

# A General Overview on Movie Data with a Look at Oscar Nominees, the Building of a Recommendation System and an Analysis on the Relationship Between Genre and Plot

Group name: MIC

Presentation by:

Sergio Gentilini  
Gianluca Caronte  
Michele Inchingolo  
Bill Mono

481089  
479367  
482748  
479379

# Content

- *Introduction*
- *Data*
- *Hypotheses and Methodology*
- *Exploratory Analysis*
- *Hypotheses Explanation*
- *Conclusion*

# Introduction

- *Movies are an important piece of culture in our society. The best films can make an unknown actor's career, while bad movies can have the opposite effect.*
- *With our analysis we are going to examine a movie dataset to learn about patterns through the construction of a recommendation system and an analysis on the relationship between genre and plot to be able to understand what makes a film a “good” one.*

# Data

- *Dataset 1 - Movie metadata*
- *Dataset 2 - Movie Plot Synopsis with Tags (MPST)*
- *Dataset 3 - MovieLens ratings*
- *Dataset 4 - The Oscar Award*

# Hypothesis and Methodology

- *Hypothesis A: "Movies belonging to the 'historical' genre have their mean vote influenced by the costume designer(s)."*
- *Hypothesis B: "There is a relationship between the genre of a movie and its plot such that it should be possible to predict the former from the latter."*
- *Metrics A: "Average votes"*
- *Metric B: "Model accuracy"*

# Hypothesis and Methodology

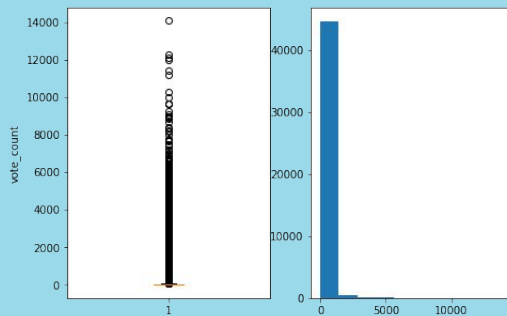
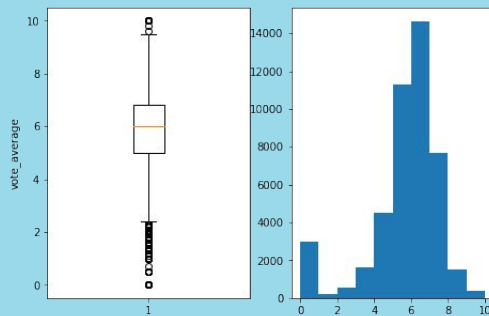
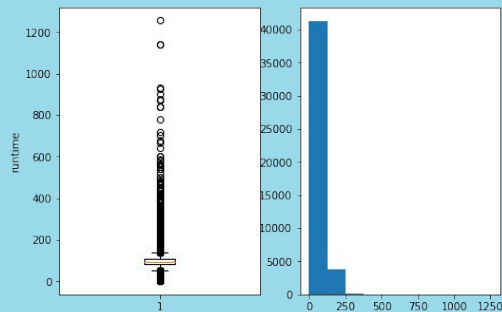
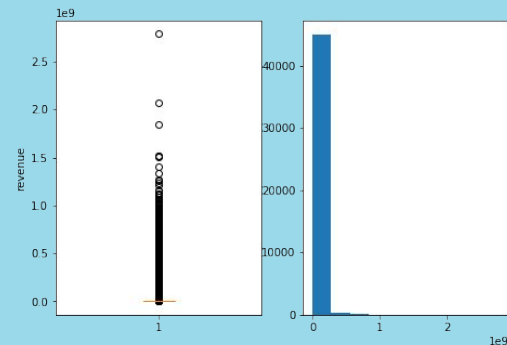
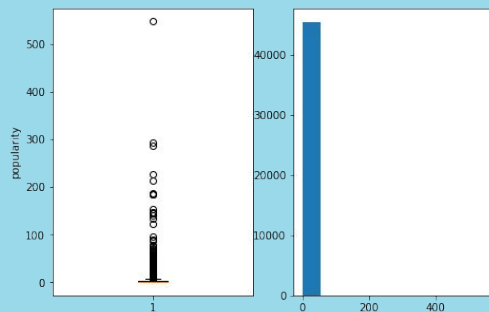
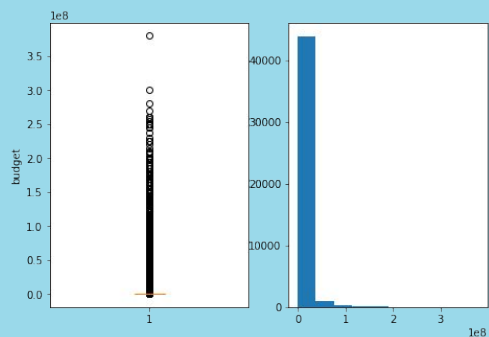
- *Hypothesis C: "Users that have shown similar tastes in the past will show similar tastes in the future."*
- *Metric C: "Users ratings"*
- *Hypothesis D: "An actor who has fulfilled more roles has participated in more films with high marks."*
- *Metric D: "Ratio between the number of actors with a good career and the total number of actors"*

# Hypothesis and Methodology

- *Hypothesis E: "The best movies have the best actors."*
- *Metric E: "Average votes of films in which an actor participated."*
- *Metric E.1: "Only movies released before the observed one."*
- *Metric F: "At least one actor who has won at least one Oscar."*

# Exploratory Analysis - Dataset Overview

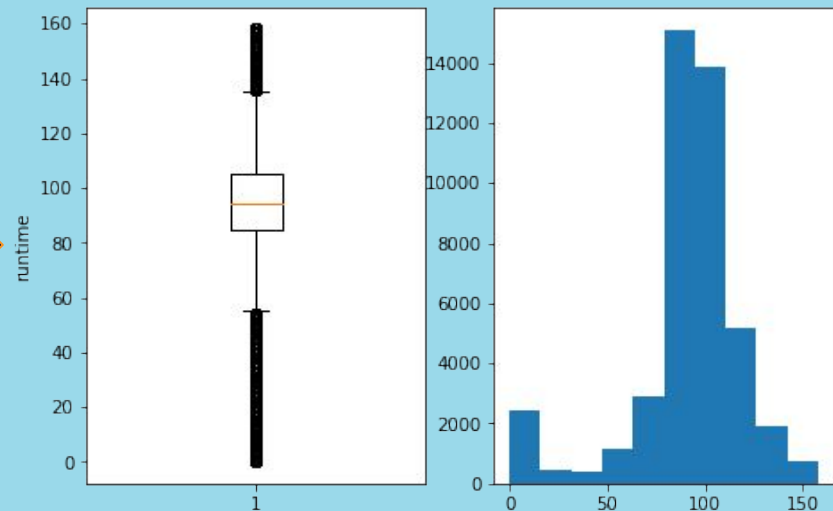
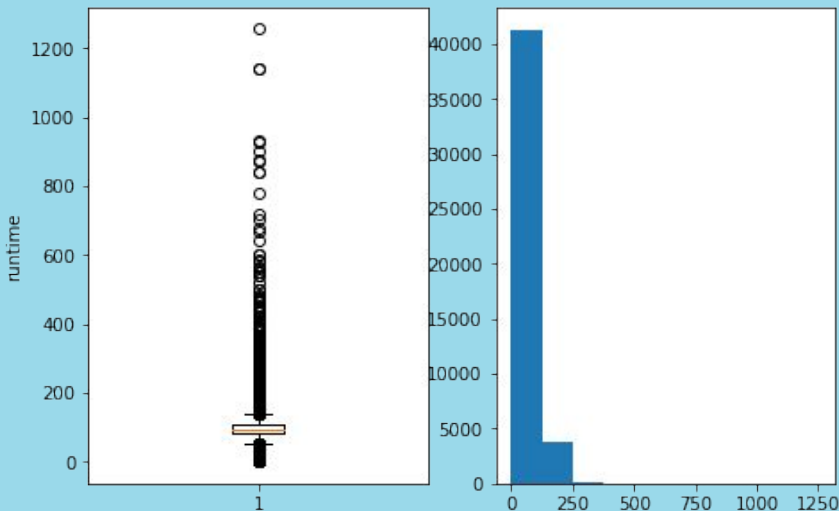
- *Movie metadata dataset.*
- Analyses through boxplots and histograms.





# Exploratory Analysis - Dataset Overview

- Data transformation: outliers removal.
- 2.5% on each side.
- It only makes sense in some cases.



# Exploratory Analysis - Dataset Overview

- Data transformation: null values substitution.
- All the null values were replaced with the mean.
- Next step: correlation.



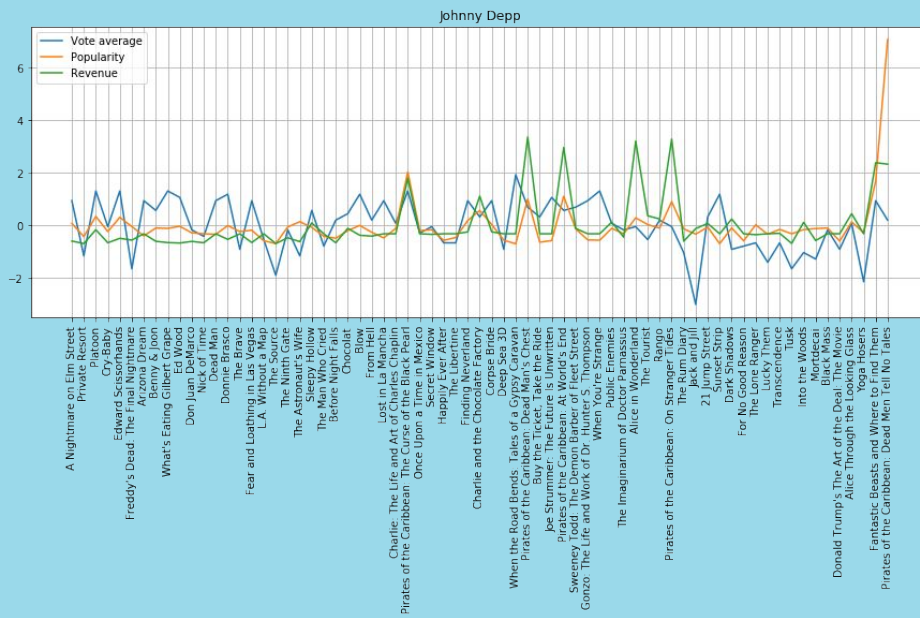
# Exploratory Analysis - Actor Analysis

- Visual representation of an actor's career.
- Movies (in chronological order) against *vote\_average*, *popularity* and *revenue*.
- Null values substitution through the median value over all the non-missing movies played by the actor.
- The correlation analysis depends on the actor.
- Correlation between revenue and popularity on the whole dataset.
- *Revenue* and *popularity* are correlated.

	vote_average	popularity	revenue
vote_average	1.000	0.154	0.084
popularity	0.154	1.000	0.506
revenue	0.084	0.506	1.000

# Exploratory Analysis - Actor Analysis

- Career analysis on Johnny Depp.
- Standardised data.
- High *revenue* often means high *popularity*.
- No visible trend.



