In [ ]:

```
import pandas as pd
import numpy as np
```

In [ ]:

```
#import datasets
ratings_df = pd.read_csv("data/movie_lense/ratings.csv")
movies_coefficients_df = pd.read_csv("data/movie_coefficents.csv", ind
ex_col=[0])
movies_w_genres = pd.read_csv("data/movies_w_genre_after96.csv", index
_col=14)
movies_variables_after96_df = pd.read_csv("data/movie_variables_after_
1996.csv", index_col=0)
```

In [ ]:

```
movies_combined_df = movies_coefficients_df
grouped_ratings_df = ratings_df.groupby([ratings_df['movieId']]).ratin
g.count()
movies_combined_df['total_ratings'] = grouped_ratings_df[movies_combin
ed_df.index]

cols = movies_combined_df.columns.tolist()
cols.insert(0, cols.pop())
movies_combined_df = movies_combined_df[cols]
movies_combined_df = movies_combined_df[movies_combined_df.index.isin(
movies_w_genres.index)]
movies_numfeatures_df = movies_combined_df
movies_numfeatures_df
```

Adding Genre, Coefficents, and Genome Tags to movies.csv

In [ ]:

```
movies_combined_df.columns
cols = movies_w_genres.columns.tolist()
for c in cols:
    movies_combined_df[c] = movies_w_genres[c]

movies_combined_df
```

In [ ]:

```
cols = np.array(movies_combined_df.columns.tolist())
toMove = cols[np.where(cols=='1')[0][0]:np.where(cols=='x**6')[0][0]+1]
toMove
cols = np.delete(arr=cols, obj=range(np.where(cols=='1')[0][0],np.where(cols=='x**6')[0][0]+1))
cols = list(cols)
cols.extend(toMove)
movies_combined_df = movies_combined_df[cols]
movies_combined_df = movies_combined_df.drop(labels=['title_lower', 'new_title'], axis=1)
movies_combined_df['release'] = movies_combined_df['release'].astype(int)
movies_combined_df['num_months'] = movies_combined_df['num_months'].astype(int)
movies_combined_df
```

In [ ]:

```
movies_variables_after96_df

movies_combined_df = movies_combined_df[movies_combined_df.index.isin(movies_variables_after96_df.index)]
movies_combined_df.columns
cols = movies_variables_after96_df.columns.tolist()
for c in cols:
    movies_combined_df[c] = movies_variables_after96_df[c]

movies_combined_df
```

In [ ]:

```
genome_scores_pca_df = pd.read_csv("data/genome-scores-pca.csv", index_col=[0])
cols = genome_scores_pca_df.columns.tolist()
cols = [c + '_pca' for c in cols]
genome_scores_pca_df.columns = cols
for c in cols:
    movies_combined_df[c] = genome_scores_pca_df[c]
```

Adding Genre, Coefficents, and Genome Tags for Dataframe with movies.csv parsed with movie_industry.csv

In [ ]:

```
industry_movies_combined_df = movies_numfeatures_df

def movie_variables(filename, to_add_df):
    movies_variables_df = pd.read_csv(filename, index_col=[0])
    to_add_df = to_add_df[to_add_df.index.isin(movies_variables_df.ind
ex)]
    cols = movies_variables_df.columns.tolist()
    for c in cols:
        #print(c)
        to_add_df[c] = movies_variables_df[c]

    return to_add_df
    #industry_movies_combined_df
```

In [ ]:

```
industry_movies_w_genres = pd.read_csv("data/movies_w_genre_after96.cs
v", index_col=14)

industry_movies_combined_df = industry_movies_combined_df[industry_mov
ies_combined_df.index.isin(industry_movies_w_genres.index)]
industry_movies_combined_df.columns
cols = industry_movies_w_genres.columns.tolist()
for c in cols:
    industry_movies_combined_df[c] = industry_movies_w_genres[c]

industry_movies_combined_df
```

In [ ]:

```
genome_scores_pca_df = pd.read_csv("data/genome-scores-pca.csv", index
_col=[0])
cols = genome_scores_pca_df.columns.tolist()
cols = [c + '_pca' for c in cols]
genome_scores_pca_df.columns = cols
for c in cols:
    industry_movies_combined_df[c] = genome_scores_pca_df[c]

industry_movies_combined_df.fillna(0, inplace=True)
industry_movies_combined_df.tail(100)
```

In [ ]:

```
industry_movies_combined_5ratings_df = industry_movies_combined_df[industry_movies_combined_df['total_ratings']/industry_movies_combined_df['num_months'] >= 5]
industry_movies_combined_df.to_csv('data/industry_time_data/industry_movies_combined.csv')
industry_movies_combined_5ratings_df.to_csv('data/industry_time_data/industry_movies_5ratingspermonth.csv')

#4 months
industry_movies_combined_5ratings_4months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 4]
industry_movies_combined_5ratings_4months_df = movie_variables("data/movie_variables/movie_variables_4_months.csv", industry_movies_combined_5ratings_4months_df)
industry_movies_combined_5ratings_4months_df.to_csv('data/industry_time_data/industry_movies_5ratings_4months.csv')

print(len(industry_movies_combined_5ratings_4months_df))

#12 months
industry_movies_combined_5ratings_12months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 12]
industry_movies_combined_5ratings_12months_df = movie_variables("data/movie_variables/movie_variables_12_months.csv", industry_movies_combined_5ratings_12months_df)
industry_movies_combined_5ratings_12months_df.to_csv('data/industry_time_data/industry_movies_5ratings_12months.csv')

print(len(industry_movies_combined_5ratings_12months_df))


#24 months
industry_movies_combined_5ratings_24months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 24]
industry_movies_combined_5ratings_24months_df = movie_variables("data/movie_variables/movie_variables_24_months.csv", industry_movies_combined_5ratings_24months_df)
industry_movies_combined_5ratings_24months_df.to_csv('data/industry_time_data/industry_movies_5ratings_24months.csv')

print(len(industry_movies_combined_5ratings_24months_df))

#60 months
industry_movies_combined_5ratings_60months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 60]
industry_movies_combined_5ratings_60months_df = movie_variables("data/movie_variables/movie_variables_60_months.csv", industry_movies_combined_5ratings_60months_df)
industry_movies_combined_5ratings_60months_df.to_csv('data/industry_time_data/industry_movies_5ratings_60months.csv')

print(len(industry_movies_combined_5ratings_60months_df))
```

```
#90 months
industry_movies_combined_5ratings_90months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 90]
industry_movies_combined_5ratings_90months_df = movie_variables("data/movie_variables/movie_variables_90_months.csv", industry_movies_combined_5ratings_90months_df)
industry_movies_combined_5ratings_90months_df.to_csv('data/industry_time_data/industry_movies_5ratings_90months.csv')

print(len(industry_movies_combined_5ratings_90months_df))

#120 months
industry_movies_combined_5ratings_120months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 120]
industry_movies_combined_5ratings_120months_df = movie_variables("data/movie_variables/movie_variables_120_months.csv", industry_movies_combined_5ratings_120months_df)
industry_movies_combined_5ratings_120months_df.to_csv('data/industry_time_data/industry_movies_5ratings_120months.csv')

print(len(industry_movies_combined_5ratings_120months_df))

#150 months
industry_movies_combined_5ratings_150months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 150]
industry_movies_combined_5ratings_150months_df = movie_variables("data/movie_variables/movie_variables_150_months.csv", industry_movies_combined_5ratings_150months_df)
industry_movies_combined_5ratings_150months_df.to_csv('data/industry_time_data/industry_movies_5ratings_150months.csv')

print(len(industry_movies_combined_5ratings_150months_df))

#180 months
industry_movies_combined_5ratings_180months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 180]
industry_movies_combined_5ratings_180months_df = movie_variables("data/movie_variables/movie_variables_180_months.csv", industry_movies_combined_5ratings_180months_df)
industry_movies_combined_5ratings_180months_df.to_csv('data/industry_time_data/industry_movies_5ratings_180months.csv')

print(len(industry_movies_combined_5ratings_180months_df))

#210 months
industry_movies_combined_5ratings_210months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 210]
industry_movies_combined_5ratings_210months_df = movie_variables("data/movie_variables/movie_variables_210_months.csv", industry_movies_combined_5ratings_210months_df)
industry_movies_combined_5ratings_210months_df.to_csv('data/industry_time_data/industry_movies_5ratings_210months.csv')
```

```
print(len(industry_movies_combined_5ratings_210months_df))

#240 months
industry_movies_combined_5ratings_240months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 240]
industry_movies_combined_5ratings_240months_df = movie_variables("data/movie_variables/movie_variables_240_months.csv", industry_movies_combined_5ratings_240months_df)
industry_movies_combined_5ratings_240months_df.to_csv('data/industry_time_data/industry_movies_5ratings_240months.csv')

print(len(industry_movies_combined_5ratings_240months_df))

#270 months
industry_movies_combined_5ratings_270months_df = industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 270]
industry_movies_combined_5ratings_270months_df = movie_variables("data/movie_variables/movie_variables_270_months.csv", industry_movies_combined_5ratings_270months_df)
industry_movies_combined_5ratings_270months_df.to_csv('data/industry_time_data/industry_movies_5ratings_270months.csv')

print(len(industry_movies_combined_5ratings_270months_df))


# industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 4].to_csv('data/industry_time_data/industry_movies_5ratings_4month.csv')
# industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 12].to_csv('data/industry_time_data/industry_movies_5ratings_12month.csv')
# industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 24].to_csv('data/industry_time_data/industry_movies_5ratings_24month.csv')
# industry_movies_combined_5ratings_df[industry_movies_combined_5ratings_df['num_months'] >= 60].to_csv('data/industry_time_data/industry_movies_5ratings_60month.csv')
```

In [ ]:

```
movies_combined_df[(movies_combined_df['num_months'] >= 12) & (movies_combined_df['total_ratings'] >= 60)]
```

In [ ]:

```
movies_combined_df.to_csv("data/master_dataset.csv", index=True) #Eshaan: 30mb Rajen: 26.9mb Seung-Hyun: 27mb
```

In [ ]: