The Man Who Loved His Little Girl

Presented by The Crazy Saturdays
Timothy Rivers, Charlie Lee, Maria Ng, Sabrinia Crouch
Client: Potential Movie Producer-William Moneybags
18 Dec 2019

INTRODUCTION

- Here we will show you that not only is our movie going to be profitable, but we have performed an extensive analysis to show that we have optimized the language, genre, release date, popularity and runtime to yield the highest revenue
- We have also performed an analysis of previous titles across genres to optimize titling of the movie
- Following this presentation, we hope that you agree to provide us the *money* needed to fund our sure-fire venture

PREPARING THE DATASET

- The Movies DataSet was used to perform our analysis
  - These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017.
  - Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.
  - This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

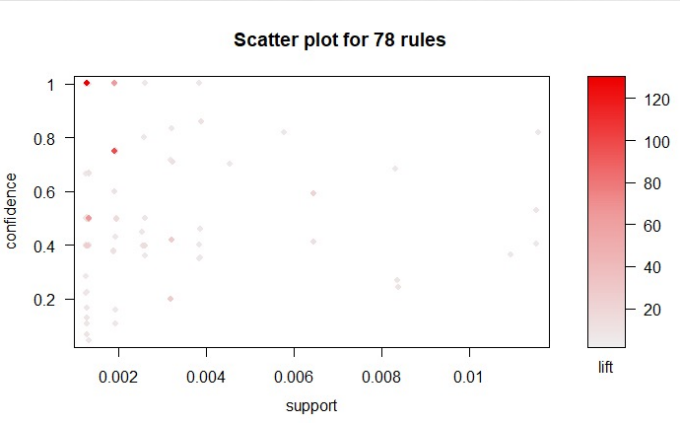
Selecting the Genre

- We built a linear model to identify which genres had a significant correlation with revenue
- We identified that music and animation are the most significant
- However neither model had a good R^2 (Both ~0.02)

```
lm(formula = revenue ~ Action + Adventure
+ Animation + Comedy + Crime +
Documentary + Drama + Family + Fantasy +
Foreign + History + Horror + Music +
Mystery + Romance + Science.Fiction +
Thriller + TV + Western, data = genre)
```

Text mining relationships

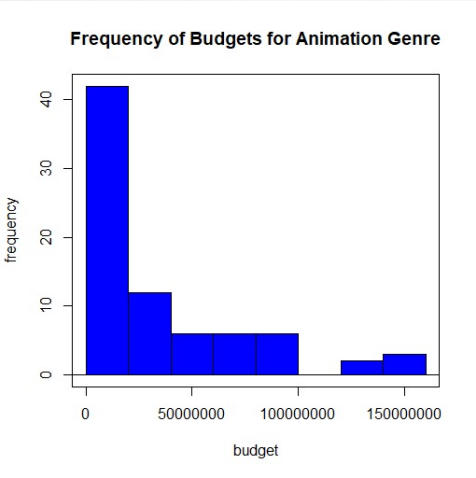
- We also wanted to know if animation and music occurred frequently together
- In none of the top 78 rules for our data set where lift >5, animation and music do not appear together