# Exploratory Analysis - Production Companies

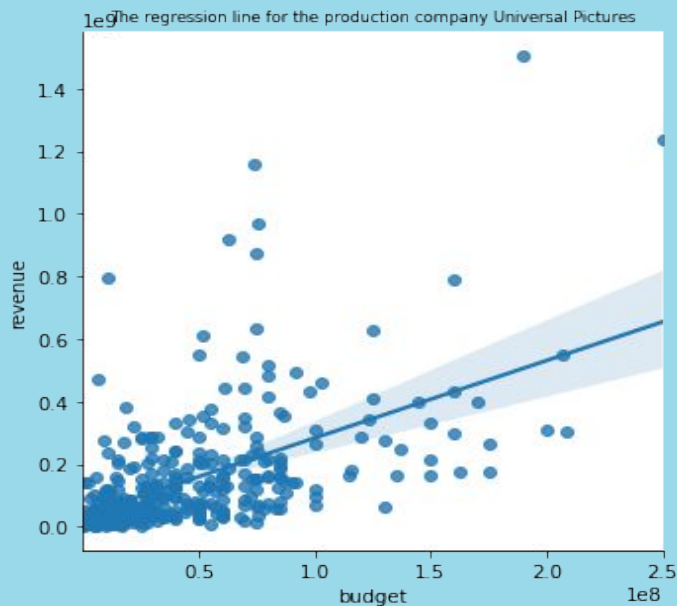
- Is there a relation between the budget spent on the film and the revenue?
- Is it a linear relationship?

# Exploratory Analysis - Production Companies

- Analysis on the production companies *Warner Bros* and *Universal Pictures*.
- Correlation depends on the production company.
- Lowest in *Universal Pictures* (0.55), highest in *Warner Bros* (0.71).

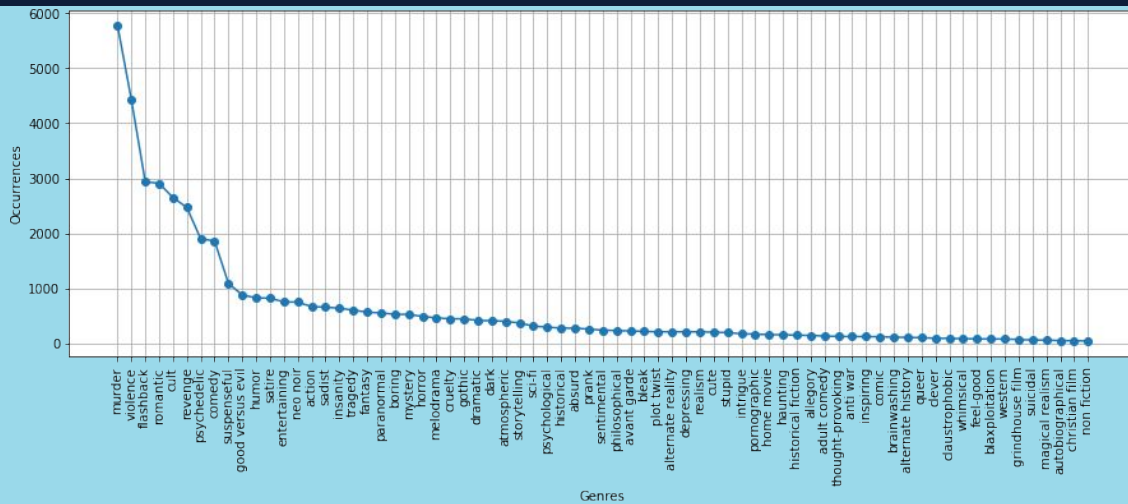
# Exploratory Analysis - Production Companies

- Linear relationship.
- Not enough data.



# Exploratory Analysis - Plots and Genres/Tags

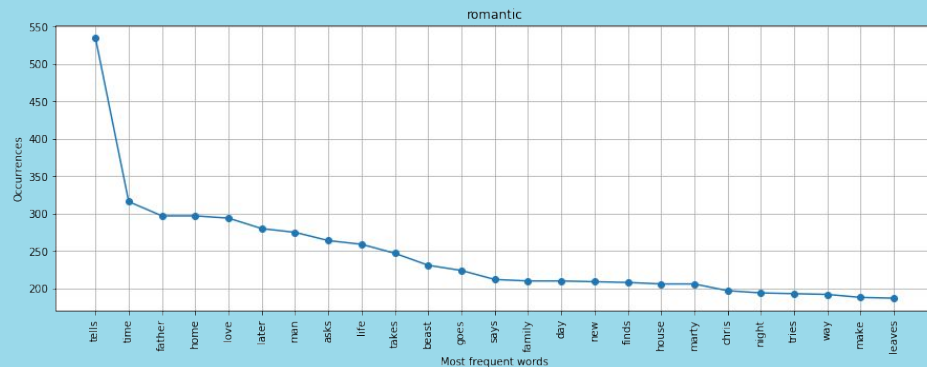
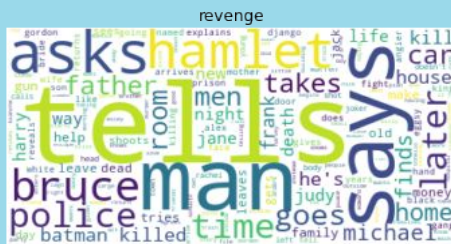
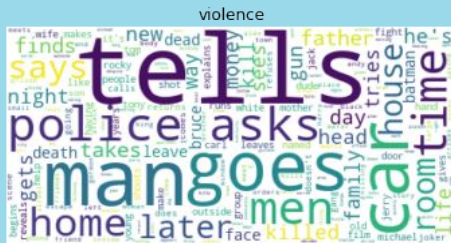
- Subdivision of movies by *genre* (or *tag*) and analysis on the most-occurring words in the plots of well-received films.
- Preliminary analysis: the most frequent tags.
- The vast majority of tags occur very rarely, less than 1000 times each.





# Exploratory Analysis - Plots and Genres/Tags

- Subdivision of movies by *genre* (or *tag*) and analysis on the most-occurring words in the plots of well-received films.
- Only “good” movies (average vote  $\geq 7.5$ ).
- Use of *stopwords*.
- Certain genres are very similar; *romantic* is rather generic

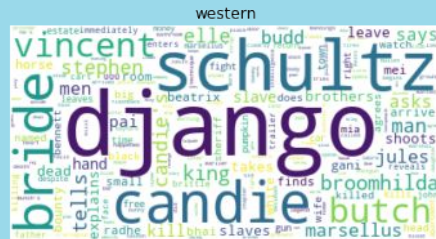
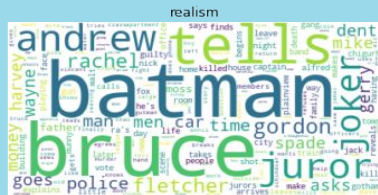
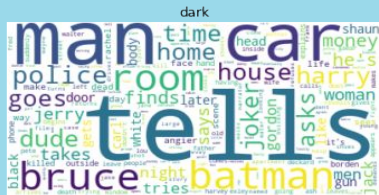
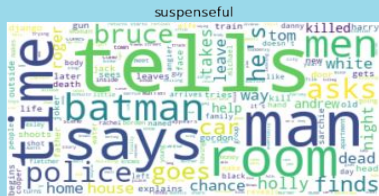
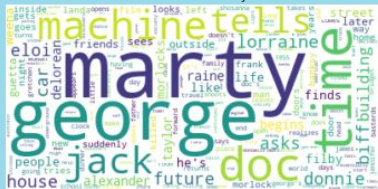


- *Comedy*: Marty Feldman?
- Presence of niche tags.
- Certain tags have been underutilised.

comedy



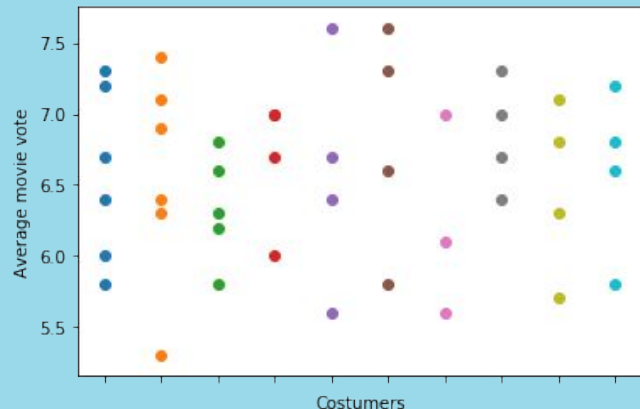
alternate history



Western has been underutilised: the most relevant movie seems to be *Django*; no mention of well-known spaghetti-westerns.

# Hypothesis A - Historical Movies

- Costumes are more important in *historical* films than in other kinds of movies.
- Historical movies with “good” costumer should have a better average vote.
- First analysis: the most prolific costumers.



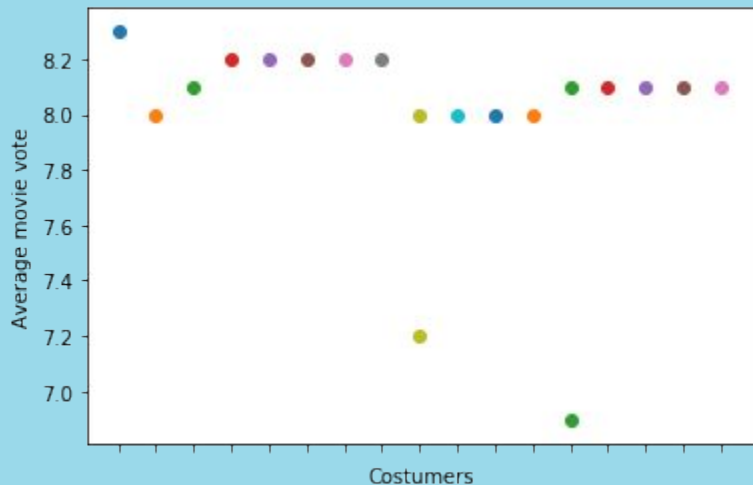
Good movie:  $\geq 7$   
Bad movie:  $< 6$

No visible relationship

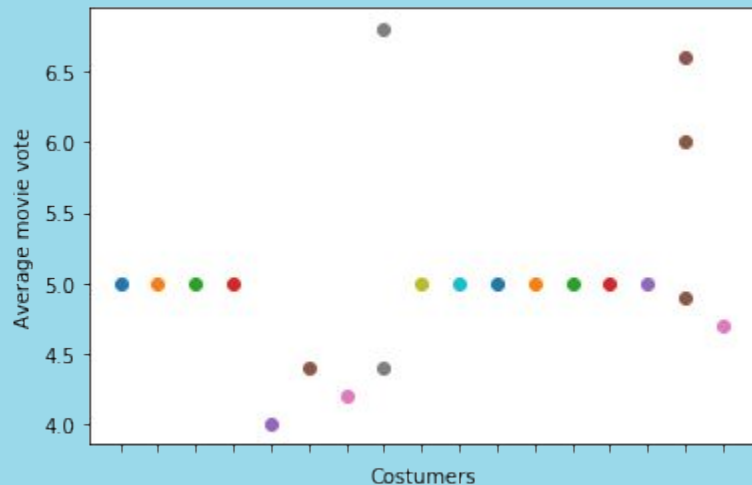
# Hypothesis A - Historical Movies

- Second analysis: a look at the 20 best and worst historical movies.
- No visible relationship. Most costumers have only worked on a single movie.

