Week 3: *Centrality & Variability*

EMSE 4575: Exploratory Data Analysis

John Paul Helveston

January 27, 2021

Thanks for the heros 😄

		
Kareemot	**Ebun**	**Alexa**
		
Kyara	Katie	Helena
		
Matthew	Carolyne	Kazi

Tip of the week:

`theme_set()`

Add "global" settings to all plots

```r
library(knitr)
library(tidyverse)
library(here)
knitr::opts_chunk$set(
    warning = FALSE,
    message = FALSE,
    comment = "#>",
    fig.path = "figs/", # Plot save path
    fig.width = 7.252, # Plot dimensions
    fig.height = 4,
    fig.retina = 3 # Better plot resolution
)

theme_set(theme_bw(base_size = 20)) # Set theme for all ggplots
```

Week 3: *Centrality & Variability*

1. Data Types
2. Measures of Centrality & Variability
3. Visualizing Centrality & Variability

BREAK

4. Relationships Between 2 Variables
5. Exploratory Data Analysis

Week 3: *Centrality & Variability*

1. Data Types

2. Measures of Centrality & Variability

3. Visualizing Centrality & Variability

BREAK

4. Relationships Between 2 Variables

5. Exploratory Data Analysis

24,901

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

Types of Data

Categorical

Subdivide things into *groups*

- What type?
- Which category?

Numerical

Measure things with numbers

- How many?
- How much?

Categorical (discrete) variables

Nominal

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

e.g. `country` in TB case data:

```
#> # A tibble: 6 x 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 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

e.g.: `Placement` 2017 Boston marathon:

```
#> # A tibble: 6 x 3
#>   Placement `Official Time` Name
#>       <dbl> <time>          <chr>
#> 1         1 02:09:37        Kirui, Geo
#> 2         2 02:09:58        Rupp, Gale
#> 3         3 02:10:28        Osako, Sug
#> 4         4 02:12:08        Biwott, Sh
#> 5         5 02:12:35        Chebet, Wi
#> 6         6 02:12:45        Abdirahman
```

Numerical data

Interval

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

e.g.: `temp` in Beaver data

```
#>   day time  temp activ
#> 1 346  840 36.33     0
#> 2 346  850 36.34     0
#> 3 346  900 36.35     0
#> 4 346  910 36.42     0
#> 5 346  920 36.55     0
#> 6 346  930 36.69     0
```

Ratio

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

e.g.: `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
```

Key Questions

Categorical

Does the order matter?

Yes	No
Ordinal	Nominal

Numerical

Is there a "baseline"?

Yes	No
Ratio	Interval

Be careful of how variables are encoded!

When numbers are categories

- "Dummy coding": e.g., `passedTest` = `1` or `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. 1", "Jan. 2", "Jan. 3", etc.
- As *ratio* scale: "Day 1", "Day 2", "Day 3", etc.

Quick practice: What's the data type?

Decide here (link also in #classroom)

```
wildlife_impacts %>%
  filter(!is.na(cost_repairs_infl_adj)) %>%
  select(incident_date, time_of_day, species, cost_repairs_infl_adj)
```

```
#> # A tibble: 615 x 4
#>    incident_date       time_of_day species                  cost_repairs_infl_adj
#>    <dttm>              <chr>       <chr>                                    <dbl>
#>  1 2018-10-25 00:00:00 Day         Unknown bird - large                      1000
#>  2 2018-09-05 00:00:00 <NA>        Unknown bird - medium                      200
#>  3 2018-08-09 00:00:00 Day         Semipalmated sandpiper                   10000
#>  4 2018-06-24 00:00:00 Day         Unknown bird - large                    100000
#>  5 2018-02-18 00:00:00 Day         Rough-legged hawk                        20000
#>  6 2018-01-05 00:00:00 Night       Brant                                   487000
#>  7 2017-10-31 00:00:00 Day         Unknown bird - small                        51
#>  8 2017-10-12 00:00:00 <NA>        Swainson's thrush                         5120
#>  9 2017-09-17 00:00:00 Day         Cattle egret                            531763
#> 10 2017-09-16 00:00:00 <NA>        Unknown bird - medium                      102
#> # … with 605 more rows
```

Week 3: *Centrality & Variability*

1. Data Types
2. Measures of Centrality & Variability
3. Visualizing Centrality & Variability

BREAK

4. Relationships Between 2 Variables
5. Exploratory Data Analysis

Summary Measures:

This week: Centrality & Variability

Next week: Correlation

Centrality (a.k.a. The "Average" Value)

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 (usually for categorical data)

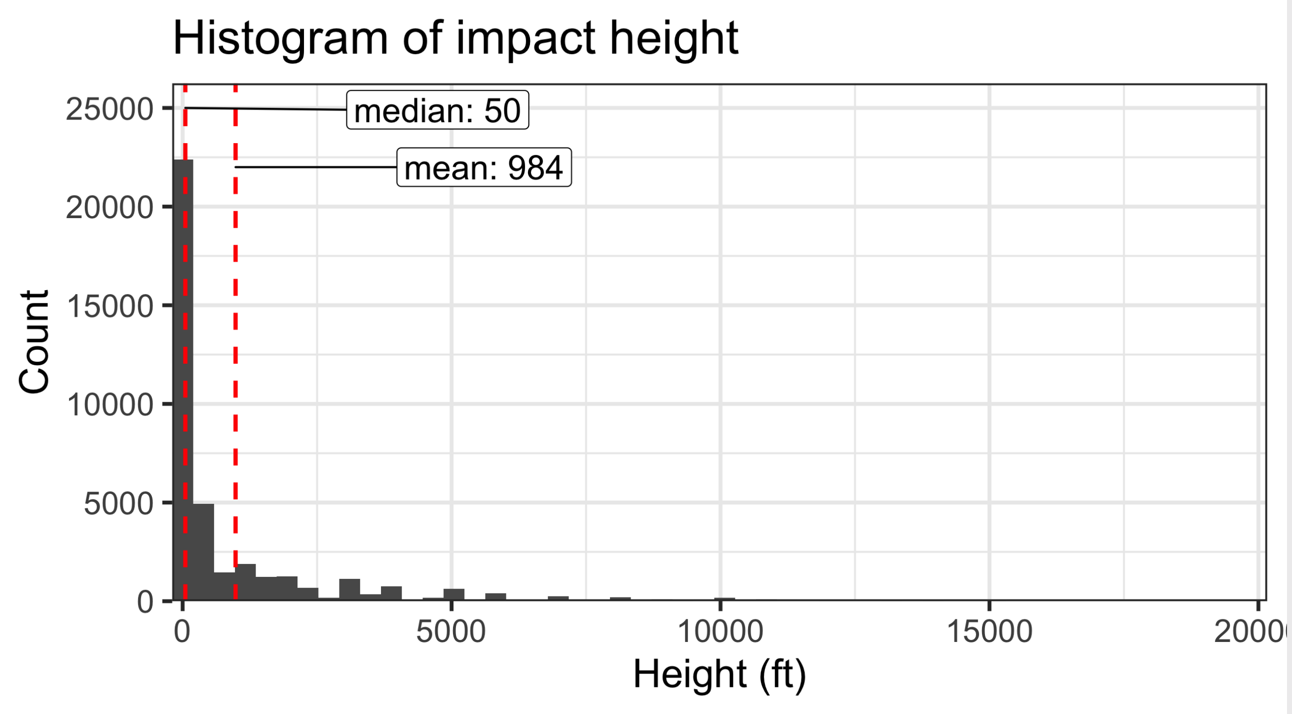
Mean isn't always the "best" choice

```
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:

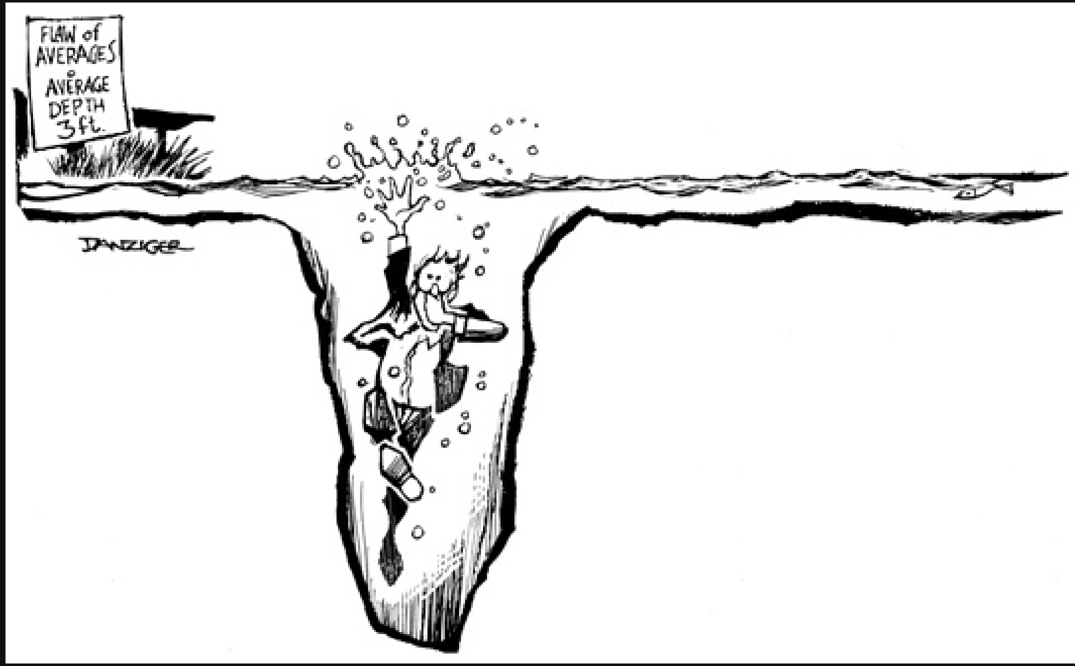
```
#> [1] "73.9%"
```



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:

```
daysToShip
```

```
#>    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:

```r
daysToShip
```

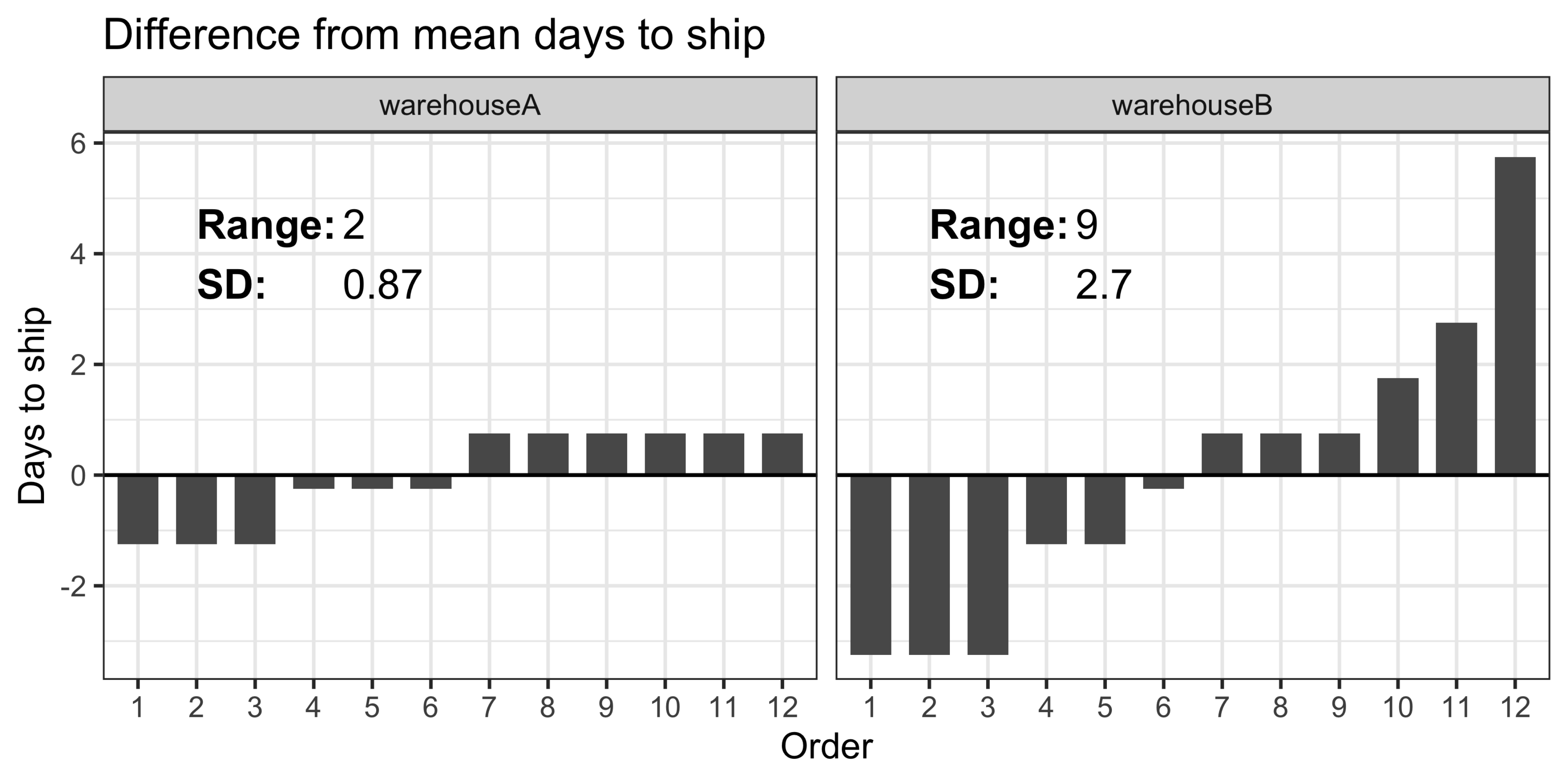
```
#>    order warehouseA warehouseB
#> 1      1          3          1
#> 2      2          3          1
#> 3      3          3          1
#> 4      4          4          3
#> 5      5          4          3
#> 6      6          4          4
#> 7      7          5          5
#> 8      8          5  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

Variability reveals difference in days to ship:

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

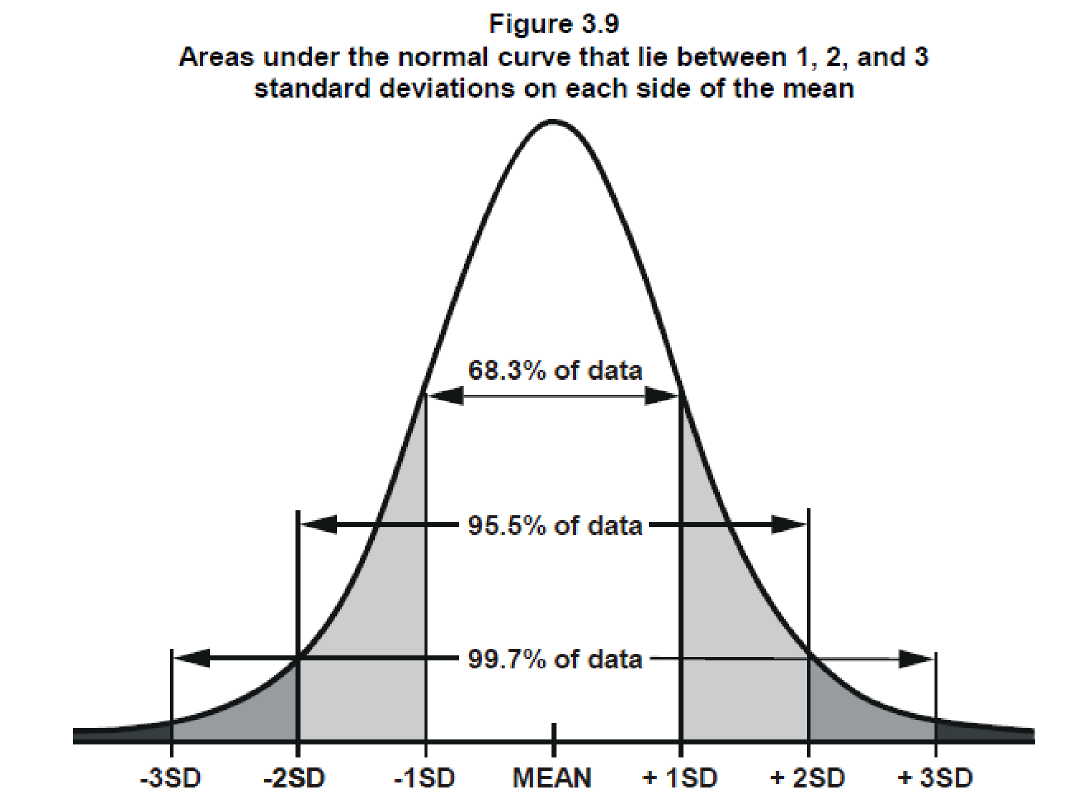
```
#> # A tibble: 2 x 5
#>   warehouse   mean median range    sd
#>   <chr>      <dbl>  <dbl> <dbl> <dbl>
#> 1 warehouseA  4.25    4.5     2 0.866
#> 2 warehouseB  4.25    4.5     9 2.70
```

Example: Days to ship



Interpreting the standard deviation

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



Outliers

Mean & **Standard Deviation** are sensitive to outliers

Outliers: $Q_1 \pm 1.5IQR$

Extreme values: $Q_1 \pm 3.0IQR$

