Task2

Understand data using statistics and visualizations.

CODE:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load dataset

df = pd.read\_csv("Titanic-Dataset.csv")

# Display first few rows

print(df.head())

# Summary for numeric columns

print("Summary Statistics:")

print(df.describe())

# Summary for categorical columns

print("\nCategorical Summary:")

print(df.describe(include='object'))

# Select numeric columns

numeric\_cols = df.select\_dtypes(include=np.number).columns.tolist()

# Histograms

df[numeric\_cols].hist(figsize=(12, 10), bins=15)

plt.suptitle("Histograms of Numeric Features")

plt.tight\_layout()

plt.show()

# Boxplots

for col in numeric\_cols:

    plt.figure(figsize=(6, 4))

    sns.boxplot(y=df[col])

    plt.title(f'Boxplot of {col}')

    plt.show()

# Correlation matrix

plt.figure(figsize=(10, 6))

sns.heatmap(df[numeric\_cols].corr(), annot=True, cmap='coolwarm', fmt=".2f")

plt.title("Correlation Matrix")

plt.show()

# Pairplot (only selected features to avoid clutter)

selected\_cols = ['Age', 'Fare', 'Pclass', 'Survived']

sns.pairplot(df[selected\_cols], hue="Survived", diag\_kind="hist")

plt.suptitle("Pairplot of Selected Features", y=1.02)

plt.show()

# Highly correlated features

cor\_matrix = df[numeric\_cols].corr()

high\_corr = cor\_matrix[(cor\_matrix > 0.7) & (cor\_matrix < 1.0)]

print("Highly Correlated Pairs (corr > 0.7):")

print(high\_corr.dropna(how='all').dropna(axis=1, how='all'))

# Outlier Detection using IQR

for col in numeric\_cols:

    Q1 = df[col].quantile(0.25)

    Q3 = df[col].quantile(0.75)

    IQR = Q3 - Q1

    outliers = df[(df[col] < Q1 - 1.5 \* IQR) | (df[col] > Q3 + 1.5 \* IQR)]

    if not outliers.empty:

        print(f"\nOutliers in {col}:")

        print(outliers[[col]])

# Correlation with target (Survived)

print("\nCorrelation of Numeric Features with 'Survived':")

for col in numeric\_cols:

    if col != 'Survived':

        corr = df['Survived'].corr(df[col])

        print(f"{col}: Correlation = {corr:.2f}")

# Range of features

print("\nFeature Ranges:")

for col in numeric\_cols:

    print(f"{col}: Range = {df[col].max() - df[col].min():.2f}")

OUTPUT:

PassengerId Survived Pclass Name Sex ... Parch Ticket Fare Cabin Embarked

0 1 0 3 Braund, Mr. Owen Harris male ... 0 A/5 21171 7.2500 NaN S

1 2 1 1 Cumings, Mrs. John Bradley (Florence Briggs Th... female ... 0 PC 17599 71.2833 C85 C

2 3 1 3 Heikkinen, Miss. Laina female ... 0 STON/O2. 3101282 7.9250 NaN S

3 4 1 1 Futrelle, Mrs. Jacques Heath (Lily May Peel) female ... 0 113803 53.1000 C123 S

4 5 0 3 Allen, Mr. William Henry male ... 0 373450 8.0500 NaN S

[5 rows x 12 columns]

Summary Statistics:

PassengerId Survived Pclass Age SibSp Parch Fare

count 891.000000 891.000000 891.000000 714.000000 891.000000 891.000000 891.000000

mean 446.000000 0.383838 2.308642 29.699118 0.523008 0.381594 32.204208

std 257.353842 0.486592 0.836071 14.526497 1.102743 0.806057 49.693429

min 1.000000 0.000000 1.000000 0.420000 0.000000 0.000000 0.000000

25% 223.500000 0.000000 2.000000 20.125000 0.000000 0.000000 7.910400

50% 446.000000 0.000000 3.000000 28.000000 0.000000 0.000000 14.454200

75% 668.500000 1.000000 3.000000 38.000000 1.000000 0.000000 31.000000

max 891.000000 1.000000 3.000000 80.000000 8.000000 6.000000 512.329200

Categorical Summary:

