1 Lecture Notes Week 1: Intro to data

Read through Sections 1.2.1 – 1.2.5 in the course textbook prior to coming to class on Friday using the reading guides at the beginning of week 1 material.

Data basics: Sections 1.2.1 – 1.2.2

Data: _____ used to answer research questions

Observational unit or case: the people or things we _____ data from

Variable: what is measured on each _____ or _____.

Types of variables

- Categorical variable:

  Ordinal: levels of the variable have a _____ ordering

  Examples: 'Scale' questions, Years of schooling completed

  Nominal:levels of the variable do _____ have a natural ordering

  Examples: hair color, eye color, zipcode

- Quantitative variable:

  Continuous variables: value can be any _____ within a range.

  Examples: percentage of students who are nursing majors, average hours of exercise per week; distance or time (measured with enough precision)

  Discrete variables: can only be _____ values, with jumps between

  Examples: years of schooling completed; SAT score, number of car accidents

Example: The Bureau of Transportation Statistics collects data on all forms of public transportation. The data set seen here includes several variables collect on flights departing on a random sample of 150 US airports in December of 2019.

```
airport <- read.csv("data/airport_delay.csv")
glimpse(airport)
#> Rows: 150
#> Columns: 19
#> $ airport            <chr> "ABI", "ABY", "ACV", "ACY", "ADQ", "AEX", "ALB", "~
#> $ city               <chr> "Abilene", "Albany", "Arcata/Eureka", "Atlantic Ci~
#> $ state              <chr> " TX", " GA", " CA", " NJ", " AK", " LA", " NY", "~
#> $ airport_name       <chr> " Abilene Regional", " Southwest Georgia Regional"~
```

```
#> $ hub                 <chr> "no", "no", "no", "no", "no", "no", "no", "no", "n~
#> $ international       <chr> "no", "no", "no", "yes", "no", "yes", "yes", "yes"~
#> $ elevation_1000      <dbl> 1.7906, 0.1932, 0.2223, 0.0748, 0.0787, 0.0881, 0.~
#> $ latitude            <dbl> 32.4, 31.5, 41.0, 39.5, 57.7, 31.3, 42.7, 35.2, 45~
#> $ longitude           <dbl> -99.7, -81.2, -124.1, -74.6, -152.5, -92.5, -73.8,~
#> $ arr_flights         <int> 195, 81, 215, 293, 54, 282, 943, 410, 53, 32314, 6~
#> $ perc_delay15        <dbl> 16.410256, 13.580247, 23.255814, 15.358362, 12.962~
#> $ perc_cancelled      <dbl> 0.5128205, 0.0000000, 4.1860465, 0.6825939, 14.814~
#> $ perc_diverted       <dbl> 0.00000000, 0.00000000, 2.32558139, 0.68259386, 0.~
#> $ arr_delay           <int> 1563, 1244, 4763, 2905, 329, 1293, 15127, 9705, 25~
#> $ carrier_delay       <int> 459, 890, 1613, 476, 180, 302, 5627, 2253, 439, 10~
#> $ weather_delay       <int> 21, 43, 549, 124, 1, 58, 2346, 168, 1236, 13331, 2~
#> $ nas_delay           <int> 257, 39, 154, 771, 51, 112, 2096, 616, 746, 45674,~
#> $ security_delay      <int> 0, 0, 0, 25, 0, 0, 44, 0, 0, 375, 0, 83, 0, 23, 0,~
#> $ late_aircraft_delay <int> 826, 272, 2447, 1509, 97, 821, 5014, 6668, 108, 10~
```

- What are the observational units?

- Identify which variables are categorical.

- Identify which variables are quantitative.

Exploratory data analysis (EDA)

Summary statistic: a number which _____ an entire data set

- Also called the _____ _____

  Examples:

  proportion of people who had a stroke

  mean (or average) age

- Summary statistic and type of plot used depends on the type of variable(s)!

Roles of variables: Sections 1.2.3 – 1.2.5

Explanatory variable: predictor variable

- The variable researchers think *may be* _____ the other variable.

- In an experiment, what the researchers _____ or _____.
- The groups that we are comparing from the data set.

Response variable:

- The variable researchers think *may be* _____ by the other variable.
- Always simply _____ or _____; never controlled by researchers.

Examples:

Can you predict a criminal's height based on the footprint left at the scene of a crime?

- Identify the explanatory variable:

- Identify the response variable:

Does marking an item on sale (even without changing the price) increase the number of units sold per day, on average?

- Identify the explanatory variable:

- Identify the response variable:

In the Physician's Health Study, male physicians participated in a study to determine whether taking a daily low-dose aspirin reduced the risk of heart attacks. The male physicians were randomly assigned to the treatment groups. After five years, 104 of the 11,037 male physicians taking a daily low-dose aspirin had experienced a heart attack while 189 of the 11,034 male physicians taking a placebo had experienced a heart attack.

- Identify the explanatory variable:

- Identify the response variable:

Relationships between variables

- Association: the _____ between variables create a pattern; knowing something about one variable tells us about the other.
  - Positive association: as one variable _____, the other tends to _____ also.
  - Negative association: as one variable _____, the other tends to _____.
- Independent: no clear pattern can be seen between the _____.

2 Lecture Notes Week 2: Study Design

Sampling Methods: Section 2.1 in the course textbook

The method used to collect data will impact

- Target population: all _____ or _____ of interest
- Sample:_____ or _____ from which data is collected

Example: Many high schools moved to partial or fully online schooling in Spring of 2020. Did students who graduated in 2020 tend to have a lower GPA during freshman year of college than the previous class of college freshmen? A nationally representative sample of 1000 college students who were freshmen in AY19-20 and 1000 college students who were freshmen in AY20-21 was taken to answer this question.

- What is the target population?

- What is the sample?

Good vs. bad sampling

GOAL: to have a sample that is _____ of the _____ _____ on the variable(s) of interest

- Unbiased sample methods:

  Simple random sample

- Biased sampling method:

Types of Sampling Bias

- Selection bias:

Example: Newspaper article from 1936 reported that Landon won the presidential election over Roosevelt based on a poll of 10 million voters. Roosevelt was the actual winner. What was wrong with this poll? Poll was completed using a telephone survey and not all people in 1936 had a telephone. Only a certain subset of the population owned a telephone so this subset was over-represented in the telephone survey. The results of the study, showing that Landon would win, did not represent the target population of all US voters.

- Non-response bias:

- To calculate the non-response rate:

$$\frac{\text{number of people who do not respond}}{\text{total number of people selected for the sample}} \times 100\%$$

- For non-response bias to occur must first select people to participate and then they choose not to.

Example: Selected to complete review of online purchase but choose to not respond.

- Response bias:

Example(s): Police officer pulls you over and asks if you have been drinking. Expect people to say no, whether they have been drinking or not.

- Need to be able to predict how people will respond.

Words of caution:

- Convenience samples: gathering data for those who are easily accessible; online polls

  Selection bias?

  Non-response bias?

  Response bias?

- Random sampling reduces _____ bias, but has no impact on _____ or _____ bias.

Examples

A radio talk show asks people to phone in their views on whether the United States should pay off its debt to the United Nations.

- Selection?

- Non-response?

- Response?

The Wall Street Journal plans to make a prediction for the US presidential election based on a survey of its readers and plans to follow-up to ensure everyone responds.

- Selection?

- Non-response?

- Response?

A police detective interested in determining the extent of drug use by high school students, randomly selects a sample of high school students and interviews each one about any illegal drug use by the student during the past year.

- Selection?

- Non-response?

- Response?