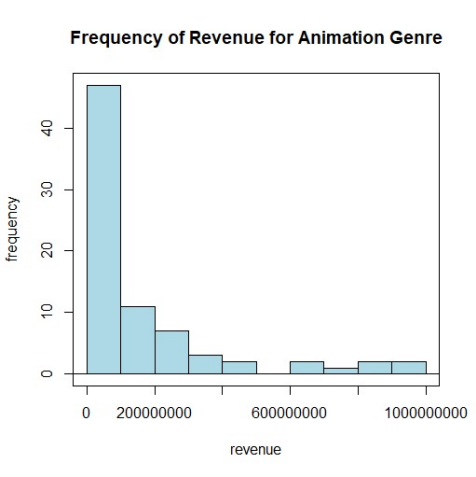


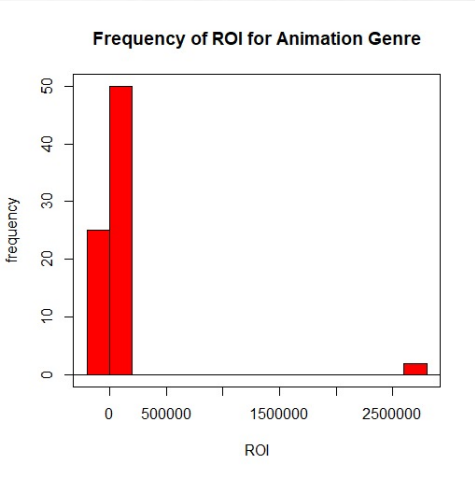
	lhs		rhs	support	confidence	lift	count
[1]	{Thriller}	=>	{Drama}	0.003211304	0.41666667	25.950000	5
[2]	{Drama}	=>	{Thriller}	0.003211304	0.20000000	25.950000	5
[3]	{Thriller}	=>	{Foreign}	0.001284522	0.16666667	8.370968	2
[4]	{Foreign}	=>	{Thriller}	0.001284522	0.06451613	8.370968	2
[5]	{History}	=>	{War}	0.008349390	0.24074074	7.809028	13
[6]	{War}	=>	{History}	0.008349390	0.27083333	7.809028	13
[7]	{Horror, Thriller}	=>	{Drama}	0.001284522	0.40000000	24.912000	2
[8]	{Drama, Horror}	=>	{Thriller}	0.001284522	1.00000000	129.750000	2
[9]	{Crime, Thriller}	=>	{Drama}	0.001926782	1.00000000	62.280000	3
[10]	{Crime, Drama}	=>	{Thriller}	0.001926782	0.75000000	97.312500	3
[11]	{Action, Thriller}	=>	{Drama}	0.001284522	0.40000000	24.912000	2
[12]	{Action, Drama}	=>	{Thriller}	0.001284522	0.50000000	64.875000	2
[13]	{Science.Fiction, Thriller}	=>	{Adventure}	0.001284522	1.00000000	7.344340	2

```
goodrules_md_TDM1<-
arules_md_TDM1[quality(rules_md_TDM1)$lift >
5]
```

Budget, Revenue and ROI- Animation







- Large differences in the data in how lower values are reported for Budget and revenue
- This skews the ROI calculation
- Future projections were made on Revenue

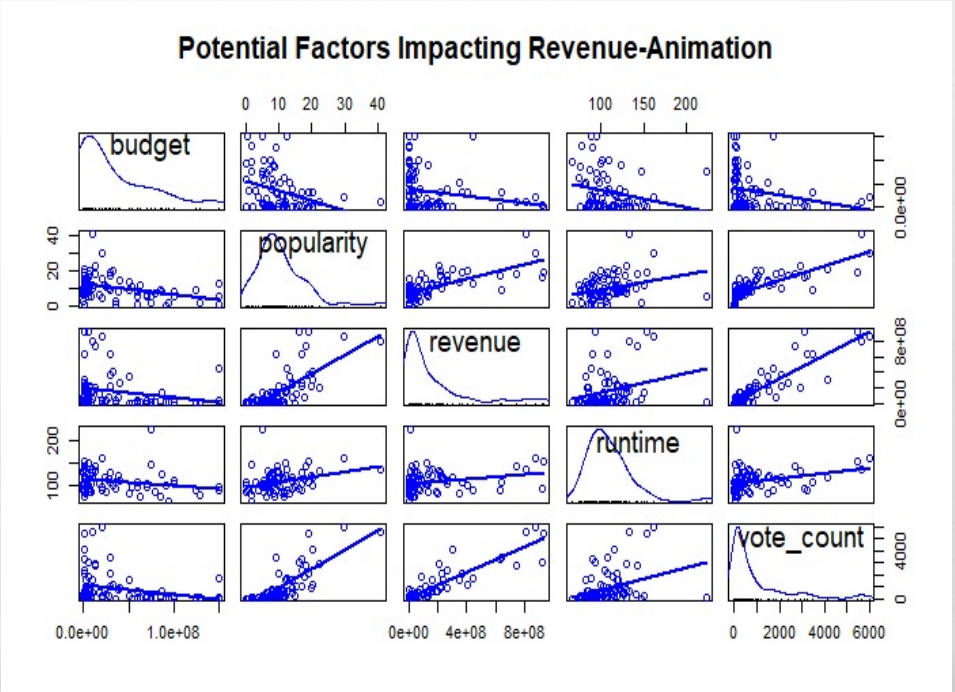
What factors impact Revenue?