Name Sex Ticket Cabin Embarked

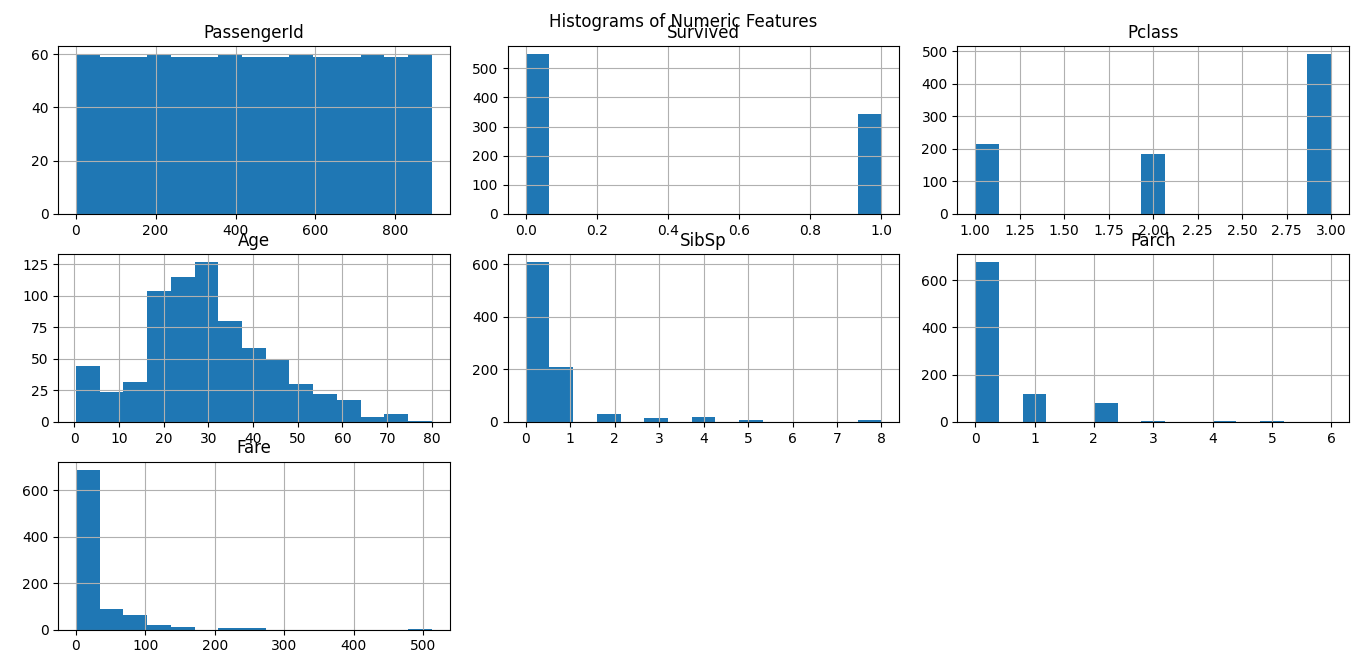
count 891 891 891 204 889

unique 891 2 681 147 3

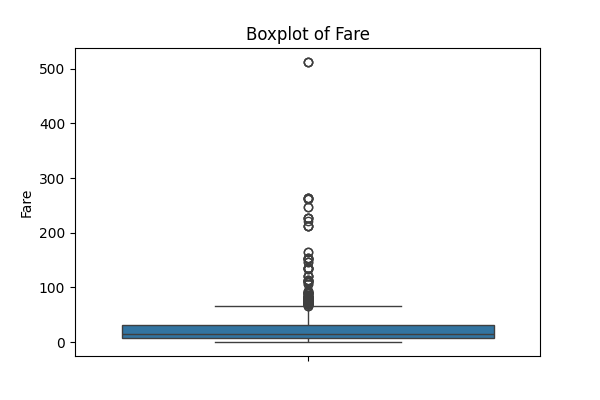
top Dooley, Mr. Patrick male 347082 G6 S

