CISC7201 INTRODUCTION TO DATA SCIENCE PROGRAMMING

Project Report

Student Name: Chen Zhilong, Weng Zhenkun, Chao Hou Kit
Student ID: MB95522, MB95548, MB95542
Email: mb95522@connect.um.edu.mo
mb95548@connect.um.edu.mo
mb95542@connect.um.edu.mo
Date: 11th December, 2019

Faculty of Science and Technology
UNIVERSITY OF MACAU

1. Introduction

This project is to design some data analytical process on a dataset (>100 MB or above).

2. Dataset

The dataset used in this project is from Kaggle (The Movies Dataset) and is the following files:
movies_metadata.csv (32.8MB): The main Movies Metadata file. It contains information on 45,000 movies include budget, revenue, release dates, popularity, id, countries and etc.
credits.csv (181MB): Consists of Cast and Crew information for movies.
The overall columns as below:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45466 entries, 0 to 45465
Data columns (total 24 columns):
adult                     45466 non-null object
belongs_to_collection     4494 non-null object
budget                    45466 non-null object
genres                    45466 non-null object
homepage                  7782 non-null object
id                        45466 non-null object
imdb_id                   45449 non-null object
original_language         45455 non-null object
original_title            45466 non-null object
overview                  44512 non-null object
popularity                45461 non-null object
poster_path               45080 non-null object
production_companies      45463 non-null object
production_countries      45463 non-null object
release_date              45379 non-null object
revenue                   45460 non-null float64
runtime                   45203 non-null float64
spoken_languages          45460 non-null object
status                    45379 non-null object
tagline                   20412 non-null object
title                     45460 non-null object
video                     45460 non-null object
vote_average              45460 non-null float64
vote_count                45460 non-null float64
dtypes: float64(4), object(20)
 <class 'pandas.core.frame.DataFrame'>
 RangeIndex: 45476 entries, 0 to 45475
 Data columns (total 3 columns):
 cast    45476 non-null object
 crew    45476 non-null object
 id      45476 non-null int64
 dtypes: int64(1), object(2)
```

Fig.2.1 Columns Information about Datasets

3. Library

3.1 AST

Abstract Syntax Tree (AST) is a very strong features in Python. Python AST module allows us to interact with Python code itself and modify it.
In this project, some fields in dataset is JSON type, however, it is used apostrophe in JSON format, therefore, when we usd json.load to read the related information in there fields, it will appear error. In order to solve this problem, 'ast.literal_eval' is be used, which is safely evaluate an expression node or a Unicode or Latin-1 encoded string containing a Python expression. The string or node provided may only consist of the following Python literal structures: strings, numbers, tuples, lists, dicts, booleans, and None.
This can be used for safely evaluating strings containing Python expressions from untrusted sources without the need to parse the values oneself.
Related code is as following:

```
movies[i]=movies[i].apply(ast.literal_eval).apply(json.dumps)
movies[i]=movies[i].apply(json.loads)
```

3.2 SEABORN

Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures.
Seaborn aims to make visualization a central part of exploring and understanding data. Its dataset-oriented plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.
In this project, seaborn is used to get the Regression Line in scatterplot by using function **replot()**. SEABORN also provides other useful function, such as **implot()**, **catplot()** and etc.

4. Data Cleaning

4.1 Data Type Conversion

The column, 'crew' in the dataset credits is json type and the columns, 'genres' and 'production_countries' in dataset movies_metadata are also json type and they need to convert to string type. The step of data type conversion is as following.

i. In the column 'production_countries', some fields are not json type and need to make these fields be '[]'.

```
for i in range(len(movies)):
  if str(movies.loc[i,'production_countries'])[0]!='[':
      movies.loc[i,'production_countries']='[]'
```

ii. json.dumps() : convert a python object to a json string
json.loads() : convert a json string to a python object
Because in columns 'crew', 'genres' and 'production_countries' are using apostrophe in json format, which is not suitable to json type. Function 'ast.literal_eval' is used to change these columns to right json type.

```
l=['genres','production_countries']
for i in l:
    movies[i]=movies[i].apply(ast.literal_eval).apply(json.dumps)
    movies[i]=movies[i].apply(json.loads)
credits['crew']=credits['crew'].apply(ast.literal_eval).apply(json.dumps)
credits['crew']=credits['crew'].apply(json.loads)
```

iii. Extract related fields in these columns.

iv. Change the type of 'id' and etc. from string to float.

```
movies['id']=movies['id'].apply(pd.to_numeric,errors = 'coerce')
df['budget'] = df['budget'].apply(pd.to_numeric,errors = 'coerce')
df['popularity'] = df['popularity'].apply(pd.to_numeric,errors = 'coerce')
```

v. Merge two datasets, movies and credits.

```
df = pd.merge(movies,credits,how='left',on='id')
```

4.2 Drop Useless Column and Rename Column

```
credits.rename(columns={'crew':'director'},inplace = True)
del credits['cast']
df.drop(['homepage','original_title','adult','belongs_to_collection','imdb_id','poster_pat
h','production_companies','tagline','spoken_languages','overview',
'status','video'],axis=1,inplace=True)
```

4.3 Nan Values Processing

```
df['runtime']=df['runtime'].fillna(df.runtime.mean())
df=df.dropna(axis=0,how='any',subset=['director','release_date','original_language'])
df.isnull().sum().sort_values(ascending=False)
```

```
director                0
vote_count              0
vote_average            0
title                   0
runtime                 0
revenue                 0
release_date            0
production_countries    0
popularity              0
original_language       0
id                      0
genres                  0
budget                  0
dtype: int64
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 44605 entries, 0 to 45541
Data columns (total 13 columns):
budget                  44605 non-null float64
genres                  44605 non-null object
id                      44605 non-null float64
original_language       44605 non-null object
popularity              44605 non-null float64
production_countries    44605 non-null object
release_date            44605 non-null object
revenue                 44605 non-null float64
runtime                 44605 non-null float64
title                   44605 non-null object
vote_average            44605 non-null float64
vote_count              44605 non-null float64
director                44605 non-null object
dtypes: float64(7), object(6)
```

Fig.4.3.1 Result of Data Processing

4.4 Fields Explanation

1) title: movie title
2) director: director
3) budget: budget (USD)
4) genres: style list, movie type
5) id: identification number
6) popularity: relative page views on Movie Database
7) production_countries: production countries
8) release_date: release time
9) revenue: revenue (USD)
10) runtime: movie duration
11) vote_average: average rating
12) vote_count: number of ratings
13) original_language: movie language

5. Data Analysis

5.1 Purpose

1) The amount of movie over year.
2) The amount movie in different genres.
3) The amount movie in different genres over year.
4) The amount of movie from different countries.
5) The most profitable and popularity movie genres.
6) The relationship between revenue and budget and vote average(rating).
7) The director who shoots the most movies.
8) The director who has the highest rating.
9) The director who get the highest box office.

5.2 The Amount of Movie over Year

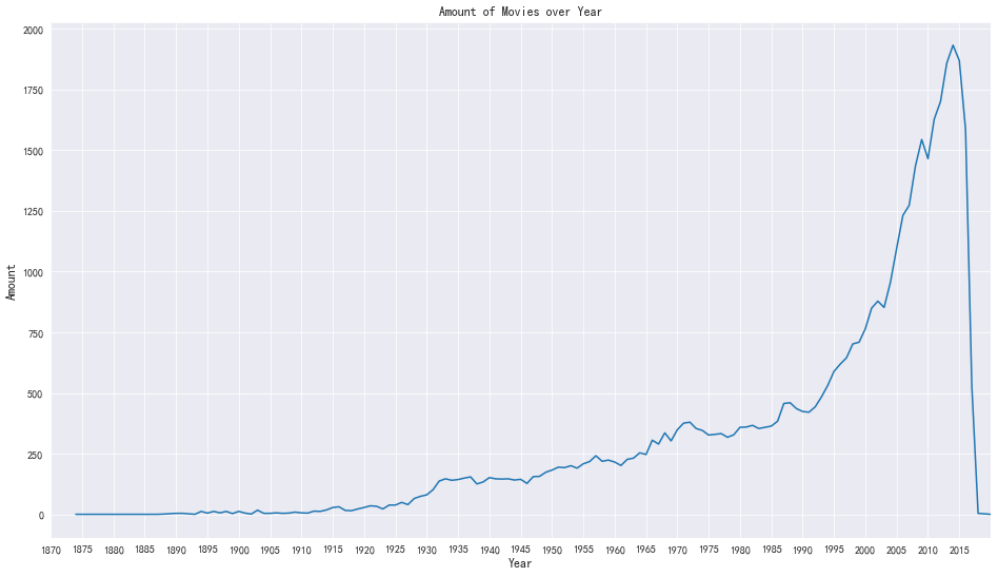

Fig.5.2.1

According to Fig.5.2.1, the amount of movie has a significant increasing from 1990.

5.3 **The Amount of Movie in different Genres**

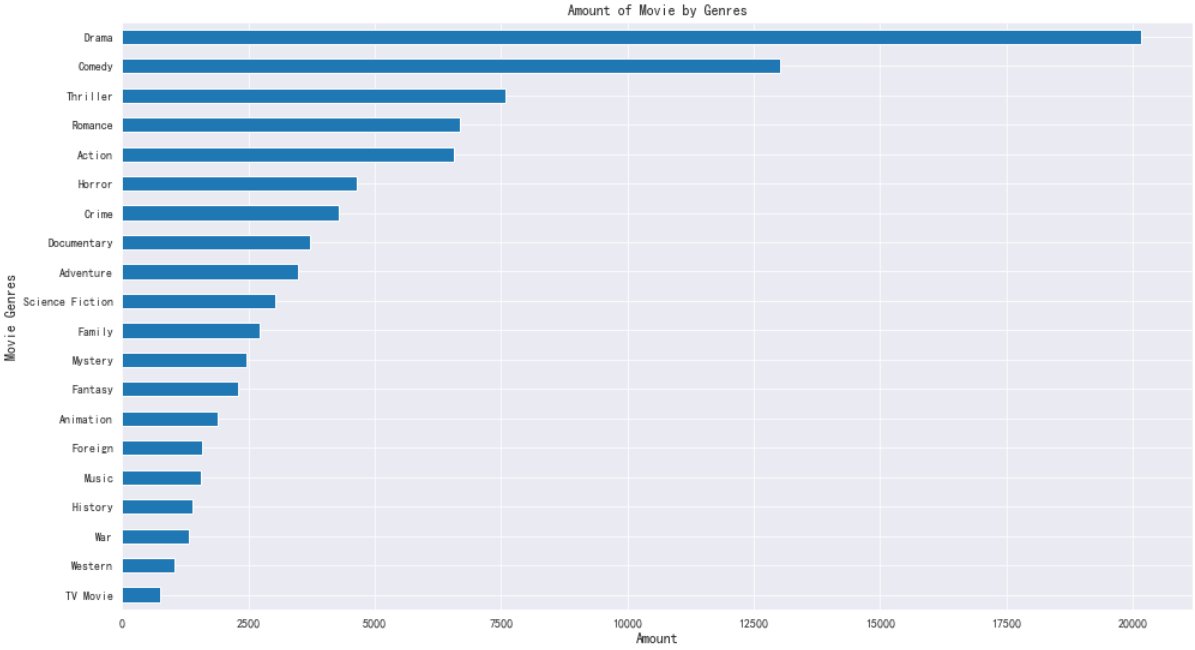

Fig.5.3.1

The total amount of Drama and Comedy is the most.

5.4 The Amount of Movie in different Genres over Year

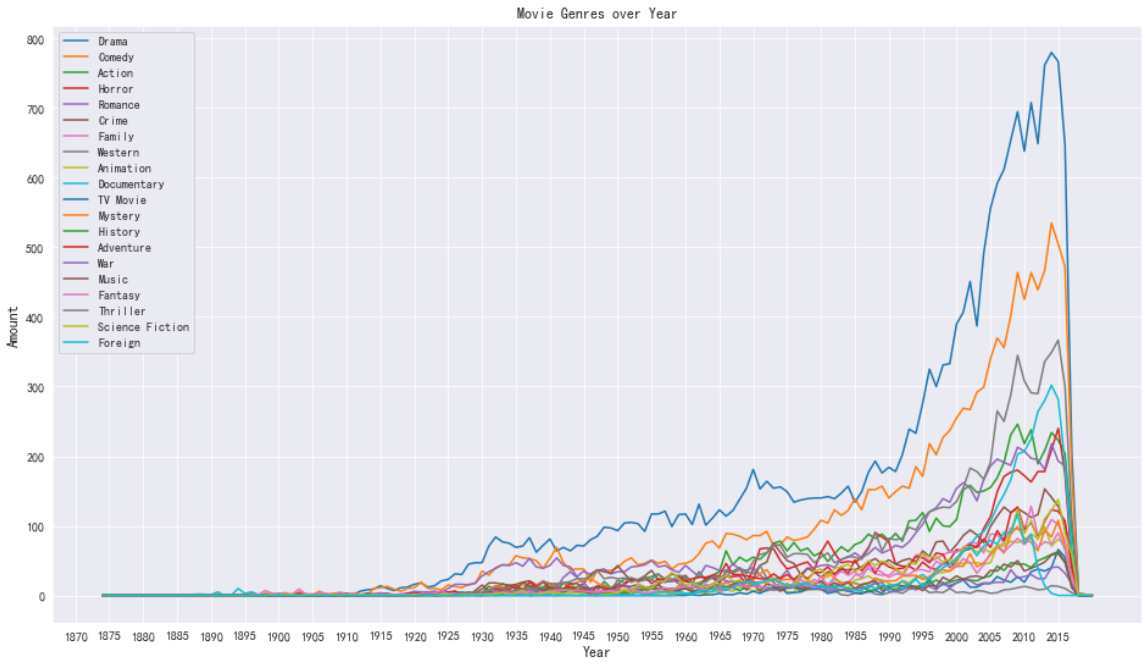

Fig.5.4.1

According to Fig.5.3.1 and Fig.5.4.1, it shows the change of the amount of different movie genres over year, and Drama and Comedy has a significant variation over year which has a higher growth rate.

5.5 The Proportion of Movie Amount in different Countries

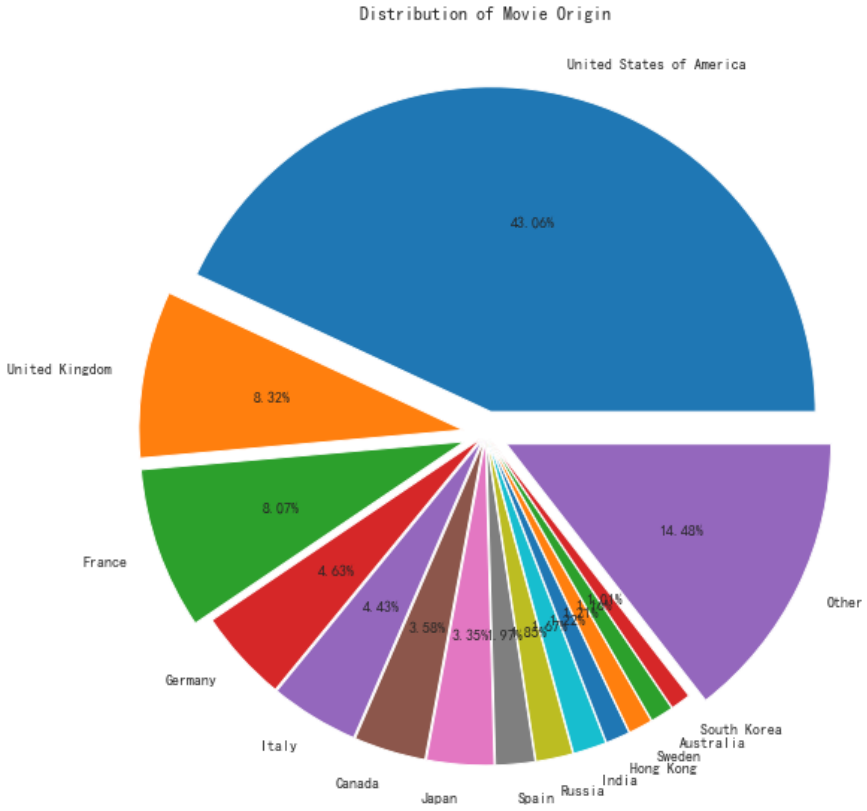

Fig.5.5.1

According to Fig.5.4.1, United States of America has the most movie production and it is 43.06% of the total.

5.6 The most Profitable and Popularity Movie Genres

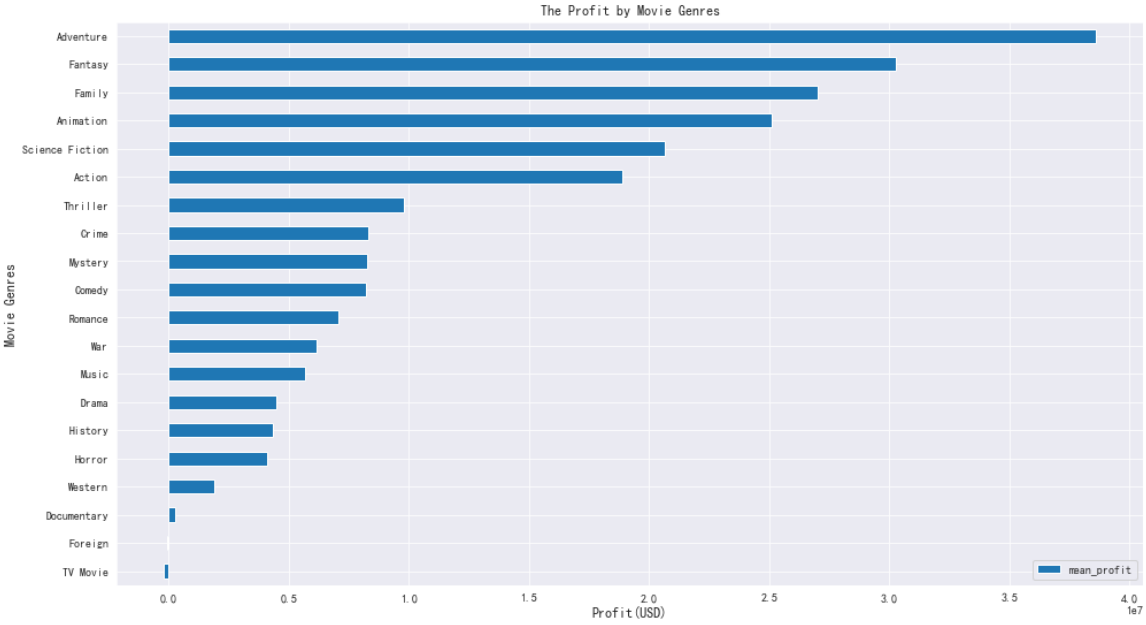

Fig.5.6.1

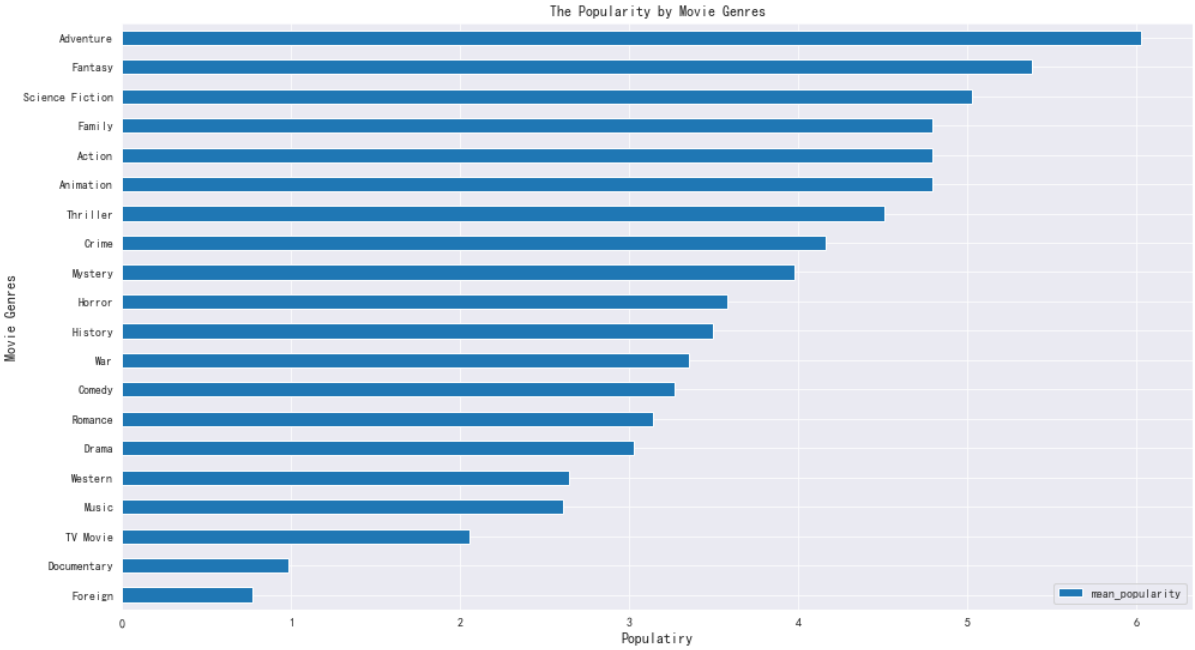

Fig.5.6.2

According to Fig.5.6.1, it shows that Adventure, Fantasy and Family are the most profitable, Foreign is not profitable and TV Movie is loss.
According to Fig.5.6.2, it shows that the most popularity movie genres are Adventure, Fantasy, Science Fiction and Family.

5.7 The Relationship between Revenue and Budget and Vote Average

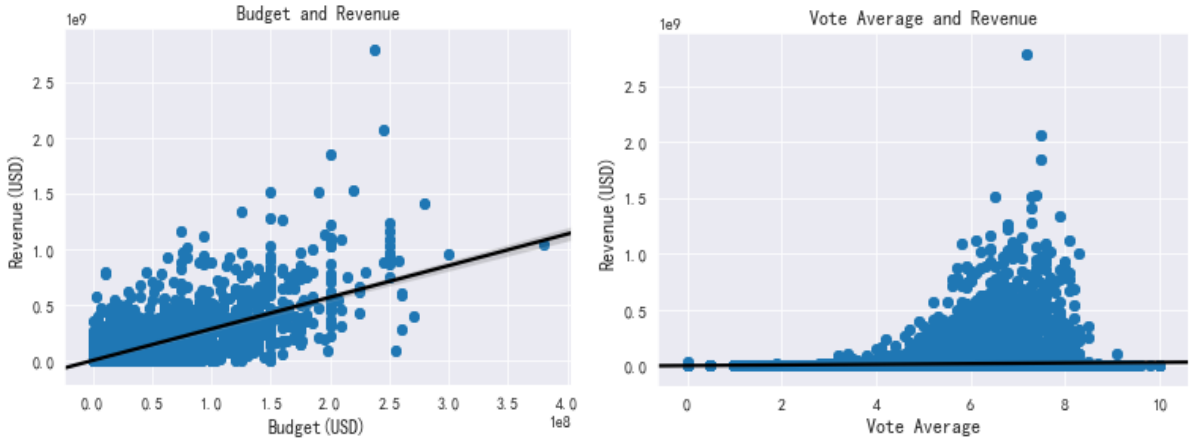

Fig.5.7.1

According to Fig.5.7.1, it shows that the relationship between Budget and Revenue is positive correlation, and the Vote Average (Rating) and Revenue is no correlation.

5.8 The Director who Shoots the Most Movies

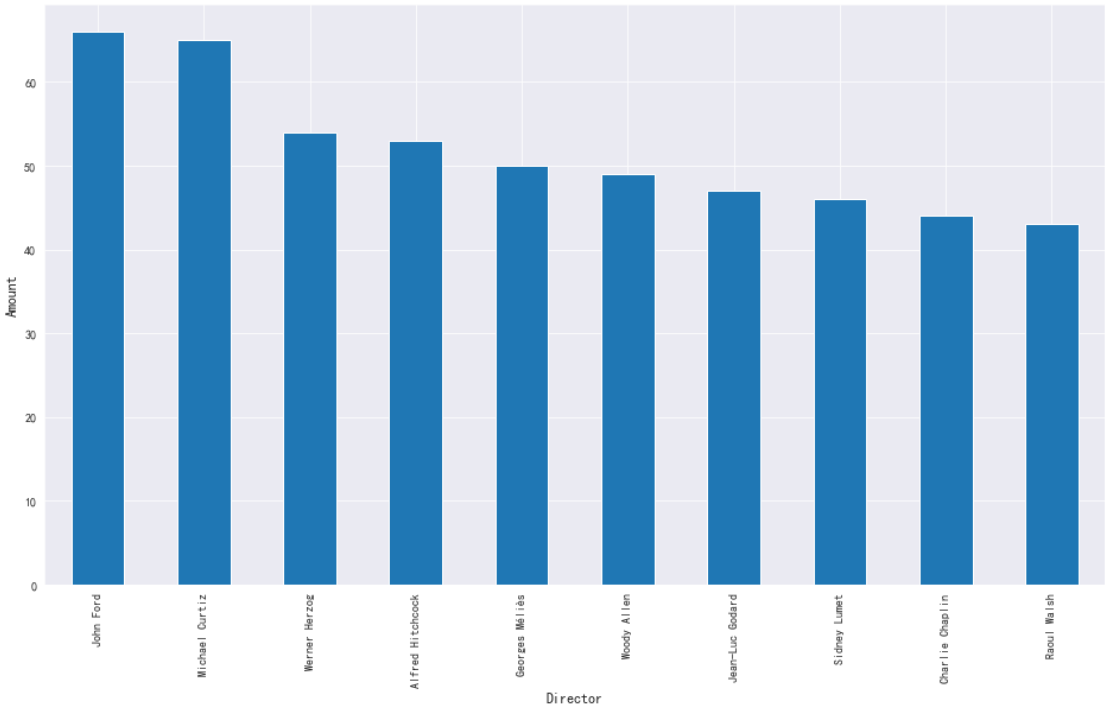

Fig.5.8.1

According to Fig.5.8.1, it is the Top 10 rank of director who shoot the most movies and Top 3 are John Ford, Michael Curtiz and Werner Herzog.

5.9 The Director who has the Highest Rating

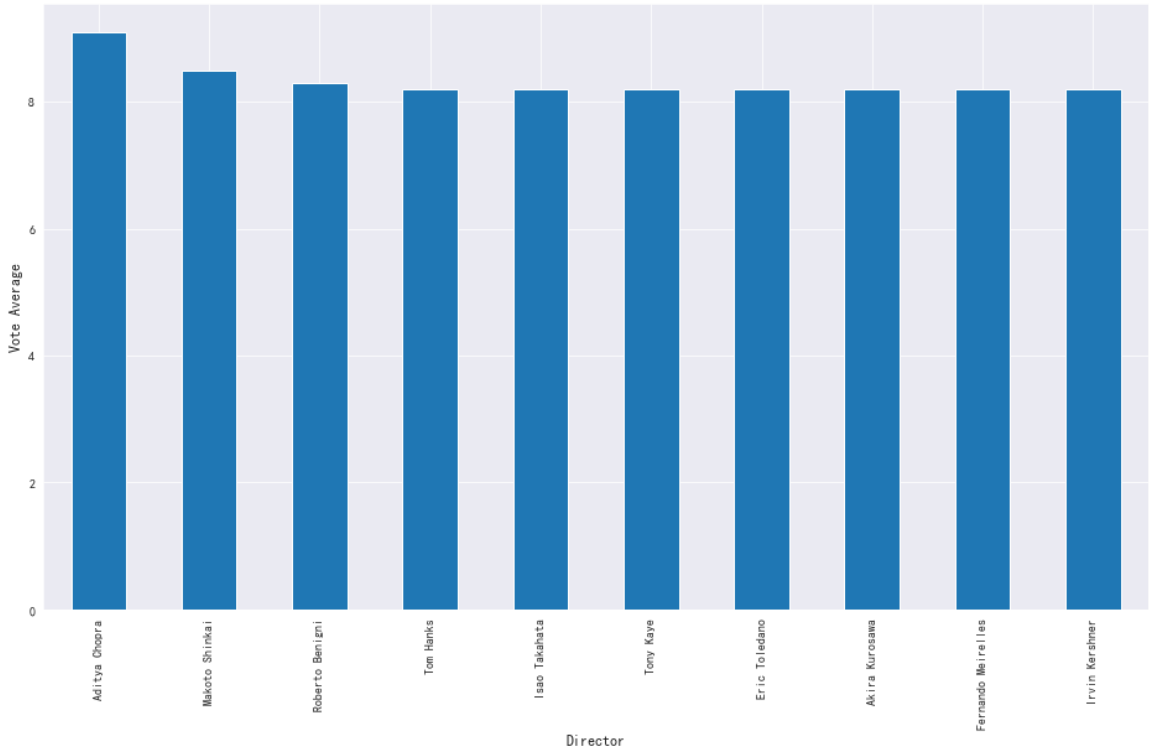

Fig.5.9.1

According to Fig.5.9.1, it is the Top 10 rank of director who has the highest rating and Top 3 are Aditya Chopra, Makoto Shinkai and Roberto Benigni.

5.10 The Director who Get the Highest Revenue

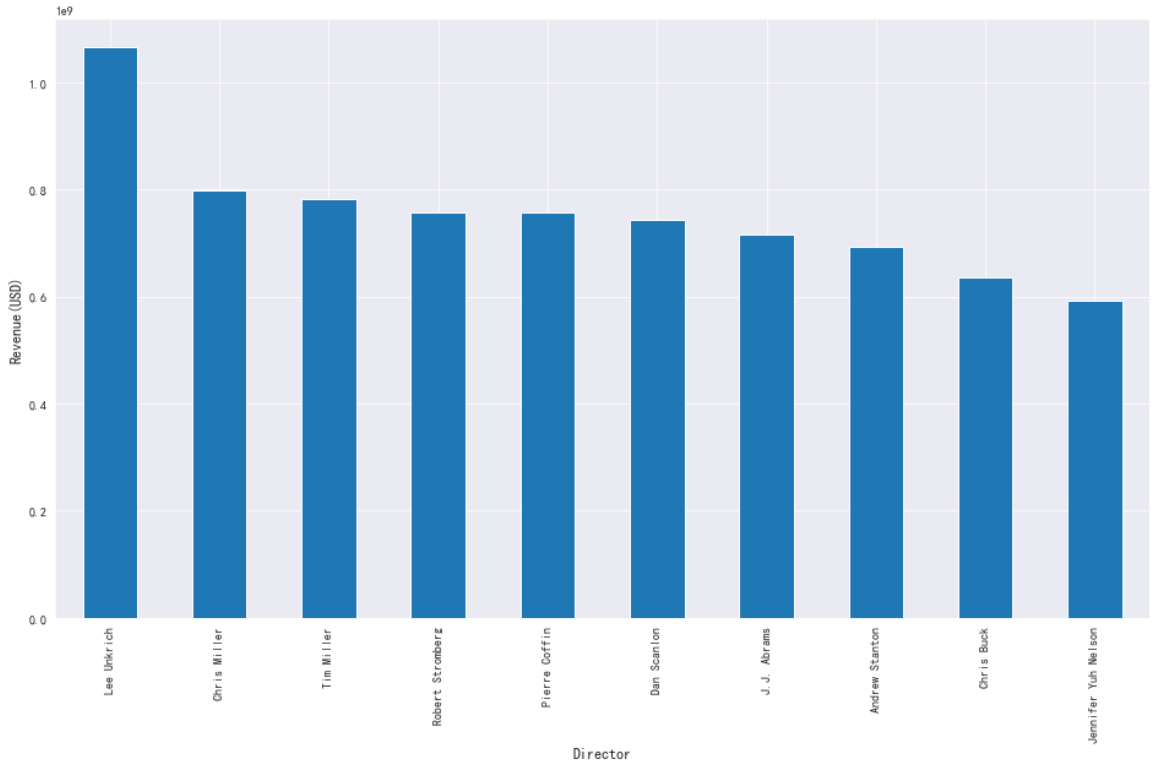

Fig.5.10.1

According to Fig.5.10.1, it is the Top 10 rank of director who get the highest revenue and Top 3 are Lee Unkrich, Chris Miller and Tim Miller,

6 Conclusion

1. The movies industry has grown rapidly since 1990, and Drama, Comedy have a significant increasing and have the most amount. America is the most has the most movie production.
2. Adventure, Fantasy and Family are Top 3 of the most profitable movie genres, and the most popular movie genres are Adventure, Fantasy, Science Fiction and Family. Therefore, Adventure, Fantasy and Family are recommended for company to shoot.
3. The relationship between revenue and budget are positive correlation, therefore, movie company should increase the budget which is beneficial for improving the movie quality.
4. For the director, John Ford shoots the most movies, Aditya Chopra has the highest rating and Lee Unkrich get the highest revenue.

Reference

[1] https://kite.com/python/docs/ast.literal_eval
[2] https://seaborn.pydata.org/
[3] https://xbuba.com/questions/53042478
[4] https://www.kaggle.com/rounakbanik/the-movies-dataset
[5] https://blog.csdn.net/xc_zhou/article/details/81458388