CENG313 Introduction to Data Science

Fall 2021-2022

Lecturer: Dr. Duygu Sarıkaya

Teaching Assistant: Dr. Begüm Mutlu Bilge

Gazi University, Department of Computer Engineering

Assignment 1 due on 10th of November 2021, 23:59 (to be submitted on guzem)

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. The file will be uploaded on lms (guzem). Note: As güzel does not accept ipynb file extension, you can zip your file and submit it that way.

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/) and Moss (Measure Of Software Similarity) 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 **last 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 and dimension of the dataset. 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 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 are there in the comments.
   Important Note: For the rest of the homework, you can delete the instances that have NaN value for specific attributes (columns) asked in the question.
5. Please create a pie chart which shows the **percentage** of the passengers that were traveling in the 1st, 2nd and 3rd classes. Explain in your comments which class has more passengers.
6. Please create a bar chart that shows the **number** of passengers who survived and the number of passengers who did not survive (You should have two bars referring to survived and did not survive)
7. Please create a plot that shows the **percentage** of passengers for each ticket class who survived and who did not survive in **pie chart** format. (You should have three pie 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 a bar chart that shows the survival **rates** of each ticket class. (You should have three bars referring to each ticket class) Explain your observations in your comments, are there more class 1, 2 or 3 passengers in total? Which class had the highest survival rate? What might be the reason?
9. 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 of each gender have survived etc. Please indicate which gender has the most number of survivors? Which gender has the lowest number of survivors? Compute the rate for each gender: number of survivors for that gender/all passengers of the gender.

Survived	0	1	All
Gender			
male	x	x	x
female	x	x	x
All	x	x	x

10. Please create a bar chart that shows the **number** of passengers who survived and who didn't survive for each Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton). (You should have 6 bars in total)
11. Please create a cross table as shown below (x will be computed and included in your answer). (Similar to Q9 but age group information is added. For the age group, we will have [0-18] as child, [19-60] as adult, [60 and over] as old). Explain which age group had a higher survival rate? What might be the reason?

	Age group	child	adult	old	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

12. Please create a heatmap for correlation between whether a passenger survived and the **strongest** three attributes. (You can remove the weakly correlated attributes and show the strongest three attributes)
13. What is the most commonly used title (you can get the title information from the Name attribute)?
14. How many distinct titles are used? Please list these titles. (see Q14)
15. What is the average age of the passengers?
16. Please create a strip chart (You can use the library seaborn for this) that plots age (y-axis), and survival (x-axis).
17. What is the age of the oldest person who survived?
18. Please create a pie chart that shows the **percentage** of people who **survived** according to the their age group. For the age group, we will have [0-18] as child, [19-60] as adult, [60 and over] as old.
19. Please create a pie chart that shows the **percentage** of people who **did not survive** according to the their age group. For the age group, we will have [0-18] as child, [19-60] as adult, [60 and over] as old. Please explain your findings relating the Q20 and Q21 in writing.
20. What is the number of the siblings of the passenger who has the highest number of siblings?