

Data and Artificial Intelligence

Cyber Shujaa Program

Week 3 Assignment

Assignment 3: Titanic Exploratory Data Analysis

Student Name: Austin Githinji

Student ID: CS-EH03-25417

Contents

Data and Artificial Intelligence	1
Cyber Shujaa Program.....	1
Week 3 Assignment Assignment 3: Titanic Exploratory Data Analysis	1
Introduction	1
Objectives.....	2
Tasks Completed	2
Conclusion.....	5

Introduction

1. This project involves performing **Exploratory Data Analysis (EDA)** on the **Titanic dataset**, a classic dataset available on Kaggle.
The main goal is to explore, clean, and analyze the dataset to uncover hidden patterns and relationships between passenger attributes and survival outcomes.
2. EDA is an essential phase in data science projects as it helps develop a deep understanding of the data's structure, detect anomalies, and prepare it for model training or business insights.
3. Store structured data into a Pandas DataFrame.
4. Export the final dataset to a .csv file.

```
Kaggle Notebook Editor +
```

kaggle.com/code/austingithini/austin-githini-cs-eh03-25417/edit

Austin_Githini_cs-eh03-25... Draft saved
File Edit View Run Settings Add-ons Help

+ - × ↺ ↻ ▶ ▶▶ Run All Code -

Draft Session (3/16) | | | | | ⏮ ⏪ ⏩ ⏭

```
[34]: # Check for missing values in each column
missing_values = df.isnull().sum()
missing_percent = (missing_values / len(df)) * 100

missing_data = pd.DataFrame({'Missing Values': missing_values, 'Percentage (%)': missing_percent})
missing_data.sort_values(by='Missing Values', ascending=False)
```

```
[35]:
```

	Missing Values	Percentage (%)
Cabin	887	71.54177
Age	177	19.98332
Embarked	2	0.21467
PassengerId	0	0.00000
Survived	0	0.00000
Pclass	0	0.00000
Name	0	0.00000
Sex	0	0.00000
SibSp	0	0.00000
ParCh	0	0.00000
Ticket	0	0.00000
Fare	0	0.00000

* Code * Markdown

```
> # Drop Cabin column if it exist[
df.drop(columns='Cabin', inplace=True, errors='ignore')

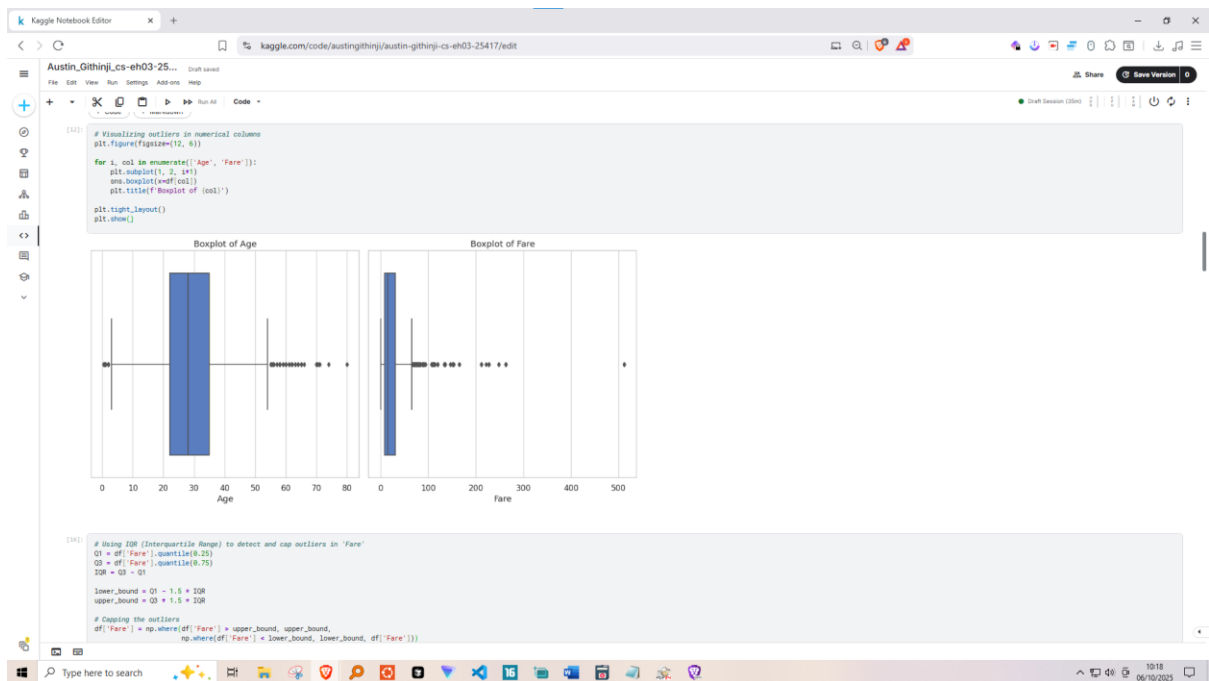
# Fill missing 'Age' values with median
df['Age'] = df['Age'].fillna(df['Age'].median())

# Fill missing 'Embarked' values with mode (most frequent value)
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])

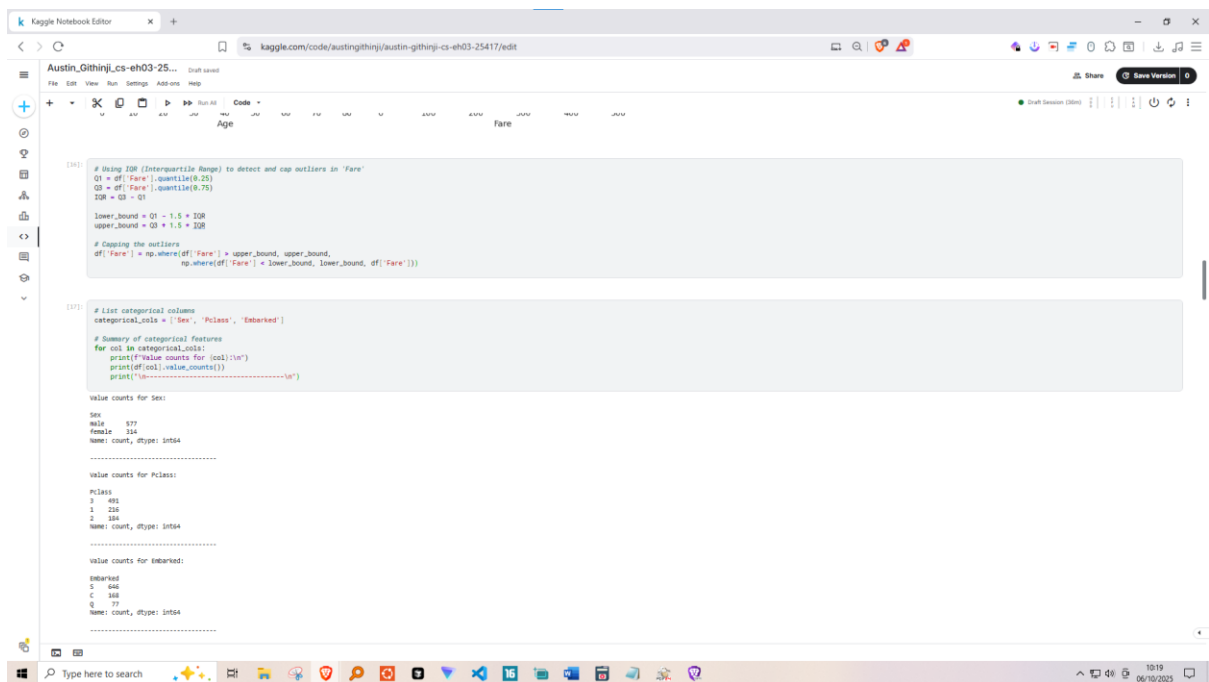
# Check again for missing values
df.isnull().sum()
```

```
[36]: PassengerId 0
Survived    0
Pclass      0
Name        0
Sex         0
Age         0
SibSp       0
ParCh       0
Ticket      0
Fare        0
Embarked    0
dtype: object
```

