python-internship-project

September 27, 2024

[31]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[32]:
```
csv = pd.read_csv('Entertainer - Basic Info.csv')
csv1 = pd.read_csv('Entertainer - Breakthrough Info.csv')
csv2 = pd.read_csv('Entertainer - Last work Info.csv')
```

[33]:
```
csv.shape
```

[33]:
```
(70, 3)
```

[34]:
```
csv.isnull().sum()
```

[34]:
```
Entertainer             0
Gender (traditional)    0
Birth Year              0
dtype: int64
```

[35]:
```
csv = csv.dropna()
```

[36]:
```
csv.shape
```

[36]:
```
(70, 3)
```

[37]:
```
csv1.shape
```

[37]:
```
(70, 4)
```

[38]:
```
csv1.isnull().sum()
```

[38]:
```
Entertainer                                   0
Year of Breakthrough/#1 Hit/Award Nomination    0
Breakthrough Name                             0
Year of First Oscar/Grammy/Emmy               6
dtype: int64
```

```
[39]: csv1 = csv1.dropna()
```

```
[40]: csv1.shape
```

```
[40]: (64, 4)
```

```
[41]: csv2.shape
```

```
[41]: (70, 3)
```

```
[42]: csv2.isnull().sum()
```

```
[42]: Entertainer                            0
      Year of Last Major Work (arguable)     0
      Year of Death                         40
      dtype: int64
```

```
[43]: csv2 = csv2.dropna()
```

```
[44]: csv2.shape
```

```
[44]: (30, 3)
```

```
[45]: merged_df = csv.merge(csv1, on='Entertainer').merge(csv2, on='Entertainer')
```

```
[46]: merged_df.isnull().sum()
```

```
[46]: Entertainer                                     0
      Gender (traditional)                            0
      Birth Year                                      0
      Year of Breakthrough/#1 Hit/Award Nomination    0
      Breakthrough Name                               0
      Year of First Oscar/Grammy/Emmy                 0
      Year of Last Major Work (arguable)              0
      Year of Death                                   0
      dtype: int64
```

0.1 What is the correlation between the birth year and the year of breakthrough?

```
[47]: correlation = merged_df['Birth Year'].corr(merged_df['Year of Breakthrough/#1␣
        ↪Hit/Award Nomination'])
      print('The correlation between the birth year and the year of breakthrough␣
        ↪is',np.round(correlation,2))
```

```
The correlation between the birth year and the year of breakthrough is 0.87
```

0.2 What is the average age at which entertainers had their breakthrough?

[48]:
```
merged_df['Age at Breakthrough'] = merged_df['Year of Breakthrough/#1 Hit/Award␣
 ↪Nomination'] - merged_df['Birth Year']
average_age_breakthrough = merged_df['Age at Breakthrough'].mean()
print('The Average age for entertainers to have a breakthrough is',np.round(␣
 ↪average_age_breakthrough,),'years')
```

```
The Average age for entertainers to have a breakthrough is 30.0 years
```

0.3 Who has had the longest career based on the difference between breakthrough year and last major work?

[49]:
```
merged_df['Career Length'] = merged_df['Year of Last Major Work (arguable)'] -␣
 ↪merged_df['Year of Breakthrough/#1 Hit/Award Nomination']
longest_career_entertainer = merged_df.loc[merged_df['Career Length'].idxmax()]
longest_career_entertainer
```

[49]:
```
Entertainer                                             Katherine Hepburn
Gender (traditional)                                                    F
Birth Year                                                           1907
Year of Breakthrough/#1 Hit/Award Nomination                         1933
Breakthrough Name                                           Morning Glory
Year of First Oscar/Grammy/Emmy                                    1933.0
Year of Last Major Work (arguable)                                   1994
Year of Death                                                      2003.0
Age at Breakthrough                                                    26
Career Length                                                          61
Name: 19, dtype: object
```

0.4 What is the average career length of entertainers?

[50]:
```
average_career_length = merged_df['Career Length'].mean()
print(f"Average Career Length: {average_career_length:.2f} years")
```

```
Average Career Length: 33.88 years
```