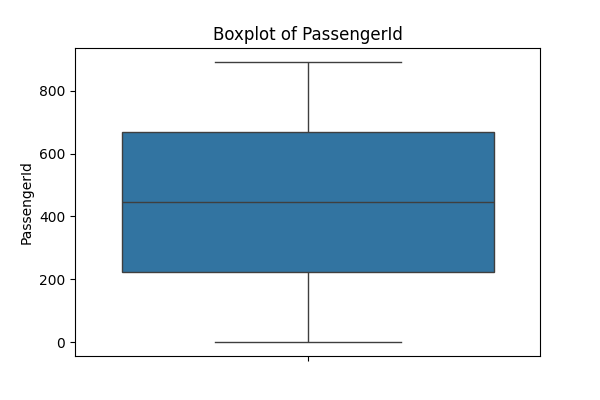
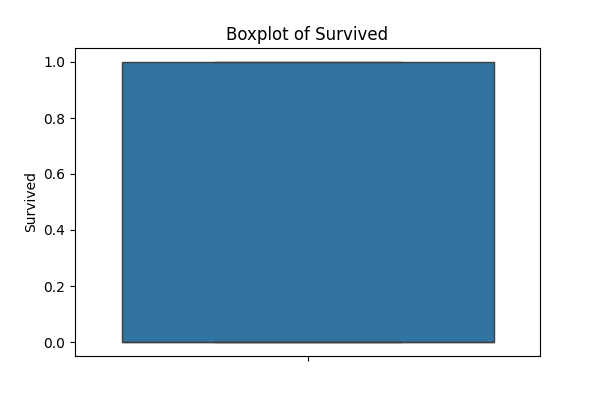
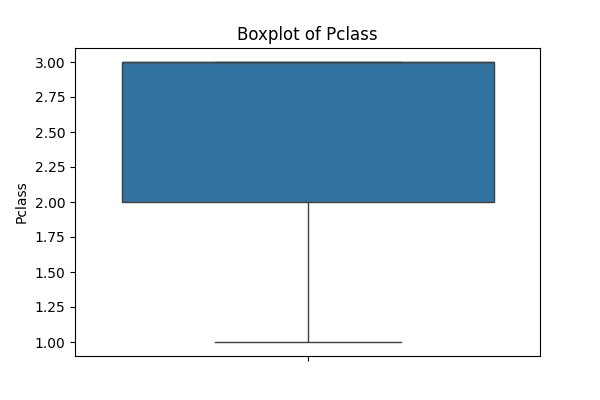
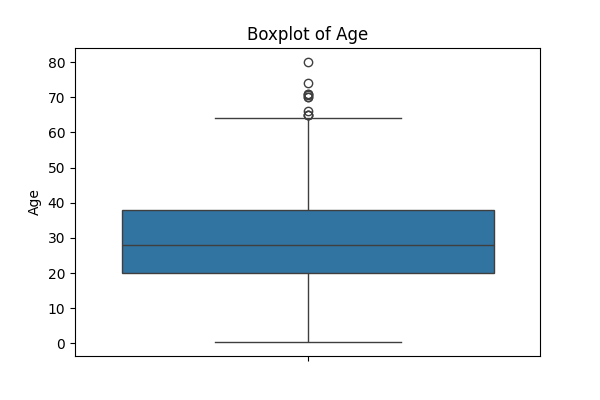
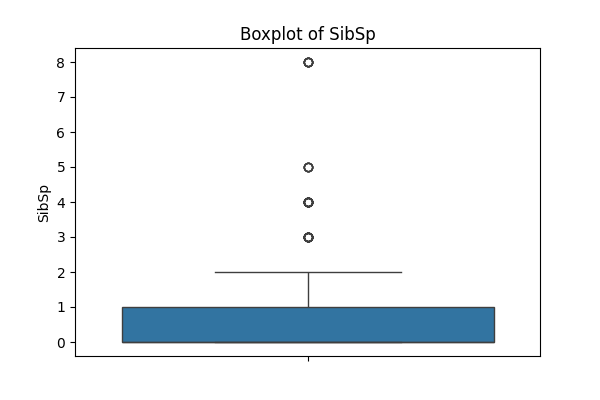
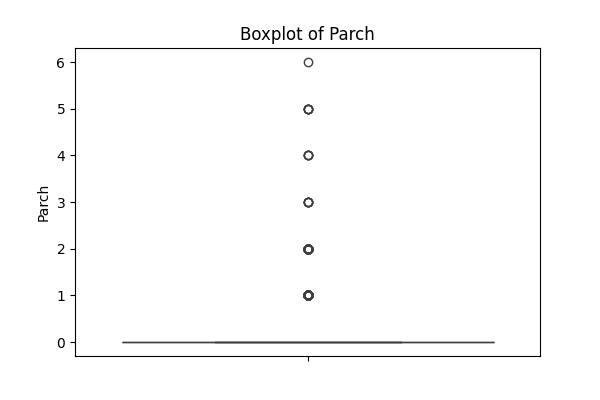
freq 1 577 7 4 644

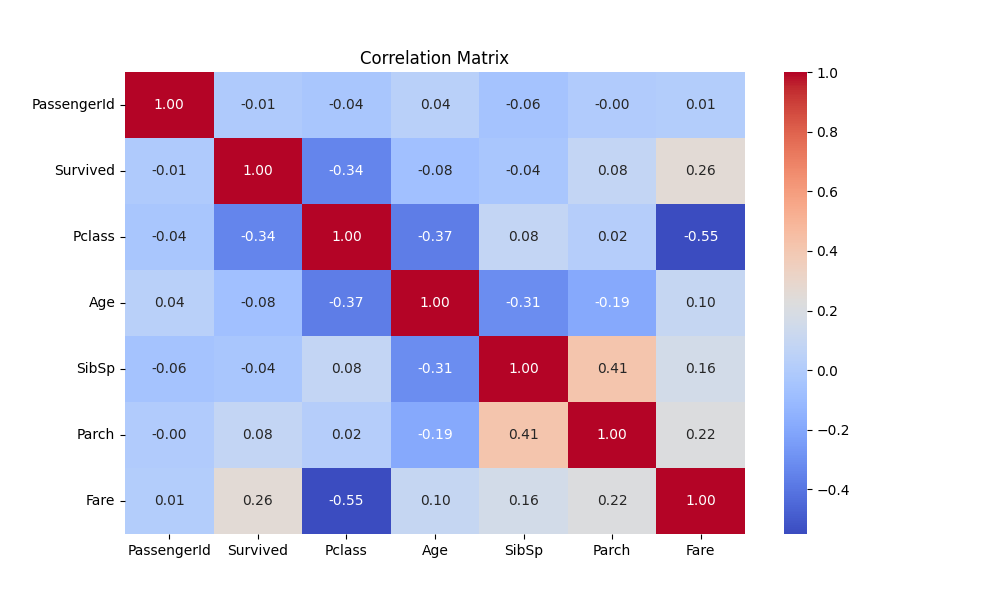
HISTOGRAM:



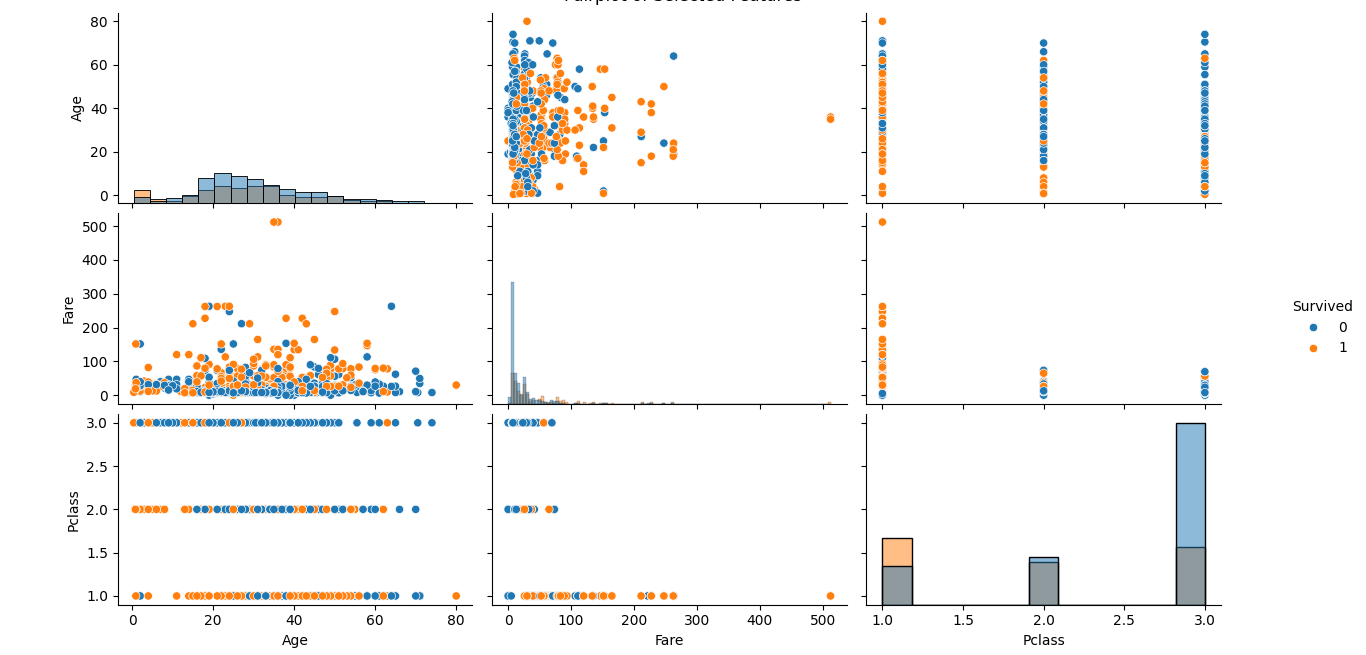
BOXPLOT:







Pairplot:



Highly Correlated Pairs (corr > 0.7):

Empty DataFrame

Columns: []

Index: []

Outliers in Age:

Age

33 66.0

54 65.0

96 71.0

116 70.5

280 65.0

456 65.0

493 71.0

630 80.0

672 70.0

745 70.0

851 74.0

Outliers in SibSp:

SibSp

7 3

16 4

24 3

27 3

50 4

59 5

63 3

68 4

71 5

85 3

88 3

119 4

159 8

164 4

171 4

176 3

180 8

182 4

201 8

229 3

233 4

261 4

266 4

278 4

324 8

341 3

374 3

386 5

409 3

480 5

485 3

541 4

542 4

634 3

642 3

683 5

686 4

726 3

787 4

792 8

813 4

819 3

824 4

846 8

850 4

863 8

Outliers in Parch:

Parch

7 1

8 2

10 1

13 5

16 1

.. ...

871 1

879 1

880 1

885 5

888 2

[213 rows x 1 columns]

Outliers in Fare:

Fare

1 71.2833

27 263.0000

31 146.5208

34 82.1708

52 76.7292

.. ...

846 69.5500

849 89.1042

856 164.8667

863 69.5500

879 83.1583

[116 rows x 1 columns]

Correlation of Numeric Features with 'Survived':

PassengerId: Correlation = -0.01

Pclass: Correlation = -0.34

Age: Correlation = -0.08

SibSp: Correlation = -0.04

Parch: Correlation = 0.08

Fare: Correlation = 0.26

856 164.8667

863 69.5500

879 83.1583

[116 rows x 1 columns]

Correlation of Numeric Features with 'Survived':

PassengerId: Correlation = -0.01

Pclass: Correlation = -0.34

Age: Correlation = -0.08

SibSp: Correlation = -0.04

Parch: Correlation = 0.08

Fare: Correlation = 0.26

863 69.5500

879 83.1583

[116 rows x 1 columns]

Correlation of Numeric Features with 'Survived':

PassengerId: Correlation = -0.01

Pclass: Correlation = -0.34

Age: Correlation = -0.08

SibSp: Correlation = -0.04

Parch: Correlation = 0.08

Fare: Correlation = 0.26

879 83.1583

[116 rows x 1 columns]

Correlation of Numeric Features with 'Survived':

PassengerId: Correlation = -0.01

Pclass: Correlation = -0.34

Age: Correlation = -0.08

SibSp: Correlation = -0.04

Parch: Correlation = 0.08

Fare: Correlation = 0.26

[116 rows x 1 columns]

Correlation of Numeric Features with 'Survived':

PassengerId: Correlation = -0.01

Pclass: Correlation = -0.34

Age: Correlation = -0.08

SibSp: Correlation = -0.04

Parch: Correlation = 0.08

Fare: Correlation = 0.26

PassengerId: Correlation = -0.01

Pclass: Correlation = -0.34

Age: Correlation = -0.08

SibSp: Correlation = -0.04

Parch: Correlation = 0.08

Fare: Correlation = 0.26

Pclass: Correlation = -0.34

Age: Correlation = -0.08

SibSp: Correlation = -0.04

Parch: Correlation = 0.08

Fare: Correlation = 0.26

Fare: Correlation = 0.26

Feature Ranges:

Feature Ranges:

PassengerId: Range = 890.00

Survived: Range = 1.00

Pclass: Range = 2.00

PassengerId: Range = 890.00

Survived: Range = 1.00

Pclass: Range = 2.00

Pclass: Range = 2.00

Age: Range = 79.58

SibSp: Range = 8.00

Age: Range = 79.58

SibSp: Range = 8.00

SibSp: Range = 8.00

Parch: Range = 6.00

Fare: Range = 512.33