PREDICTING EXCELLENT US MOVIES

By Qian Xu, Julie Wang, Brittany Field

Summary

- Introduction
- Data Preprocessing
- Observations using Tableau
- Data Modeling and Classification
- Conclusions

Introduction

INTRODUCTION

Background

- Dataset containing IMDB information for over 5000 movies globally.

Target

- To predict whether a US. movie made during 2005~2016 is good or not from this dataset.

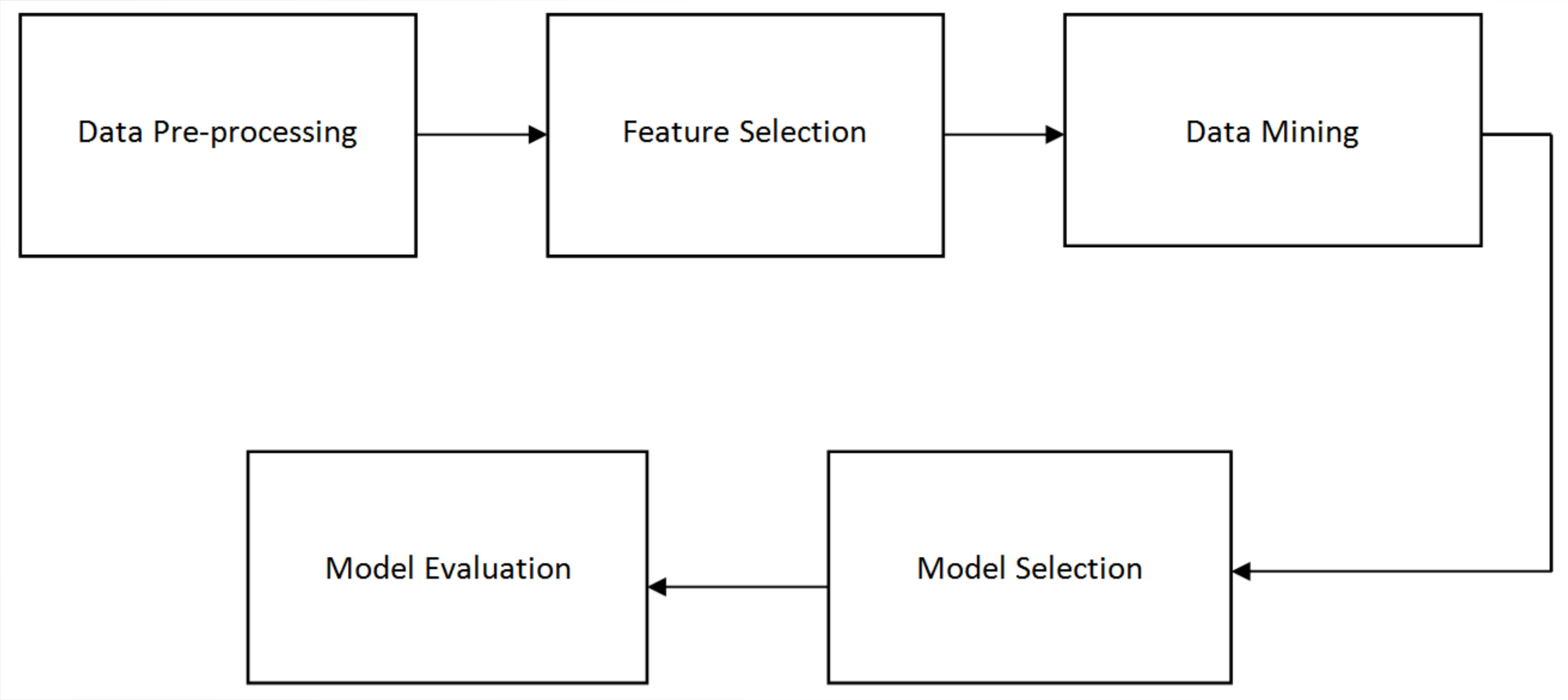
Data Source

- https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

CHALLENGES

- Definition of an Excellent/Not Excellent US. Movie
- Data Preprocessing
- Appropriate Algorithms

Process of Classification



Data Preprocessing

DATA DESCRIPTION

Original data :	28 variables for 5043 movies, spanning across 100 years 66 countries
Data used for Project:	16 variables for 1477 movies, From 2005 to 2016 US Movies Only

TARGET VARIABLE

IMDB SCORE

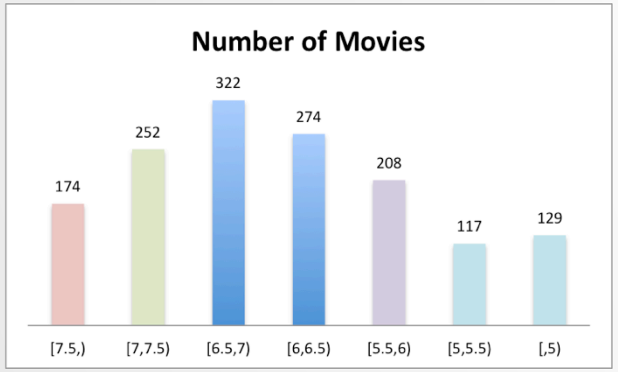
PREDICTOR VARIABLES

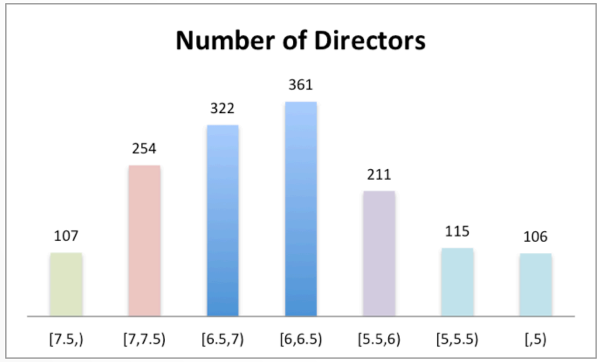
- Quality of Director
- Number of users voted
- Number of critics for reviews
- Director facebook likes
- Actor 1 facebook likes
- Quality of the Actor1
- Movie facebook likes
- Title year
- Cast total Facebook likes
- Num user for reviews
- Gross after CPI
- Budget after CPI

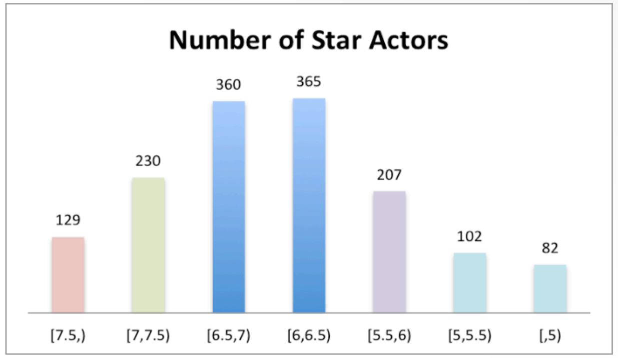
DATA PREPROCESSING

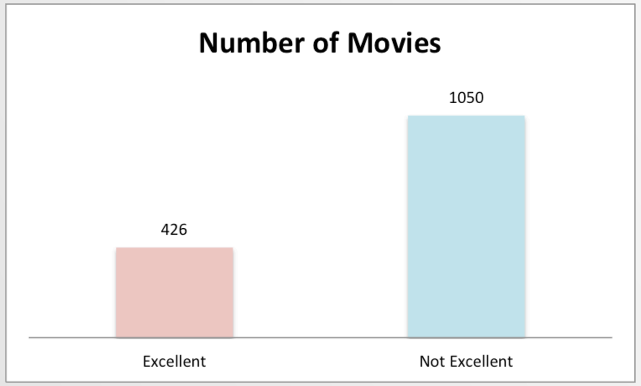
- Delete duplicate data and movies
- Delete columns not useful to classification(date, content rating, duration…)
- Discretization of **IMDB score, directors** and **leading actors**
- Recalculate **gross** and **budget** according to **CPI**

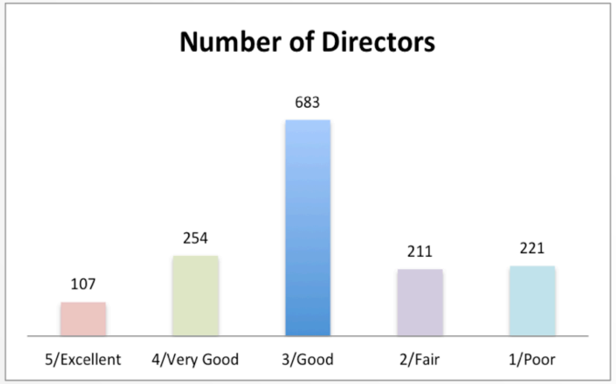
DATA DISCRETIZATION

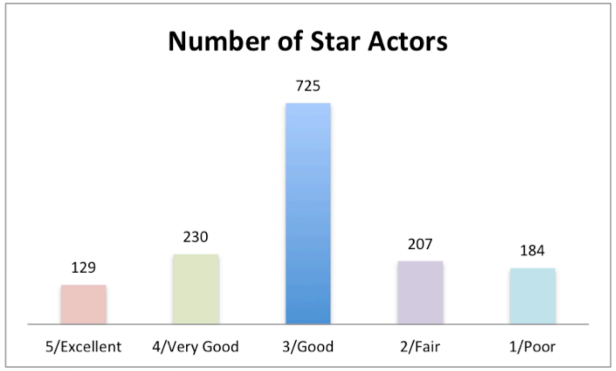




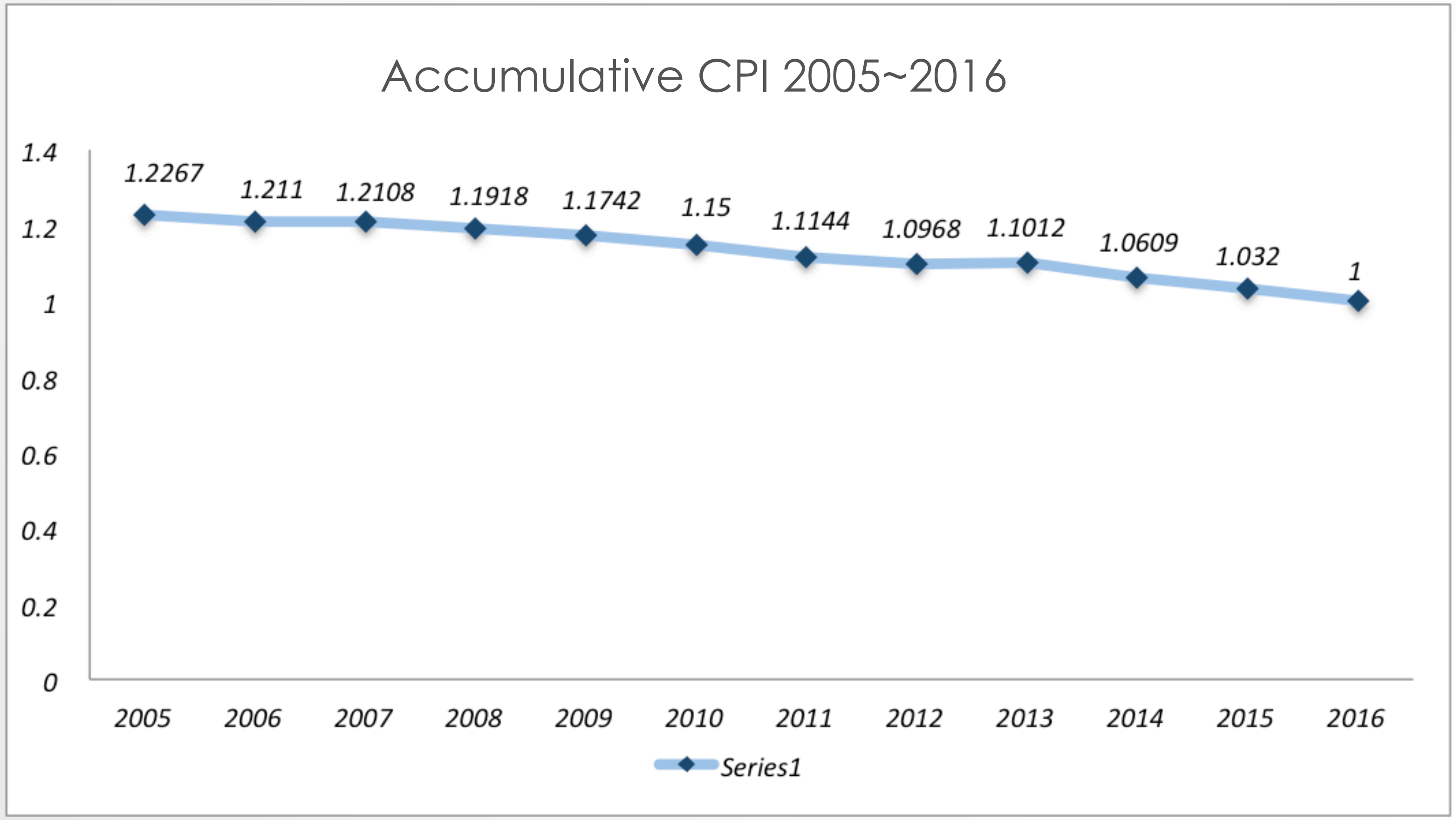






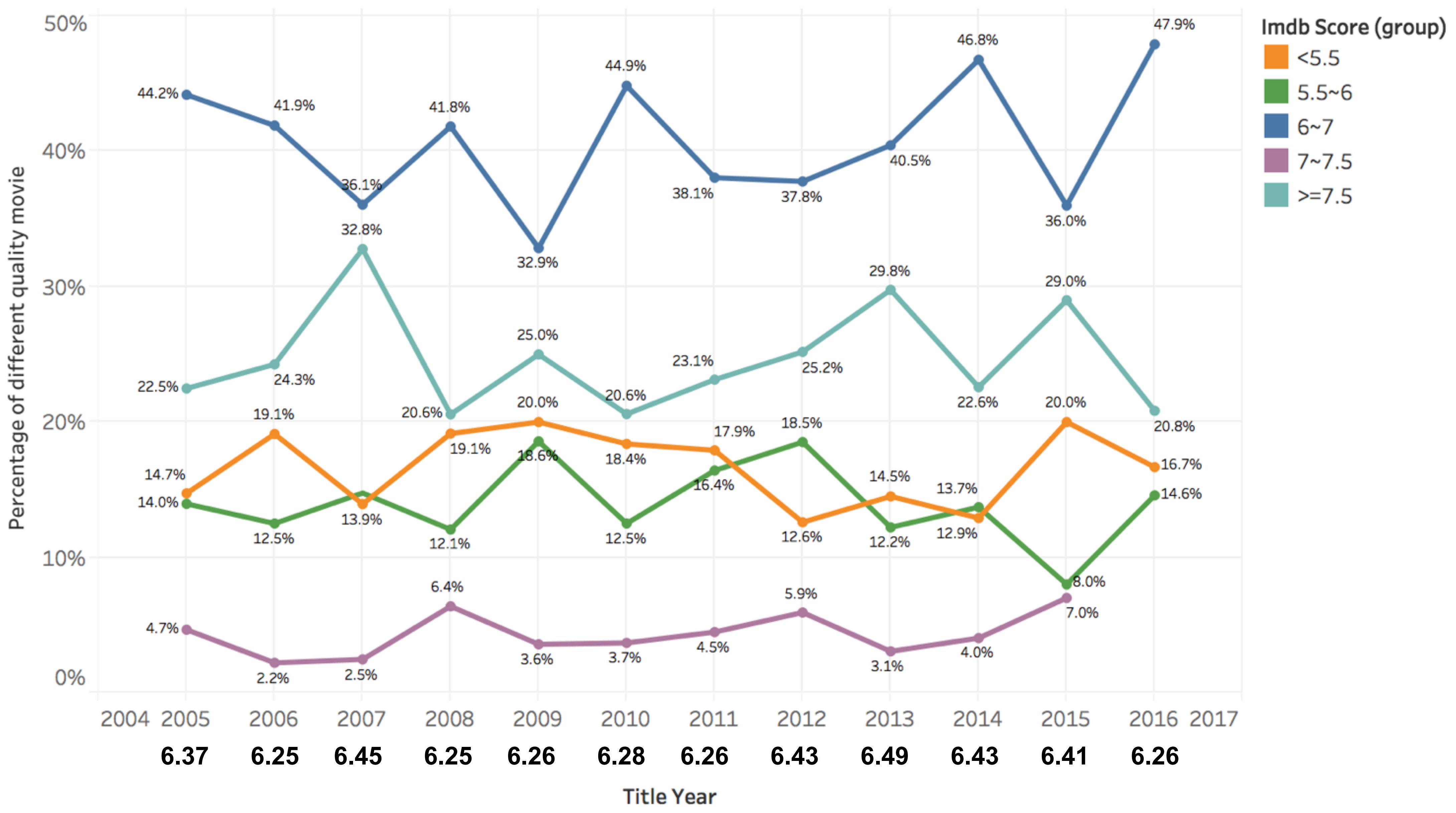


RECALCULATION WITH CPI



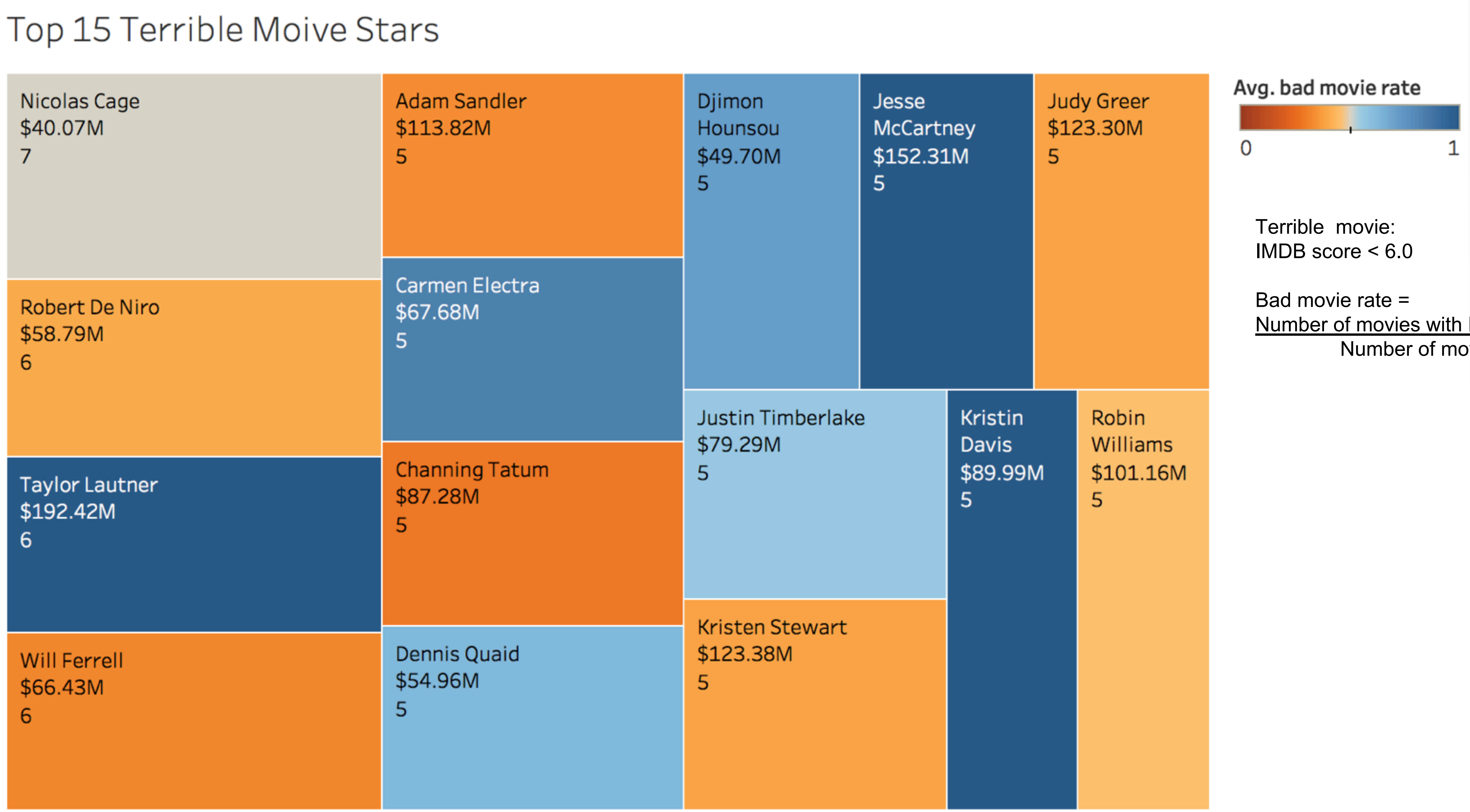
Observations using Tableau

Percentage of Different Quality Movies 2005~2016



The trend of average of number of different quality movie for Title Year. Color shows details about Imdb Score (group). The marks are labeled by sum of number of different quality movie.

ACTORS WHO OFTEN STAR IN TERRIBLE MOVIES



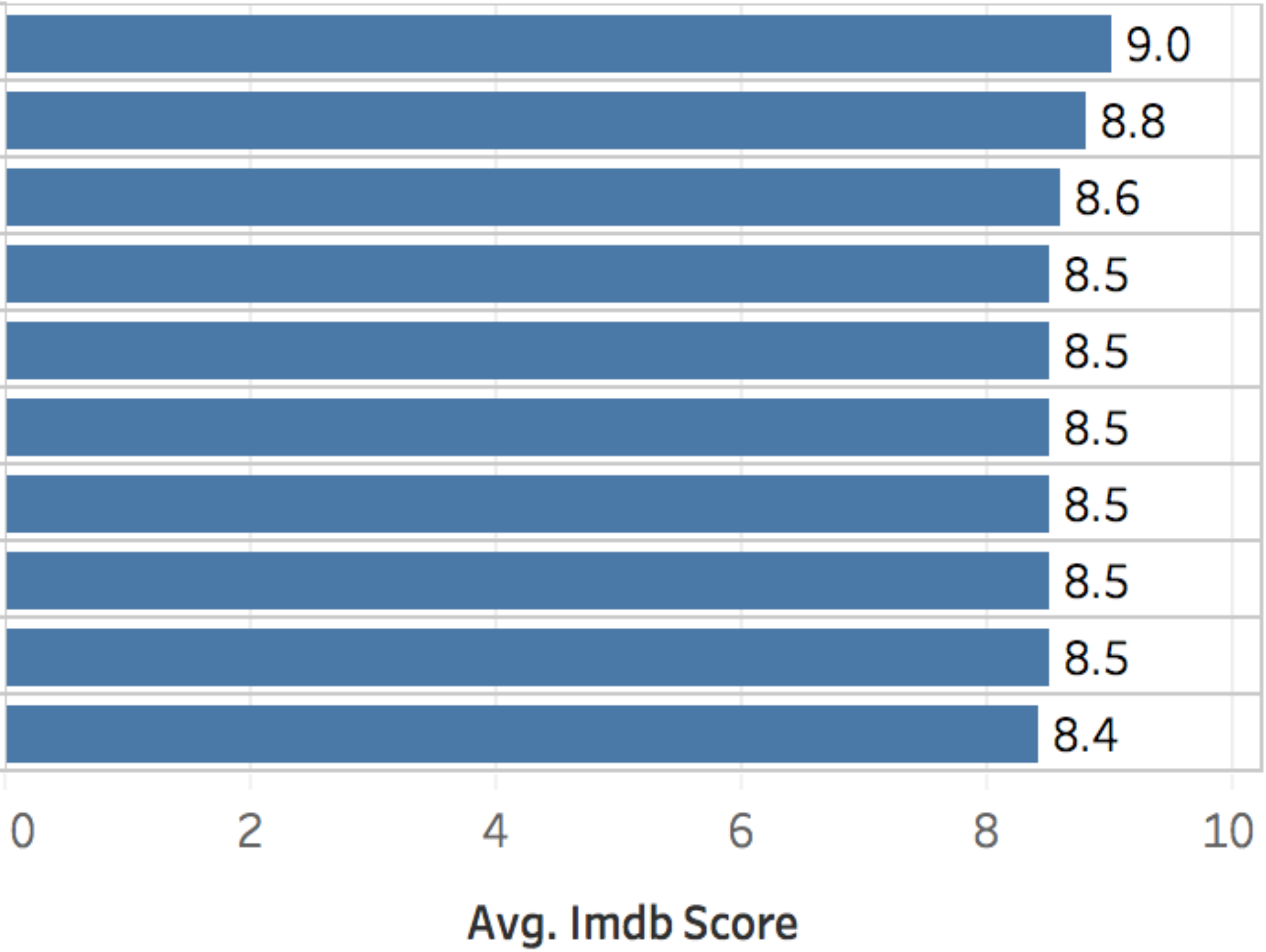
Actor All Name, average of Avg Gross and average of Total Num Movie <6 All. Color shows average of bad movie rate. Size shows average of Total Num Movie <6 All. The marks are labeled by Actor All Name, average of Avg Gross and average of Total Num Movie <6 All. The view is filtered on Actor All Name and average of bad movie rate. The Actor All Name filter keeps 15 of 1,007 members. The average of bad movie rate filter ranges from 0 to 1 and keeps Null values.

Terrible movie:
IMDB score < 6.0

Bad movie rate = $\dfrac{\text{Number of movies with IMDB <6.0 per moive star}}{\text{Number of movies per movie star}}$

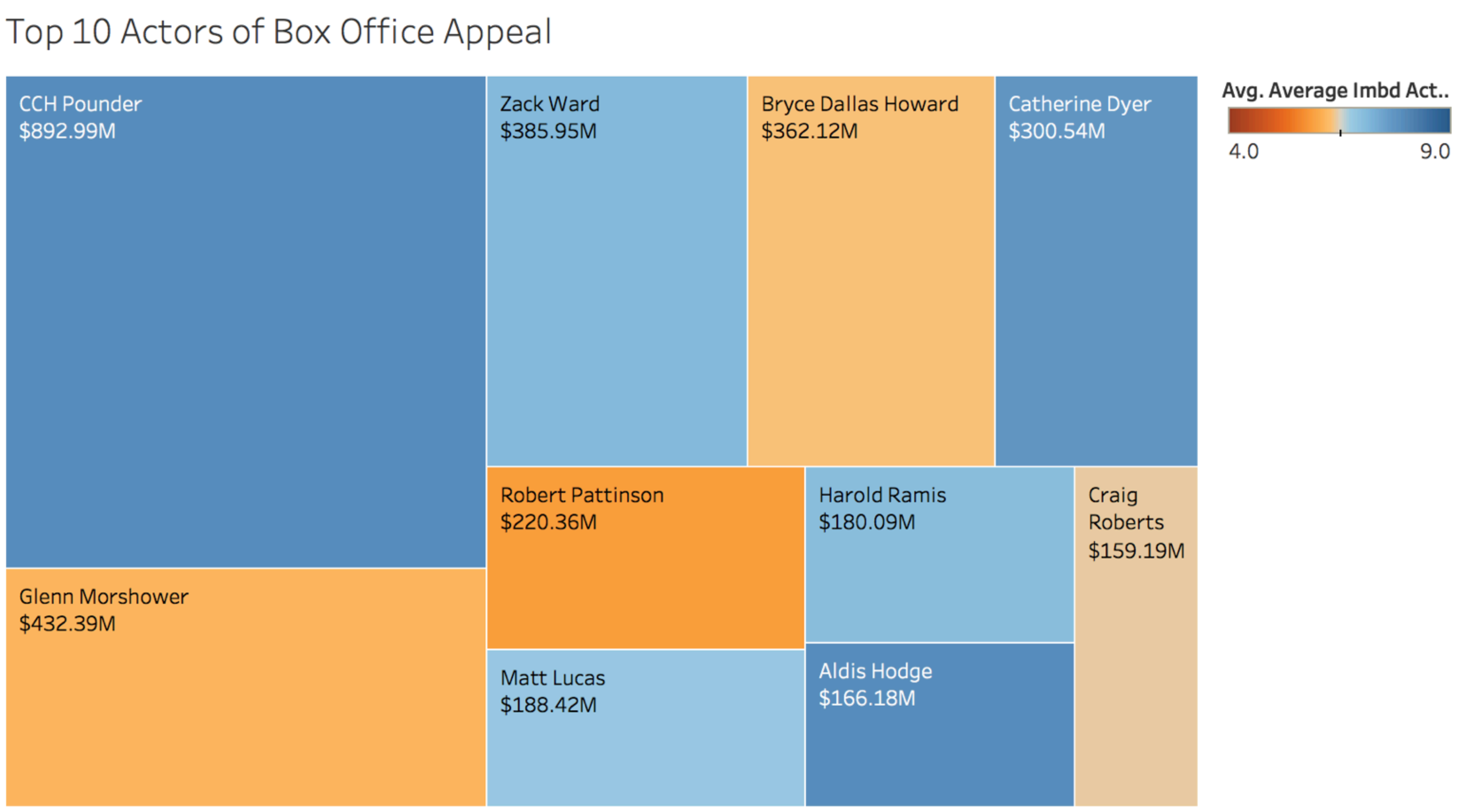
Top 10 Moives by IMBD Score 2005~2016

Movie Title	Director Name	Avg. Imdb Score
Batman Begins	Christopher Nolan	9.0
Inception	Christopher Nolan	8.8
Interstellar	Christopher Nolan	8.6
Django Unchained	Quentin Tarantino	8.5
Hugo	Martin Scorsese	8.5
Samsara	Ron Fricke	8.5
The Dark Knight	Christopher Nolan	8.5
The Dark Knight Rises	Christopher Nolan	8.5
Whiplash	Damien Chazelle	8.5
WALL ·E	Andrew Stanton	8.4



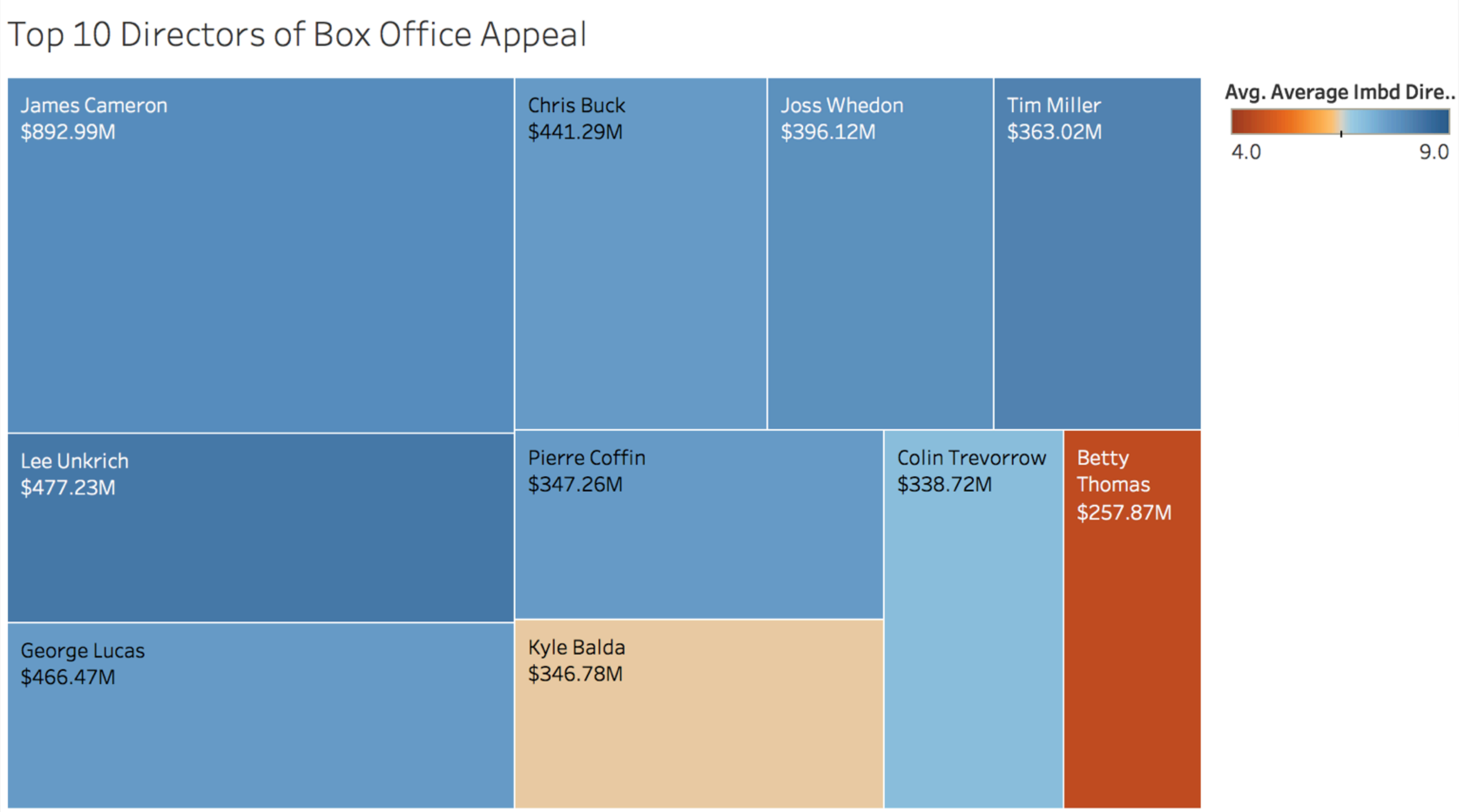
Average of Imdb Score for each Director Name broken down by Movie Title. The marks are labeled by sum of Imdb Score. The view is filtered on Movie Title, which has multiple members selected.

WHO HAS MAJOR BOX-OFFICE APPEAL?



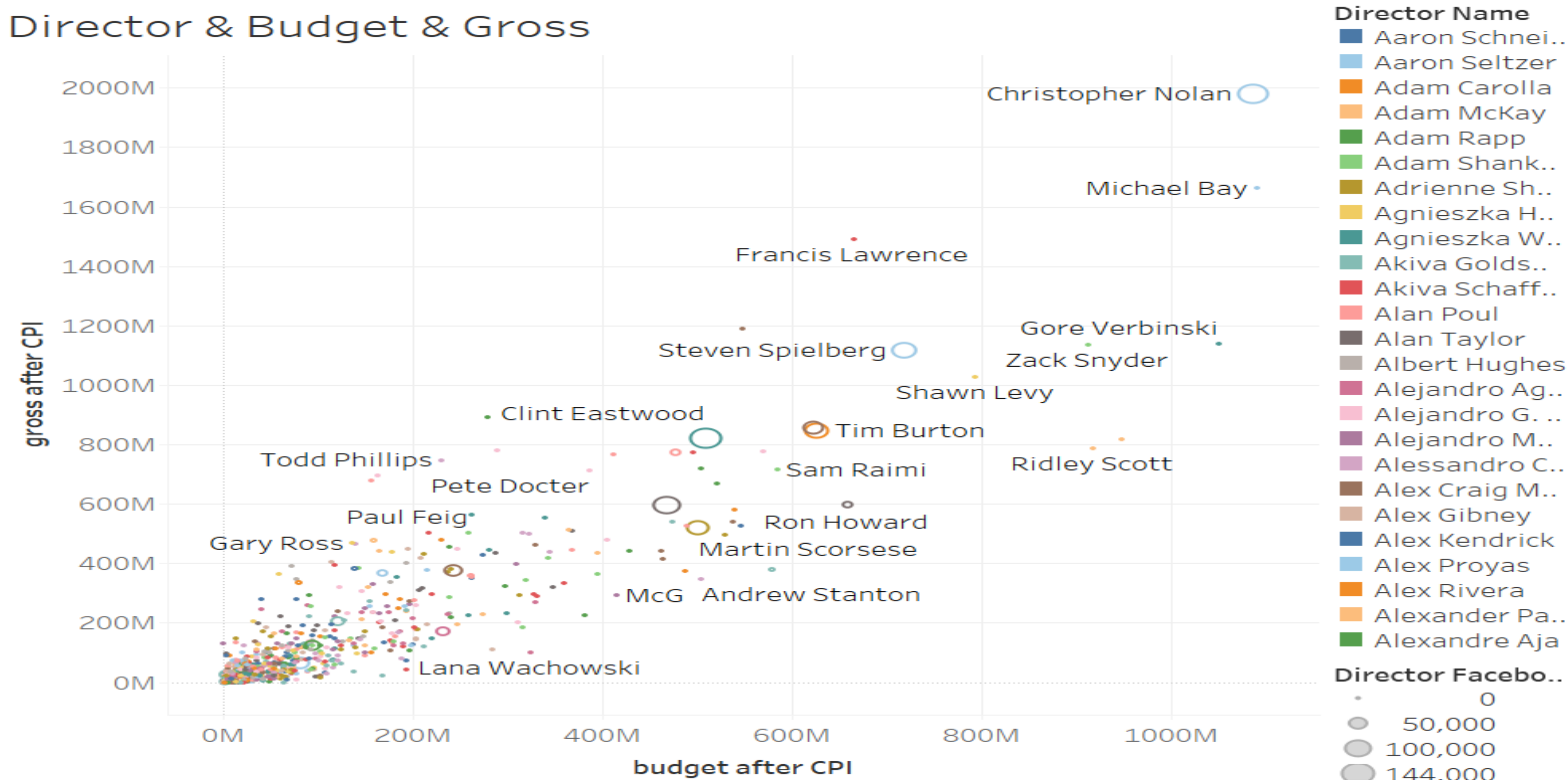
Actor 1 Name and average of gross after CPI. Color shows average of Average Imbd Actor1. Size shows average of gross after CPI. The marks are labeled by Actor 1 Name and average of gross after CPI. The view is filtered on Actor 1 Name, which has multiple members selected.

WHO HAS MAJOR BOX-OFFICE APPEAL?

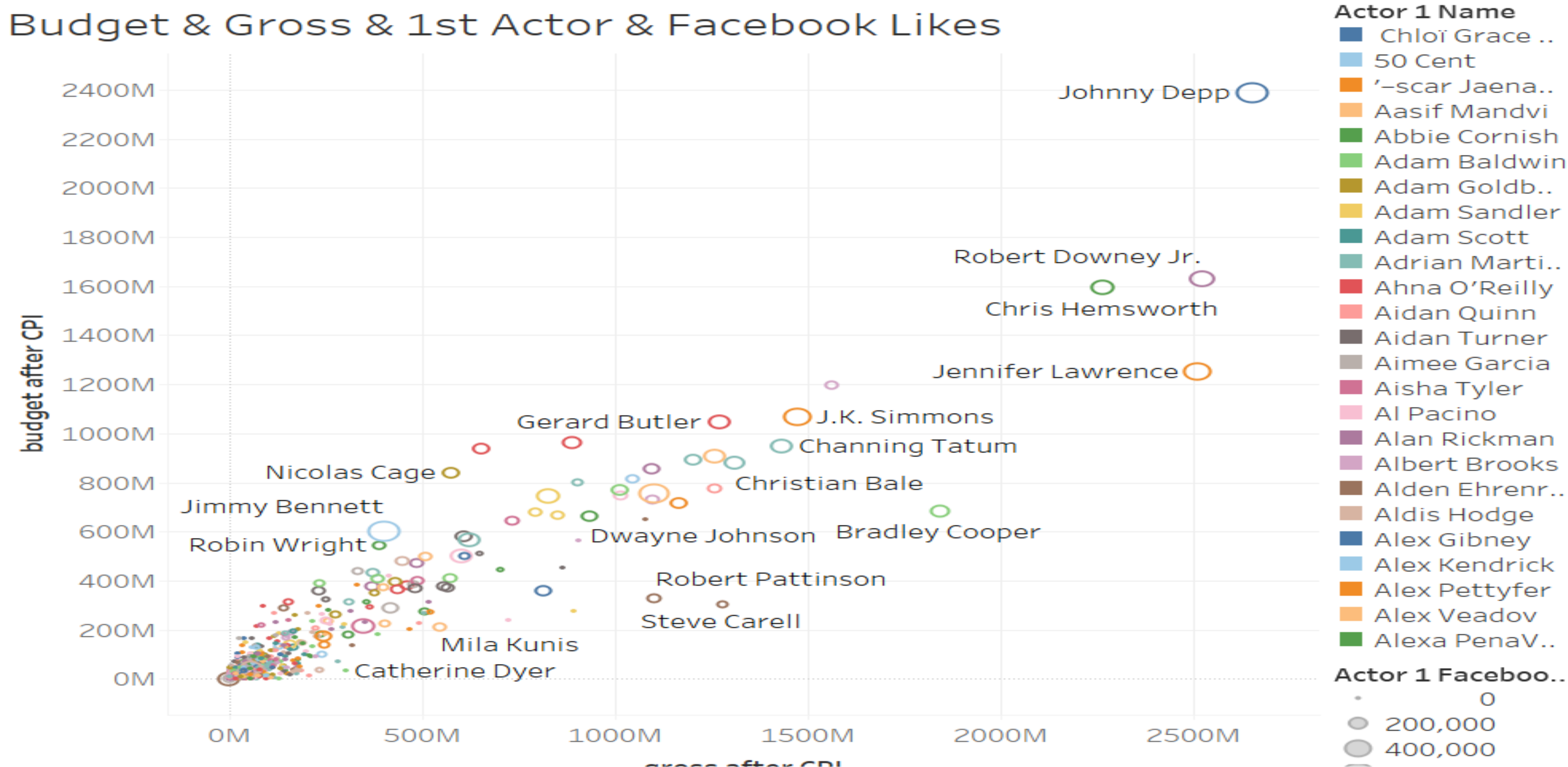


Director Name and average of gross after CPI. Color shows average of Average Imbd Director. Size shows average of gross after CPI. The marks are labeled by Director Name and average of gross after CPI. The view is filtered on Director Name, which keeps 10 of 886 members.

Correlation between Director with Gross, Budget, Facebook❤️



Correlation between 1st Actor with Gross, Budget, Facebook



Data Classification

LOGISTIC REGRESSION

	Observed	
Predicted	0	1
0	385	46
1	28	132

Convert IMDB score to Binomial

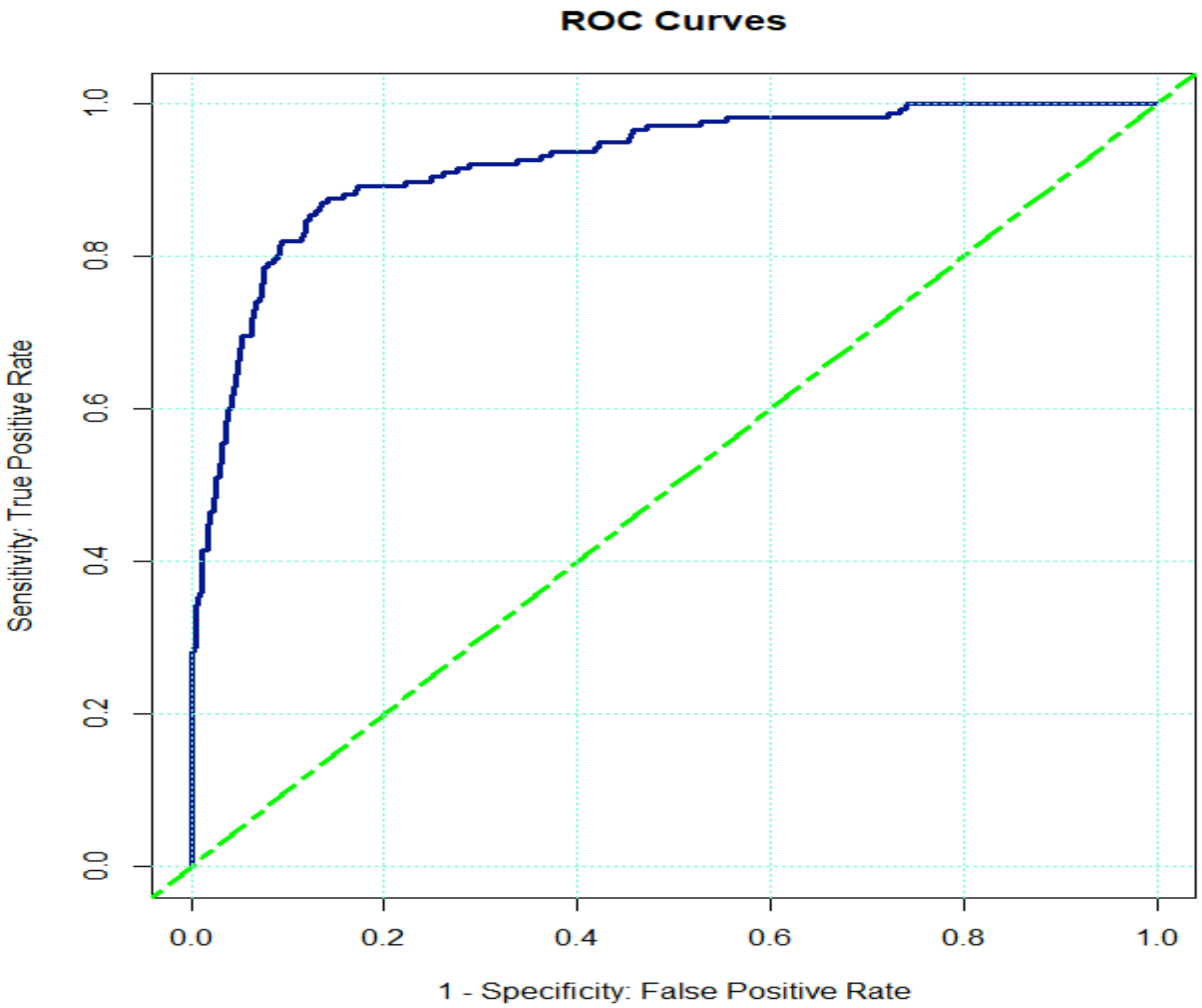
1: Excellent Movie>= 7.0

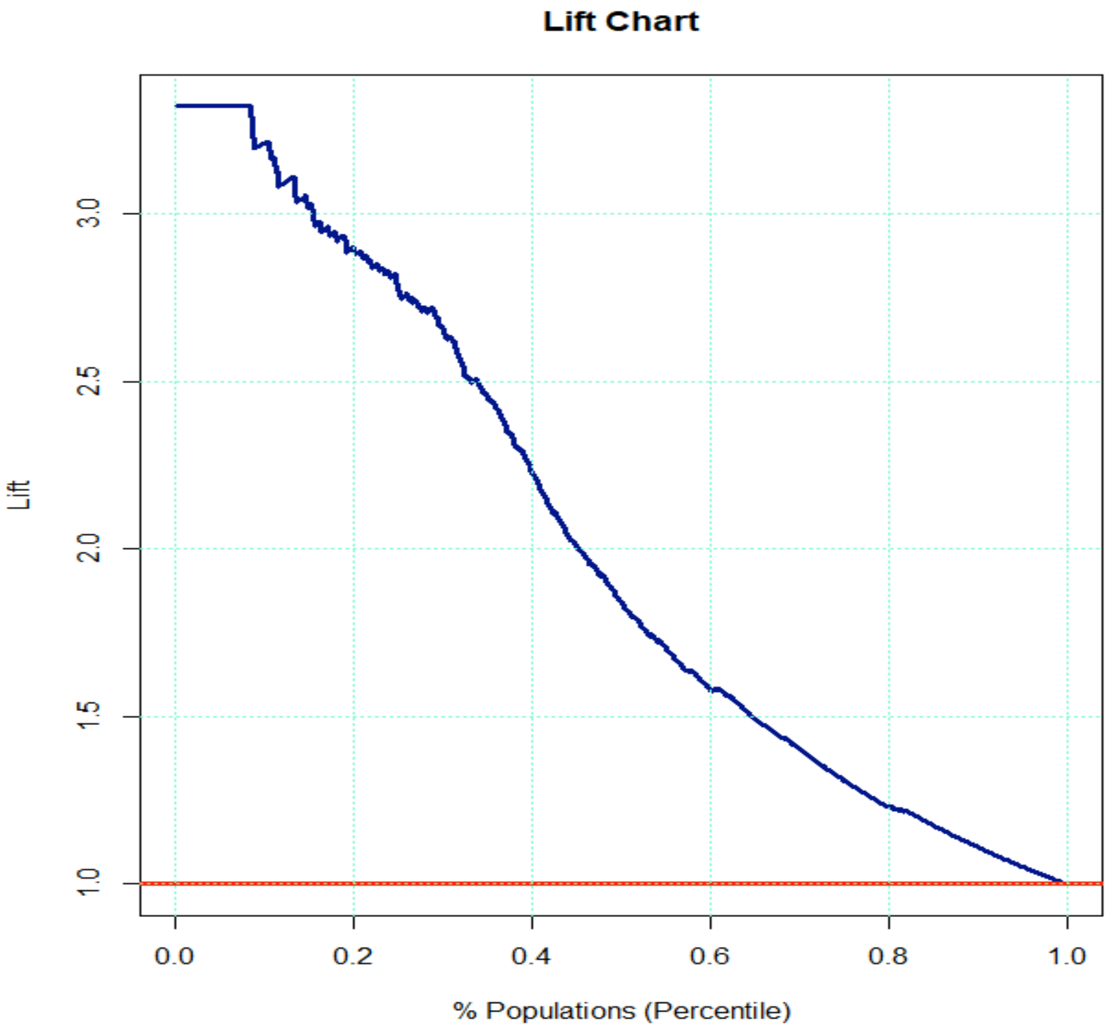
0: Non Excellent Movie<7.0

Overall accuracy rate= 87.48%

Overall error rate =12.52%

LOGISTIC REGRESSION





KNN

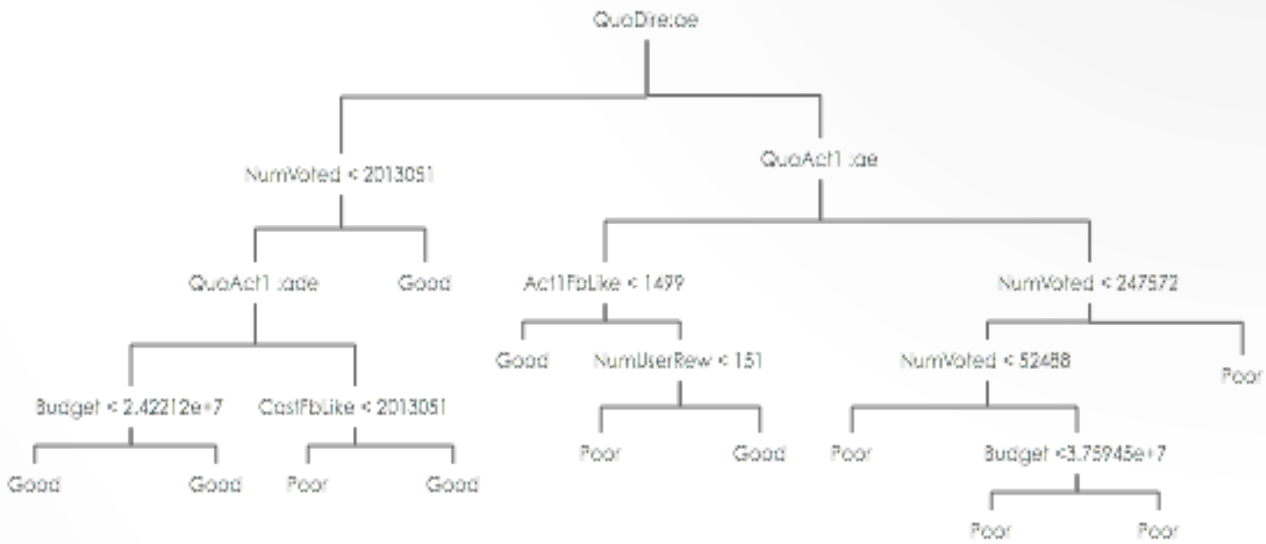
	Observed	
Predicted	Excellent	Not Excellent
Excellent	3	2
Not Excellent	420	1030

K	1	2	3	4	5	6	7	8	9	10
Correct Classification Rate	63.32%	63.25%	68.34%	67.59%	68.14%	68.07%	69.08%	69.36%	69.9%	69.76%

- Sample T : 20
- K = 9
- Overall accuracy rate= 71%
- Overall error rate =29%

DECISION TREE----1

	Observed	
Predicted	Excellent	Not Excellent
Excellent	145	36
Not Excellent	38	371

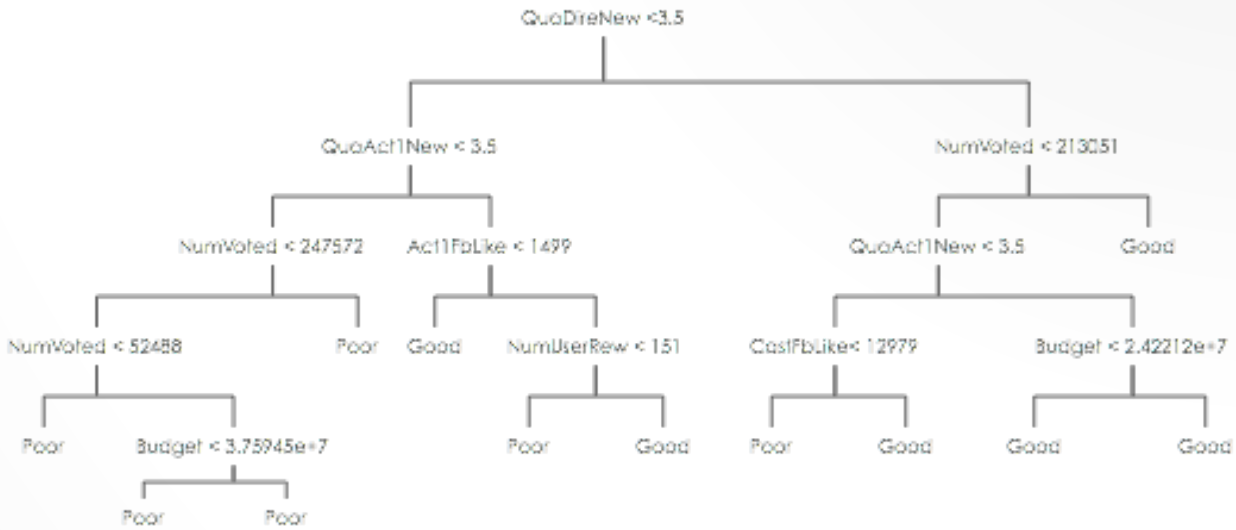


ae:(excellent, very good),(good, fair,poor)   ace:(excellent,poor,very good),(good,fair)

- Binomial variable for levels of directors and actor1s : (,7.5]: Excellent; [7,7.5):Very Good; [6,7):Good; [5.5,6):Fair; [5.5,):Poor
- Overall accuracy rate= 87.46%
- Overall error rate =12.54%

DECISION TREE----2

	Observed	
Predicted	Excellent	Not Excellent
Excellent	86	45
Not Excellent	97	362



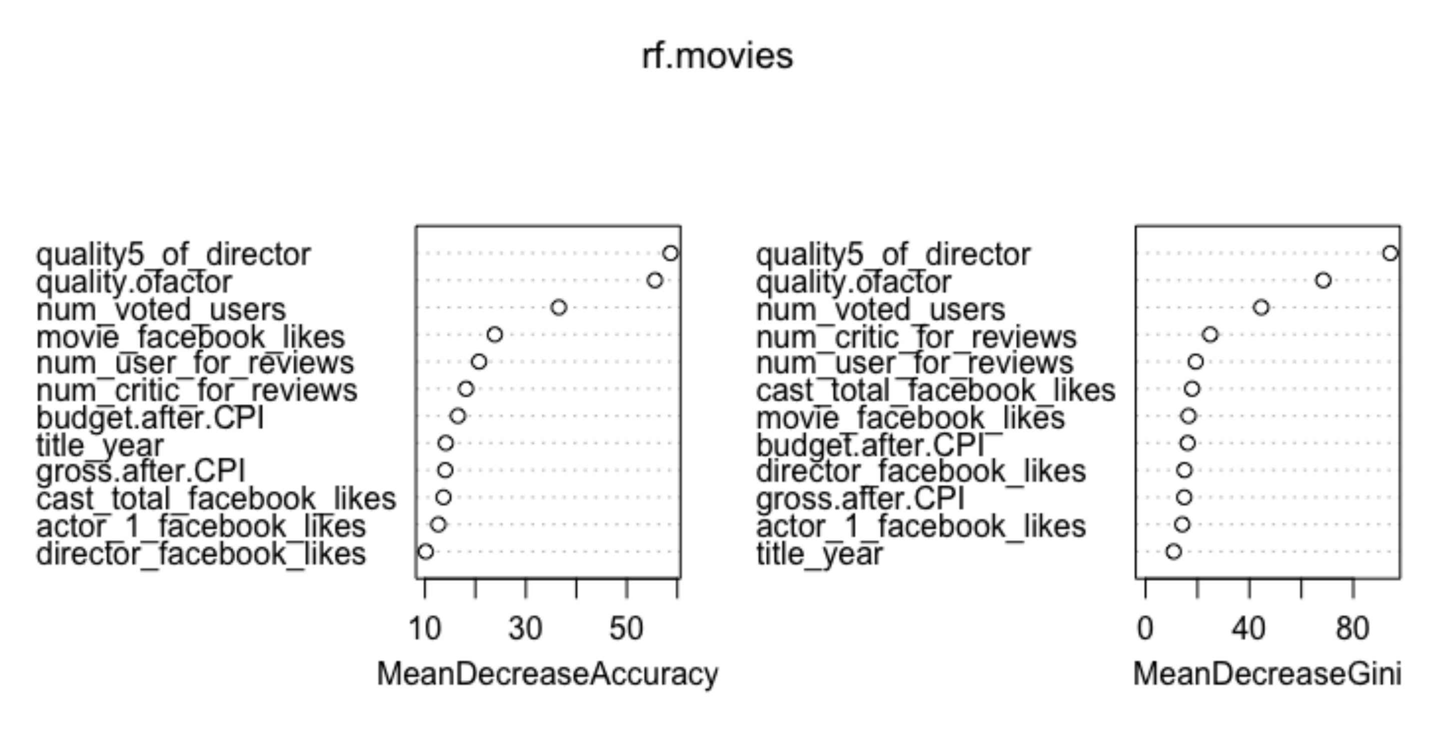
- Numerical Variable for levels of directors and actor1s : (,7.5]:5; [7,7.5):4; [6,7):3; [5.5,6):2; [5.5,):1
- Overall accuracy rate= 87.8%
- Overall error rate =12.2%

RANDOM FOREST METHOD

	Observed	
Predicted	Excellent	Not Excellent
Excellent	154	26
Not Excellent	24	387

- Overall accuracy rate = 91.5%
- Misclassification error rate= 8.5%
- Binomial variable for levels of directors and actors : (,7.5]: Excellent; [7,7.5):Very Good; [6,7):Good; [5.5,6):Fair; [5.5,):Poor

VARIABLE IMPORTANCE PLOT



- According to this plot, quality of directors, quality of actors and the number of users that voted on IMDB for particular movie are the three best indicators of a movie being “good” or “poor"

Model Comparison

Model	Logistic Regression	KNN	Decision Tree--1	Decision Tree--2	Random Forest
Accuracy Rate	87.48%	71%	87.46%	87.8%	91.5%

Conclusions

- When comparing the three models, we can see that the random forest model has the highest accuracy rate and the lowest misclassification rate.

- We believe this to be the model that will best predict whether a movie is ”Excellent” or “Not Excellent”

- One flaw is that the way we categorized the actor and director quality was based on the average IMDB score for their movies. This means that their quality and the quality of the movie are highly correlated because of the way we decided to categorize them.

In conclusion

- While this dataset appeared relatively “clean,” the majority of time was spent scrubbing the data to be able to create useful models

- The way we categorized our data, impacted our models a great deal

- While these models may be able to predict whether a movie is a success by our definition, there are other ways to determine if a movie a success or not

Thank You!