0.5 Which decade had the most breakthroughs?

[51]:
```
merged_df['Breakthrough Decade'] = (merged_df['Year of Breakthrough/#1 Hit/
 ↪Award Nomination'] // 10) * 10
decade_most_breakthroughs = merged_df['Breakthrough Decade'].value_counts().
 ↪idxmax()
decade_most_breakthroughs_count = merged_df['Breakthrough Decade'].
 ↪value_counts().max()
decade_most_breakthroughs, decade_most_breakthroughs_count
```

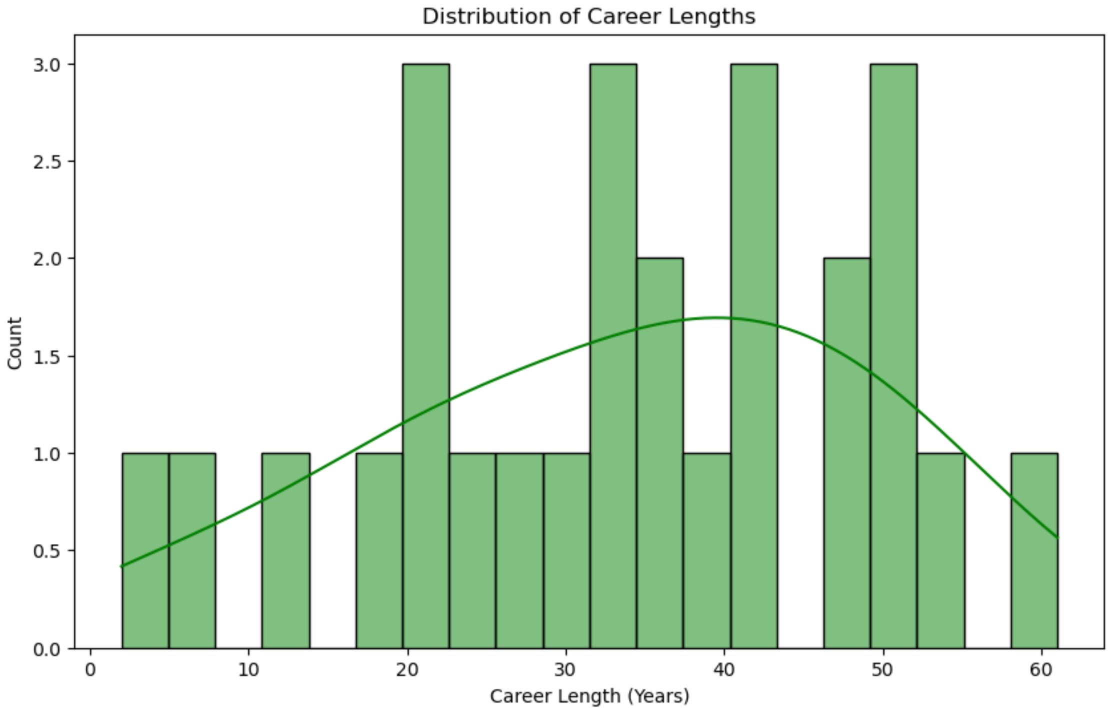
[51]:
```
(1930, 10)
```

[ ]:
```

```

0.6 Distribution of Career Lengths

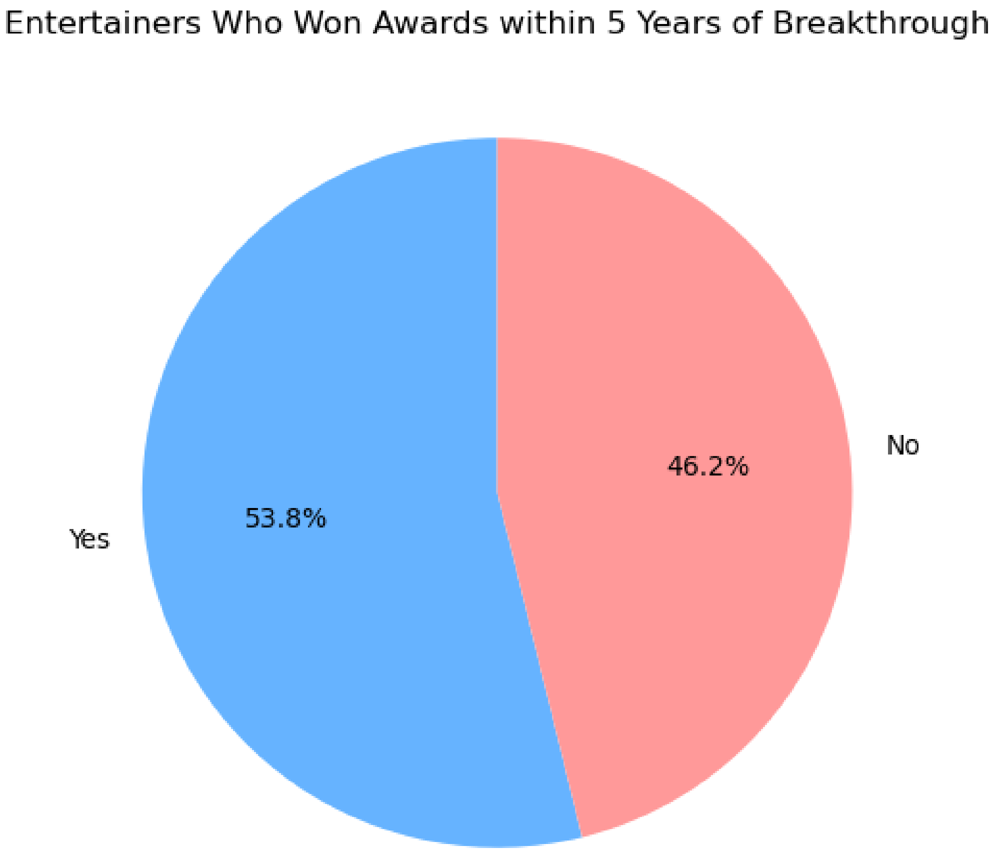
[52]:
```
plt.figure(figsize=(10, 6))
sns.histplot(merged_df['Career Length'], bins=20, kde=True, color='green')
plt.title('Distribution of Career Lengths')
plt.xlabel('Career Length (Years)')
plt.ylabel('Count')
plt.show()
```



[53]:
```
merged_df['Award within 5 years'] = (merged_df['Year of First Oscar/Grammy/
  ↪Emmy'] - merged_df['Year of Breakthrough/#1 Hit/Award Nomination']) <= 5

# Plot the distribution
award_dist = merged_df['Award within 5 years'].value_counts()

plt.figure(figsize=(8, 6))
plt.pie(award_dist, labels=['Yes', 'No'], autopct='%1.1f%%', colors=['#66b3ff',␣
  ↪'#ff9999'], startangle=90)
plt.title('Entertainers Who Won Awards within 5 Years of Breakthrough')
plt.show()
```

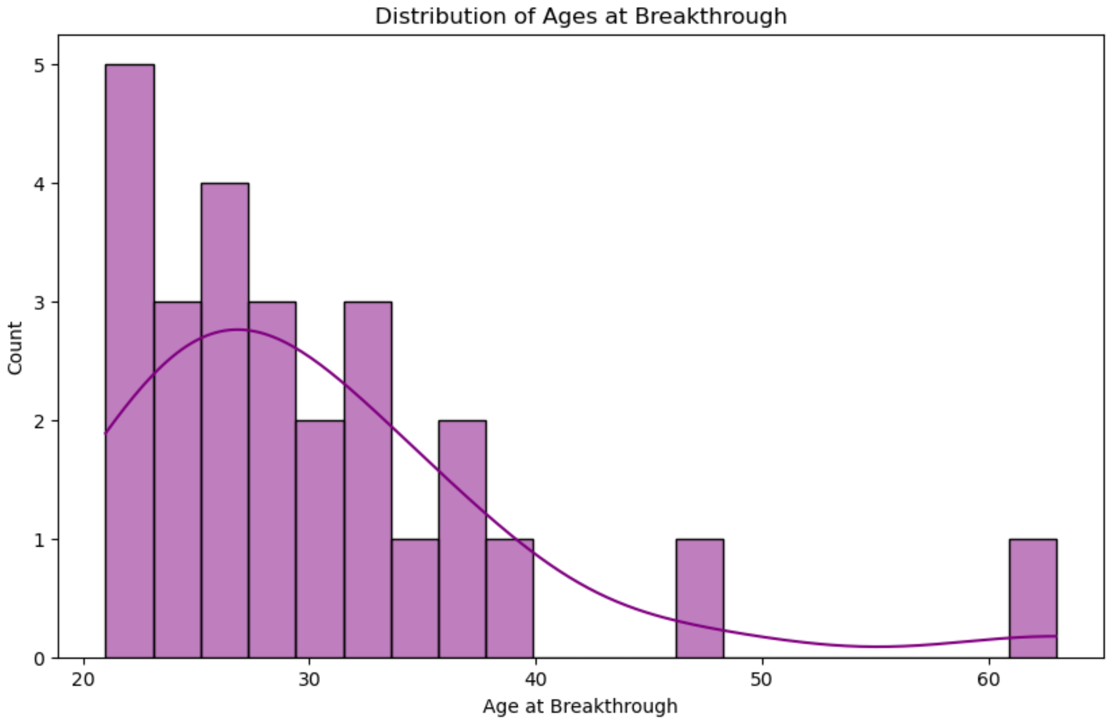


0.7 Distribution of Ages at Breakthrough

[54]:

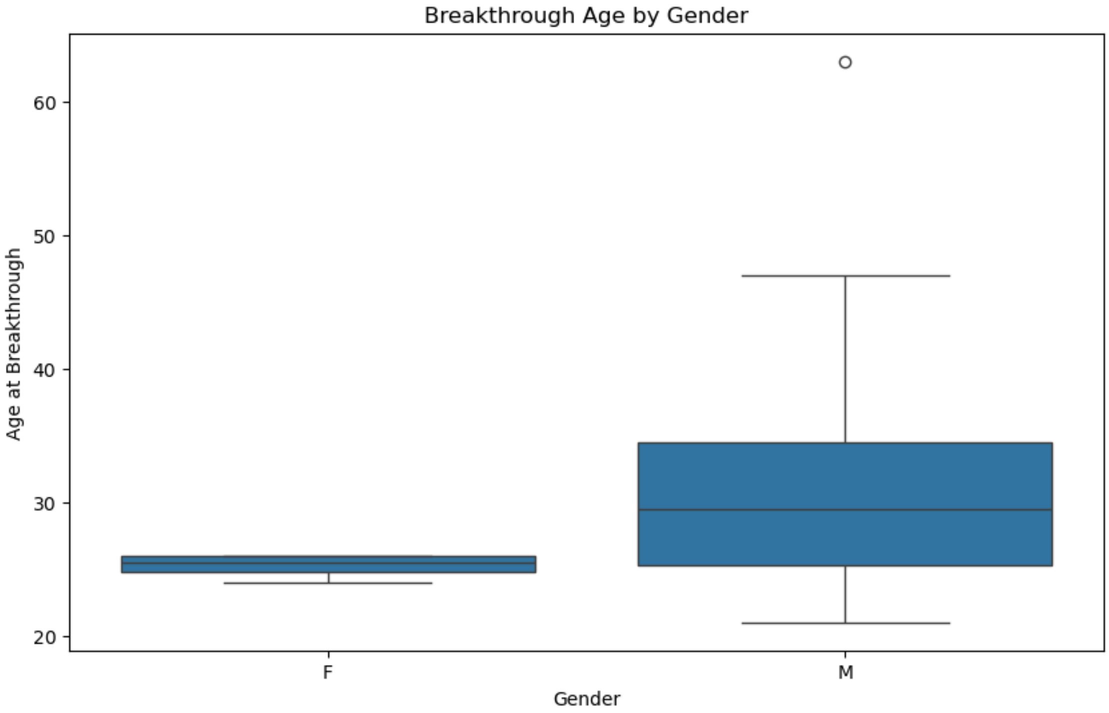
```
merged_df['Age at Breakthrough'] = merged_df['Year of Breakthrough/#1 Hit/Award␣
 ↪Nomination'] - merged_df['Birth Year']

plt.figure(figsize=(10, 6))
sns.histplot(merged_df['Age at Breakthrough'], bins=20, kde=True,␣
 ↪color='purple')
plt.title('Distribution of Ages at Breakthrough')
plt.xlabel('Age at Breakthrough')
plt.ylabel('Count')
plt.show()
```