- What factors were most impactful on Revenue for Animation and Music?
  - Budget
  - Popularity- incorporates number of votes, views per day, positive reviews, release date etc
  - Runtime- length of the film
  - Vote Count- the number of people who watched the film-good or bad

- What models are most predictive of Revenue?
  - Linear models
  - SVM/KSVM
  - Neural nets

Factors impacting Revenue- animation

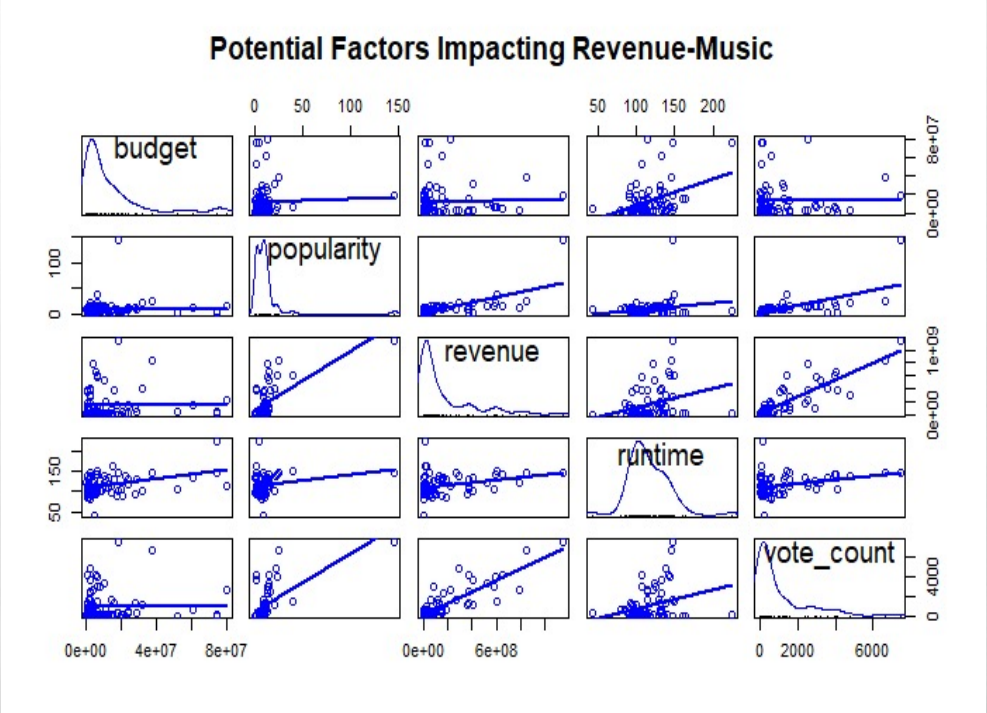
- **Looking at our top parameters:**
  - **Budget**
  - **Popularity**
  - **Runtime**
  - **Vote count**
- **Strong linear correlation between revenue and vote counts**
  - **Less strong with popularity and runtime**



Ideally we would like to predict on Budget or Runtime- things we can control!!

Factors impacting Revenue- music

- **Looking at our top parameters:**
  - **Budget**
  - **Popularity**
  - **Runtime**
  - **Vote count**
- **Strong linear correlation between revenue and vote counts**
  - **Less strong with popularity and runtime**



Ideally we would like to predict on Budget or Runtime- things we can control!!

Predicting Revenue- Budget_linear model

- As suggested by the previous scatterplots, the linear model is a poor fit. Although it is marginally significant, it has a very poor R^2

```
lm(formula = revenue ~ budget, data = animation)

Residuals:
       Min          1Q      Median          3Q         Max
-191229250 -132784285  -74693120   29587311  742126730

Coefficients:
                Estimate     Std. Error t value     Pr(>|t|)
(Intercept) 191901634.4917  33806374.9684   5.676 0.000000246 ***
budget             -1.1528         0.6379  -1.807       0.0748 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 229100000 on 75 degrees of freedom
Multiple R-squared:  0.04172,    Adjusted R-squared:  0.02895
F-statistic: 3.266 on 1 and 75 DF,  p-value: 0.07476
```

Predicting Revenue- Budget_linear model

- **As suggested by the previous scatterplots, the linear model is a poor fit with a very poor R^2 and p-value**

```
lm(formula = revenue ~ budget, data = music)

Residuals:
       Min          1Q      Median          3Q         Max
-158516998 -145440046 -115955404    27155725 1000071032

Coefficients:
                  Estimate      Std. Error t value  Pr(>|t|)
(Intercept) 149007083.6325  34277790.0601   4.347 0.0000446 ***
budget              0.2348         1.5750   0.149     0.882
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 237700000 on 72 degrees of freedom
Multiple R-squared:  0.0003086,  Adjusted R-squared:  -0.01358
F-statistic: 0.02222 on 1 and 72 DF,  p-value: 0.8819
```

SVM models- Budget and runtime

Support Vector Machine object of class "ksvm"

