📄 Exploratory Data Analysis (EDA) Report

Name :Indirala Varshith

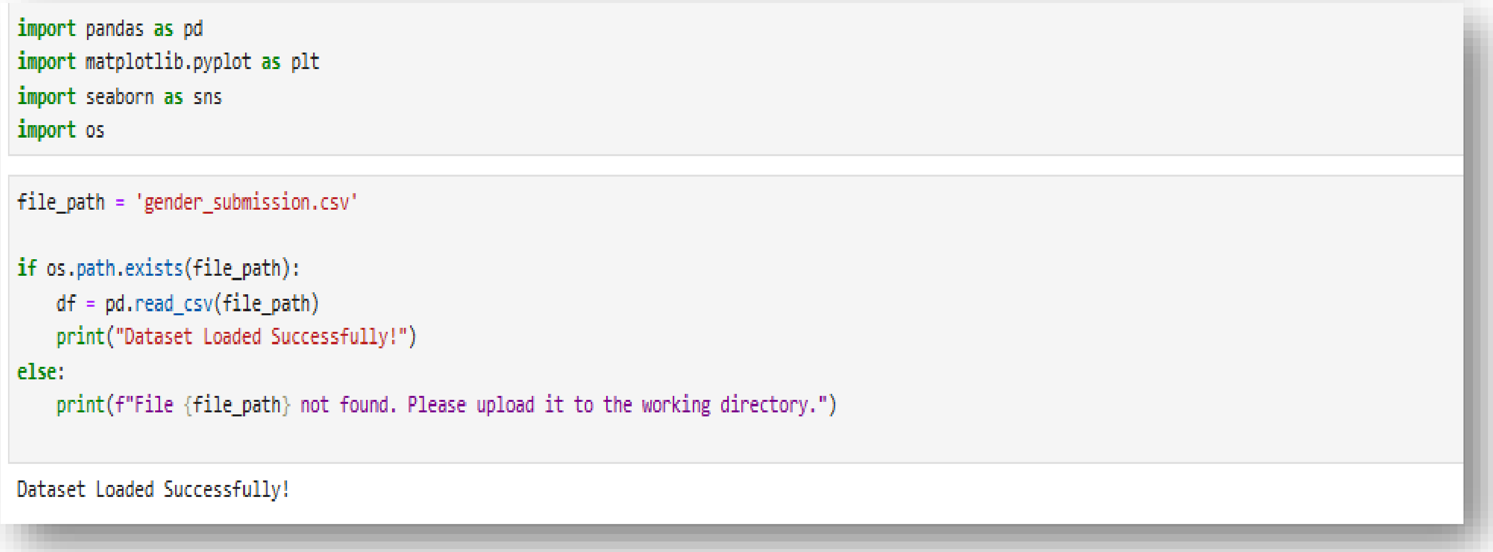
Dataset: gender_submission.csv
Tools Used: Python (Pandas, Matplotlib, Seaborn)

1. Introduction:

The objective of this EDA is to explore the dataset gender_submission.csv, understand its structure, visualize the data, identify any patterns or anomalies, and summarize key findings.

