

Data and Artificial Intelligence

Cyber Shujaa Program

Week 3 Assignment

Titanic Exploratory Data Analysis

Student Name: Salome Kungu

Student ID: CS-DA03-26054

Contents

Data and Artificial Intelligence	1
Cyber Shujaa Program.....	1
Week 3 Assignment Titanic Exploratory Data Analysis	1
Introduction	2
The key objectives of the assignment were to:.....	2
1. Initial Data Exploration.....	2
2. Handling Missing Values and Outliers.....	2
3. Univariate Analysis.....	2
4. Bivariate Analysis	2
5. Multivariate Analysis.....	2
6. Target Variable Analysis	2
Tasks Completed	2
Conclusion.....	3
Question 1: Load the Dataset and	3
Question 2: Univariate Analysis.	7
Question 3: Bivariate Analysis.....	11
Question 5: Outlier detection and handling:	19
Question 6: Target Variable Exploration.....	20

Introduction

The objective of this assignment was to develop hands-on experience in **Exploratory Data Analysis (EDA)** using a real-world dataset sourced from Kaggle. The Titanic dataset was used to systematically explore, clean, and analyze passenger data in order to understand patterns, relationships, and factors influencing survival outcomes.

The assignment focused on applying core EDA techniques including initial data exploration, handling missing values and outliers, univariate, bivariate, and multivariate analysis, as well as in-depth target variable exploration. Using Python and the Pandas, Matplotlib, and Seaborn libraries within a Kaggle Notebook environment enabled reproducible and well-documented analysis.

Through this exercise, practical skills were developed in data profiling, visualization, statistical reasoning, and interpretation of insights, all of which are essential for data-driven decision-making and machine learning preparation.

The key objectives of the assignment were to:

1. Initial Data Exploration
2. Handling Missing Values and Outliers
3. Univariate Analysis
4. Bivariate Analysis
5. Multivariate Analysis
6. Target Variable Analysis

Tasks Completed

The assignment began with loading the Titanic dataset from Kaggle into a Pandas DataFrame and conducting an initial exploration to understand its structure, size, and feature composition. Key exploratory functions were used to inspect data types, identify missing values, detect duplicate records, and generate summary statistics for numerical features.

Subsequently, data quality issues were addressed by handling missing values in features such as Age and Embarked using appropriate imputation techniques, while features with excessive missing data were treated cautiously. Outlier detection was performed on numerical variables like Fare, and decisions were made to retain outliers where they represented meaningful real-world variations.

Univariate analysis was carried out to examine the distribution of individual variables such as age, passenger class, embarkation point, and fare. This was followed by bivariate analysis to explore relationships between survival and key features including gender, passenger class, age, and embarkation location. Multivariate analysis further examined the combined effects of multiple variables to uncover interaction patterns influencing survival outcomes.

Finally, the target variable **Survived** was analyzed to assess class balance and identify factors associated with higher or lower survival rates. Relevant visualizations were created to support findings, and the complete analysis was documented and published in a Kaggle Notebook with supporting screenshots as evidence of task completion.

Link to Code: <https://www.kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-titanic-eda-report>

Conclusion

This assignment provided a comprehensive introduction to Exploratory Data Analysis using a real-world dataset. By systematically exploring and visualizing the Titanic dataset, meaningful insights were uncovered regarding passenger survival patterns and influencing factors such as gender, class, and age.

The exercise strengthened practical skills in data cleaning, visualization, statistical reasoning, and analytical storytelling. These competencies are foundational for advanced data analytics and machine learning workflows. The completed Kaggle Notebook and report also serve as valuable additions to my professional data science portfolio.

Screenshots of the project

The screenshots illustrate the Exploratory Data Analysis conducted on the Titanic dataset, including data cleaning, handling missing values, and univariate and multivariate visualizations. These steps enabled the identification of survival trends across demographic and socioeconomic variables, supporting data-driven insights and model readiness.

Question 1: Load the Dataset and Initial Data Exploration

Load the Dataset:

What:

Import the Titanic CSV file using `pd.read_csv()` and display the first few rows using `df.head()`.

Why:

Loading the dataset into Pandas provides a structured format for analysis. Viewing the first rows helps verify successful loading and gives an initial understanding of the data.

Initial Data Exploration:

What:

Use functions such as:

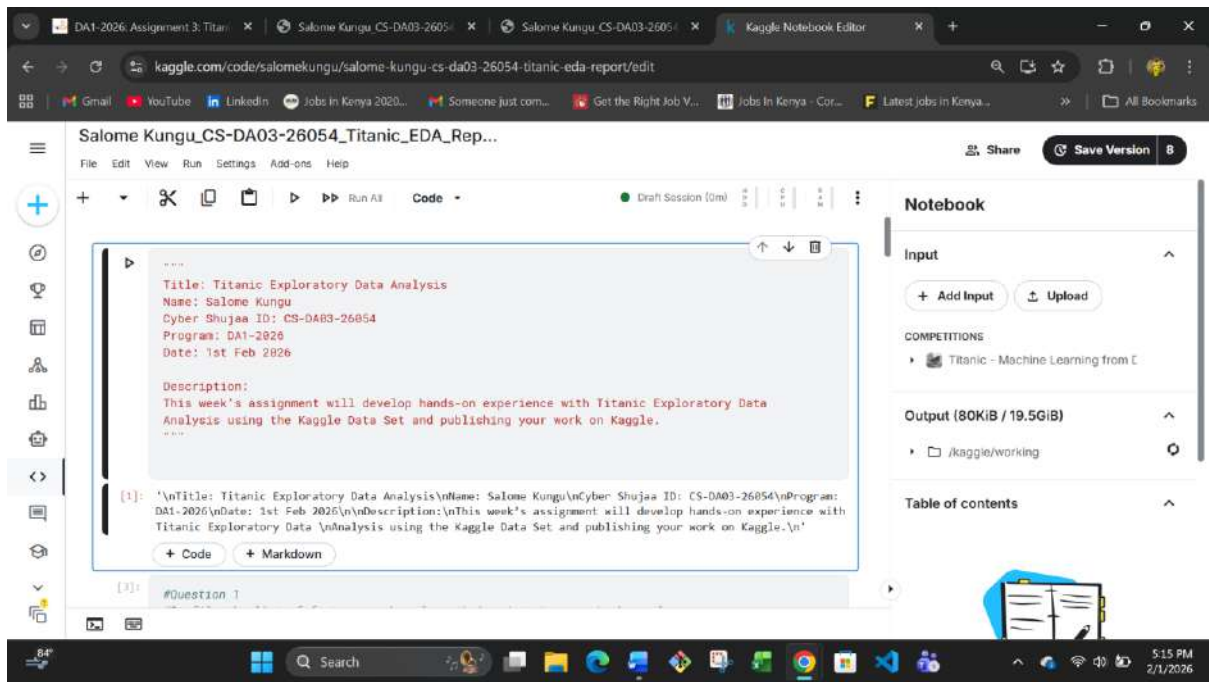
- `df.shape`
- `df.info()`
- `df.describe()`
- `df.columns`
- `df.nunique()`
- `df.duplicated().sum()`

Why:

This step helps identify:

- The size of the dataset
- Data types of each feature
- Missing values and duplicates
- Numerical distributions and potential anomalies

Understanding these aspects early prevents incorrect assumptions during analysis.



DAI-2026 Assignment 3: Titanic | Salome Kungu CS-DA03-26054 | Salome Kungu CS-DA03-26054 | Kaggle Notebook Editor

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-titanic-eda-report/edit

Salome Kungu_CS-DA03-26054_Titanic_EDA_Rep...

File Edit View Run Settings Add-ons Help

Share Save Version 8

Code

Draft Session (0m)

Notebook

Input

+ Add Input Upload

COMPETITIONS

Titanic - Machine Learning from E

Output (80KiB / 19.5GiB)

/kaggle/working

Table of contents

[1]:

```

Title: Titanic Exploratory Data Analysis
Name: Salome Kungu
Cyber Shujaa ID: CS-DA03-26054
Program: DAI-2026
Date: 1st Feb 2026

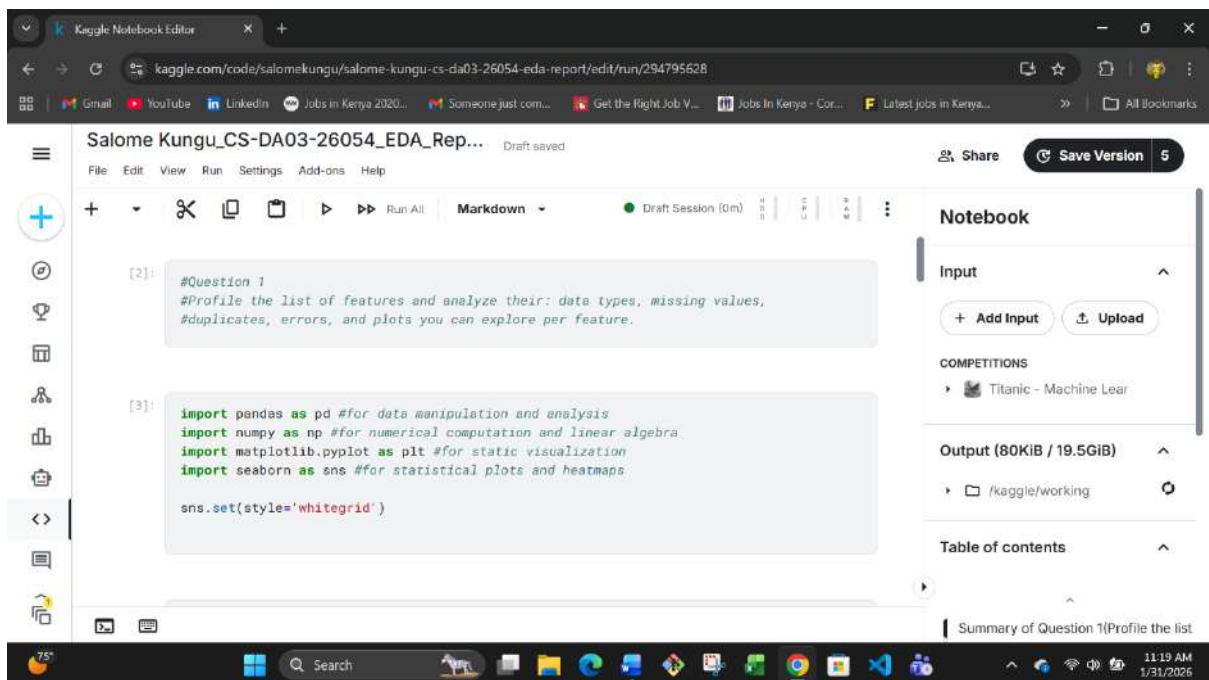
Description:
This week's assignment will develop hands-on experience with Titanic Exploratory Data
Analysis using the Kaggle Data Set and publishing your work on Kaggle.

```

+ Code + Markdown

[1]: #Question 1

64° Search 5:15 PM 2/1/2026



Kaggle Notebook Editor

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit/run/294795628

Salome Kungu_CS-DA03-26054_EDA_Rep... Draft saved

File Edit View Run Settings Add-ons Help

Share Save Version 5

Markdown

Draft Session (0m)

Notebook

Input

+ Add Input Upload

COMPETITIONS

Titanic - Machine Lear

Output (80KiB / 19.5GiB)

/kaggle/working

Table of contents

[2]:

```

#Question 1
#Profile the list of features and analyze their: data types, missing values,
#duplicates, errors, and plots you can explore per feature.

```

[3]:

```

import pandas as pd #for data manipulation and analysis
import numpy as np #for numerical computation and linear algebra
import matplotlib.pyplot as plt #for static visualization
import seaborn as sns #for statistical plots and heatmaps

sns.set(style='whitegrid')

```

Summary of Question 1(Profile the list

75° Search 11:19 AM 1/31/2026

Kaggle Notebook Editor

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit/run/294795628

Salome Kungu_CS-DA03-26054_EDA_Rep... Draft saved

File Edit View Run Settings Add-ons Help

Run All Markdown Draft Session (1m)

```
[4]: import os
os.listdir('/kaggle/input')
!ls /kaggle/input/titanic

gender_submission.csv test.csv train.csv
```

```
[5]: #Load Titanic dataset

df = pd.read_csv('/kaggle/input/titanic/train.csv')
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S

Summary of Question 1(Profile the list

Kaggle Notebook Editor

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit/run/294795628

Salome Kungu_CS-DA03-26054_EDA_Rep... Draft saved

File Edit View Run Settings Add-ons Help

Run All Markdown Draft Session (1m)

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Summary of Question 1(Profile the list

Kaggle Notebook Editor

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit/run/294795628

Salome Kungu_CS-DA03-26054_EDA_Rep... Draft saved

File Edit View Run Settings Add-ons Help

Run All Markdown Draft Session (1m)

```
[6]:
# shape and info

df.shape
df.info()
df.describe()
df.columns
df.nunique()
df.duplicated().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  -
 0   PassengerId    891 non-null    int64
 1   Survived       891 non-null    int64
 2   Pclass        891 non-null    int64
 3   Name           891 non-null    object
 4   Sex            891 non-null    object
 5   Age            714 non-null    float64
 6   SibSp          891 non-null    int64
 7   Parch          891 non-null    int64
 8   Ticket         891 non-null    object
 9   Fare           891 non-null    float64
10   Cabin         204 non-null    object
11   Embarked       889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Notebook

Input

+ Add Input Upload

COMPETITIONS

Titanic - Machine Lear

Output (80KiB / 19.5GiB)

/kaggle/working

Table of contents

Summary of Question 1(Profile the list

75° Search 11:20 AM 1/31/2026

Kaggle Notebook Editor

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit/run/294795628

Salome Kungu_CS-DA03-26054_EDA_Rep... Draft saved

File Edit View Run Settings Add-ons Help

Run All Markdown Draft Session (1m)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  -
 0   PassengerId    891 non-null    int64
 1   Survived       891 non-null    int64
 2   Pclass        891 non-null    int64
 3   Name           891 non-null    object
 4   Sex            891 non-null    object
 5   Age            714 non-null    float64
 6   SibSp          891 non-null    int64
 7   Parch          891 non-null    int64
 8   Ticket         891 non-null    object
 9   Fare           891 non-null    float64
10   Cabin         204 non-null    object
11   Embarked       889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Notebook

