Data and Artificial Intelligence
Cyber Shujaa Program

Week 3 Assignment
Exploratory Data Analysis

Student Name: Faith Jeptoo

Student ID: CS-DA02-25005

Table of Contents

Introduction

This assignment focuses on performing **Exploratory Data Analysis (EDA)** on the famous **Titanic dataset** from Kaggle.
The goal is to gain hands-on experience in exploring, cleaning, and visualizing data to uncover meaningful patterns and insights.

EDA involves:

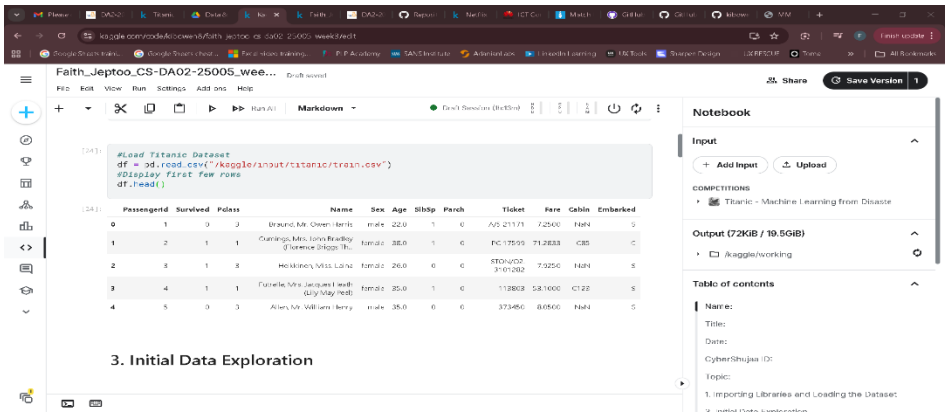
- Understanding the structure and quality of the data
- Handling missing values and outliers
- Performing univariate, bivariate, and multivariate analyses
- Exploring the target variable (**Survived**) to identify influencing factors

Tasks Completed

Step 1: Importing Libraries and Loading the Dataset

```
import pandas as pd                 # Data manipulation and analysis
import numpy as np                  # Numerical computations
import matplotlib.pyplot as plt     # Data visualization
import seaborn as sns               # Advanced plots
sns.set(style="whitegrid")

# Load the Titanic dataset
df = pd.read_csv("/kaggle/input/titanic/train.csv")# Display first few rows
df.head()
```



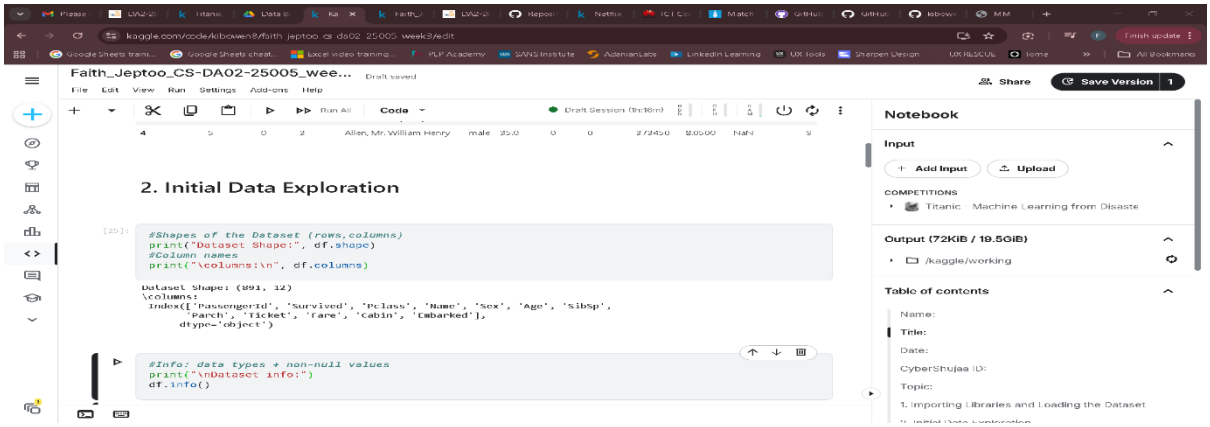
Step 2: Initial Data Exploration

#Shapes of the Dataset (rows,columns)

```
print("Dataset Shape:", df.shape)
```

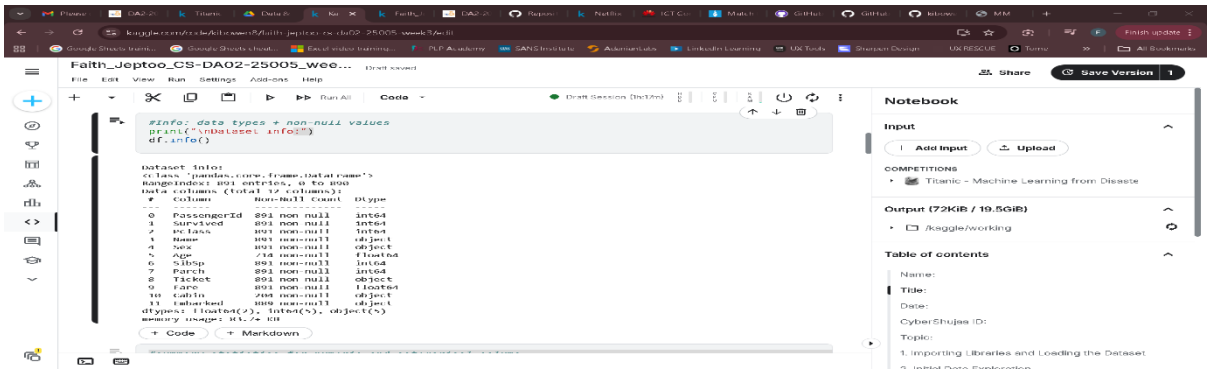
#Column names

```
print("\columns:\n", df.columns)
```



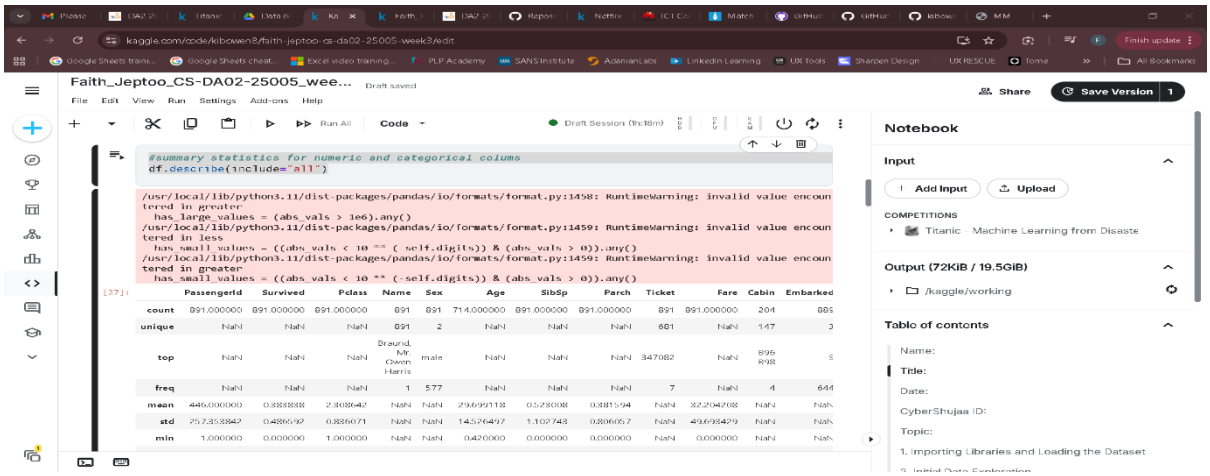
#Info: data types + non-null values

```
print("\nDataset info:")
df.info()
```



#summary statistics for numeric and categorical colums

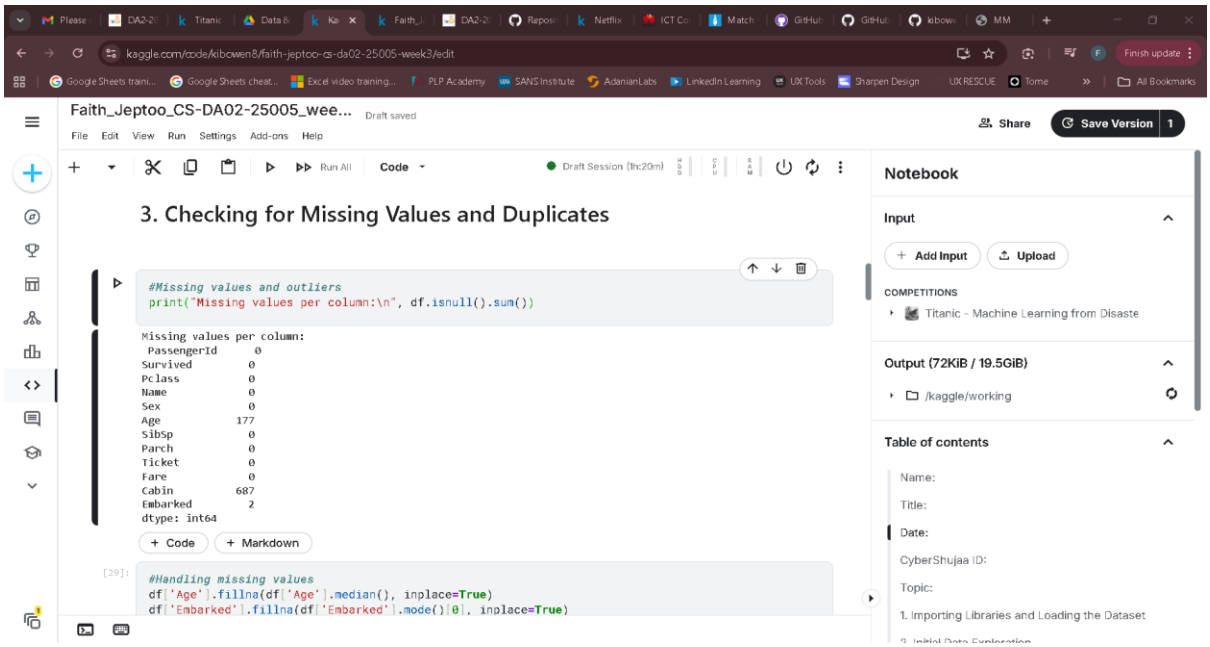
```
df.describe(include="all")
```



Step 3: Checking for Missing Values and Duplicates

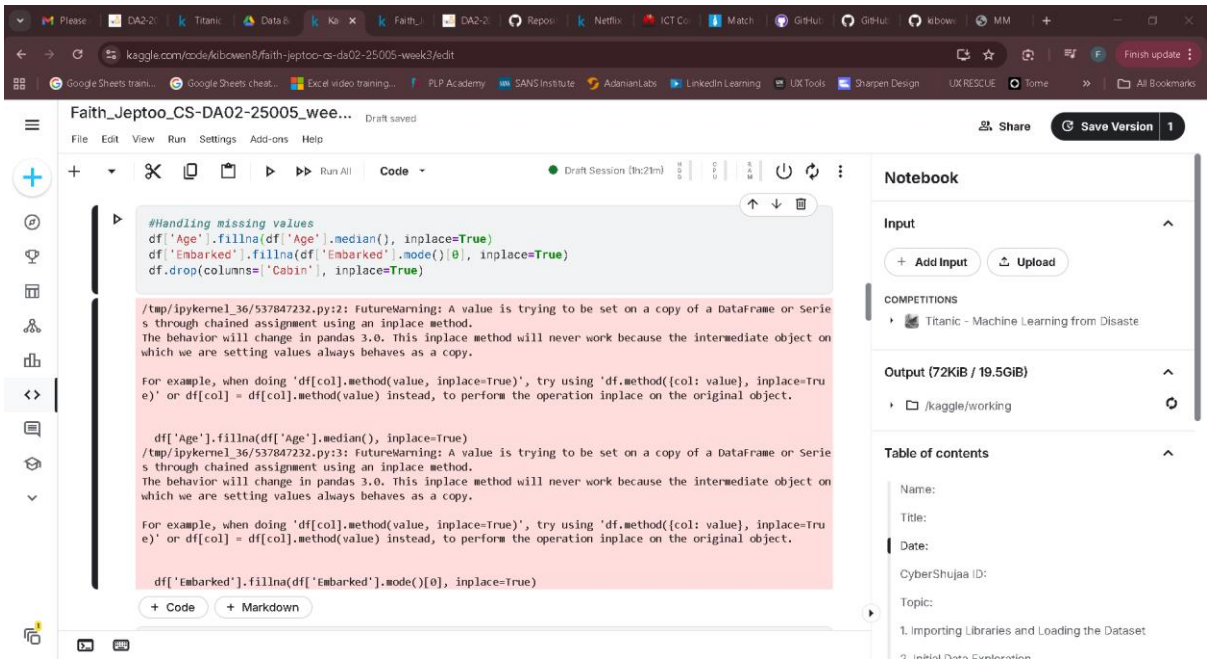
#Missing values and outliers

```
print("Missing values per column:\n", df.isnull().sum())
```

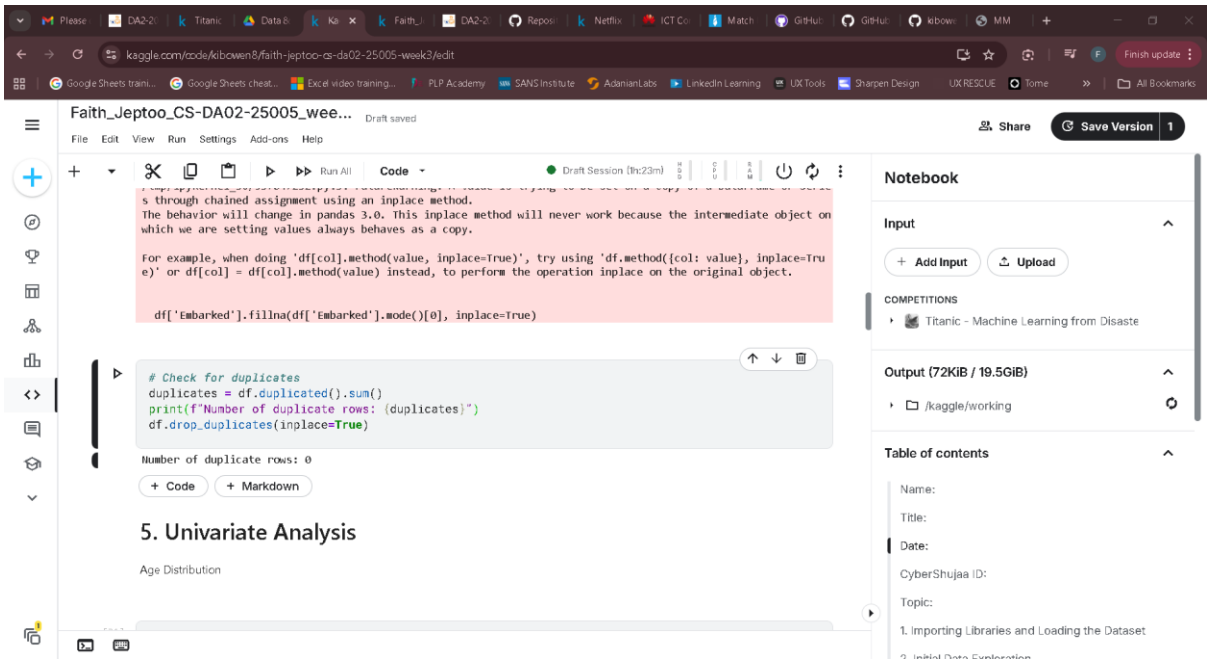


#Handling missing values

```
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
df.drop(columns=['Cabin'], inplace=True)
```



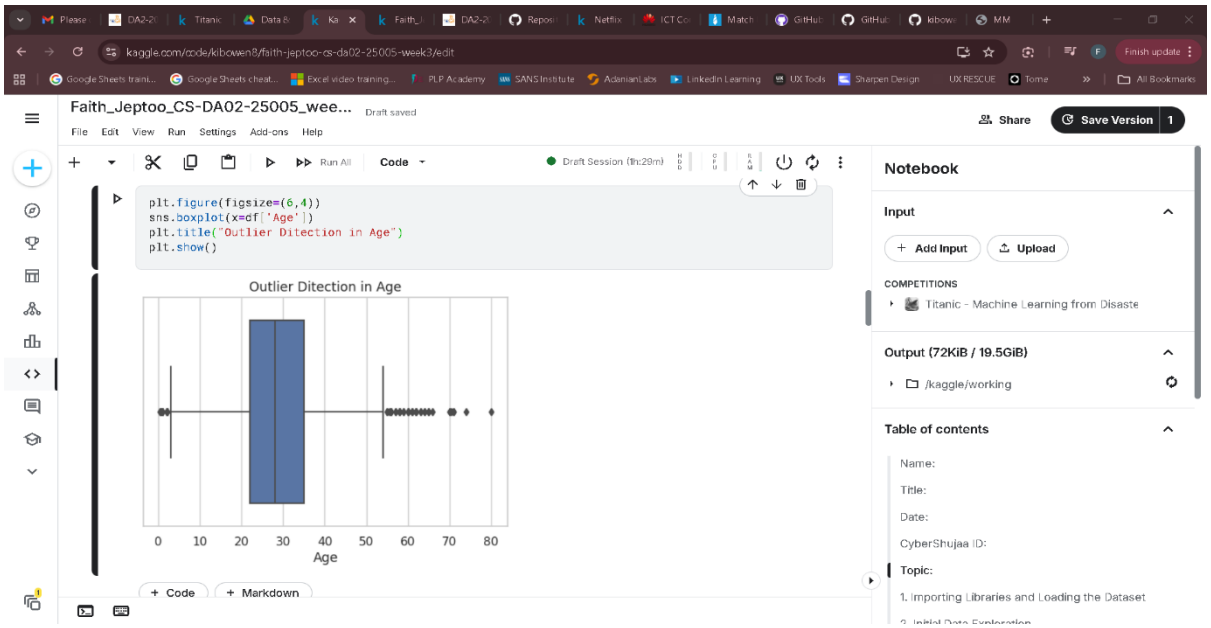
```
# Check for duplicates
duplicates = df.duplicated().sum()
print(f"Number of duplicate rows: {duplicates}")
df.drop_duplicates(inplace=True)
```



Step 4: Univariate Analysis

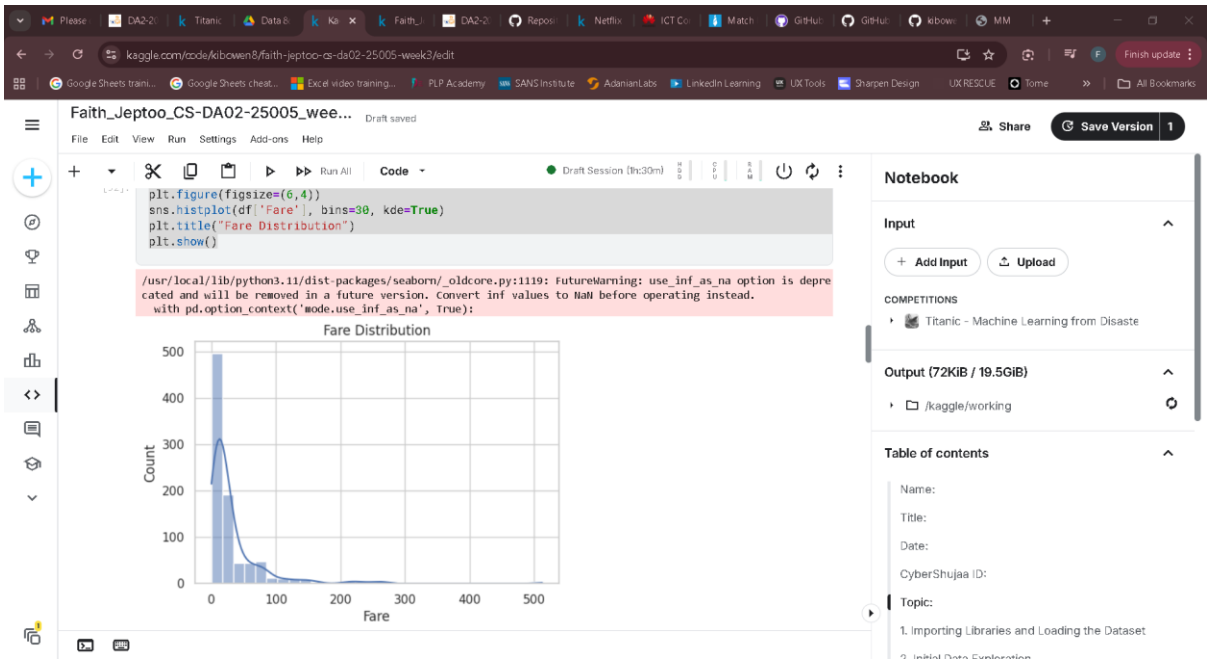
Age Distribution

```
plt.figure(figsize=(6,4))
sns.boxplot(x=df['Age'])
plt.title("Outlier Ditection in Age")
plt.show()
```



Fare Distribution

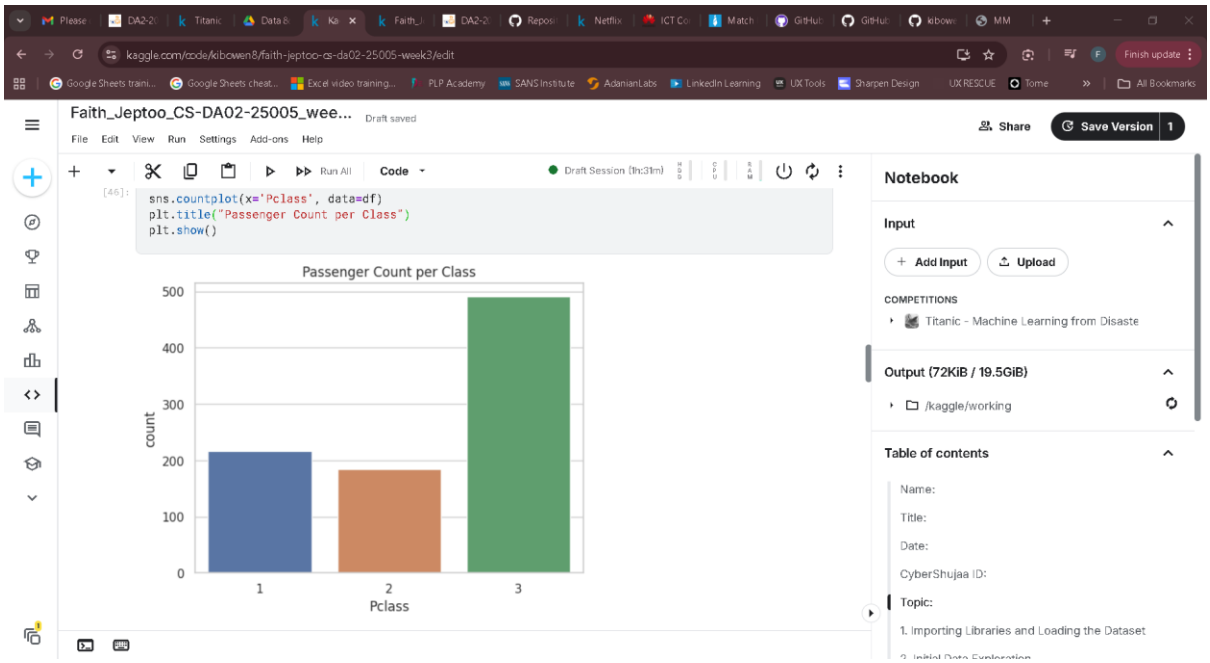
```
plt.figure(figsize=(6,4))
sns.histplot(df['Fare'], bins=30, kde=True)
plt.title("Fare Distribution")
plt.show()
```



Passenger Class, Gender, and Embarked

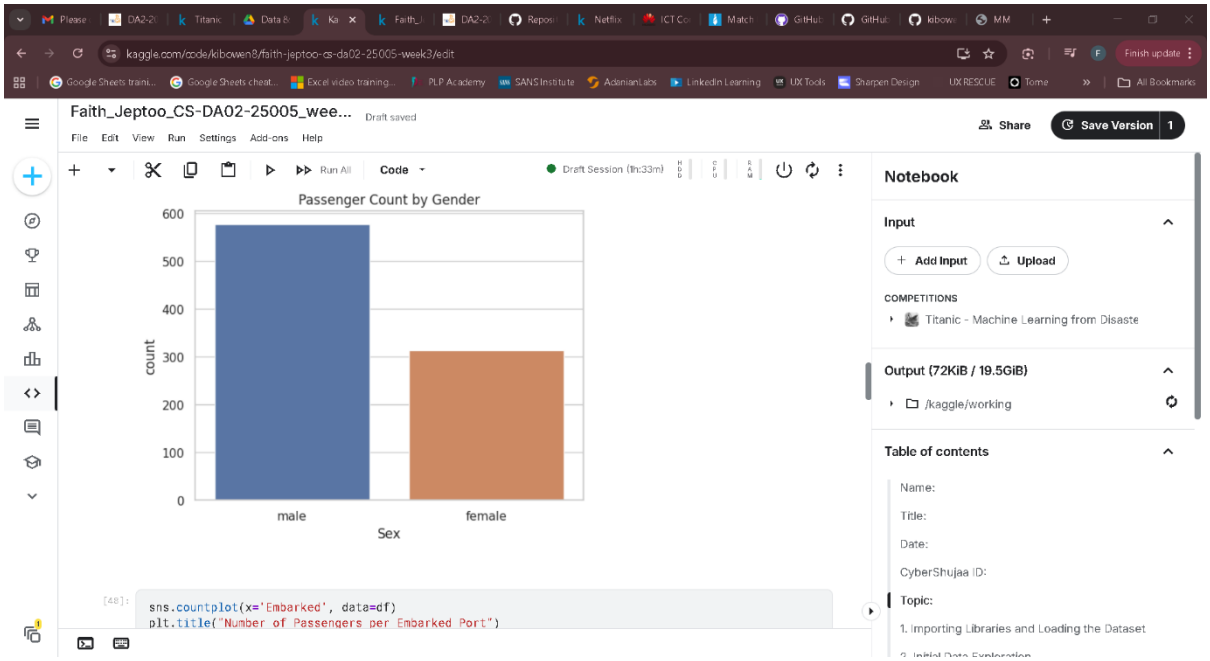
Class

```
sns.countplot(x='Pclass', data=df)
plt.title("Passenger Count per Class")
plt.show()
```



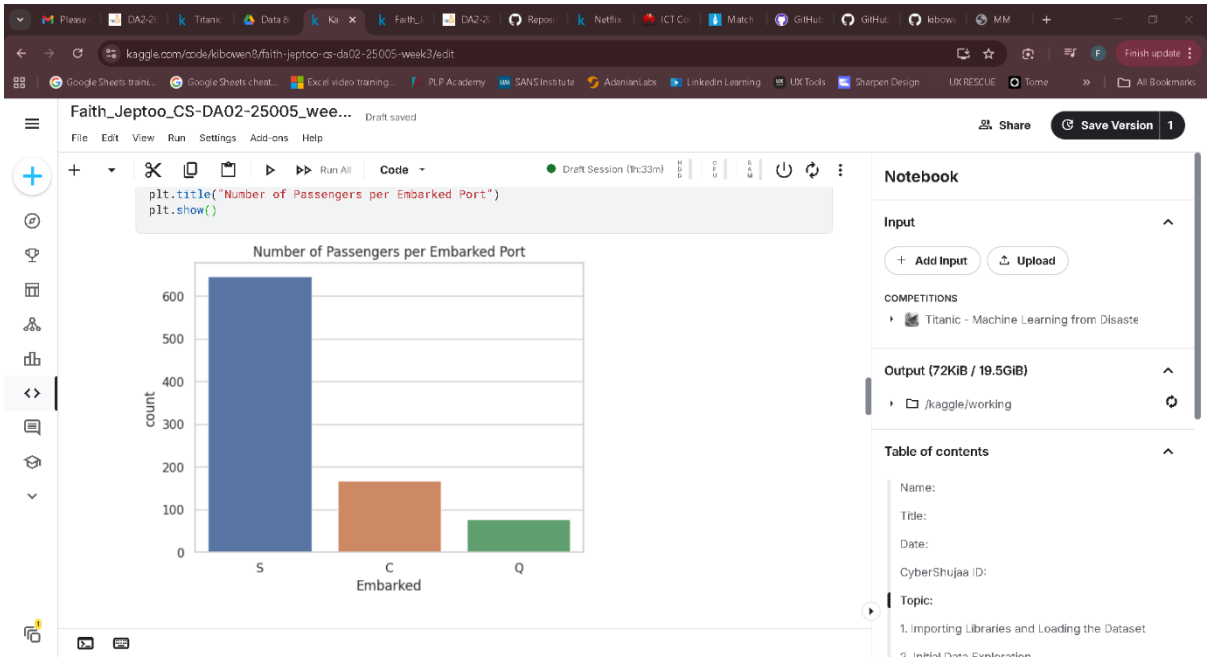
Gender

```
sns.countplot(x='Sex', data=df)
plt.title("Passenger Count by Gender")
plt.show()
```



