COGS9: Introduction to Data Science
Assignment #1: Data Visualization
Submit through Gradescope individually

First Name	Last Name	PID

Part I: **The Largest Vocabulary in Hip Hop** (contains adult language):

For the first part of this assignment, you are to explore how to visualize the results from an analysis that uses primarily textual data. You will read and critically engage with a data journalism project. Answer the questions concisely (please no more than a few sentences per point, concision is key!).

Navigate to The Pudding's piece "The Largest Vocabulary in Hip Hop":
http://poly-graph.co/vocabulary.html
And check out this updated version too: https://pudding.cool/projects/vocabulary/index.html

a. **(2 pts)** Choose 10 of the artists visualized on the page linked above. You can choose your favorite artists or pick randomly along the spectrum.

Then, in a Google Sheets, make three columns of data: in the first column, record the **name of the artist/group**, in the second, record each of your 10 artists' **number of unique words used** (from above website). In the third column, find (from Google and/or Wikipedia) **the year of each artists' first studio album release** (e.g., 1995).

REPLACE THIS TEXT WITH A COPY/SCREENSHOT OF YOUR TABLE FROM GOOGLE SHEETS.

Create a scatterplot of the **number of unique words** against the **year of their album release** (highlight the data > insert chart > select scatter chart.). Paste your table in your submission and attach the resulting plot.

REPLACE THIS TEXT WITH A COPY/IMG OF YOUR SCATTERPLOT.

Is the plot of points slanted up, down, or flat? What does this tell us about the number of unique words used vs. the year of each artists' first studio album release?

REPLACE THIS TEXT WITH THE INTERPRETATION OF YOUR SCATTERPLOT.

b. **(2 pt)** Repeat the above but plot their **number of unique words** against the **age at which they first released their album**. (If it's a group, take the average/mean across all group members.) Paste your new table and graph, and then describe your results.

REPLACE THIS TEXT WITH A COPY OF YOUR TABLE FROM GOOGLE SHEETS, YOUR SCATTERPLOT AND YOUR INTERPRETATION.

c. **(1 pt)** How do you think hip hop lyrics compare to rock and country lyrics? Find a few popular country and rock song lyrics (e.g. using Holler's lists - https://holler.country/playlists/the-most-popular-country-songs/) and think about the differences compared to hip hop lyrics. E.g. word choices, sentence lengths, and so on. *Specific examples to highlight/support your points are best here*.

REPLACE THIS TEXT WITH YOUR ANSWER.

d. **(1 pt)** When it comes to unique words, what do the data suggest happened from early on around approx the 1990s to approx late 2010s (from the sample set you selected)?

REPLACE THIS TEXT WITH YOUR ANSWER.

e. **(1 pt)** Suggest 3 additional variables that the author could have extracted and analyzed from the lyrics data, other than the # of unique words used.

REPLACE THIS TEXT WITH YOUR ANSWER.

f. **(1 pts)** Suppose two lyrics have the same number of words and the same type of words (i.e. vocabulary). In what aspect could the lyrics still differ? Explain with a single example (not of lyrics necessarily, but simple 1-2 lines of text).

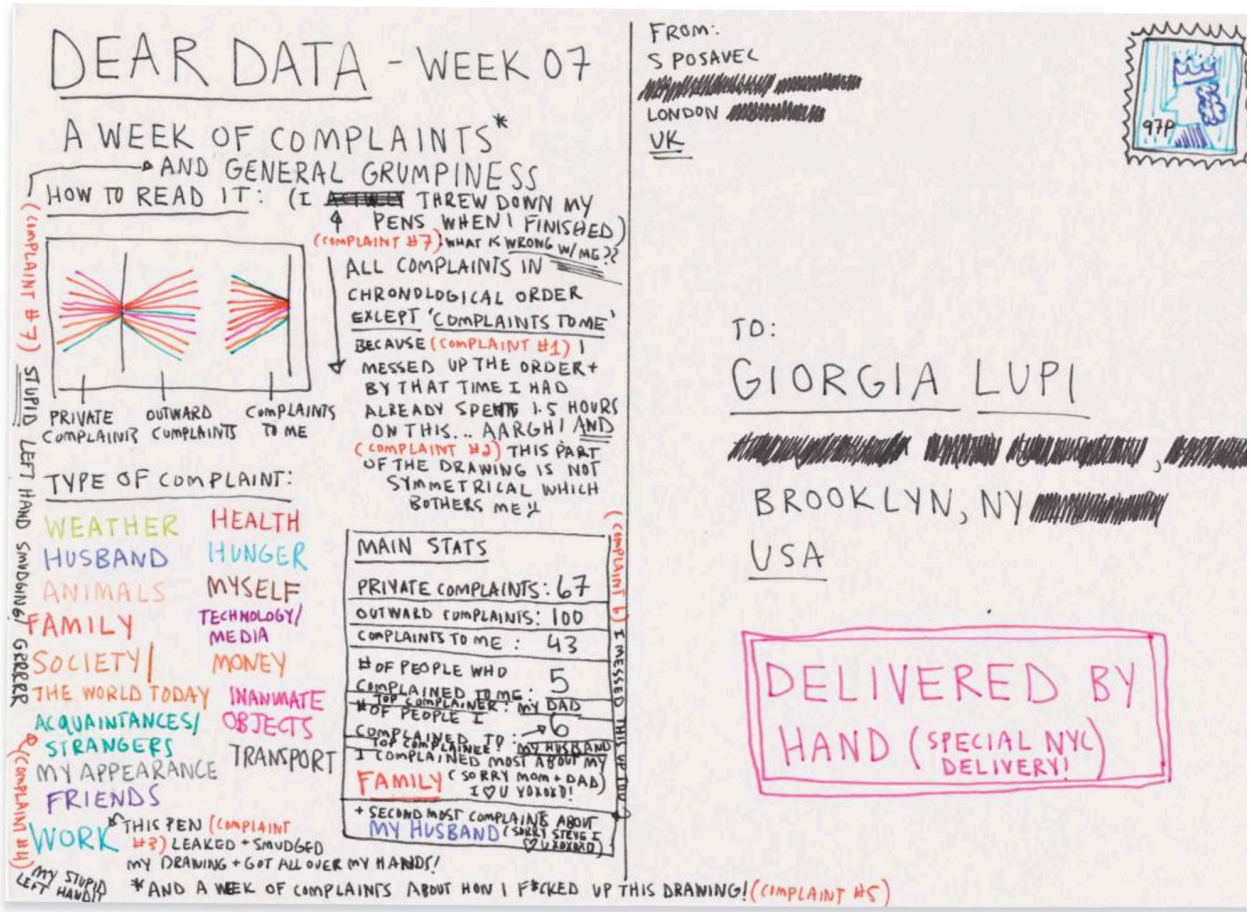
REPLACE THIS TEXT WITH YOUR ANSWER.

Part II: **Your Visualization**

The goal of this part of the assignment is to collect data about something in your daily life or something that you're interested in and effectively visualize that information.

You need at least 10 data points (but can have more). For example, if you collect data about something you do once a day, and you only have 5 days left to complete this assignment, you'll need to collect data from at least 2 people to get 10 data points.

You are free to collect data on any classroom-appropriate topic, but we encourage you to be creative. You could plot the time you brush your teeth every day (kinda boring.) or you could track all the compliments you overhear others giving over the course of the week (more interesting!).
If you need some inspiration, consider the topics visualized in the Dear Data (http://www.dear-data.com/theproject) project. For example, here is a page of visualization from *Dear Data* that illustrates the number and types of complaints recorded by the author during a particular week,



Your visualization will likely not *look* like a Dear Data visualization, but it will likely help inspire you on the type of data you may want to collect.

Data (4 pts):

Include a table with the data you've collected here or a link to the data in Google Sheets. (If you include a link to Google Sheets because your data are too large to be pasted here, be sure the link is viewable by others.) Be sure that these data are stored in a tidy data format and follow the best practices for information stored in tables/spreadsheets.

REPLACE THIS TEXT WITH YOUR TABLE OR A LINK TO THE DATA IN GOOGLE SHEETS.

Data Visualization (4 pts):

Generate an effective explanatory visualization of the data you collected. What you use to create this visualization is up to you. You could draw it on a piece of paper, create it on a drawing app, generate it in Excel/Google Sheets, or use a programming language (R, Python, JavaScript, etc.) to generate the visualization - it's totally up to you. This visualization should be appropriate given the type of data you've collected and the message you want to convey. It should follow the best practices for visualization discussed in lectures.

REPLACE THIS TEXT WITH YOUR VISUALIZATION.

Visualization Interpretation (2 pt):

Explain in a few sentences what you want the viewer to take away from your visualization.

REPLACE THIS TEXT WITH YOUR INTERPRETATION.

Design Explanation (2 pts):

Explain in a few sentences why you made the design choices you did. Why were you interested in visualizing these data? Why that type of plot? Why those colors? How did you decide on your title?

REPLACE THIS TEXT WITH YOUR DESIGN CHOICES EXPLANATION.

Once complete, download as a PDF and submit on Gradescope.