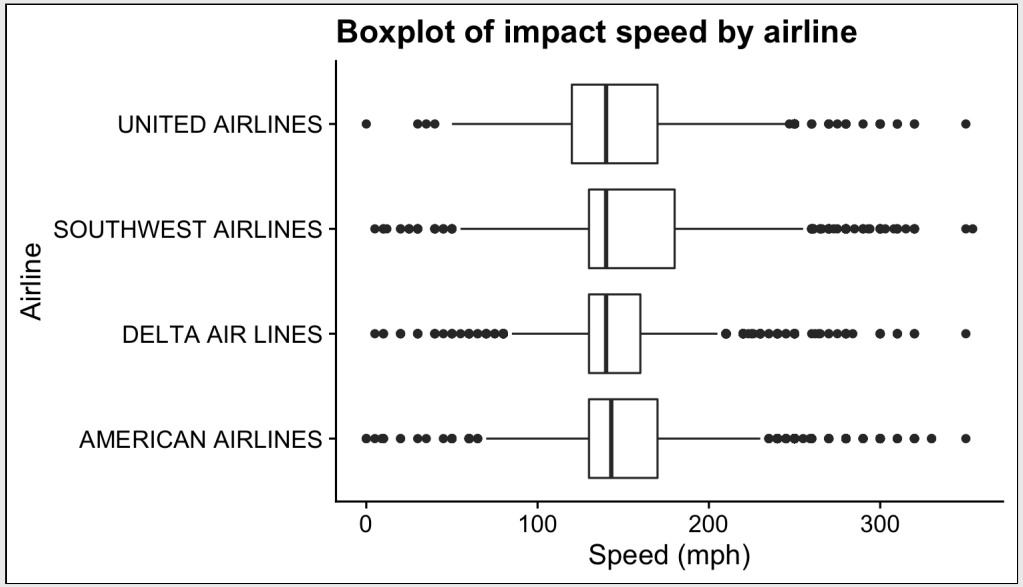
One continuous, one categorical

Summarise with mean, median, sd, range, & IQR:

```
## # A tibble: 4 x 6
##   operator            mean median    sd range   IQR
##   <chr>              <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 AMERICAN AIRLINES    155    143  41.3   350    40
## 2 DELTA AIR LINES      149    140  41.1   345    30
## 3 SOUTHWEST AIRLINES   156    140  42.7   349    50
## 4 UNITED AIRLINES      149    140  44.3   350    50
```

Visualize with **boxplot**

```
wildlife_impacts %>%
    ggplot() +
    geom_boxplot(aes(x=operator, y=speed)) +
    coord_flip()
```



Practice with visualizing relationships

1) Read in the following data sets:

- `marathon.csv`
- `wildlife_impacts.csv`

2) Visualize the *relationships* between the following variables using an appropriate chart (bar plots, scatterplots, and / or box plots):

- marathon: `Age` & `` `Official Time` ``
- marathon: `` `M/F` `` & `` `Official Time` ``
- wildlife_impacts: `state` & `operator`

3) Share what you learned about each variable with your neighbor.

Outliers

Outliers (continuous data)

Outliers: $Q_1 \pm 1.5IQR$

Extreme values: $Q_1 \pm 3.0IQR$

Outliers can have strong effect on the **mean** and **standard deviation**

```
data = c(7,4,6,5,6,5,3,3,8,9)
```

- Mean: 5.6
- Standard Deviation: 2.01
- Median: 5.5
- IQR: 2.5

```
data = c(7,4,6,5,6,5,3,3,9,20)
```

- Mean: 6.8
- Standard Deviation: 4.98
- Median: 5.5
- IQR: 2.5

Robust statistics for continuous data

Centrality: Use *median* rather than *mean*

Variability: Use *IQR* rather than *standard deviation*

Doing EDA

EDA is an iterative process that helps you understand your data:

1. Generate questions about your data
2. Search for answers by visualising, transforming, and/or modelling your data
3. Use what you learn to refine your questions and/or generate new questions