Observational studies, experiments, and scope of inference: Sections 2.2 – 2.4 in the course textbook

- Review
  - Explanatory variable: the variable researchers think *may be* effecting the other variable.
  - Response variable: the variable researchers think *may be* influenced by the other variable.
- Confounding variable:
  - associated with both the explanatory and the response variable
  - explains the association shown by the data

Example:

Study design

- Observational study:

- Experiment:

Principles of experimental design

- Control

- Randomization

- Replication

- Blocking

Example: It is well known that humans have more difficulty differentiating between faces of people from different races than people within their own race. A 2018 study published in the Journal of Experimental Psychology: Human Perception and Performance investigated a similar phenomenon with gender. In the study, volunteers were shown several pictures of strangers. Half the volunteers were randomly assigned to

rate the attractiveness of the individuals pictured. The other half were told to rate the distinctiveness of the faces seen. Both groups were then shown a slideshow of faces (some that had been rated in the first part of the study, some that were new to the volunteer) and asked to determine if each face was old or new. Researchers found people were better able to recognize faces of their own gender when asked to rate the distinctiveness of the faces, compared to when asked to rate the attractiveness of the faces.

- What is the study design?

Example: In the Physician's Health Study, male physicians participated in a study to determine whether taking a daily low-dose aspirin reduced the risk of heart attacks. The male physicians were randomly assigned to the treatment groups. After five years, 104 of the 11,037 male physicians taking a daily low-dose aspirin had experienced a heart attack while 189 of the 11,034 male physicians taking a placebo had experienced a heart attack.

- What is the study design?

- Assuming these data provide evidence that the low-dose aspirin group had a lower rate of heart attacks than the placebo group, is it valid for the researchers to conclude the lower rate of heart attacks was caused by the daily low-dose aspirin regimen?

Scope of Inference

1. How was the sample selected?

- Random sample with no sampling bias:

- Non-random sample with sampling bias:

2. What is the study design?

- Randomized experiment:

- Observational study:

Scope of Inference Table:

	Study Design	
Selection of Cases	Randomized experiment	Observational study
Random sample (and no other sampling bias)		
Non-random sample (or other sampling bias)		

Example: It is well known that humans have more difficulty differentiating between faces of people from different races than people within their own race. A 2018 study published in the Journal of Experimental Psychology: Human Perception and Performance investigated a similar phenomenon with gender. In the study, volunteers were shown several pictures of strangers. Half the volunteers were randomly assigned to rate the attractiveness of the individuals pictured. The other half were told to rate the distinctiveness of the faces seen. Both groups were then shown a slideshow of faces (some that had been rated in the first part of the study, some that were new to the volunteer) and asked to determine if each face was old or new. Researchers found people were better able to recognize faces of their own gender when asked to rate the distinctiveness of the faces, compared to when asked to rate the attractiveness of the faces.

- What is the scope of inference for this study?

Purpose of random assignment:

Purpose of random selection:

3 Lecture Notes Week 3: Exploratory Data Analysis

Summarizing categorical data

- A _____ is calculated on data from a sample
- The parameter of interest is what we want to know from the population.
- Includes:
  - Population word (true, long-run, population)
  - Summary measure (depends on the type of data)
  - Context
    - Observational units
    - Variable(s)

Categorical data can be numerically summarized by calculating a _____ from the data set.

Notation used for the population proportion:

- Single categorical variable:

- Two categorical variables:

  - Subscripts represent the _____ variable groups

Notation used for the sample proportion:

- Single categorical variable:

- Two categorical variables

Categorical data can be reported in a _____table, which plots counts or a _____ frequency table, which plots the proportion.

When we have two categorical variables we report the data in a _____ or two-way table with the _____ variable on the columns and the _____ variable on the rows.

Example: Gallatin Valley is the fastest growing county in Montana. You'll often hear Bozeman residents complaining about the 'out-of-staters' moving in. A local real estate agent recorded data on a random sample of 100 home sales over the last year at her company and noted where the buyers were moving from as well as the age of the person or average age of a couple buying a home. The variable age was binned into two categories, "Under30" and "Over30." Additionally, the variable, state the buyers were moving from, was created as a binary variable, "Out" for a location out of state and "In" for a location in state.

The following code reads in the data set, `moving_to_mt` and names the object moving.

```
moving <- read.csv("data/moving_to_mt.csv")
```

The R function glimpse was used to give the following output.

```
glimpse(moving)
#> Rows: 100
#> Columns: 4
#> $ From      <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", ~
#> $ Age_Group <chr> "Under30", "Under30", "Under30", "Under30", "Under30", "Unde~
#> $ Age       <int> 25, 26, 27, 27, 29, 29, 35, 37, 49, 63, 65, 77, 22, 24, 24, ~
#> $ InOut     <chr> "Out", "Out", "Out", "Out", "Out", "Out", "Out", "Out", "Out~
```

- What are the observational units in this study?

- What type of variable is Age?

- What type of variable is Age_Group?

To further analyze the categorical variable, From, we can create either a frequency table:

```
moving %>%
    count(From)
#>   From  n
#> 1   CA 12
#> 2   CO  8
#> 3   MT 61
#> 4   WA 19
```

Or a relative frequency table:

```
moving %>%
  count(From) %>%
  mutate(freq = n/sum(n))
#>   From  n freq
#> 1   CA 12 0.12
#> 2   CO  8 0.08
#> 3   MT 61 0.61
#> 4   WA 19 0.19
```

- How many buyers are from WA?

- What proportion of sampled buyers are from WA?

- What notation is used for the proportion of buyers that that are from WA?

To look at the relationship between the variable, `Age_Group` and the variable, `From` create the following two-way table using the `R` output below. Note, we are using `From` as the explanatory variable to predict whether a buyer is over or under the age of 30.

```
moving %>%
    group_by(Age_Group) %>% count(From) %>% print(n=8)
```

```
#> # A tibble: 8 x 3
#> # Groups:   Age_Group [2]
#>   Age_Group From      n
#>   <chr>     <chr> <int>
#> 1 Over30    CA        6
#> 2 Over30    CO        2
#> 3 Over30    MT       47
#> 4 Over30    WA       10
#> 5 Under30   CA        6
#> 6 Under30   CO        6
#> 7 Under30   MT       14
#> 8 Under30   WA        9
```

	State				
Age Group	CA	CO	MT	WA	Total
Over30					
Under30					
Total					

- Using the table above, how many of sampled buyers were under 30 years old and from Montana?

If we want to know what proportion of each age group is from each state, we would calculate the proportion of buyers from each _____ within each _____. In other words, divide the number of buyers from each state that are over 30 by the total for row 1, the total number of buyers over 30.

- What proportion of under 30-year-old sampled buyers were from California?

- What notation should be used for this value?

Additionally, we could find the proportion of buyers in each state for each age group. Here we would calculate the proportion of buyers in each _____ within each _____. Divide the number of buyers in each age group from CA by the total for column 1, the total number of buyers from CA.
Fill in the following table, to find the column proportions.

	State			
Age Group	CA	CO	MT	WA
Over30				
Under30				

- What does the value 0.770 represent?

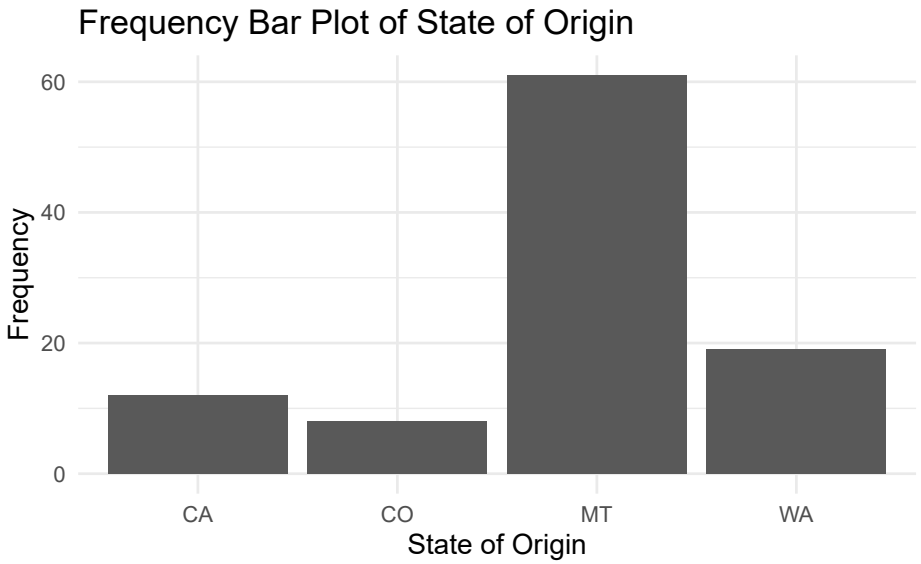
Displaying categorical variables

- Types of plots for a single categorical variable

- Types of plots for two categorical variables

The following code in `R` will create a frequency bar plot of the variable, `From`.

