Name: katherine Alas Date:1/29/2020

**Lab 1B: Get the Picture?**

**Directions: Record your responses to the lab questions in the space provided. Use Google Colab to help you solve the answers.**

* In this lab, we'll learn about different types of variables.
  + Such as those that are measured by numbers and others that have values that are categories.
* We'll also look at ways to visualize these different types of data using plots (A word data scientists use interchangeably with the word graph).

**Variable Types**

* Numerical variables have values that are measured in units.
* Categorical Variables have values that describe or categorize our observations.
* View your cdc data and find the columns for height and gender
  + **Is height a numerical or categorical variable? Why?**

Height is a numerical variable because is measured in units

* + **Is gender a numerical or categorical variable? Why?**

Gender is a categorical variable because is a value that categorizes observations**.**

* + **List either the different categories or what you think the measuring units are for height and gender.**

The measuring units for height are units such as feet or meters, and gender is categories could be male, female or other.

**Which is which?**

• Run the code you used in the previous lab to display the names of your cdc data's variables (**len(cdc.columns)**). Use the code's output to help you complete the following:

**– Write down 3 variables that you think are categorical variables and why.**

Race, bully in school, sexuality is categorical variables because they are not measures in united and are values that categorize an aspect of the student.

**– Write down 3 variables that you think are numerical variables and why.**

Age, height, drink juice is numerical because the values are measured by units.

**Data Structures**

* One way to get a good summary of your data is to look at the data's structure.
  + One way would be to run the following in the console:
    - cdc.info()
* Look at the structure of your CDC data and answer:
  + **List all the types of info the .info() function outputs**

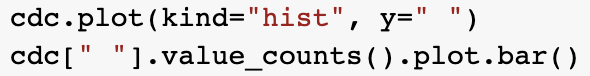
Object, float64 and inter64

* + **Were you able to correctly guess which variables were categorical and numeric? Which ones did you mislabel?**

I was able to guess some but i didn't guess the age wasn't a numerical variable I thought because the age was in numbers it would be numerical.

**Visualizing data**

* Visualizing data is a really helpful way to learn about our variables.
  + Making a picture of the distribution of a variable is a good way to begin visualizing data.
  + Remember: A distribution gives us the values of the variable and tells us how many of these values we have in our data set.
* Choose one numeric and one categorical variable from the data and create both a bar graph and a histogram for each variable.
* Use your Lab from yesterday to help you make a bargraph or histogram.

****

* **Which function, either bargraph or a histogram is better at visualizing categorical variables? Which is better at visualizing numerical variables?**

**cdc[''].value\_counts().plot.bar() is better for visualizing categorical data and cdc.plot(kind="hist", y="")is better for visualizing numerical variables and this also gave me an error when I tried using it for a categorical variable .**

**How often do people text & drive?**

* Make a graph that shows how often people in our data texted while driving.
  + **What does the y-axis represent?**

The Y-axis represents the number of students.

* + **What does the x-axis tell us?**

The X-axis tells us the number of days the students drive and text.

* + **Would you say that most people never texted while driving? What does the word most mean?**

Yes, most students never texted while driving . Most means the majority of students meaning the number of students who didn’t text and drive than students who did.

* + **Approximately what percent of the people texted while driving for 20 or more days? (Hint: There are 15,624 students in our data.**

0.0323%

**Does texting and driving differ by gender?**

**Fill in the blanks with the correct variables to create a side-by-side bargraph:**

**sns.countplot(y="drive\_text", hue="gender", data= cdc)**

* **Write a sentence explaining how boys and girls differ in their use texting while driving.**

Boys don't text while driving as often as girls do.

* **Would you say that most girls never text and drive? Would you say that most boys never text and drive?**

I wouldn't say either girls or boys text and drive because there is still some percentage that does.

* **How did including the group’s argument in your code change the graph?**

It shows both arguments in one graph and it changes the color for the gender to make the visualization easier to understand.

**Do males/females have similar heights?**

* **Groups use color to differentiate between groups.**

**Type the following into Google Colab Cell**

**sns.pairplot(cdc, hue="gender")**

* **Do you think males & females have similar heights? Use the plot you create to justify your answer.**

I think that males have a higher height than females base on the graph