```
Pearson's product-moment correlation data: scaa hrs and scaa gpa t = 3.275, df = 324, p-value = 0.001171 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.0717779 0.2821460 sample estimates: cor 0.1790072
```

Clustering & Predictive Analysis of Kaggle's TMDb 5000 Movie Dataset

Level 2 Scholar: Qing Li

(M.B.A. & Candidate of M.S. in Statistics, Computational Statistics/Data Science Option)

Department of Statistics & Biostatistics; qing.li@csueastbay.edu

Introduction

The research dataset was collected from Kaggle.com, which is a data publication website for data science research and contests, and it was originally complied by TMDb (The Movie Database API).

The original data has 4803 observations with 23 variables. Important variables include movie names, their release years, production companies, popularity, their budgets, revenues, vote averages, vote counts, genres, casting information, etc. My research only used some of the numerical variables as the explanatory/predictor variables.

My goal of the research is to analyze what predictors contribute to high profit/profit rate and predict the profit/profit rate based on the information got.

My hypothesis for the research are: 1) there are some correlations between profit/profit rate & some or all of the predictors in the dataset; different predictors contribute differently to the profit/profit rates; the profit/profit rate are predictable by the predictors.

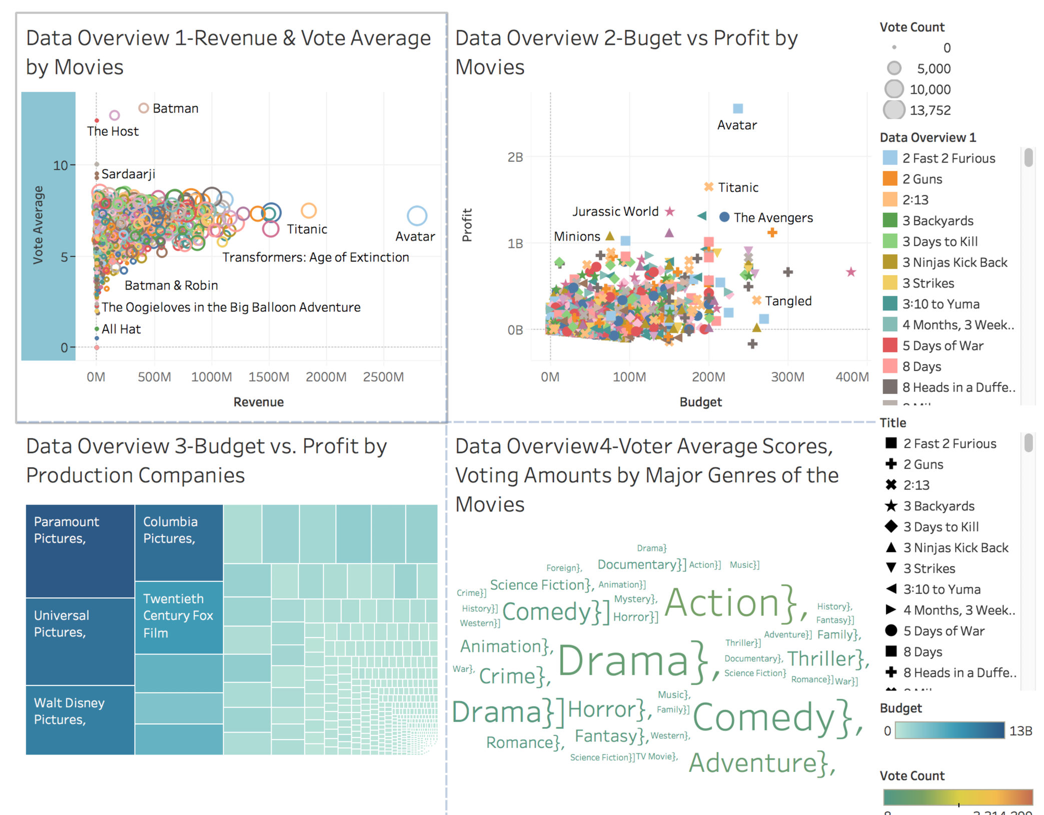
Research Method & Steps

My research process followed three steps:
1) data overview through visualizing data on Tableau(a data visualization software);
2) data preparing & cleaning through R;
3) building clustering and predictive statistical & machine learning models through R, including:
(a) using simple linear regression for getting predictive information;
(b) building a k-means model as the clustering tool;
(c) predicting by a neural network model using the following machine learning steps:
(1) preparing data,
(2) training a model on the data,
(3) evaluating model performance
(4) improving model performance.

Conclusions & limitations of this research was stated followed by the research results.

Results

- **(Partial)Overview & Visualization by Tableau(Figure 1)**



- ***Data Cleaning Process via R:***
- *Checked if unusual values exist by sorting the variables in ascending order under the logic that zero values are missing values;*
- *replaced missing values with the mean values of that variable;*
- *applied the above data cleaning steps to budget, revenue, popularity, run time, vote average & vote count.*
- *generated two new variables to the dataset: 1) Profit=Revenue-Budget; 2) Profit Rate=Profit/Budget.*

- **Linear regression model information(Figure 2-3)**

```
Call:
lm(formula = Profit ~ popularity + vote_average + budget, data = movie)

Residuals:
       Min         1Q     Median         3Q        Max
-830498425  -50479451   -8118100   33844188 2026191839

Coefficients:
                        Estimate        Std. Error  t value              Pr(>|t|)
(Intercept)  -39925980.67944835 10080456.12761365 -3.96073           0.000075803 ***
popularity     1559448.70990560    54633.03960918 28.54406 < 0.0000000000000000222 ***
vote_average   6899467.08199164  1618849.00270066  4.26196           0.000020653 ***
budget               1.18322101        0.04418528 26.77862 < 0.0000000000000000222 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 102975800 on 4799 degrees of freedom
Multiple R-squared:  0.384536,  Adjusted R-squared:  0.3841512
F-statistic: 999.4563 on 3 and 4799 DF,  p-value: < 0.000000000000000022204
Call:
lm(formula = Profit.Rate ~ popularity + vote_average + budget,
    data = movie)

Residuals:
     Min       1Q   Median       3Q       Max
 -507230  -287159  -216582  -119690 116938309

Coefficients:
                    Estimate       Std. Error  t value Pr(>|t|)
(Intercept)  -50501.464190838 419156.373327575 -0.12048 0.904105
popularity    -2279.527538325   2271.701444513 -1.00345 0.315697
vote_average  67582.128002322  67313.509264546  1.00399 0.315434
budget           -0.003785624      0.001837272 -2.06046 0.039408 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4281846 on 4799 degrees of freedom
Multiple R-squared:  0.001973716,       Adjusted R-squared:  0.001349819
F-statistic: 3.163531 on 3 and 4799 DF,  p-value: 0.02352139
```

- ***The Interpretation of R results in Figure 4:***
- *The results from Figure 2-3 suggested that budget, Popularity & vote-rate all had positive influence to profit and the three predictors together explained about 38% of all the influences; however, when it turns to profit rate, only budget generated some significant negative influences on the profit, which only explained about 0.13% of all the influences.*