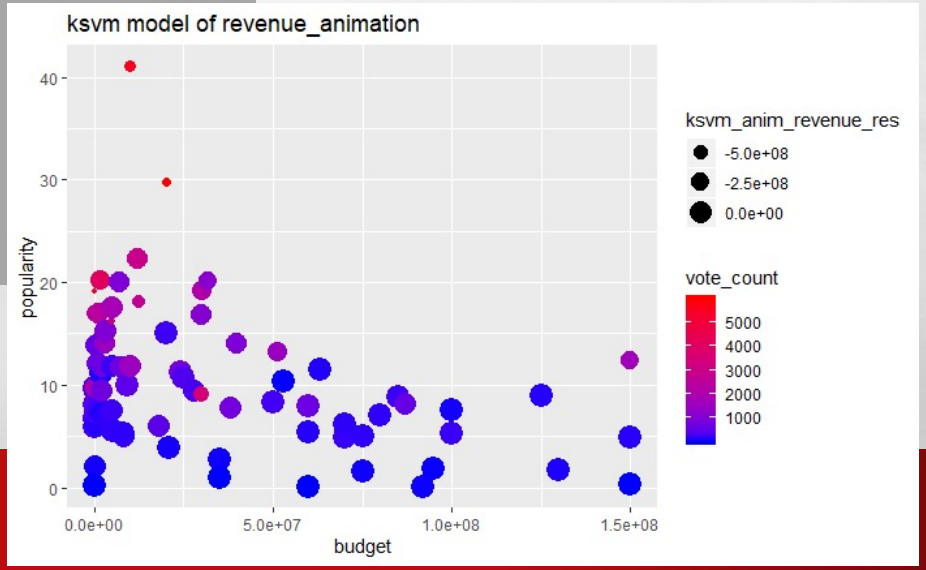
SV type: eps-svr (regression)
 parameter : epsilon = 0.1 cost C = 3

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma = 0.47022135669666

Number of Support Vectors : 40

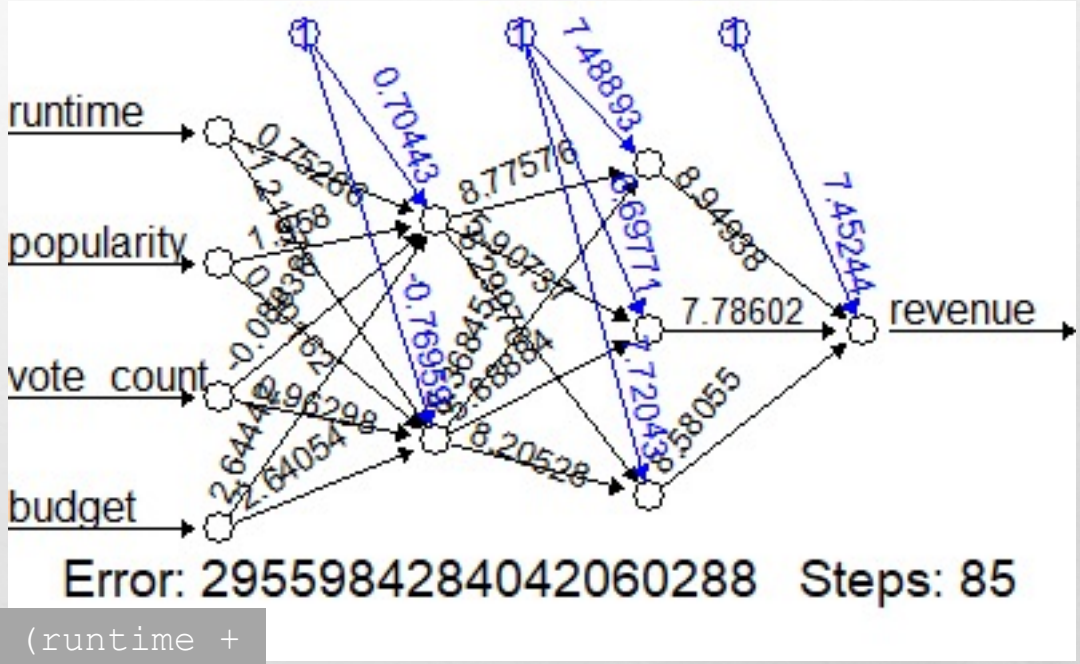
Objective Function Value : -21.3946
Training error : 0.059256
Cross validation error : 2.815407e+16
Laplace distr. width : 120186909

- Poor model fit with SVM models
  - Lower budget -> vote count
  - No obvious relationship to vote count
  - Animations are less expensive to make but very popular



Neural nets- predicting revenue

- **Using more complex modeling with neural nets did not improve our predictivity of revenue**




```
movie_net<-neuralnet(revenue ~ (runtime +
popularity+vote_count +budget), PRRB_1,
hidden = c(2, 3), lifesign="minimal",
linear.output =FALSE, threshold =0.0001)
```