Embarked

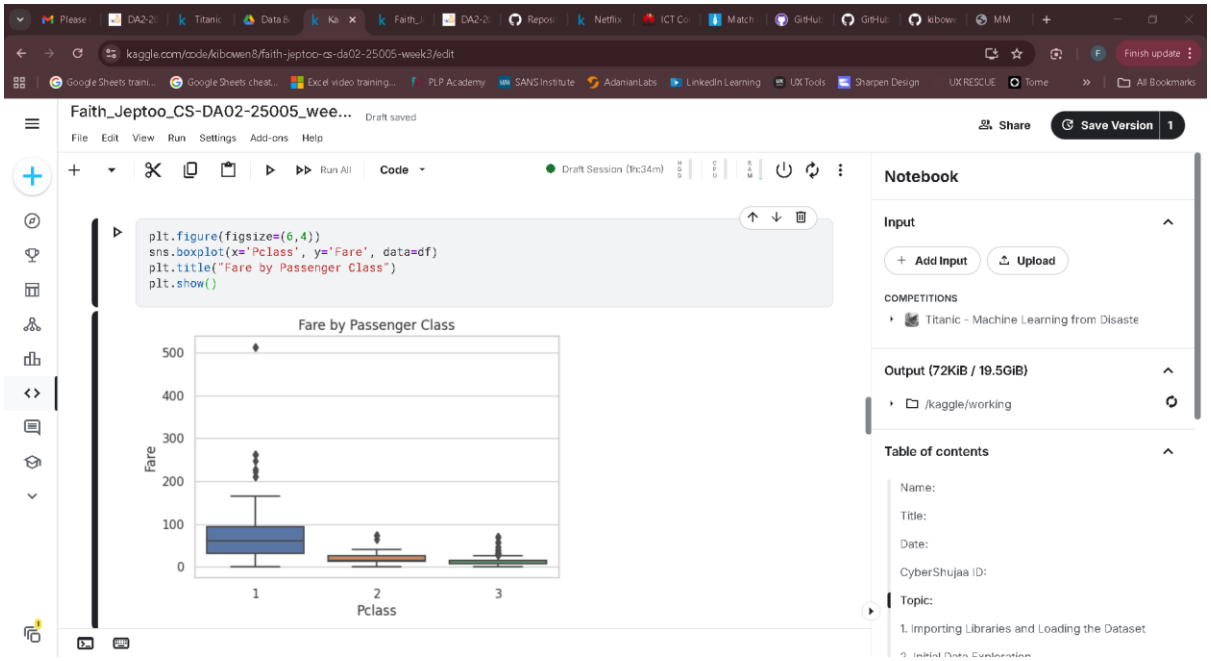
```
sns.countplot(x='Embarked', data=df)
plt.title("Number of Passengers per Embarked Port")
plt.show()
```



Step 5: Bivariate Analysis

Fare vs Pclass

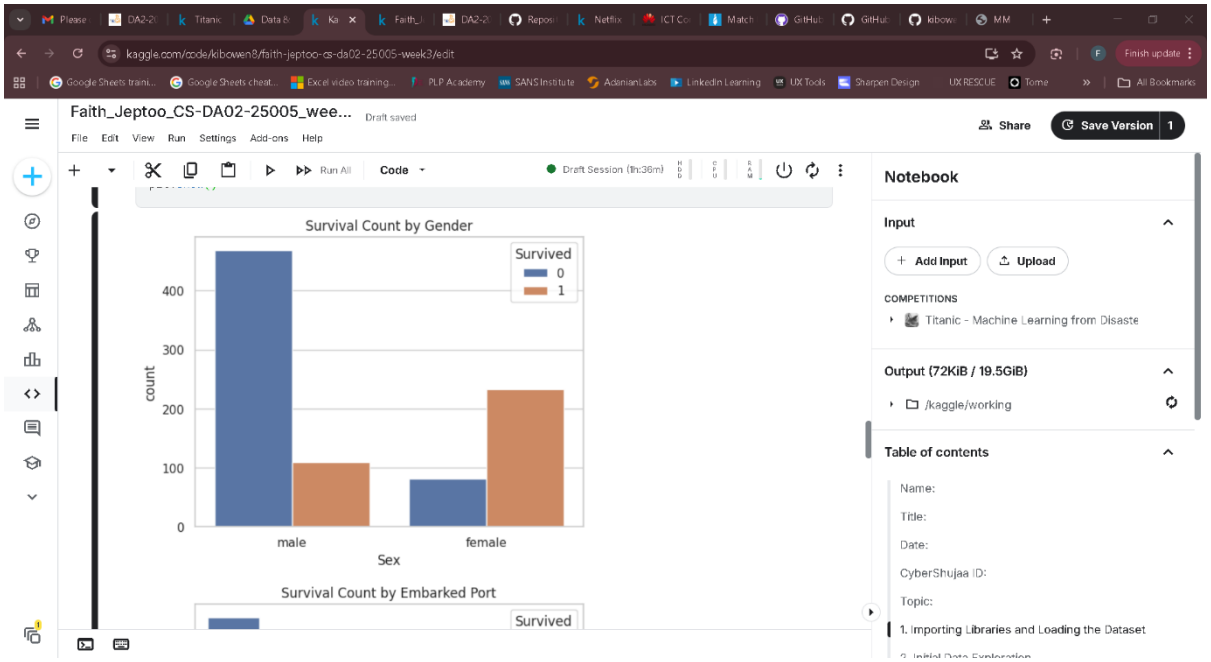
```
plt.figure(figsize=(6,4))
sns.boxplot(x='Pclass', y='Fare', data=df)
plt.title("Fare by Passenger Class")
plt.show()
```



Survival by Gender and Embarked

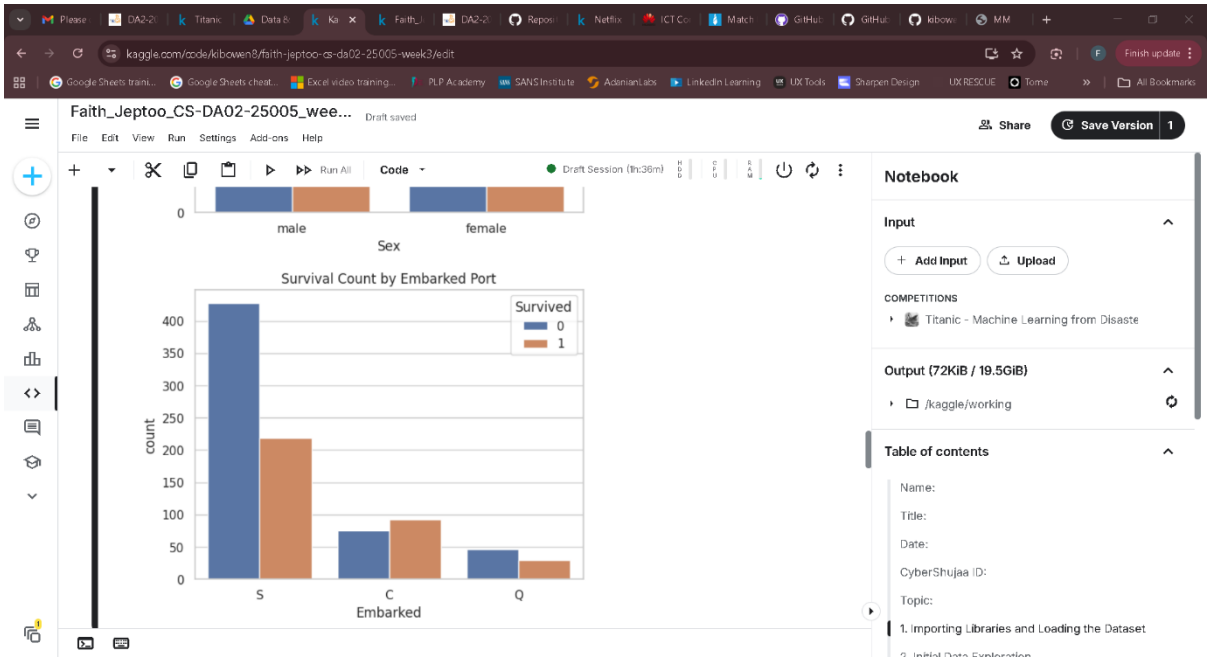
Gender

```
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title("Survival Count by Gender")
plt.show()
```



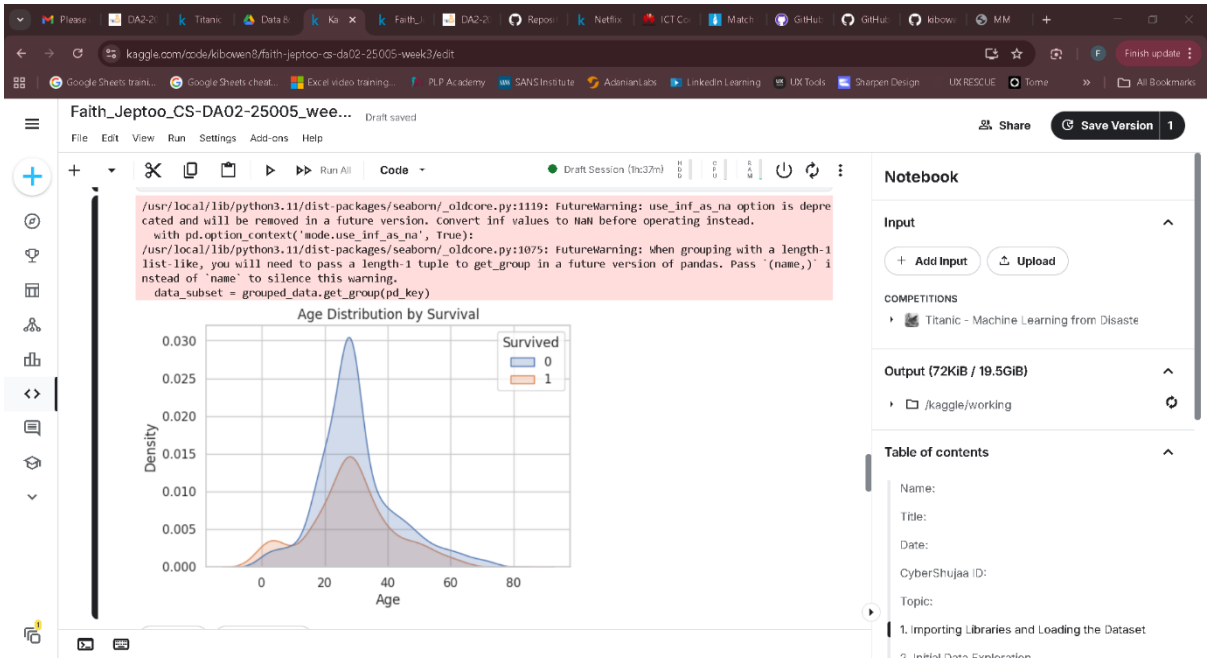
Emberked

```
sns.countplot(x='Embarked', hue='Survived', data=df)
plt.title("Survival Count by Embarked Port")
plt.show()
```



Age and Survival

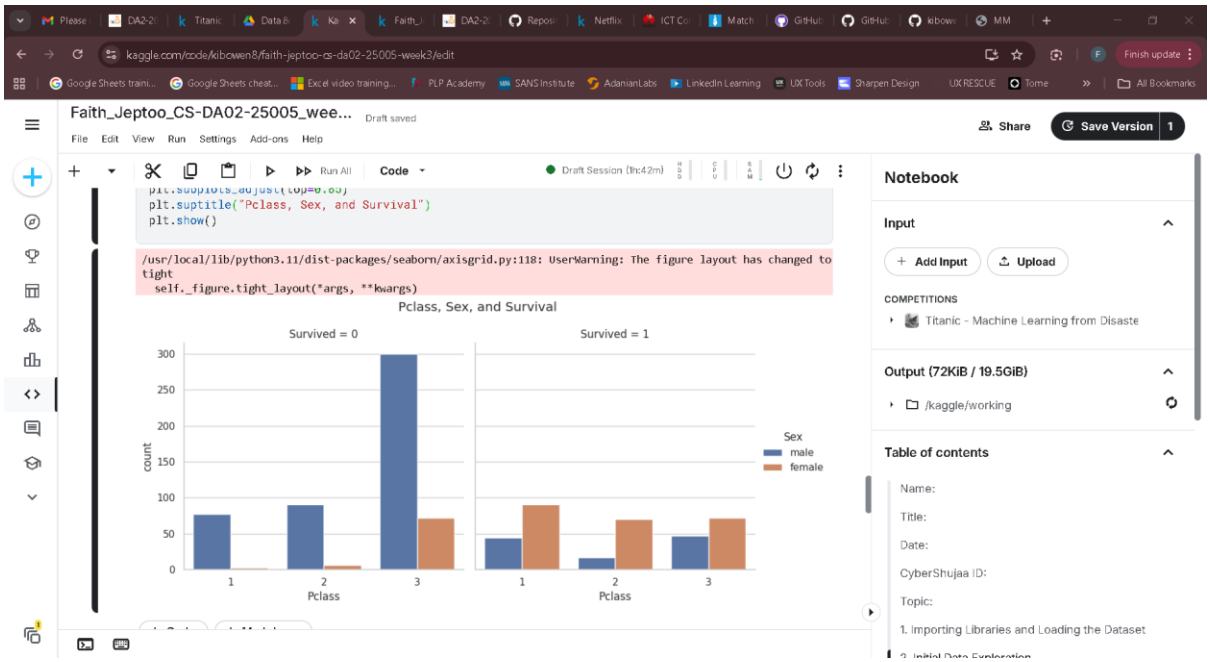
```
plt.figure(figsize=(6,4))
sns.kdeplot(data=df, x='Age', hue='Survived', fill=True)
plt.title("Age Distribution by Survival")
plt.show()
```



Step 6: Multivariate Analysis

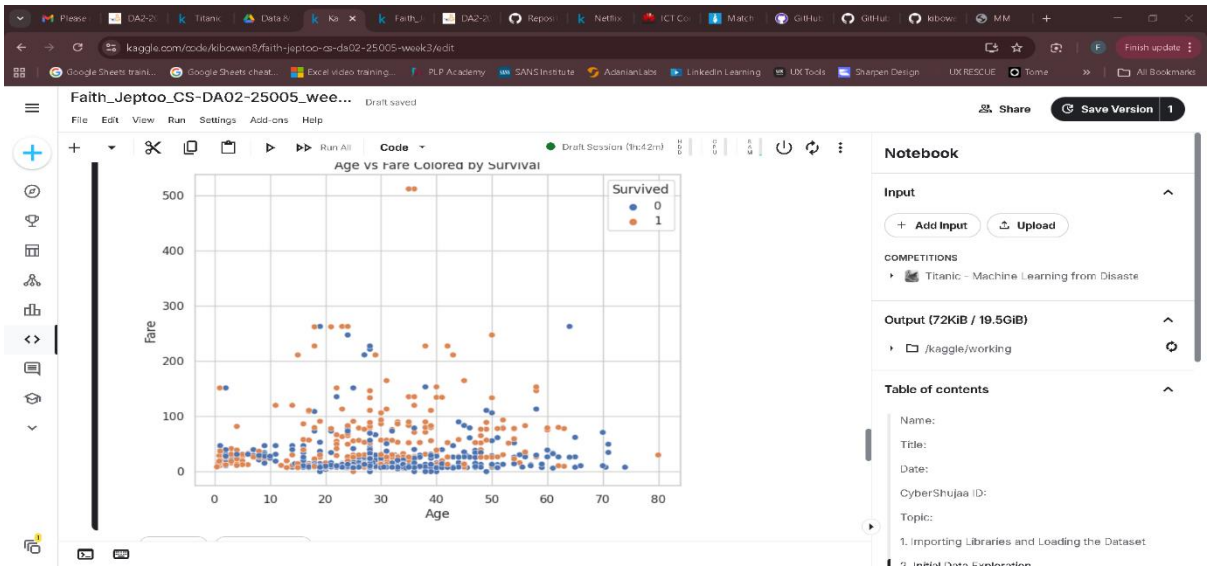
Class, Gender, and Survival

```
sns.catplot(x='Pclass', hue='Sex', col='Survived', data=df, kind='count')
plt.subplots_adjust(top=0.85)
plt.suptitle("Pclass, Sex, and Survival")
plt.show()
```