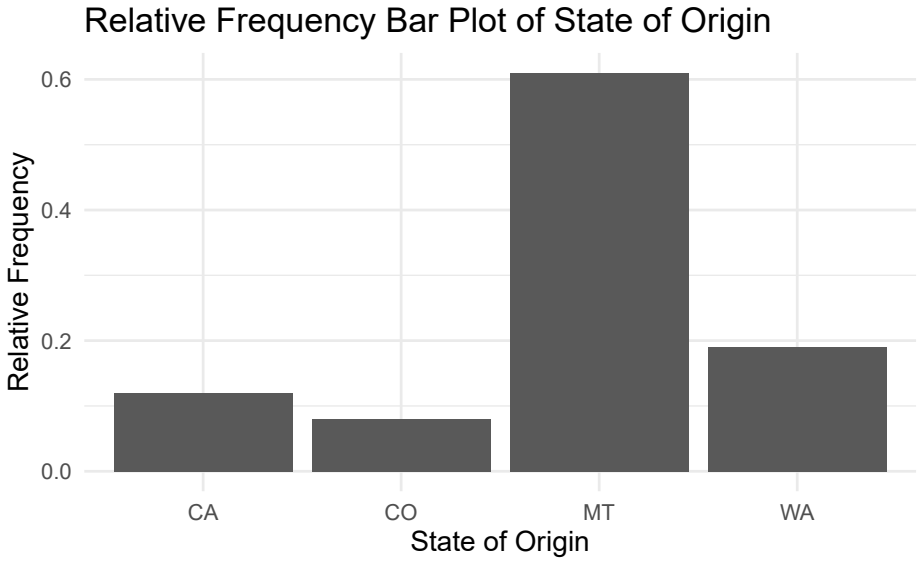
```
moving %>%
    ggplot(aes(x = From))+ #Enter the variable to plot
    geom_bar(stat = "count") +
    labs(title = "Frequency Bar Plot of State of Origin", #Title your plot
       y = "Frequency", #y-axis label
       x = "State of Origin") #x-axis label
```



- What can we see from this plot?

Additionally, we can create a relative frequency bar plot.

```
moving %>%
  ggplot(aes(x = From))+ #Enter the variable to plot
  geom_bar(aes(y = after_stat(prop), group = 1)) +
  labs(title = "Relative Frequency Bar Plot of State of Origin", #Title your plot
       y = "Relative Frequency", #y-axis label
       x = "State of Origin") #x-axis label
```

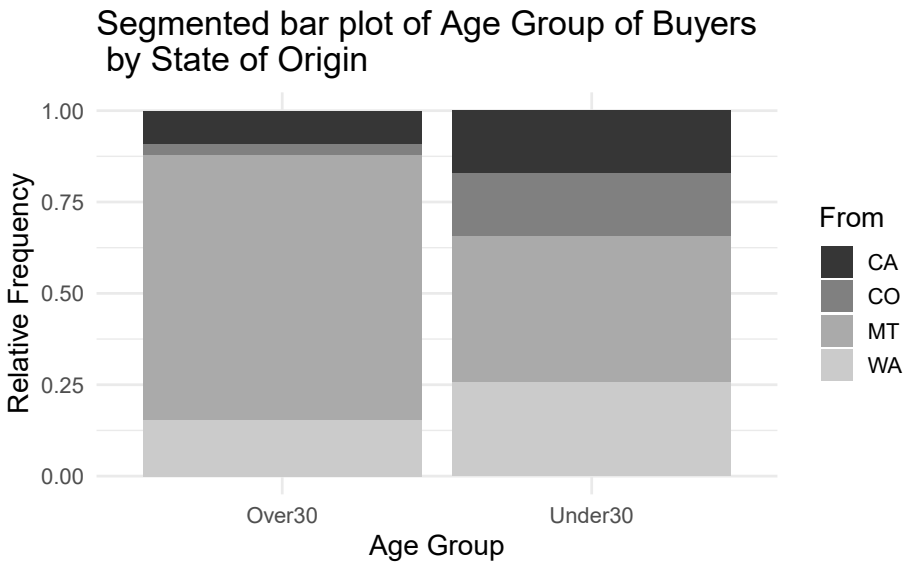


- Note: the x-axis is the _____ between the frequency bar plot and the relative frequency bar plot. However, the _____ differs. The scale for the frequency bar plot

goes from _____ and the scale for the relative frequency bar plot is from _____.

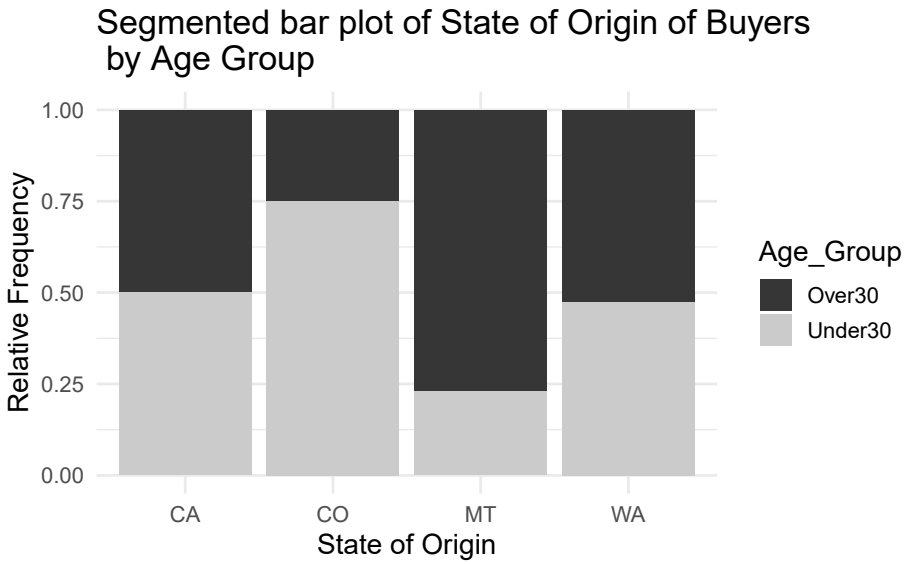
In a segmented bar plot, the bar for each category will sum to 1. In this first plot, we are plotting the row proportions calculated conditional on the age group.

```
moving %>%
  ggplot(aes(x = Age_Group, fill = From))+ #Enter the variables to plot
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of Age Group of Buyers \n by State of Origin", #Title your plot
       y = "Relative Frequency", #y-axis label
       x = "Age Group") + #x-axis label
  scale_fill_grey()
```



