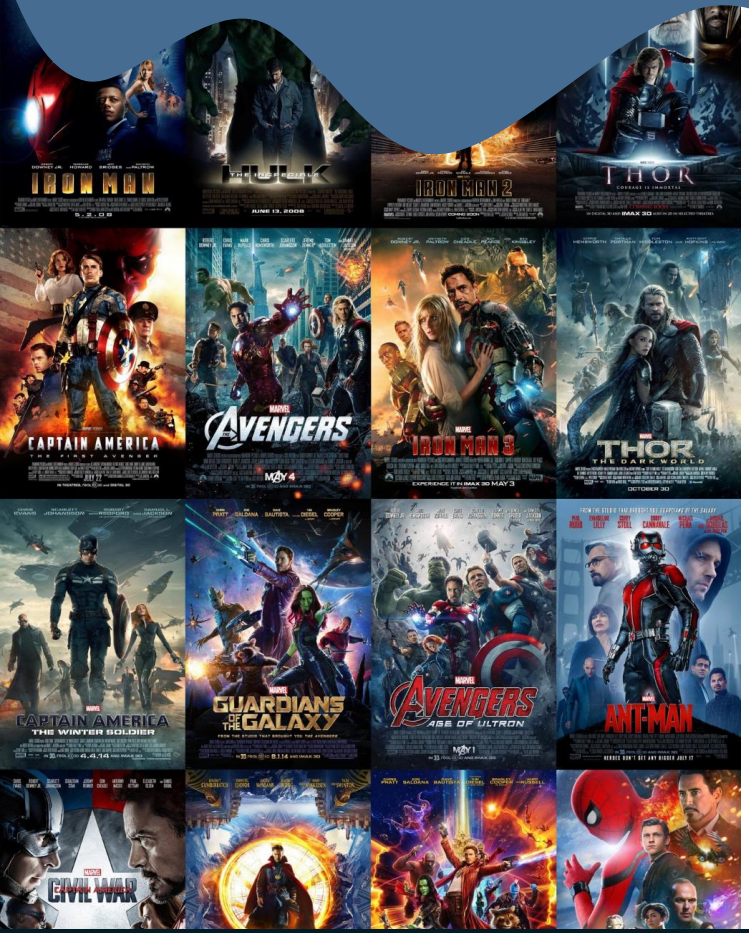
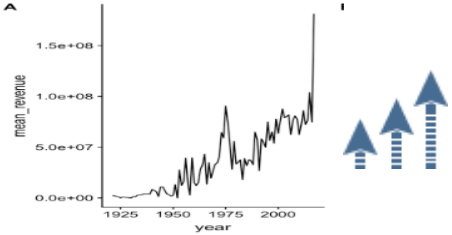
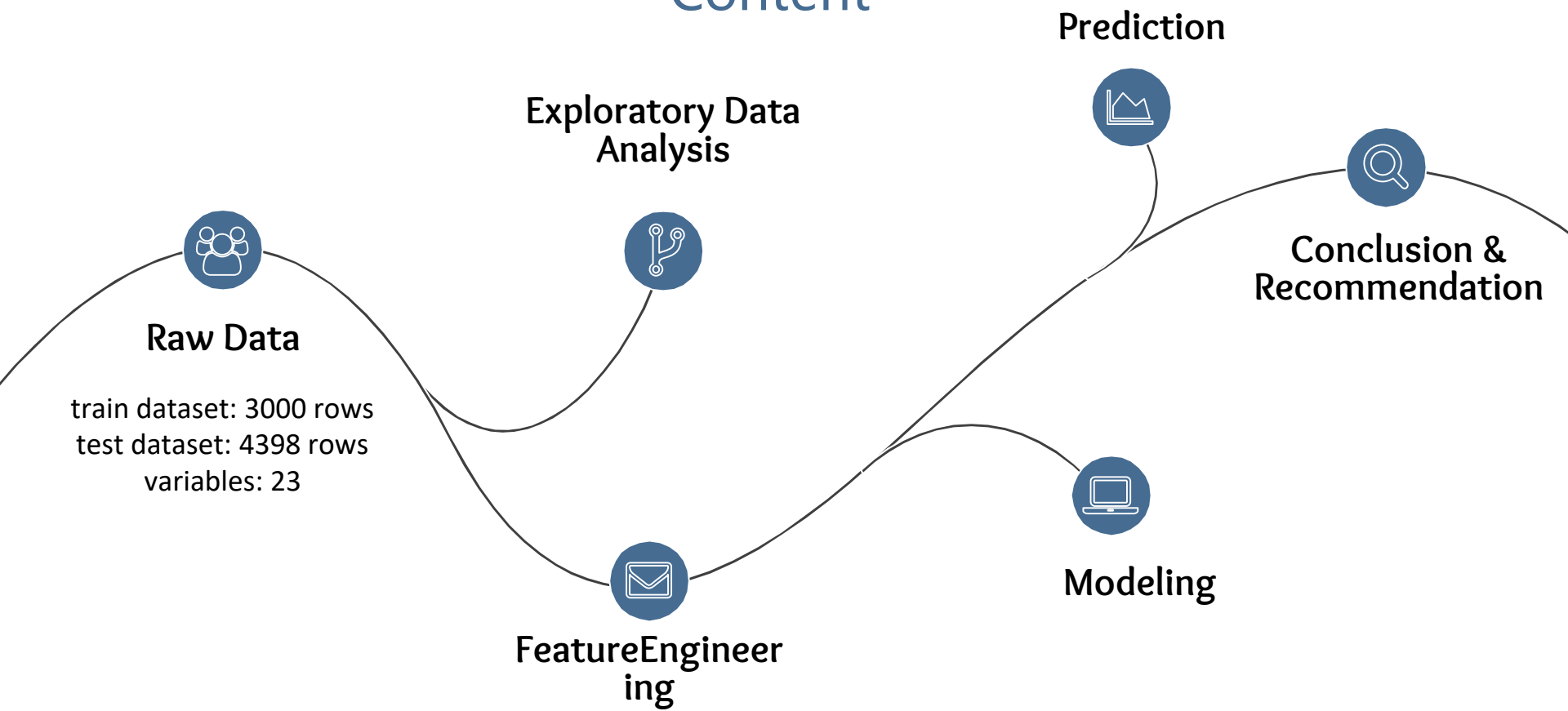


