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

# Load the data
basic_info = pd.read_excel('/Users/vandana/Downloads/Entertainer Data Analysis 2/
Entertainer - Basic Info.xlsx')
breakthrough_info = pd.read_excel('/Users/vandana/Downloads/Entertainer Data
Analysis 2/Entertainer - Breakthrough Info.xlsx')
last_work = pd.read_excel('/Users/vandana/Downloads/Entertainer Data Analysis 2/
Entertainer - Last work Info.xlsx')

# Display the first few rows of each dataframe
print("Entertainment Basic")
print(basic_info.head())
print("Entertainment Breakthrough")
print(breakthrough_info.head())
print("Entertainment Last Work")
print(last_work.head())

# Display information about the dataframes to understand data types and missing
values
print("Basic Info:")
print(basic_info.info())
print("Breakthrough Info:")
print(breakthrough_info.info())
print("Last Work Info:")
print(last_work.info())

# Check for missing values
missing_basic_info = basic_info.isnull().sum()
missing_breakthrough_info = breakthrough_info.isnull().sum()
missing_last_work_info = last_work.isnull().sum()

print("Missing values in Basic Info Dataset:\n", missing_basic_info)
print("\nMissing values in Breakthrough Info Dataset:\n", missing_breakthrough_info)
print("\nMissing values in Last Work Info Dataset:\n", missing_last_work_info)

# Merge the dataframes on the 'Entertainer' column
combined_df = pd.merge(basic_info, breakthrough_info, on='Entertainer',
how='outer')
combined_df = pd.merge(combined_df, last_work, on='Entertainer', how='outer')

# Check initial data types
print("Initial data types:\n", combined_df.dtypes)

# Remove duplicates
combined_df = combined_df.drop_duplicates()

# Ensure numeric columns are of the correct type
```

```
numeric_columns = ['Year of Breakthrough/#1 Hit/Award Nomination', 'Year of First
Oscar/Grammy/Emmy',
                'Year of Last Major Work (arguable)', 'Year of Death', 'Birth Year']

for column in numeric_columns:
    combined_df[column] = pd.to_numeric(combined_df[column], errors='coerce')

# Ensure string columns are of the correct type
string_columns = ['Entertainer', 'Gender (traditional)', 'Breakthrough Name']

for column in string_columns:
    combined_df[column] = combined_df[column].astype(str)

# Fill missing values in 'Year of Death' with 'Alive'
combined_df['Year of Death'] = combined_df['Year of Death'].fillna('Alive')


# Verify data types
print("Data types after conversion:\n", combined_df.dtypes)

# Display the first few rows of the combined dataframe
print("\nFirst few rows of the combined dataframe:\n", combined_df.head())

# Distribution of Entertainers by gender
plt.figure(figsize=(10, 6))
sns.countplot(data=combined_df, x='Gender (traditional)')
plt.title('Distribution of Entertainers by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()

# Age at Breakthrough and First Major Award
combined_df['Age at Breakthrough'] = combined_df['Year of Breakthrough/#1 Hit/
Award Nomination'] - combined_df['Birth Year']
combined_df['Age at First Major Award'] = combined_df['Year of First Oscar/
Grammy/Emmy'] - combined_df['Birth Year']

plt.figure(figsize=(12, 6))
sns.histplot(combined_df['Age at Breakthrough'].dropna(), bins=20, kde=True)
plt.title('Age at Breakthrough')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()

plt.figure(figsize=(12, 6))
sns.histplot(combined_df['Age at First Major Award'].dropna(), bins=20, kde=True)
plt.title('Age at First Major Award')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

```
# Career duration (from breakthrough to last major work)
combined_df['Career Duration'] = combined_df['Year of Last Major Work (arguable)']
- combined_df['Year of Breakthrough/#1 Hit/Award Nomination']

plt.figure(figsize=(12, 6))
sns.histplot(combined_df['Career Duration'].dropna(), bins=20, kde=True)
plt.title('Career Duration (from Breakthrough to Last Major Work)')
plt.xlabel('Duration (years)')
plt.ylabel('Frequency')
plt.show()

# Number of Entertainers Who Have Passed Away
passed_away_count = combined_df['Year of Death'].apply(lambda x: x !=
'Alive').sum()
alive_count = combined_df['Year of Death'].apply(lambda x: x == 'Alive').sum()

plt.figure(figsize=(8, 8))
plt.pie([passed_away_count, alive_count], labels=['Passed Away', 'Alive'],
autopct='%1.1f%%', startangle=140)
plt.title('Number of Entertainers Who Have Passed Away')
plt.show()

# Save the combined dataframe to a CSV file
combined_df.to_csv('/Users/vandana/Downloads/combined_data.csv', index=False)

# Verify that the data was saved correctly
print("\nFirst few rows of the combined dataframe saved to CSV:\n",
combined_df.head())
```