Input

+ Add Input Upload

COMPETITIONS

Titanic - Machine Lear

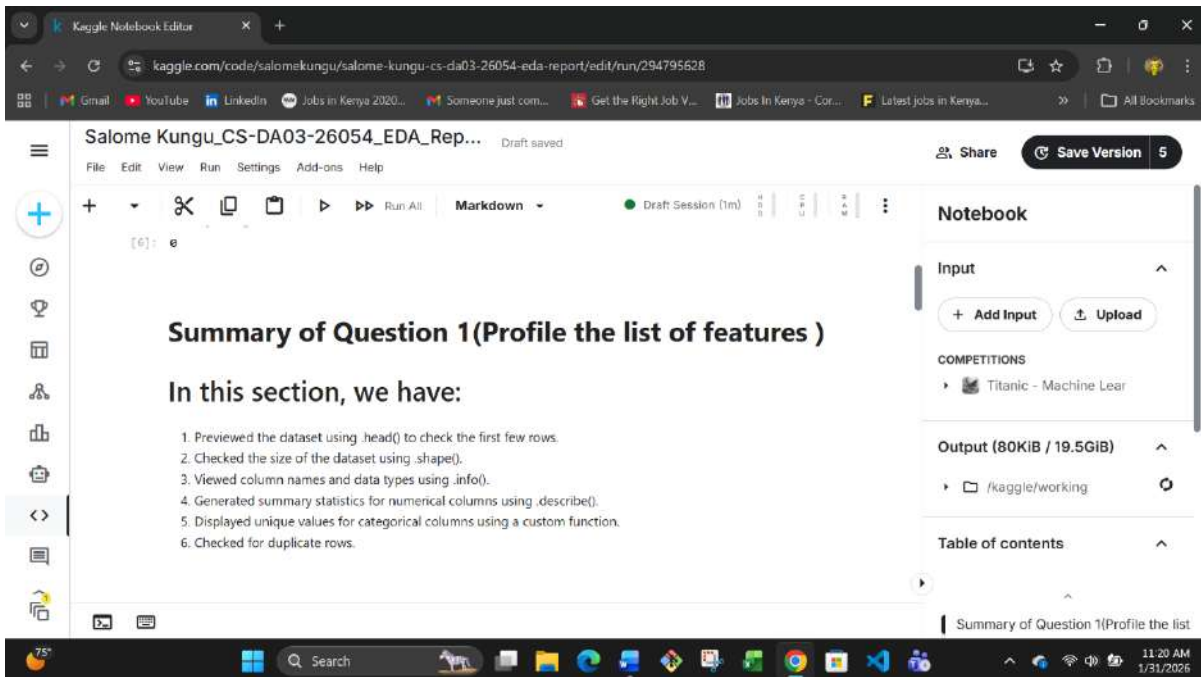
Output (80KiB / 19.5GiB)

/kaggle/working

Table of contents

Summary of Question 1(Profile the list

75° Search 11:20 AM 1/31/2026



Salome Kungu_CS-DA03-26054_EDA_Report Draft saved

File Edit View Run Settings Add-ons Help

Run All Markdown Draft Session (1m)

Summary of Question 1(Profile the list of features)

In this section, we have:

1. Previewed the dataset using `.head()` to check the first few rows.
2. Checked the size of the dataset using `.shape()`.
3. Viewed column names and data types using `.info()`.
4. Generated summary statistics for numerical columns using `.describe()`.
5. Displayed unique values for categorical columns using a custom function.
6. Checked for duplicate rows.

Notebook

Input

+ Add Input Upload

COMPETITIONS

Titanic - Machine Lear

Output (80KiB / 19.5GiB)

/kaggle/working

Table of contents

Summary of Question 1(Profile the list

Question 2: Univariate Analysis.

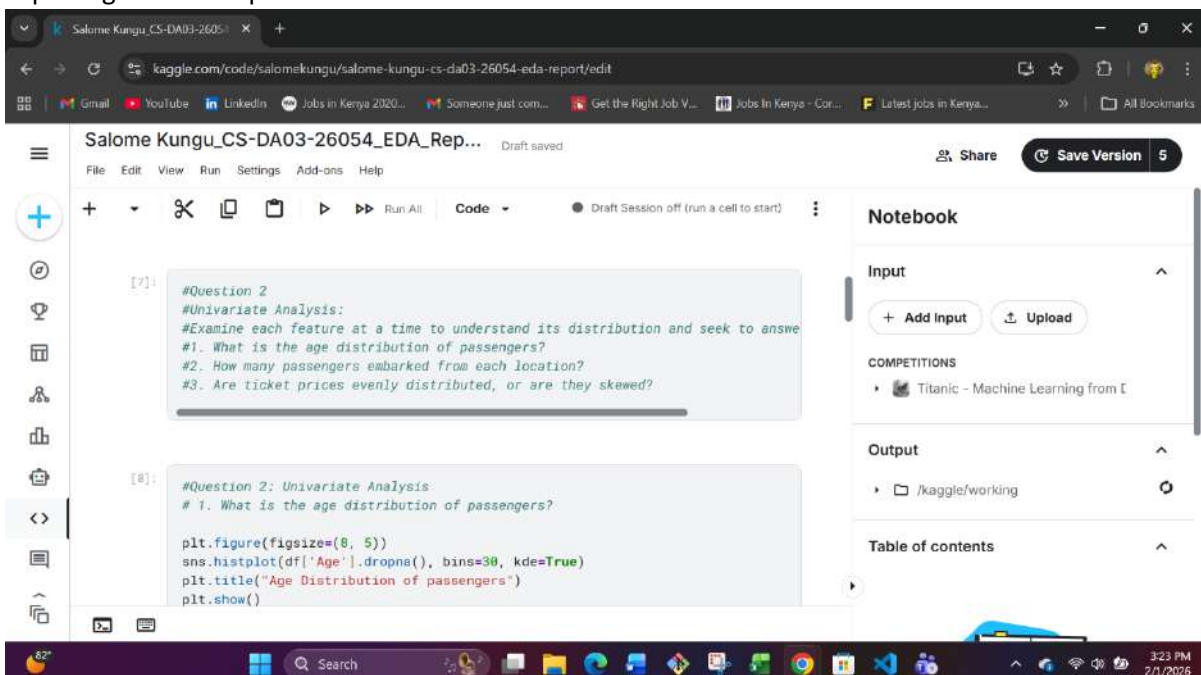
What:

Analyze individual features using histograms, count plots, and bar charts to answer questions such as:

- What is the age distribution of passengers?
- How many passengers embarked from each port?
- Is the Fare distribution skewed?

Why:

Univariate analysis provides a foundational understanding of each variable independently before exploring relationships.



Salome Kungu_CS-DA03-26054_EDA_Report Draft saved

File Edit View Run Settings Add-ons Help

Run All Code Draft Session off (run a cell to start)

```
[7]: #Question 2
#Univariate Analysis:
#Examine each feature at a time to understand its distribution and seek to answer
#1. What is the age distribution of passengers?
#2. How many passengers embarked from each location?
#3. Are ticket prices evenly distributed, or are they skewed?
```

```
[8]: #Question 2: Univariate Analysis
# 1. What is the age distribution of passengers?

plt.figure(figsize=(8, 5))
sns.histplot(df['Age'].dropna(), bins=30, kde=True)
plt.title("Age Distribution of passengers")
plt.show()
```

Notebook

Input

+ Add Input Upload

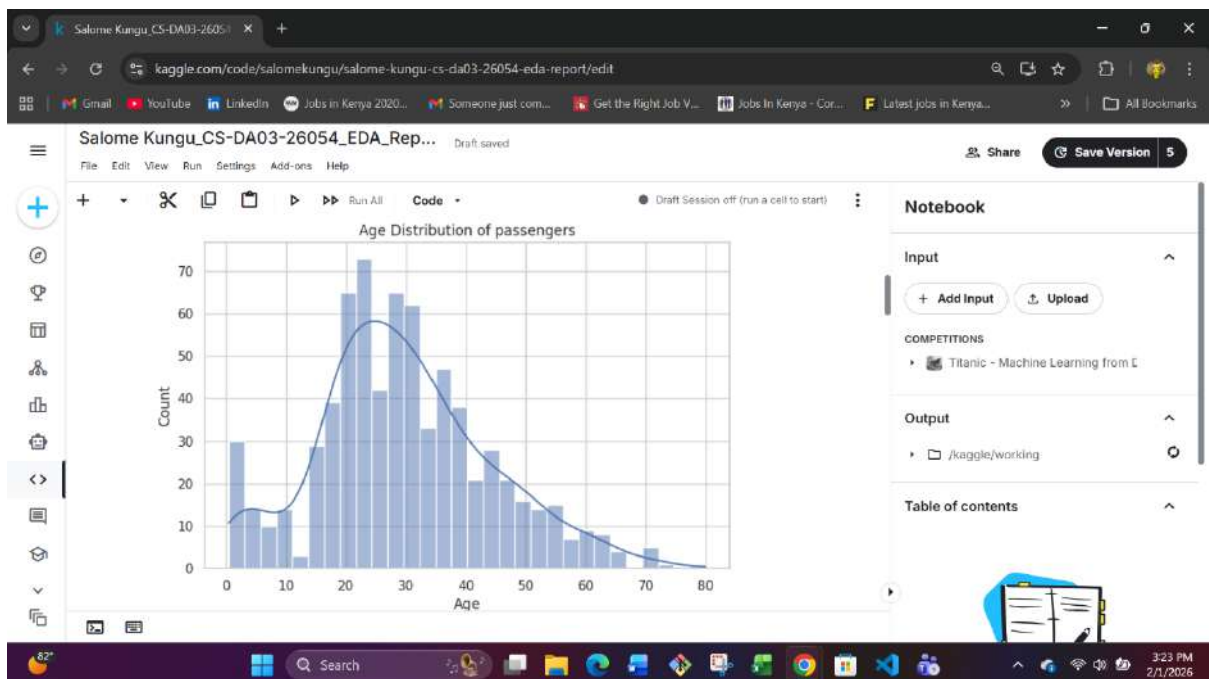
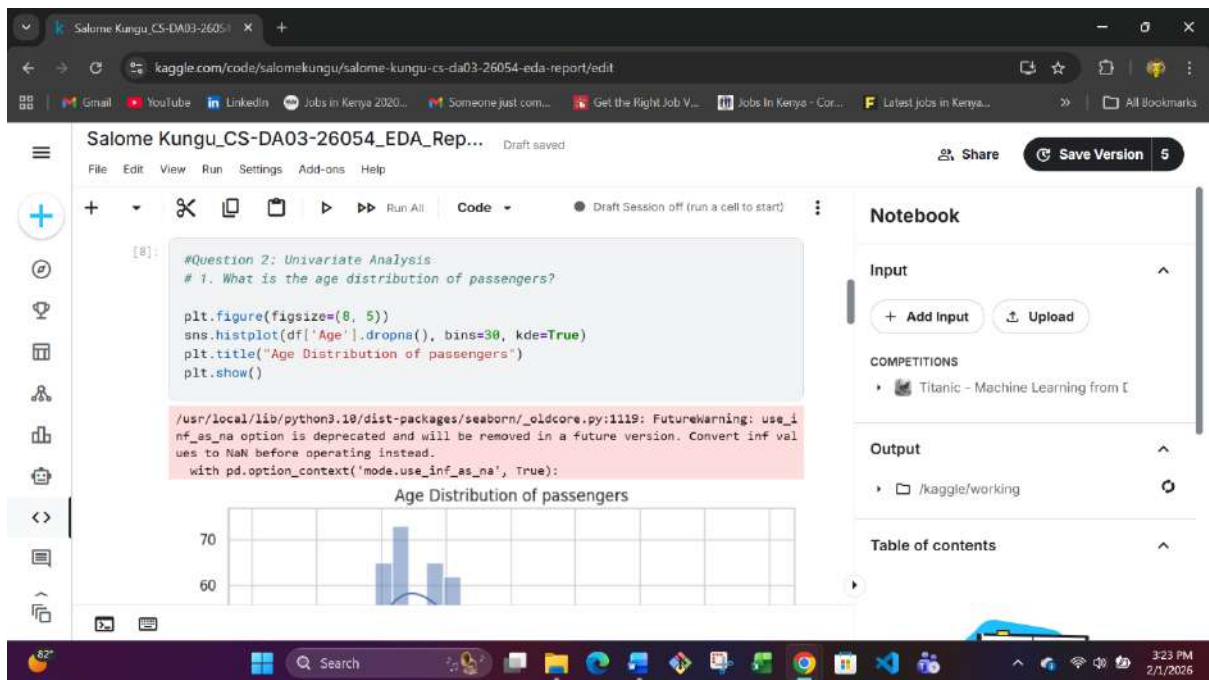
COMPETITIONS

