Detailed Project Report

Entertainer Data Analysis

Table of Contents:

1. Executive Summary

The Entertainer Data Analysis project examined career trajectories of notable entertainers across multiple decades. By processing and visualizing data from various sources, we uncovered significant trends in breakthrough timing, gender representation, and award recognition. Key findings include a peak of breakthroughs in the 1950s-60s, a substantial gender imbalance favouring male entertainers, and a strong correlation between breakthrough years and first major awards. The analysis provides valuable insights into the historical evolution of the entertainment industry and raises important questions about current trends and future directions.

2. Project Overview

This project aimed to analyse the careers of entertainers by combining data from multiple sources, cleaning and processing the data, and creating visualizations to uncover insights. The analysis focused on breakthrough moments, gender representation, age at key career milestones, and the relationship between breakthroughs and major award recognition.

3. Data Processing

3.1 Data Loading

Four Excel files were loaded using pandas:

- "Copy of Entertainer - Last work Info.xlsx"

- "Entertainer - Basic Info.xlsx"

- "Entertainer - Breakthrough Info.xlsx"

- "Entertainer - Last work Info.xlsx"

Data Loading

The code loads four Excel files containing information about entertainers using pandas.

In [3]:
```
df1 = pd.read_excel("Copy of Entertainer - Last work Info.xlsx")
df2 = pd.read_excel("Entertainer - Basic Info.xlsx")
df3 = pd.read_excel("Entertainer - Breakthrough Info.xlsx")
df4 = pd.read_excel("Entertainer - Last work Info.xlsx")
```

3.2 Data Exploration

The structure and content of each data frame were examined using. head () and .info () methods. A consistency check confirmed that df1, df2, and df4 were identical.

Data Exploration

It examines the structure and content of each dataframe using methods like .head() and .info().

In [4]:
```
df1.head()
```

Out[4]:

	Entertainer	Gender (traditional)	Birth Year
0	Adele	F	1988
1	Angelina Jolie	F	1975
2	Aretha Franklin	F	1942
3	Bette Davis	F	1908
4	Betty White	F	1922

In [5]:
```
df2.head()
```

Out[5]:

	Entertainer	Gender (traditional)	Birth Year
0	Adele	F	1988
1	Angelina Jolie	F	1975
2	Aretha Franklin	F	1942
3	Bette Davis	F	1908
4	Betty White	F	1922

In [6]:
```
df4.head()
```

Out[6]:

	Entertainer	Gender (traditional)	Birth Year
0	Adele	F	1988
1	Angelina Jolie	F	1975
2	Aretha Franklin	F	1942
3	Bette Davis	F	1908
4	Betty White	F	1922

In [7]:
```
df3.head()
```

Out[7]:

	Entertainer	Year of Breakthrough/#1 Hit/Award Nomination	Breakthrough Name	Year of First Oscar/Grammy/Emmy
0	Adele	2008	19	2009.0
1	Angelina Jolie	1999	Girl, Interrupted	1999.0
2	Aretha Franklin	1967	I Never Loved a Man (The Way I Love You)	1968.0
3	Bette Davis	1934	Of Human Bondage	1935.0
4	Betty White	1952	Life with Elilzabeth	1976.0

In [8]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70 entries, 0 to 69
Data columns (total 3 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   Entertainer           70 non-null     object
 1   Gender (traditional)  70 non-null     object
 2   Birth Year            70 non-null     int64 
dtypes: int64(1), object(2)
memory usage: 1.8+ KB
```

In [9]:
```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70 entries, 0 to 69
Data columns (total 3 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   Entertainer           70 non-null     object
 1   Gender (traditional)  70 non-null     object
 2   Birth Year            70 non-null     int64 
dtypes: int64(1), object(2)
memory usage: 1.8+ KB
```

In [10]:
```
df3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70 entries, 0 to 69
Data columns (total 4 columns):
 #   Column                                       Non-Null Count  Dtype  
---  ------                                       --------------  -----  
 0   Entertainer                                  70 non-null     object 
 1   Year of Breakthrough/#1 Hit/Award Nomination  70 non-null     int64  
 2   Breakthrough Name                            70 non-null     object 
 3   Year of First Oscar/Grammy/Emmy              64 non-null     float64
dtypes: float64(1), int64(1), object(2)
memory usage: 2.3+ KB
```

In [11]:
```
df4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70 entries, 0 to 69
Data columns (total 3 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   Entertainer           70 non-null     object
 1   Gender (traditional)  70 non-null     object
 2   Birth Year            70 non-null     int64 
dtypes: int64(1), object(2)
memory usage: 1.8+ KB
```

Data Consistency Check

The code verifies if df1, df2, and df4 are identical using the .equals() method.

In [13]:
```
print(df1.equals(df4))
```

```
True
```

3.3 Handling Missing Data

Null values were identified and removed from df3 to create a cleaned version.

Handling Missing Data

It checks for null values in the dataframes

```
In [14]: df3.isnull().sum()
```

```
Out[14]: Entertainer                                0
         Year of Breakthrough/#1 Hit/Award Nomination    0
         Breakthrough Name                          0
         Year of First Oscar/Grammy/Emmy            6
         dtype: int64
```

```
In [15]: df3_cleaned = df3.dropna()
```

```
In [16]: df3_cleaned.isnull().sum()
```

```
Out[16]: Entertainer                                0
         Year of Breakthrough/#1 Hit/Award Nomination    0
         Breakthrough Name                          0
         Year of First Oscar/Grammy/Emmy            0
         dtype: int64
```

3.4 Data Merging

df1 was merged with the cleaned df3 on the 'Entertainer' column to combine basic info with breakthrough data.

Data Merging

The code merges df1 with the cleaned df3 on the 'Entertainer' column to combine basic info with breakthrough data.

```
In [17]: # Merge df1 with df3_cleaned on the 'Entertainer' column
         merged_df = df1.merge(df3_cleaned, on='Entertainer', how='left')
```

```
In [18]: merged_df
```

Out[18]:

	Entertainer	Gender (traditional)	Birth Year	Year of Breakthrough/#1 Hit/Award Nomination	Breakthrough Name	Year of First Oscar/Grammy/Emmy
0	Adele	F	1988	2008.0	19	2009.0
1	Angelina Jolie	F	1975	1999.0	Girl, Interrupted	1999.0
2	Aretha Franklin	F	1942	1967.0	I Never Loved a Man (The Way I Love You)	1968.0
3	Bette Davis	F	1908	1934.0	Of Human Bondage	1935.0
4	Betty White	F	1922	1952.0	Life with Elilzabeth	1976.0
...	...	...	...	...	...	...
65	Tom Hanks	M	1956	1984.0	Splash	1993.0
66	Tony Bennett	M	1926	1951.0	Because of You	1963.0
67	Wayne Newton	M	1942	NaN	NaN	NaN
68	Will Smith	M	1968	1990.0	The Fresh Prince of Bel-Air	1988.0
69	Willie Nelson	M	1933	1975.0	Red Headed Stranger	1976.0

70 rows × 6 columns

3.5 Feature Engineering

New columns were created:

- 'Age': Calculated as 2024 - Birth Year

- 'Decade of Breakthrough': Derived from 'Year of Breakthrough/#1 Hit/Award Nomination'

Feature Engineering

It creates new columns for 'Age' and 'Decade of Breakthrough' to enable further analysis.

In [19]:
```
merged_df['Age'] = 2024 - merged_df['Birth Year']
merged_df['Decade of Breakthrough'] = (merged_df['Year of Breakthrough/#1 Hit/Award Nomination'] // 10) * 10
```

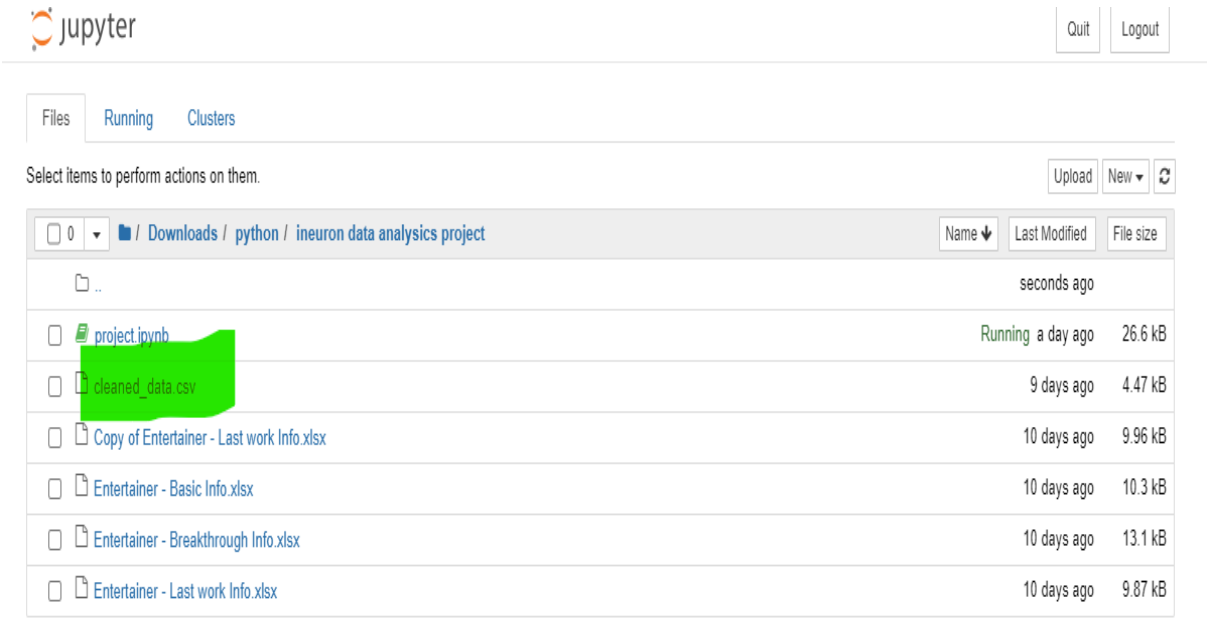
3.6 Data Export

The cleaned and processed data was exported to 'cleaned_data.csv'.

Data Export

Finally, the cleaned and processed data is exported to a CSV file for future use.

In [20]:
```
#Export Cleaned Data
merged_df.to_csv('cleaned_data.csv', index=False)
```

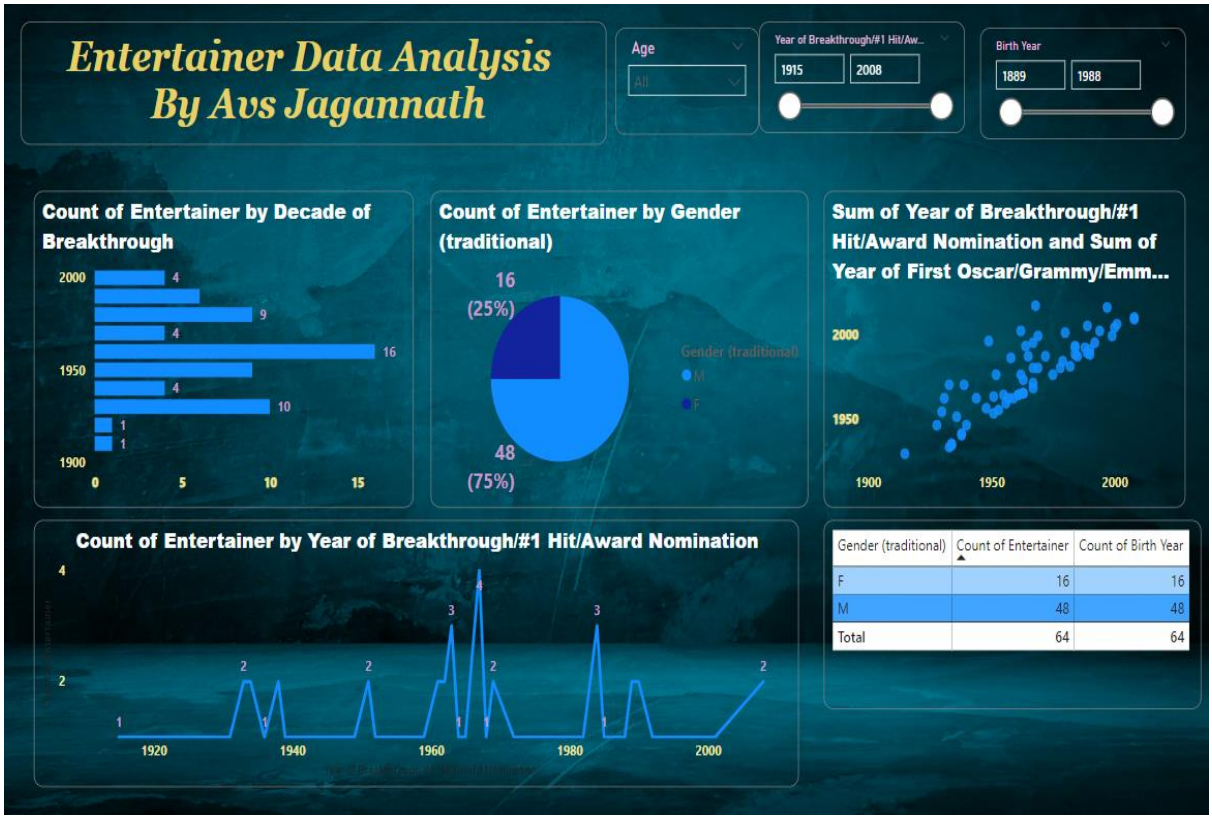


4. Data Visualization

4.1 Dashboard 1: Overview and Trends

Components:

- Filters for Age, Year of Breakthrough, and Birth Year
- Bar chart: Count of Entertainer by Decade of Breakthrough
- Pie chart: Gender distribution
- Scatter plot: Breakthrough Year vs. First Major Award Year
- Line chart: Breakthroughs by Year
- Gender breakdown table



4.2 Dashboard 2: Detailed Career Trajectories

A comprehensive table view displaying:

Entertainer, Gender, Birth Year, Breakthrough Year, First Award Year, Age, Decade of Breakthrough, Age at Breakthrough, Age at First Award, Breakthrough Name

Entertainer	Gender (traditional)	Birth Year	Year of Breakthrough/#1 Hit/Award Nomination	Year of First Oscar/Grammy /Emmy	Age	Decade of Breakthrough	Age at Breakthrough	Age at First Award	Breakthrough Name	Decade of Breakthrough1
Adele	F	1988	2008	2009	36	2000	20	21	19	2000
Angelina Jolie	F	1975	1999	1999	49	1990	24	24	Girl, Interrupted	1990
Aretha Franklin	F	1942	1967	1968	82	1960	25	26	I Never Loved a Man (The Way I Love You)	1960
Bette Davis	F	1908	1934	1935	116	1930	26	27	Of Human Bondage	1930
Betty White	F	1922	1952	1976	102	1950	30	54	Life with Elilzabeth	1950
Bing Crosby	M	1903	1931	1962	121	1930	28	59	Several Songs	1930
Bob Hope	M	1903	1938	1940	121	1930	35	37	The Big Broadcast of 1938	1930
Carol Burnett	F	1933	1959	1962	91	1950	26	29	The Garry Moore Show	1950

5. Key Insights and Analysis

5.1 Breakthrough Trends Over Time

- Peak in the 1950s and 1960s with 16 breakthroughs in the 1960s alone
- Decline in recorded breakthroughs in recent decades
- Sporadic peaks in specific years (e.g., 1960 and 1980)

5.2 Gender Representation in Entertainment

- 75% male (48 entertainers) vs. 25% female (16 entertainers)
- Reflects historical gender inequalities in the industry

5.3 Correlation Between Breakthroughs and Major Awards

- Strong positive correlation between breakthrough year and first major award year
- Most entertainers received their first major award close to their breakthrough year
- Notable outliers suggest interesting stories of late recognition or early exceptional success

5.4 Age and Career Milestones

- Many entertainers had breakthroughs in their mid-20s (e.g., Charlie Chaplin, Joan Crawford)
- Varied time gaps between breakthrough and first major award (e.g., 14 years for Chaplin, same year for Hepburn)

5.5 Historical Context and Industry Evolution

- 1950s-60s peak coincides with the golden age of Hollywood and rise of rock and roll
- Decline in recent breakthroughs may indicate changing industry dynamics or data collection bias

6. Conclusions

The analysis reveals significant patterns in entertainment careers over the past century. The 1950s-60s emerge as a golden era for breakthroughs, likely due to technological advancements and cultural shifts. The substantial gender imbalance highlights historical inequalities that persist in the industry. The strong correlation between breakthroughs and first major awards underscores the importance of initial success in an entertainer's career trajectory. The varied career paths observed demonstrate that success in entertainment can take many forms, from early recognition to late-blooming talent.

7. Appendix: Technical Notes

- Data processing and analysis were conducted using Python and pandas

- Visualizations were created using an unspecified dashboard tool (PowerBI)

- The project's code and data files are stored separately and can be accessed for further analysis or replication of results

This comprehensive report provides a multifaceted view of entertainer careers, offering insights into historical trends, industry dynamics, and individual career trajectories. It serves as a foundation for understanding the evolution of the entertainment industry and can inform future studies on talent development, industry representation, and the changing nature of success in entertainment.