20 best historical movies



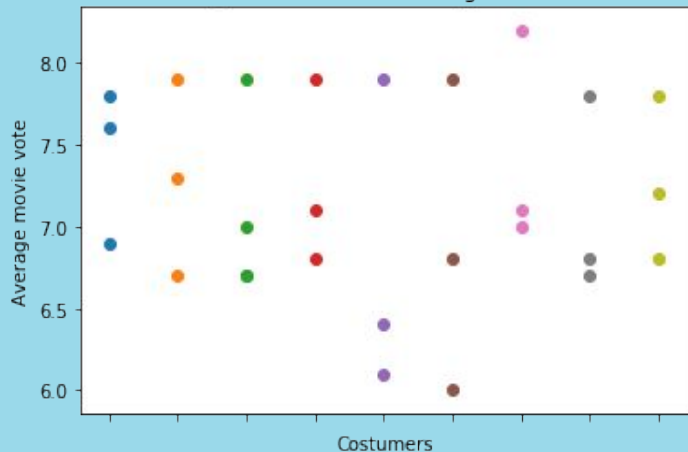
20 worst historical movies



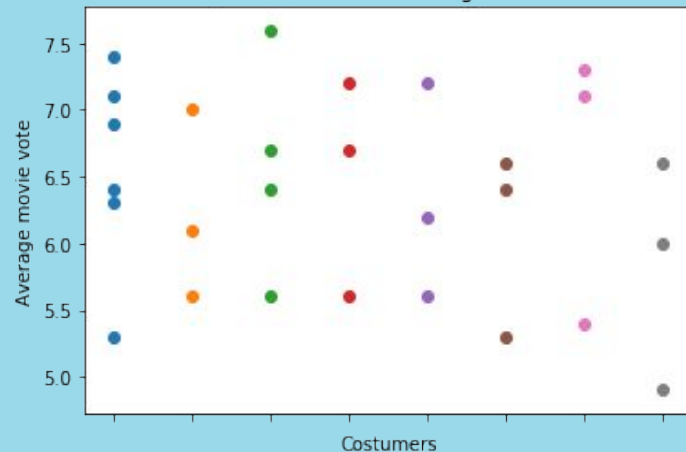
# Hypothesis A - Historical Movies

- Third analysis: a look at some of the best and worst historical movies by only considering customer who have worked on at least three films.
- The hypothesis is acceptable for good movies, not so much for bad ones.

Costumers who have worked on at least 3 historical movies, one of which is among the best films



Costumers who have worked on at least 3 historical movies, one of which is among the worst films

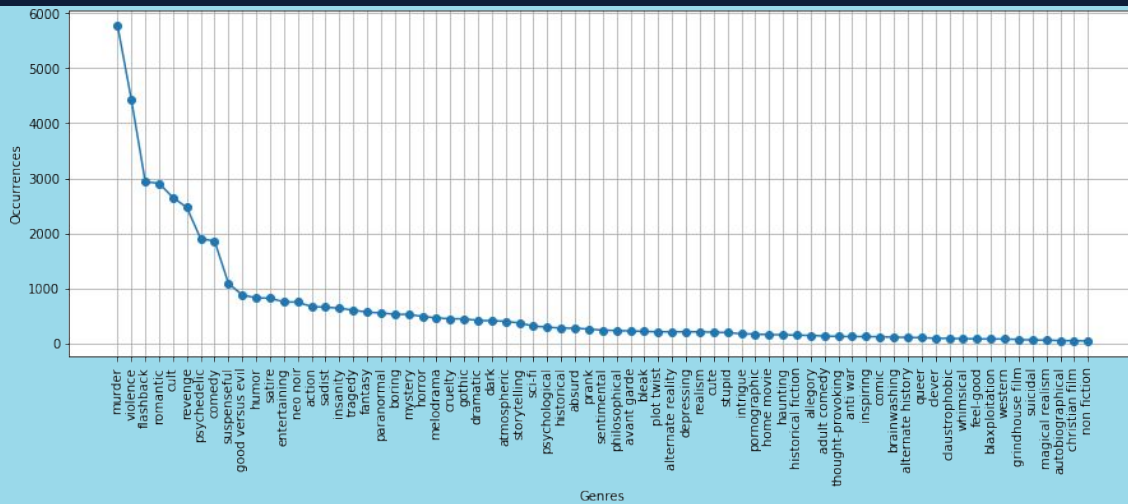


# Hypothesis A - Historical Movies

- Conclusion: no relationship between the reception of a historical movie and the involved costumer.
- The costume crew is not enough even when considering this niche.
- More data could lead to further considerations.

# Hypothesis B - Predicting Genres from Plots

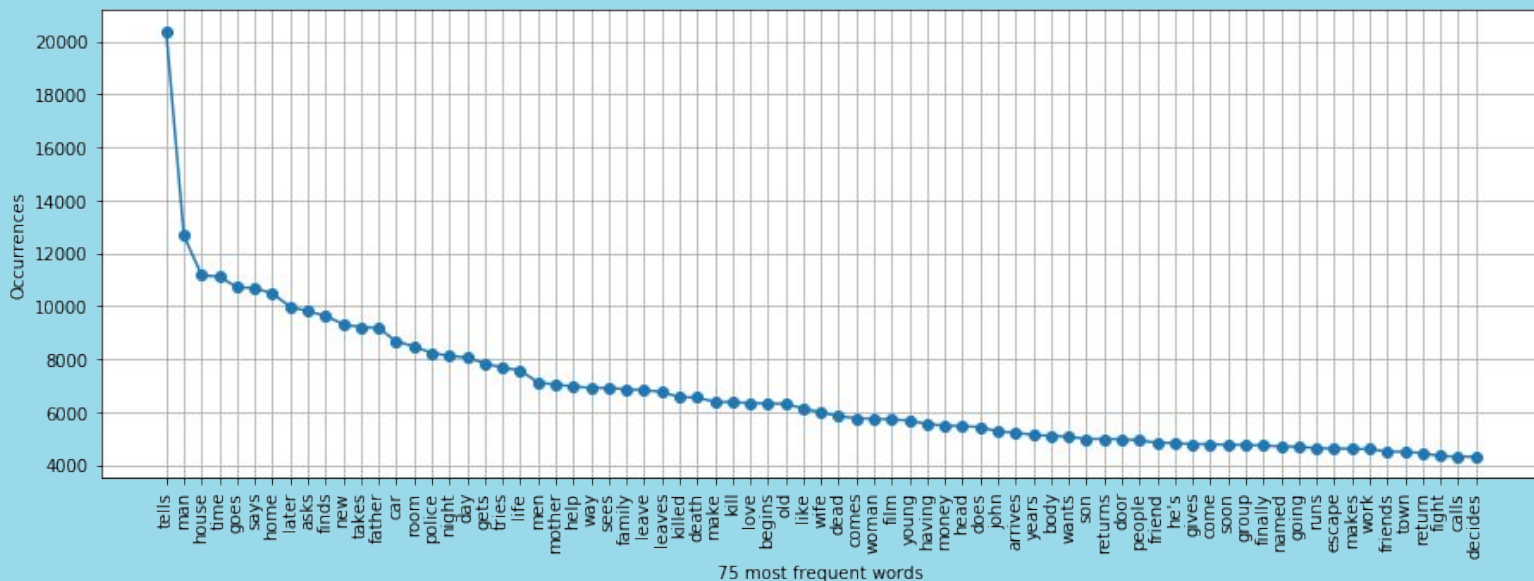
- Movies belonging to the same genre are similar, so it should be possible to perform a prediction through their plots.
- *MPST dataset*, ~12500 movies.
- 71 tags that may act as genres or classes.
- Many of them appear rarely.



# Hypothesis B - Predicting Genres from Plots

## General Case

- All the dataset has been considered.
- In case of multiple tags, only the first one is counted.
- *Stopwords* removal and exploratory analysis.





# Hypothesis B - Predicting Genres from Plots

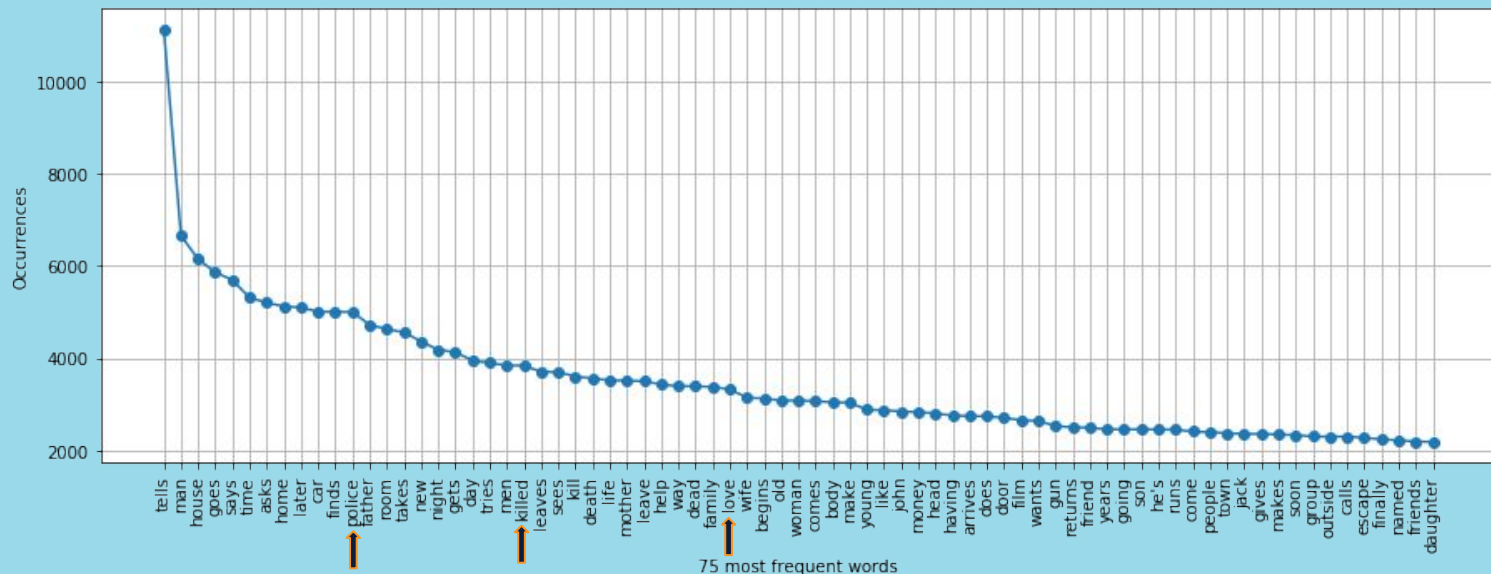
## General Case

- Features via *bag-of-words*; dictionary of 10000 terms.
- *Classifier*: naive Bayes.
- *Accuracy*: 19.1%.
- Many genres are misclassified because of overspecificity or lack of tag data.
- Conclusion: too many vague, unusable tags that only label a small subgroup of the dataset lead to a poor performance.

