



Movie Analysis for Microsoft

By Benjamin Dean



Introduction

This project is Phase 1

Students were given a scenario



Outline

- Business Problem
- Data Exploration
- Methods
- Results
- Conclusions



Data

There were eleven datasets initially provided and for this project we focus on six of them from IMDB.com, TheMovieDb.org, and the-numbers.com. After determining the profit and ratings as the variables we wanted to target we selected these datasets because they contained data on production budgets, worldwide grosses, director names, movie titles, etc. We then created a profit column by subtracting the production budgets series from the worldwide gross series.



Methods

Once we had selected appropriate data it was time to prepare and model our graphs. We first cleaned the data by getting rid of columns and rows that we did not want. We also removed movie genres in which the count was only one. We then renamed a number of columns so that we could easily merge datasets. The first three datasets were merged on the tconst column, a column that associated movies with a code. The following three datasets were merged on the columns 'nconst', 'primary_title', and 'primary_title' and 'release_date'.

Results





