COGS9: Introduction to Data Science
Assignment #1: Data Visualization
Due date: Friday 2024 October 18 23:59:59

Download the editable version of this document and add your responses in the locations indicated. Please respond using the blue font color used in the response text, as it makes the assignments easier to grade. This is assignment is to be completed individually. Once completed, save the document as a PDF and submit on Gradescope. Be sure to assign pages to each answer when you submit (see the Gradescope instructions here.)

Student Name: Jaden Goelkel
PID: A18247795

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

For the first part of this assignment, you are to explore how to visualize the results from an analysis that used primarily textual data. You will read and critically engage with a data journalism project. Answer the questions concisely (no more than 1 sentence per point, less is fine).

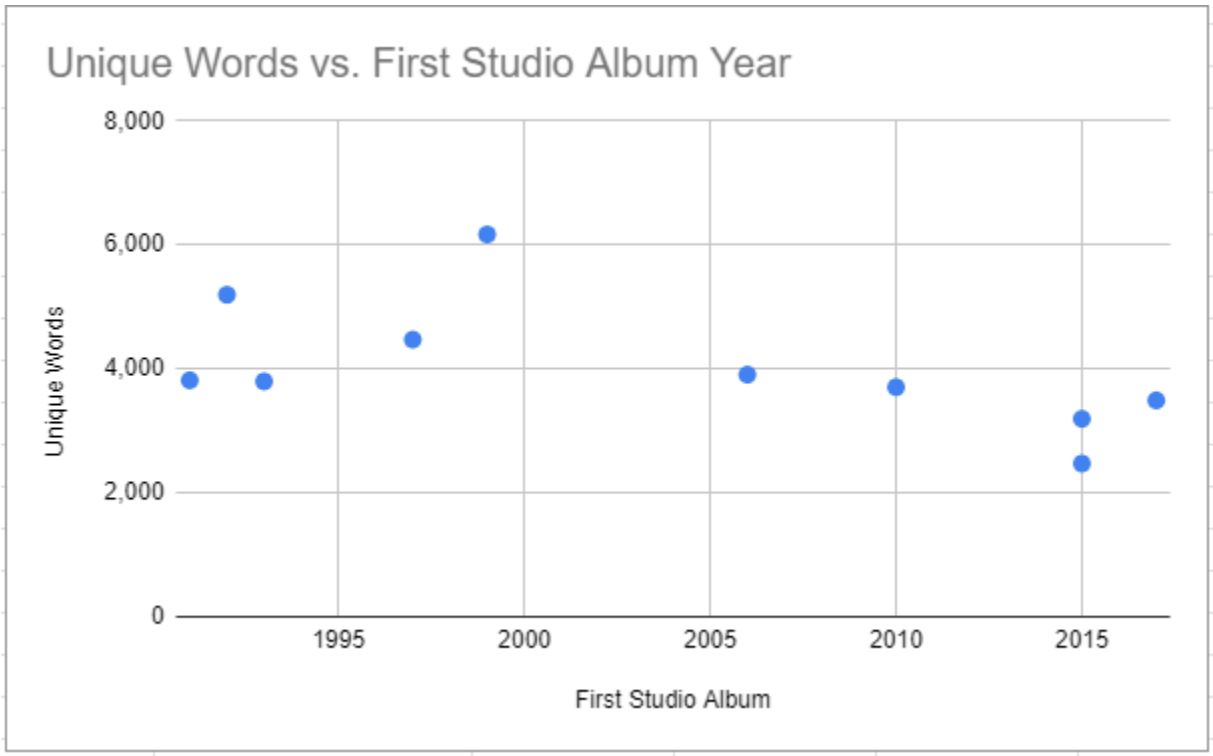
Navigate to The Pudding's piece "The Largest Vocabulary in Hip Hop":
https://pudding.cool/projects/vocabulary/index.html

a. (6 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 Google Sheets, make two columns of data: in the first column, record each of your 10 artists' number of unique words used (from above website). In the second column, find (from Google and/or Wikipedia) the year of each artists' first studio album release (e.g., 1995).

Artist	Unique Words	First Studio Album
Snoop Dogg	3,797	1993
Redman	5,196	1992
Murs	4,472	1997
MF Doom	6,169	1999
NF	2,472	2015
2Pac	3,815	1991
Migos	3,193	2015
Rick Ross	3,907	2006
Russ	3,491	2017
BoB	3,704	2010

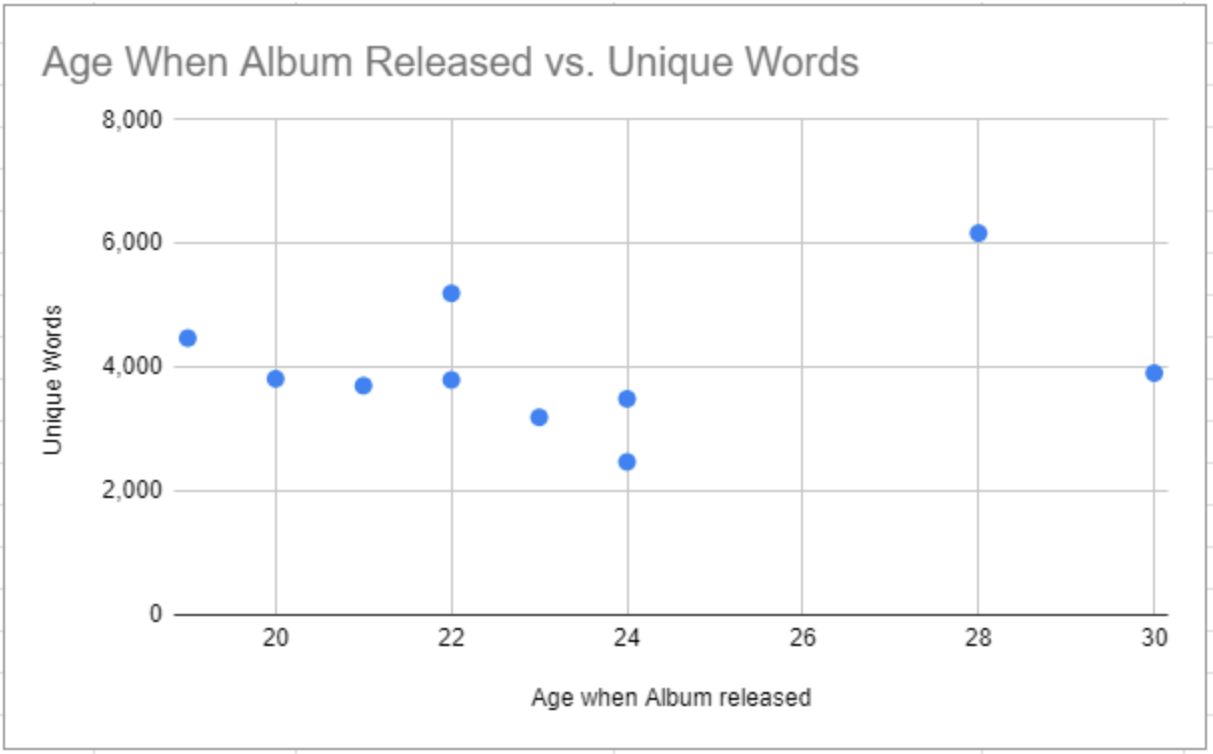
In Google Sheets, create a scatterplot of the number of unique words on the y-axis and the year of their album release on the x-axis (highlight the data, insert chart, select scatter chart). Paste your table in your submission and attach the resulting plot. Make sure to label your axes.



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?

The plot is slanted slightly down, and the amount of unique words is getting lower over time.

b. (4 pts) Repeat the above but now plot their number of unique words against the age at which they first released their album (note this doesn't work for a band or multi-artist group, so either pick a single artist, rather than the group, or choose a specific artist from that group and put their age, and explain that below your graph). Paste your new table and graph, and then describe your results. Make sure to label your axes.



Artist	Unique Words	Age when Album realese
Snoop Dogg	3,797	22
Redman	5,196	22
Murs	4,472	19
MF Doom	6,169	28
NF	2,472	24
2Pac	3,815	20
Migos	3,193	23
Rick Ross	3,907	30
Russ	3,491	24
BoB	3,704	21

There is little association between age when the album was released and number of unique words.

c. (2 pts) How do hip hop lyrics compare to rock and country lyrics?

Hip hop lyrics contain more unique words on average than Country.

d. (2 pts) When it comes to unique words used in hip hop, what do the data suggest happened between 2000s and 2010s?

The number of unique words decreased from 2000 to 2010.

e. (2 pts) Suggest 3 additional variables that the author could have extracted and analyzed from the lyrics data, other than # of unique words used. This is an act in creativity where you are required to come up with new ideas.

Number of rhymes, Average characters in each word, and percentage of unique words in each song.

f. (4 pts) What is “token analysis” and how might it bias, i.e., interfere with, our interpretation of comparing hip-hop artists’ vocabularies to that of Shakespeare and Melville? (We haven’t covered token analysis in lecture, so you will have to do some reading on your own. This is directly discussed in the article, and there is also an informative link to another article that covers token analysis in a bit more detail. You answer only needs to be a sentence or two.)

Token analysis is the process of break the text down into smaller sub units called tokens, this creates a difference between Shakespeare and Melville because token analysis has subword bias.

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 (mundane, but acceptable), 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 project. 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):

* My Garmin Data

Day	Steps Taken	Time spent walking (Minutes)
Day 1	7,512	45
Day 2	8,310	50
Day 3	5,923	35
Day 4	10,201	60
Day 5	6,716	40
Day 6	12,521	75
Day 7	9,024	55
Day 8	8,801	52
Day 9	7,211	43
Day 10	11,634	65

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.

* My Garmin Data

Day	Steps Taken	Time spent walking (Minutes)
Day 1	7,512	45
Day 2	8,310	50
Day 3	5,923	35
Day 4	10,201	60
Day 5	6,716	40
Day 6	12,521	75
Day 7	9,024	55
Day 8	8,801	52
Day 9	7,211	43
Day 10	11,634	65

Data Visualization (10 pts):

Generate an effective 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 lecture.

Visualization Interpretation (2 pts):

The more steps taken leads to a great amount of time spent walking.

Design Explanation (4 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?

I choose to use a scatter plot because it visually shows the change of steps taken and the effects that it has on the time spent walking. I tried following the principle that we learned in class to show the whole picture of the data. I used simple colors to not distract from the message the graph is trying to convey.