# Hypothesis B - Predicting Genres from Plots

## Binary Case

- Classification based on two abundant genres that are easy to tell apart: *murder* and *romantic* (~7000 movies).
- *Stopwords* removal and exploratory analysis.



# Hypothesis B - Predicting Genres from Plots

## Binary Case

- Features via *bag-of-words*; dictionary of 1000 terms.
- *Classifiers*: naive Bayes, logistic regression, SVM.

	Naive Bayes	Logistic Regression	Support Vector Machine
Accuracy	83.0%	85.9%	87.4%

- *Naive Bayes*: quick and rough result; it works well with *bag-of-words* features.
- *Logistic regression*: standard for binary classification.
- *SVM*: improve the model through non-linear boundaries.

# Hypothesis B - Predicting Genres from Plots

## Binary Case

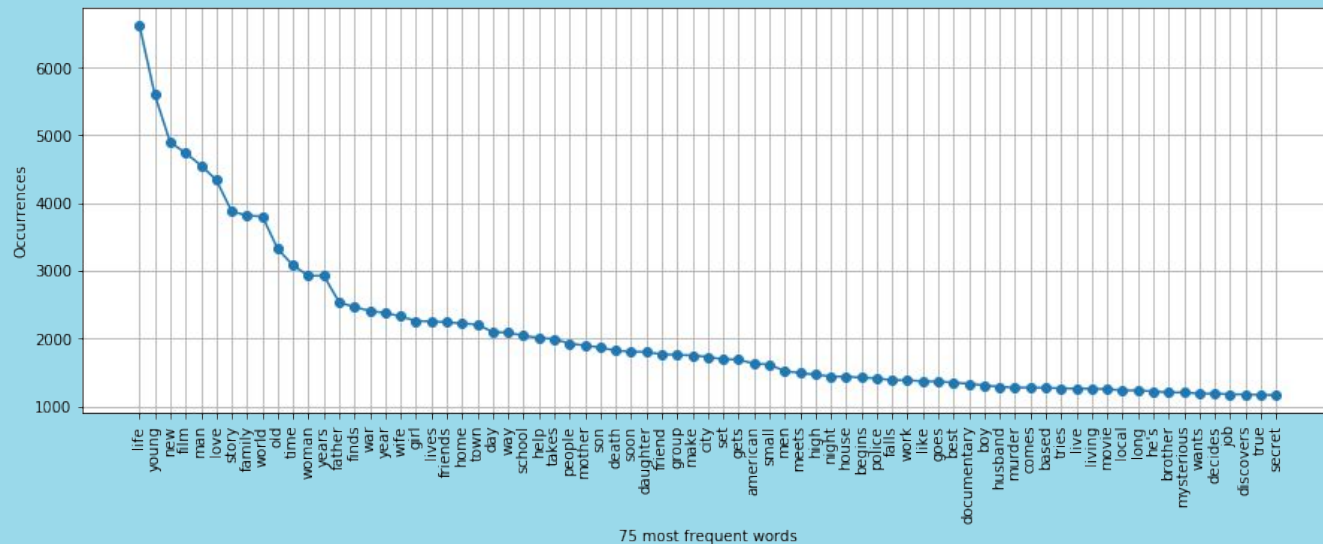
	Naive Bayes	Logistic Regression	Support vector machine
Accuracy	83.0%	85.9%	87.4%

- *Naive Bayes*: fast, good accuracy and performance despite class imbalance.
- *Logistic regression*: higher accuracy and fewer false-negatives, but more *romantic* movies are mistakenly classified as *murder* movies.
- *SVM*: highest accuracy, better performance against imbalance than logistic regression.
- Conclusion: working on genres that are clearly identifiable leads to valid results.

# Hypothesis B - Predicting Genres from Plots

## Genre-Overview Case

- Analysis on the *movie metadata* dataset.
- 32 *genres* and shorter plot descriptions (*overviews*).
- Given the change in dataset, the post-stopwords word count is visibly different.



# Hypothesis B - Predicting Genres from Plots

## Genre-Overview Case

- Features via *bag-of-words*; dictionary of 1000 terms.
- *Classifiers*: naive Bayes, logistic regression, SVM.

	Naive Bayes	Logistic Regression	Support Vector Machine
Accuracy	42.2%	42.4%	43.8%

- ~38500 observations for training, ~4300 for testing.
- Logistic regression and SVM via *one versus rest*.
- The performances are similar, therefore only the naive Bayes case is analysed, as it's simpler and easy to train.

# Hypothesis B - Predicting Genres from Plots

## Genre-Overview Case

	Naive Bayes	Logistic Regression	Support Vector Machine
Accuracy	42.2%	42.4%	43.8%

- The genres with the most observations in the dataset are *drama*, *action* and *comedy*.
- The genres that are predicted the most accurately are documentary (71.6%), drama (53.8%) and comedy (51.3%).
- Even with a larger quantity of data, the results are underwhelming.
- Conclusion: Good classification requires few genres that are clearly separated. If these criteria are not met, the results will be unsatisfying.

# Hypothesis D - The Number of Roles Determines an Actor's Career

- Goal: We want to see if there is a relationship between the number of roles a person has fulfilled in their entire career, such as *actor* or *director*, and the ability to participate in more films that have obtained high marks from reviewers, hence having a better career.

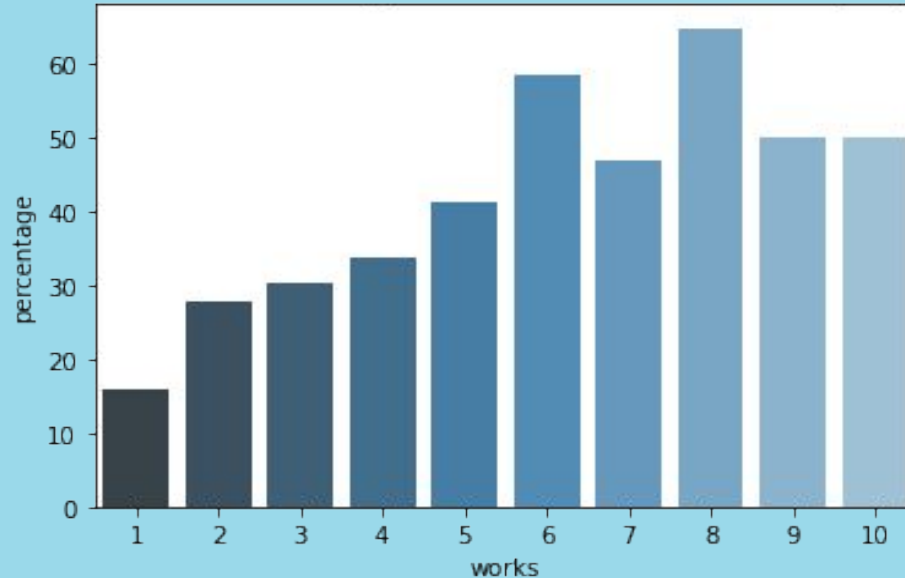


# Hypothesis D - The Number of Roles Determines an Actor's Career

- People were separated according to the number of roles - starting from *actor* - played in their entire career
- From this subdivision, 10 categories were created.
- For each person, the total number of movies and highly-rated films they had participated in were calculated.
- A ratio was made between the number of people belonging to a category and the people (still in that category) who had worked in movies with a rating higher than 7.5 for at least 10% of their career.

# Hypothesis D - The Number of Roles Determines an Actor's Career

Percentage that an actor has good career due to the number of jobs performed



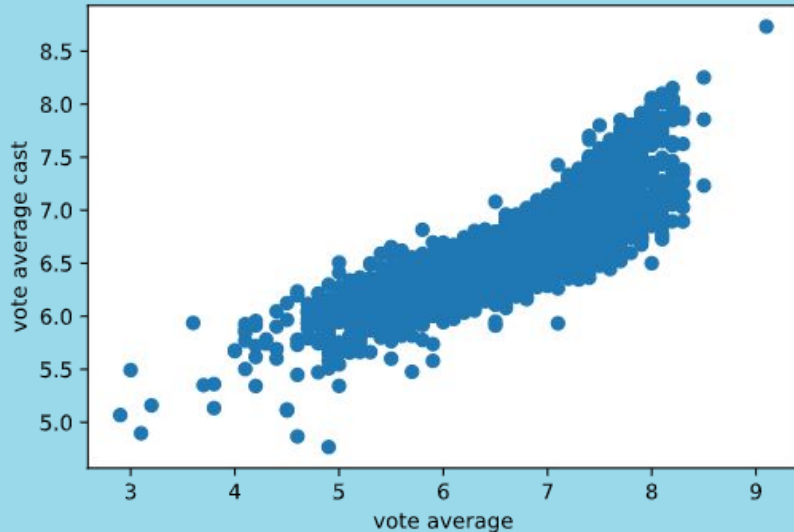
# Hypothesis D - The Number of Roles Determines an Actor's Career

- Conclusion: Actors who have fulfilled multiple roles in their career are more likely to have a good career, rather than actors who have only had one or few positions.

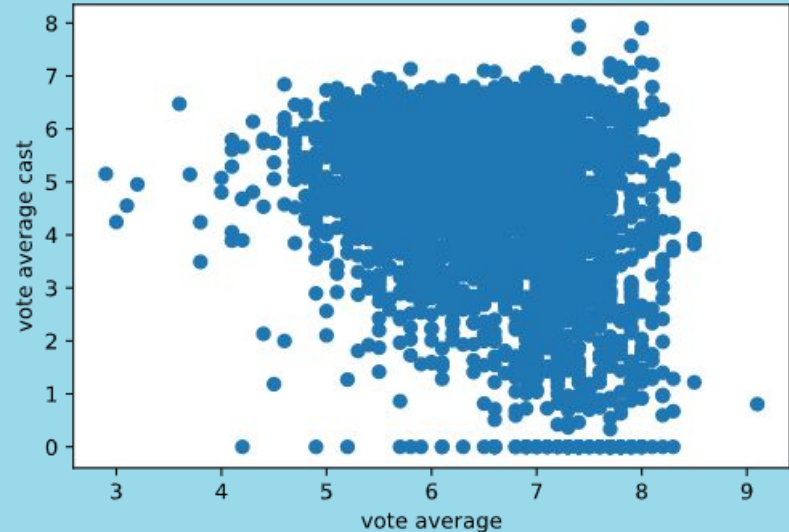
# Hypothesis E - Movies and Actors

- Goal: Prove that better actors correspond to better movies.
- Problem: Metric definition for the adjective “better”, both for movies and actors.
- Solution:
  - Movies have a score from the dataset.
  - The actors' skill score is calculated on the movies' vote average.

M1. Actors talent based on average vote of all film



M2. Actors talent based on average vote of previous films



# Hypothesis E - Movies and Actors

## Oscar Collection Integration

- Goal: Discover if Oscar information could help a movie or a cast to have a higher score, i.e. to be better for our metrics.
- Problem: How to integrate Oscar information with the previous dataset.
- Solution: Add to each movie in the dataset a binary variable to indicate whether at least one Oscar-winning actor is present in the movie cast or not.
  - 1) We consider Oscars won in the whole career.
  - 2) We consider Oscars won in the past career only.

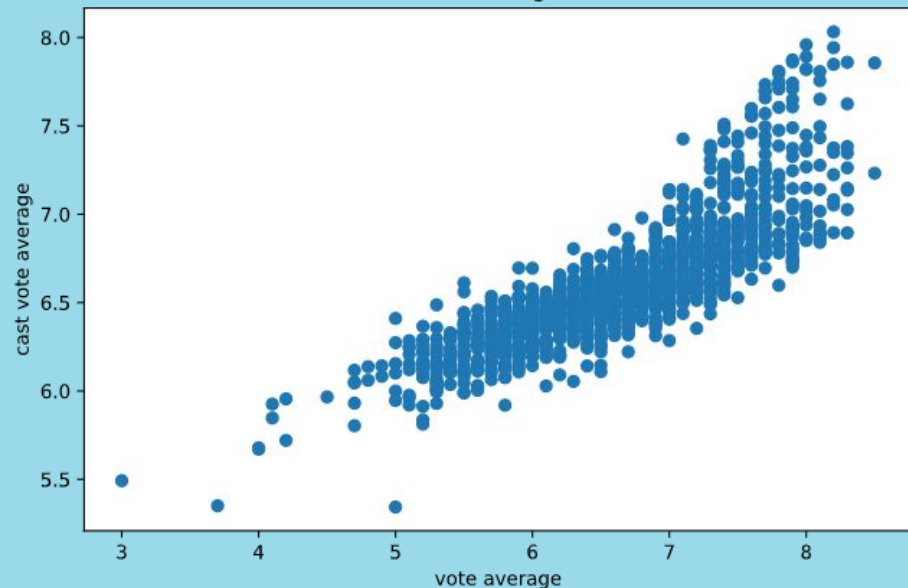
# Hypothesis E - Movies and Actors

## Oscar Collection Integration

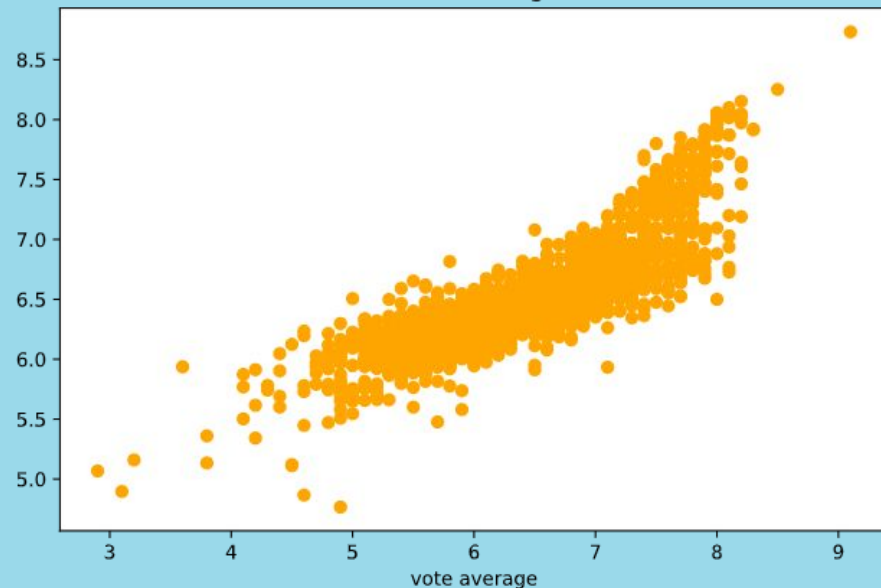
Oscar Metric 1: Oscars won in the whole career.

Relation between movie vote and cast vote

Oscar-winning actor



Not oscar-winning actor



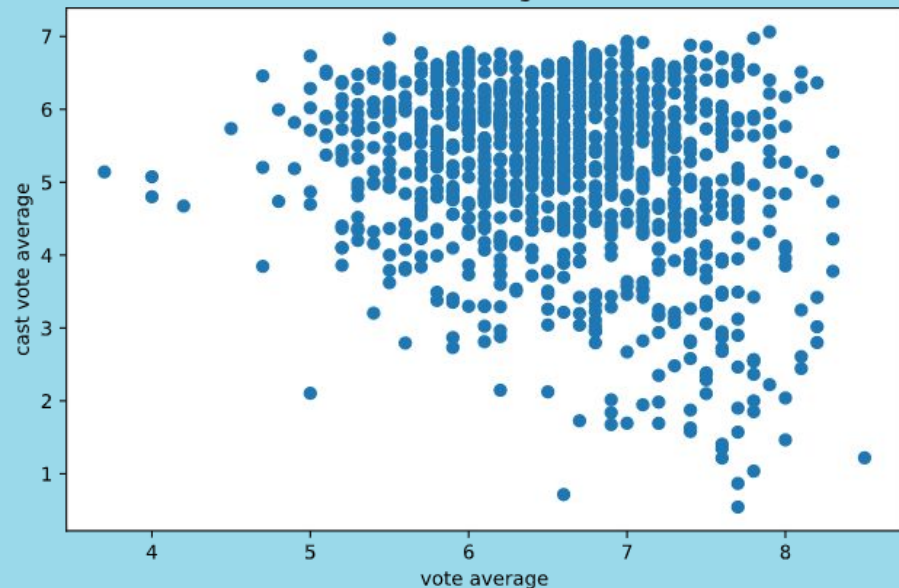
# Hypothesis E - Movies and Actors

## Oscar Collection Integration

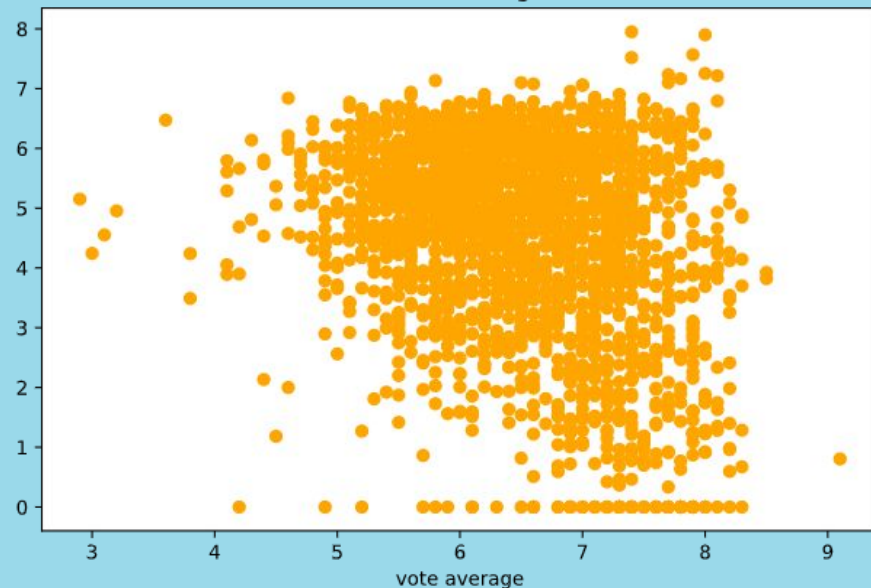
Oscar Metric 2: Oscars won in the past career only.

Relation between movie vote and cast vote  
(considering past career)

Oscar-winning actor



Not oscar-winning actor

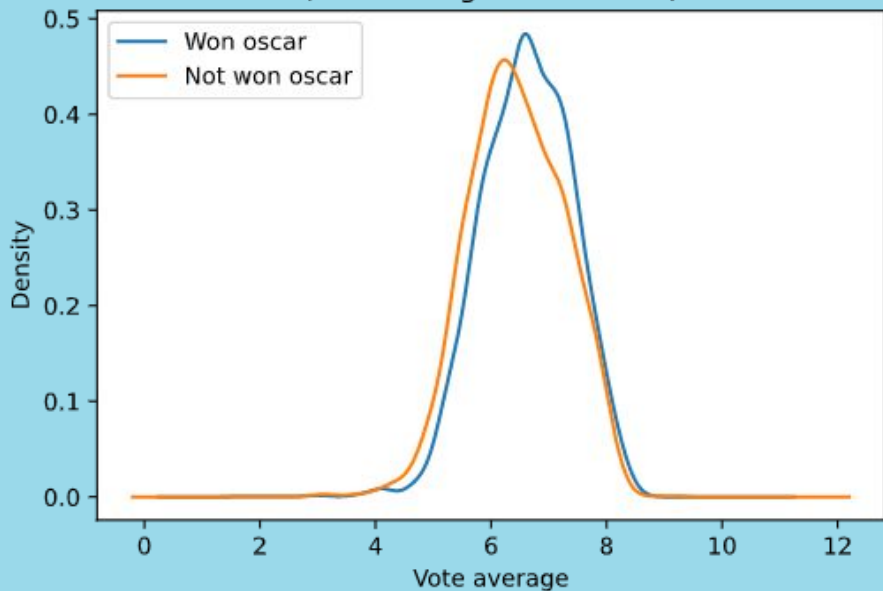


# Hypothesis E - Movies and Actors

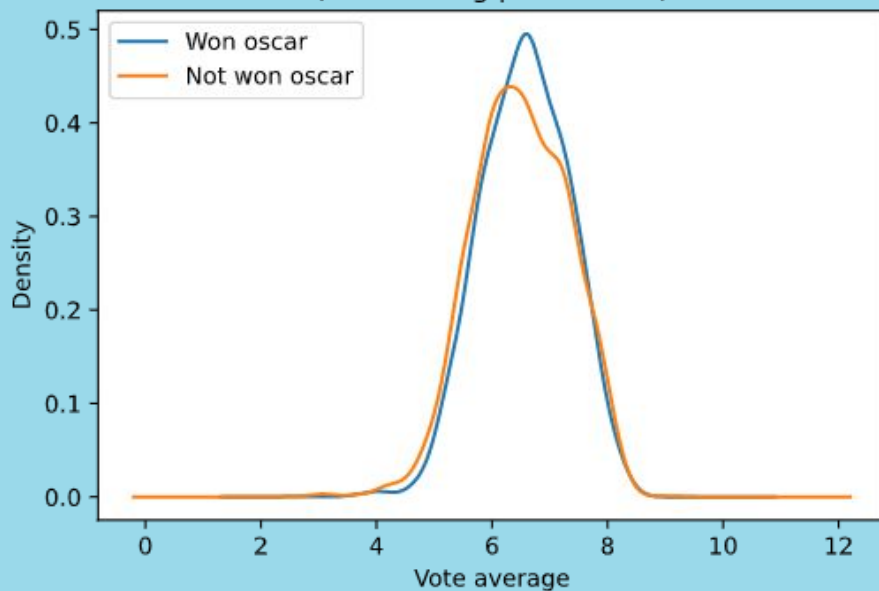
## Oscar Collection Integration

Movie vote density function.