```r
data1 <- c(3,3,4,5,5,6,6,7,8,9)
```

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

```r
data2 <- c(3,3,4,5,5,6,6,7,8,20)
```

- Mean: 6.7
- Standard Deviation: 4.95
- Median: 5.5
- IQR: 2.5

Robust statistics for continuous data

Centrality: Use *median* rather than *mean*

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

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

1. Centrality

2. Variability

Week 3: *Centrality & Variability*

1. Data Types
2. Measures of Centrality & Variability
3. Visualizing Centrality & Variability

BREAK

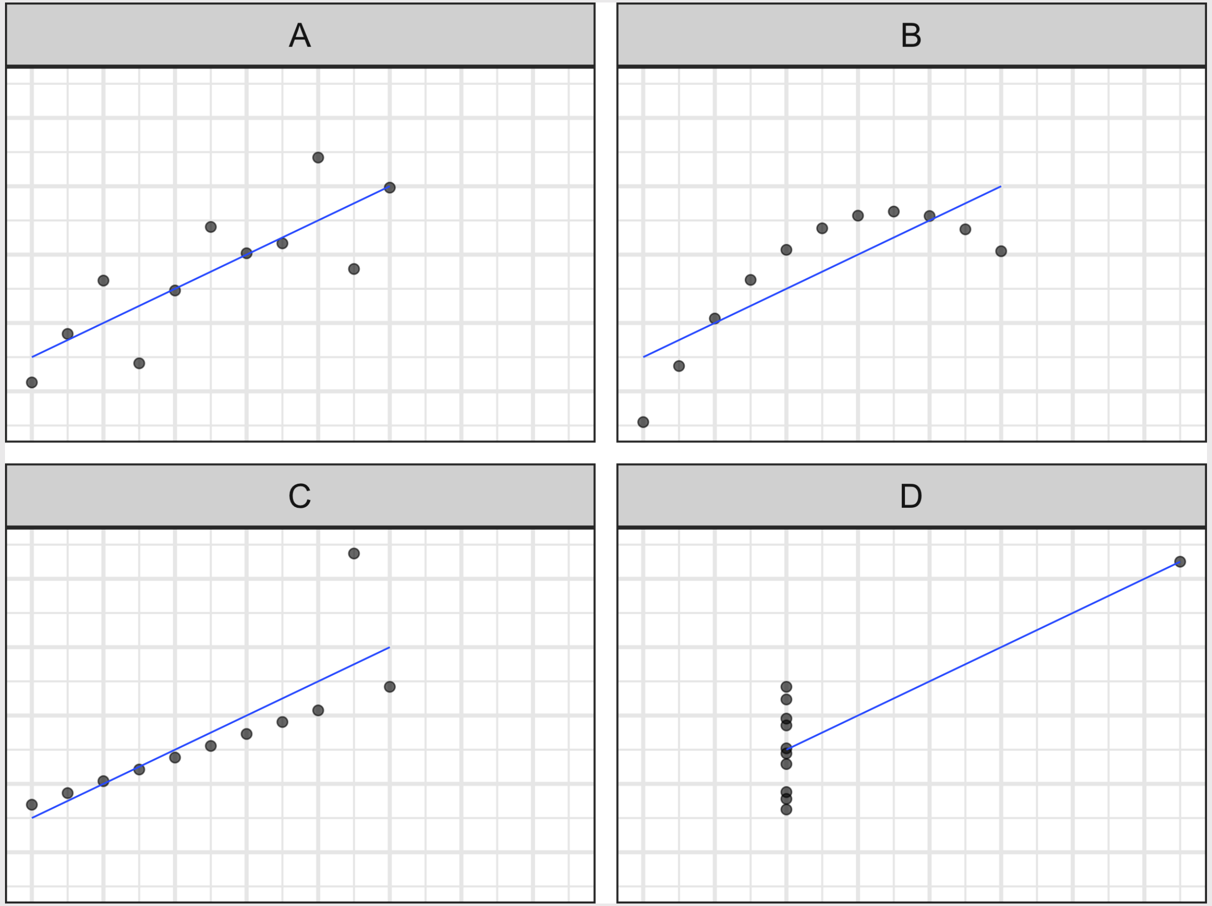
4. Relationships Between 2 Variables
5. Exploratory Data Analysis

"Visualizing data helps us think"

	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

Stephen Few (2009, pg. 6)

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

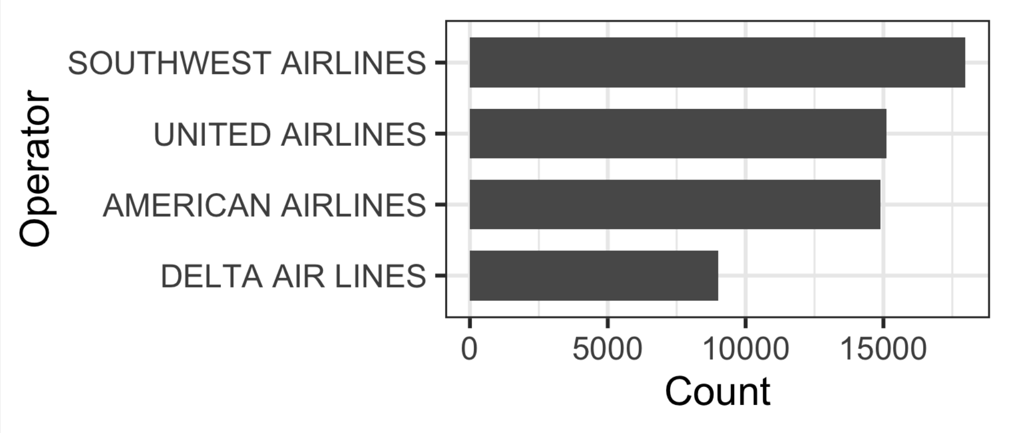
Summarize with counts / percentages

```
wildlife_impacts %>%
    count(operator, sort = TRUE) %>%
    mutate(p = n / sum(n))
```

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

Visualize with bars

```
wildlife_impacts %>%
    count(operator, sort = TRUE) %>%
    ggplot() +
    geom_col(aes(x = n, y = reorder(operator, n)),
             width = 0.7) +
    labs(x = "Count", y = "Operator")
```



Summarizing **Ordinal** data

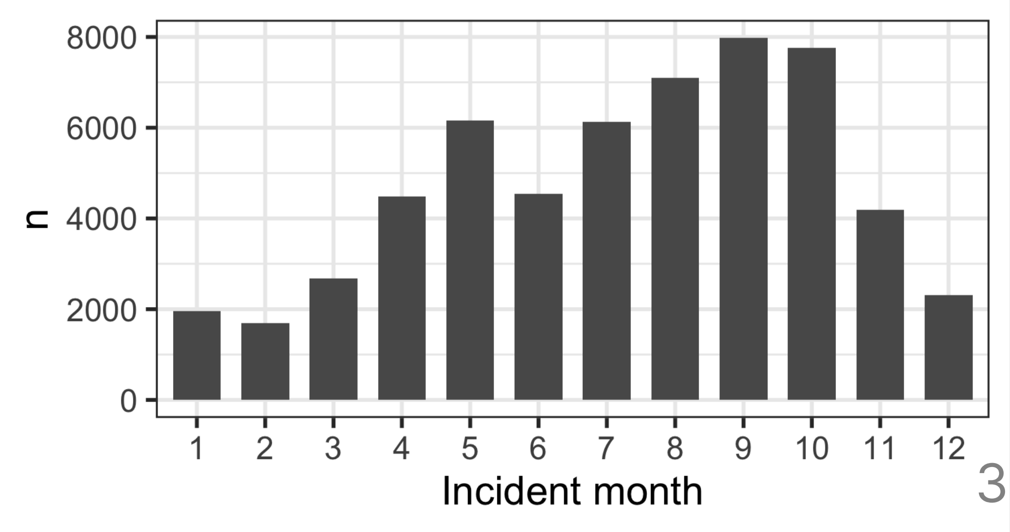
Summarize: Counts / percentages

```
wildlife_impacts %>%
    count(incident_month, sort = TRUE) %>%
    mutate(p = n / sum(n))
```

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

Visualize: Bars

```
wildlife_impacts %>%
    count(incident_month, sort = TRUE) %>%
    ggplot() +
    geom_col(aes(x = as.factor(incident_month),
                 y = n), width = 0.7) +
    labs(x = "Incident month")
```



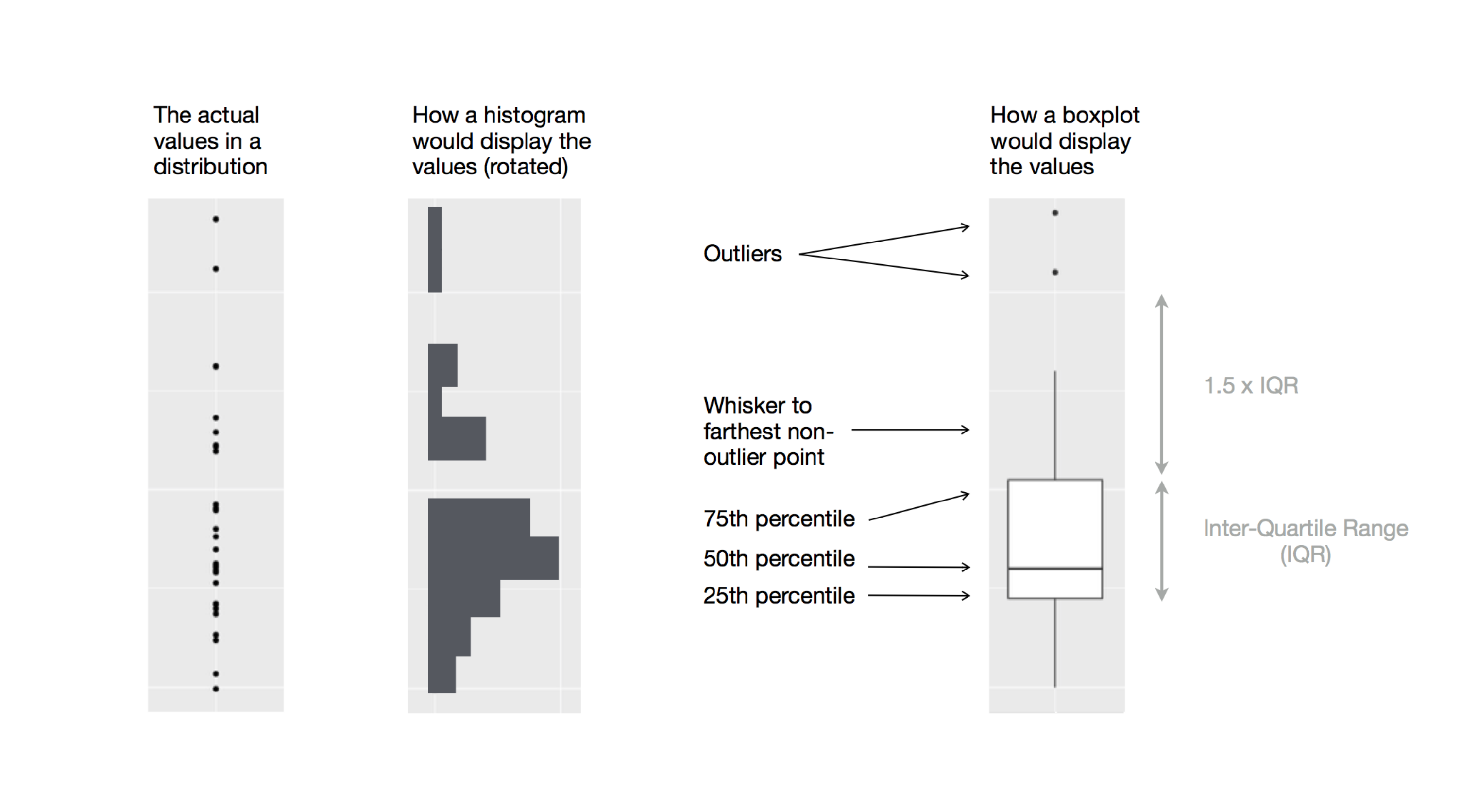
Summarizing **continuous** variables

Histograms:

- Identifying skewness
- Identifying # of modes

Boxplots:

- Identifying outliers
- Comparing distributions across groups



Histogram: Identify Skewness & # of Modes

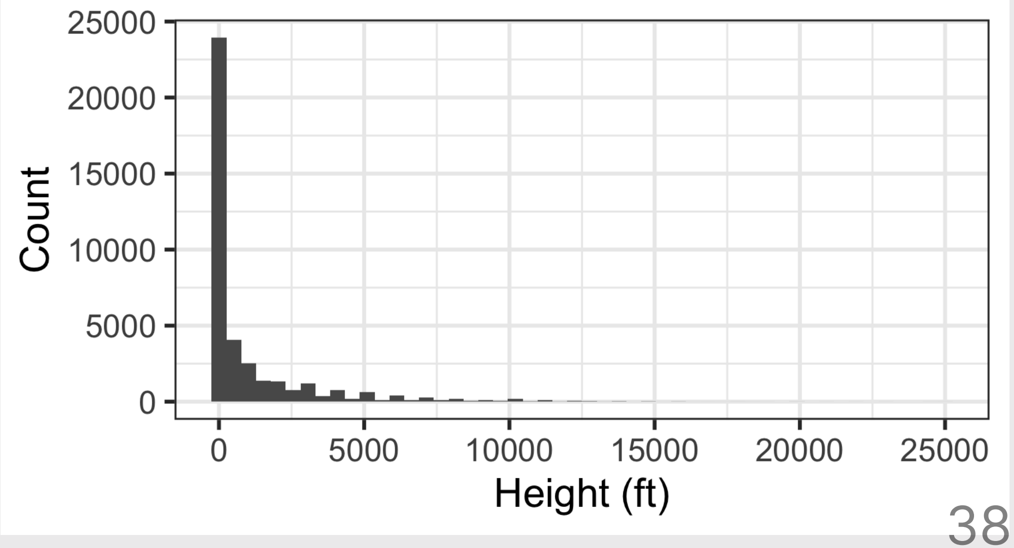
Summarise:
Mean, median, sd, range, & IQR:

```
wildlife_impacts %>%
    filter(! is.na(height)) %>%
    summarise(
        mean   = mean(height),
        median = median(height),
        sd     = sd(height),
        range  = max(height) - min(height),
        IQR    = IQR(height))
```

```
#> # A tibble: 1 x 5
#>    mean median    sd range   IQR
#>   <dbl>  <dbl> <dbl> <dbl> <dbl>
#> 1  984.     50 2027. 25000  1000
```

Visualize:
Histogram (identify skewness & modes)

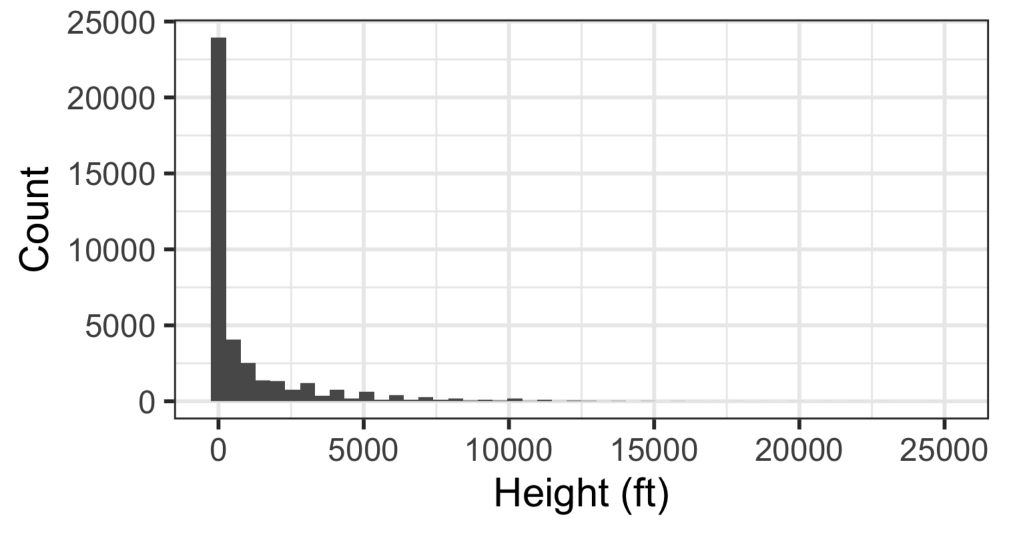
```
ggplot(wildlife_impacts) +
  geom_histogram(aes(x = height), bins = 50)
  labs(x = 'Height (ft)', y = 'Count')
```



Histogram: Identify Skewness & # of Modes

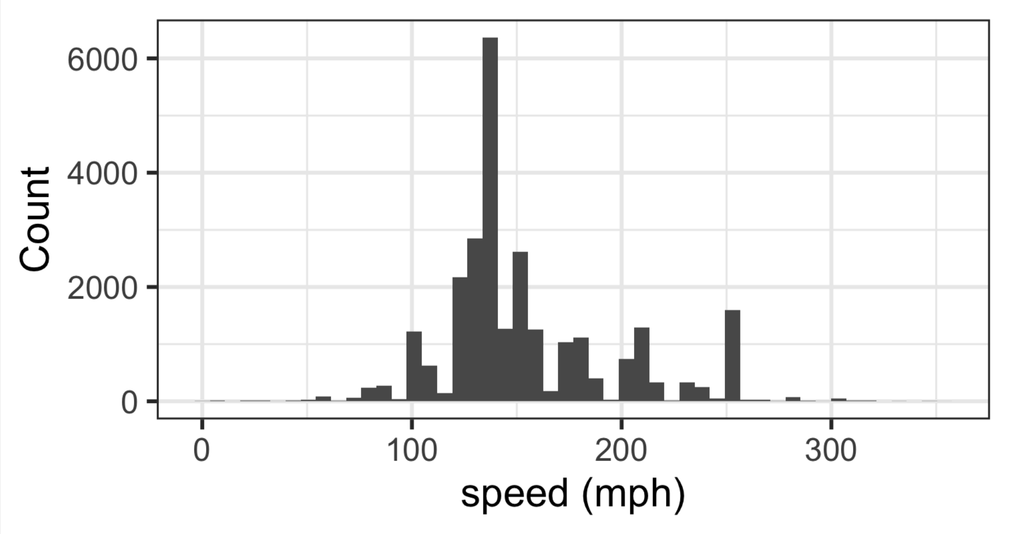
Height

```
ggplot(wildlife_impacts) +
  geom_histogram(aes(x = height), bins = 50)
  labs(x = 'Height (ft)', y = 'Count')
```



Speed

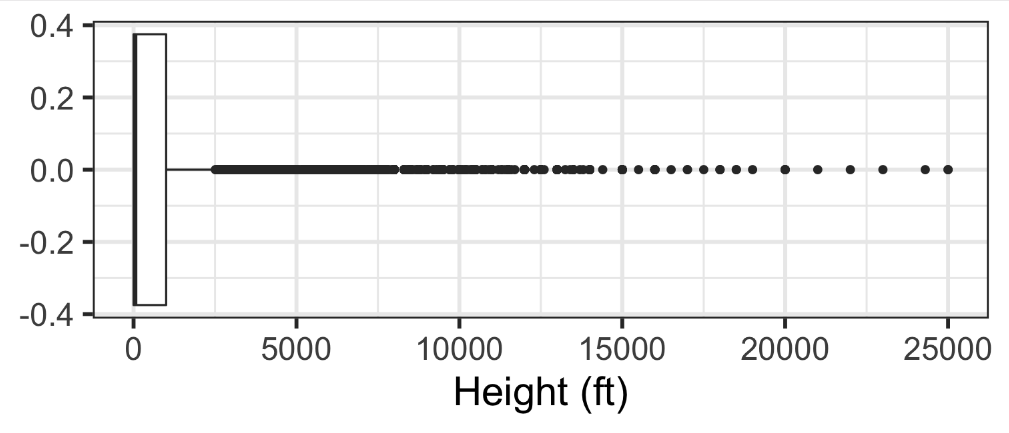
```
ggplot(wildlife_impacts) +
  geom_histogram(aes(x = speed), bins = 50)
  labs(x = 'speed (mph)', y = 'Count')
```



Boxplot: Identify outliers

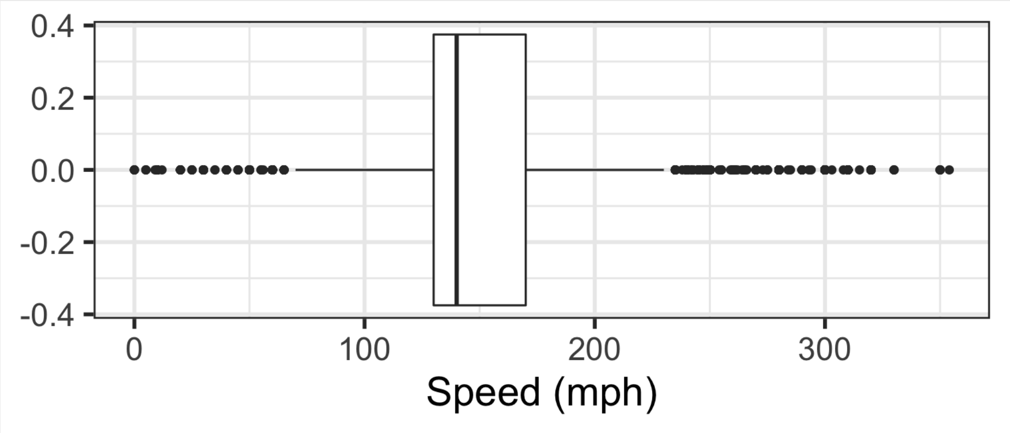
Height

```
ggplot(wildlife_impacts) +
    geom_boxplot(aes(x = height)) +
    labs(x = 'Height (ft)', y = NULL)
```



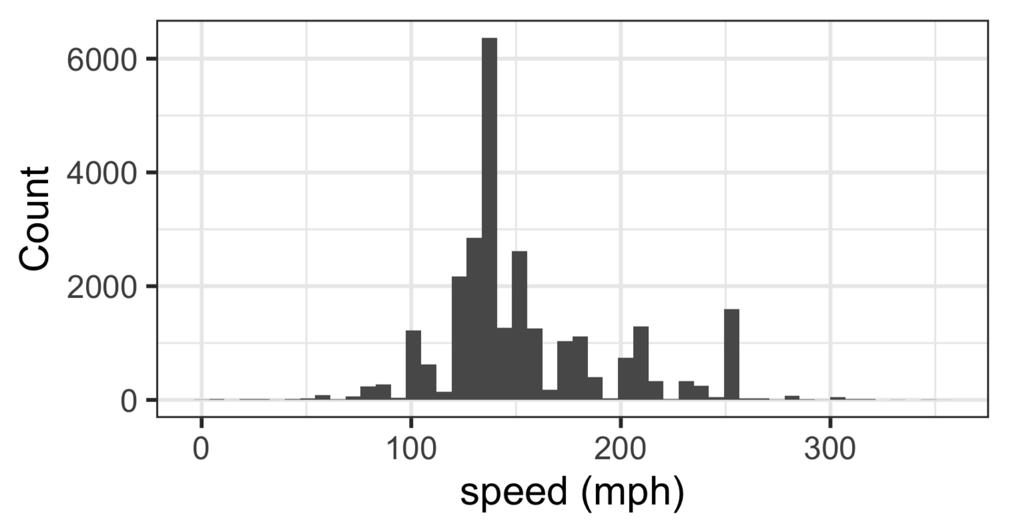
Speed

```
ggplot(wildlife_impacts) +
    geom_boxplot(aes(x = speed)) +
    labs(x = 'Speed (mph)', y = NULL)
```



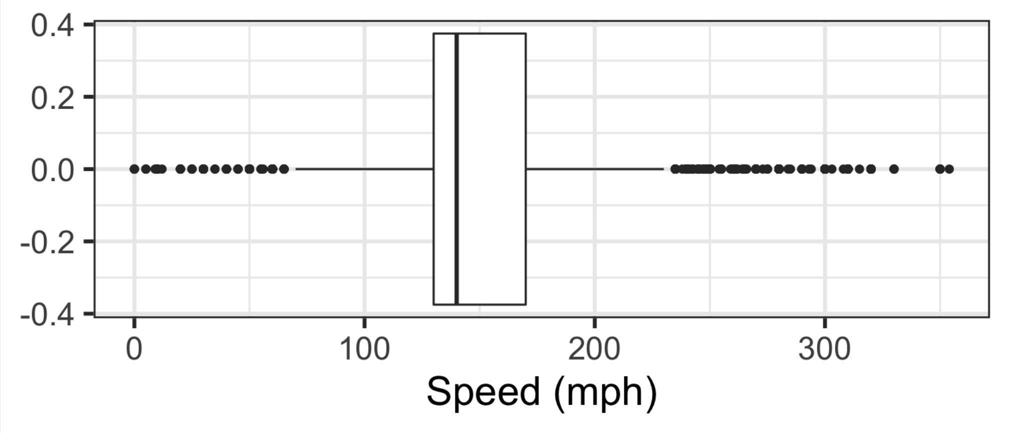
Histogram

- Skewness
- Modes



Boxplot

- Outliers



Practicing 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` ``

Break!

Stand up, Move around, Stretch!

05:00

Week 3: *Centrality & Variability*

1. Data Types

2. Measures of Centrality & Variability

3. Visualizing Centrality & Variability

BREAK

4. Relationships Between 2 Variables

5. Exploratory Data Analysis

Two **Categorical** Variables

Summarize with a table of counts

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

```
#> # A tibble: 20 x 3
#>    operator            time_of_day     n
#>    <chr>               <chr>       <int>
#>  1 AMERICAN AIRLINES   Dawn          458
#>  2 AMERICAN AIRLINES   Day          7809
#>  3 AMERICAN AIRLINES   Dusk          584
#>  4 AMERICAN AIRLINES   Night        3710
#>  5 AMERICAN AIRLINES   <NA>         2326
#>  6 DELTA AIR LINES     Dawn          267
#>  7 DELTA AIR LINES     Day          4846
#>  8 DELTA AIR LINES     Dusk          353
#>  9 DELTA AIR LINES     Night        2090
#> 10 DELTA AIR LINES     <NA>         1449
#> 11 SOUTHWEST AIRLINES  Dawn          394
#> 12 SOUTHWEST AIRLINES  Day          9109
```

Two **Categorical** Variables

Convert to "wide" format with `spread()` to make it easier to compare values

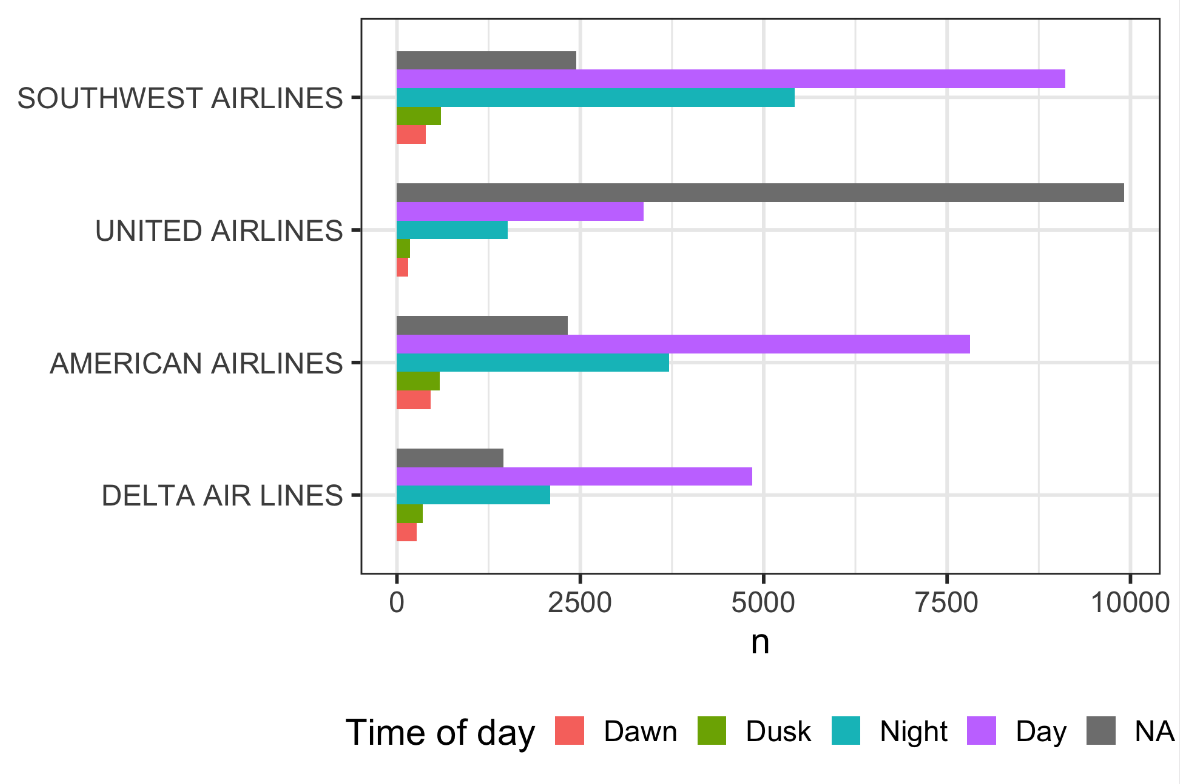
```
wildlife_impacts %>%
    count(operator, time_of_day) %>%
    spread(key = time_of_day, value = n)
```

```
#> # 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
```

Two **Categorical** Variables

Visualize with bars:
map **fill** to denote 2nd categorical var

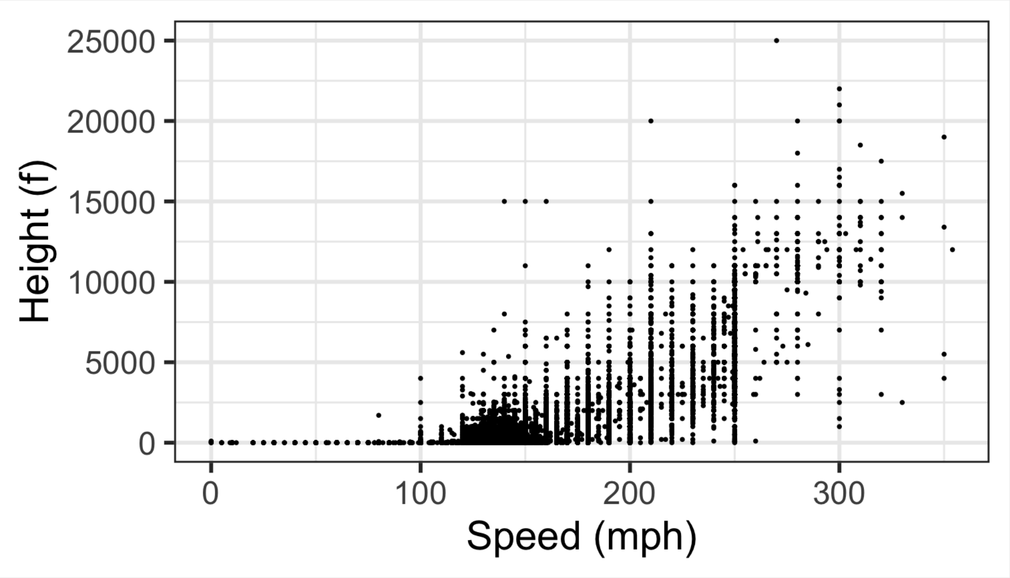
```
wildlife_impacts %>%
  count(operator, time_of_day) %>%
  ggplot() +
  geom_col(aes(x = n,
               y = reorder(operator, n),
               fill = reorder(time_of_d
           width = 0.7,
           position = 'dodge') +
  theme(legend.position = "bottom") +
  labs(fill = "Time of day", y = NULL)
```



Two **Continuous** Variables

Visualize with scatterplot - looking for *correlational* relationship

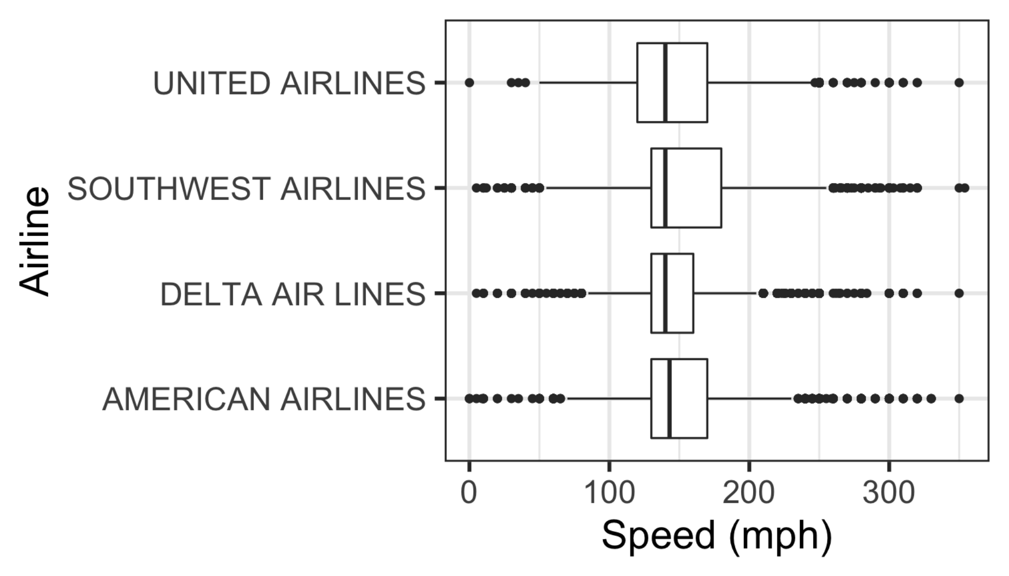
```
ggplot(wildlife_impacts) +
  geom_point(aes(x = speed, y = height),
             size = 0.5) +
  labs(x = 'Speed (mph)',
       y = 'Height (f)')
```



One **Continuous**, One **Categorical**

Visualize with **boxplot**

```
ggplot(wildlife_impacts) +
  geom_boxplot(aes(x = speed,
                   y = operator)) +
  labs(x = 'Speed (mph)',
       y = 'Airline')
```



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: Country & `Official Time`
- wildlife_impacts: `state` & `operator`

Week 3: *Centrality & Variability*

1. Data Types
2. Measures of Centrality & Variability
3. Visualizing Centrality & Variability

BREAK

4. Relationships Between 2 Variables
5. Exploratory Data Analysis

Exploratory Analysis

Goal: **Form** hypotheses.

Improves quality of **questions**.

(do this in THIS class)

Confirmatory Analysis

Goal: **Test** hypotheses.

Improves quality of **answers**.

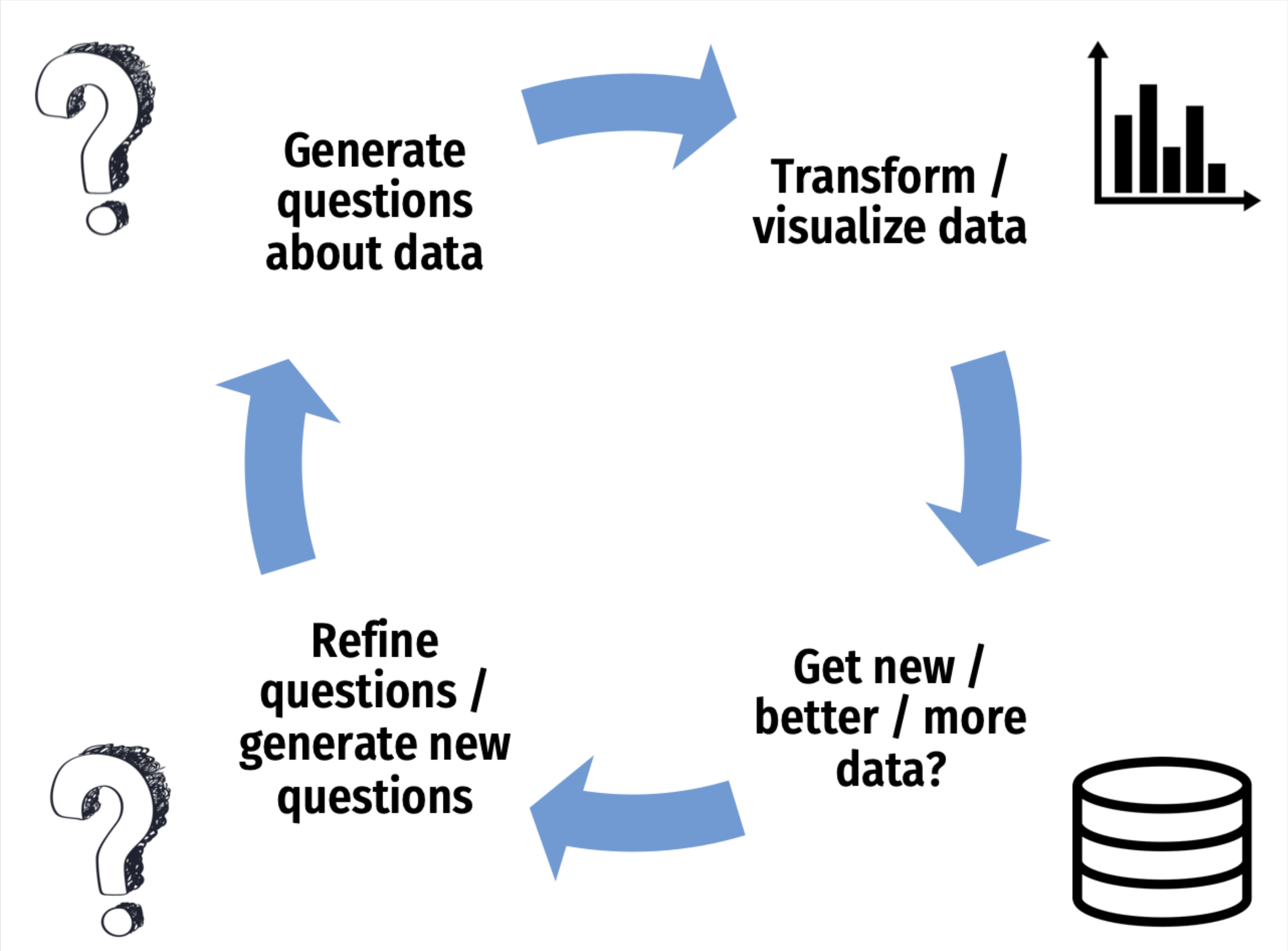
(do this in your stats classes)

Don't be Icarus

"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

EDA is an iterative process to help you *understand* your data and ask better questions

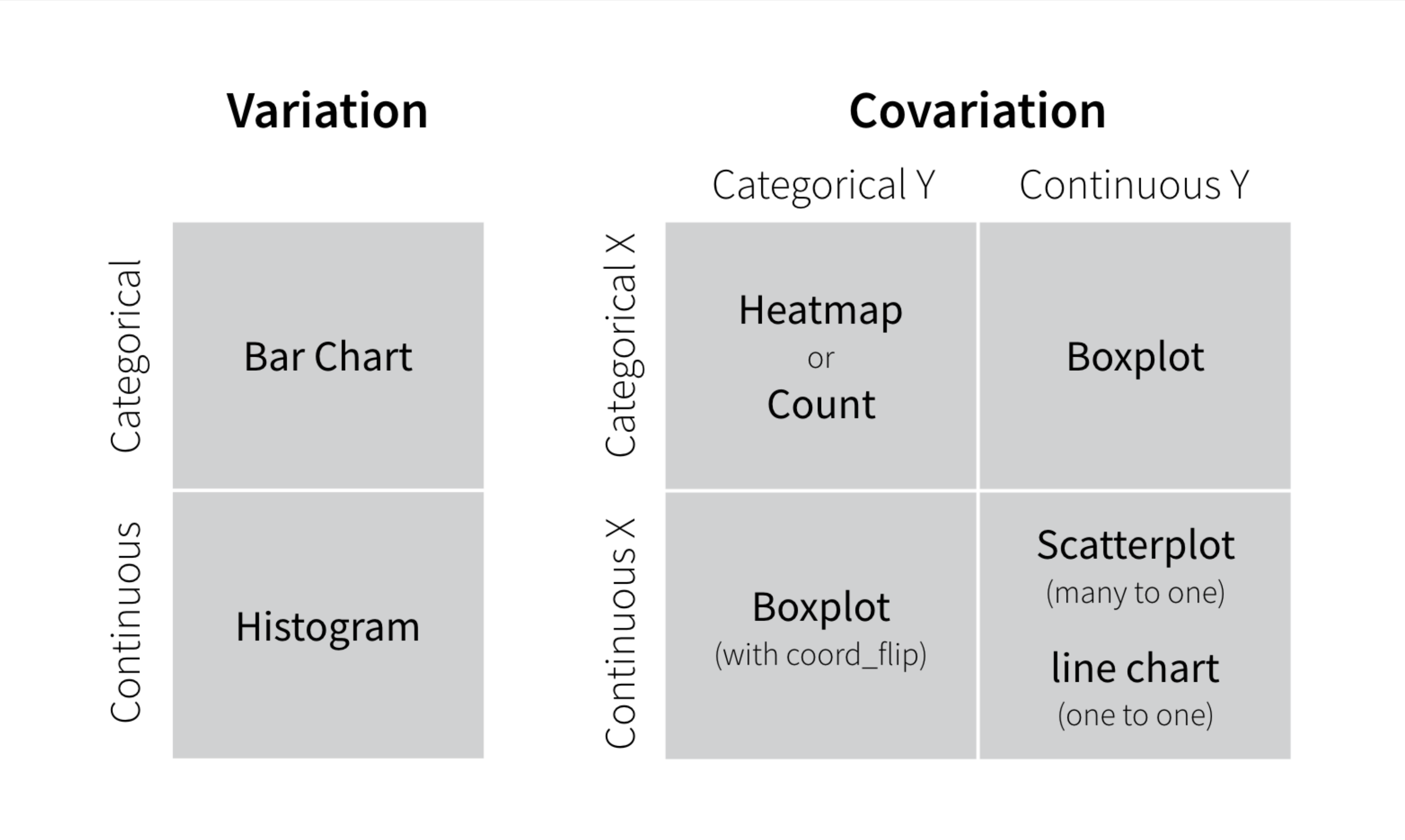


Visualizing variation

Ask yourself:

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

Check out these guides



Practice doing EDA

1) Read in the candy_rankings.csv data sets

2) Preview the data, note the data types and what each variable is.

3) Visualize (at least) three *relationships* between two variables (guided by a question) using an appropriate chart:

- Bar chart
- Scatterplot
- Boxplot

Start thinking about research questions

Writing a research question

Follow these guidelines - your question should be:

- **Clear**: your audience can easily understand its purpose without additional explanation.
- **Focused**: it is narrow enough that it can be addressed thoroughly with the data available and within the limits of the final project report.
- **Concise**: it is expressed in the fewest possible words.
- **Complex**: it is not answerable with a simple "yes" or "no," but rather requires synthesis and analysis of data.
- **Arguable**: its potential answers are open to debate rather than accepted facts (do others care about it?)

Writing a research question

Bad question: Why are social networking sites harmful?

- Unclear: it does not specify *which* social networking sites or state what harm is being caused; assumes that "harm" exists.

Improved question: How are online users experiencing or addressing privacy issues on such social networking sites as Facebook and Twitter?

- Specifies the sites (Facebook and Twitter), type of harm (privacy issues), and who is harmed (online users).

Other good examples: See the Example Projects Page page

Start self-organizing

Find your topic / teammate(s) here (link also in #classroom)