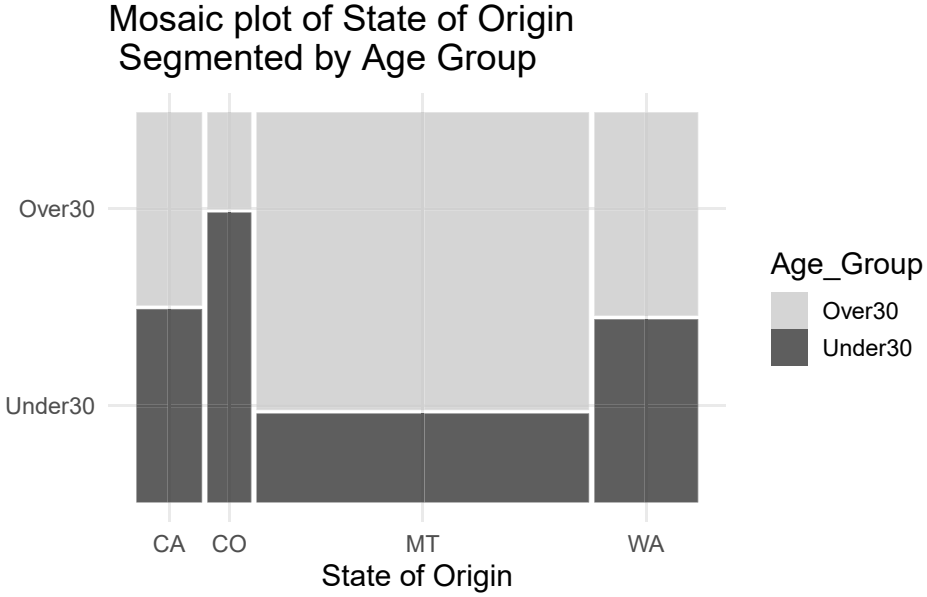
In this second plot, we are plotting the column proportions calculated conditional on the state of origin for the buyer.

```
moving %>%
  ggplot(aes(x = From , fill = Age_Group))+ #Enter variables to plot
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of State of Origin of Buyers \n by Age Group", #Title your plot
       y = "Relative Frequency", #y-axis label
       x = "State of Origin") + #x-axis label
  scale_fill_grey()
```



Mosaic plot:

```
moving$Age_Group <- factor(moving$Age_Group, levels = c("Under30", "Over30"))
moving %>% # Data set piped into...
  ggplot() +   # This specifies the variables
  geom_mosaic(aes(x=product(From), fill = Age_Group)) +
    # Tell it to make a mosaic plot
  labs(title = "Mosaic plot of State of Origin \n Segmented by Age Group",
       # Make sure to title your plot
       x = "State of Origin",   # Label the x axis
       y = "") +  # Remove y axis label
    scale_fill_grey(guide = guide_legend(reverse = TRUE)) # Make figure color
```



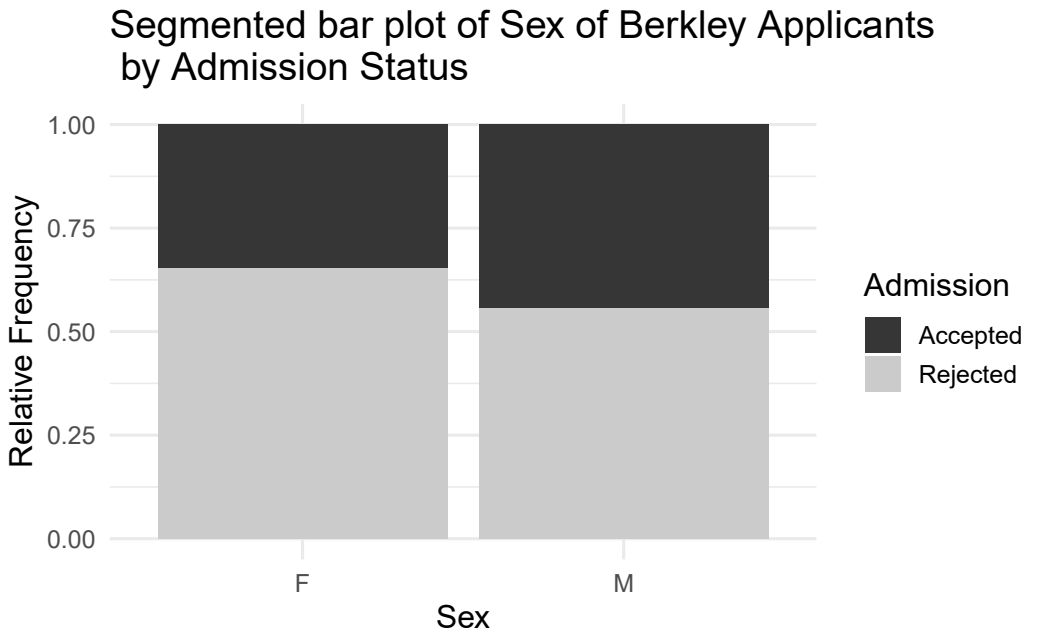
- Why is the bar for MT the widest on the mosaic plot?

Simpson's paradox

- When an apparent _____ between explanatory and response variables reverses when accounting for _____ variable.

Example: The "Berkeley Dataset" contains all 12,763 applicants to UC-Berkeley's graduate programs in Fall 1973. This dataset was published by UC Berkeley researchers in an analysis to understand the possible gender bias in admissions and has now become a classic example of Simpson's Paradox.

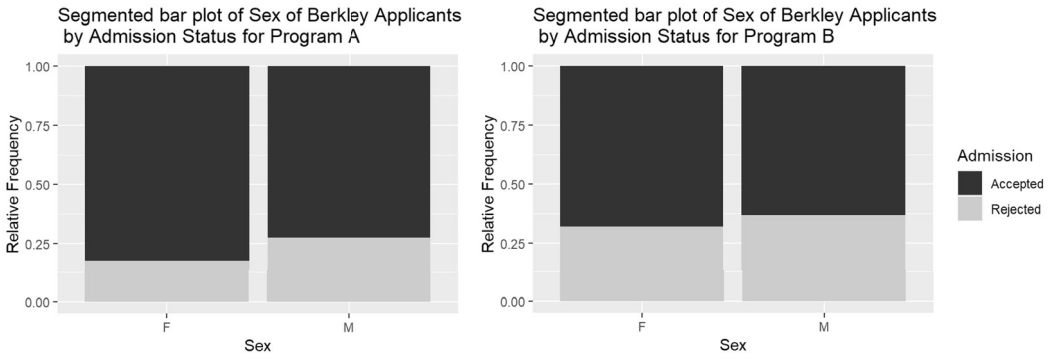
```
discrim <- read.csv ("https://waf.cs.illinois.edu/discovery/berkeley.csv")

discrim %>%
  ggplot(aes(x =Gender, fill = Admission))+
  geom_bar(stat = "count", position = "fill") +
  labs(title = "Segmented bar plot of Sex of Berkley Applicants \n by Admission Status",
       y = "Relative Frequency",
       x = "Sex") +
  scale_fill_grey()
```



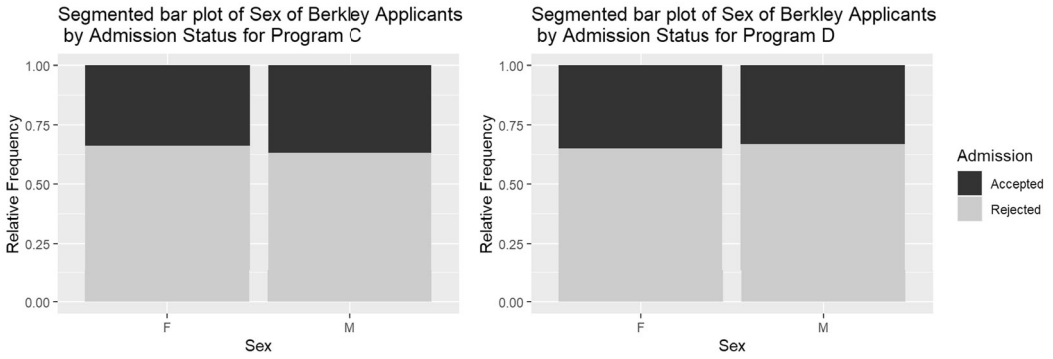
The data showed that 44% of male applicants were accepted and 35% of female applicants were accepted. Does it appear that the female students are discriminated against?

We can break down the data by major. A major code (either A, B, C, D, E, F, or Other) was used.

