Republic of the Philippines
POLYTECHNIC UNIVERSITY OF THE PHILIPPINES
College of Engineering
COMPUTER ENGINEERING DEPARTMENT

CMPE 363: BIG DATA ANALYTICS

Exploratory Data Analysis of the Titanic Dataset

Submitted by:
Balangao, Samantha A.
BSCpE 3-2

Submitted to:
Engr. Edcel Artificio

Introduction and Purpose of the Analysis

Even though a century has passed since the sinking of the Titanic, it is still a widely known topic around the world, and many individuals-regardless of their profession performed analysis in the different aspects of the Titanic. In this activity, we will be exploring the Titanic dataset acquired from Kaggle which is based on the information about the passengers of the said ship. It includes extensive details such as their survival status, passenger class, gender, age, familial connections, their fares, and from their embarkation points.

The purpose of this analysis is to explore the factors that may have influenced survival during the unfortunate disaster. Specifically, we aim to understand how certain variables like age, gender, passenger class, and town affected their likelihood of survival. This will be attained through statistical summaries and visualizations in order to see and identify possible patterns or outliers in survival outcomes.

Data Dictionary

This part of the paper contains the variables (columns) that are utilized in the dataset which enabled us to produce insights and queries:

Variable	Data Type	Value	Definition
PassengerId	Categorical	1, 2, 3, … 890	For identification or identifier of passengers
Survived	Categorical	1 or 0	1 = Survived 0 = Did not survive
Pclass	Categorical	1, 2, 3	This may pertain to the economic classes of the passengers (1st, 2nd, or 3rd)
Name	Object	Braund, Mr. Owen Harris; Cumings, Mrs. John Bradley; Heikkinen, Miss Laina	Names of passengers
Sex	Object	Male or Female	Sex/gender of the passengers
Age	Continuous	0 to n	Ages of the passengers
SibsSp	Categorical	0 to n	Number of siblings/spouses aboard

Parch	Categorical	0 to n	Number of parents/children aboard
Ticket	Object	A/5 21171; PC 17599; STON/02.3101282	Their ticket number
Fare	Continuous	7.2500; 71.2833; 7.9250	Fare amount
Cabin	Object	C85; C123; E46	Cabin assignment
Embarked	Object	C (Cherbourg); Q (Queenstown); S (Southampton)	The town/port they boarded the ship from.

Analysis Process

In order to approach this EDA, we employed a combination of statistical and graphical techniques to produce meaningful insights and identify possible patterns within the dataset. Specifically, our analysis focuses on understanding the relationship between survival and various factors/variables such as the passengers' sex, the town they came from , and their ages.

Firstly, we simply "explored" the data by examining the csv file; or in the case of importing the Titanic dataset from the Seaborn library, we can run head() to get a preview of what we are working on. Additionally, by utilizing **shape**, **info()**, **isnull().sum()**, and **nunique()**, we were able to summarize key elements of the dataset- its size, the data type, the count of null, and unique values. However, just by doing that may still produce vague or unsatisfactory insights about their relation so we have also utilized a **heatmap** to observe correlations between numerical variables. Here is the result of the correlation:

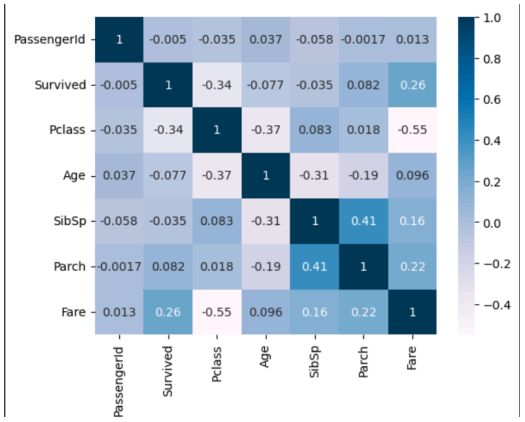


Image 1. The correlation of the numerical variables within the dataset

This initial process of "exploring" helped us generate specific questions focused on the passengers' survival aspect, which is the focus of our EDA. These queries will be then answered through statistical and graphical techniques for better understanding and can recommend further steps for a much more in-depth analysis. The statistical techniques primarily focused on statistical insights. Through the use of **describe()**, it allows us to view the descriptive statistics of our data frame since most of our values are numerical; allowing us to look into the statistical insights of our dataset which aids us in our queries. Afterwards, we made use of average/mean (**mean()**), the total count (**count()**), and formulas that calculate the supposed outcome of the questions. **Pandas**, **Matplotlib**, **Seaborn**, and **Plotly** libraries were also used for us to visualize these statistical outcomes in the form of charts, tables, and a box plot.

The guiding questions are as follows:

1. How many **survived** the Titanic sinking?
   1.1. What is the average **survival** rate?
2. Based on **sex** (gender), which group had a higher survival rate?
3. From which **town** (port) did people embark from, and how many survived from each location?
4. Are there more survivors across particular **classes** (1, 2, 3)?
5. What is the average **age** of the survivors?
   5.1. How does the age distribution differ between survivors and non-survivors?

Analysis and Insights

From the questions above, they certainly produced meaningful and essential outcomes for further analysis or use. To elaborate, the questions are answered through the following statistical and graphical techniques/visualizations.

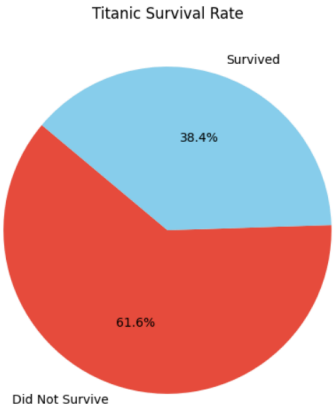
| Question 1. How many passengers survived the Titanic sinking?

Titanic Survival Count Table

Survival Status	Count
Did Not Survive	549
Survived	342

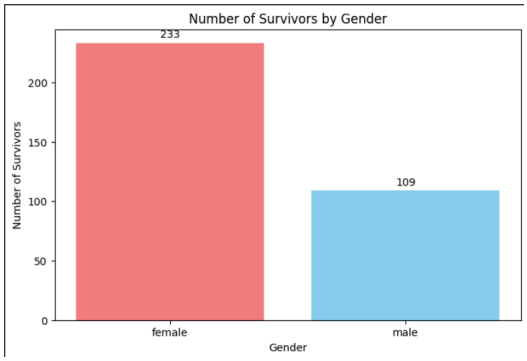
This table shows a clear count of the passengers who survived and those who did not.

|Question 1.1. What is the survival rate of the Titanic?

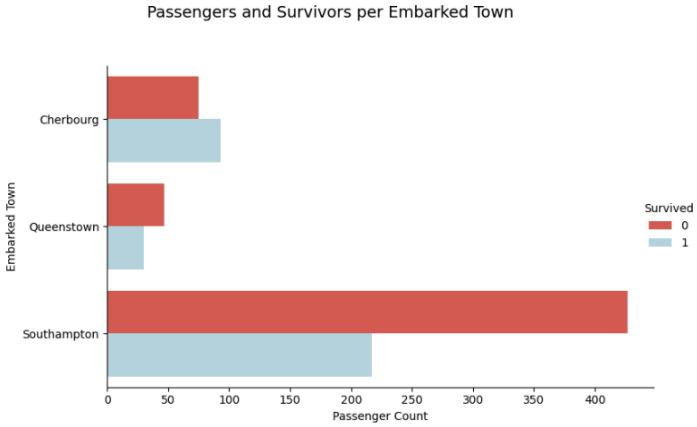


This chart shows that the survival rate is lesser than the death rate when the ship sank.

|Question 2. Based on sex (gender), which had more survivors?

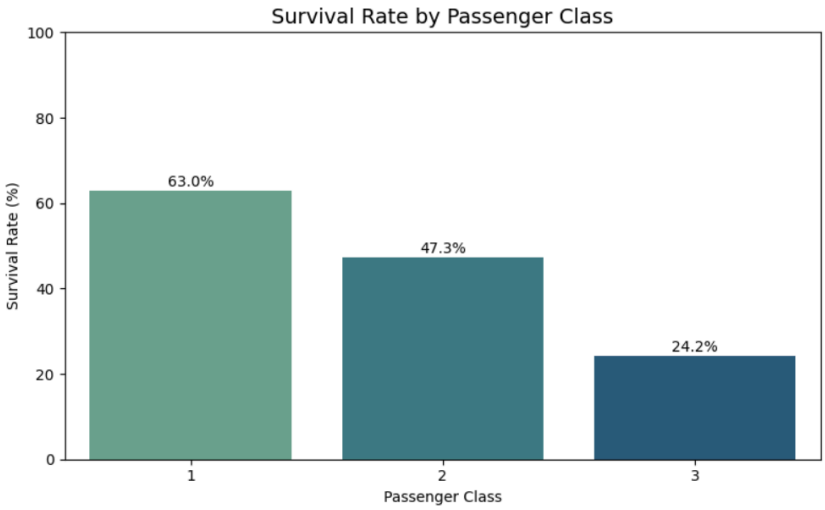


|Question 3. From which town (port) did people embark from, and how many survived from each location?



This shows that the majority of the passengers boarded from Southampton, thus also having more records on the deaths and survival of these individuals.

| Question 4. Are there more survivors across particular classes (1, 2, 3)?



We can see that the first class (probably of the nobility class) had a higher survival rate compared to the other two classes. This can possibly raise questions like "were 1st class individuals prioritized to be rescued first before the common people?" which opens a new discussion that may need a more in-depth analysis.

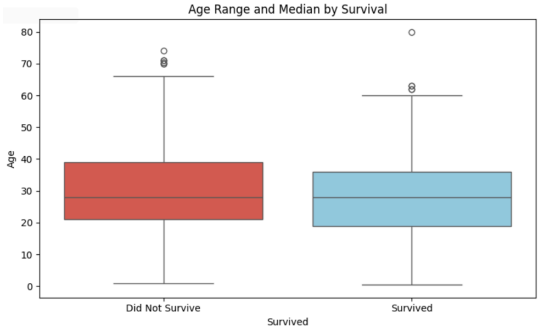
| Question 5. What is the average age of the survivors?

```
#What is the average age of the survivors?
avg_age = df[df['Survived'] == 1]['Age'].mean()
print(f"The average age of the survivors is: {avg_age:.2f} years")
```

```
The average age of the survivors is: 28.34 years
```

For this question, we only want to get the mean of the ages of all passengers to learn if most of the survivors were younger or closer to adult age. This opens to a possibility of questions on age's relation to survival such as

| Question 5.1. How does the age distribution differ between survivors and non-survivors?



The boxplot displays that the survivors generally had a slightly lower median age, but their outliers were significant which suggests that age alone does not really have a strong correlation with their survival.

Thus, these insights only provide an overview of what is "in" the dataset. It tells a brief story of the Titanic through the lens of survival and its relation to the variables that may influence the survival rate. These few questions open to more queries, especially those that are necessary for in-depth analysis or predictions- such as in big data analytics or machine learning.

Conclusion and Recommendation

To conclude, this exploratory data analysis teaches us how to effectively explore our data before doing any in-depth processes that significantly affect how we see the data given to us. It is important that we impose our curiosity as we approach our data, in ways such as asking even the simple questions early as this helps us in having a more extensive and meaningful analysis as we dive deeper into it.

As big data analytics majors, it is imperative that we learn to be curious when we try to analyze datasets, regardless of their size. By having this mindset, it allows us to get deeper insights and understand the data given to us better. Data exploration is essential in helping us formulate ideas on how to tell the story behind the data, whether we focus on one subject and relate it to the variables like we did in this activity, or tailoring it based on another individual's request (e.g., client requests). That being said, we get to plan how we can present the information to our audience in a clear and impactful way.

In this EDA, we were not able to explore other factors such as the ticket information and cabin details. Thus, this leaves an opportunity for future analysts to explore the different factors/variables, either in relation to survival rates or through a different lens to focus on. There are still remaining factors that are yet to be discovered in which analysts can provide a fresh perspective on.

References:

M, Yasser. H. (2021). *Titanic Dataset.* Kaggle. https://www.kaggle.com/datasets/yasserh/titanic-dataset

Tikkanen, A. (2025). *Titanic.* Britannica. https://www.britannica.com/topic/Titanic/Aftermath-and-investigation

Annex

Scripts and Codes

```
#BALANGAO, Samantha A. (BSCpE 3-2)
#Exploratory Data Analysis

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go

df = pd.read_csv('Titanic-Dataset.csv')
df.head()
```

```
df.shape
```

```
df.info()
```

```
df.isnull().sum()
```

```
df.nunique()
```

```
#Using a heatmap, we can visualize the correlation of the NUMERICAL
columns to each other.
df_num = df.select_dtypes(include=['number'])
sns.heatmap(df_num.corr(),
            annot=True,
            cmap='PuBu')
plt.show()
```

```
#Using describe(), we are able to get a summary of descriptive analytics
```

```
df.describe()
```

```
#1. How many survived the Titanic sinking?
#Calculating the total number of survivors and non-survivor in the column
'Survived'.
survived_count = df['Survived'].value_counts()

#Using a table
fig = go.Figure(data=[go.Table(
    header=dict(values=['Survival Status', 'Count'],
                fill_color='lightblue',
                align='left'),
    cells=dict(values=[['Did Not Survive', 'Survived'],  # Survival Status
                       [survived_count[0], survived_count[1]]],  # Counts
               fill_color='lavender',
               align='left'))
])

fig.update_layout(title_text="Titanic Survival Count Table")
fig.show()
```

```
#1.1. What is the survival rate of the titanic?
#Calculating the total number of survivors and non-survivor in the column
'Survived'.
survived_count = df['Survived'].value_counts()

#Using a pie chart
plt.figure(figsize=(6, 6))
plt.pie(survived_count,    labels=['Did    Not    Survive',    'Survived'],
colors=['#e74c3c', 'skyblue'], autopct='%1.1f%%', startangle=140)
plt.title('Titanic Survival Rate')
plt.show()
```

```
#2. Based on sex (gender), which had more survivors?
#Calculating the number of survivors grouped by Sex
gender_survivors              =              df[df['Survived']              ==
1].groupby('Sex')['Survived'].count()

#Using bar chart
```

```
fig, ax = plt.subplots(figsize=(8, 5))
bars = ax.bar(gender_survivors.index,
              gender_survivors.values,
              color=['lightcoral','skyblue' ])
ax.set_xlabel('Gender')
ax.set_ylabel('Number of Survivors')
ax.set_title('Number of Survivors by Gender')
for bar in bars:
    height = bar.get_height()
    ax.annotate(f'{height}',
                xy=(bar.get_x() + bar.get_width() / 2, height),
                xytext=(0, 3),
                textcoords="offset points",
                ha='center', va='bottom')
```

```
#3. From which town did most people came from? and how many survivors are
there per town?
#Calculating the count for Survivors per Embarked Town
embarked_survivors                =                df.groupby(['Embarked',
'Survived'])['Survived'].count().reset_index(name='Count')
embarked_survivors.columns = ['Embarked', 'Survived', 'Count']
embarked_survivors['Embarked                    Town']                    =
embarked_survivors['Embarked'].map({'S':  'Southampton', 'C':  'Cherbourg',
'Q': 'Queenstown'})

#Using horizontal barchart
town = sns.catplot(
    data=embarked_survivors,
    kind='bar',
    y='Embarked Town',
    x='Count',
    hue='Survived',
    palette={0: '#e74c3c', 1: 'lightblue'},
    height=5,
    aspect=1.5)

town.set_axis_labels("Passenger Count", "Embarked Town")
town._legend.set_title("Survived")
town._legend.set_bbox_to_anchor((1.05, 0.5))
```

```
town.fig.suptitle('Passengers and Survivors per Embarked Town',
fontsize=14, y=1.05)
plt.tight_layout()
plt.show()
```

```
#4. Are there more number of survivors across particular passenger classes
(1, 2, 3)?

#Calculating the total number of passengers, then survivors per class
class_total = df.groupby('Pclass')['Survived'].count()
class_survivors = df[df['Survived'] ==
1].groupby('Pclass')['Survived'].count()

#Calculating the survival rate
survival_rate = (class_survivors / class_total *
100).reset_index(name='Survival Rate')

#Using a bargraph
plt.figure(figsize=(8, 5))
sns.barplot(data=survival_rate,
            x='Pclass',
            y='Survival Rate',
            palette='crest')

for index, row in survival_rate.iterrows():
          plt.text(index, row['Survival Rate'] + 1, f"{row['Survival
Rate']:.1f}%", ha='center')

plt.title('Survival Rate by Passenger Class', fontsize=14)
plt.xlabel('Passenger Class')
plt.ylabel('Survival Rate (%)')
plt.ylim(0, 100)
plt.tight_layout()
plt.show()
```

```
#5. What is the average age of the survivors?
avg_age = df[df['Survived'] == 1]['Age'].mean()
print(f"The average age of the survivors is: {avg_age:.2f} years")
```

```
#5.1. How does the age distribution differ between survivors and non-survivors?
#Using a boxplot
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x='Survived', y='Age', palette={'0': '#e74c3c', '1': 'skyblue'})
plt.title('Age Range and Median by Survival')
plt.xlabel('Survived')
plt.ylabel('Age')
plt.xticks([0, 1], ['Did Not Survive', 'Survived'])
plt.tight_layout()
plt.show()
```

https://colab.research.google.com/drive/1TX17Fd9WpyAhkZ0dfd7FdpOVi6j8X_MB?usp=sharing