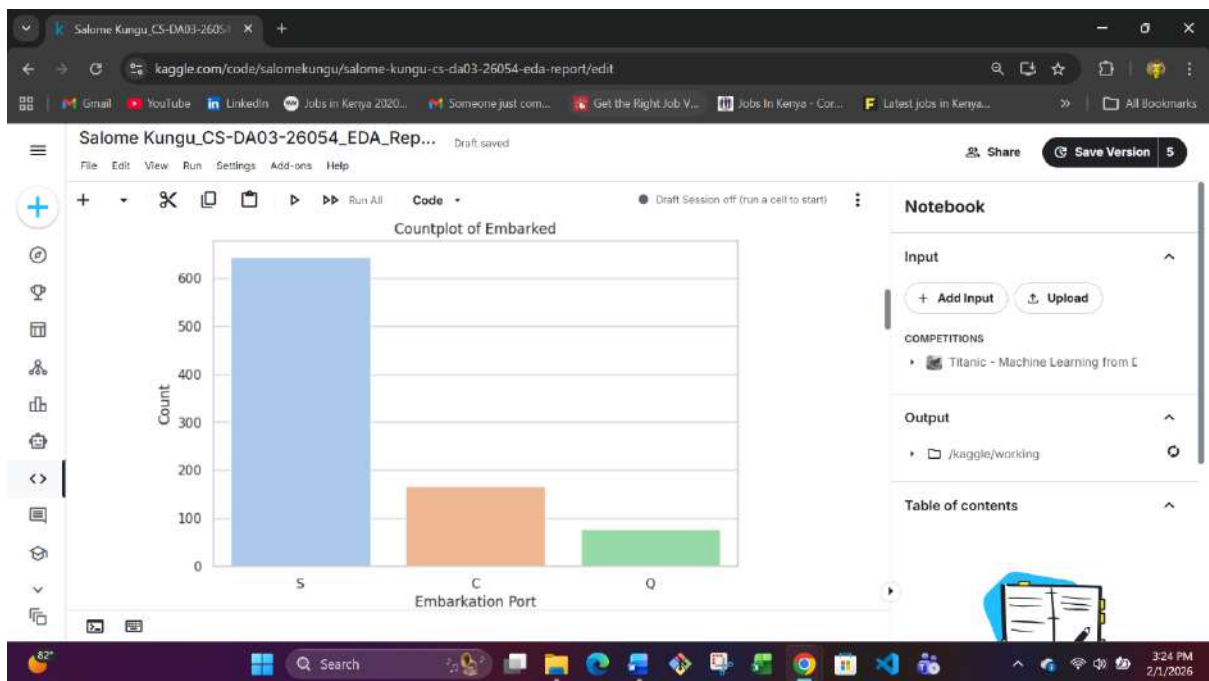
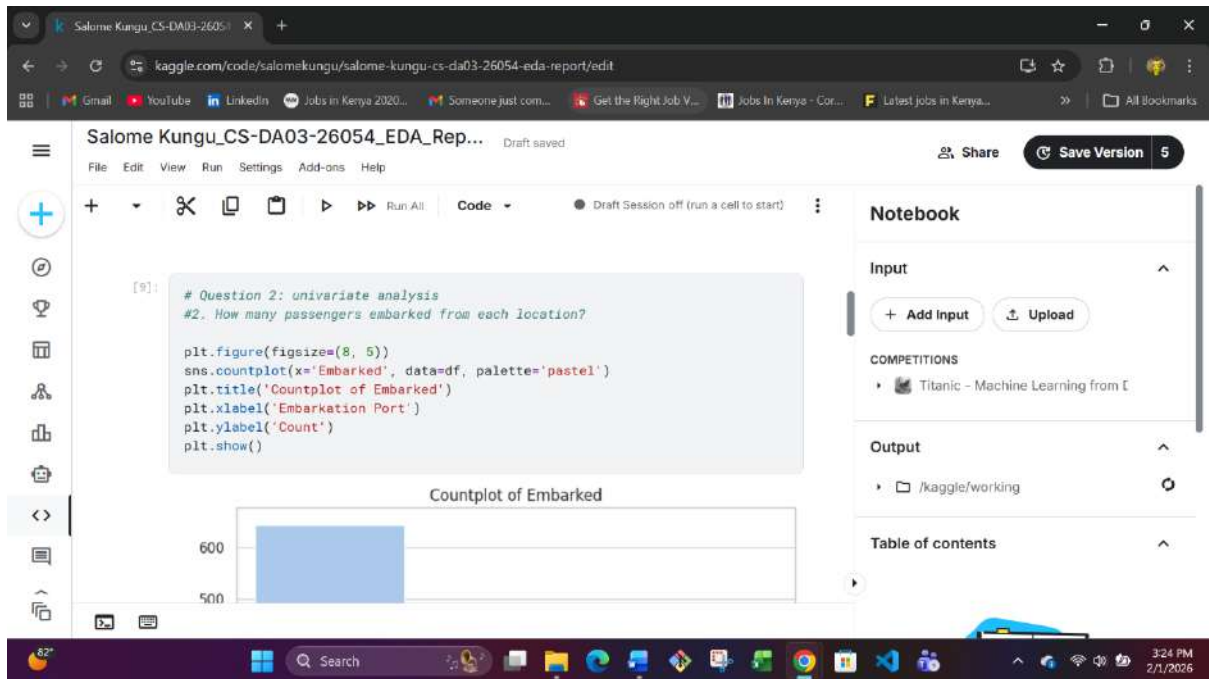
Titanic - Machine Learning from E

Output

/kaggle/working

Table of contents





Salome Kungu_CS-DA03-26054

[kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit](#)

Salome Kungu_CS-DA03-26054_EDA_Rep...

Draft saved

Share

Save Version 5

+

File Edit View Run Settings Add-ons Help

Run All

Code

Draft Session off (run a cell to start)

```
[10]: # Question 2: univariate analysis
#3. Are ticket prices evenly distributed, or are they skewed?

plt.figure(figsize=(8, 5))
sns.kdeplot(df['Fare'], shade=True)
plt.title('KDE Plot of Fare')
plt.xlabel('Fare')
plt.show()
```

<ipython-input-10-2ab684f25e06>:5: FutureWarning:

'shade' is now deprecated in favor of 'fill'; setting 'fill=True'.


This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(df['Fare'], shade=True)

/usr/local/lib/python3.10/dist-packages/seaborn/_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.

with pd.option_context('mode.use_inf_as_na', True):

KDE Plot of Fare



Notebook

Input

+ Add Input Upload

COMPETITIONS

Titanic - Machine Learning from E

Output

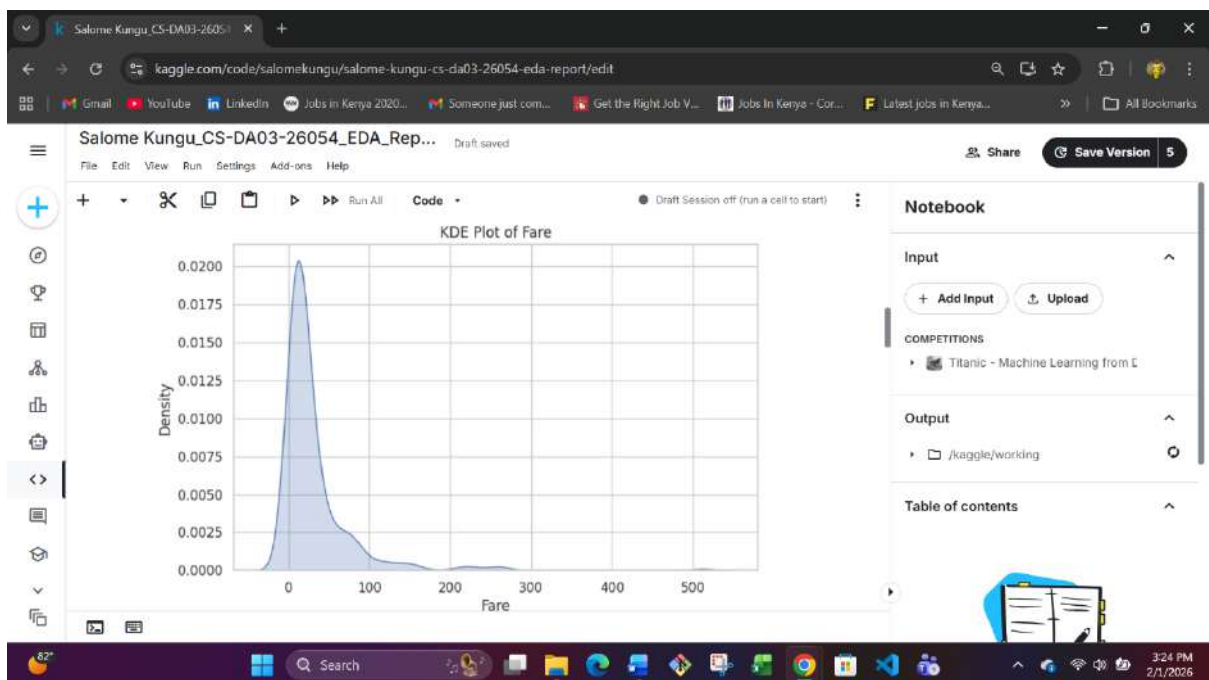
/kaggle/working

Table of contents

62°

Search

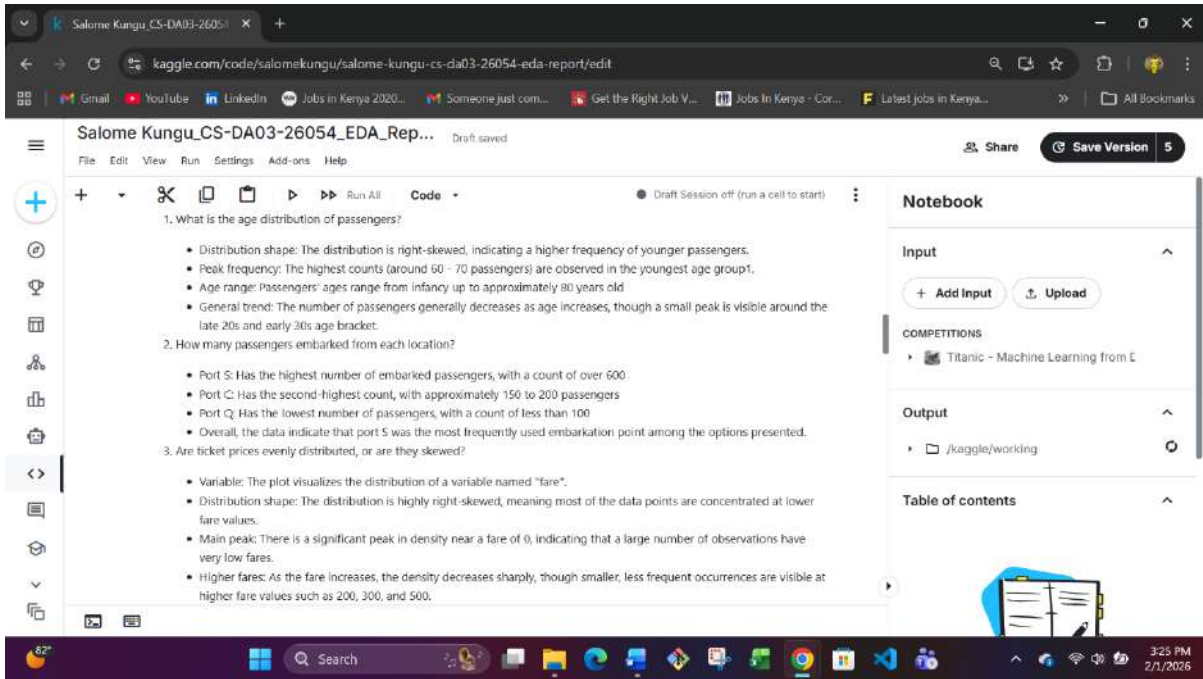
3:24 PM 2/1/2026



62°

Search

3:24 PM 2/1/2026



Salome Kungu_CS-DA03-26054_EDA_Report Draft saved

File Edit View Run Settings Add-ons Help

Code Run All Draft Session off (run a cell to start)

1. What is the age distribution of passengers?

- Distribution shape: The distribution is right-skewed, indicating a higher frequency of younger passengers.
- Peak frequency: The highest counts (around 60 - 70 passengers) are observed in the youngest age group.
- Age range: Passengers' ages range from infancy up to approximately 80 years old.
- General trend: The number of passengers generally decreases as age increases, though a small peak is visible around the late 20s and early 30s age bracket.

2. How many passengers embarked from each location?

- Port S: Has the highest number of embarked passengers, with a count of over 600.
- Port C: Has the second-highest count, with approximately 150 to 200 passengers.
- Port Q: Has the lowest number of passengers, with a count of less than 100.
- Overall, the data indicate that port S was the most frequently used embarkation point among the options presented.

3. Are ticket prices evenly distributed, or are they skewed?

- Variable: The plot visualizes the distribution of a variable named "fare".
- Distribution shape: The distribution is highly right-skewed, meaning most of the data points are concentrated at lower fare values.
- Main peak: There is a significant peak in density near a fare of 0, indicating that a large number of observations have very low fares.
- Higher fares: As the fare increases, the density decreases sharply, though smaller, less frequent occurrences are visible at higher fare values such as 200, 300, and 500.

Notebook

Input

+ Add Input Upload

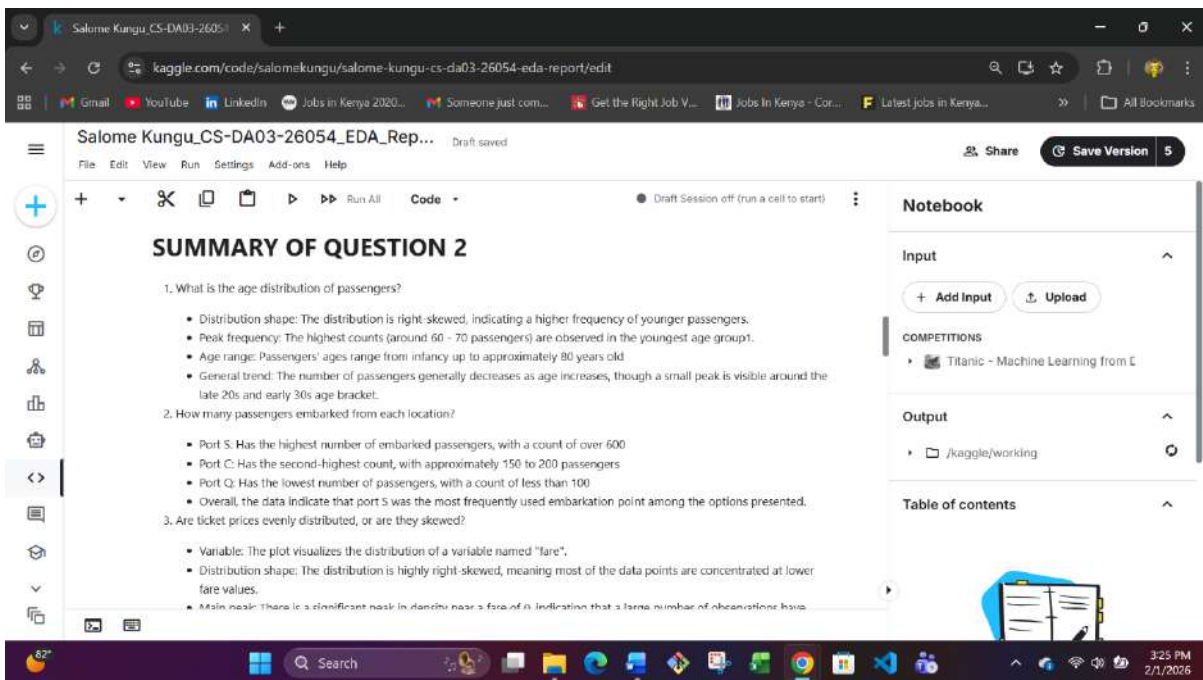
COMPETITIONS

Titanic - Machine Learning from E

Output

/kaggle/working

Table of contents



Salome Kungu_CS-DA03-26054_EDA_Report Draft saved

File Edit View Run Settings Add-ons Help

Code Run All Draft Session off (run a cell to start)

SUMMARY OF QUESTION 2

1. What is the age distribution of passengers?

- Distribution shape: The distribution is right-skewed, indicating a higher frequency of younger passengers.
- Peak frequency: The highest counts (around 60 - 70 passengers) are observed in the youngest age group.
- Age range: Passengers' ages range from infancy up to approximately 80 years old.
- General trend: The number of passengers generally decreases as age increases, though a small peak is visible around the late 20s and early 30s age bracket.

