Statistics for Data Science
UE21CS241A
PES University,
Bangalore

# Case Study for the Datathon

Teaching Assistants:
Ananya Jha
Ananya Mahishi
Ananya J

# Chapter 1

# Introduction

The two sections below, Background and Case Study provide context for the data science hackathon. This exercise will allow you to test your skills in using the Python programming language to effectively explore the characteristics of a dataset and analyse the features using descriptive statistics such as summary statistics, tables, and graphs. Happy coding!

# Background

The Titanic's tragic sinking stands as one of history's most infamous maritime disasters. During her maiden voyage on April 15, 1912, the Titanic struck an iceberg, resulting in her demise and the loss of over 1,500 lives out of the 2,224 passengers and crew on board. This unforgettable catastrophe remains etched in the collective memory of the world. The construction of the Titanic cost a staggering $7.5 million, yet her fate was sealed by a collision with an iceberg.

# Case Study

Embarking on a quest into the past, a researcher aims to study the effects of various factors on survival rates of passengers aboard the Titanic. He records information about passengers including demographics, ticket class, cabin allocation, and survival outcomes. Through the lens of statistical analysis, we will use the dataset thus obtained to shed light on the conditions that favoured survival or led to tragedy in that fateful maritime disaster.

# Dataset Description

The Titanic dataset contains information about passengers who were on board the Titanic when it sank on April 15, 1912. The variables in the dataset include:

1. PassengerId: a unique identifier for each passenger
2. Survived: a binary variable indicating whether a passenger survived (1) or not (0)
3. Pclass: the passenger class (1 = first class, 2 = second class, 3 = third class)
4. Name: the name of the passenger
5. Sex: the gender of the passenger (male or female)
6. Age: the age of the passenger
7. SibSp: the number of siblings/spouses aboard the Titanic for the passenger
8. Parch: the number of parents/children aboard the Titanic for the passenger
9. Ticket: the ticket number for the passenger

10. Fare: the fare paid by the passenger
11. Cabin: the cabin number for the passenger
12. Embarked: the port of embarkation for the passenger (C = Cherbourg, Q = Queenstown, S = Southampton)

# Chapter 2

# Problem Set

1. Classify the features in the Titanic dataset into their appropriate data types (ordinal, nominal, interval, or ratio). Provide a rationale for each classification.
2. A summary statistic provides a numerical summary of a specific feature within the dataset. There are two commonly used categories of summary statistics: those that indicate the central tendency and those that indicate the spread of the data. Identify the most appropriate measure of central tendency for each attribute in the dataset and state its corresponding value. Additionally, calculate the standard deviation and range of values for each column.
3. Identify and describe any data quality issues or inconsistencies within the Titanic dataset. What steps would you take to clean and preprocess the data to ensure its accuracy and reliability for further analysis?
4. Using a histogram and box plot, assess the presence of outliers in the 'Age' and 'Fare' variables. Describe the visualisations, identify any potential outliers, and explain how you determined their presence or absence.
5. What actions would you take to resolve the presence of outliers? Visualise the changes. Hint: Use boxplot and histogram
6. Examine the normal probability plot (Q-Q plot) for the 'Fare' variable in the Titanic dataset. Based on the shape and trend of the plot, what conclusions can be drawn? Provide a rationale for your conclusions.
7. Calculate the correlation between age and other numerical variables (e.g., fare or the number of siblings/spouses). Set a correlation threshold and create a heatmap to visualise the relationships.
8. Generate a pairplot that includes the variables 'Age,' 'Fare,' and 'Sex' while using 'Survived' as the hue in the Titanic dataset. What insights can be gained from the pair plot, and how does it help in visualising the relationships between age, fare, gender, and survival status on the Titanic?

9. Use hypothesis testing to answer the following. Define a null and alternate hypothesis. Use a T-test to investigate. Does the amount of Fare paid by the passengers have a significant impact on their survival chances in the Titanic disaster? Plot a histogram to visualise the results. Assume the significance level as 0.05.
10. Calculate the margin of error to quantify the precision of the analysis done previously and what you can infer from the results.