**CENG313 Introduction to Data Science**

Fall 2020-2021

Lecturer: Dr. Duygu Sarıkaya

Teaching Assistant: Kevser Özdem

Gazi University, Department of Computer Engineering

**Assignment 1 due on 27th of November, Friday 23:59**

**Assignment 1: Exploratory Data Analysis of the Titanic: Machine Learning From Disaster dataset**

**Titanic: Machine Learning From Disaster dataset**

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. You will find the dataset (`train.csv`) that include passenger information like name, age, gender, class, etc.

In this assignment you are asked some questions which will guide your explonatory data analysis of the dataset. You will submit a jupyter notebook (ipynb file) with executable Python script and in each executable section you should answer the related question. Please do not forget to indicate the number of the question (Q1,Q2,Q3 etc.) at the top of the related section(comment line). You will write Python scripts, and you will use the libraries we covered in class (pandas, numpy, matplotlib, scikit-learn). You should import all the libraries you will use at the top of your notebook. Please refer to course slides, tutorials and practicals to set up a running Python environment, Jupyter notebook and to import these libraries. You can check the documentation of each library (available online) to get more information about the functions you will use.

**Important Note:**

You are not asked to answer the questions manually, you will submit the executable script that allows you to answer the questions. You will receive points only if your script executes, shows the correct answer, and includes the explanation (text in comment section at the top of each section) if asked in the question.

This is an individual assignment, meaning that you will be working on it alone (please check the Class Rules and Expectations below, also available in the syllabus)

**Submission:**

You will submit a jupyter notebook (ipynb file) with executable Python script and comments (explanations) for each question.

**Grading:**

Each question is 5 points and the total of the 20 questions is 100 points. You will receive points only when your script 1)executes, 2)gives the correct answer, and when 3)the explainations are provided.

**Course Rules and Expectations**

All work on programming assignments must be done individually unless stated otherwise. You are encouraged to discuss with your classmates about the given assignments, however, everything that is turned in for each assignment must be your own work. In particular, it is not acceptable to: submit another person's assignment as your own work (in part or in its entirety), get someone else to do all or a part of the work for you, submit a previous work that was done for another course in its entirety (self- plagiarism), submit material found on the web as is etc. These acts are in violation of academic integrity (plagiarism), and these incidents will not be tolerated. Homeworks, programming assignments, exams and projects are subject to Turnitin (https://www.turnitin.com/) checks.

**Data Dictionary**

| Variable | Definition | Key |
|---|---|---|
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Gender | |
| Age | Age in years | |
| sibsp | # of siblings / spouses aboard the Titanic | |
| parch | # of parents / children aboard the Titanic | |
| ticket | Ticket number | |
| fare | Passenger fare | |
| cabin | Cabin number | |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

For starters, you can open the csv (comma separated value) file and create a data frame using the pandas function read_csv:

import pandas as pd

df_titanic = pd.read_csv('../input/titanic/train.csv') // you should replace the path with your own

df_titanic.info() // shows information about the data frame you have just created for the Titanic dataset

**Questions:**

1. Please show all the information that belongs to the first five passengers. You should have 5 rows each referring to a passenger, and the values of 12 features (columns) for each passenger.
2. Please show the size and dimension of the dataset: (number of passangers, number of features). Do not forget to write what the output of your script refers to.
3. Please check how many missing values there are in the dataset for each feature column. Missing values will have a null value (NaN). Do not forget to write which classes have missing values, and how many missing values in the comments.
4. Please create a pie chart which shows the **percentage** of passengers that survived and the percentage of the passengers that did not survive. Explain in your comments if more people have survived or did not survive.
5. Please create a bar chart that shows the **number** of female passengers and the total number of male passengers (You should have two bars referring to female and male)
6. Please create a bar chart that shows the **number** of females and males who survived and who did not survive. (You should have four bars referring females who survived and didn't survive and males who survived and didn't survive)
7. Please create a bar chart that shows the surival **rates** of females and males. (You should have two bars referring to female and male) Explain your observations in your comments, are there more female or male passengers in total? Did more females or males survive? What might be the reason?
8. Please create a cross table as shown below (x will be computed and included in your answer). The cross table makes it possible to get information about how many people in the 2nd class have survived etc. Please indicate which class has the most number of survivors? Which class has the lowest number of survivors? Manually compute the rate for each class: number of survivors in that class/all passengers in that class.

| Survived | 0 | 1 | All |
|---|---|---|---|
| Pclass | | | |
| 1 | x | x | x |
| 2 | x | x | x |
| 3 | x | x | x |
| All | x | x | x |

9. Please create a bar chart that shows the number of passengers who survived and who didn't survive for each class. (You should have 6 bars in total)
10. Please create a cross table as shown below (x will be computed and included in your answer). (Simiar to Q8 but gender information is added). Explain which gender had a higher survival rate? What might be the reason?

| | Pclass | 1 | 2 | 3 | All |
|---|---|---|---|---|---|
| Sex | Survived | | | | |
| female | 0 | x | x | x | x |
| | 1 | x | x | x | x |
| male | 0 | x | x | x | x |
| | 1 | x | x | x | x |
| All | | x | x | x | x |

11. What is the age of the olders passenger?
12. What is the age of the youngest passenger?
13. What is the average age of the passengers?
14. Please plot the histogram that shows the age distribution of the passengers **who survived**. (You should have 10 bins for the range of the ages.)
15. Please plot the histogram that shows the age distibution of the passengers who didn't survive. (You should have 10 bins for the range of the ages.)
    Please explain your findings relating the Q14 and Q15 in writing.
16. How much is the the lowest fare?
17. How much is the highest fare?
18. How much is the average fare?
19. Please plot the histogram that shows the distribution of passengers according to the fare they paid. (You should have 10 bins for the range of the ages.)
20. Are there any children under the age of 10 traveling without their parents? What might this indicate?