Report 1: Preliminary Analysis and Visualization

Due Date: October 29, 23:59

Overview

In this report, your group will analyze the different types of data that is in your dataset. You will also visualize your dataset in different ways. There should not be any noticeable screenshots for any parts of this report.

Analysis

1. Calculate the range, mean, and mode of your attributes where it makes sense to do so.
2. Using two statistical metrics, calculate values for your attributes
3. Describe why you selected those two metrics
4. Write a paragraph on what this analysis describes about your data.

Visualization

1. Visualize your dataset using five different visualization techniques. Describe what each visualization element is displaying (what do you want the reader to understand about the image). The depiction of your dataset must be easy to see and comprehensive. Points will be deducted for any parts that reduces the quality of the visualization (i.e. noticeable screenshots, small font, etc.).

Code

You will be submitting a python script then will allow me to generate the same type of information on a subset of your dataset. Follow these steps for the coding aspect:

1. Create a subset of your dataset that contains 100 samples. If your dataset consists of running text, the subset will contain 1000 words. Save this subset data as its own file, because I will be running your code on it.
2. Create a python script report_1.py that will take the subset file as an argument (i.e. I will call the script as “python report_1.py <subset_file>)”.
3. report_1.py should display **only** the following
    a. “ANALYSIS” followed by each statistical metric and its value on different rows for attributes represented by columns.
    b. The five visualization figures

Example output of your code.

ANALYSIS

	Attr_1	Attr_2	Attr3
Range	[2,50]	[3,9]	[-4,2]
Mean	33.4	4	0.2
Mode	33	4	0
M_a	2	5	-1
M_b	2	34	111

"Followed by 5 graphs"

Attr are your attributes, M_a, and M_b are your extra metrics used for evaluation. Keep in mind that not all metrics should be applied to all attributes (i.e. mean car model).

The code should be implemented in python 3, because I will be running it in python 3. I will be running your code on your submitted subset and it must run to get full points. I will not get the same results that are in your report, because I will run the script on a small subset. The results in your report should represent your entire dataset.

You can use any python packages that can be installed through python's pip.

This and all future reports **must** be written using provided latex style file and bibliography file. Only one member of the group is required to submit the report.

What to Submit

1. A pdf file that was created using the provided latex style file that includes your analysis and visualization. Make sure you include descriptions within those sections.
2. Python script report_1.py
3. A small subset of your dataset that I will run the python script on.

Rubric Guide

This is a rough guide to how the report will marked, but it might differ a little.

Points	Category
2	Latex format
3	Grammar
5	Analysis
15	Visualization (3 points for each technique)
10	Code