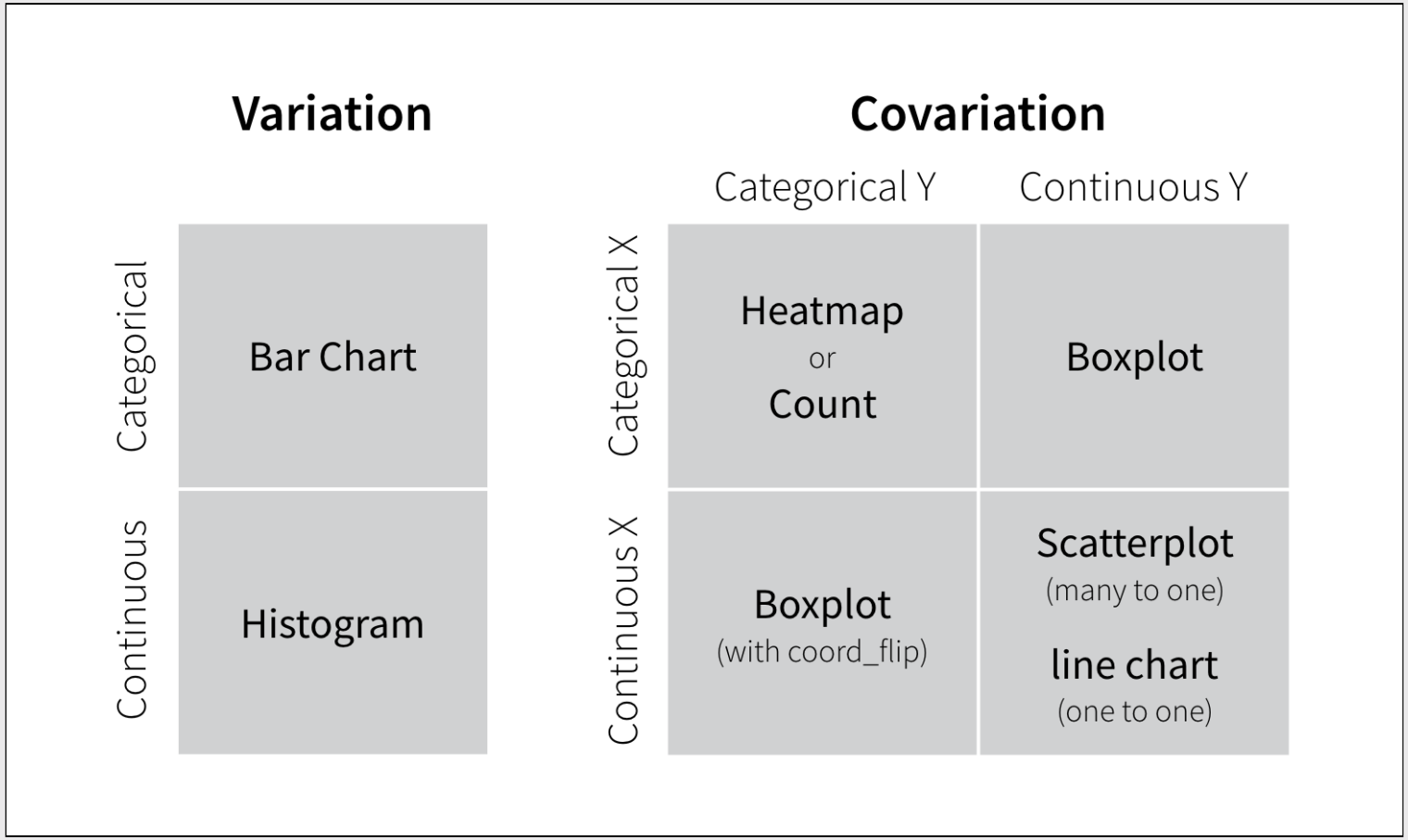
EDA is a tool for *discovery*, not *confirmation*

Visualizing variation

Ask yourself:

- What type of **variation** occurs within my variables?
- What type of **covariation** occurs between my variables?

Check out these guides



"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

— John Tukey

Practice doing EDA: Groups of 3

1) Read in the following data sets:

- `avengers.csv`
- `candy_rankings.csv`
- `college_all_ages.csv`

2) For each variable, note the data type:

Categorical	Numerical
Nominal	Interval
Ordinal	Ratio

3) For each variable, if possible, summarize its

- *Centrality*
- *Variability*

4) Summarize some of the variables using an appropriate chart:

- Bar chart
- Histogram
- Boxplot

5) Visualize a *relationship* between two variables using an appropriate chart:

- Bar chart
- Scatterplot
- Boxplot