Project 3 - Investigate a dataset

Dataset to be analysed

The dataset to be analysed in this project is the TMDb movie data.

Posed Question

Which genres are most popular from year to year? Has this changed over time?

Description of Investigation

The following is a summary of the steps undertaken in this investigation:

1) Read in file 'tmdb-movies.csv' as a Data Frame.
2) Filtered the data to include only variables of interest – 'popularity', 'release_year' & 'genres'.
3) Determined an arbitrary definition of popularity – Top 1% of all movies in list.
4) Filtered data to only include popular movies.
5) Separated genres defined separated by '|' into separate columns.
6) Stacked the columns of genres while preserving their release_year.
7) Removed genres with nulls.
8) Created a count variable to allocate 1 for each genre present.
9) Summed up the counts by grouping over years and then genres.
10) Pivoted data to get frequencies of each genre as a column.
11) Replaced all null values with a 0.
12) Created various visualisations to explore the data.

The variables of Interest

The three dependent variables of interest in this investigation were the release year, the popularity, and the genre of the movie. The independent variable of interest was the count corresponding to a release year and genre, given that the movie passed a popularity test.

The Statistics and Visualisations that were used

The 99^{th} percentile of popularity was used as a definition of popularity because I wanted to be precise on my definition of 'popular'. If we had less data, we could have used the 50^{th} percentile or average as a cut-off for this value.

A stacked vertical bar chart that had counts for the various combinations of years was used to inspect the composition of genres overs time.

A stacked line chart was used to capture the trend over time in genre composition.

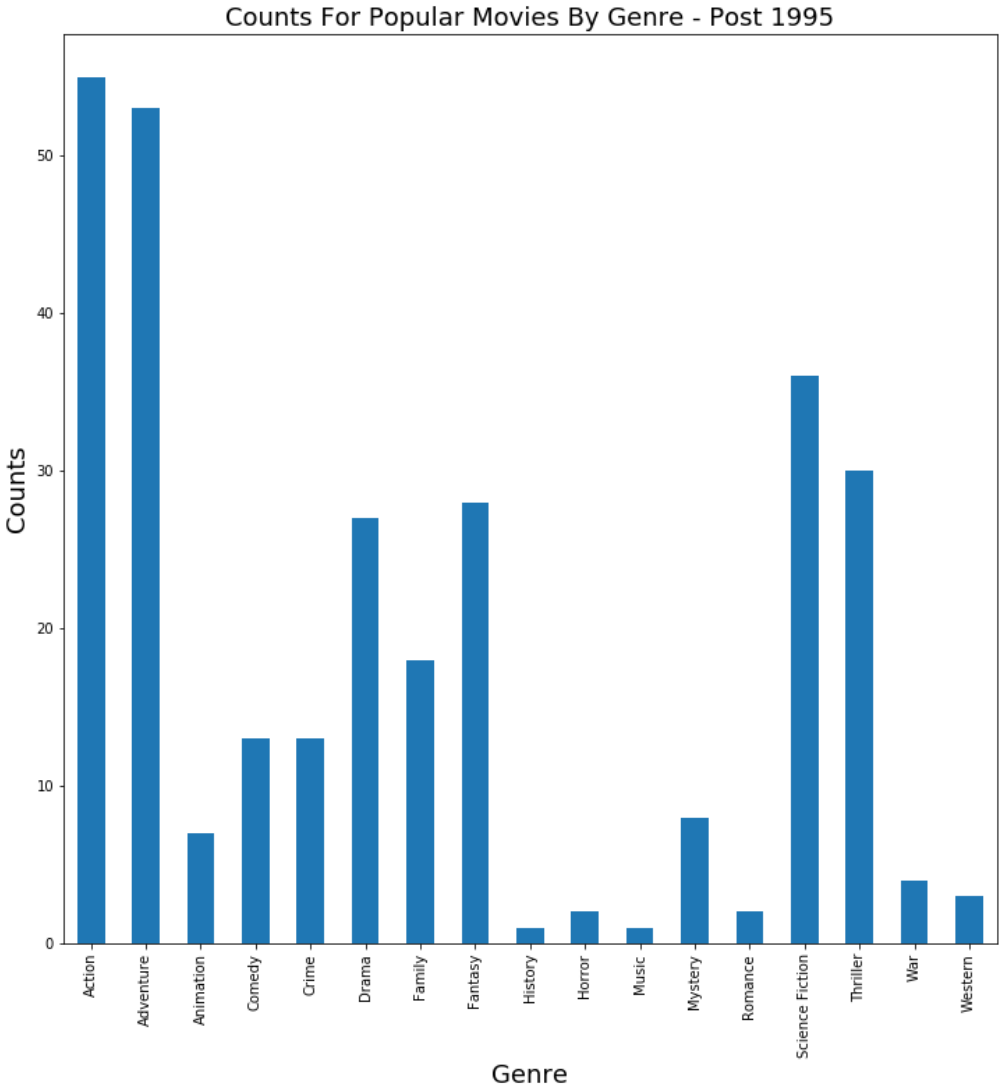
Bar charts containing counts for each genre before 1995 and after 1995 were using to capture changes in genre composition. The year 1995 was chosen as a midpoint between 1975 and 2015.

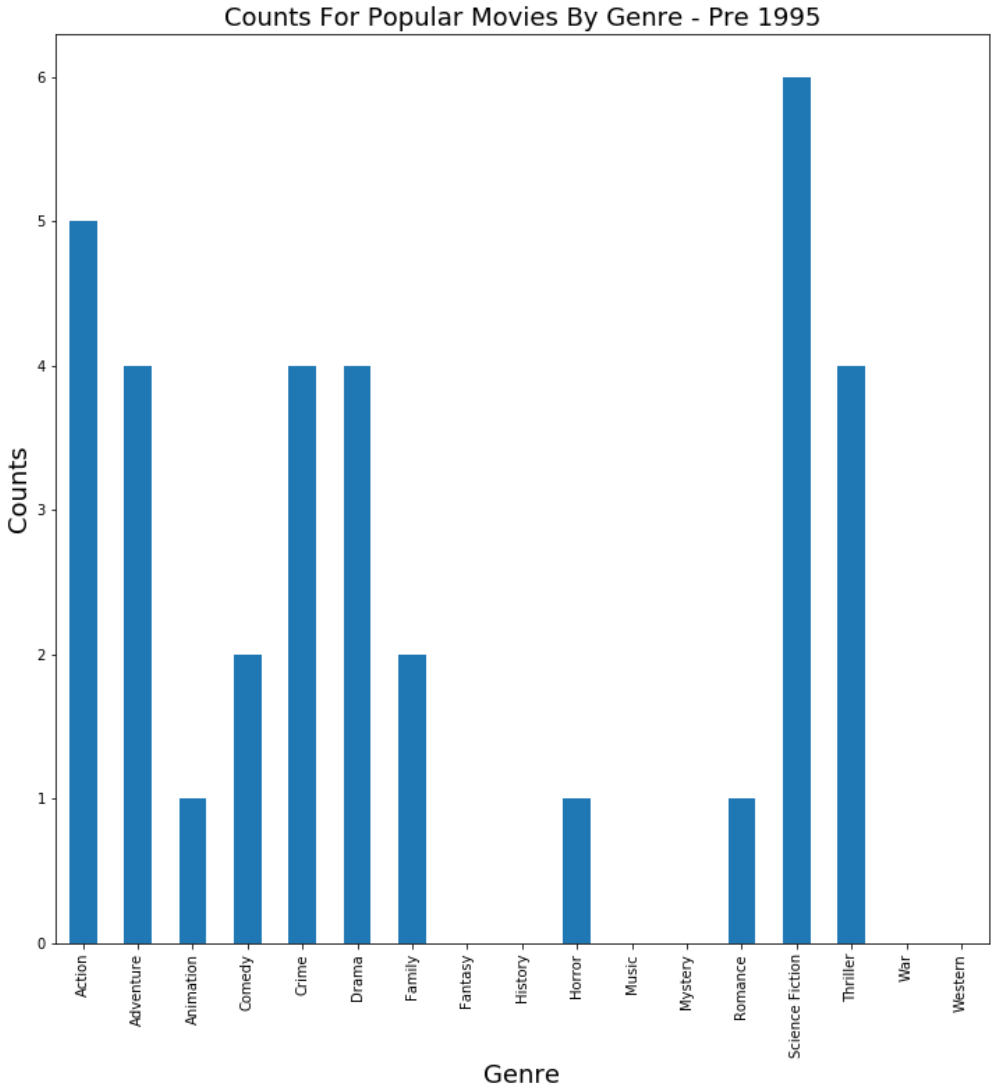
Conclusions

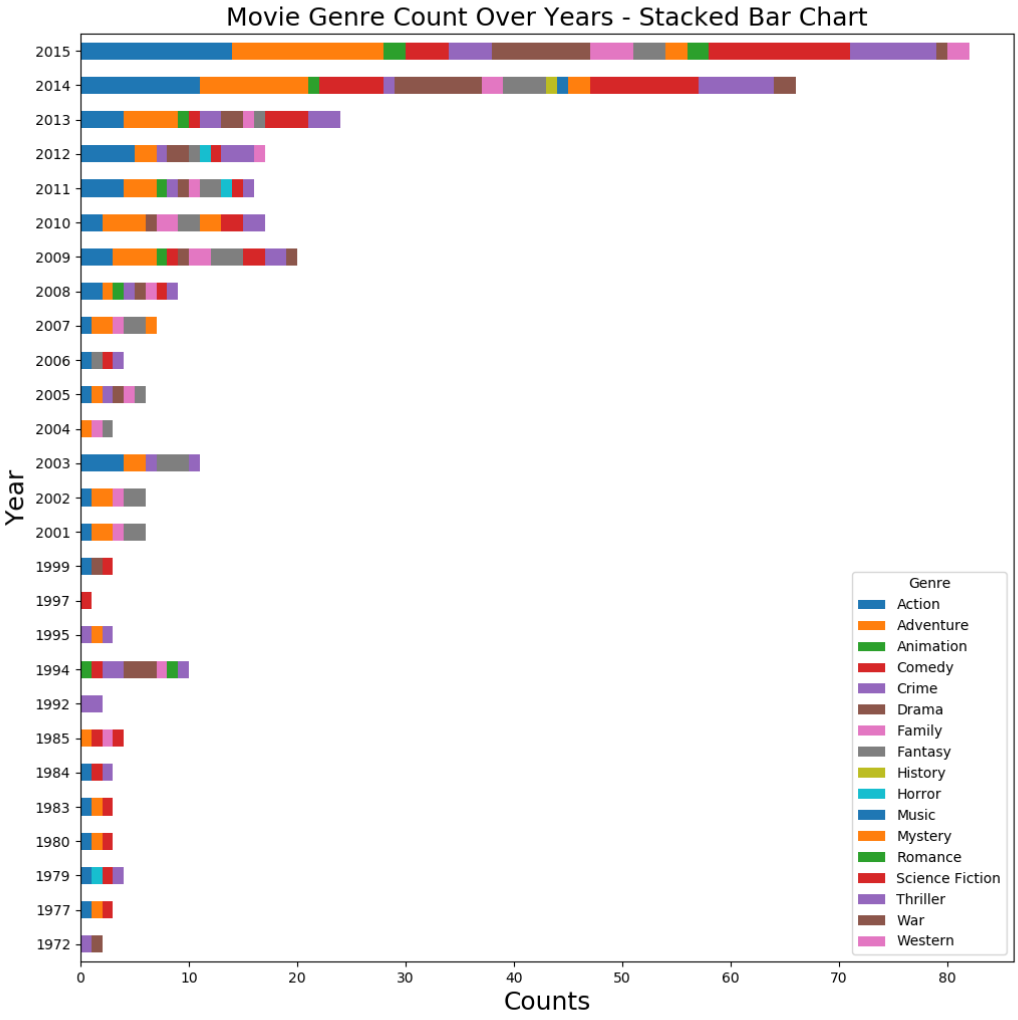
The results of this exploration can be summarised in the visualisations attached. As shown, it seems that more genres have become increasingly popular over time. Fantasy, History, Music, Mystery, War, and Western genres did not make the top 1% of movies in popularity prior to 95. Action and Adventure have gained popularity while Thrillers and Science fiction have lost popularity. It should be noted that newer movies are considered more popular than older movies. A Bias is likely to exist in terms of the measurement of popularity. Newer movies also tend to have more multiple genres than older movies. These factors may make the results of this analysis less meaningful.

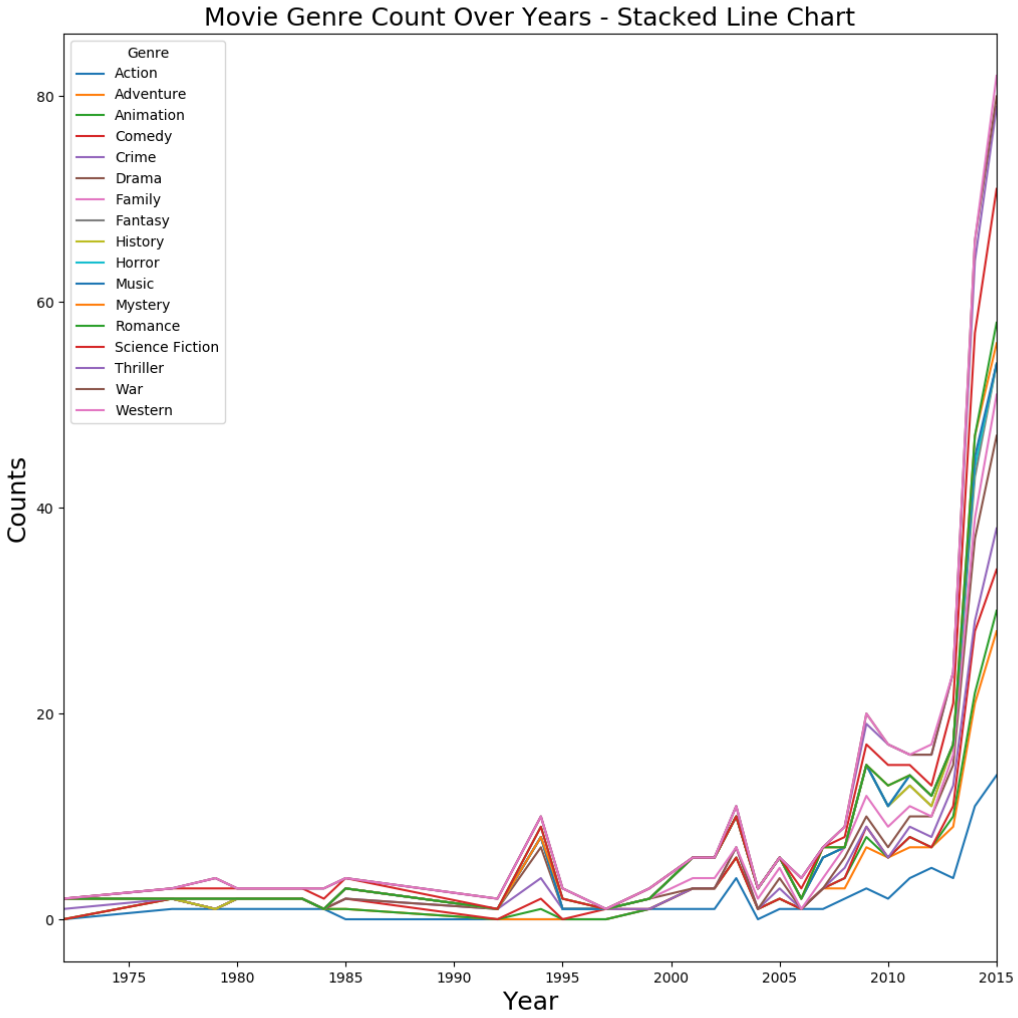
Code and Visualisations

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv('tmdb-movies.csv')

df = df[['popularity','release_year','genres']]
popular = np.percentile(df['popularity'],99)
df = df.query('popularity > {}'.format(popular))
df['genre_1'], df['genre_2'],df['genre_3'],df['genre_4'],df['genre_5'], = df['genres'].str.split('|').str

year = df['release_year'].append(df['release_year']).append(df['release_year']).append(df['release_year']).append(df['release_year'])
genre = df['genre_1'].append(df['genre_2']).append(df['genre_3']).append(df['genre_4']).append(df['genre_5'])

df = pd.DataFrame()
df['Year'] = year
df['Genre'] = genre
df.dropna(inplace = True)
df['count']= 1

df = df.groupby(['Year','Genre'],as_index=False).sum()
df = df.pivot(index='Year',columns='Genre')
df.fillna(0, inplace = True)
df.columns = df.columns.droplevel(0)

df.query('Year >= 1995').sum().plot(kind='bar',figsize=(12,12))
plt.xlabel('Genre',fontsize=18)
plt.ylabel('Counts',fontsize=18)
plt.title('Counts For Popular Movies By Genre - Post 1995',fontsize=18)
plt.savefig('Counts For Popular Movies By Genre - Post 1995.png');

df.query('Year < 1995').sum().plot(kind='bar',figsize=(12,12))
plt.xlabel('Genre',fontsize=18)
plt.ylabel('Counts',fontsize=18)
plt.title('Counts For Popular Movies By Genre - Pre 1995',fontsize=18);
plt.savefig('Counts For Popular Movies By Genre - Pre 1995.png');

df.plot(kind='barh',stacked=True,figsize=(12,12));
plt.xlabel('Counts',fontsize=18)
plt.ylabel('Year',fontsize=18)
plt.title('Movie Genre Count Over Years - Stacked Bar Chart',fontsize=18)
plt.savefig('Movie Genre Count Over Years - Stacked Bar Chart.png');

df.plot(kind='line',stacked=True,figsize=(12,12));
plt.xlabel('Year',fontsize=18)
plt.ylabel('Counts',fontsize=18)
plt.title('Movie Genre Count Over Years - Stacked Line Chart',fontsize=18)
plt.savefig('Movie Genre Count Over Years - Stacked Line Chart.png');
```


Counts For Popular Movies By Genre - Post 1995
Counts
0
10
20
30
40
50
Action
Adventure
Animation
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Mystery
Romance
Science Fiction
Thriller
War
Western
Genre


Counts For Popular Movies By Genre - Pre 1995
Counts
0
1
2
3
4
5
6
Action
Adventure
Animation
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Mystery
Romance
Science Fiction
Thriller
War
Western
Genre


Movie Genre Count Over Years - Stacked Bar Chart
Year
Counts
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
1999
1997
1995
1994
1992
1985
1984
1983
1980
1979
1977
1972
0
10
20
30
40
50
60
70
80
Genre
Action
Adventure
Animation
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Mystery
Romance
Science Fiction
Thriller
War
Western


Movie Genre Count Over Years - Stacked Line Chart
Genre
Action
Adventure
Animation
Comedy
Crime
Drama
Family
Fantasy
History
Horror
Music
Mystery
Romance
Science Fiction
Thriller
War
Western
Counts
0
20
40
60
80
1975
1980
1985
1990
1995
2000
2005
2010
2015
Year