2. How many passengers embarked from each location?

- Port S: Has the highest number of embarked passengers, with a count of over 600.
- Port C: Has the second-highest count, with approximately 150 to 200 passengers.
- Port Q: Has the lowest number of passengers, with a count of less than 100.
- Overall, the data indicate that port S was the most frequently used embarkation point among the options presented.

3. Are ticket prices evenly distributed, or are they skewed?

- Variable: The plot visualizes the distribution of a variable named "fare".
- Distribution shape: The distribution is highly right-skewed, meaning most of the data points are concentrated at lower fare values.
- Main peak: There is a significant peak in density near a fare of 0, indicating that a large number of observations have very low fares.

Notebook

Input

+ Add Input Upload

COMPETITIONS

Titanic - Machine Learning from E

Output

/kaggle/working

Table of contents

Question 3: Bivariate Analysis.

What:

Examine relationships between two variables, such as:

- Survival vs Gender
- Survival vs Passenger Class
- Survival vs Age
- Fare vs Pclass

Why:

This analysis helps identify factors that may influence survival and justifies feature importance for predictive modeling.

Salome Kungu_CS-DA03-26054 | DAT-2026: Assignment 3: Tiber | +

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit

Salome Kungu_CS-DA03-26054_EDA_Rep... Draft saved

File Edit View Run Settings Add-ons Help

Code

```
[11]: # Question 3: Bivariate Analysis
#Examine pairs of features of interest.
#Justify which features you would like to pair in the analysis and seek to answer questions like?
#1. Does the Fare change depending on the Pclass?
#2. Are younger passengers more likely to survive on the Titanic?
#3. Does the Embarked location affect the survival rate?

[12]: #Question 3
#1. Does the Fare change depending on the Pclass?

# Boxplot of Fare grouped by Pclass
plt.figure(figsize=(8, 5))
sns.boxplot(x='Pclass', y='Fare', data=df, palette='Set2')
plt.title('Boxplot of Fare by Pclass')
plt.xlabel('Passenger Class')
plt.ylabel('Fare')
plt.show()
```

Notebook

Input

+ Add Input Upload

COMPETITIONS

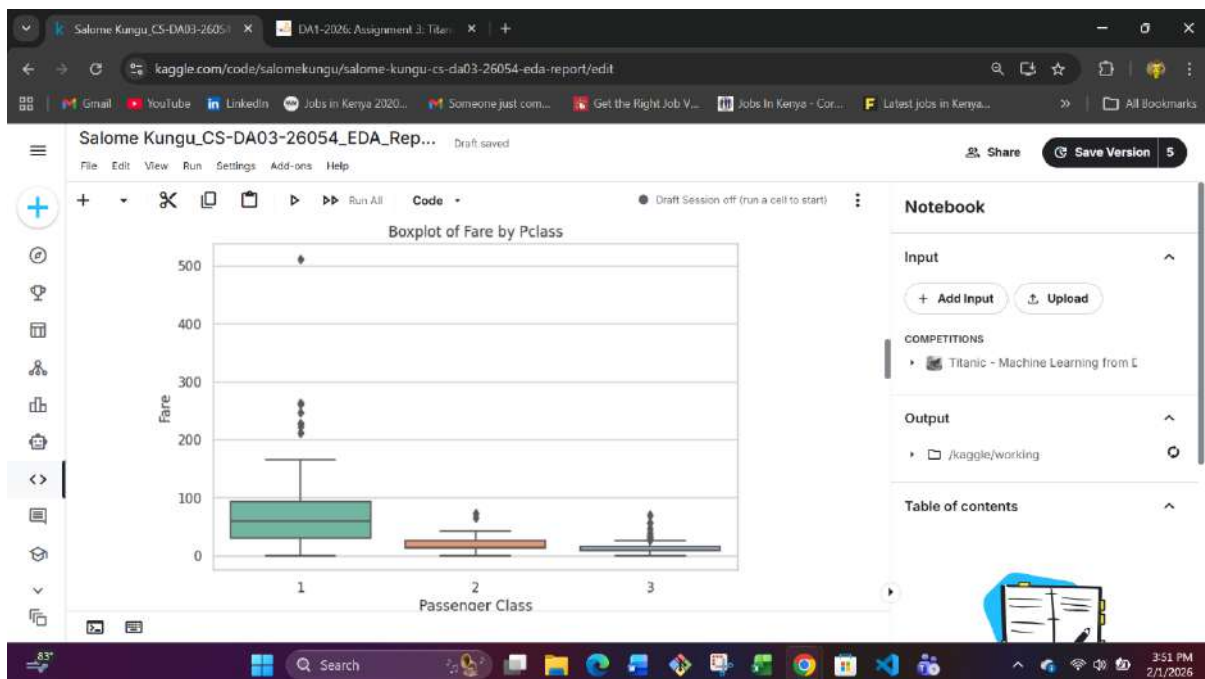
Titanic - Machine Learning from E

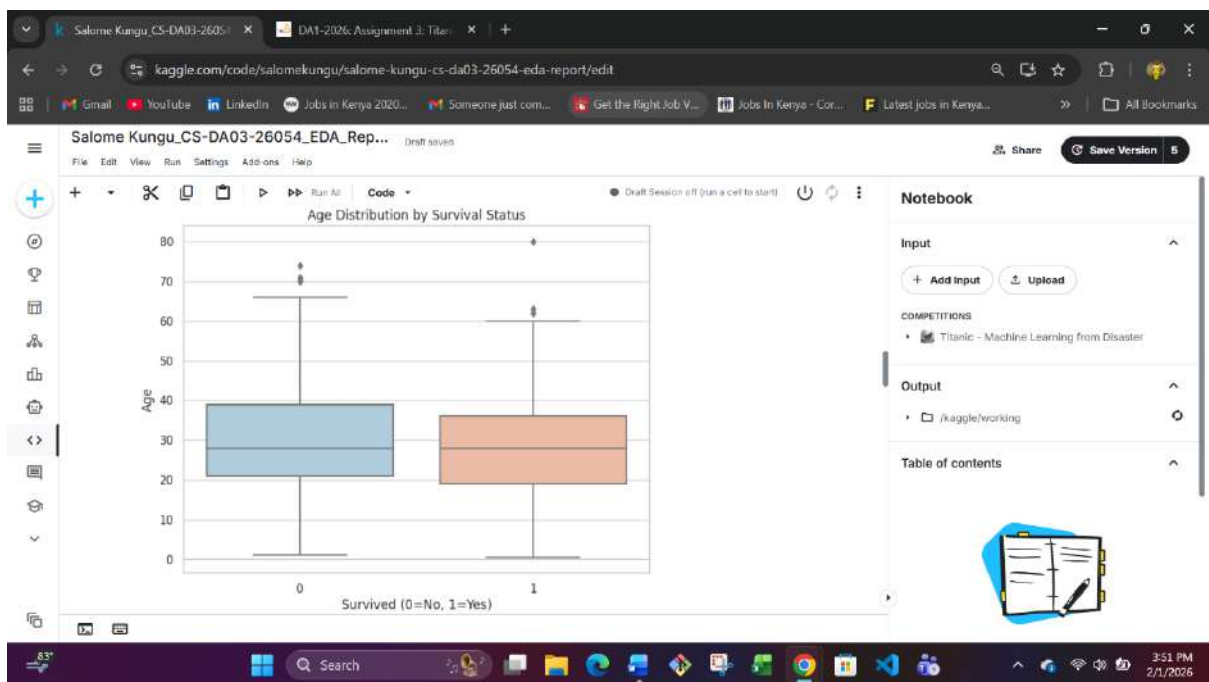
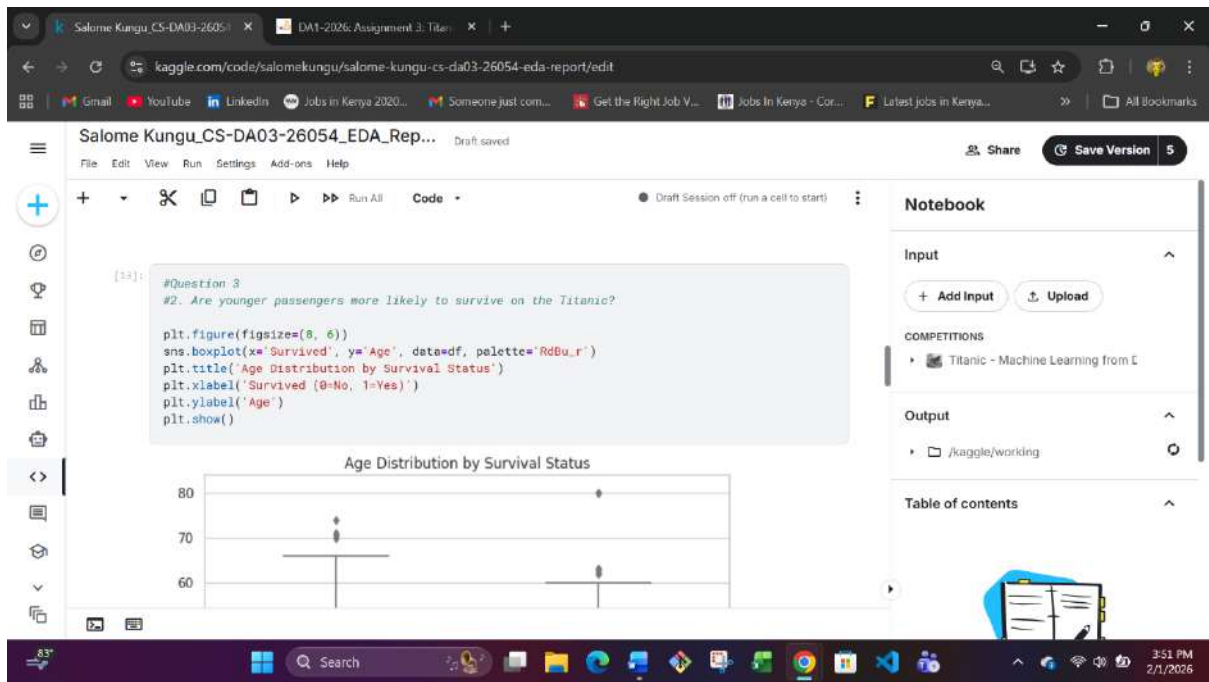
Output

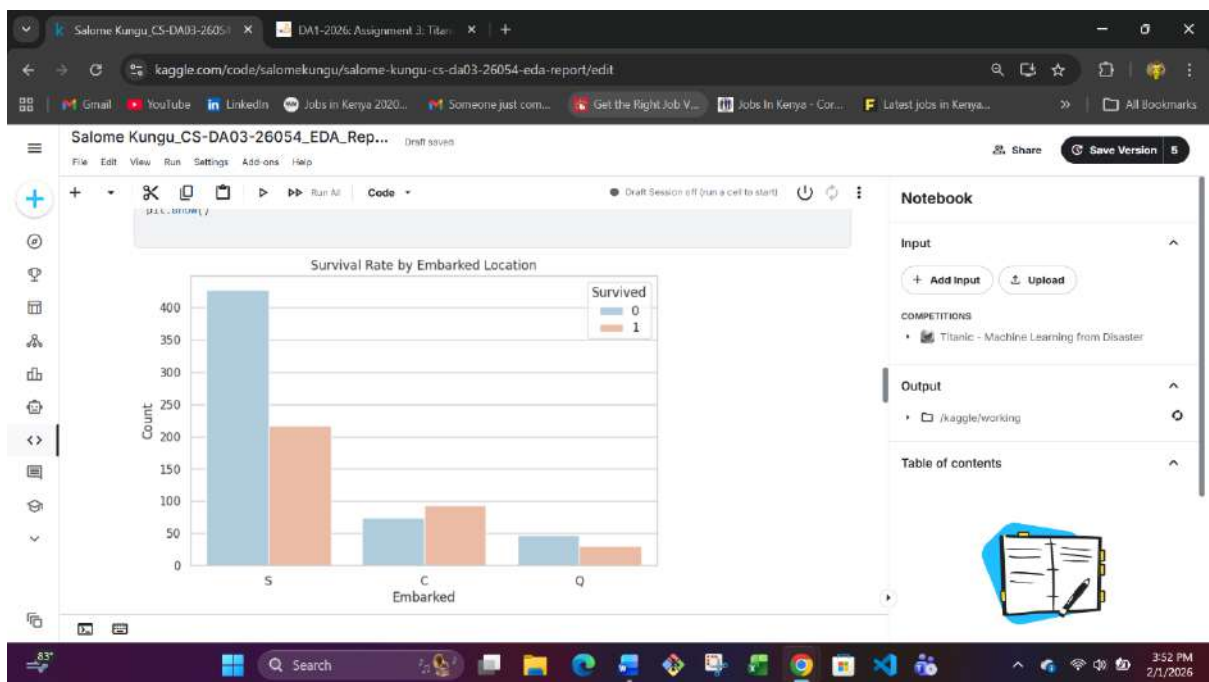
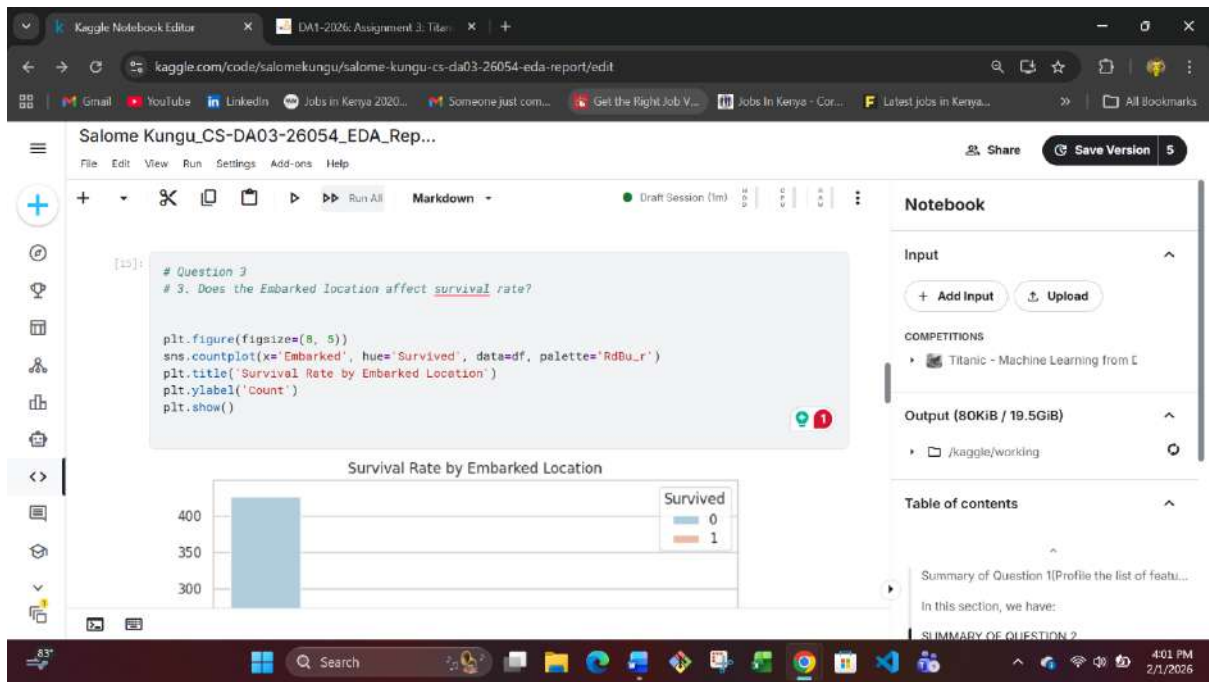
/kaggle/working

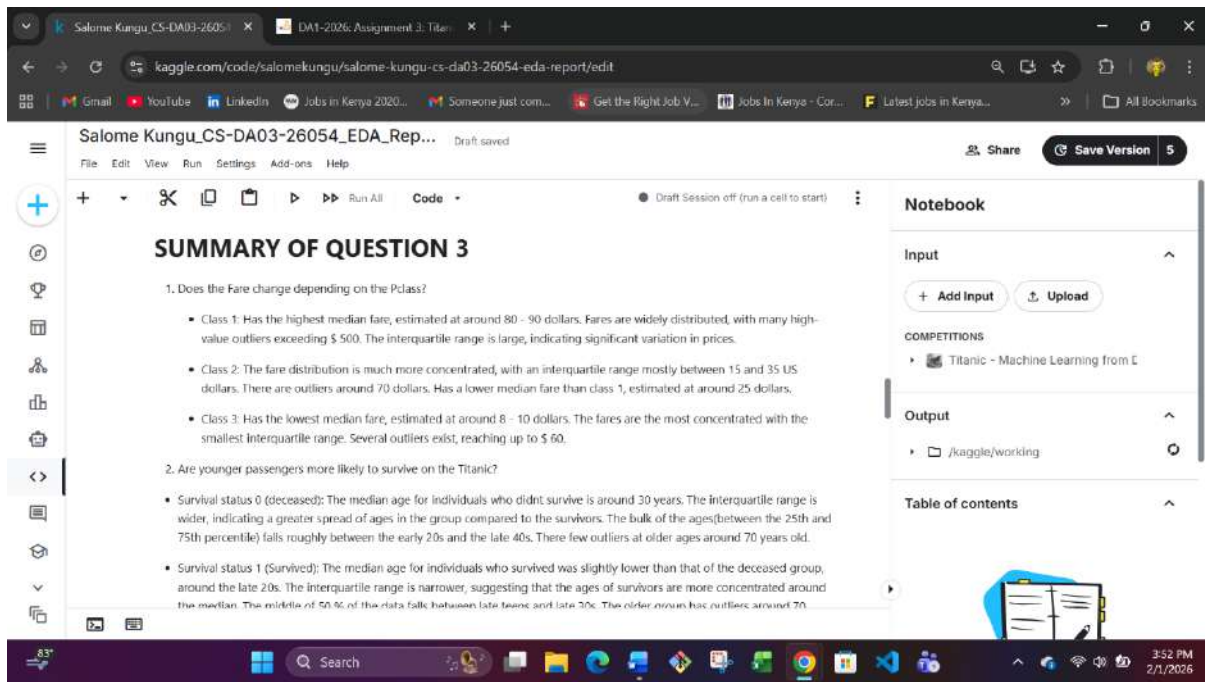
Table of contents

63° Search 3:50 PM 2/1/2026



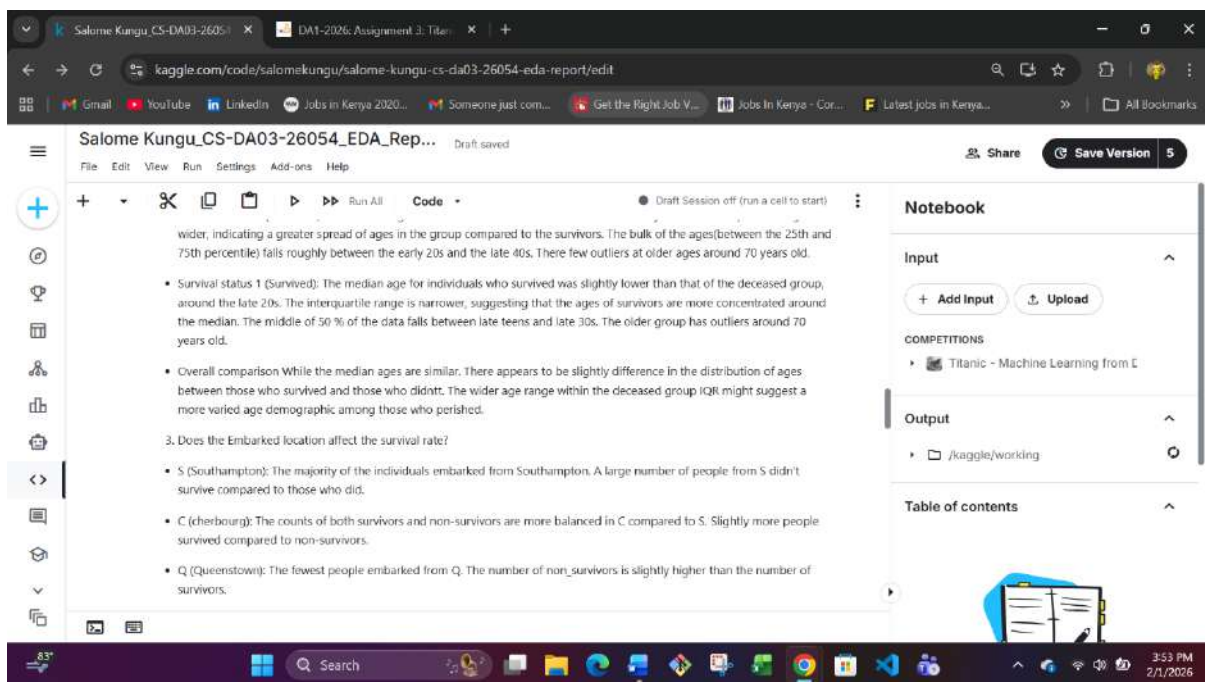






SUMMARY OF QUESTION 3

- Does the Fare change depending on the Pclass?
 - Class 1: Has the highest median fare, estimated at around 80 - 90 dollars. Fares are widely distributed, with many high-value outliers exceeding \$ 500. The interquartile range is large, indicating significant variation in prices.
 - Class 2: The fare distribution is much more concentrated, with an interquartile range mostly between 15 and 35 US dollars. There are outliers around 70 dollars. Has a lower median fare than class 1, estimated at around 25 dollars.
 - Class 3: Has the lowest median fare, estimated at around 8 - 10 dollars. The fares are the most concentrated with the smallest interquartile range. Several outliers exist, reaching up to \$ 60.
- Are younger passengers more likely to survive on the Titanic?
 - Survival status 0 (deceased): The median age for individuals who didn't survive is around 30 years. The interquartile range is wider, indicating a greater spread of ages in the group compared to the survivors. The bulk of the ages (between the 25th and 75th percentile) falls roughly between the early 20s and the late 40s. There are few outliers at older ages around 70 years old.
 - Survival status 1 (Survived): The median age for individuals who survived was slightly lower than that of the deceased group, around the late 20s. The interquartile range is narrower, suggesting that the ages of survivors are more concentrated around the median. The middle of 50 % of the data falls between late teens and late 30s. The older group has outliers around 70 years old.



SUMMARY OF QUESTION 4

- Does the Embarked location affect the survival rate?
 - S (Southampton): The majority of the individuals embarked from Southampton. A large number of people from S didn't survive compared to those who did.
 - C (Cherbourg): The counts of both survivors and non-survivors are more balanced in C compared to S. Slightly more people survived compared to non-survivors.
 - Q (Queenstown): The fewest people embarked from Q. The number of non_survivors is slightly higher than the number of survivors.
- Overall comparison
 - While the median ages are similar, there appears to be a slight difference in the distribution of ages between those who survived and those who didn't. The wider age range within the deceased group (IQR) might suggest a more varied age demographic among those who perished.

Question 4: Multivariate analysis.

What:

Analyze three or more variables simultaneously, for example:

- Survival by Pclass and Gender
- Survival by Age, Fare, and Pclass
- Survival across Embarked locations while considering class

Why:

Multivariate analysis reveals interaction effects and complex patterns not visible in simpler analyses.

Kaggle Notebook Editor
DA1-2026: Assignment 3: Title...
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit

Salome Kungu_CS-DA03-26054_EDA_Rep...
Draft saved
Share
Save Version 5

+
-
Run All
Markdown

Draft Session (13m)

```

[16]: #Question 4
#Multivariate analysis
#Explore more complex relationships between three or more variables simultaneously.
#Detect interactions, combined effects, and hidden patterns that may not be visible in bivariate ar
#This can help answer complex questions, such as:
#1. How do Pclass, Age, and Fare jointly affect survival?
#2. Are survival rates different for Embarked locations when considering Pclass?

[17]: #Question 4
#1. How do Pclass, Age, and Fare jointly affect survival?
# Pair plot for numerical columns

plt.figure(figsize=(10, 10))
sns.pairplot(data=df, vars=['Pclass', 'Age', 'Fare'], hue='Survived', diag_kind='kde', palette='S
plt.show()

```

Notebook

Input
Add Input
Upload

COMPETITIONS
Titanic - Machine Learning from E

Output (80KiB / 19.5GiB)
/kaggle/working

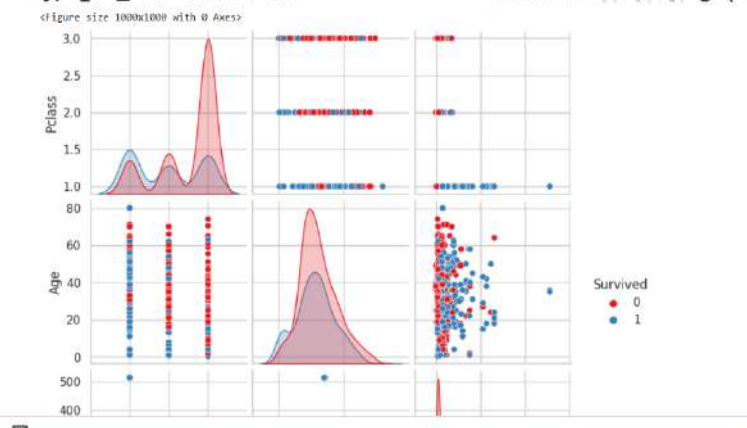
Table of contents
Summary of Question 1(Profile the list of featu...
In this section, we have:
SUMMARY OF QUESTION 2

Kaggle Notebook Editor
DA1-2026: Assignment 3: Title...
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit

Salome Kungu_CS-DA03-26054_EDA_Rep...
Draft saved
Share
Save Version 5

+
-
Run All
Code

Draft Session (14m)



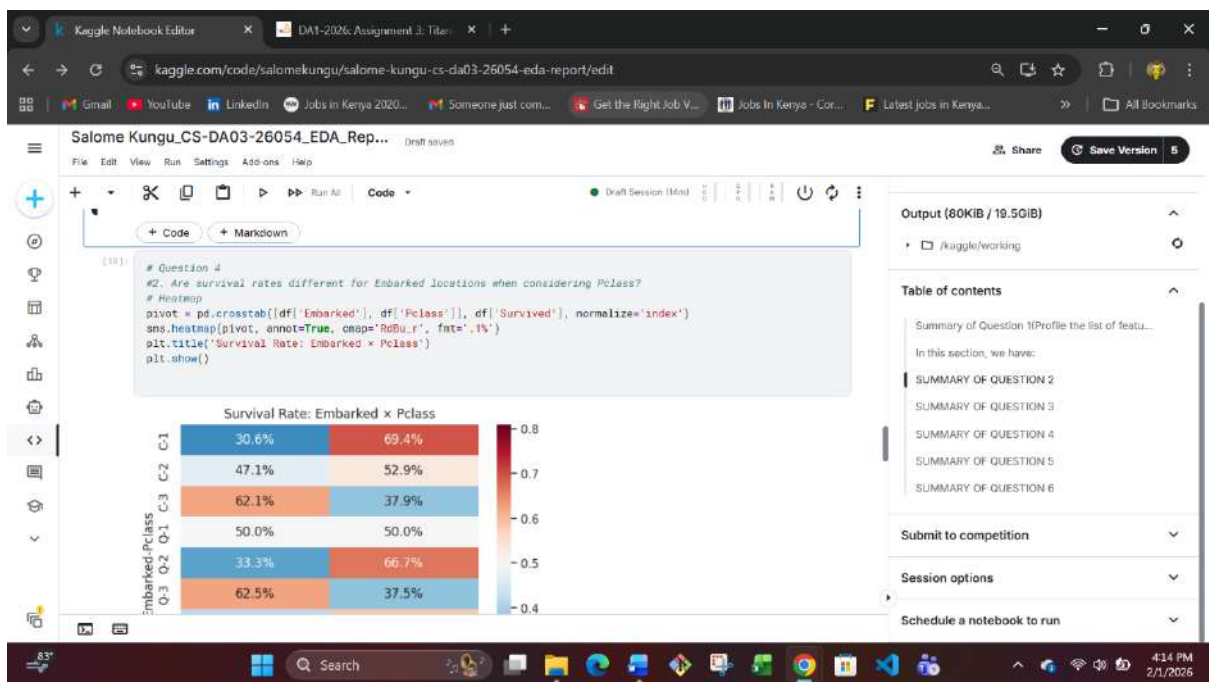
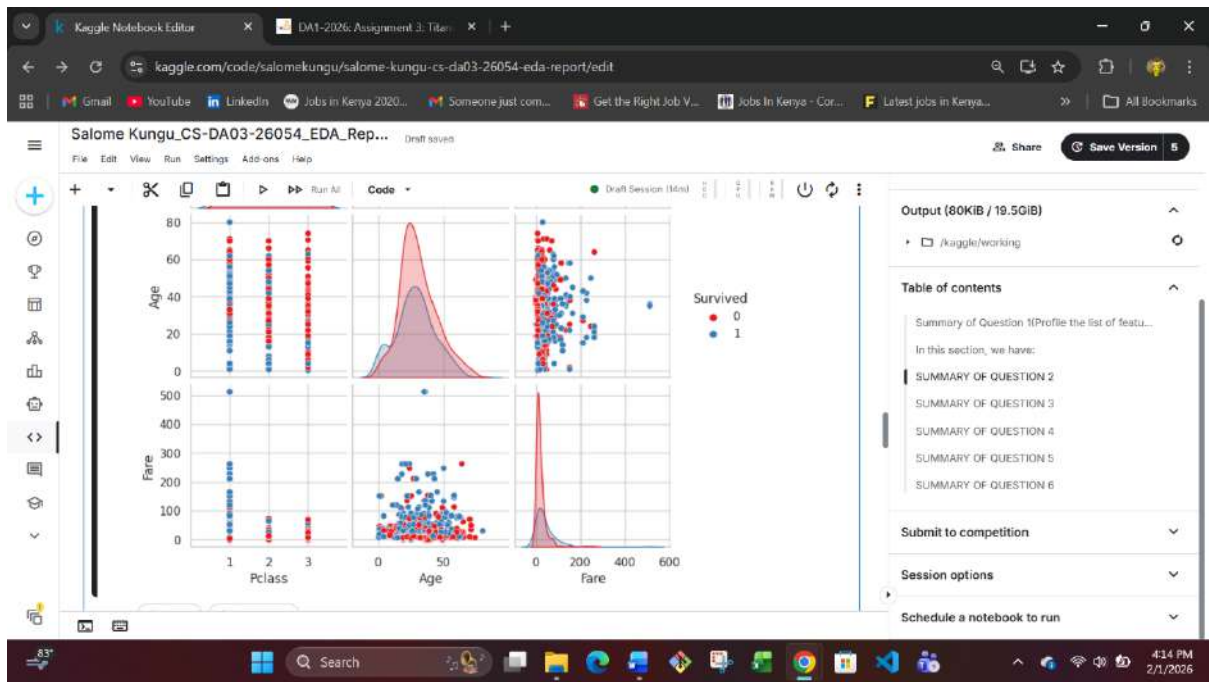
```

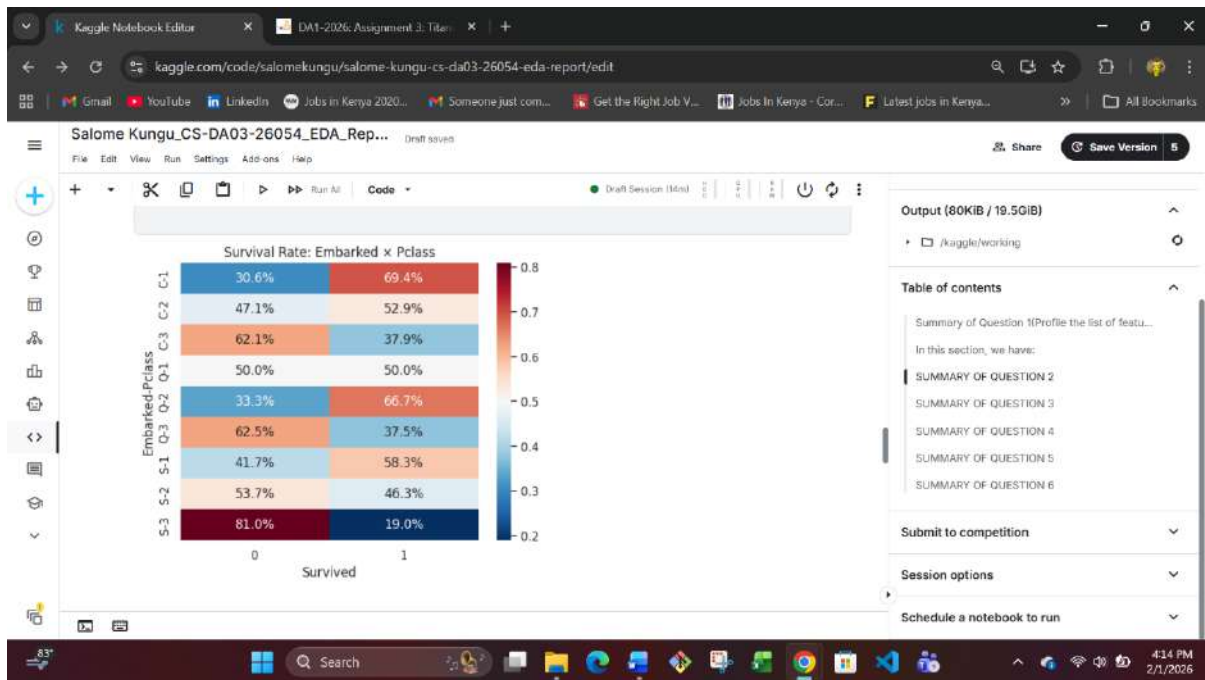
<figure size 1000x1000 with 0 Axes>

```

Output (80KiB / 19.5GiB)
/kaggle/working

Table of contents
Summary of Question 1(Profile the list of featu...
In this section, we have:
SUMMARY OF QUESTION 2
SUMMARY OF QUESTION 3
SUMMARY OF QUESTION 4
SUMMARY OF QUESTION 5
SUMMARY OF QUESTION 6
Submit to competition
Session options
Schedule a notebook to run





Kaggle Notebook Editor | DA1-2026: Assignment 3: Titanic | x +

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit

Salome Kungu_CS-DA03-26054_EDA_Rep... | Draft saved | Share | Save Version 5

File Edit View Run Settings Add-ons Help

Code

Draft Session (14m)

SUMMARY OF QUESTION 4

1. How do Pclass, Age, and Fare jointly affect survival?

- Distributions(Diagonal plots)**
 - Pclass: The majority of passengers were in class 3, with fewer in class 1 and class 2.
 - Age: The distribution is somewhat normal, with a peak around 20 - 40 years old, and a long tail for older ages.
 - Fare: The distribution is heavily skewed right, with most fares being low and a few very high fares.
- Relationships (off_diagonal plots)**
 - Age vs Pclass: Younger passengers are present across all classes, while older passengers are in Pclass 1
 - Fare vs Pclass: Higher fares are exclusively associated with Pclass 1, while Pclass 3 has the lowest fares.
 - Fare vs Age: There isn't a strong linear correlation, but higher fares tend to be paid by a range of ages, mostly in Pclass 1.

2. Are survival rates different for Embarked locations when considering Pclass?

Output (80KiB / 19.5GiB)

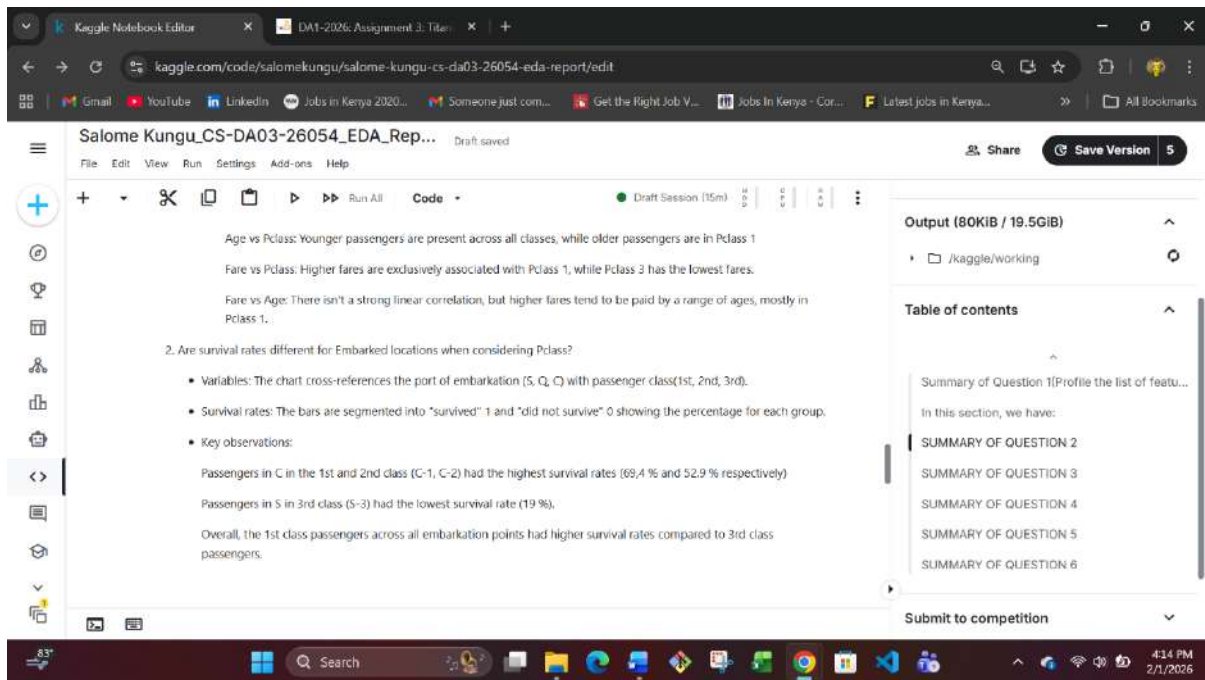
/kaggle/working

Table of contents

- Summary of Question 1(Profile the list of featu...
- In this section, we have:
- SUMMARY OF QUESTION 2
- SUMMARY OF QUESTION 3
- SUMMARY OF QUESTION 4
- SUMMARY OF QUESTION 5
- SUMMARY OF QUESTION 6

Submit to competition

63° Search 4:14 PM 2/1/2026



Salome Kungu_CS-DA03-26054_EDA_Report

File Edit View Run Settings Add-ons Help

Code

Age vs Pclass: Younger passengers are present across all classes, while older passengers are in Pclass 1

Fare vs Pclass: Higher fares are exclusively associated with Pclass 1, while Pclass 3 has the lowest fares.

Fare vs Age: There isn't a strong linear correlation, but higher fares tend to be paid by a range of ages, mostly in Pclass 1.

2. Are survival rates different for Embarked locations when considering Pclass?

- Variables: The chart cross-references the port of embarkation (S, Q, C) with passenger class (1st, 2nd, 3rd).
- Survival rates: The bars are segmented into "survived" 1 and "did not survive" 0 showing the percentage for each group.
- Key observations:
 - Passengers in C in the 1st and 2nd class (C-1, C-2) had the highest survival rates (68.4 % and 52.9 % respectively)
 - Passengers in S in 3rd class (S-3) had the lowest survival rate (19 %).
 - Overall, the 1st class passengers across all embarkation points had higher survival rates compared to 3rd class passengers.

Table of contents

- Summary of Question 1(Profile the list of featu...
- In this section, we have:
- SUMMARY OF QUESTION 2
- SUMMARY OF QUESTION 3
- SUMMARY OF QUESTION 4
- SUMMARY OF QUESTION 5
- SUMMARY OF QUESTION 6

Submit to competition

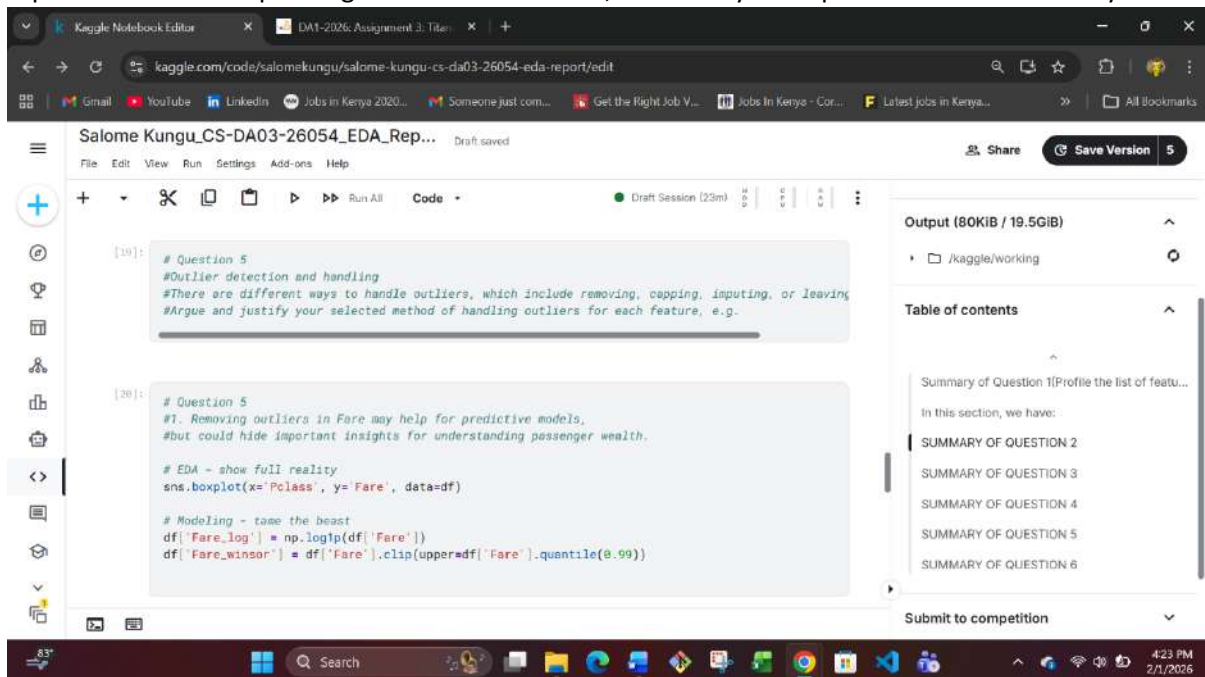
Question 5: Outlier detection and handling:

What:

Identify outliers in numerical features such as **Fare** using boxplots and statistical summaries.

Why:

Outliers can skew analysis and model performance. In this case, outliers were retained as they represent real-world passenger wealth differences, which may be important for survival analysis.



Salome Kungu_CS-DA03-26054_EDA_Report

File Edit View Run Settings Add-ons Help

Code

```
[19]: # Question 5
#Outlier detection and handling
#There are different ways to handle outliers, which include removing, capping, imputing, or leaving
#Argue and justify your selected method of handling outliers for each feature, e.g.
```

```
[20]: # Question 5
#1. Removing outliers in Fare may help for predictive models,
#but could hide important insights for understanding passenger wealth.

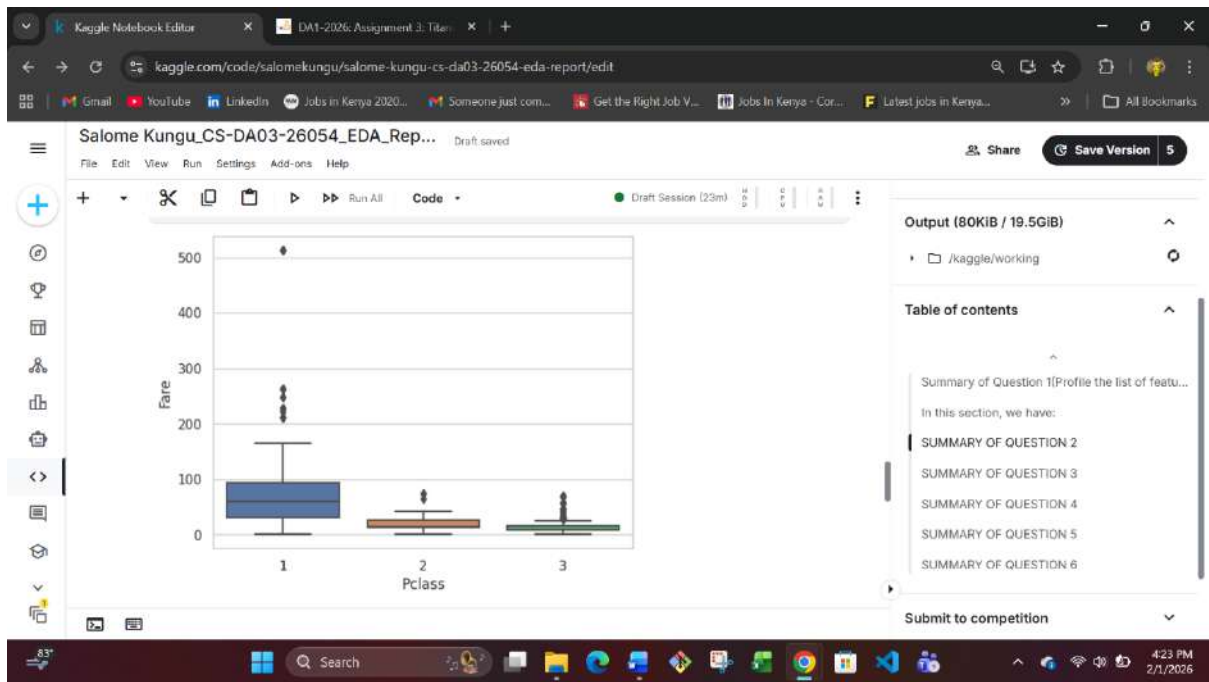
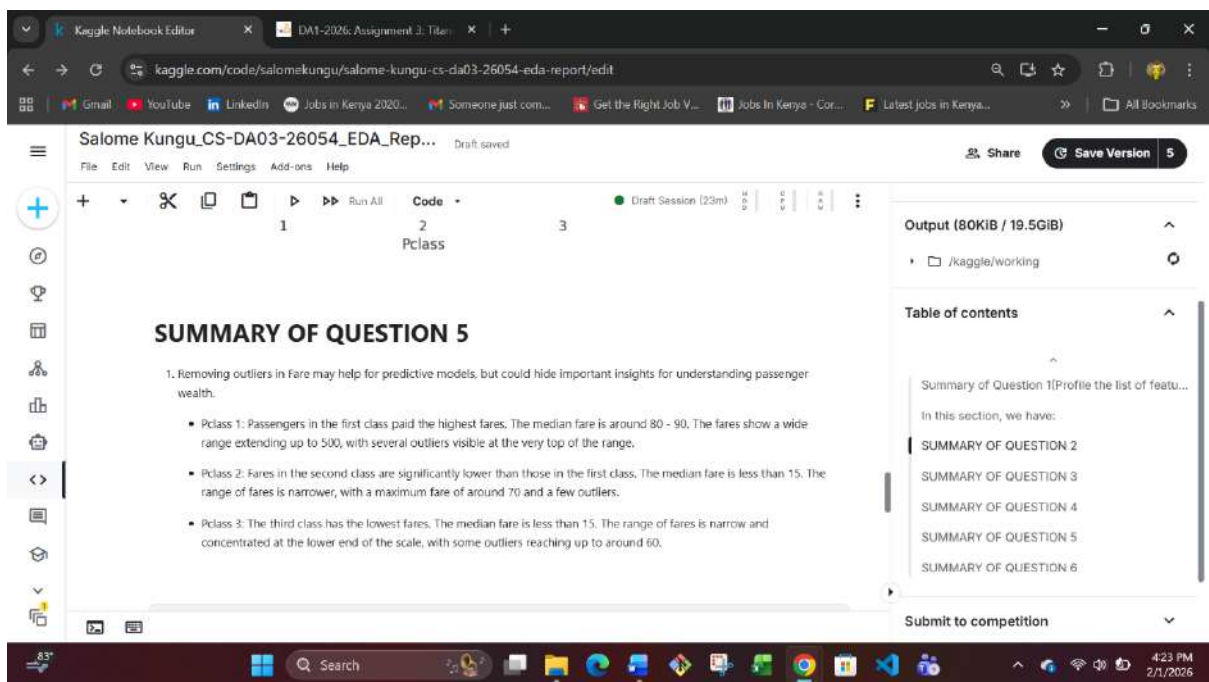
# EDA - show full reality
sns.boxplot(x='Pclass', y='Fare', data=df)

# Modeling - tame the beast
df['Fare_log'] = np.log1p(df['Fare'])
df['Fare_winsor'] = df['Fare'].clip(upper=df['Fare'].quantile(0.99))
```

Table of contents

- Summary of Question 1(Profile the list of featu...
- In this section, we have:
- SUMMARY OF QUESTION 2
- SUMMARY OF QUESTION 3
- SUMMARY OF QUESTION 4
- SUMMARY OF QUESTION 5
- SUMMARY OF QUESTION 6

Submit to competition

SUMMARY OF QUESTION 5

1. Removing outliers in fare may help for predictive models, but could hide important insights for understanding passenger wealth.

- Pclass 1: Passengers in the first class paid the highest fares. The median fare is around 80 - 90. The fares show a wide range extending up to 500, with several outliers visible at the very top of the range.
- Pclass 2: Fares in the second class are significantly lower than those in the first class. The median fare is less than 15. The range of fares is narrower, with a maximum fare of around 70 and a few outliers.
- Pclass 3: The third class has the lowest fares. The median fare is less than 15. The range of fares is narrow and concentrated at the lower end of the scale, with some outliers reaching up to around 60.

Question 6: Target Variable Exploration

What:

- Visualize the distribution of the **Survived** variable.
- Assess class balance.
- Analyze how survival varies across age, gender, class, and embarkation point.

Why:

Understanding the target variable is essential for framing analytical insights and preparing the dataset for machine learning tasks.

Kaggle Notebook Editor

DA1-2026: Assignment 3: Title

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit

Salome Kungu_CS-DA03-26054_EDA_Rep... Draft saved

File Edit View Run Settings Add-ons Help

Code

Draft Session (28m)

```
[21]: # Question 6
# Target Variable Exploration
#Analyze the Target/Dependent Variable Survived and explore:
#1. The distribution of the target variable (Survived) using countplots and bar plots.
#2. How balanced or imbalanced the dataset is.
#3. What factors (like age, gender, class, or embarkation point) may influence survival?
#4. Use combined plots to detect interaction effects

[22]: # Question 6
#1. The distribution of the target variable (Survived) using countplots and bar plots.

plt.figure(figsize=(8, 5))
sns.countplot(x='Survived', data=df, palette='Set2')
plt.title('Survival Count')
plt.xlabel('Survived (0 = No, 1 = Yes)')
plt.ylabel('Count')
plt.show()
```

Output (80KiB / 19.5GiB)

/kaggle/working

Table of contents

Summary of Question 1(Profile the list of featu...

In this section, we have:

SUMMARY OF QUESTION 2

SUMMARY OF QUESTION 3

SUMMARY OF QUESTION 4

SUMMARY OF QUESTION 5

SUMMARY OF QUESTION 6

Submit to competition

84° Search 4:27 PM 2/1/2026

Kaggle Notebook Editor

DA1-2026: Assignment 3: Title

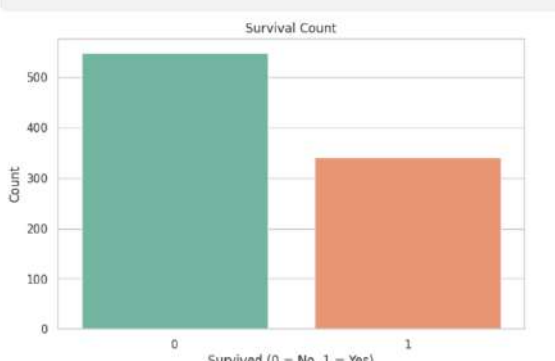
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit

Salome Kungu_CS-DA03-26054_EDA_Rep... Draft saved

File Edit View Run Settings Add-ons Help

Code

Draft Session (28m)



Survival Count

Count

Survived (0 = No, 1 = Yes)

Output (80KiB / 19.5GiB)

/kaggle/working

Table of contents

Summary of Question 1(Profile the list of featu...

In this section, we have:

SUMMARY OF QUESTION 2

SUMMARY OF QUESTION 3

SUMMARY OF QUESTION 4

SUMMARY OF QUESTION 5

SUMMARY OF QUESTION 6

Submit to competition

Session options

Schedule a notebook to run

84° Search 4:27 PM 2/1/2026

Kaggle Notebook Editor
DA1-2026: Assignment 3: Titanic

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit

Salome Kungu_CS-DA03-26054_EDA_Rep...

```

[77]: # Question 6
#2. How balanced or imbalanced the dataset is.
print("Exact counts:")
print(df['Survived'].value_counts())
print("\nPercentages:")
print(df['Survived'].value_counts(normalize=True) * 100)

plt.figure(figsize=(6, 6))
plt.pie(df['Survived'].value_counts(), labels=['Died (62%)', 'Survived (38%)'],
        autopct='%1.1f%%', colors=['lightcoral', 'lightblue'])
plt.title('Target Variable Imbalance')
plt.show()

Exact counts:
Survived
0    549
1    342
Name: count, dtype: int64

Percentages:
Survived
0    61.616162
1     38.383838
Name: proportion, dtype: float64

```

Output (80KiB / 19.5GiB)

Table of contents

Summary of Question 1(Profile the list of featu...
In this section, we have:
SUMMARY OF QUESTION 2
SUMMARY OF QUESTION 3
SUMMARY OF QUESTION 4
SUMMARY OF QUESTION 5
SUMMARY OF QUESTION 6

Submit to competition

Session options

Schedule a notebook to run

Kaggle Notebook Editor
DA1-2026: Assignment 3: Titanic

kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054-eda-report/edit

Salome Kungu_CS-DA03-26054_EDA_Rep...

```

Percentages:
Survived
0    61.616162
1     38.383838
Name: proportion, dtype: float64

Target Variable Imbalance

Died (62%)
61.6%

Survived (38%)
38.4%

```

Output (80KiB / 19.5GiB)

