Name: Date:

**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?**
  + **Is gender a numerical or categorical variable? Why?**
  + **List either the different categories or what you think the measuring units are for height and gender.**

**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.**

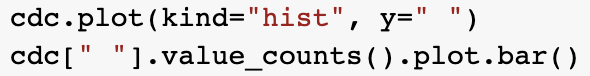
**– Write down 3 variables that you think are numerical variables and why.**

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

**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 bargraph 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?**

**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?**
  + **What does the x-axis tell us?**
  + **Would you say that most people never texted while driving? What does the word most mean?**
  + **Approximately what percent of the people texted while driving for 20 or more days? (Hint: There are 15,624 students in our data.**

**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="\_\_\_\_\_", hue="gender", data=\_\_\_\_)**

* **Write a sentence explaining how boys and girls differ in their use texting while driving.**
* **Would you say that most girls never text and drive? Would you say that most boys never text and drive?**
* **How did including the group’s argument in your code change the graph?**

**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.**
* **Just like we did for the histogram, is it possible to create a split bargraph? Try to create a bargraph of drive\_text that's split by gender to find out.**