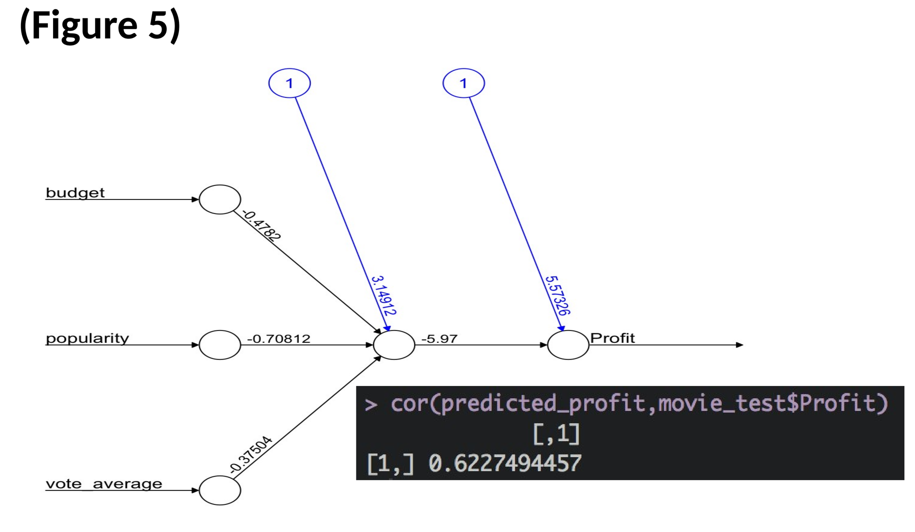
- **K-means clustering model and its Result (Figure 4)**

	budget	popularity	Profit.Rate	runtime	vote_average	vote_count	Profit
1	2.7536038462	3.0843804104	-0.04137724187	1.0079171383	0.9798967599	4.0177946902	4.0222301594
2	-0.3408259689	-0.1780573986	0.03091517200	0.2833063281	0.5502799439	-0.2528845837	-0.2748238833
3	-0.1959396605	-0.3660266247	-0.01985593835	-0.5998771410	-0.9045077032	-0.3940409330	-0.2063966333
4	1.1029745062	0.9453891546	-0.04137780003	0.4818641427	0.4243421036	1.0711064053	0.6184214692

- ***The Interpretation & Conclusion of R results in Figure 4:***
- I used a z-standardization method to make all the variables numbers not far from 0, which is their standardized mean values. The more positive they are, the more they are from the mean; the more negative they are, the less they are from the mean.
- With my settings of k=4, the model clustered the movies into four groups: Group 1 had the highest reputation (vote average), budget, profit and good popularity, but with comparatively low profit rate; Group 2 was the winner in terms the profit rate and reputation, but it had the lowest (both below average)budget and profit; Group 3 was the loser group with everything below average; Group 4 had every other variables in the middle (ranked either 2 or 3 out of 4 groups) but had the lowest profit rate.

- ***In other words, from a investor's point of view, the high profit/return rate is mainly associated to low budget but not high revenue; and although good reputations of the movies are associated with high profits, it is also very likely to have a negative effect on the profit rate due to their high budgets.***

- **Prediction based on the Neural Network Model (Figure 5)**



- ***The Interpretation of R results in Figure 5:***
- *This neural network model suggested a moderate towards strong correlation in predicting profit by budget, popularity and vote average. But the same three predictors failed to predict profit rate, which matched the information gain from linear regression model and k-means model.*

Conclusions & Limitations

- Movies with high budgets are more likely to bonded with high profit and high reputation/vote scores, but it could also increase the risk of getting low profit rate because of the comparatively high cost or budget invested to the movies.
- Profit rate might not be the only goal every movie makers want to reach, it could be possible that they value reputation more, which is not the focus of this research.
- Since all the categorical variables or factors are in JSON forms, I was not able to clean the data in that sense. In other words, since the numerical predictors can only say no more than 38% of the associations, more predictable information that the categorical variables like genres, production companies and casting/staffing may offer was not analyzed in my models.
- Not all of the data was collected accurately due to the limitation of TMDb data collecting process, which would also make the analysis and prediction less accurate.

Helpful References

1. Lantz, Brett (2015). *Machine Learning with R*(2nd Ed.). PACKT Publishing . ISBN 978-1-78439-390-8
2. Yan, Nathan (2011), *Visualize This, The FollowingData Guide to Design*(2nd *Ed)*, Wiley Publishing, Inc., ISBN: 978-0-470-94488-2

Acknowledgments

Great appreciation to my faculty mentor Prof. Eric A Suess, who have offered constructive recommendations and knowledgeable support to this research, during office hours and within his classes.