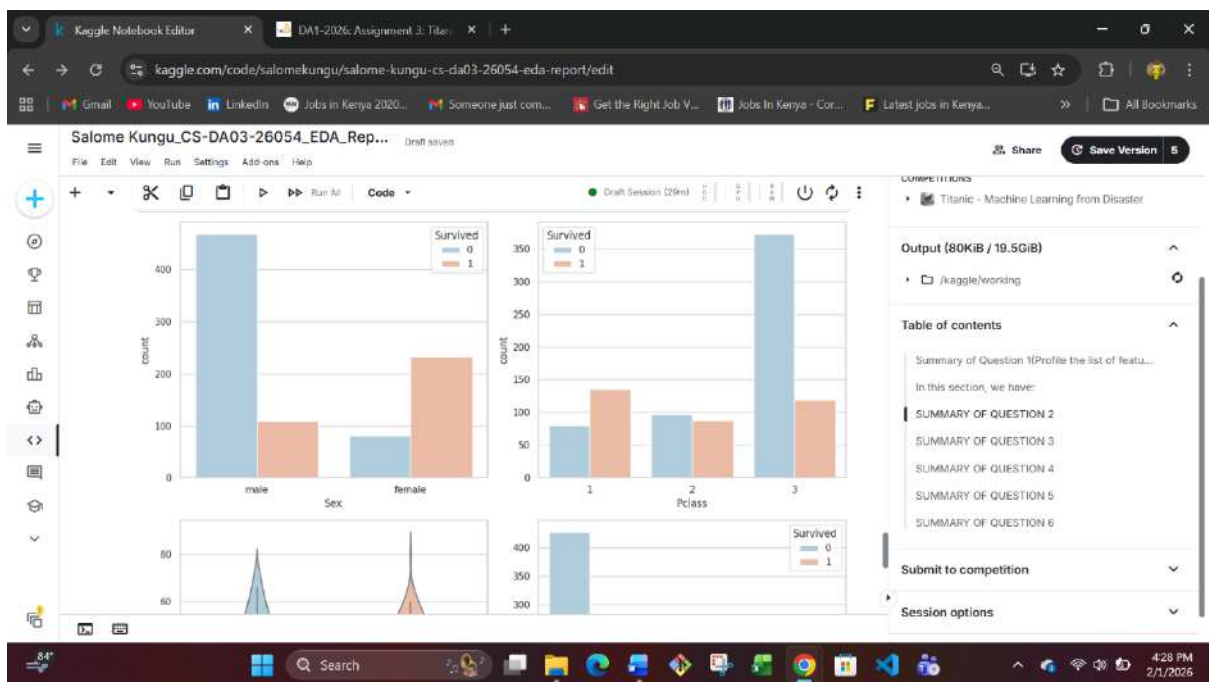
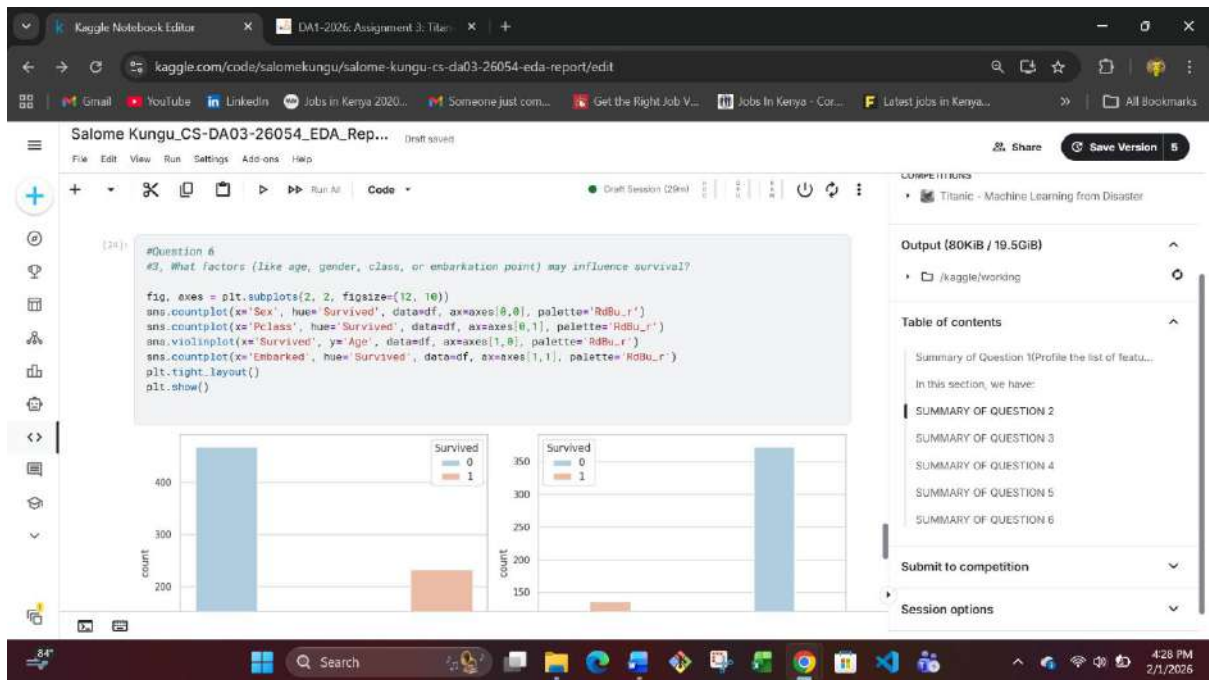
Table of contents

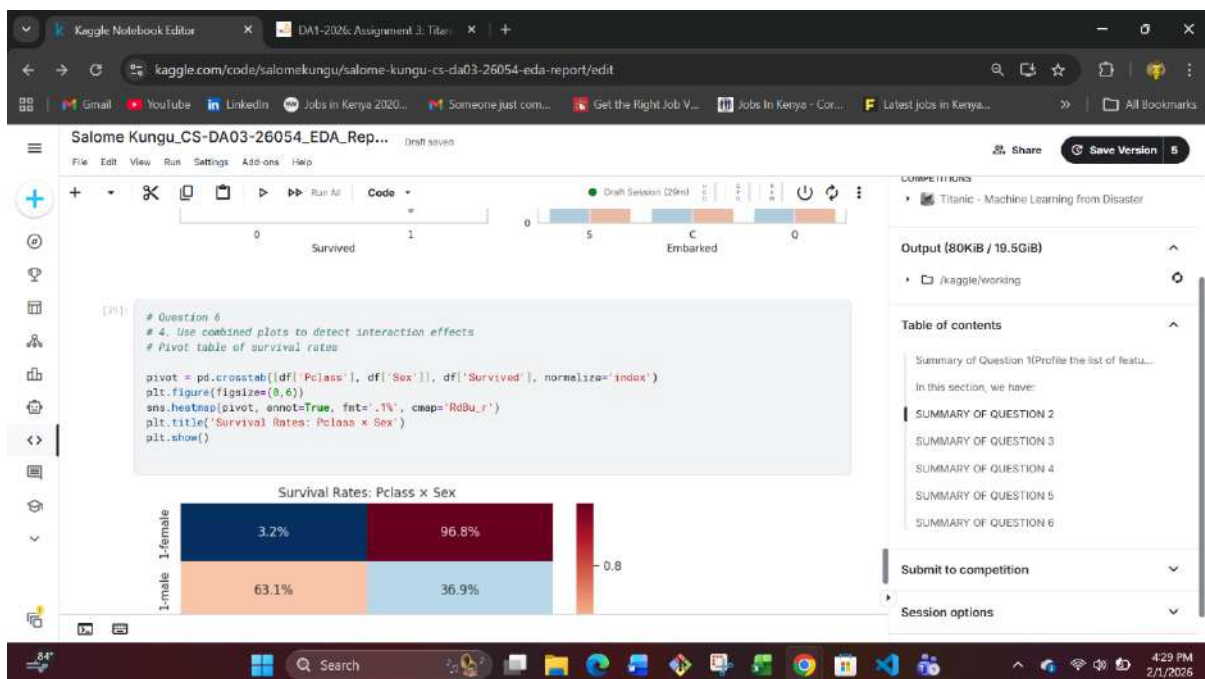
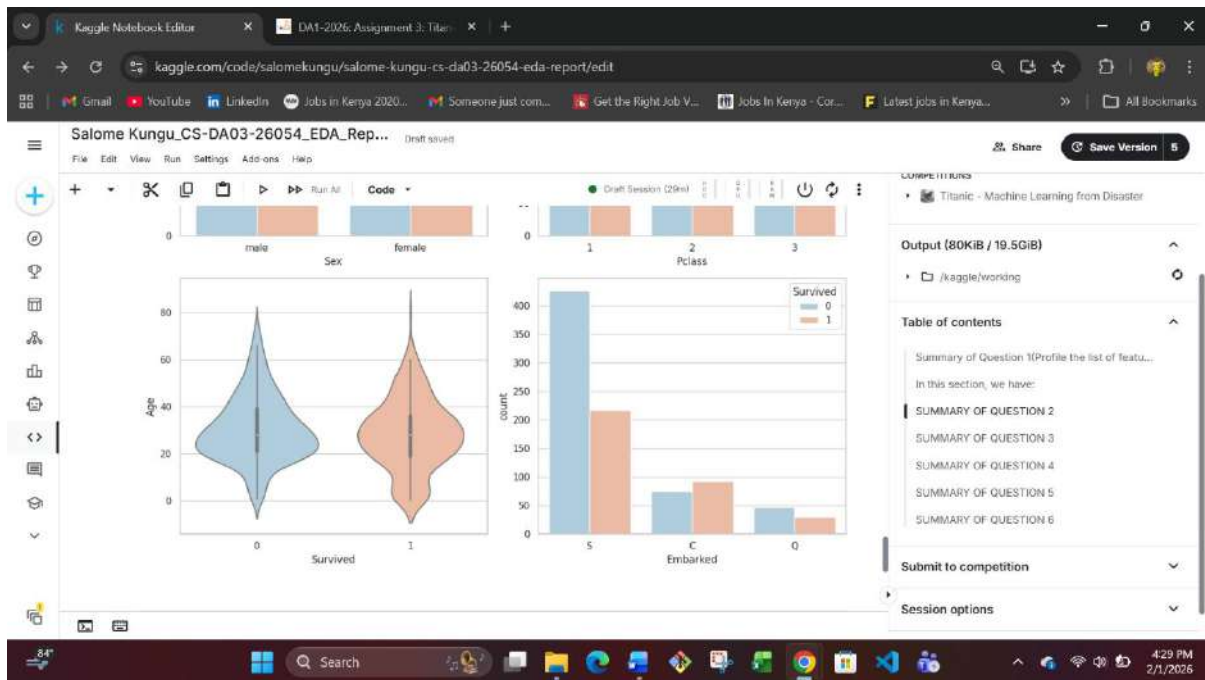
Summary of Question 1(Profile the list of featu...
In this section, we have:
SUMMARY OF QUESTION 2
SUMMARY OF QUESTION 3
SUMMARY OF QUESTION 4
SUMMARY OF QUESTION 5
SUMMARY OF QUESTION 6

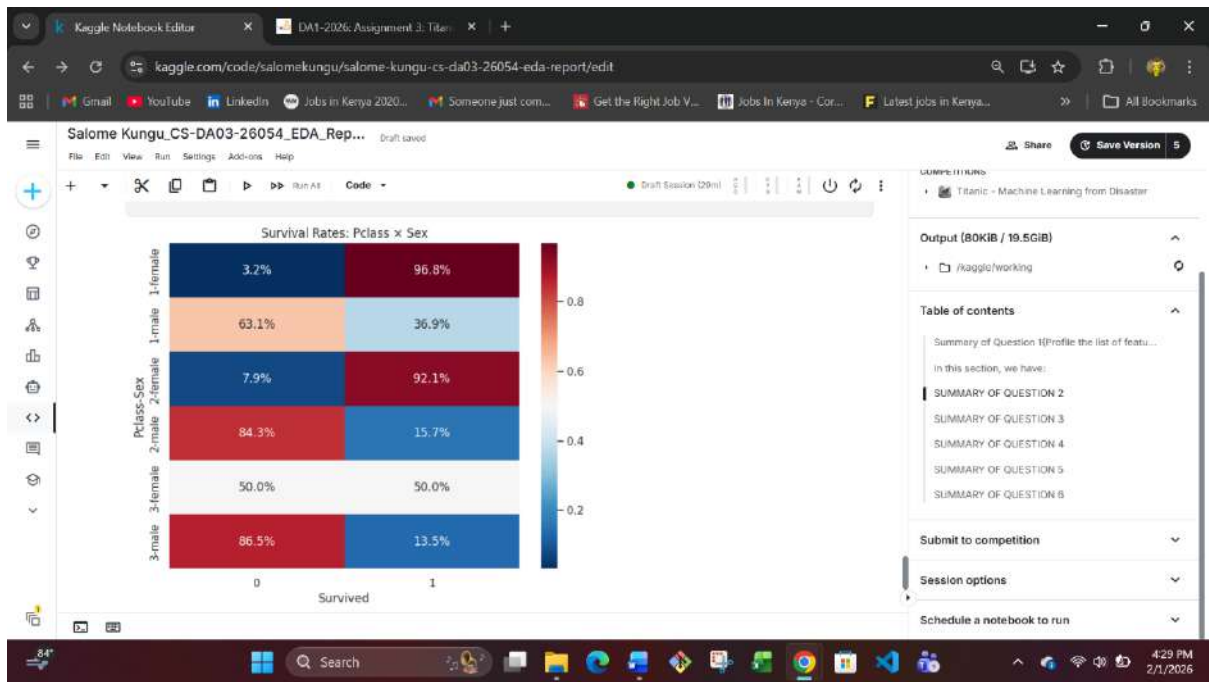
Submit to competition

Session options

Schedule a notebook to run







Kaggle Notebook Editor | DA1-2026: Assignment 3: Titanic | Draft saved

Salome Kungu_CS-DA03-26054_EDA_Rep... | Draft Session (30m) | Save Version 5

SUMMARY OF QUESTION 6

- The distribution of the target variable (Survived) using countplots and bar plots.
 - Non-survivors (0): There were approximately 500 people who didn't survive.
 - Survivors (1): There were approximately 300 people who survived.
- How balanced or imbalanced the dataset is.
 - Non-survivors (0): The majority of passengers died, with 61.6 %.
 - Survivors (1): Accounts for the remaining 38.4 % survived.
- What factors (like age, gender, class, or embarkation point) may influence survival?
 - Sex vs Survival: There were more male passengers overall than female passengers. A higher number of male passengers didn't survive compared to those who did. Female passengers had a higher survival rate than male passengers.
 - Pclass vs Survival: The majority of passengers were in the third class (Pclass 3). The second class (Pclass 2) had the fewest passengers. Most non-survivors were from Pclass 3. Passengers in Pclass 1 had the highest proportion of survivors relative to non-survivors among the classes.

Table of contents:

- Summary of Question 1(Profile the list of featu...
- In this section, we have:
- SUMMARY OF QUESTION 2
- SUMMARY OF QUESTION 3
- SUMMARY OF QUESTION 4
- SUMMARY OF QUESTION 5
- SUMMARY OF QUESTION 6

4:29 PM 2/1/2026

