
Welcome to Week 2

Understanding Data

Dataset vs Summary Table

	Country	BMI	Population	Life_expectancy_female	Life_expectancy	CHE GDP(%)	GDP
1	Afghanistan	22.9	34413603	64.7	63.2	10.1	1.990711e+
2	Albania	26.7	2880703	78.2	76.1	NA	1.138693e+
3	Algeria	25.5	39728025	77.2	76.2	7.0	1.659780e+
4	Andorra	27.0	78011	NA	NA	10.3	2.811489e+(
5	Angola	22.9	27884381	64.5	62.2	2.6	1.161940e+
6	Antigua and Barbuda	26.6	93566	77.4	75.0	4.7	1.336693e+(
7	Argentina	27.6	43131966	80.1	76.8	8.8	5.947490e+
8	Armenia	26.5	2925553	77.9	74.6	10.1	1.055334e+
9	Australia	27.4	23815995	84.5	82.6	9.3	1.351690e+
10	Austria	26.0	8642699	83.8	81.4	10.4	3.818180e+
11	Azerbaijan						

Dataset vs Summary Table

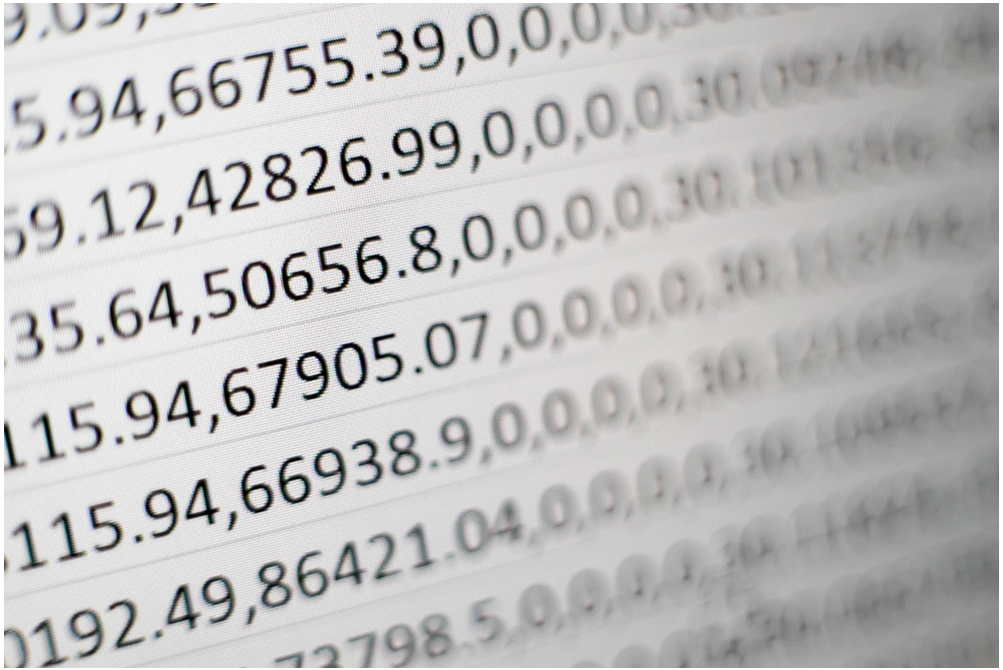
	Country	BMI	Population	Life_expectancy_female	Life_expectancy	CHE GDP(%)	GDP
1	Afghanistan	22.9	34413603	64.7	63.2	10.1	1.990711e+
2	Albania	26.7	2880703	78.2	76.1	NA	1.138693e+
3	Algeria	25.5	39728025	77.2	76.2	7.0	1.659780e+
4	Andorra	27.0	78011	NA	NA	10.3	2.811489e+(
5	Angola	22.9	27884381	64.5	62.2	2.6	1.161940e+
6	Antigua and Barbuda	26.6	93566	77.4	75.0	4.7	1.336693e+(
7	Argentina	27.6	43131966	80.1	76.8	8.8	5.947490e+
8	Armenia	26.5	2925553	77.9	74.6	10.1	1.055334e+
9	Australia	27.4	23815995	84.5	82.6	9.3	1.351690e+
10	Austria	26.0	8642699	83.8	81.4	10.4	3.818180e+
11	Azerbaijan						

Dataset vs Summary Table

	Africa (n=564)	Americas (n=270)	Asia (n=354)	Europe (n=315)	Oceania (n=19)	Overall (n=1704)
Life Expectancy						
Mean (SD)	48.8 (9.03)	64.6 (9.16)	60.2 (11.8)	71.9 (5.51)	75.2 (4.08)	59.5 (12.9)
Median [Min, Max]	47.6 [23.6, 76.4]	66.8 [37.6, 80.7]	62.0 [28.8, 82.2]	72.3 [43.6, 81.8]	76.3 [69.1, 81.2]	60.9 [23.6, 82.2]
Missing	64 (11.3%)	27 (10.0%)	40 (11.3%)	39 (12.4%)	2 (10.5%)	182 (10.7%)
Population						
Mean (SD)	10.2 (15.8)	23.8 (50.6)	72.5 (193)	16.2 (19.8)	9.66 (6.75)	28.6 (102)
Median [Min, Max]	4.73 [0.0600, 135]	5.97 [0.765, 301]	14.3 [0.120, 1230]	8.31 [0.148, 82.4]	9.20 [1.99, 20.4]	7.00 [0.0600, 1320]
Missing	63 (11.2%)	29 (10.7%)	42 (11.9%)	33 (10.5%)	1 (5.3%)	190 (11.2%)
Gdp Per Capita						
high	87 (15.4%)	176 (65.2%)	134 (37.9%)	273 (86.7%)	15 (78.9%)	764 (44.8%)
low	420 (74.5%)	64 (23.7%)	177 (50.0%)	19 (6.0%)	0 (0%)	767 (45.0%)
Missing	57 (10.1%)	30 (11.1%)	43 (12.1%)	23 (7.3%)	4 (21.1%)	173 (10.2%)

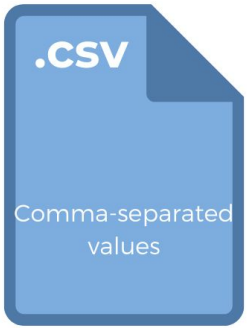
Variable Types

- Numeric
- Character
- Date
- Logical/Boolean



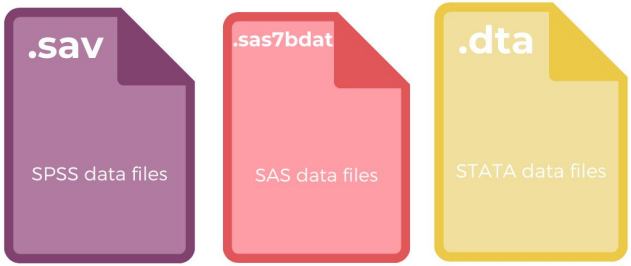
File Extensions

Comma-separated values (.csv) files



- Each row of data contains values for each column, separated by commas
- Can be opened using Microsoft Excel, Notepad, and other programs
- In R, we will use the **read_csv( )** function to import data

Other File Extensions



Import data using the tidyverse package **haven**

Advanced file types

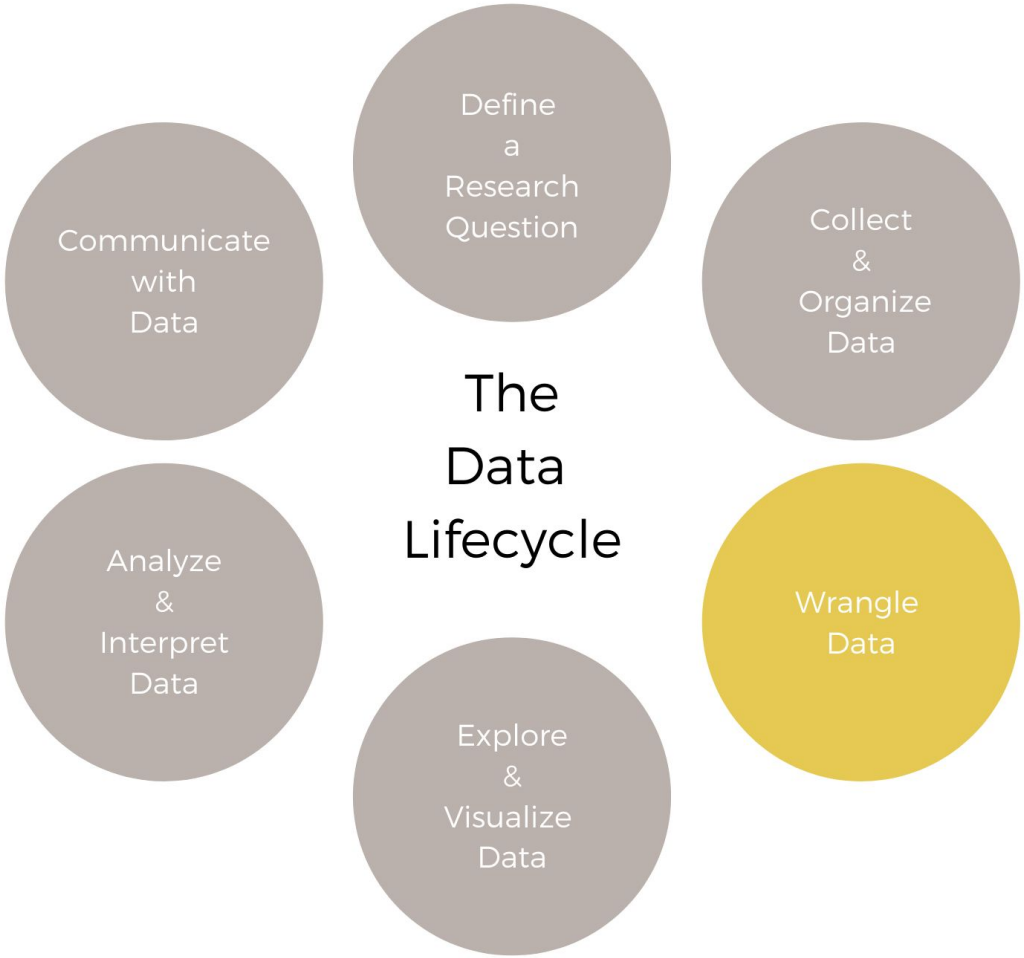

Importing Data into R

Getting to Know your Dataset


Data Wrangling

"No data is perfect, nor objective. And if we recognize this, we can start seeing data as the beginning of the conversation, not the end."

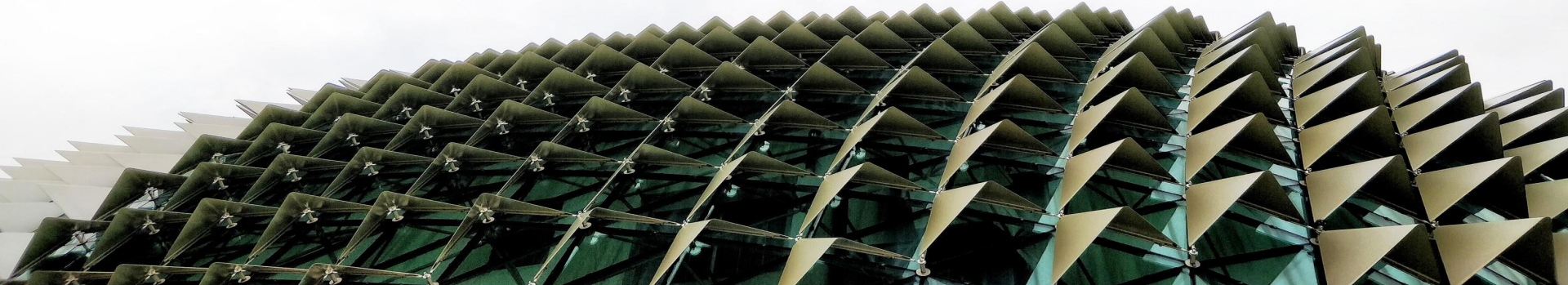
- Giorgia Lupi, co-author of Dear Data

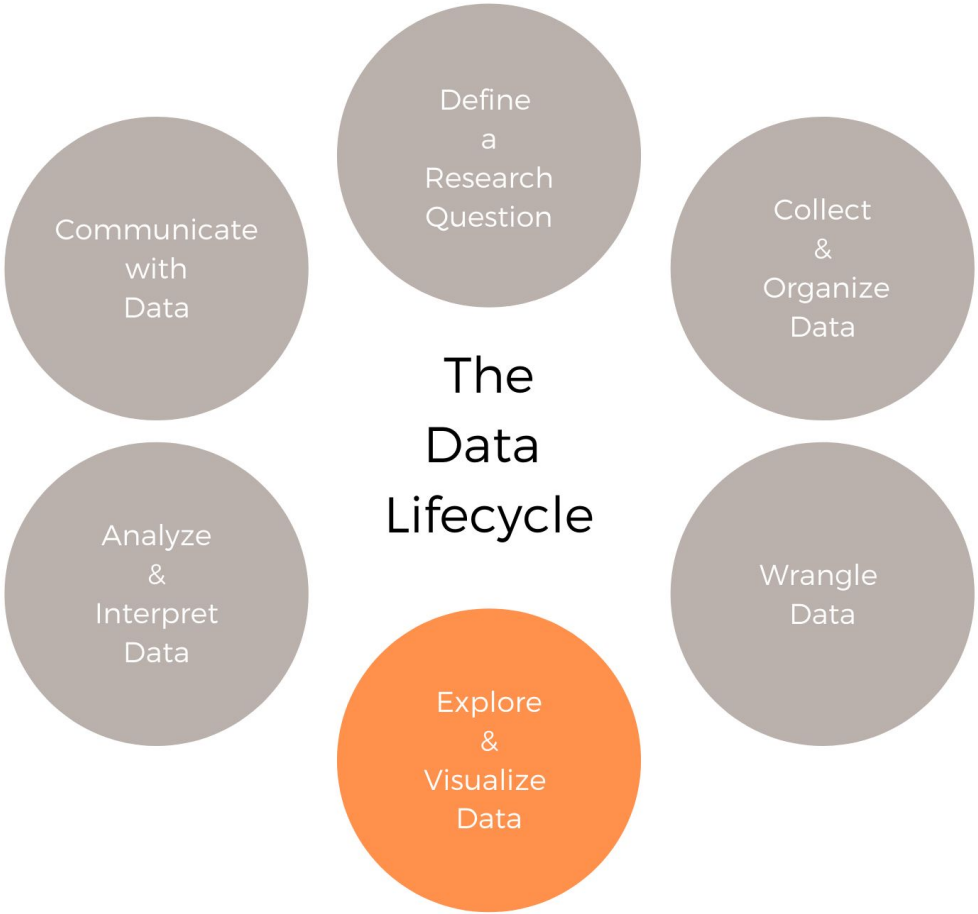

Define a Research Question
Collect & Organize Data
Wrangle Data
Explore & Visualize Data
Analyze & Interpret Data
Communicate with Data
The Data Lifecycle

How do we do it?

- Wrangling Variables
  - Using the names() function to access a list of variable names
    - Ordered by their appearance in the dataset
  - Changing the name of a variable
  - Changing the format of a variable
  - Creating a new variable
- Wrangling Data Values
  - Fixing typos in the dataset
  - Deleting missing data
  - Dropping anomalous data values

Explore
&
Visualize Data




Define a Research Question
Collect & Organize Data
Wrangle Data
Explore & Visualize Data
Analyze & Interpret Data
Communicate with Data
The Data Lifecycle

Explore & Visualize Data - Approach

- Examining variables **one at a time**
  - Starting with target variable
  - Numeric summaries
  - Plots
- Comparing **two** variables
  - Numeric summaries
  - Plots
- Comparing **more than two** variables
  - Pivoted numeric summaries
  - Plots with additional aesthetics (size, alpha, color, facet)

Looking Ahead

Image Credits:

Office Supply Photo by STIL on Unsplash
Blue Painted Wall Photo by Sharon McCutcheon on Unsplash
Numbers Photo by Mika Baumeister on Unsplash
Filing Cabinet Photo by Adam Birkett on Unsplash
Tablet on Wood Table Photo by Kelly Sikkema on Unsplash
Laptop 1 Photo by Adam Birkett on Unsplash
Laptop 2 Photo by Artem Sapegin on Unsplash
Apples in a Row Photo by twinsfisch on Unsplash
Spiked Singapore Building Photo by timj on Unsplash
Desert Floor Photo by Vincent Burkhead on Unsplash

Data Wrangling - Activity - College Scorecard Data

- Open and set up a new R script
  - Comments
  - Load packages
  - Save the script (and save as you go)
- Import dataset
  - Using read_csv
- Getting to know the dataset
  - About the data
    - Source/Description
    - Data Dictionary
  - Ask questions like:
    - How many numeric variables, categorical , etc.
    - How many rows and columns

Data Wrangling - Activity - College Scorecard Data

- Wrangling variable names
  - Using the names() function to access a list of variable names
    - Ordered by their appearance in the dataset
  - Changing the name of a variable
    - Using the rename() function
  - Changing the name/appearance of multiple variables
    - Removing special characters
    - Changing the letter case
- Wrangling data values
  - Changing the format of a variable
    - Making dates appear as dates (lubridate function)
    - Converting strings to numeric variables
  - Creating a new variable and Transforming existing variables
    - mutate
  - Fixing typos in the dataset
  - Dealing with missing data

Data Wrangling - Activity - College Scorecard Data

- Wrapping Up
  - Writing a dataset back out to csv
    - write_csv(df, “new_dataset.csv”)
    - Naming convention (use same descriptive terms as original dataset, and add on _cleaned)
    - You may name yours however you like, but remember how you named it when it’s time to re-import.
    - Make sure you use the right filepath (navigate to the right folder)
  - Save your script!
  - In the next activity,
    - Instead of reproducing all the steps we did in this activity,
    - We will read in our wrangled dataset and resume.

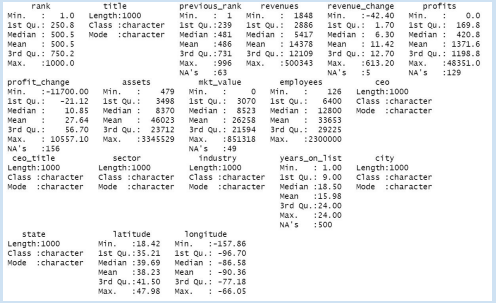
Exploring Data - Activity - College Scorecard Data

- Recall the research question
    - Use that question to develop other closed ended questions
- Examining variables one at a time
    - Numeric summaries
    - Plots
    - Start with target variable (cost of attendance)
        - summary(df$var_name)
        - Df %>% ggplot(aes(x= var_name)) + geom_density()
    - Summarize the st_abbr variable
        - “How many colleges per state?” df %>% count(st_abbr)

Exploring Data - Activity - College Scorecard Data

- Comparing two variables
  - Numeric summaries
  - Plots
  - Cost of attendance by state (numeric vs. character)
    - df %>% group_by(st_abbr) %>% summarize(avg_cost = mean(cost))
    - Take same code as above, and pipe on a ggplot(aes(x = st_abbr, y= avg_cost)) + geom_col()
  - Cost of attendance compared to avg_sat (numeric vs numeric)
    - Scatter plot only
    - Is there a linear relationship?
  - Tell learners to try comparing cost of attendance to another variable
- Comparing more than two variables
  - Cost of attendance, state, and avg_sat
  - Pivoted numeric summaries
    - df %>% group_by(st_abbr) %>% summarize(avg_cost = mean(cost), avg_avg_sat = mean(avg_sat))
  - Plots with additional aesthetics (size, alpha, color, facet)

Understanding Data - Quiz



```
      rank            title           previous_rank      revenues      revenue_change      profits      
 Min.   :   1.0   Length:1000        Min.   :   1   Min.   :  1848   Min.   :-42.40   Min.   :    0.0  
 1st Qu.: 250.8   Class :character   1st Qu.:239   1st Qu.:  2886   1st Qu.:  1.70   1st Qu.:  169.8  
 Median : 500.5   Mode  :character   Median :481   Median :  5417   Median :  6.30   Median :  420.8  
 Mean   : 500.5                      Mean   :486   Mean   : 14378   Mean   : 11.42   Mean   : 1371.6  
 3rd Qu.: 750.2                      3rd Qu.:731   3rd Qu.: 12109   3rd Qu.: 12.70   3rd Qu.: 1198.8  
 Max.   :1000.0                      Max.   :996   Max.   :500343   Max.   :613.20   Max.   :48351.0  
                                     NA's   :63                     NA's   :5        NA's   :129      
 profit_change          assets          mkt_value         employees          ceo           
 Min.   :-11700.00   Min.   :   479   Min.   :    0   Min.   :   126   Length:1000       
 1st Qu.:   -21.12   1st Qu.:  3498   1st Qu.:  3070   1st Qu.:  6400   Class :character  
 Median :    10.85   Median :  8370   Median :  8523   Median : 12800   Mode  :character  
 Mean   :    27.64   Mean   : 46023   Mean   : 26258   Mean   : 33653                     
 3rd Qu.:    56.70   3rd Qu.: 23712   3rd Qu.: 21594   3rd Qu.: 29225                     
 Max.   : 10557.10   Max.   :3345529   Max.   :851318   Max.   :2300000                    
 NA's   :156                           NA's   :49                                          
  ceo_title             sector           industry         years_on_list        city          
 Length:1000        Length:1000        Length:1000        Min.   : 1.00   Length:1000       
 Class :character   Class :character   Class :character   1st Qu.: 9.00   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Median :18.50   Mode  :character  
                                                          Mean   :15.98                     
                                                          3rd Qu.:24.00                     
                                                          Max.   :24.00                     
                                                          NA's   :500                       
    state              latitude       longitude       
 Length:1000        Min.   :18.42   Min.   :-157.86  
 Class :character   1st Qu.:35.21   1st Qu.: -96.70  
 Mode  :character   Median :39.69   Median : -86.58  
                    Mean   :38.23   Mean   : -90.36  
                    3rd Qu.:41.50   3rd Qu.: -77.18  
                    Max.   :47.98   Max.   : -66.05  
```

- Question 1

  Is this a dataset or a summary table?

  (Should be set up either as a dropdown, or single choice radio button question)

  (Use fortune_summary.png from Dropbox images folder)

  Feedback if “dataset” is chosen: Remember, a dataset is an organized collection of data values structured across rows and columns.

  Answer(“summary table”): This is a summary table. It’s not an organized collection of data values across rows and columns.

Understanding Data - Quiz

- Question 2

  Based on the descriptions of the fictional variables listed below, which of the following represent **numeric** variables?

shirt_color: indicates the color of the shirt the person was wearing when interviewed; can take on values like "red", "gray", "black", "white", and "blue"
item_price: indicates the price of an item purchased, in US dollars; can take on any value between 0 and 200.00
wearing_hat: indicates whether or not the person was wearing a hat when interviewed; only takes on the values "TRUE" or "FALSE"
age_group: subject's age, classified into groups (e.g., "Under 18", "18-29", "30-49", "50-69", "70 and Older")
num_students: indicates the total number of students enrolled in a high school (e.g., 348, 2712, 1949, 514, 3192)

Correct responses: (item_price, num_students)
[explanation]
Numeric variables can take on data values across a continuous range of numbers.
[explanation]

Wrangle Data - Quiz

- <Single choice question>
- Which of the following statements regarding data wrangling is **incorrect**?

It’s the easiest step in the data lifecycle. (CORRECT ANSWER)

It allows you to quickly identify data values that are missing or don’t make sense.

It can speed up the process of getting started with subsequent steps of the data lifecycle.

<EXPLANATION>There is no “easiest step” of the data lifecycle.

Wrangle Data - Quiz

- Which of the following special characters do we try to avoid when defining variables, datasets, and other R objects? Check any that apply.

Percent signs (%) CORRECT ANSWER

Underscores (_)

Parentheses (( )) CORRECT ANSWER

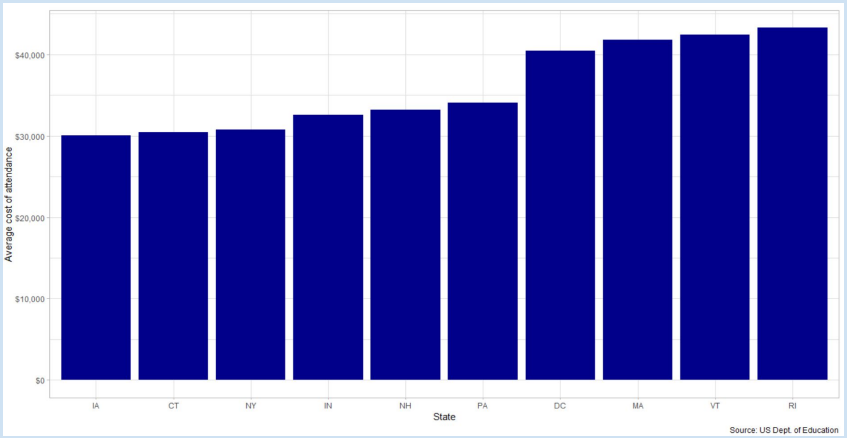
Dollar signs ($) CORRECT ANSWER

[explanation]

Underscores are okay to include in variable names, but we try to avoid using any of the other characters.

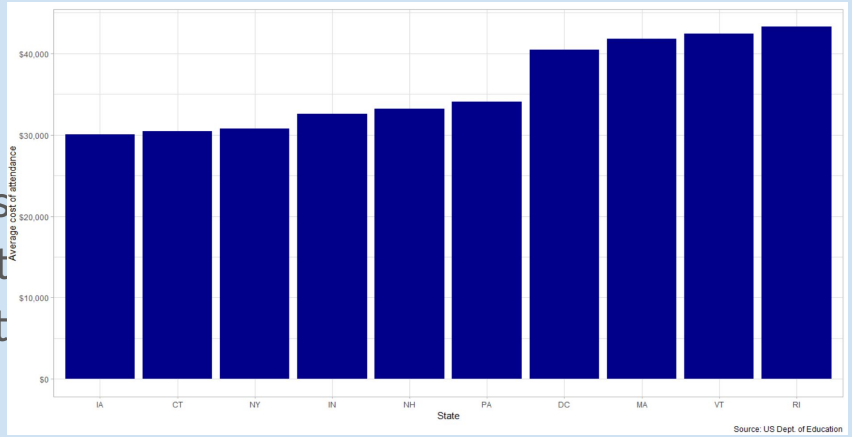
Explore and Visualize Data - Quiz

- <Use cs_avgcost_top10state_bar.png from Dropbox > images folder>
- <single choice>
- The graph below displays the ten US states with the highest average cost of attendance. Based on this graph, which of the following is closest to the average cost of attendance of institutions in Pennsylvania (PA)?
    - $40,000
    - $34,000 (CORRECT ANSWER)
    - $29,000
    - $15,000



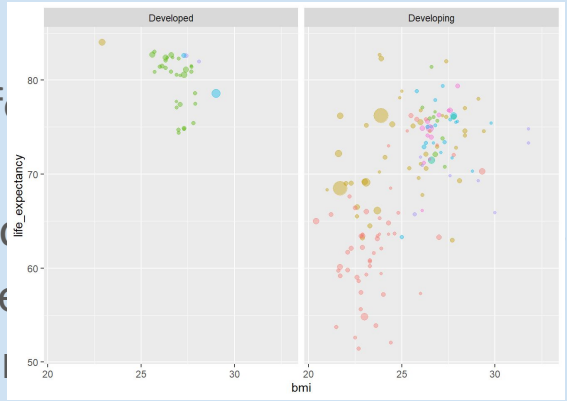
Explore and Visualize Data - Quiz

- <Use cs_avgcost_top10state_bar.png from
- <single choice>
- The graph below displays the ten US states
  attendance. Based on this graph, which of t
  difference between the average cost of att
  (RI) and Iowa (IA)?
  - $20,000
  - $17,000
  - $13,000 (CORRECT ANSWER)
  - $10,000



Explanation: The average cost of attendance for Rhode Island institutions lies between $40,000 and $45,000, while the average cost of attendance for Iowa institutions is about $30,000. So it makes sense that the difference between these two might be closest to

Explore and Visualize Data - Quiz



- <Use who_facet_multvars.png from Dropbox > images f
- <single choice>
- The plot below compares average life expectancy an
  around the world. The size of each point varies by the
  population (larger sized points indicate higher countr
  countries classified as Developed are plotted separately from countries classified as Developing. Which of the following statements regarding the plot is **true** ?
  - On average, life expectancy in developed countries was **lower** than in developing countries.
  - The countries with the **highest** populations had an average BMI greater than 25.
  - All developed countries had an average life expectancy above 60 years. (CORRECT ANSWER)

End of Week 2 Quiz

- Is this a dataset or summary table?
- Which function will print numeric summaries for all non-character variables in a dataset?
- Which data viz is most appropriate for comparing variables X and Y?
- Interpreting visualizations

End of Week 2 Quiz

- Is this a dataset or a summary table? <Should be formatted as either a dropdown or radio button single choice question>

	Africa (n=564)	Americas (n=270)	Asia (n=354)	Europe (n=315)	Oceania (n=19)	Overall (n=1704)
Life Expectancy						
Mean (SD)	48.8 (9.03)	64.6 (9.16)	60.2 (11.8)	71.9 (5.51)	75.2 (4.08)	59.5 (12.9)
Median [Min, Max]	47.6 [23.6, 76.4]	66.8 [37.6, 80.7]	62.0 [28.8, 82.2]	72.3 [43.6, 81.8]	76.3 [69.1, 81.2]	60.9 [23.6, 82.2]
Missing	64 (11.3%)	27 (10.0%)	40 (11.3%)	39 (12.4%)	2 (10.5%)	182 (10.7%)
Population						
Mean (SD)	10.2 (15.8)	23.8 (50.6)	72.5 (193)	16.2 (19.8)	9.66 (6.75)	28.6 (102)
Median [Min, Max]	4.73 [0.0600, 135]	5.97 [0.765, 301]	14.3 [0.120, 1230]	8.31 [0.148, 82.4]	9.20 [1.99, 20.4]	7.00 [0.0600, 1320]
Missing	63 (11.2%)	29 (10.7%)	42 (11.9%)	33 (10.5%)	1 (5.3%)	190 (11.2%)
Gdp Per Capita						
high	87 (15.4%)	176 (65.2%)	134 (37.9%)	273 (86.7%)	15 (78.9%)	764 (44.8%)
low	420 (74.5%)	64 (23.7%)	177 (50.0%)	19 (6.0%)	0 (0%)	767 (45.0%)
Missing	57 (10.1%)	30 (11.1%)	43 (12.1%)	23 (7.3%)	4 (21.1%)	173 (10.2%)

Answer: summary table

End of Week 2 Quiz

```
  Country               Year          Status           Life_expectancy Adult_Mortality
Length:2938         Min.   :2000   Length:2938        Min.   :36.30   Min.   :  1.0
Class :character    1st Qu.:2004   Class :character   1st Qu.:63.10   1st Qu.: 74.0
Mode  :character    Median :2008   Mode  :character   Median :72.10   Median :144.0
                    Mean   :2008                      Mean   :69.22   Mean   :164.8
                    3rd Qu.:2012                      3rd Qu.:75.70   3rd Qu.:228.0
                    Max.   :2015                      Max.   :89.00   Max.   :723.0
                                                      NA's   :10      NA's   :10
infant_deaths       Alcohol        percentage_expenditure  Hepatitis_B       Measles
Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000      Min.   : 1.00   Min.   :     0.0
1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685      1st Qu.:77.00   1st Qu.:     0.0
Median :   3.0   Median : 3.7550   Median :   64.913      Median :92.00   Median :    17.0
Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251      Mean   :80.94   Mean   :  2419.6
3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534      3rd Qu.:97.00   3rd Qu.:   360.2
Max.   :1800.0   Max.   :17.8700   Max.   :19479.912      Max.   :99.00   Max.   :212183.0
                 NA's   :194                              NA's   :553
```

Based on the summary above, how many values are missing for the variable `Hepatitis_B`?
Answer: 553

End of Week 2 Quiz

```
  Country               Year          Status           Life_expectancy Adult_Mortality
Length:2938         Min.   :2000   Length:2938        Min.   :36.30   Min.   :  1.0
Class :character    1st Qu.:2004   Class :character   1st Qu.:63.10   1st Qu.: 74.0
Mode  :character    Median :2008   Mode  :character   Median :72.10   Median :144.0
                    Mean   :2008                      Mean   :69.22   Mean   :164.8
                    3rd Qu.:2012                      3rd Qu.:75.70   3rd Qu.:228.0
                    Max.   :2015                      Max.   :89.00   Max.   :723.0
                                                      NA's   :10      NA's   :10
infant_deaths       Alcohol        percentage_expenditure  Hepatitis_B       Measles
Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000      Min.   : 1.00   Min.   :     0.0
1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685      1st Qu.:77.00   1st Qu.:     0.0
Median :   3.0   Median : 3.7550   Median :   64.913      Median :92.00   Median :    17.0
Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251      Mean   :80.94   Mean   :  2419.6
3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534      3rd Qu.:97.00   3rd Qu.:   360.2
Max.   :1800.0   Max.   :17.8700   Max.   :19479.912      Max.   :99.00   Max.   :212183.0
                 NA's   :194                              NA's   :553
```

Based on the summary above, what was the median life expectancy across the entire dataset?
Answer: 72.10 years

End of Week 2 Quiz

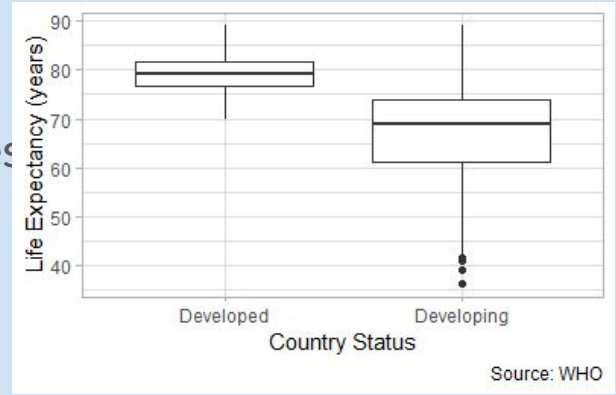
```
  Country               Year          Status           Life_expectancy Adult_Mortality
Length:2938         Min.   :2000   Length:2938        Min.   :36.30   Min.   :  1.0
Class :character    1st Qu.:2004   Class :character   1st Qu.:63.10   1st Qu.: 74.0
Mode  :character    Median :2008   Mode  :character   Median :72.10   Median :144.0
                    Mean   :2008                      Mean   :69.22   Mean   :164.8
                    3rd Qu.:2012                      3rd Qu.:75.70   3rd Qu.:228.0
                    Max.   :2015                      Max.   :89.00   Max.   :723.0
                                                      NA's   :10      NA's   :10
infant_deaths       Alcohol        percentage_expenditure  Hepatitis_B       Measles
Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000      Min.   : 1.00   Min.   :     0.0
1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685      1st Qu.:77.00   1st Qu.:     0.0
Median :   3.0   Median : 3.7550   Median :   64.913      Median :92.00   Median :    17.0
Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251      Mean   :80.94   Mean   :  2419.6
3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534      3rd Qu.:97.00   3rd Qu.:   360.2
Max.   :1800.0   Max.   :17.8700   Max.   :19479.912      Max.   :99.00   Max.   :212183.0
                 NA's   :194                              NA's   :553
```

Based on the summary above, is there data for the year 1999?
Answer: No. The minimum value for year is 2000, so we don't have data for the year 1999.

End of Week 2 Quiz

Recall the variables `life_expectancy` and `status` from the WHO Life Expectancy dataset. Which type of plot(s) would be most appropriate for comparing these two variables?

- Tree map (feedback: a tree map is best suited for summarizing two categorical variables)
- Box plots
- Histogram (feedback: a histogram is bes eric variable)

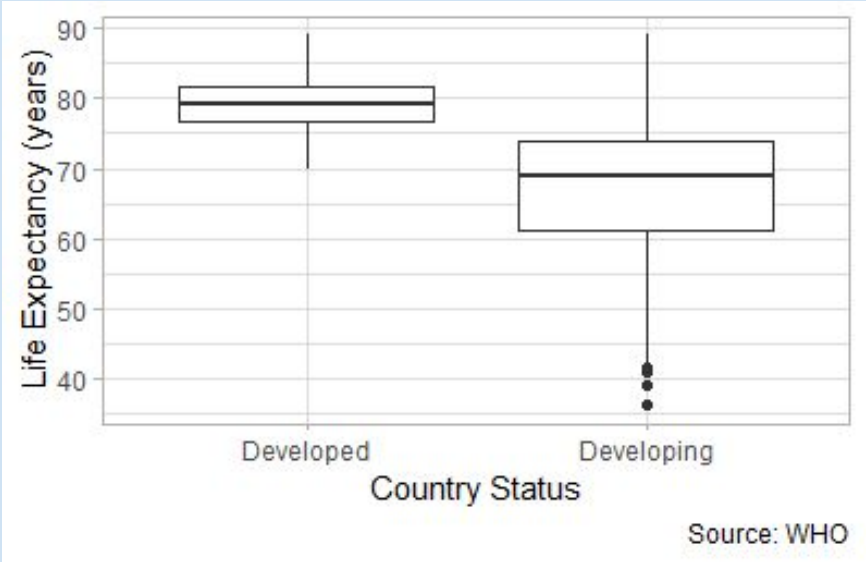


Answer: Box plots are best suited for comparing a numeric variable by a categorical variable

End of Week 2 Quiz

Based on the plot below, what was the median life expectancy among developing countries?

A. 78 years
B. 74 years
C. 69 years
D. 61 years



Answer: The median life expectancy among developing countries was approximately 69 years. This value is represented by the solid black line in the middle of the box for

Example: WHO Life Expectancy

- [Instructions for learners: download the dataset from edX and save it to a folder on your computer related to this course. If using RStudio cloud, upload the dataset to your project folder.]
- First, make sure you know where the dataset is saved. You will need to refer to the correct filepath in your script
- Screenshot/screencast of R code to import dataset using script **import_who_data.R**
    - First load tidyverse package
    - Use read_csv to import and then assign to a name (e.g., 'who')
    - Aim for a name that is short but informative. Try not to use names like 'data' - too generic.

Course Handout - Data Wrangling Resources

- R Studio data import cheatsheet
- (importing data) Tidyverse readr reference page https://readr.tidyverse.org/reference/index.html
- (wrangling/exploring) Tidyverse dplyr reference page https://dplyr.tidyverse.org/reference/index.html

Course Handout - R Data Visualization Resources

- From-data-to-viz
- R Graph Gallery
- Tidyverse ggplot reference page: https://ggplot2.tidyverse.org/reference/index.html
- https://socviz.co/
- Leaflet : https://rstudio.github.io/leaflet/
- Storytelling with Data (book, blog, and podcast) http://www.storytellingwithdata.com/
- Alberto Cairo books: “the truthful art” and “the functional art”

In the handout, consider categorizing by topic

- R code examples
- Theory of data visualization
-