TMDB Box Office Prediction

Team: Yuting Gong, Cijun Sun,
Mianchun Lu, Miller Luo



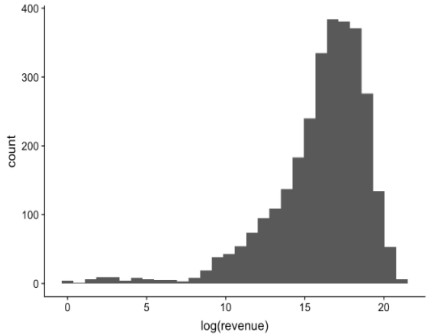
Content


Prediction
Exploratory Data Analysis
Conclusion & Recommendation
Raw Data
train dataset: 3000 rows
test dataset: 4398 rows
variables: 23
Modeling
FeatureEngineering

EDA and Feature Engeering

Variable type	Variable name
numeric	Budget, popularity, runtime, revenue
categorical	original_language, production_countries, spoken_languages, status
JSON format	Genres, production_companies, belongs_to_collection, cast, crew,
text	Original_title, tagline, title, keywords, overview
date	release_date

dependent variable
(transfer to log)



"[{'id': 35, 'name': 'Comedy'}]"
"[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'name': 'Drama'}, {'id': 10751, 'name': 'Family'}, {'id': 10749, 'name': 'Romance'}]"
"[{'id': 18, 'name': 'Drama'}]"

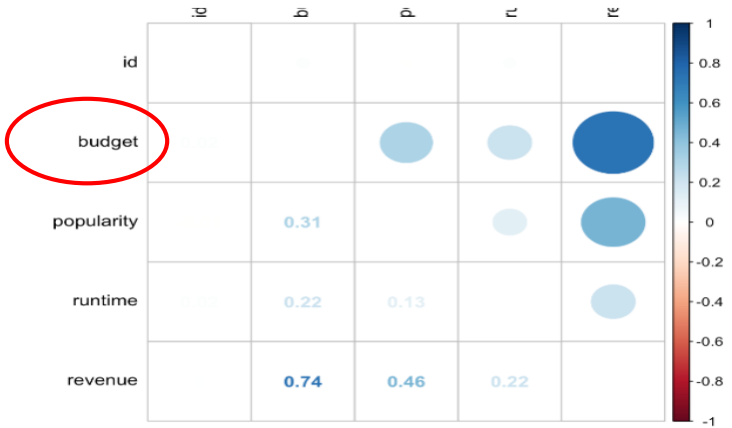
"A live-action adaptation of Disney's version of the classic 'Beauty and the Beast' tale of a cursed prince and a beautiful young woman who helps him break the spell."

