CENG313 Introduction to Data Science

Fall 2022-2023

Lecturer: Dr. Duygu Sarıkaya

Teaching Assistant: Berrin İşlek

Gazi University, Department of Computer Engineering

Assignment 1 is due on the 7th of November, Monday 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 the comment section at the top of each section) if asked in the question.

You will be using the libraries we have covered in class (pandas, matplotlib, and if needed you may use numpy, scipy, sci-kit learn (won't be necessary for this assignment). **For specific questions, you may use seaborn or plotly libraries** if it is indicated so.

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 an executable Python script and comments (explanations) for each question. The file will be uploaded on lms (guzem).

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 explanations 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 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. Homework, 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 six passengers**. You should have 6 rows each referring to a passenger, and the values of 12 features (columns) for each passenger.
2. Please list the attributes (column titles).
3. Please show the size of the dataset: (the number of passengers **only**). Do not forget to write what the output of your script refers to.
4. Please check how many missing values there are in the dataset for the **columns "Age","Cabin" and "Embarked".** Missing values will have a null value (NaN). Do not forget to write how many missing values there are for each of these three columns in the comments.

   Important Note: For the rest of the homework, you can delete the instances that have NaN values for **specific attributes (columns) asked in the related question**.

5. Please create a pie chart that shows the **percentage** of passengers that embarked from each port (Southampton, Cherbourg, Queenstown). Explain in your comments which port the most number of people embarked from and which port the least amount of people embarked from.
6. Please create a bar chart that shows the **number** of passengers traveling for each class. (You should have three bars referring to each class). Please explain which class had the most number of passengers.
7. Please create a plot that shows the **number** of **female passengers** for each ticket class who survived and who did not survive in **bar chart** format. (You should have three bar charts (for each ticket class) referring to passengers of ticket class 1 who survived and didn't survive, passengers of ticket class 2 who survived and didn't survive, and so on)
8. Please create **two boxplots** that show the **key age statistics** of **female and male passengers who survived and who did not survive**. (You should have two boxplots, one for female passengers and one for male passengers, each of these box plots having two boxes for the passengers who survived and who didn't, and each of these boxes should communicate key age statistics related to each group (e.g. female who survived)
9. Please create a **cross table** as shown below (x will be computed and included in your answer).
10. The cross table makes it possible to get information about how many people for each gender have survived etc. Please indicate which gender has the most number of survivors. Explain which gender had a higher survival rate. What might be the reason?

Survived	0	1	All
Sex			
female	x	x	x
male	x	x	x
All	x	x	x

11. Please create a **heatmap** for the correlation between the attributes of survival, age, sex, fare, and ticket class. You may use seaborn or plotly libraries for this question.
12. Please calculate the **"Pearson" standard correlation coefficient** for each correlation mentioned in Q11 and show it on the heatmap you created for Q11. You may use seaborn or plotly libraries for this question. How is survival is correlated to ticket class? Please explain.
13. What is the age of the **youngest** passenger?
14. How much is the **average** fare?
15. What is the age of the **oldest passenger who survived**?
16. What is the age of the **oldest female passenger who survived**?
17. Are there any children **under the age of 10** traveling **without their parents**? What might this indicate?
18. Please plot the histograms that show the age distribution of the female and male passengers **who survived**. (You can arrange 10 bins for the range of ages [0 to 100].) You should have two histograms overlayed on top of each other in different colors. You may use seaborn or plotly libraries for this question. Please explain which gender had a higher survival chance by looking at these histograms.
19. What is the number of siblings of the passenger who has **the highest number of siblings**?
20. Please plot **a scatterplot** that shows the **age–fare correlation** for **passengers who survived and who did not survive.** (You should have age on the x-axis and fare on the y-axis, and you should color code the passengers according to their survival) You may use seaborn or plotly libraries for this question.