2. Importing Libraries
Loading the Dataset:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
```



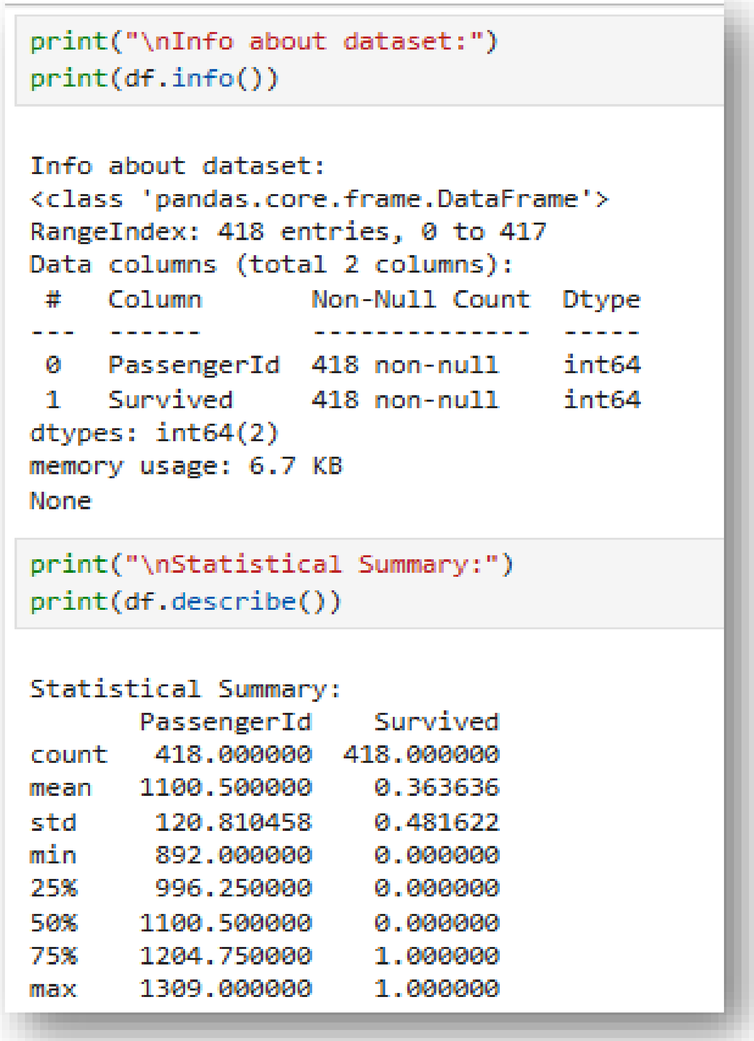
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
```

```
file_path = 'gender_submission.csv'

if os.path.exists(file_path):
    df = pd.read_csv(file_path)
    print("Dataset Loaded Successfully!")
else:
    print(f"File {file_path} not found. Please upload it to the working directory.")
```

```
Dataset Loaded Successfully!
```

3.Data Overview:

- print(df.head())
- print(df.shape)
- print(df.info())
- print(df.describe())



```
print("\nInfo about dataset:")
print(df.info())
```

```
Info about dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Survived     418 non-null    int64
dtypes: int64(2)
memory usage: 6.7 KB
None
```

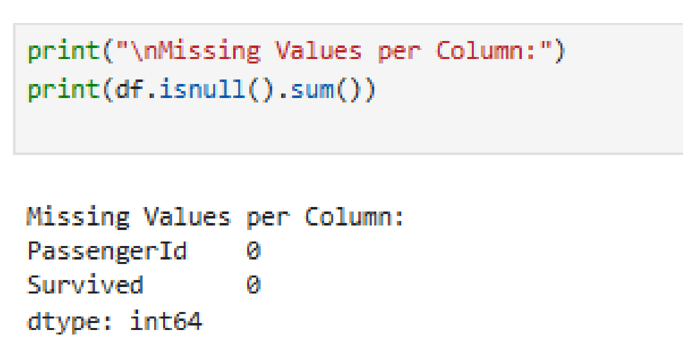
```
print("\nStatistical Summary:")
print(df.describe())
```

```
Statistical Summary:
       PassengerId    Survived
count   418.000000  418.000000
mean   1100.500000    0.363636
std     120.810458    0.481622
min     892.000000    0.000000
25%     996.250000    0.000000
50%    1100.500000    0.000000
75%    1204.750000    1.000000
max    1309.000000    1.000000
```

Observations:

- Dataset contains Passenger Id and Survived columns.
- Passenger Id is a unique identifier.
- Survived indicates survival status (0 = No, 1 = Yes).
- No missing values are observed.

4. Missing Values Check:



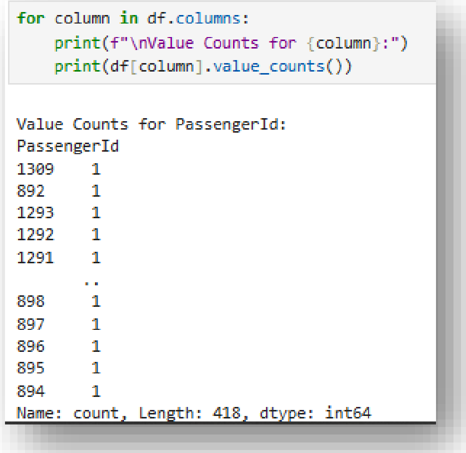
```
print("\nMissing Values per Column:")
print(df.isnull().sum())
```

```

Missing Values per Column:
PassengerId    0
Survived       0
dtype: int64
```

Observations:

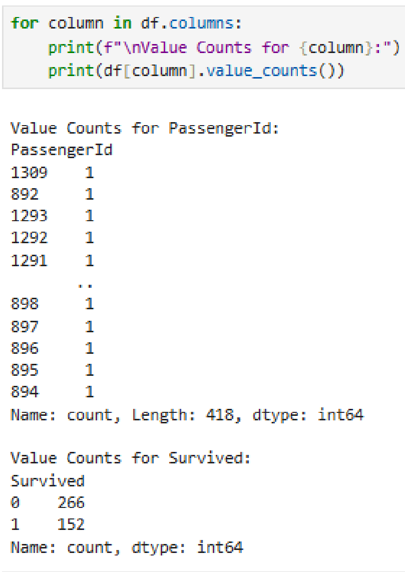
•There are **no missing values** in the dataset.