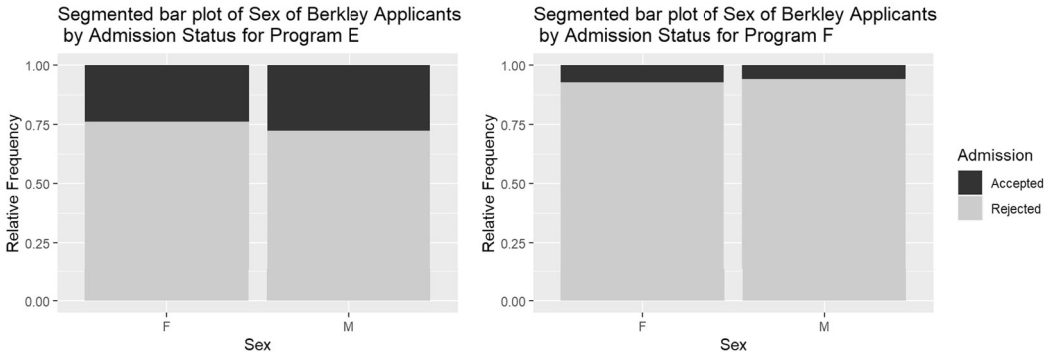
Here we look at the relationship between admission status and sex for Program A and for Program B.



Showing Program C and Program D.



And finally, Program E and F.



We can see in several programs the acceptance rate is higher for females than for males.

Summarizing quantitative data

Quantitative data can be numerically summarized by finding:

Two measures of center:

- Mean:

- Median:

Two measures of spread:

- Standard deviation:

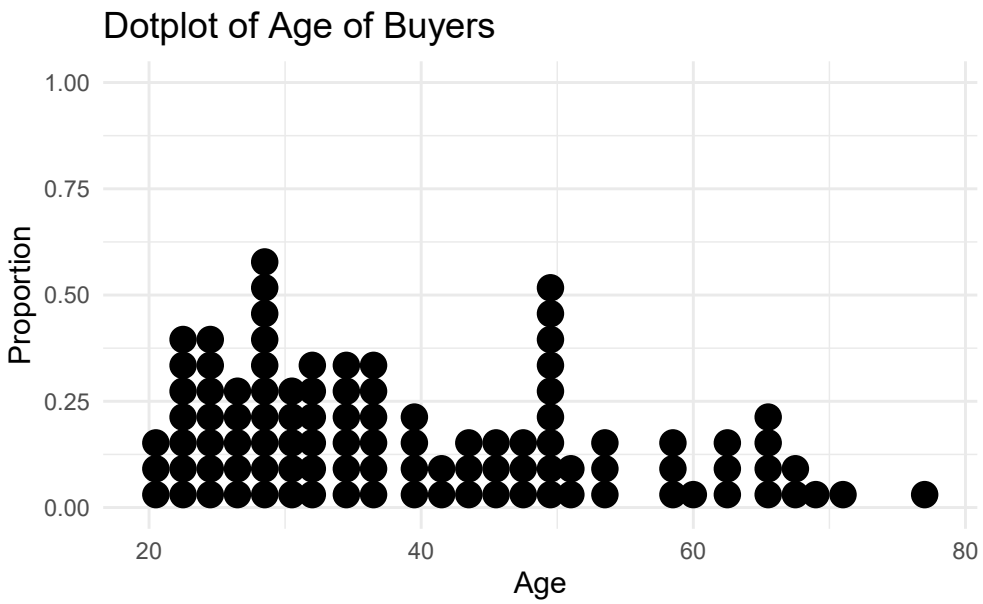
- Interquartile range:

Types of plots

We will revisit the moving to Montana data set and plot the age of the buyers.

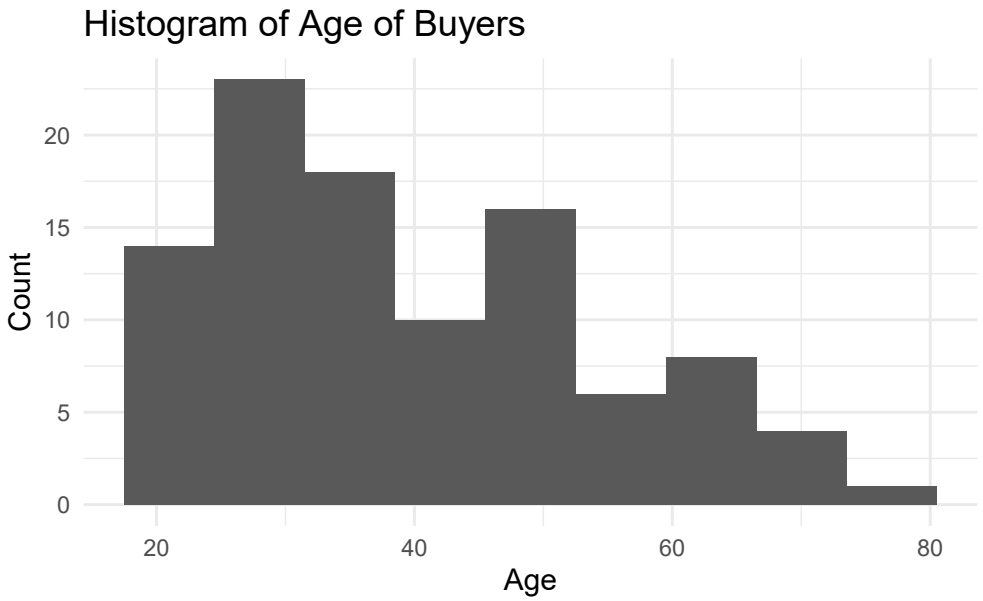
- Dotplot:

```
moving %>%
  ggplot(aes(x = Age))+ #Enter variable to plot
  geom_dotplot() +
  labs(title = "Dotplot of Age of Buyers", #Title your plot
       x = "Age", #x-axis label
       y = "Proportion") #y-axis label
```



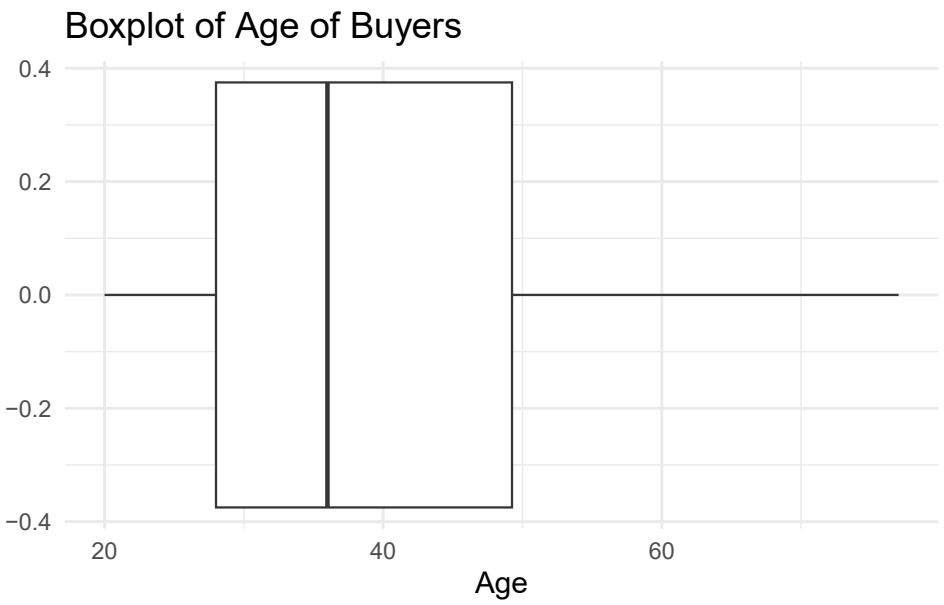
- Histogram

```
moving %>%
  ggplot(aes(x = Age))+
  geom_histogram(binwidth = 7) +
  labs(title = "Histogram of Age of Buyers",
       x = "Age",
       y = "Count")
```



- Boxplot

```
moving %>%
  ggplot(aes(x = Age))+ #Enter variable to plot
  geom_boxplot() +
  labs(title = "Boxplot of Age of Buyers", #Title your plot
       x = "Age", #x-axis label
       y = "") #y-axis label
```




```
favstats(moving$Age)
#>  min Q1 median    Q3 max  mean       sd   n missing
#>   20 28     36 49.25  77 39.77 14.35471 100       0
```

Interpret the value of Q_3 for the age of buyers.

Interpret the value of s for the age of buyers.

Calculate the IQR for the age of buyers.

Four characteristics of plots for quantitative variables

- Shape:

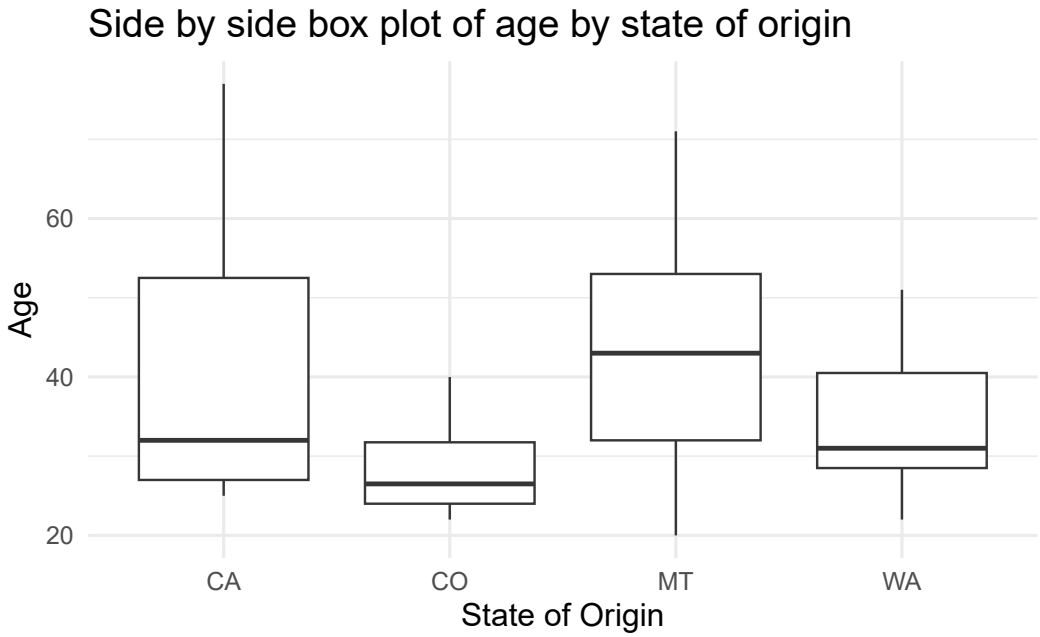
- Center:

- Spread (or variability):

- Outliers?

Let's look at side-by-side boxplot of the variable age by state of origin moved from.

```
moving %>%  # Data set piped into...
  ggplot(aes(y = Age, x = From))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Side by side box plot of age by state of origin",  # Title
       x = "State of Origin",    # x-axis label
       y = "Age")  # y-axis label
```



- Which state of origin had the oldest median age of sampled buyers?

- Which state of origin had the most variability in age of sampled buyers?

- Which state of origin had the most symmetric distribution of ages of sampled buyers?

- Which state of origin had outliers?

Robust statistics

Let's review the summary statistics and histogram of age of buyers.


```
#>  min Q1 median    Q3 max  mean       sd   n missing
#>   20 28     36 49.25  77 39.77 14.35471 100       0
```

Notice that the _____ has been pulled in the direction of the _____.

- The _____ is a _____ measure of center.
- The _____ is a _____ measure of spread.
- Robust means not _____ by.

When the distribution is symmetric use the _____ as the measure of center and the _____ as the measure of spread.

When the distribution is skewed with outliers use the _____ as the measure of center and the _____ as the measure of spread.

4 Lecture Notes Week 4: Regression and Correlation

Summary measures and plots for two quantitative variables

A _____ is used to display the relationship between two _____ variables.

Four characteristics of the scatterplot:

- Form:

- Direction:

- Strength:

- Outliers:

    - Influential points: outliers that change the regression line; far from the line of regression
    - High leverage points: outliers that are extreme in the x- axis; far from the mean of the x-axis

The summary measures for two quantitative variables are:

- _____
- _____
- _____

- Least-squares regression line: $\hat{y} = b_0 + b_1 \times x$ (put y and x in the context of the problem) or $\widehat{response} = b_0 + b_1 \times$ explanatory
- $\hat{y}$ or $\widehat{\text{response}}$ is

- b_0 is

- b_1 is

- x or explanatory is

- The estimates for the linear model output will give the value of the _____ and the _____.
- Interpretation of slope: an increase in the _____ variable of 1 unit is associated with an increase/decrease in the _____ variable by the value of slope, on average.
- Interpretation of the y-intercept: for a value of 0 for the _____ variable, the predicted value for the _____ variable would be the value of y-intercept.
- We can predict values of the _____ variable by plugging in a given _____ variable value using the least squares equation line.
- A prediction of a response variable value for an explanatory value outside the range of x values is called _____.
- To find how far the predicted value deviates from the actual value we find the _____.

- To find the least squares regression line the line with the _____ SSE is found.

  SSE =

  - To find SSE, the _____ for each data point is found, squared and all the squared residuals are summed together

Correlation is always between the values of _____ and _____.

- Measures the _____ and _____ of the linear relationship between two quantitative variables.
- The stronger the relationship between the variables the closer the value of _____ is to _____ or _____.
- The sign gives the _____.

The coefficient of determination can be found using the _____ for each variable or the SSE and SST (sum of squared total)

- $r^2 = (r)^2 = \frac{SST-SSE}{SST} = \frac{s_y^2 - s_{residual}^2}{s_y^2}$
- The coefficient of determination measures the _____ of total variation in the _____ variable that is explained by the changes in the _____ variable.

Notation:

- Population slope:
- Population correlation:
- Sample slope:
- Sample correlation:

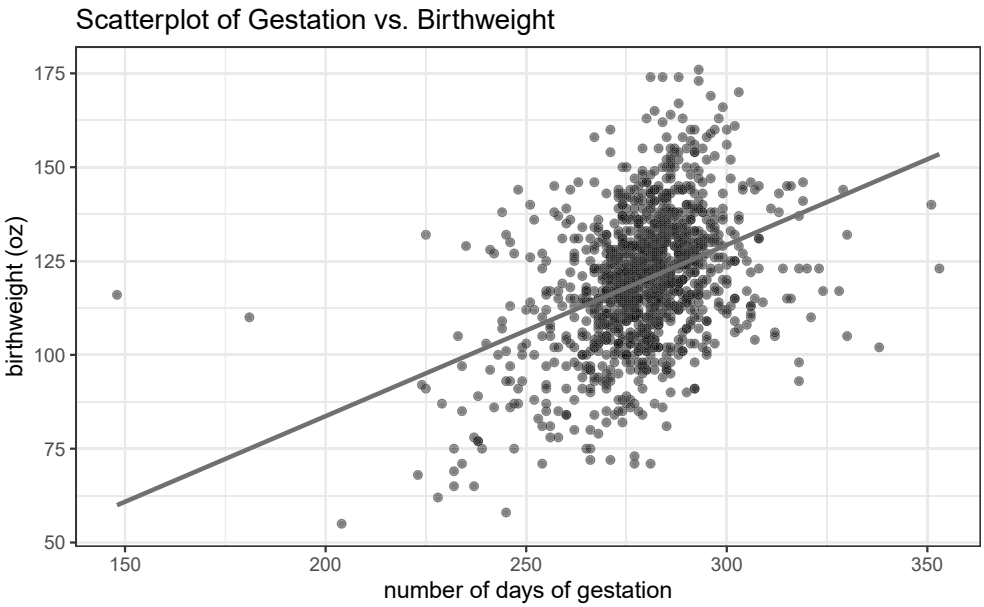
Example: Data was collected from 1236 births between 1960 and 1967 in the San Francisco East Bay area to better understand what variables contributed to child birth weight, as children with low birth weight often suffer from an array of complications later in life.

```
babies<-read.csv("data/babies.csv")
glimpse(babies)
#> Rows: 1,152
#> Columns: 7
#> $ bwt       <int> 120, 113, 128, 108, 136, 138, 132, 120, 143, 140, 144, 141, ~
#> $ gestation <int> 284, 282, 279, 282, 286, 244, 245, 289, 299, 351, 282, 279, ~
#> $ parity    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
#> $ age       <int> 27, 33, 28, 23, 25, 33, 23, 25, 30, 27, 32, 23, 30, 38, 25, ~
#> $ height    <int> 62, 64, 64, 67, 62, 62, 65, 62, 66, 68, 64, 63, 63, 63, 65, ~
#> $ weight    <int> 100, 135, 115, 125, 93, 178, 140, 125, 136, 120, 124, 128, 1~
#> $ smoke     <int> 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, ~
```

Here you see a glimpse of the data. The 1236 rows correspond to the sample size. The case variable is labeling each pregnancy 1 through 1236. Then 7 variables are recorded. Birthweight (bwt), length of gestation in days, parity is called an indicator variable telling us if the pregnancy was a first pregnancy (labeled as 0) or not (labeled as 1) were recorded about the child and pregnancy. The age, height, and weight were recorded for the mother giving birth, as was smoke, another indicator variable where 0 means the mother did not smoke during pregnancy, and 1 indicates that she did smoke while pregnant.

The following shows a scatterplot of length of gestation as a predictor of birthweight.

```
babies %>% # Data set pipes into...
ggplot(aes(x = gestation, y = bwt))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "number of days of gestation",  # Label x-axis
       y = "birthweight (oz)",  # Label y-axis
       title = "Scatterplot of Gestation vs. Birthweight") + # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE) + # Add regression line
    theme_bw()
```



Describe the scatterplot using the four characteristics.

The linear model output for this study is given below:

```
# Fit linear model: y ~ x
babiesLM <- lm(bwt ~ gestation, data=babies)
summary(babiesLM)$coefficients # Display coefficient summary
#>               Estimate Std. Error    t value     Pr(>|t|)
#> (Intercept) -7.5837816 8.64847207 -0.8768926 3.807281e-01
#> gestation    0.4562854 0.03091203 14.7607695 2.767227e-45
```

Write the least squares equation of the line.

Interpret the slope in context of the problem.

Interpret the y-intercept in context of the problem.

Predict the birthweight of a baby with 310 days gestation.

Calculate the residual for a baby with a birthweight of 151 ounces and at 310 days gestation.

Is this value (151, 310) above or below the line of regression? Did the line of regression overestimate or underestimate the birthweight?

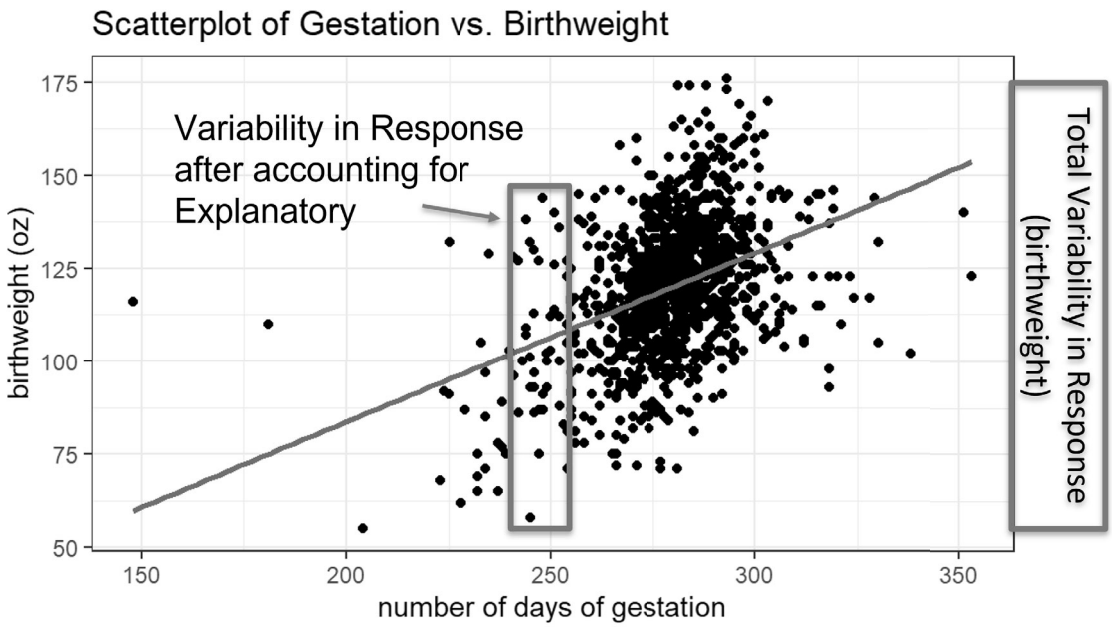
The following code finds the value of correlation between gestation and birthweight.

```
cor(bwt~gestation, data=babies, use="pairwise.complete.obs")
#> [1] 0.399103
```

This shows a _____, _____ relationship between gestation and birthweight.

The value for SST was found to be 382172.68. The value for SSE was found to be 321299.00.

Calculate and interpret the coefficient of determination between gestation and birthweight.



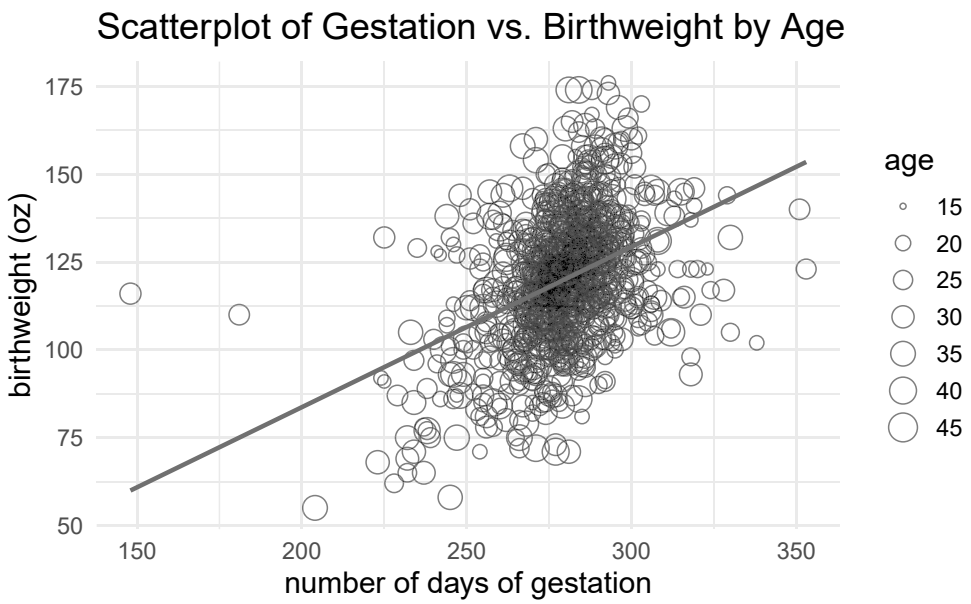
Multivariable plots

Aesthetics: visual property of the objects in your plot

- Position on the axes: groups for _____ variables, or a number line if the variable is _____
- Color or shape - to represent _____ variables
- Size - to represent _____ variables