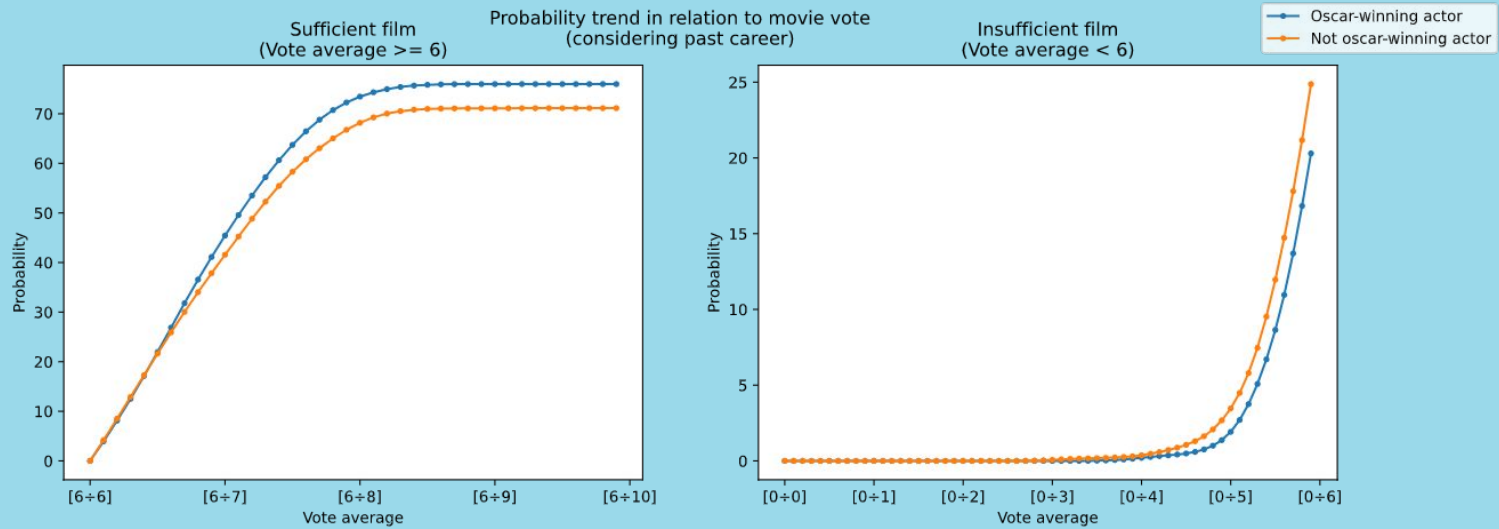
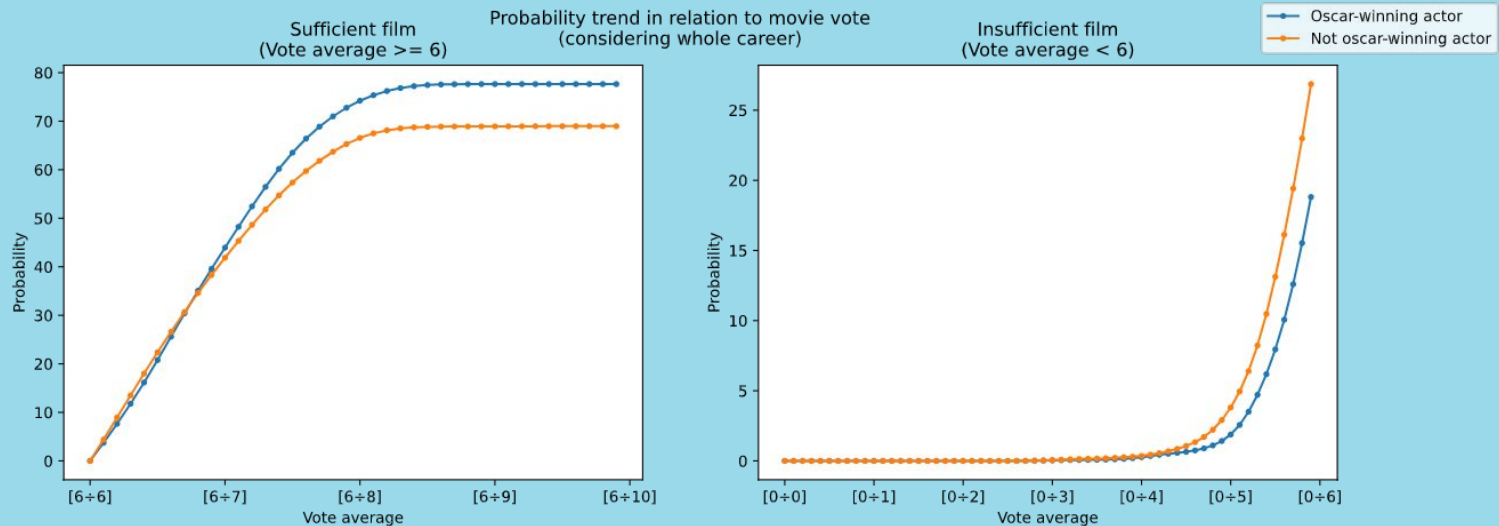
Probability density function  
(considering whole career)



Probability density function  
(considering past career)