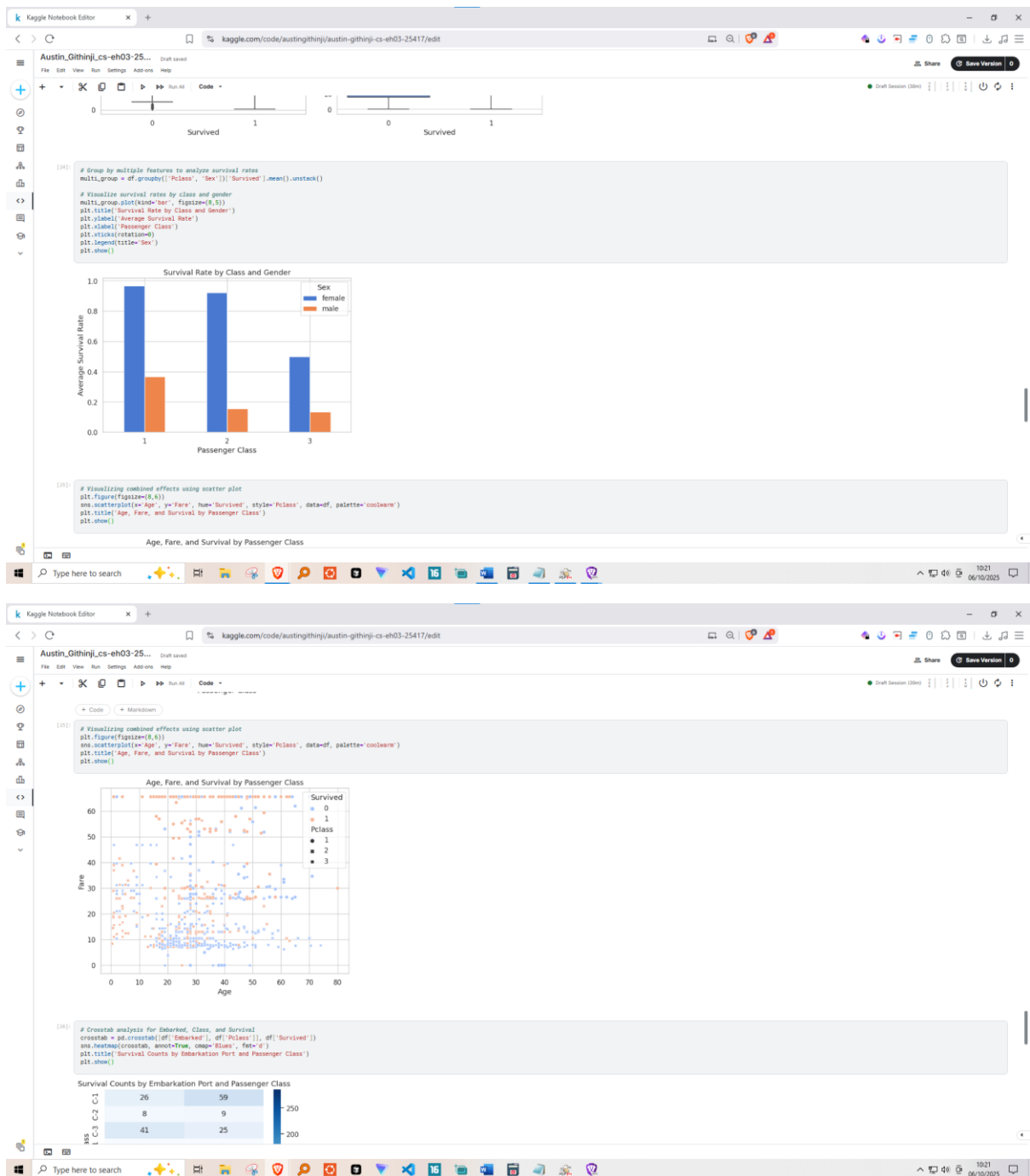
Univariate Analysis



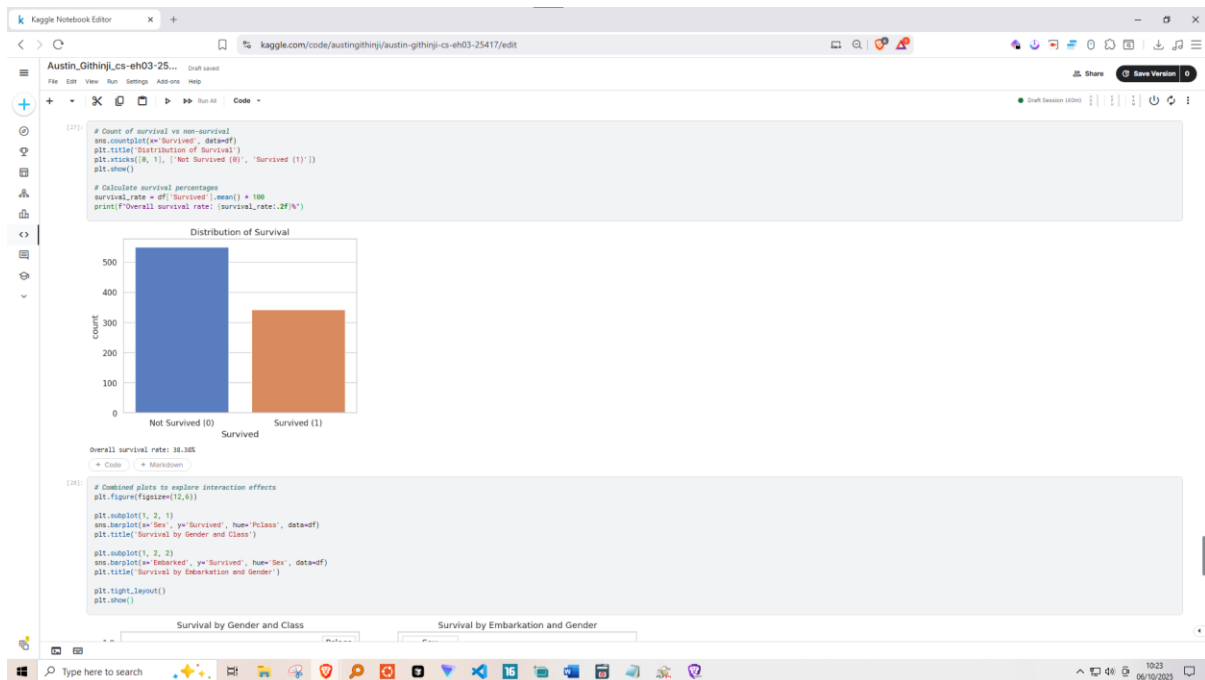
Bivariate Analysis



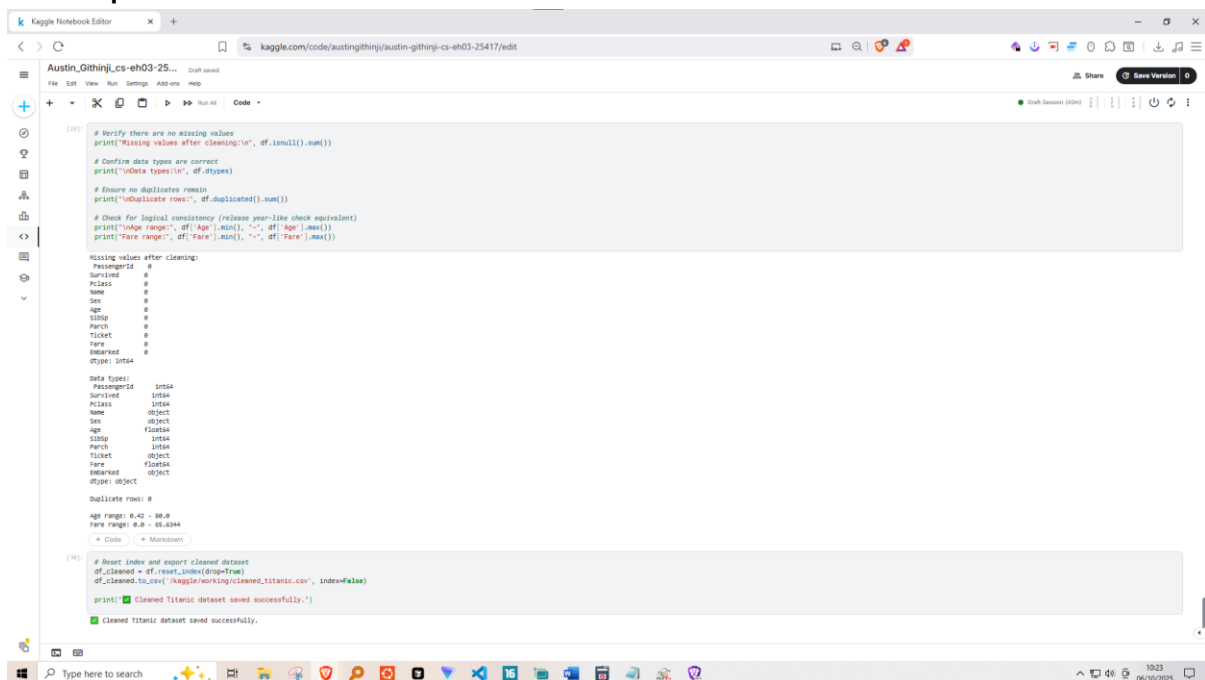
Multivariate Analysis



Target Variable Analysis:



Final Steps:



Link to Code: <https://www.kaggle.com/code/austingithinji/austin-githinji-cs-eh03-25417>

Conclusion

This EDA project successfully demonstrated the data wrangling and analytical process applied to the Titanic dataset.

Through a structured analysis, key patterns affecting survival were identified.

The dataset is now clean, validated, and ready for predictive modeling or deeper statistical inference.

EDA revealed that **social and economic factors** (gender and class) played a crucial role in survival chances.