Age, Fare, and Survival

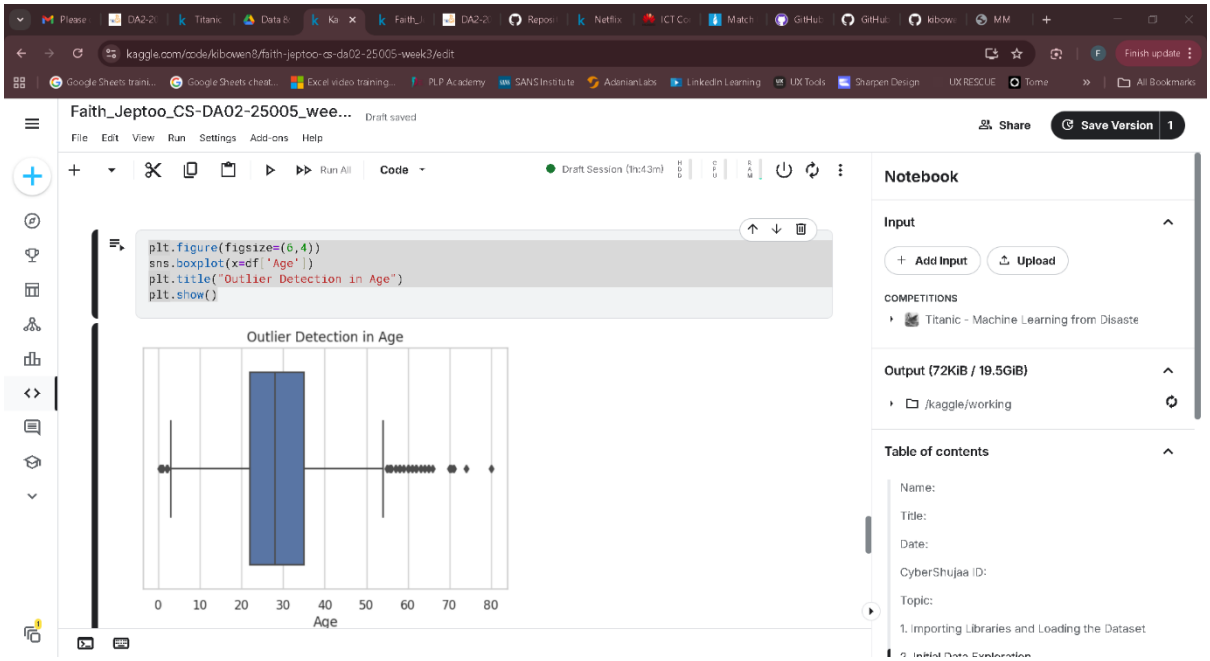
```
plt.figure(figsize=(8,6))
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
plt.title("Age vs Fare Colored by Survival")
plt.show()
```



Step 7: Outlier Detection and Handling

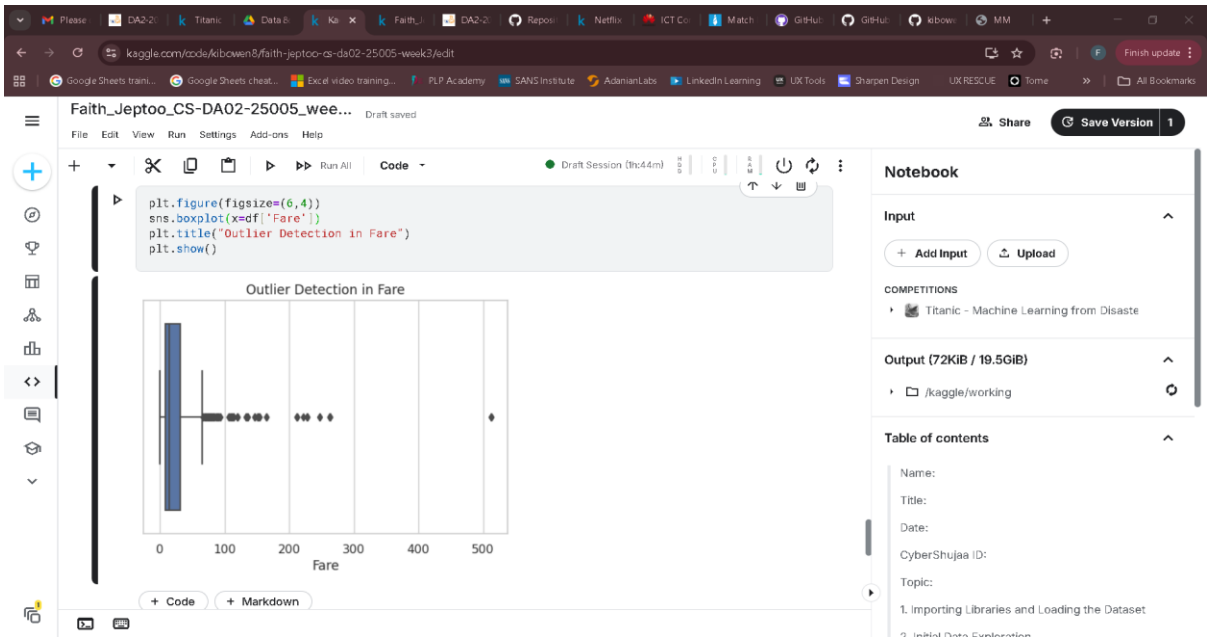
Age Outliers

```
plt.figure(figsize=(6,4))
sns.boxplot(x=df['Age'])
plt.title("Outlier Detection in Age")
plt.show()
```



Fare Outliers

```
plt.figure(figsize=(6,4))
sns.boxplot(x=df['Fare'])
plt.title("Outlier Detection in Fare")
plt.show()
```

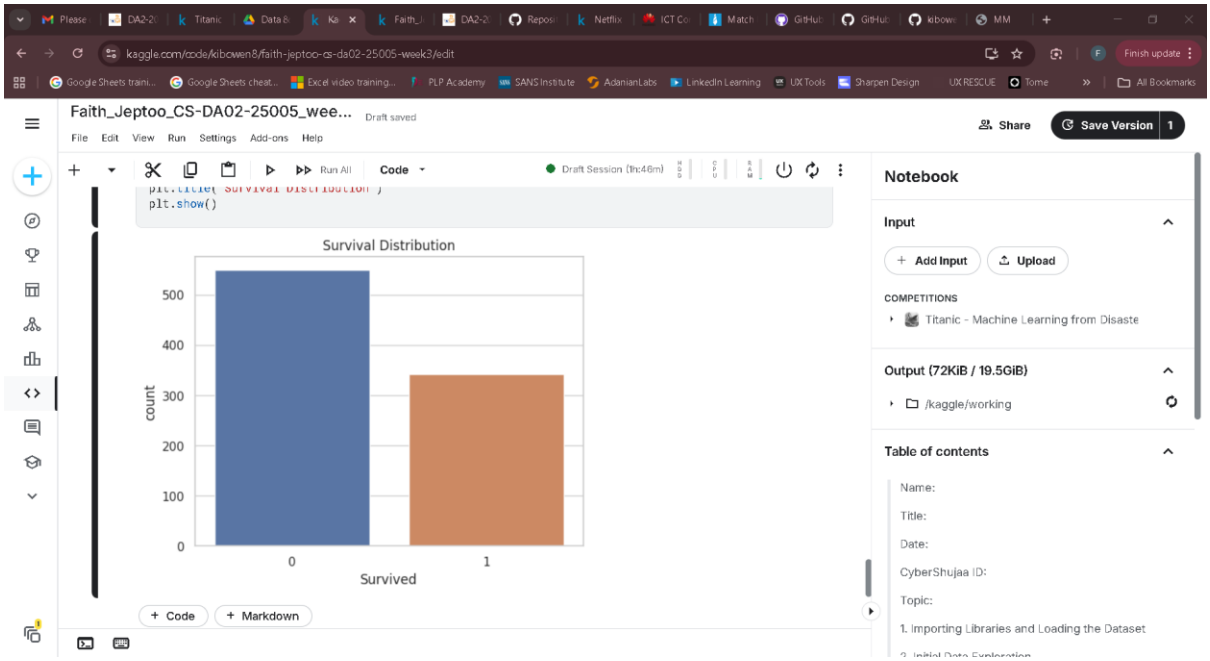


 # Cap extreme fares at 99th percentile

```
fare_cap = df['Fare'].quantile(0.99)
df['Fare'] = np.where(df['Fare'] > fare_cap, fare_cap, df['Fare'])
```

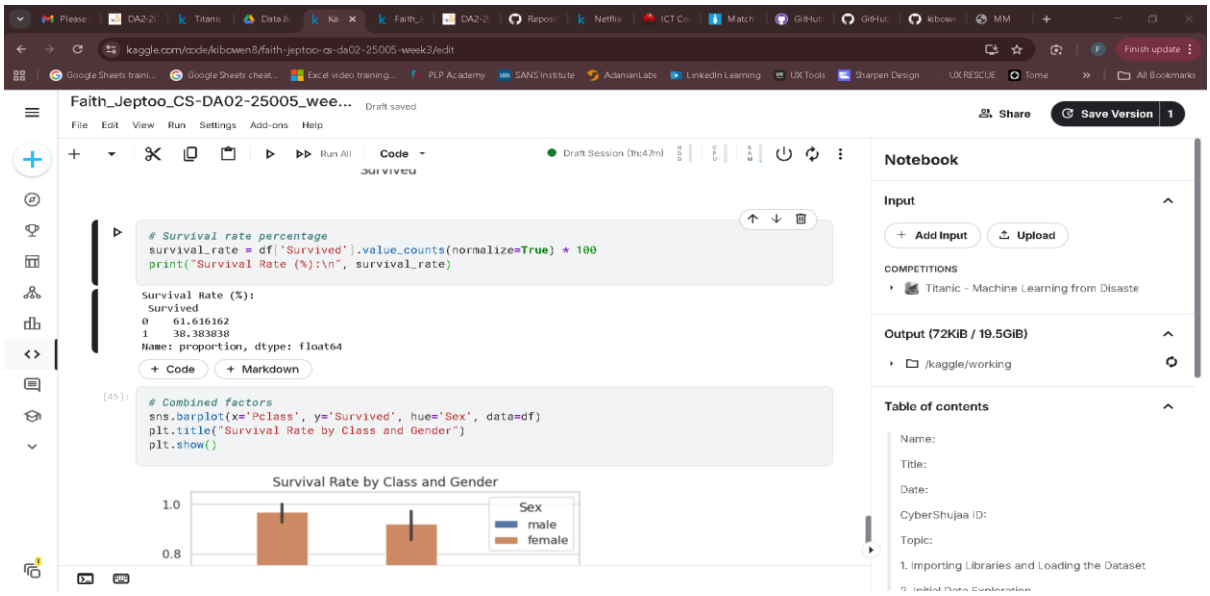
Step 8: Target Variable (Survived) Analysis

```
sns.countplot(x='Survived', data=df)
plt.title("Survival Distribution")
plt.show()
```

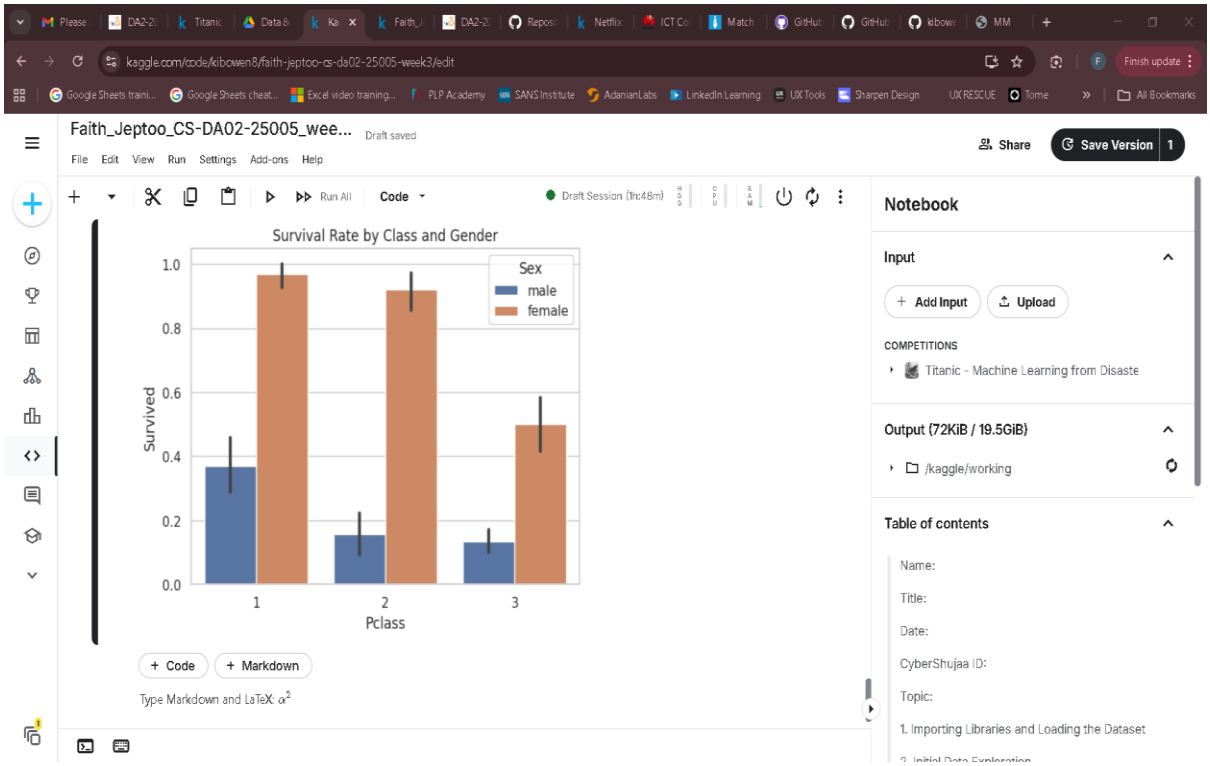


 # Survival rate percentage

```
survival_rate = df['Survived'].value_counts(normalize=True) * 100
print("Survival Rate (%):\n", survival_rate)
```



```
# Combined factors
sns.barplot(x='Pclass', y='Survived', hue='Sex', data=df)
plt.title("Survival Rate by Class and Gender")
plt.show()
```



Conclusion

The Exploratory Data Analysis on the Titanic dataset revealed the following insights:

- The dataset contained missing values in *Age*, *Cabin*, and *Embarked* which were handled using imputation and removal.
- Most passengers were male and from 3rd class.
- Fare and Age were both right-skewed distributions.
- Survival was highly influenced by **Gender**, **Class**, and **Fare**.
- Outliers in *Fare* were capped to minimize skewness.
- The target variable (*Survived*) showed moderate imbalance, which is important for modeling

Link To Notebook

https://www.kaggle.com/code/kibowen8/faith-jeptoo-cs-da02-25005-week-3