MICROSOFT MOVIES ANALYSIS

- Student name: Frida Oyucho
- Student pace: full time
- Instructor name: William Okombo/Noah Kandie

Research Questions:

- What is the average duration of a movie?
- Which are the top 10 movie genres?
- Which movie genres have the highest revenue?
- Which is the highest rated movie genre?
- What is the correlation between movies and revenue variables collected?

Data Exploration and Visualization

In [1]:

```
# Importing necessary Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
# importing the cleaned dataset and reading as a data frame

df = pd.read_csv('CleanedData.csv', index_col=0)
df.head()
```

Out[2]:

	popularity	title	vote_average	tconst	primary_title	start_year	ru
original_language							
en	3.952	Ceremony	5.2	tt1341341	Ceremony	2010	
en	3.948	Legendary	6.4	tt1563704	Legendary	2010	
en	3.948	The Stranger	5.0	tt2238470	The Stranger	2014	
en	4.270	The Stranger	4.8	tt2238470	The Stranger	2014	
en	4.270	The Stranger	4.8	tt4016718	The Stranger	2014	

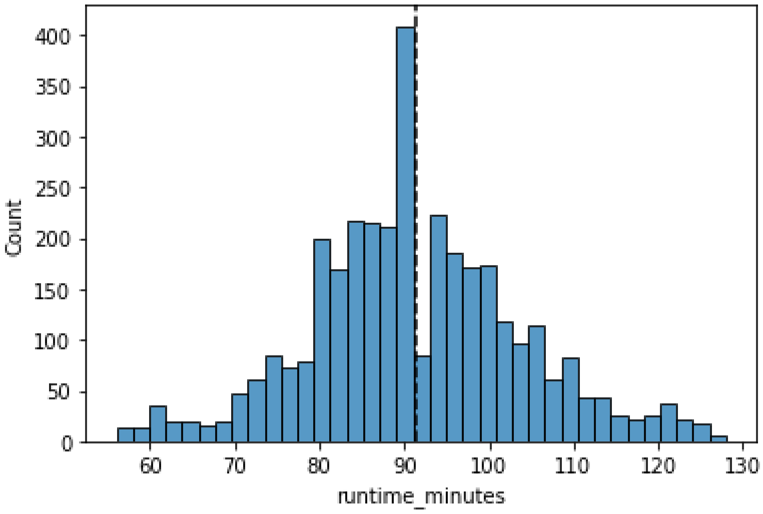
Univariate Analysis

In [3]:

```
#Average duration of a movie that people tend to watch
sns.histplot(data=df['runtime_minutes'],bins='auto')
average_runtime_minutes = df['runtime_minutes'].mean()
plt.axvline(average_runtime_minutes, color='black', linestyle='dashed', lin
```

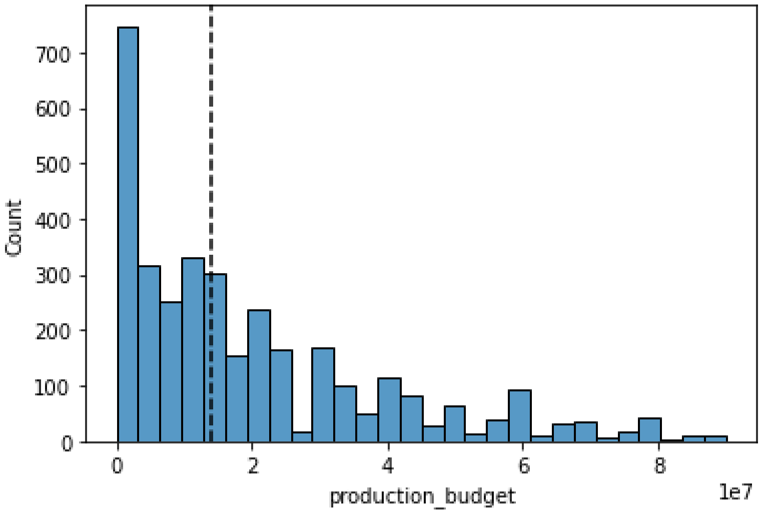
Out[3]:

```
<matplotlib.lines.Line2D at 0x267e8b565e0>
```



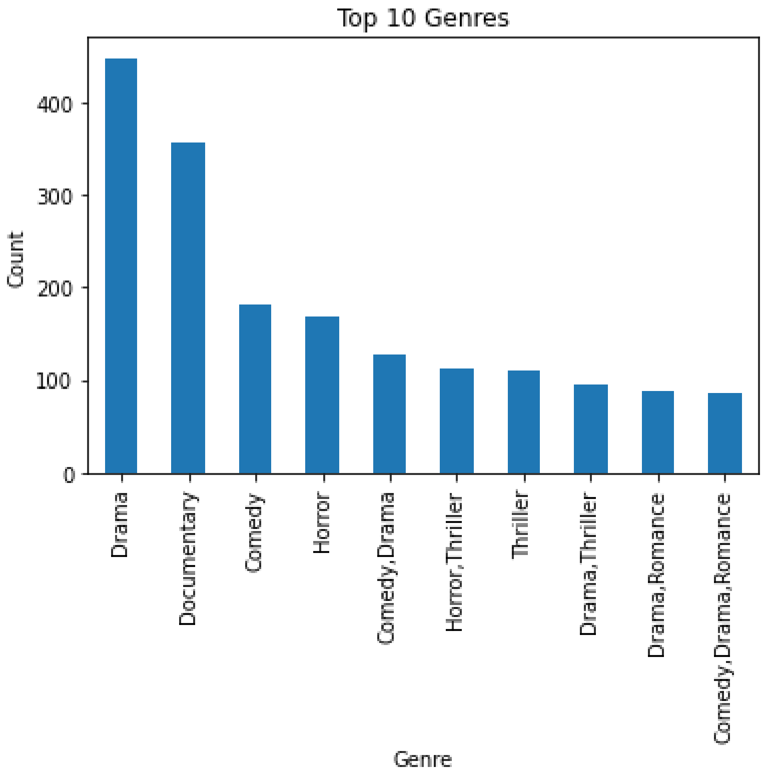
In [4]:

```
sns.histplot(data=df['production_budget'],bins='auto')
average_production_budget = df['production_budget'].median()
plt.axvline(average_production_budget, color='black', linestyle='dashed', l
```



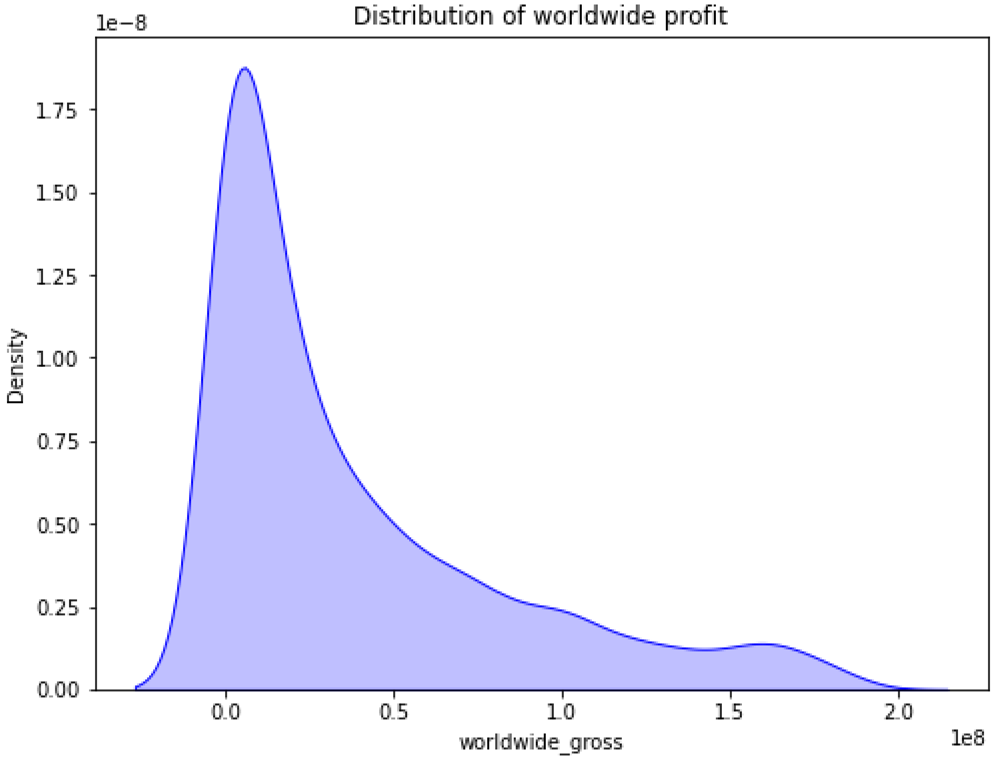
In [5]:

```
# Get the count of each genre
genre_counts = df['genres'].value_counts()

# Get the top 10 genres
top_10_genres = genre_counts.head(10)

# Create a bar plot of the top 10 genres
top_10_genres.plot(kind='bar')

# Add a title and labels to the plot
plt.title('Top 10 Genres')
plt.xlabel('Genre')
plt.ylabel('Count')

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



In [6]:

```
# Plot density plot
plt.figure(figsize=(8, 6))
sns.kdeplot(df['worldwide_gross'], shade=True, color='blue')
plt.title('Distribution of worldwide profit')
plt.xlabel('worldwide_gross')
plt.ylabel('Density')
plt.show()
```



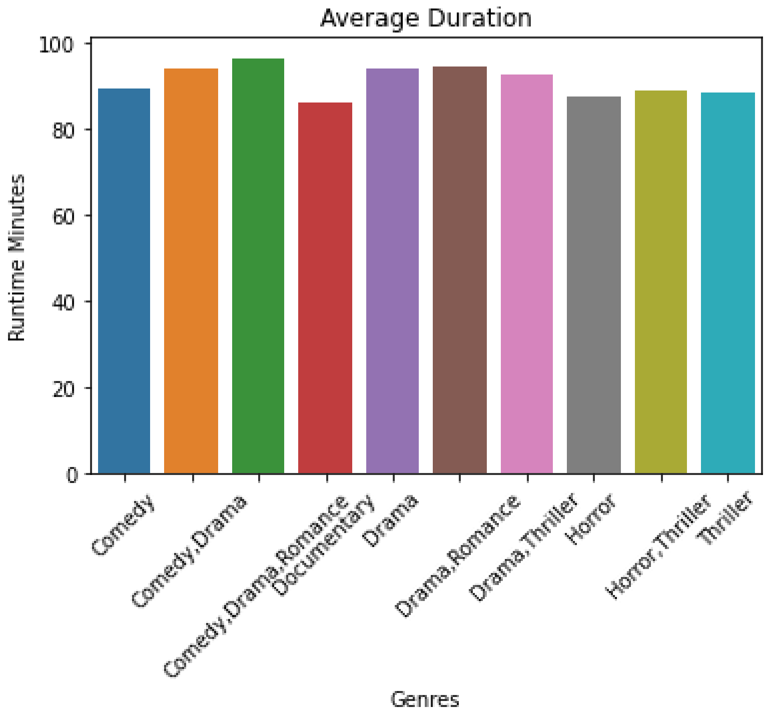
Bivariate Analysis

In [ ]:

```
#Which genres tend to have the highest revenue?
# Plotting bar plot
genre = df['genres'].head(20)
plt.figure(figsize=(12, 8))
sns.barplot(x=genre, y='worldwide_gross', data=df, estimator=max)  # Use ma
plt.title('Distribution of Revenue by Genre')
plt.xlabel('Genre')
plt.ylabel('Revenue (Worldwide Gross)')
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability
plt.show()
skk-osya-hwc
```

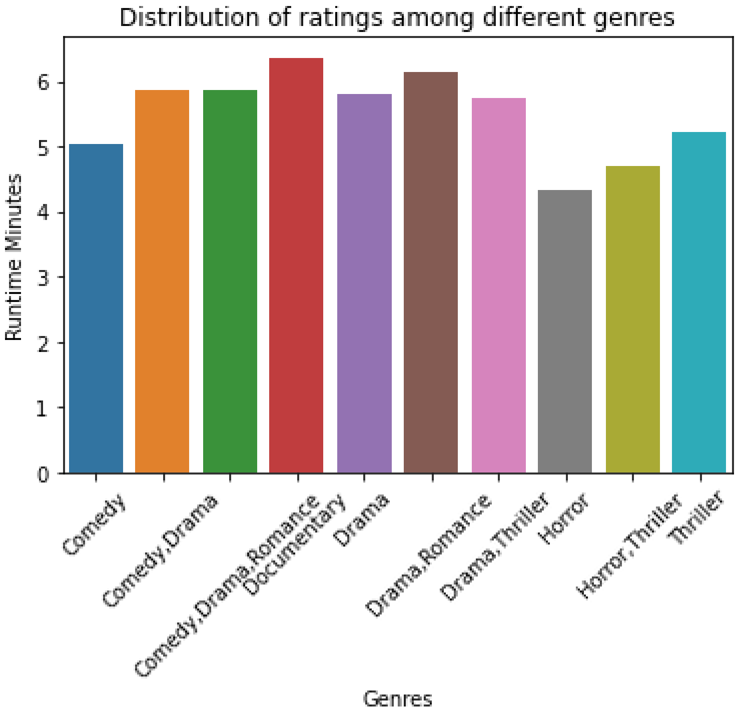
In [9]:

```
# Get the top 10 genres by frequency
top_genres = df['genres'].value_counts().head(10).index.tolist()

# Filter the dataframe to only include rows with one of the top 10 genres
top_genre_df = df[df['genres'].isin(top_genres)]

# Group the data by genres and calculate the mean runtime for each genre
genre_means = top_genre_df.groupby('genres')['runtime_minutes'].mean()

# Create a bar plot to visualize the mean runtime for each genre
sns.barplot(x=genre_means.index, y=genre_means.values)
plt.title('Average Duration')
plt.xlabel('Genres')
plt.ylabel('Runtime Minutes')
plt.xticks(rotation=45)
plt.show()
```



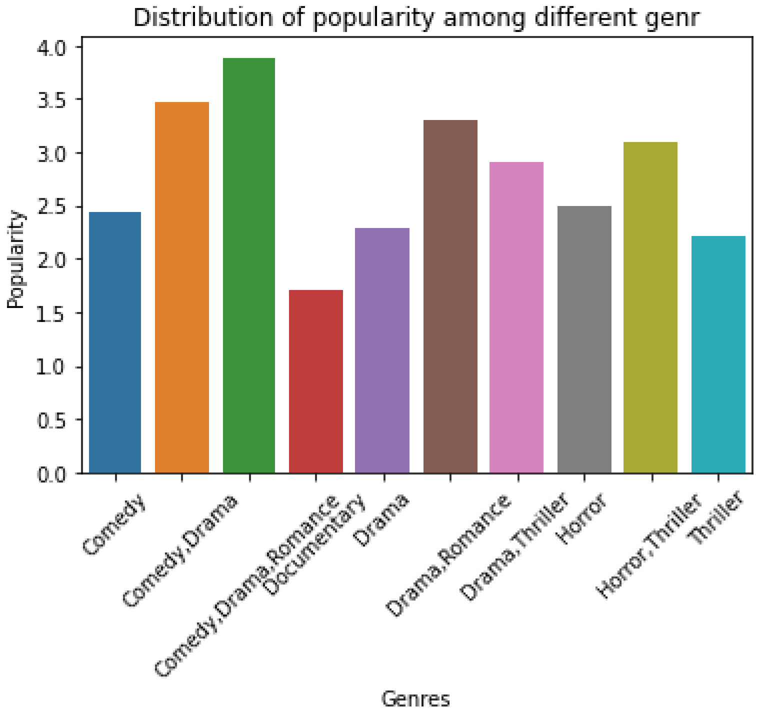
In [10]:

```
#Highest rated genre
# Get the top 10 genres by frequency
top_genres = df['genres'].value_counts().head(10).index.tolist()

# Filter the dataframe to only include rows with one of the top 10 genres
top_genre_df = df[df['genres'].isin(top_genres)]

# Group the data by genres and calculate the voteaverage for each genre
genre_means = top_genre_df.groupby('genres')['vote_average'].mean()

# Create a bar plot to visualize the mean runtime for each genre
sns.barplot(x=genre_means.index, y=genre_means)
plt.title('Distribution of ratings among different genres')
plt.xlabel('Genres')
plt.ylabel('Runtime Minutes')
plt.xticks(rotation=45)
plt.show()
```



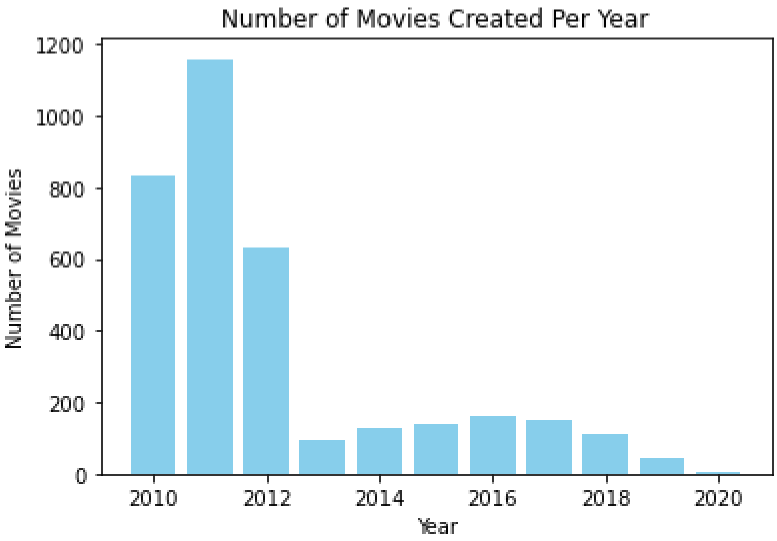
In [11]:

```
#Relationship between popularity and genre
#Highest rated genre
# Get the top 10 genres by frequency
top_genres = df['genres'].value_counts().head(10).index.tolist()

# Filter the dataframe to only include rows with one of the top 10 genres
top_genre_df = df[df['genres'].isin(top_genres)]

# Group the data by genres and calculate the voteaverage for each genre
genre_means = top_genre_df.groupby('genres')['popularity'].mean()

# Create a bar plot to visualize the mean runtime for each genre
sns.barplot(x=genre_means.index, y=genre_means.values)
plt.title('Distribution of popularity among different genr')
plt.xlabel('Genres')
plt.ylabel('Popularity')
plt.xticks(rotation=45)
plt.show()
```



In [12]:

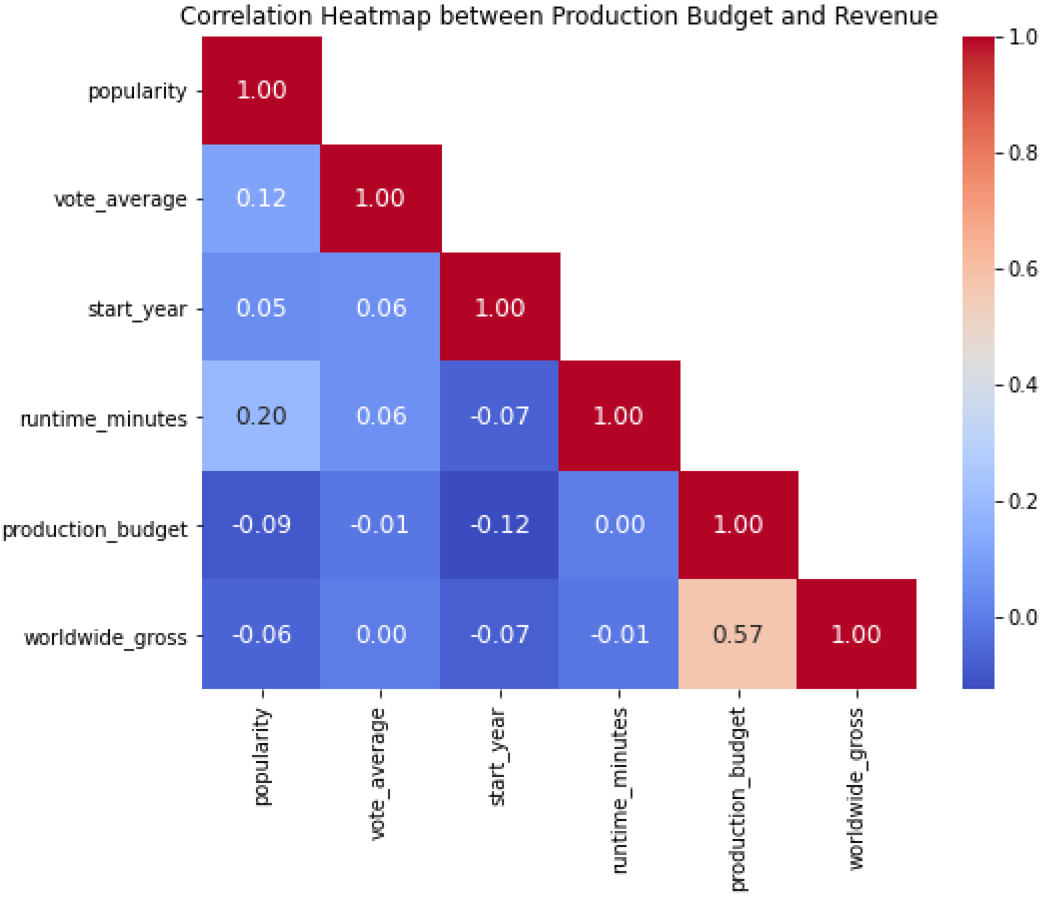
```
# Count the number of movies per year
movies_per_year = df['start_year'].value_counts()

# Create a bar plot
plt.bar(movies_per_year.index, movies_per_year.values, color='skyblue')
# Add labels and title
plt.xlabel('Year')
plt.ylabel('Number of Movies')
plt.title('Number of Movies Created Per Year')

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



Multivariate Analysis

In [13]:

```
# Calculate correlation matrix
correlation_matrix = df.corr()
# Create a mask to hide the upper triangle
mask = np.triu(np.ones_like(correlation_matrix), k=1)
# Plot heatmap
plt.figure(figsize=(8, 6))

sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", ann
plt.title('Correlation Heatmap between Production Budget and Revenue')
plt.show()
```



Conclusion

- It is advisable to create movies that last between 90 to 120 minutes. This duration seems to align well with audience preferences.
- Focus on producing more Comedy and Drama movies, as these genres appear to be the most popular among viewers. This could potentially attract a larger audience.
- There is a strong positive correlation between production budget and revenue. This suggests that increasing the production budget might lead to higher revenue. However, it's essential to carefully assess the cost-benefit relationship to ensure profitability.
- Consider prioritizing the production of Drama, Horror, and History genres, as they contribute to the highest revenue. This insight can guide content creation strategies for maximizing profitability.