5. Unique Values and Value Counts:



```
for column in df.columns:
    print(f"\nValue Counts for {column}:")
    print(df[column].value_counts())
```

```

Value Counts for PassengerId:
PassengerId
1309    1
892     1
1293    1
1292    1
1291    1
       ..
898     1
897     1
896     1
895     1
894     1
Name: count, Length: 418, dtype: int64
```



```
for column in df.columns:
    print(f"\nValue Counts for {column}:")
    print(df[column].value_counts())
```

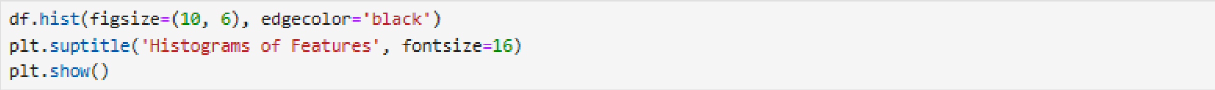
```

Value Counts for PassengerId:
PassengerId
1309    1
892     1
1293    1
1292    1
1291    1
       ..
898     1
897     1
896     1
895     1
894     1
Name: count, Length: 418, dtype: int64

Value Counts for Survived:
Survived
0    266
1    152
Name: count, dtype: int64
```

Observations:

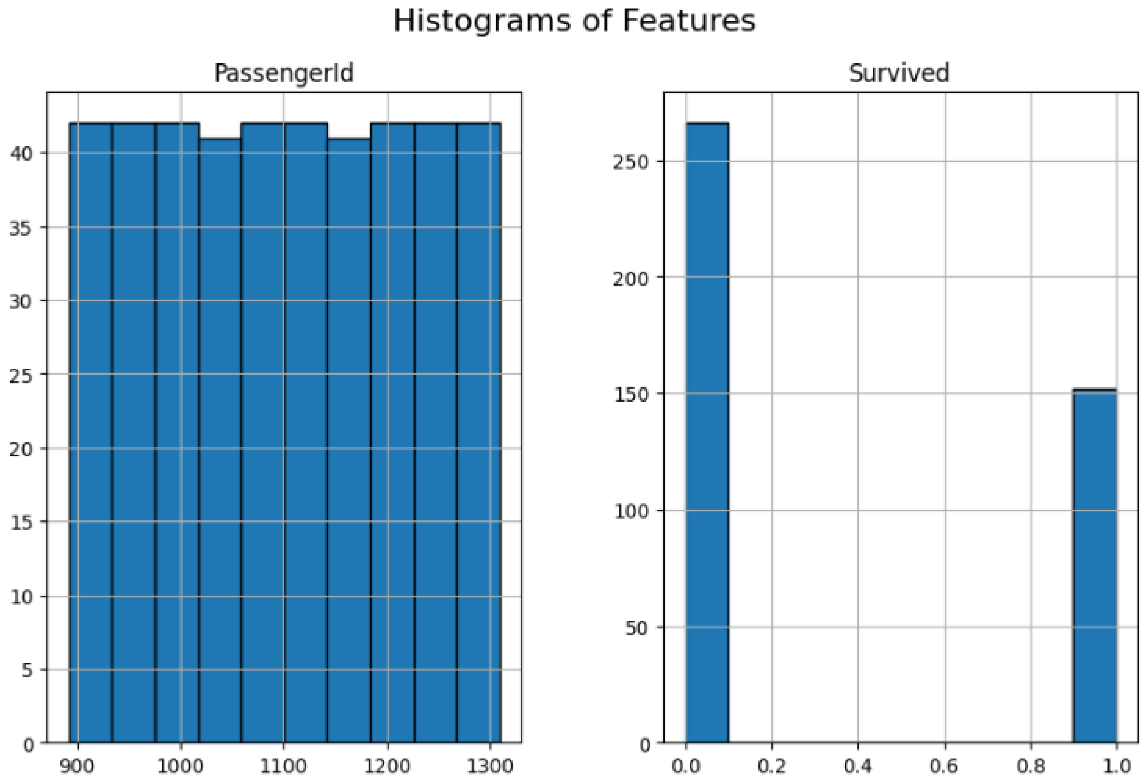
•Each Passenger Id is unique.