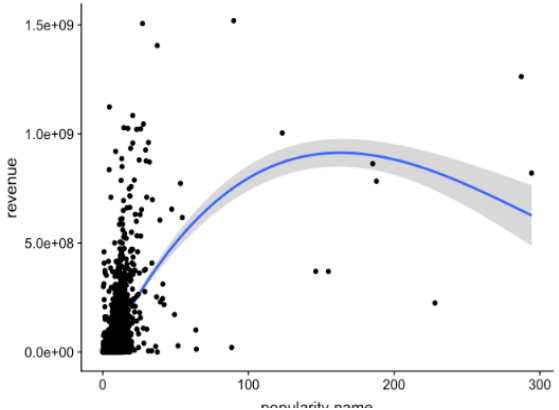
"2/20/15" "8/6/04" "10/10/14" "3/9/12"

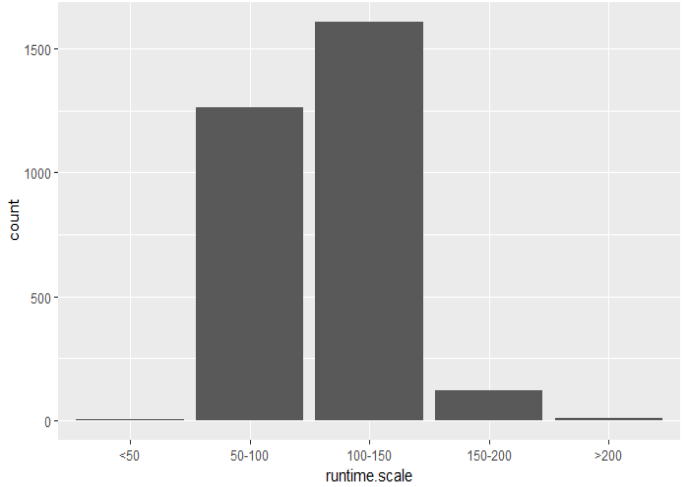
Drop

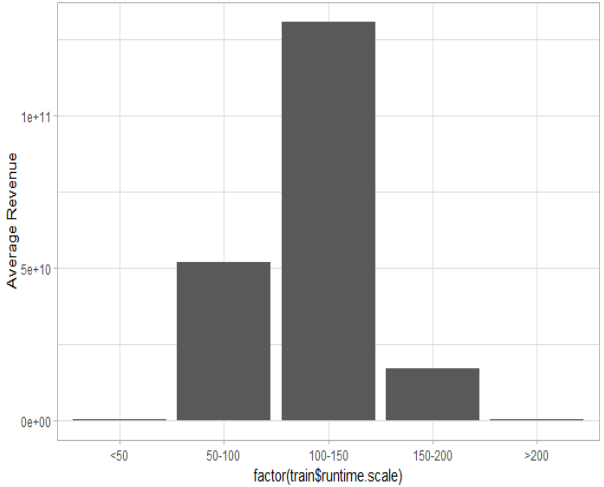
```
##  $ homepage          : chr  NA NA "http://sonyclassics.com/whiplash/" "http://kahaanithefilm.com/" ...
##  $ imdb_id           : chr  "tt2637294" "tt0368933" "tt2582802" "tt1821480" ...
```

Variable: Numeric









Budget has the highest positive correlation with revenue. And **popularity** is the second one.

