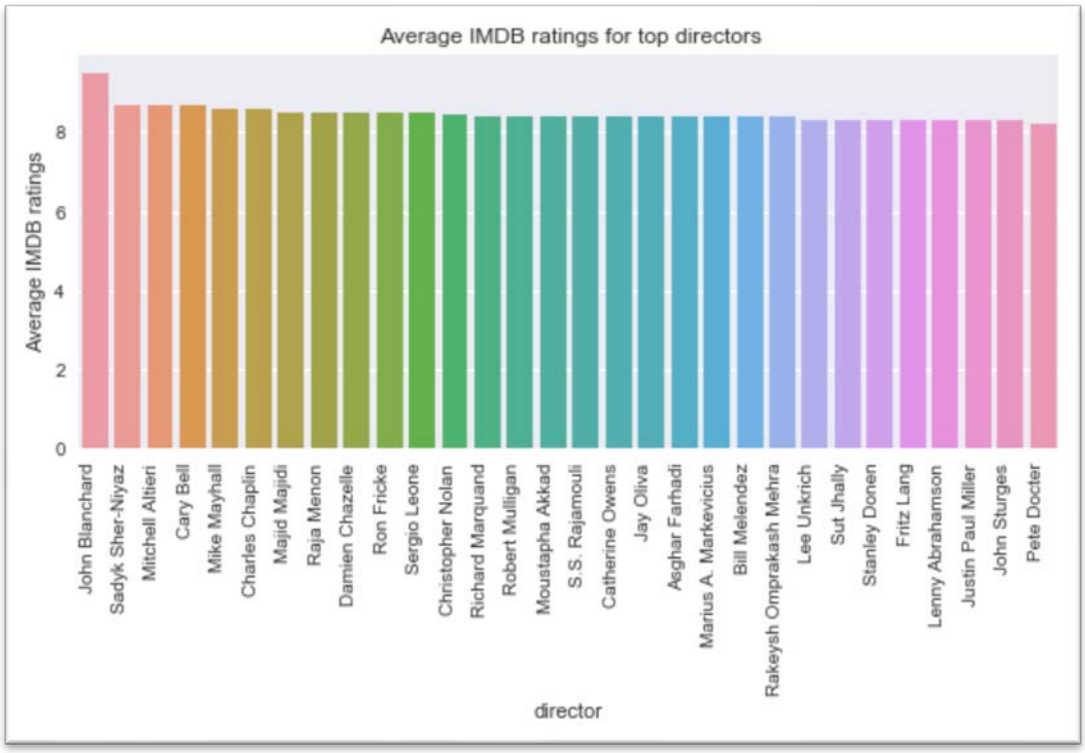
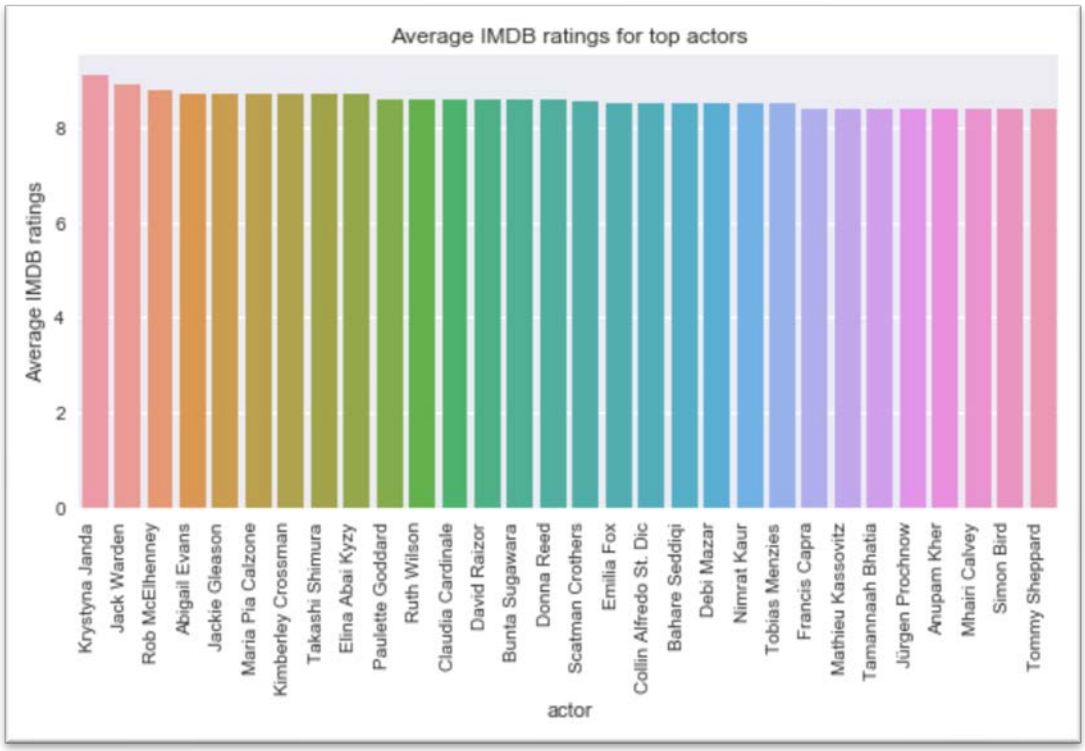
Prediction of IMDB movie ratings

By Harsh Ileshbhai Darji

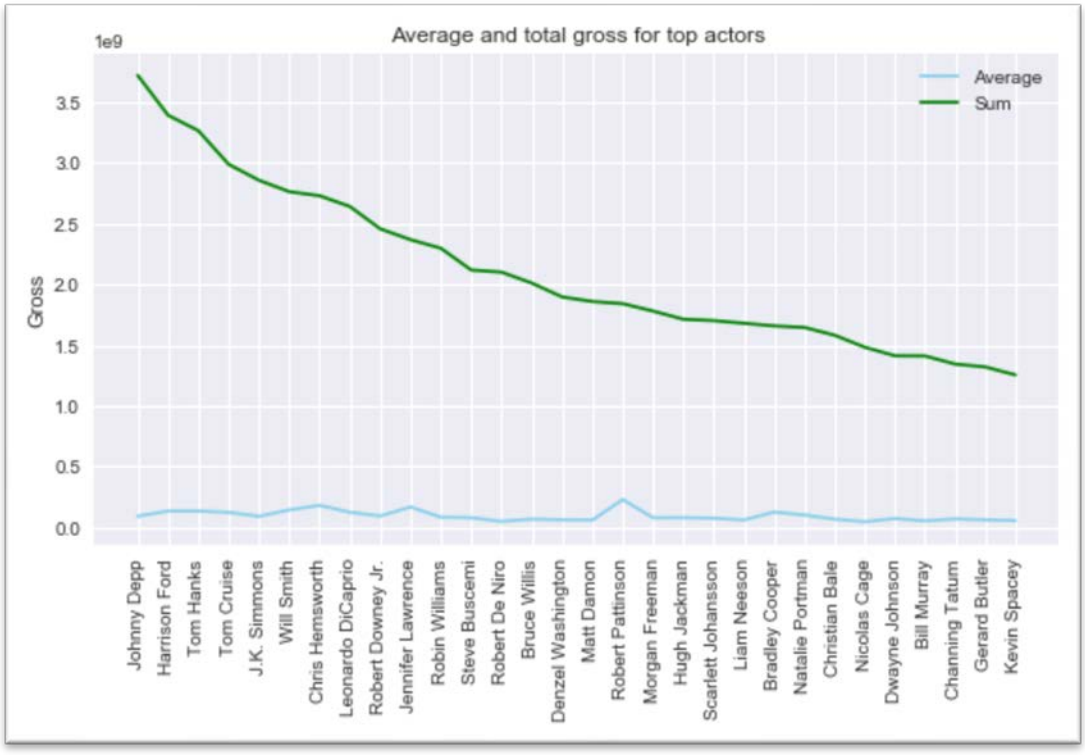
Exploratory Data Analysis:

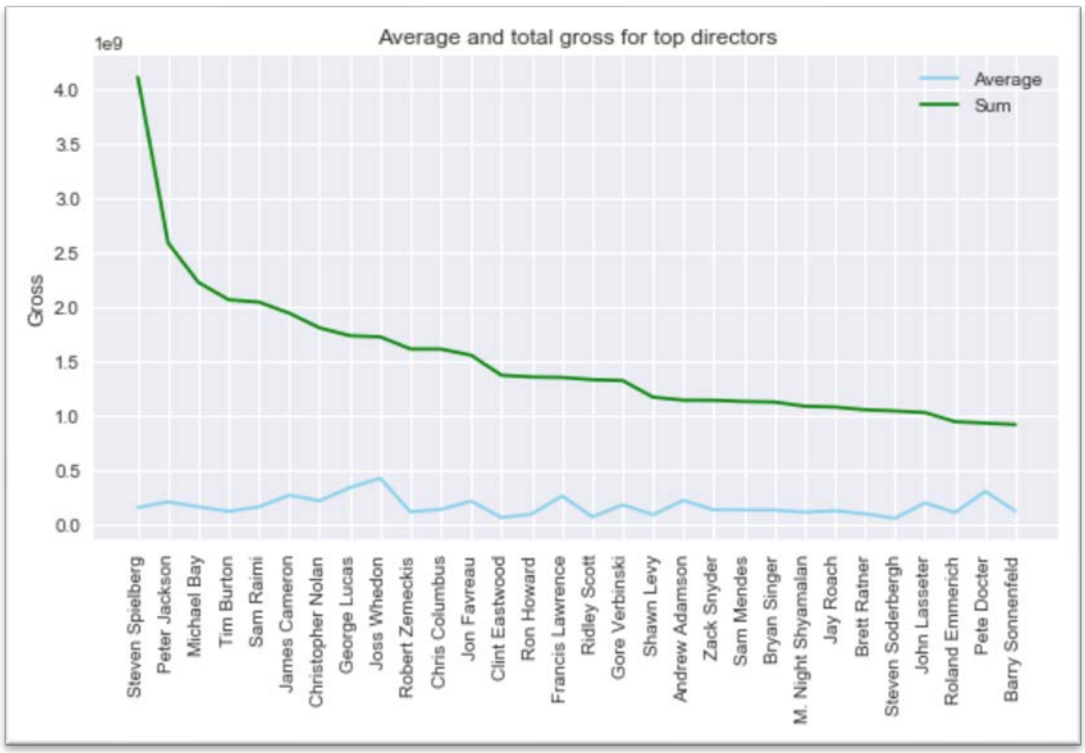
1.) Average IMDB ratings for top 30 directors and actors.



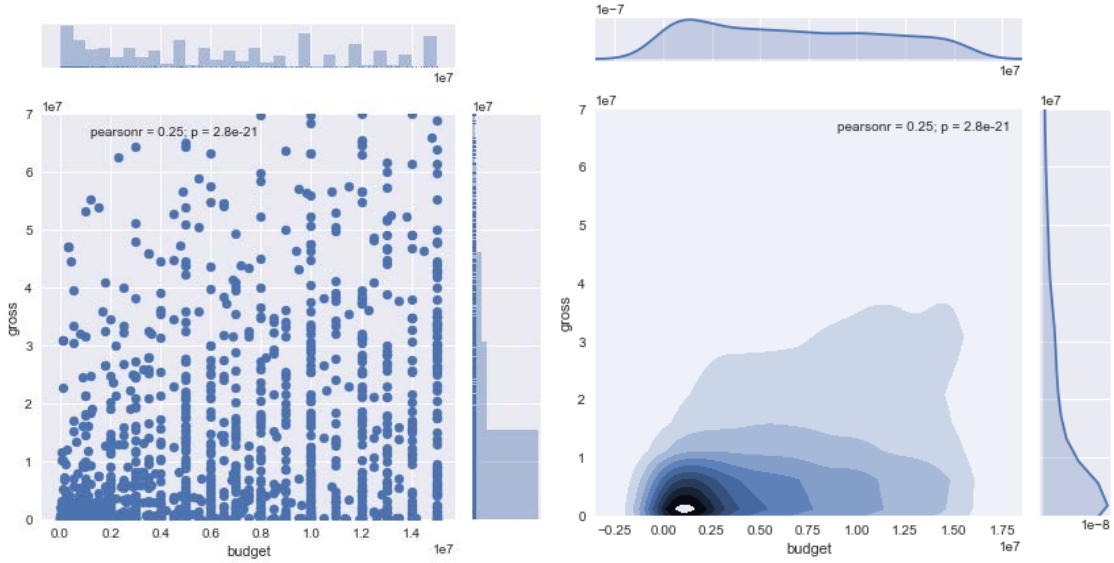


2.) Total and Average gross for top 30 actors and directors.

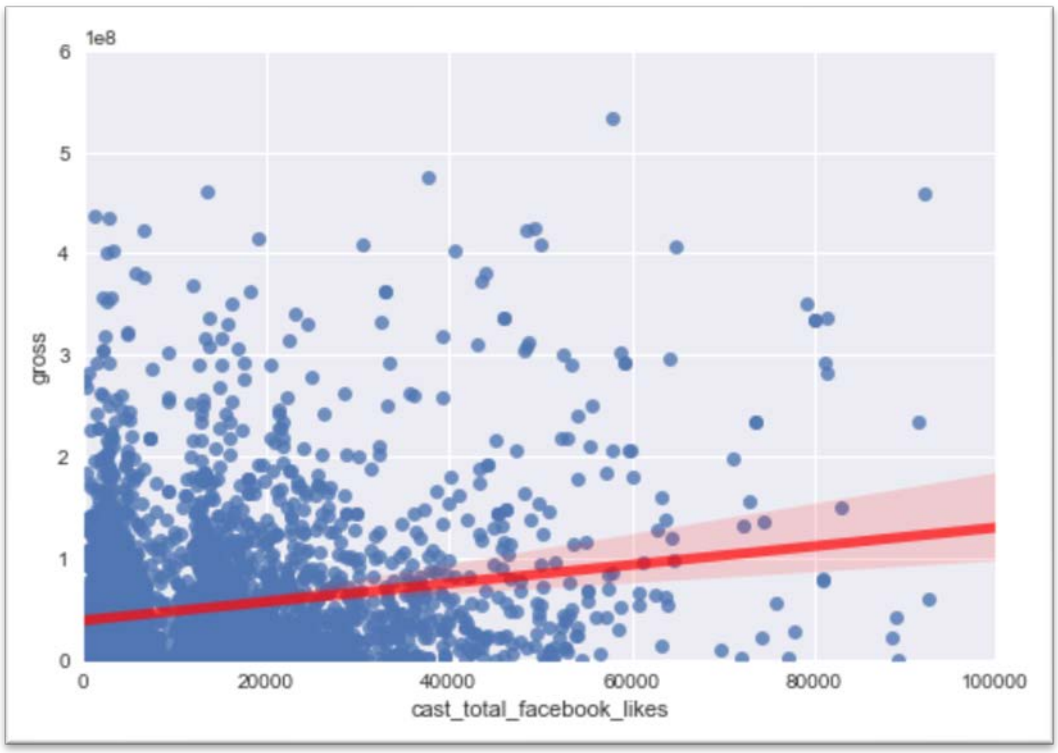




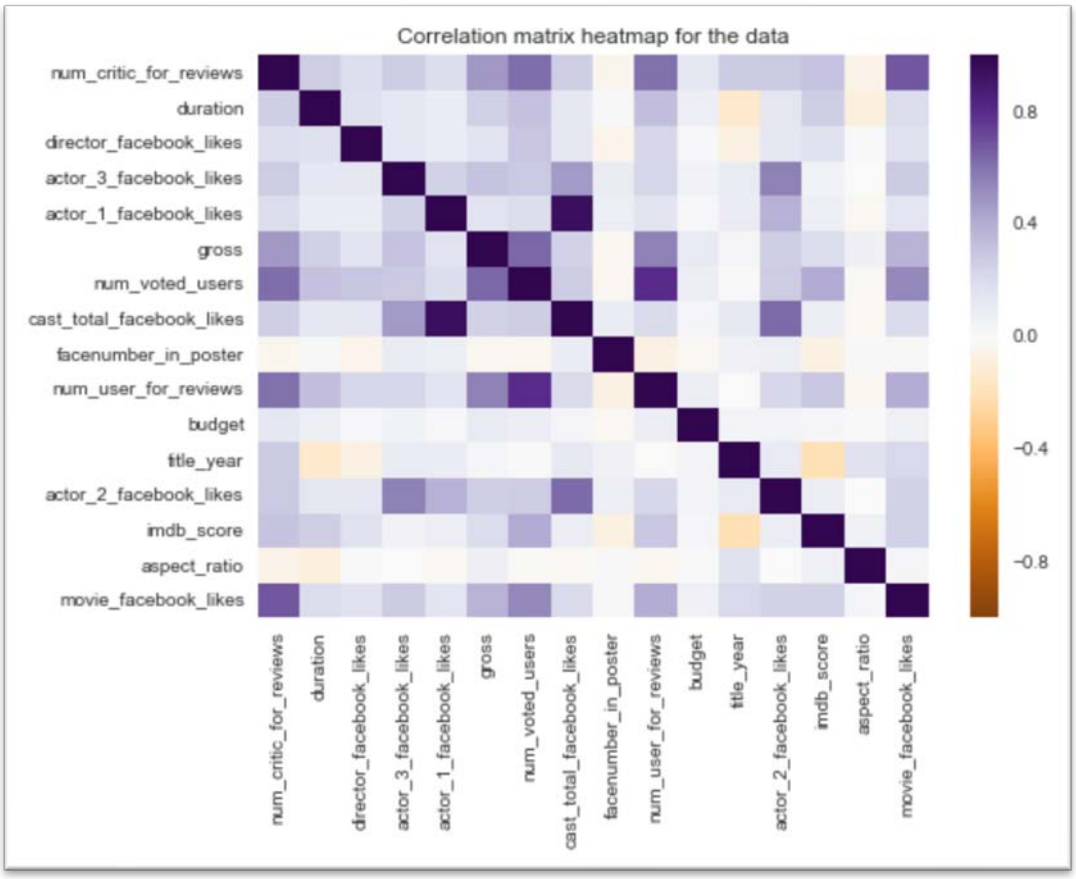
3.) Scatter plot with a marginal histogram and KDE plot with marginal density distribution for budget vs gross.



4.) Scatter plot for cast popularity versus growth with a regression line.



5.) Correlation matrix heatmap for the numerical columns in the data.



IMDB Ratings Prediction Models:

Below helper functions were used for data pre-processing:

- clean_data(): Removes null values from the data and encodes the feature values.
- get_XY_data(): Separates the features and label columns from the data. In this case imdb_score is separated as label.

These models were evaluated using mean absolute error and r2 score of the predicted values w.r.t the actual values of the imdb_score. The train-test split used for this case is 8:2.

Model	Mean Absolute Error	r2 score
SVM	0.778	0.04
Linear Bayesian Ridge	0.583	0.461
Linear Regression	0.582	0.464
Passive Aggressive Regression	1.311	-1.374
K-Nearest Neighbours	0.685	0.235
Decision Tree	0.723	0.049
Multi Layer Perceptron	0.809	0.004
Huber Regression	0.779	0.078

Conclusion:

The results suggest that the best model is the Linear Regression model with an r2 score of 0.464. Linear Bayesian Ridge model is also acceptable with an r2 score of 0.461.