•Survived has two categories (0 and 1)

6. Data Visualization:



```
df.hist(figsize=(10, 6), edgecolor='black')
plt.suptitle('Histograms of Features', fontsize=16)
plt.show()
```



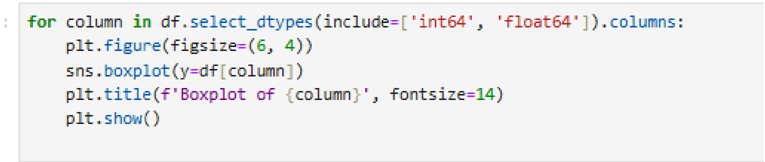
Histograms of Features

Observation:

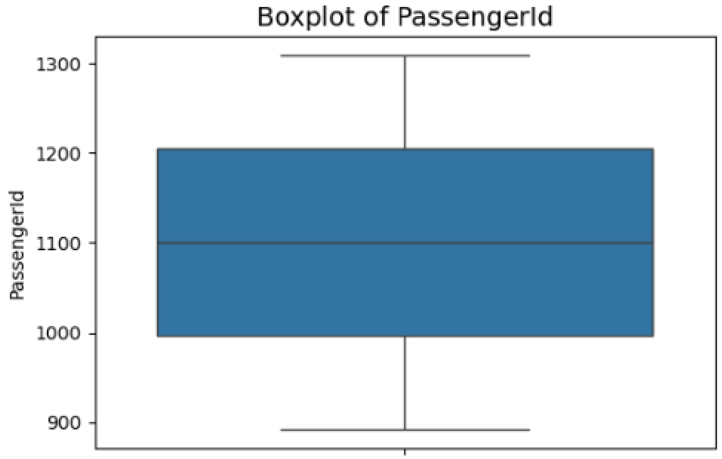
- Passenger Id is uniformly distributed (because it's just an ID).
- Survived distribution shows more non-survivors than survivors.

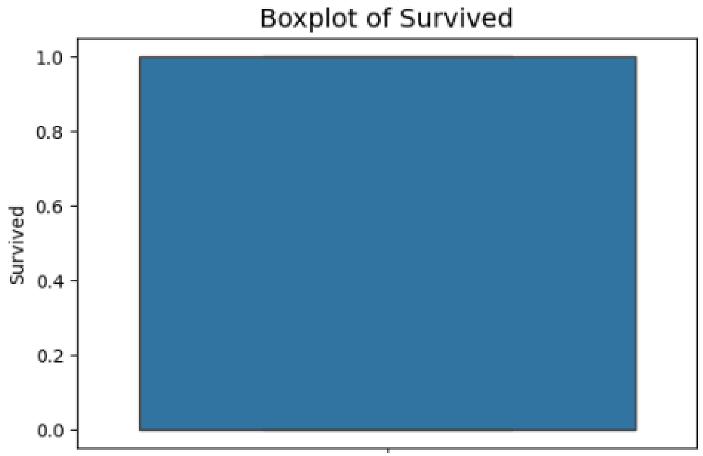
6.2 Boxplots:

- for column in df.select_dtypes(include=['int64', 'float64']).columns:
-     plt.figure(figsize=(6, 4))
-     sns.boxplot(y=df[column])
-     plt.title(f'Boxplot of {column}', fontsize=14)
-     plt.show()



```
for column in df.select_dtypes(include=['int64', 'float64']).columns:
    plt.figure(figsize=(6, 4))
    sns.boxplot(y=df[column])
    plt.title(f'Boxplot of {column}', fontsize=14)
    plt.show()
```

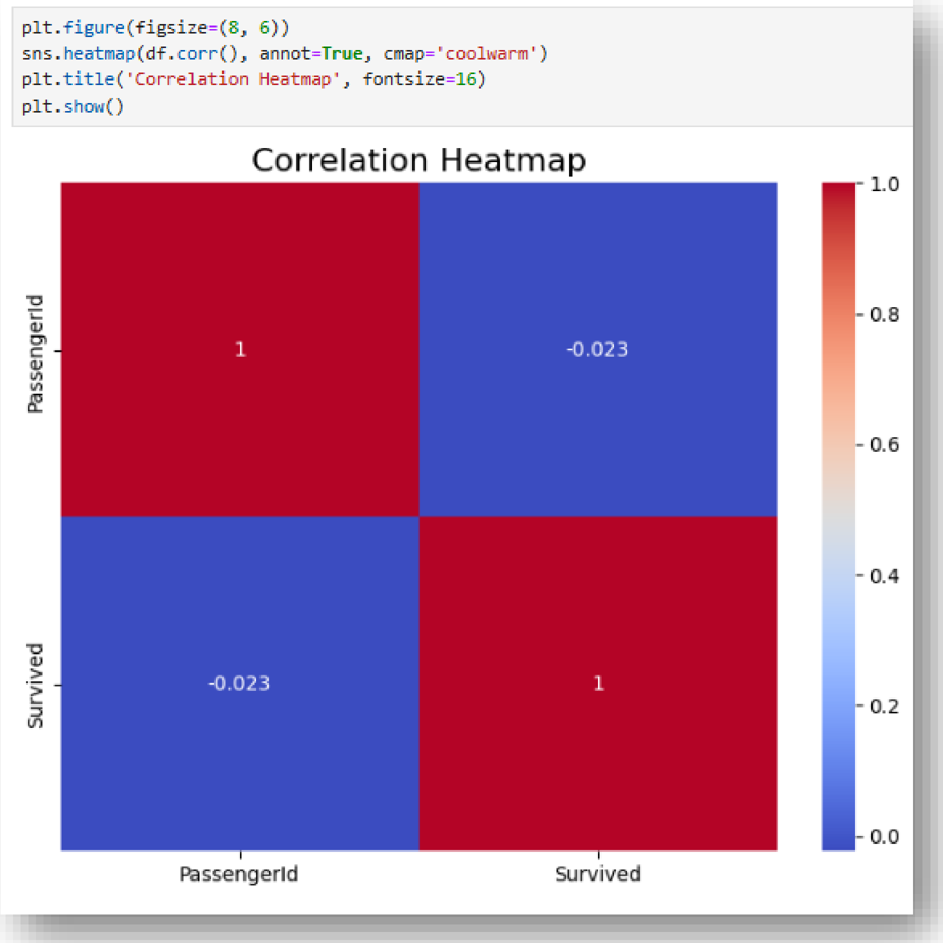




Observation:

•PassengerId and Survived show no meaningful outliers.

6.3 Correlation Heatmap:



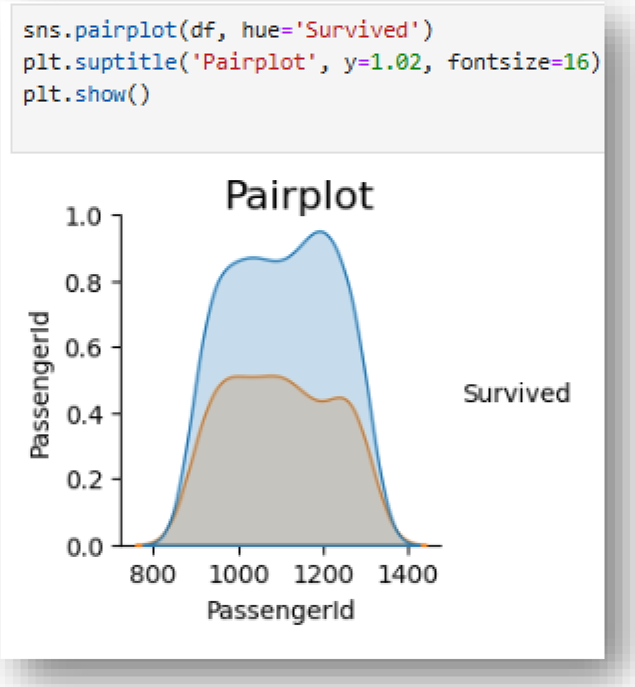
```
plt.figure(figsize=(8, 6))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap', fontsize=16)
plt.show()
```

Observation:

- Passenger Id has no correlation with Survived (expected).
- Limited correlation analysis is possible due to minimal features.

6.4 Pair-plot:

```
sns.pairplot(df, hue='Survived')
plt.suptitle('Pairplot', y=1.02, fontsize=16)
plt.show()
```



```
sns.pairplot(df, hue='Survived')
plt.suptitle('Pairplot', y=1.02, fontsize=16)
plt.show()
```

Observation:

•Pair-plot confirms limited feature interaction in the dataset.

7. Summary of Findings:

- The dataset has no missing values.
- Only two main columns are available: Passenger-Id and Survived.
- Passenger-Id serves purely as an identifier, not suitable for predictive modeling.
- Survived is a binary target variable.
- No strong patterns or relationships can be identified from this limited dataset.
- For deeper analysis, additional features (like Age, Sex, Fare, P-class) are needed.

☑ End of Report