Enrollment – 92301794018

Name – Yug Trivedi

Question1: Perform **EDA on an open dataset** (Kaggle/real-world) → highlight relationships, trends, and create three visualizations.

Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
# --- Initial Setup ---
sns.set_style("whitegrid")
plt.rcParams['figure.figsize'] = (10, 6)
# Load the Titanic training data
df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/train.csv')
# Handle missing 'Age' for both EDA and PCA by filling with the median
median_age = df['Age'].median()
df['Age_filled'] = df['Age'].fillna(median_age)
print("First 5 rows of the Data:")
print(df.head())
print("\nColumn Information (Note: Age_filled is the imputed column):")
df.info()
# --- Visualization 1: Survival Rate by Gender ---
plt.figure(figsize=(7, 5))
survival_rate_sex = df.groupby('Sex')['Survived'].mean().reset_index()
```

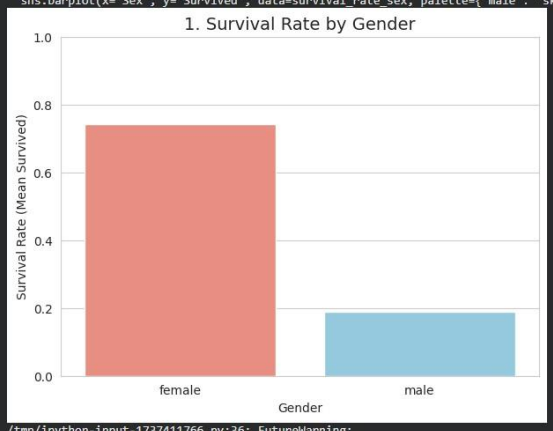
```
sns.barplot(x='Sex', y='Survived', data=survival_rate_sex, palette={'male': 'skyblue', 'female': 'salmon'})

plt.title('1. Survival Rate by Gender', fontsize=14)

plt.ylabel('Survival Rate (Mean Survived)')

plt.xlabel('Gender')

plt.ylim(0, 1)

plt.show() # Display the plot

# --- Visualization 2: Survival Rate by Passenger Class (Pclass) ---

plt.figure(figsize=(7, 5))

survival_rate_pclass = df.groupby('Pclass')['Survived'].mean().reset_index()

sns.barplot(x='Pclass', y='Survived', data=survival_rate_pclass, palette='viridis')

plt.title('2. Survival Rate by Passenger Class', fontsize=14)

plt.ylabel('Survival Rate (Mean Survived)')

plt.xlabel('Passenger Class')

plt.ylim(0, 1)

plt.xticks(ticks=[0, 1, 2], labels=['1st', '2nd', '3rd'])

plt.show() # Display the plot

# --- Visualization 3: Age Distribution of Survivors vs. Non-Survivors ---

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

sns.kdeplot(df[df['Survived'] == 0]['Age_filled'], label='Non-Survivor (0)', color='red', fill=True, alpha=0.5)

sns.kdeplot(df[df['Survived'] == 1]['Age_filled'], label='Survivor (1)', color='green', fill=True, alpha=0.5)

plt.title('3. Age Distribution of Survivors vs. Non-Survivors', fontsize=14)

plt.xlabel('Age (Median Imputed)')

plt.legend(title='Survived')

plt.show() # Display the plot
```

DSE PAC Assignment – Premium

Output:

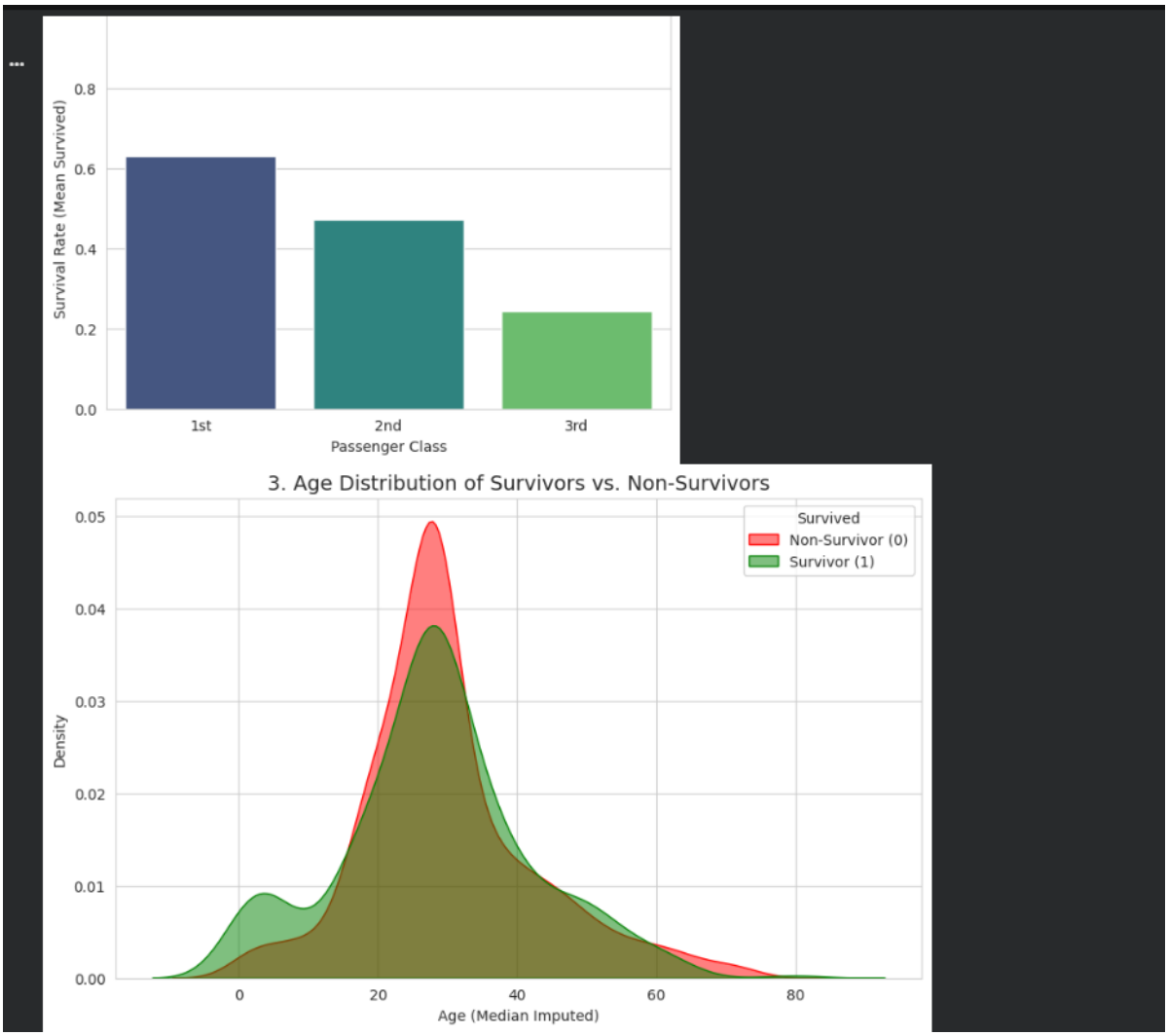
```
First 5 rows of the Data:
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked  Age_filled
0      0         A/5 21171   7.2500   NaN        S        22.0
1      0          PC 17599  71.2833   C85        C        38.0
2      0  STON/O2. 3101282   7.9250   NaN        S        26.0
3      0            113803  53.1000  C123        S        35.0
4      0            373450   8.0500   NaN        S        35.0

Column Information (Note: Age_filled is the imputed column):
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 13 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
 12  Age_filled   891 non-null    float64
dtypes: float64(3), int64(5), object(5)
memory usage: 90.6+ KB
/tmp/ipython-input-1737411766.py:26: FutureWarning:
```

```
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same ef

  sns.barplot(x='Sex', y='Survived', data=survival_rate_sex, palette={'male': 'skyblue', 'female': 'salmon'})
```




```
/tmp/ipython-input-1737411766.py:36: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.
```

DSE PAC Assignment – Premium



Question2: Implement **dimensionality reduction (PCA or feature selection)** on a dataset and explain the improvement in analysis.

Code:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

# --- Initial Setup ---

sns.set_style("whitegrid")

plt.rcParams['figure.figsize'] = (10, 6)

# Load the Titanic training data

df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/train.csv')

# Handle missing 'Age' for both EDA and PCA by filling with the median

median_age = df['Age'].median()

df['Age_filled'] = df['Age'].fillna(median_age)

print("First 5 rows of the Data:")

print(df.head())

print("\nColumn Information (Note: Age_filled is the imputed column):")

df.info()

# --- PCA Implementation ---

# 1. Select and Prepare Features (using the imputed Age_filled)

features = ['Age_filled', 'Fare', 'SibSp', 'Parch']

df_pca = df[features].copy()

# Fill any potentially remaining missing 'Fare' values with the median (for robustness)
```

```
df_pca['Fare'] = df_pca['Fare'].fillna(df_pca['Fare'].median())

# 2. Standardize the data

scaler = StandardScaler()

scaled_features = scaler.fit_transform(df_pca)

# 3. Apply PCA to reduce to 2 components

pca = PCA(n_components=2)

principal_components = pca.fit_transform(scaled_features)

# Create DataFrame for the components

pca_df = pd.DataFrame(data=principal_components, columns=['PC1', 'PC2'])

pca_df['Survived'] = df['Survived']

# Print Explained Variance

print("\n--- PCA Analysis ---")

print("Explained Variance Ratio of the two Principal Components:")

print(pca.explained_variance_ratio_)

print(f"Total Variance Explained by 2 PCs: {pca.explained_variance_ratio_.sum():.2f}")

# 4. Visualization of PCA Result

plt.figure(figsize=(8, 8))

sns.scatterplot(x='PC1', y='PC2', hue='Survived', data=pca_df,

palette=['red', 'green'], alpha=0.7, s=50)

plt.title('PCA of Numerical Titanic Features (2 Components)', fontsize=14)

plt.xlabel(f'Principal Component 1 ({pca.explained_variance_ratio_[0]*100:.1f}% Variance)')

plt.ylabel(f'Principal Component 2 ({pca.explained_variance_ratio_[1]*100:.1f}% Variance)')

plt.legend(title='Survived', labels=['No', 'Yes'])

plt.show() # Display the plot
```

DSE PAC Assignment – Premium

Output:

```
First 5 rows of the Data:
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked  Age_filled
0      0         A/5 21171   7.2500   NaN        S        22.0
1      0          PC 17599  71.2833   C85        C        38.0
2      0  STON/O2. 3101282   7.9250   NaN        S        26.0
3      0            113803  53.1000  C123        S        35.0
4      0            373450   8.0500   NaN        S        35.0

Column Information (Note: Age_filled is the imputed column):
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 13 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
 12  Age_filled   891 non-null    float64
dtypes: float64(3), int64(5), object(5)
memory usage: 90.6+ KB
```

```
--- PCA Analysis ---
Explained Variance Ratio of the two Principal Components:
[0.40662092 0.27577972]
Total Variance Explained by 2 PCs: 0.68
```