CS 240
EXPLORATORY DATA ANALYSIS
Spring 2020
Assignment 1
Due Date: Sunday, March 15th, 2020, 23:59

Assignment Submission: *Turn in your assignment by the due date through LMS. For your submissions,* ***you should use the template.ipynb file*** *and save the file as <your first name>_<your last name>.ipynb. Run all the cells and upload the file on LMS.*

All work in the questions must be your own; you must neither copy from nor provide assistance to anybody else. If you need guidance for any question, talk to the instructor or teaching assistants.

Late Assignment Policy: *You have a total of 4 days of late assignment turn-in allowance throughout this semester. For a single assignment, you can use a maximum of* ***2 late-days****. You decide which assignments you are going to use your late-days. After assignment due date/time, each 24-hours period is counted as one late date (i.e., if you submit late 1 hour or 23 hours, you use 1 late-date). It is your responsibility to keep track of your late days. If you are late more than 2 days for any assignment or you exhausted your late days, you get 0 from the late assignment (No exceptions)*

In this assignment, you will do exploratory data analysis to understand a dataset and its features. This assignment is mainly about how to use the ***datascience*** python library to do basic data exploration.

For this assignment, you are provided with two datasets: **Background.csv** and **Grades.csv.** The first dataset includes background information for a group of students, including Gender, Ethnic Background, Parental Level of Education and so on. The grades dataset includes the scores the students received in three different parts of an exam: math, reading and writing. **ID** column is unique student ID and refers to the same student in both tables.

1. **(5 pts)** Create two tables called **background** and **grades** out of the provided files.

2. **(5 pts)** Report the names of columns and the number of rows for both tables.

3. **(10 pts)** The **grades** table contains three different numerical columns representing the scores a students obtained in the different parts of the exam. Aggregate these results into a new column called **Average Score** which is the weighted average of 40% Math,

30% Reading and 30% Writing scores.

4. **(5 pts)** Sort the **grades** table according to this new column in **descending** order (highest to lowest). Report how many students received an average score of 95 and higher.

5. **(10 pts)** Plot 3 scatter plots showing the relations between the 3 different scores:math, reading and writing. For example, the first scatter plot would have math in the y-axis and reading in the x-axis. Repeat this with the other two combinations (math vs. writing, reading vs. writing) . Briefly discuss whether you see an **association/relation** between these scores.

6. **(10 pts)** Group the data according to the **gender** of the students that took the exam. Report which gender, on **average**, did better in each score type (math, writing, reading, average score).

7. **(10 pts)** Create a new column called **Letter Grade** that contains the letter grade a student would have received according to the table below and add this new column to grades table. Plot a distribution of the letter grades in the data and report the letter grade with the highest frequency.

Average Score	Letter Grade
X >= 90	A
80 <= X < 90	B
70 <= X < 80	C
60 <= X < 70	D
X < 60	F

8. **(10 pts)** Using the **Letter Grade** column created in part **7,** report which ethnic group has the highest number of failures (received a letter grade of F)

9. **(10 pts)** Find all students that **failed in the math score** but **got at least 70 in the writing score.**

10. **(10 pts)** Plot two different **histograms** for each score according to whether the student attended a preparation course or not. **Use the following bin sizes:** 0 - 50, 50 - 60, 60 - 65, 65 - 70, 70 - 75, 75 - 80, 80 - 90, 90 - 100.

11. **(10 pts)** Determine if it was more likely for a student to get an **A or B** if their parents had an education higher than high school or not. To answer this question, find the percentage of students who got A or B in both cases and compare.

12. **(5 pts)** Show in a grid format how many students got a letter grade of **A** for each gender and ethnicity combination (rows are genders and columns are ethnicities).

IMPORTANT NOTES

- Prepare and upload one Jupyter notebook file which should be named as <your first name>_<your last name>.ipynb.
- A sample Jupyter notebook file provided to you. Follow the template's structure.
- Explain your code with comments.

Wrong file name format	-10 points
Not using template	-10 points