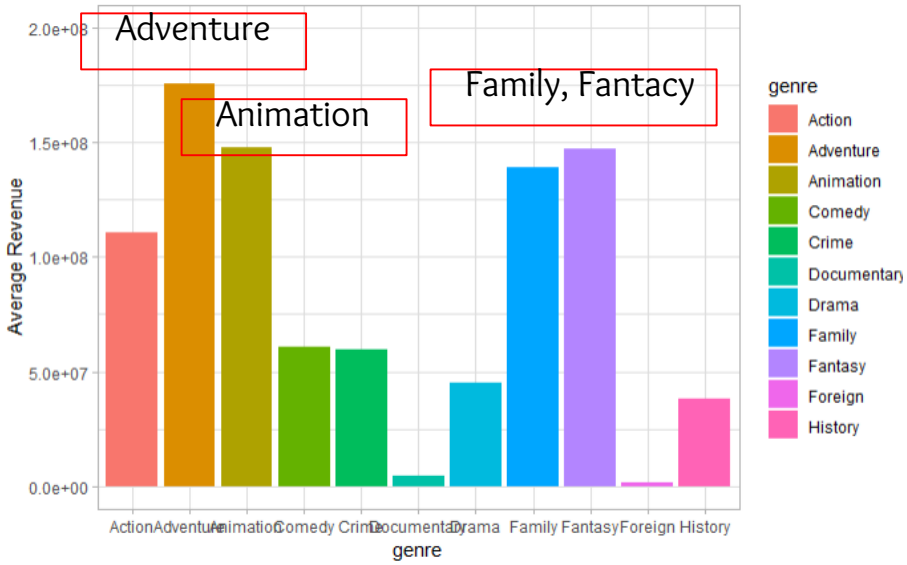
Runtime:

- ➔ 0--50 min
- ➔ 50--100 min
- ➔ 100--150 min
- ➔ 150--200 min
- ➔ 200+ min

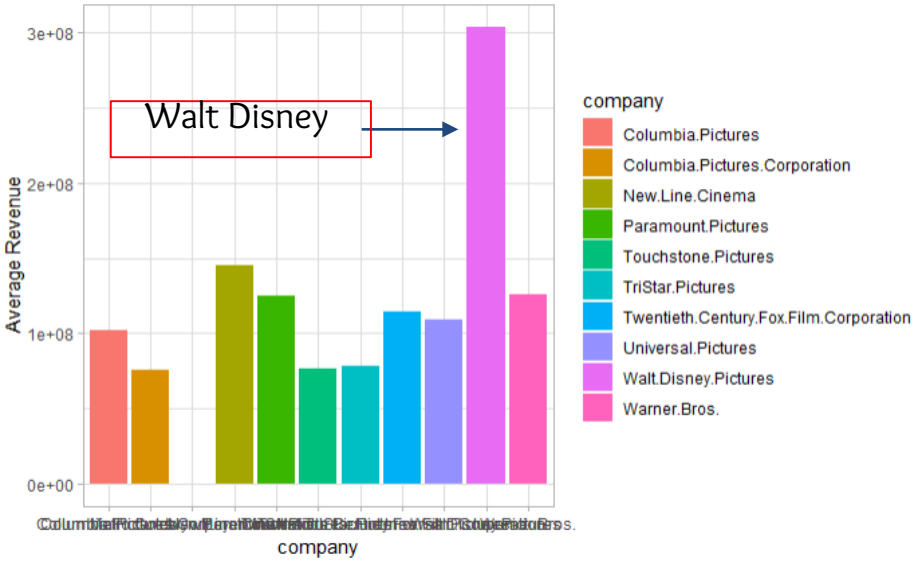
Movies with runtime between 100 minutes and 150 minutes have the highest average revenue.

Variable: JSON

Genre:



Producing Company:



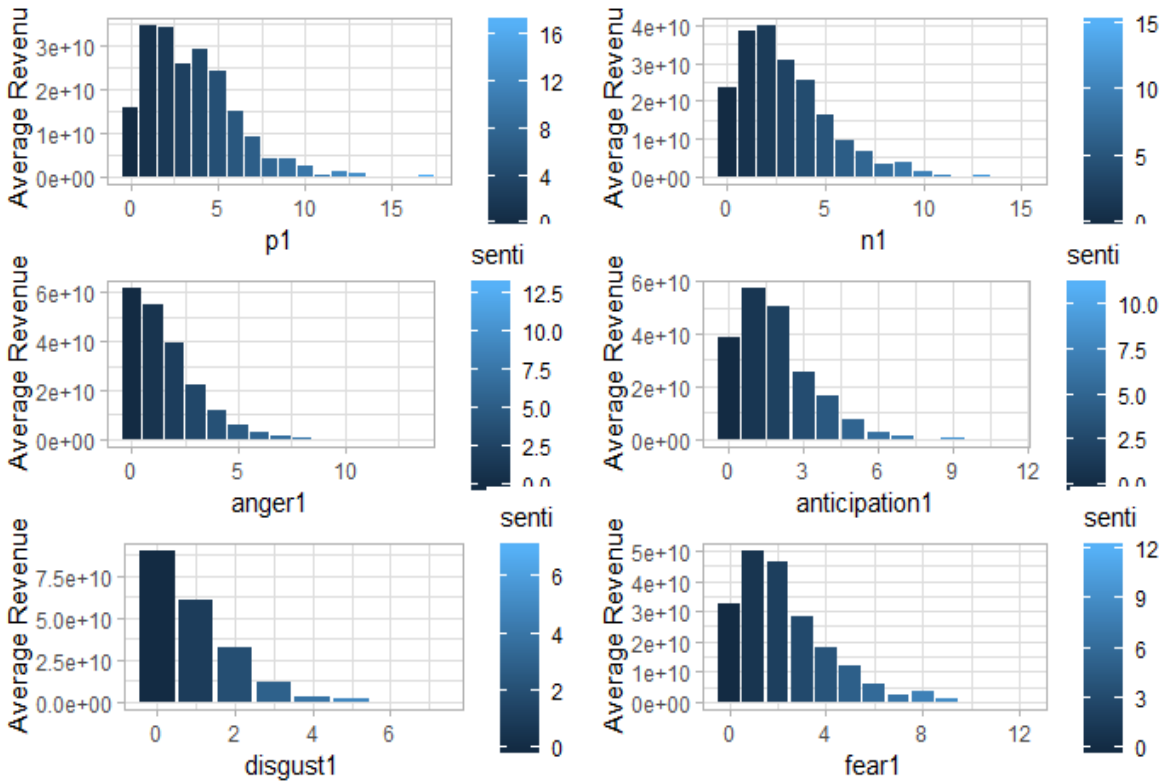
The most common types:
Drama, Comedy, Thriller, and Action

The types with higher average revenue:
Adventure, Animation, Family, and Fantasy

Companies produced more movies:
Warner Bros, Universal Pictures, and Paramount Pictures

Company having the highest revenue:
Walt Disney

Variable: text



Movies with high revenue tend to include limited number of sentimental words.
The emotion of their overview is relatively neutral.

	WORD	FREQ
1	life	1257
2	after	1165
3	new	1066
4	young	931
5	world	818
6	man	782
7	love	772
8	family	730
9	story	686
10	must	613
11	film	585
12	only	573
13	while	558
14	finds	548
15	years	525
16	where	507
17	father	476
18	help	468
19	woman	464
20	back	461
21	friends	452
22	war	429
23	lives	425
24	own	423
25	home	411

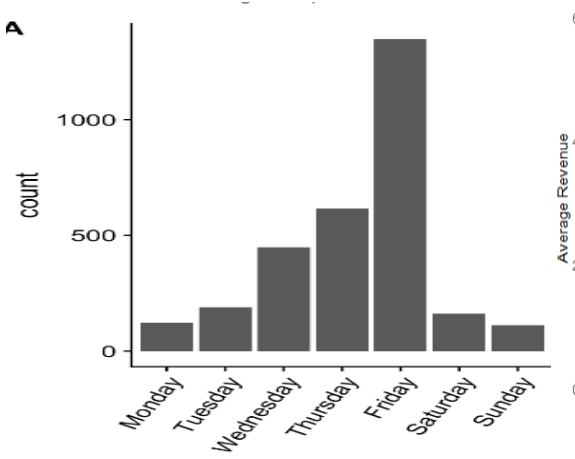
Variable: Date

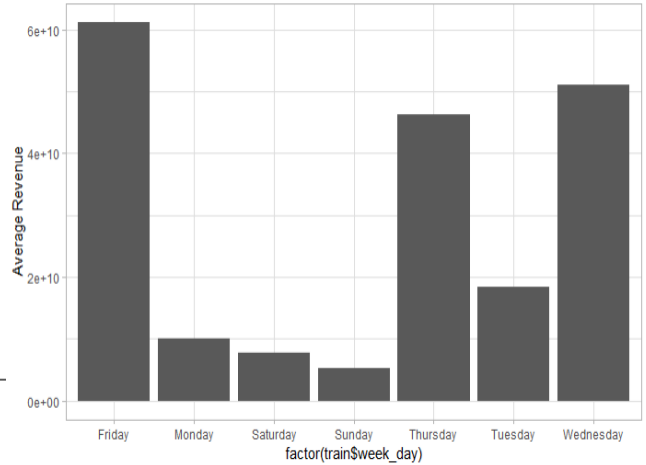
Feature Engeering

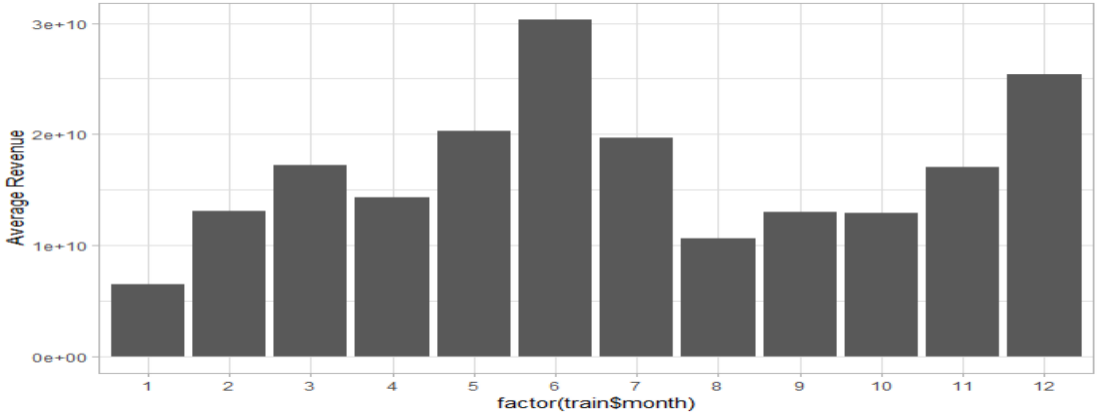
"2/20/15"

Release Date

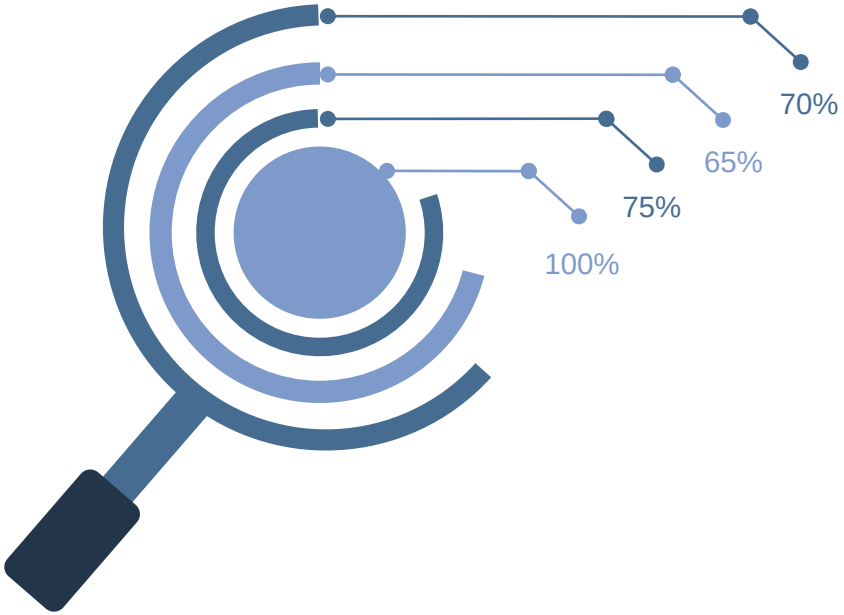
- Year
- Month
- Day
- Week Day







PART II. Model Fitting



- XGBOOST
- CLUSTERING THEN PREDICT
- DEEP LEARNING
- RANDOM FOREST
- SVM

Model Experimentation - Clustering then predict + xgboost

Process:

- Data preparation: `dummyVars` to dummy,and choose `fullRank=T` to avoid linear dependencies between the columns
- K-means clustering
- A k= 2 clusters are suggested by both Total Within Sum of squares plot and Silhouette width plot
- Built two xgboost models based on clusters and combine

Learning

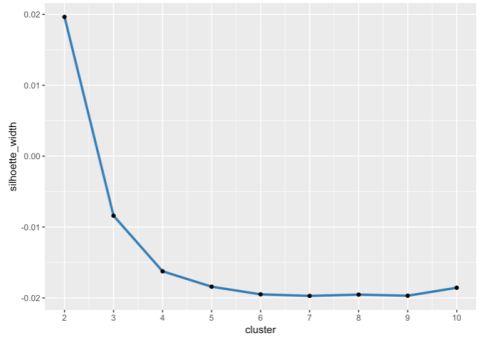
- Clustering does not work for our movie data set
- Dummy works

Package and functions we used:

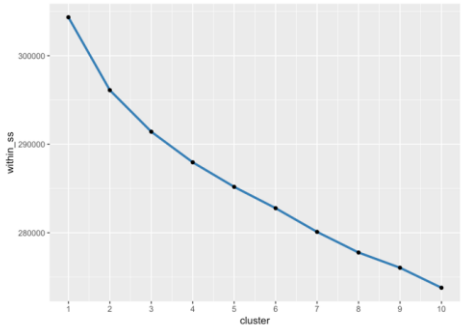
```
library(caret)
```

```
library(cluster)
library(flexclust)
```

Total Within Sum of Squares



Silhoette Width



Model	Kaggle RMSLE
Clustering then predict	2.23
No clustering	2.19

Model Experimentation - Deep learning

Process:

I tried two different library “nnet” and “keras”.

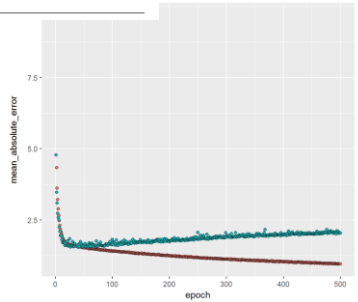
The library nnet contains a basic neural network function, nnet(), which fist single-hidden-layer neural network, possibly with skip-layer connections.

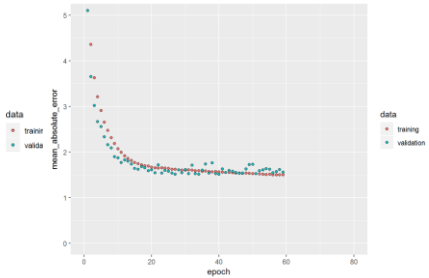
Keras: Normalization + a sequential model with two densely connected hidden layers, and an output layer that returns a single, continuous value.

```
## _________________________________________________________________
## Layer (type)                 Output Shape              Param #
## =================================================================
## dense (Dense)                (None, 64)                960
## _________________________________________________________________
## dense_1 (Dense)              (None, 64)                4160
## _________________________________________________________________
## dense_2 (Dense)              (None, 1)                 65
## =================================================================
## Total params: 5,185
## Trainable params: 5,185
## Non-trainable params: 0
## _________________________________________________________________
```

Packag

```
library(MASS)
library(nnet)
library(keras)
install_keras()
```





Learning

- Neural nets are not good models for sparse data.
- How to transform your data to make it something more neural net compatible.
- However, when we tried to scale the data or using log transformation, but due to the sparse variables from text analytics, there is still no improvements.
- Keras provides more hidder layers which gives better performance, and tuning with rmse.

Model Experimentation - SVM and RandomForest

Process:

SVM: the data points lie in between the two borders of the margin which is maximized under suitable conditions to avoid outlier inclusion;

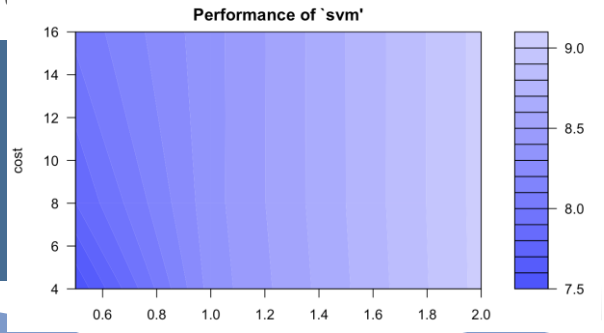
cost = 1000
gamma = 1000
tune.svm()

Random Forest: Random forests has two advantages.

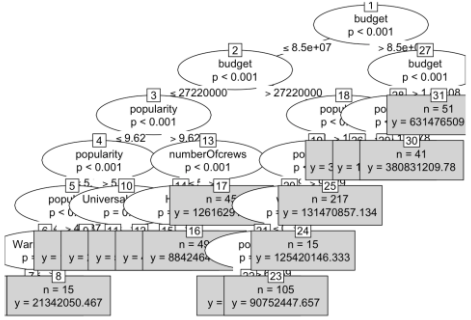
Firstly, reduction in over fitting: by averaging several trees
Secondly, less variance

Package and functions we

```
library(rpart)
library(svm)
library(RandomForest)
```



Plot for randomforest with library(party)

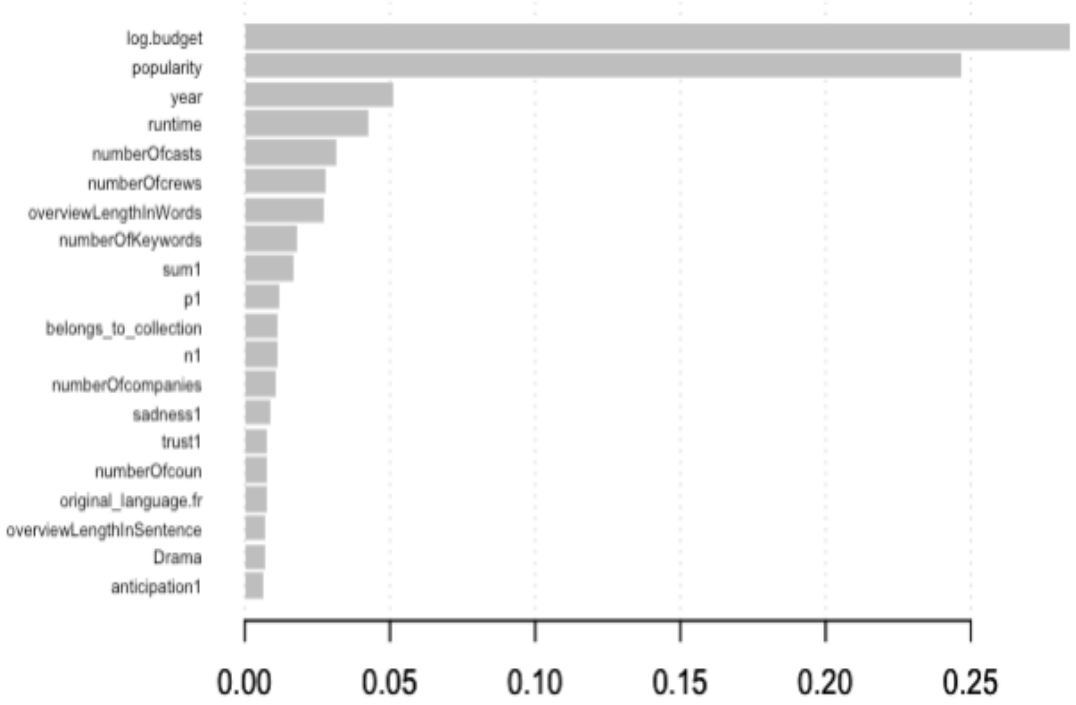


Final results:

Show 10 entries Search:

	Model name	Test score
1	XGboost	2.18
2	XGboost+Clustering	2.2
3	Keras	2.21
4	Nnet	2.33
5	SVM	2.49
6	Random Forest	2.83

The conclusion



We ran a variable importance plot based on the xgboost model as shown on the left to understand what factors influence movie revenue.

Based on the importance plot, we concluded that: the most important variables are: budget, number of casts and number of crews, popularity. Overview Length sentiment, proper runtime, and Movies belongs to a collection.

Suggestiones for producers

Movie producers to write detailed overview with proper amount of positive words to attract consumers.

Movie producers with low budget to wisely allocate their resources to marketing to increase popularity and exposure among consumers before release date.

Proper runtime and belonging to a certain collections are also suggested for producers.