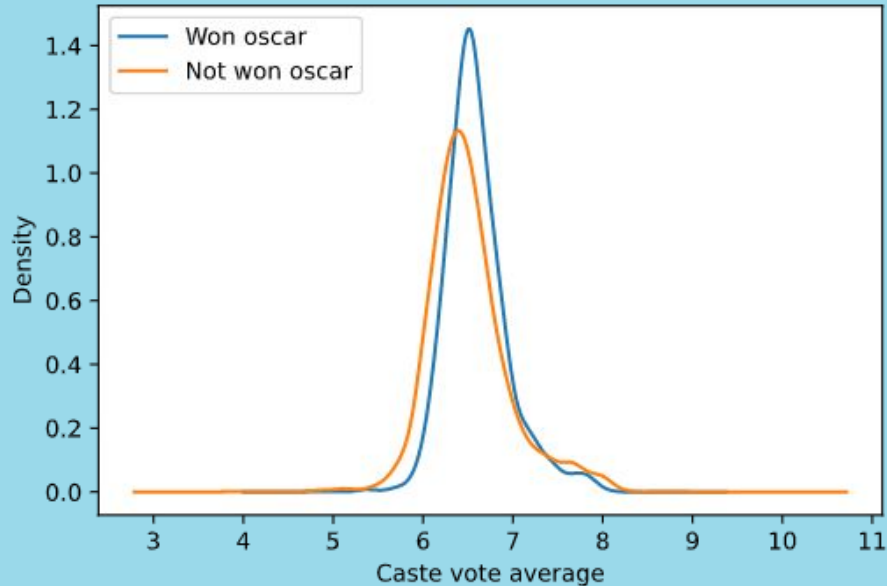


# Hypothesis E - Movies and Actors

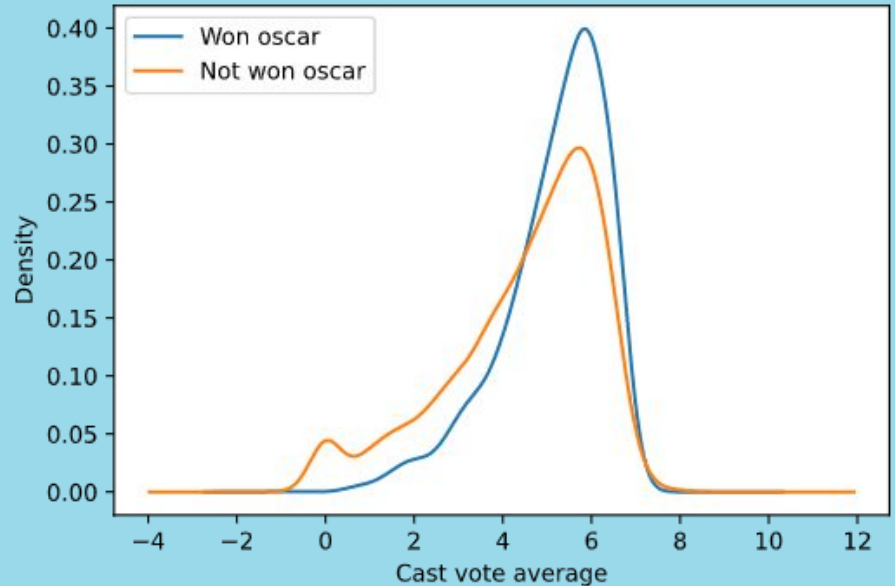
## Oscar Collection Integration

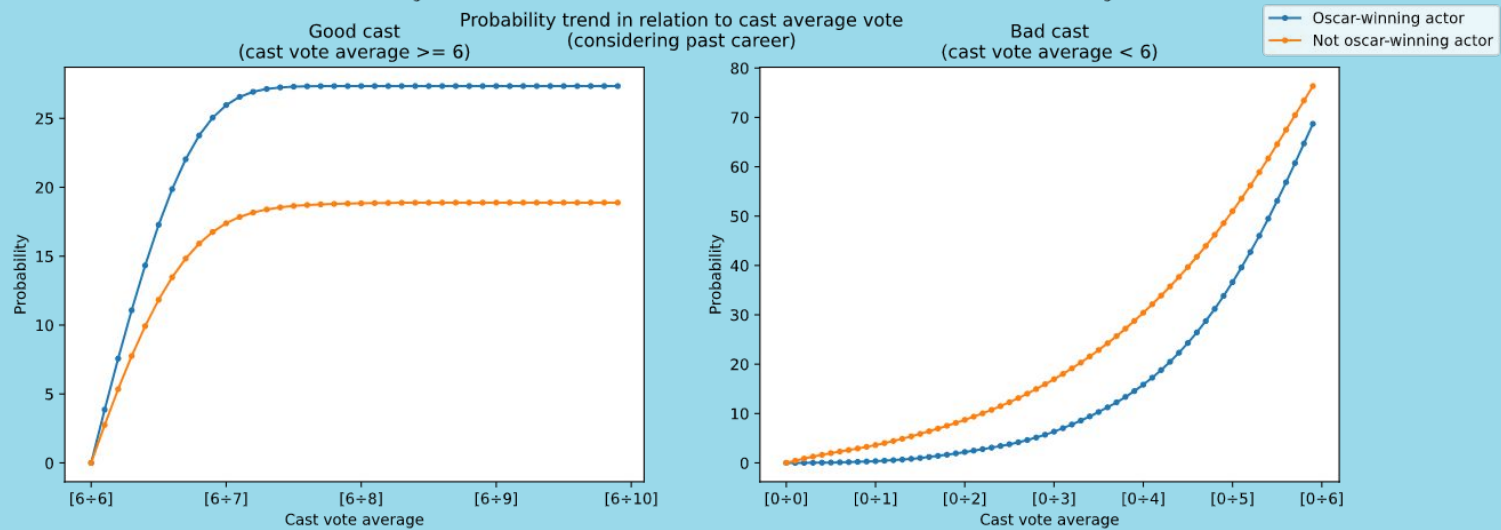
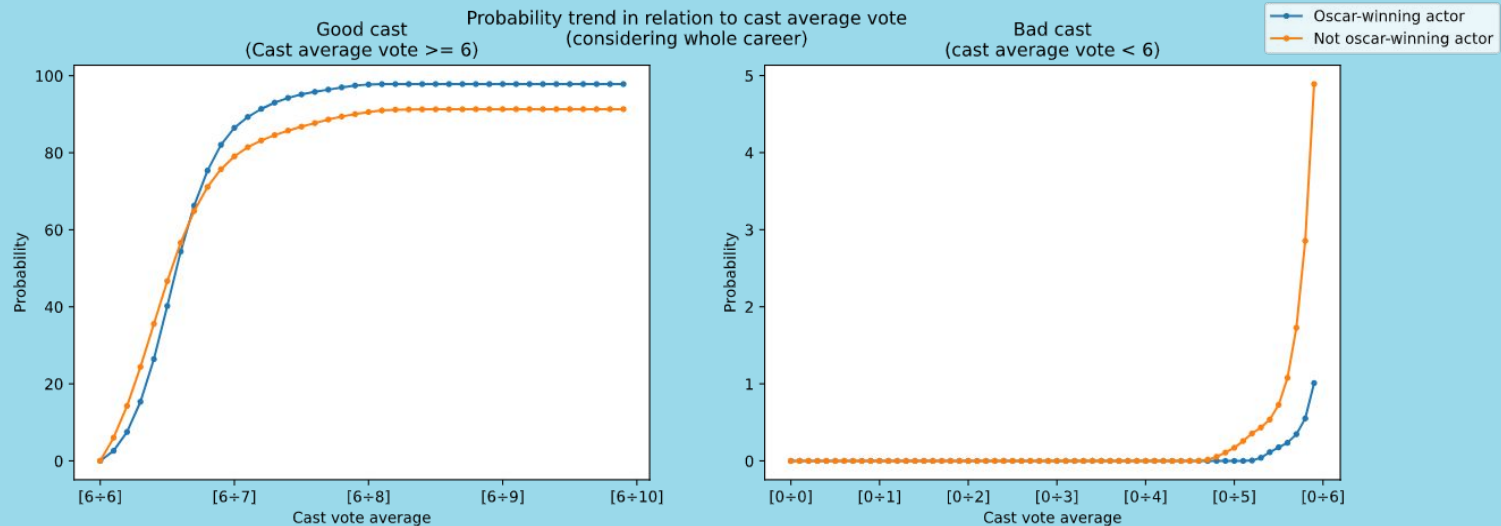
Cast vote density function.

Probability density function  
(considering whole career)

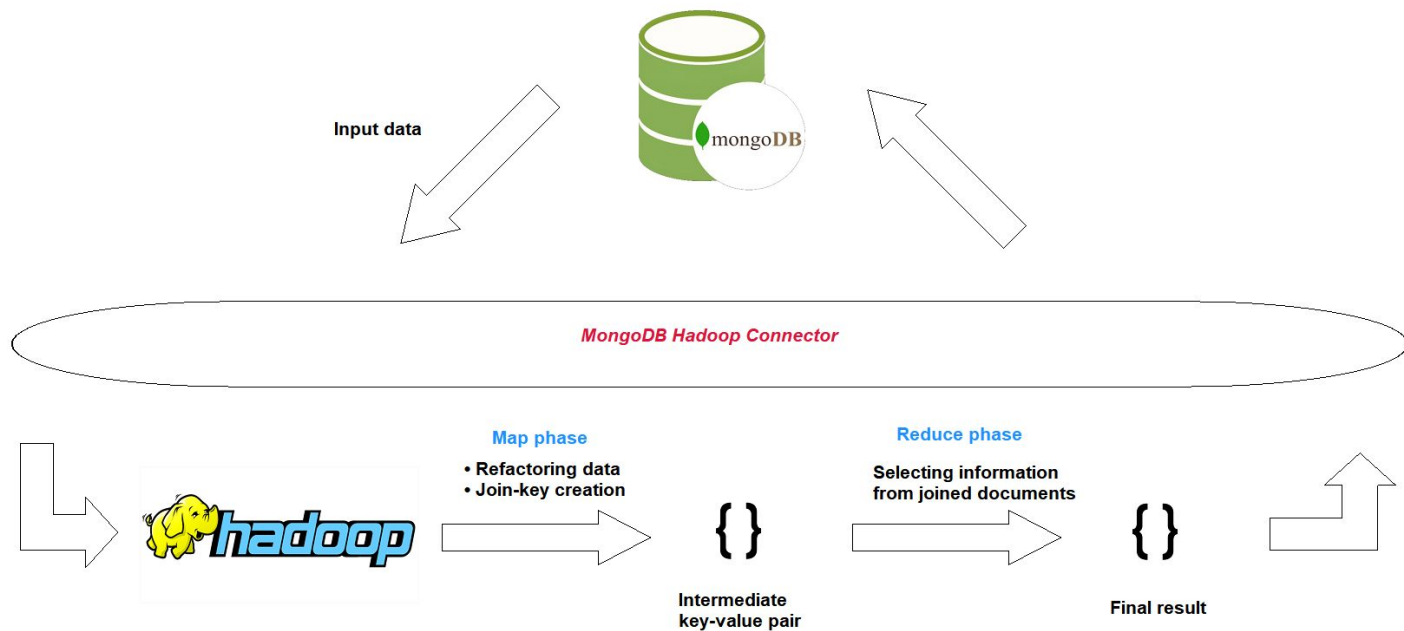


Probability density function  
(considering past career)

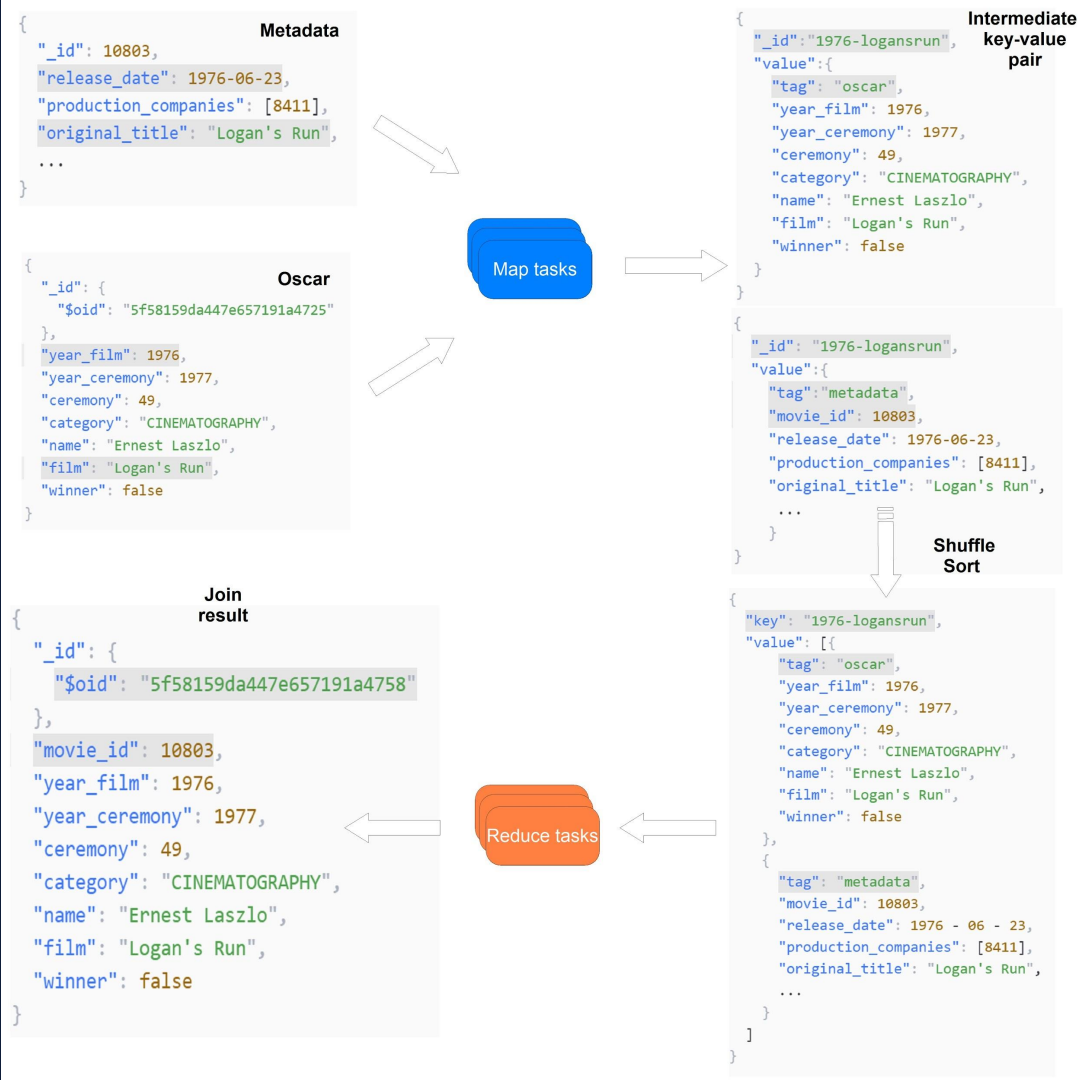




# MapReduce



# MapReduce Join Implementation



# Hypothesis C: Collaborative Filtering

Exploiting common user behaviour (collaborative) to predict the future ones (filtering).



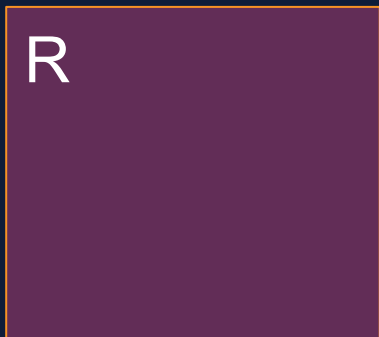
```

pipeline = [
  { "$set": {
    "rating": { "$switch": {
      "branches": [
        {
          "case": { "$lt": [ "$rating", 2.5 ] },
          "then": -1
        },
        {
          "case": { "$gte": [ "$rating", 2.5 ] },
          "then": 1
        },
      ],
    }
  } },
  { "$group": {
    "_id": "$userId",
    "ratings": {
      "$push": {
        "k": { "$toString": "$movieId" }, "v": "$rating"
      }
    }
  } },
  { "$project": { "_id": "$_id", "ratings": { "$arrayToObject": "$ratings" } } },
  { "$set": {
    "ratings._id": "$_id"
  } },
  { "$replaceRoot": { "newRoot": "$ratings" } },
  { "$out": "reshaped_ratings" }
]

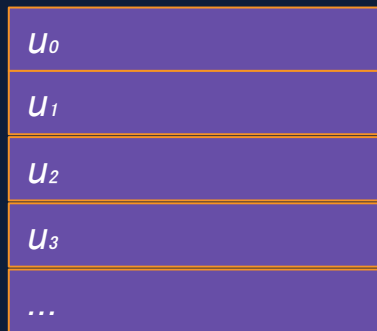
```

# Building the User/Ratings Table on MongoDB

# Slope One



=



$$f(u, v) = \frac{\langle u - \bar{u}, v - \bar{v} \rangle}{\sum_{i \in S(u) \cap S(v)} (u_i - \bar{u})^2 \sum_{i \in S(u) \cap S(v)} (v_i - \bar{v})^2}$$

Where  $S(u)$  is the set of ratings of user  $u$ .

If  $u$  and  $v$  haven't both rated at least one item,  $f(u, v)$  can't be computed.

1	$f(u_0, u_1)$	$f(u_0, u_2)$	$f(u_0, u_3)$
	1	$f(u_1, u_2)$	$f(u_1, u_3)$
		1	$f(u_2, u_3)$
			1



# Slope One

To predict the rating of movie  $i$  for user  $u$ :

1. Find the set of users ( $V$ ) that have rated movie  $i$  and have a correlation value with user  $u$ .
2. For each user  $v$ , take the movie rate minus the user average rating and multiply it by the correlation value of  $v$  with  $u$ .
3. Sum the result by every  $v$ , then normalize by the sum of the absolute value of the correlation values.
4. Add the  $u$  rate average.

$$R(u)_i = \bar{u} + \frac{\sum_{v \in V} f(u,v)(v_i - \bar{v})}{\sum_{v \in V} |f(u,v)|}$$

# Alternating Least Squares



$$R \in \mathbb{R}^{u \times m}$$

$$U \in \mathbb{R}^{u \times l}$$

$$M \in \mathbb{R}^{l \times m}$$

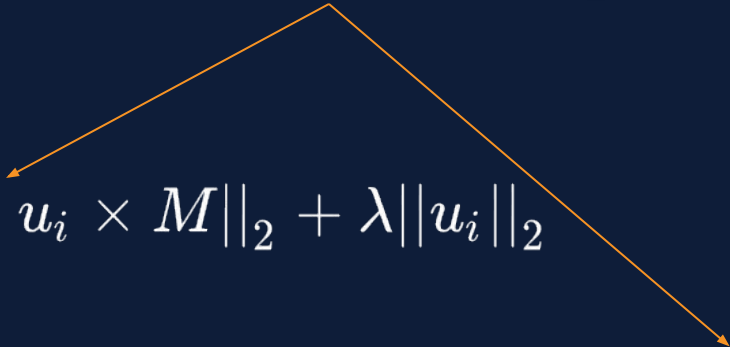
Hyper-parameters:

- $l$ : the latent dimension size
- $\lambda$ : the regularization factor

*The problem*

# Alternating Least Squares

$$J = ||R - U \times M|| + \lambda(||U||_2 + ||M||_2)$$


$$\forall u_i : J(u_i) = ||R_i - u_i \times M||_2 + \lambda ||u_i||_2$$

$$\forall m_j : J(m_j) = ||R_j - U \times m_j||_2 + \lambda ||m_j||_2$$

*This requires a dense matrix to find a solution.*

Handling missing  
values

# Alternating Least Squares

$$J = ||R - U \times M|| + \lambda(||U||_2 + ||M||_2)$$

$$W_{ij} = \begin{cases} 0 & \text{if } R_{ij} \text{ is unknown} \\ 1 & \text{otherwise} \end{cases}$$

$$J = \sum_{i,j} W_{ij} (R_{ij} - u_i \times m_j)^2 + \lambda(||U||_2 + ||M||_2)$$

*The missing values are not considered in the optimization problem.*

*The solution*

# Alternating Least Squares

$$m_j = (U^T \times (w_j \times I) \times U + \lambda I)^{-1} \times (w_j \times I) \times r_j$$

$$u_i = (M \times (w_i \times I) \times M^T + \lambda I)^{-1} \times (w_i \times I) \times r_i$$

*The rows of  $U$  (like the columns of  $M$ ) are independent from each other.*

# Alternating Least Squares

1. Initialize  $M$  and  $U$  with random values or a heuristic method.
2. Copy  $M$  in each machine, split  $U$  rows between them.
3. Assuming  $M$  constant, approximate a new value of  $U$ .
4. Merge  $U$  rows.
5. Copy  $U$  in each machine, split  $M$  columns between them.
6. Assuming  $U$  constant, approximate a new value of  $M$ .
7. Merge  $M$  columns.
8. Repeat from point 2, until the stop criterion is satisfied.

# Alternating Least Squares

The prediction for user  $i$  of movie  $j$  is given by the  $i,j$  element of the product of  $U$  and  $M$ .

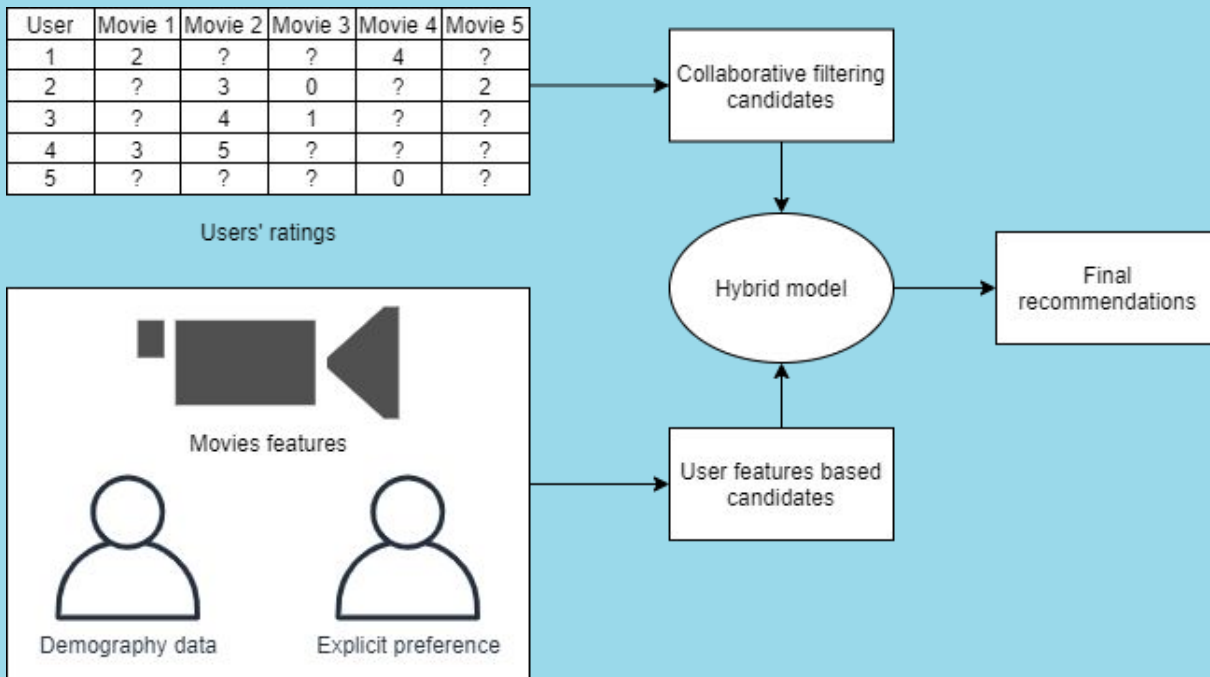
It can be computed lazily, saving:

- memory when the latent dimension is significantly smaller than the number of users and movies;
- computation time if most rating predictions are not required.

userId	movieId	rating	prediction
11	55363	3.0	3.0168087
11	33437	3.0	3.0608275
11	58559	4.5	3.5650816
11	57368	3.5	2.8809972
11	55247	4.5	3.3122718
11	53921	3.5	2.8489442
11	47	3.5	3.4671936
11	7347	3.5	2.8703792
11	48774	3.5	3.239589
11	2054	2.5	2.4438343
11	53322	4.0	3.1268735
11	49272	3.5	3.3229468
11	60126	3.0	2.8533041
11	56633	2.5	2.9591103
11	52973	3.5	3.033453
11	49130	3.5	3.097041
11	51935	4.0	3.18803
11	55729	2.0	2.973852
11	44555	4.0	3.4821699
11	61132	3.5	2.8999534

only showing top 20 rows

# Hybrid Recommendations





# Conclusion

For the Actors:

- *We have seen that there is correlation between the number of extra roles played by an actor in their career and the possibility that they are more likely to have a good career, compared to actors who have only fulfilled one or few positions.*
- *We can say that it's very difficult to find a clear metric to define how good an actor is, as it could depend on many different features that in most cases, even if using a measurable metric, depend on subjective tastes or opinions.*

# Conclusion

For the Movies:

- *We have seen that the ability of the costume designer doesn't affect the vote given by reviewers.*
- *We have seen that classifying movies by genre based on their plots can be tricky, as it depends on several factors. The number of genres, the amount of available data and the length of the provided plots can all affect the final accuracy of the classifier.*

# Conclusion

For the Reviewers:

- *We have seen that the ALS model shows good performance. Generally speaking, the results are not enough to estimate the actual rating given by the user, but they are sufficient for a practical usage: recommending a movie based on the preferences previously expressed by the user. The model could be improved by exploring the hyper-parameters space, given enough computational power and/or time.*



**Thanks for the Attention**