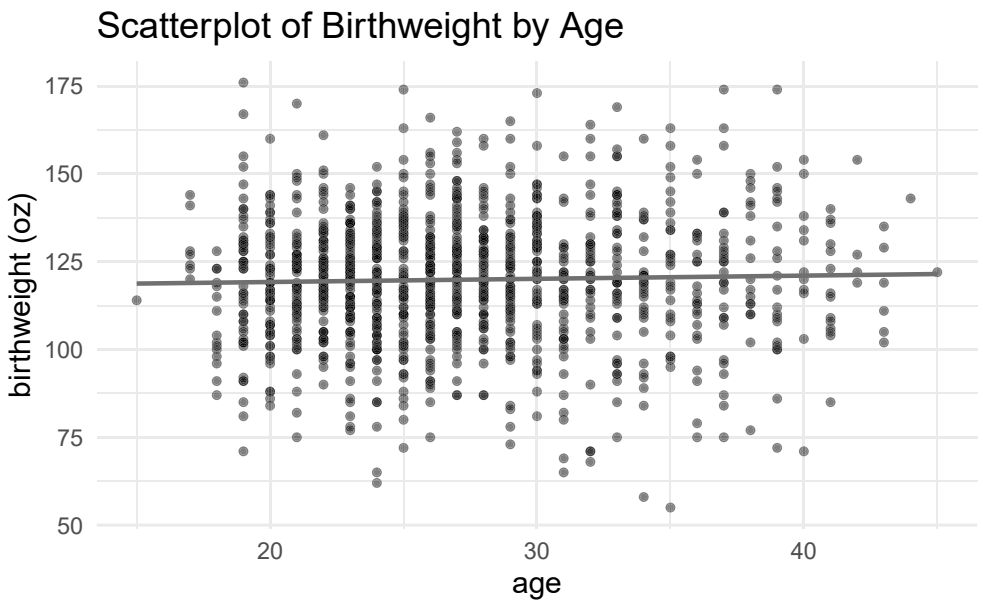
Adding the quantitative variable maternal age to the scatterplot between gestation and birthweight.

```
babies %>% # Data set pipes into...
ggplot(aes(x = gestation, y = bwt))+  # Specify variables
  geom_point(alpha=0.5, shape=1, aes(size=age)) +  # Add scatterplot of points
  labs(x = "number of days of gestation",  # Label x-axis
       y = "birthweight (oz)",  # Label y-axis
       title = "Scatterplot of Gestation vs. Birthweight by Age") + # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```



Is there an association between maternal age and birthweight?

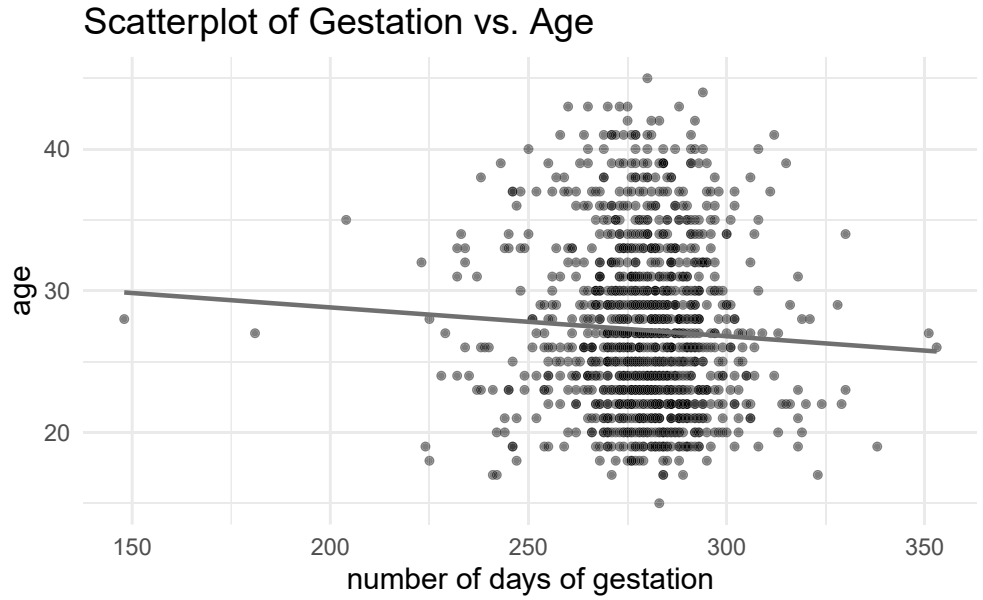
```
babies %>% # Data set pipes into...
ggplot(aes(x = age, y = bwt))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "age",  # Label x-axis
       y = "birthweight (oz)",  # Label y-axis
       title = "Scatterplot of Birthweight by Age") + # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```




```
cor(bwt~age, data=babies, use="pairwise.complete.obs")
#> [1] 0.02897539
```

Is there an association between maternal age and gestation?

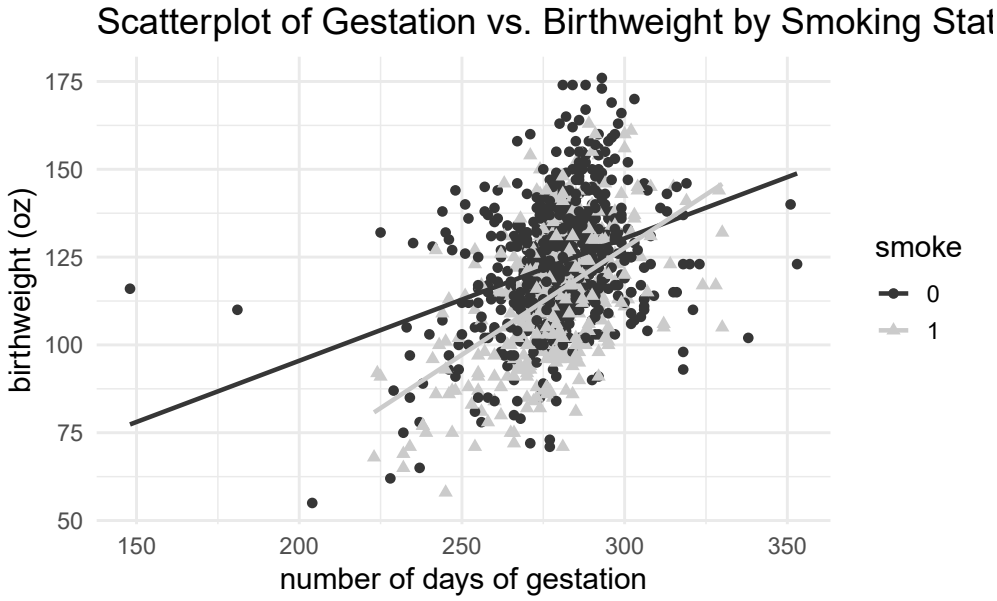
```
babies %>% # Data set pipes into...
ggplot(aes(x = gestation, y = age))+  # Specify variables
  geom_point(alpha=0.5) +  # Add scatterplot of points
  labs(x = "number of days of gestation",  # Label x-axis
       y = "age",  # Label y-axis
       title = "Scatterplot of Gestation vs. Age") + # Be sure to title your plots
  geom_smooth(method = "lm", se = FALSE)  # Add regression line
```



```
cor(age~gestation, data=babies, use="pairwise.complete.obs")
#> [1] -0.05560369
```

Let's add the categorical variable, whether a mother smoked, to the scatterplot between gestation and birthweight.

```
babies <- babies %>%
    mutate(smoke = factor(smoke)) %>%
    na.omit()

babies %>% # Data set pipes into...
    ggplot(aes(x = gestation, y = bwt, color = smoke))+  #Specify variables
    geom_point(aes(shape = smoke), size = 2) +  #Add scatterplot of points
    labs(x = "number of days of gestation",  #Label x-axis
         y = "birthweight (oz)",  #Label y-axis
         title = "Scatterplot of Gestation vs. Birthweight by Smoking Status") +
    #Be sure to title your plots
    geom_smooth(method = "lm", se = FALSE) + #Add regression line
    scale_color_grey()
```



Does the relationship between length of gestation and birthweight appear to depend upon maternal smoking status?

Is the variable smoking status a potential confounding variable?

Adding a categorical predictor:

- Look at the regression line for each level of the _____
- If the slopes are _____, the two predictor variables do not _____ to help explain the response
- If the slopes _____, there is an interaction between the categorical predictor and the relationship between the two quantitative variables.