0.8 Comparison of Breakthrough Ages by Gender

[55]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Gender (traditional)', y='Age at Breakthrough', data=merged_df)
plt.title('Breakthrough Age by Gender')
plt.xlabel('Gender')
plt.ylabel('Age at Breakthrough')
plt.show()
```



0.9 Entertainers with Longest Careers

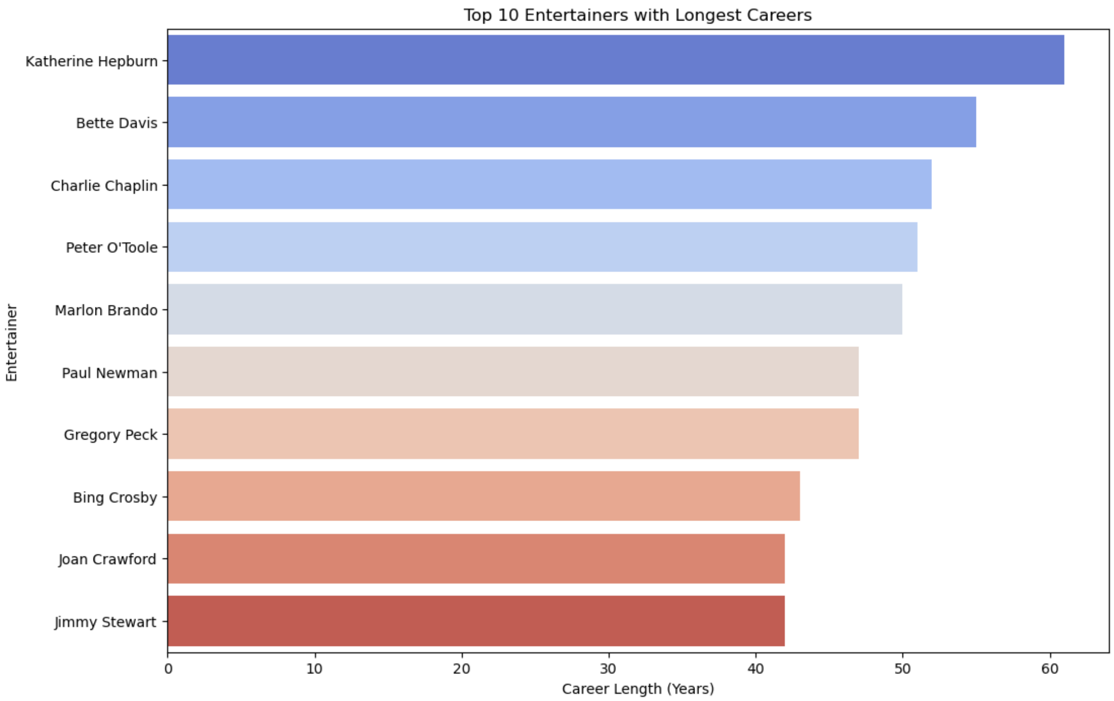
[56]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# Create a new column for career length
merged_df['Career Length'] = merged_df['Year of Last Major Work (arguable)'] -␣
↪merged_df['Year of Breakthrough/#1 Hit/Award Nomination']

# Sort the dataframe by career length and select the top 10 entertainers
top_10_longest_careers = merged_df.sort_values(by='Career Length',␣
↪ascending=False).head(10)

# Create a bar plot for the top 10 longest careers
plt.figure(figsize=(12, 8))
sns.barplot(x='Career Length', y='Entertainer', data=top_10_longest_careers,␣
↪palette='coolwarm', hue='Entertainer', dodge=False)

plt.title('Top 10 Entertainers with Longest Careers')
plt.xlabel('Career Length (Years)')
plt.ylabel('Entertainer')
plt.show()
```


Top 10 Entertainers with Longest Careers
Entertainer
Katherine Hepburn
Bette Davis
Charlie Chaplin
Peter O'Toole
Marlon Brando
Paul Newman
Gregory Peck
Bing Crosby
Joan Crawford
Jimmy Stewart
0
10
20
30
40
50
60
Career Length (Years)

[ ]:

[ ]:

[ ]: