Data Science 2024-2025

COMP5122M

Practical 1 (Week 2)

Lecturer: Duygu Sarikaya

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 exploratory data analysis of the dataset. You are not asked to answer the questions manually, please use your tools of choice for data analysis to answer the questions.

For this Practical, please work in pairs. You are asked to complete a report answering the questions below with a short explanation if the question asks for one.

If you prefer to use Python:

Please refer to the tutorial on Minerva to set up a running Python environment, Jupyter notebook, or simply log in on Google Colab, and import the libraries you will need. You can check the documentation of each library you will need such as NumPy, Pandas, Matplotlib (available online) to get more information about the functions you will use.

If you prefer to use Tableau:

Please refer to the "Tableau getting started & tutorial" on Minerva.

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

If you are using Python:

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
```

If you are using Tableau, you can simply open the csv file or import it with Tableau CSV Connector.

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 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.
3. Please list the attributes (column titles).
4. 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.
5. 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.
6. What is the age of the **youngest** passenger?
7. What is the age of the **youngest passenger who survived**?
8. How much is the **average** fare?
9. What is the age of the **oldest passenger who survived**?

10. What is the age of the **oldest female passenger who survived**?
11. Are there any children **under the age of 10** traveling **without their parents**? What might this indicate?
12. What is the number of siblings of the passenger who has **the highest number of siblings**?
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 Q12)
15. What is the **average age of the passengers**?