Predicting by vote_count

- Although not ideal, we are able to predict revenue on vote count
- Therefore we could predict potential profits early, based on early vote counts or decide to withdraw the film

Predicted Revenue	Vote count
$ 13,123,436.00	0
$ 168,117,436.00	1000
$ 323,111,436.00	2000
$ 478,105,436.00	3000
$ 633,099,436.00	4000
$ 788,093,436.00	5000
$ 943,087,436.00	6000

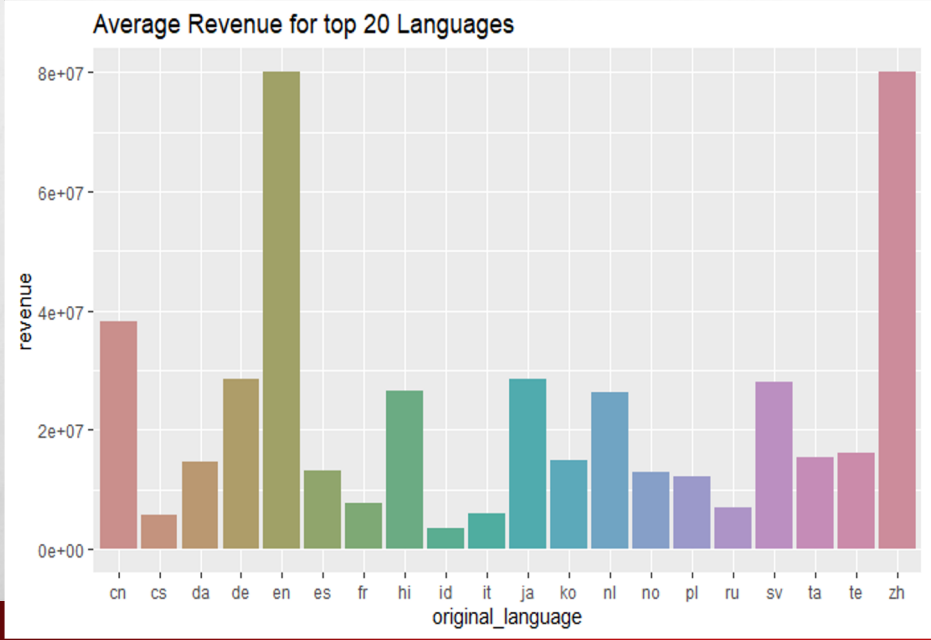
```
lm(formula = revenue ~ vote_count, data = animation)

Residuals:
       Min          1Q      Median          3Q         Max
-350724426  -38589091  -13419307   34104232  429023854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13123436   14183143   0.925    0.358
vote_count    154994       8788  17.636   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 103100000 on 75 degrees of
freedom
Multiple R-squared:  0.8057,  Adjusted R-squared:  0.8031
F-statistic:   311 on 1 and 75 DF,  p-value: < 2.2e-16
```

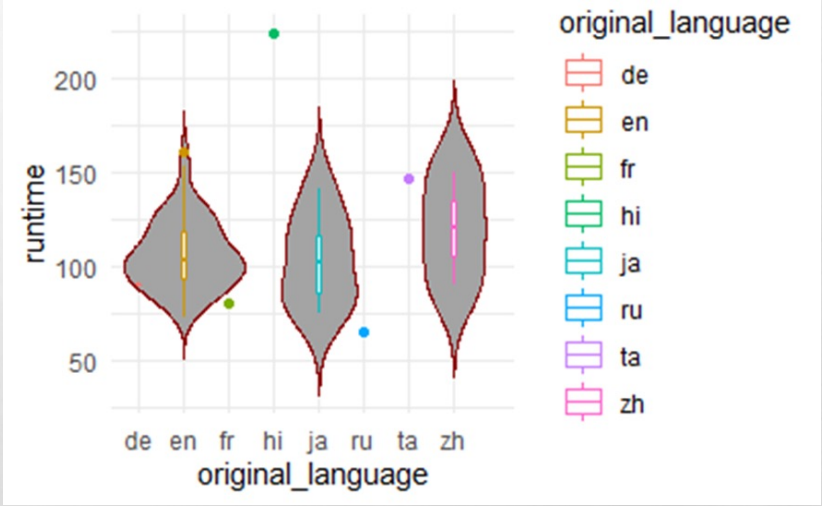
RELEASE DATE

- **Based on months that generate the most revenue, our movie should be released in May- June**

LANGUAGE/RUNTIME

We will launch our movie in English then in chinese, german and japanese with a runtime of ~100minutes and no longer than 160 minutes

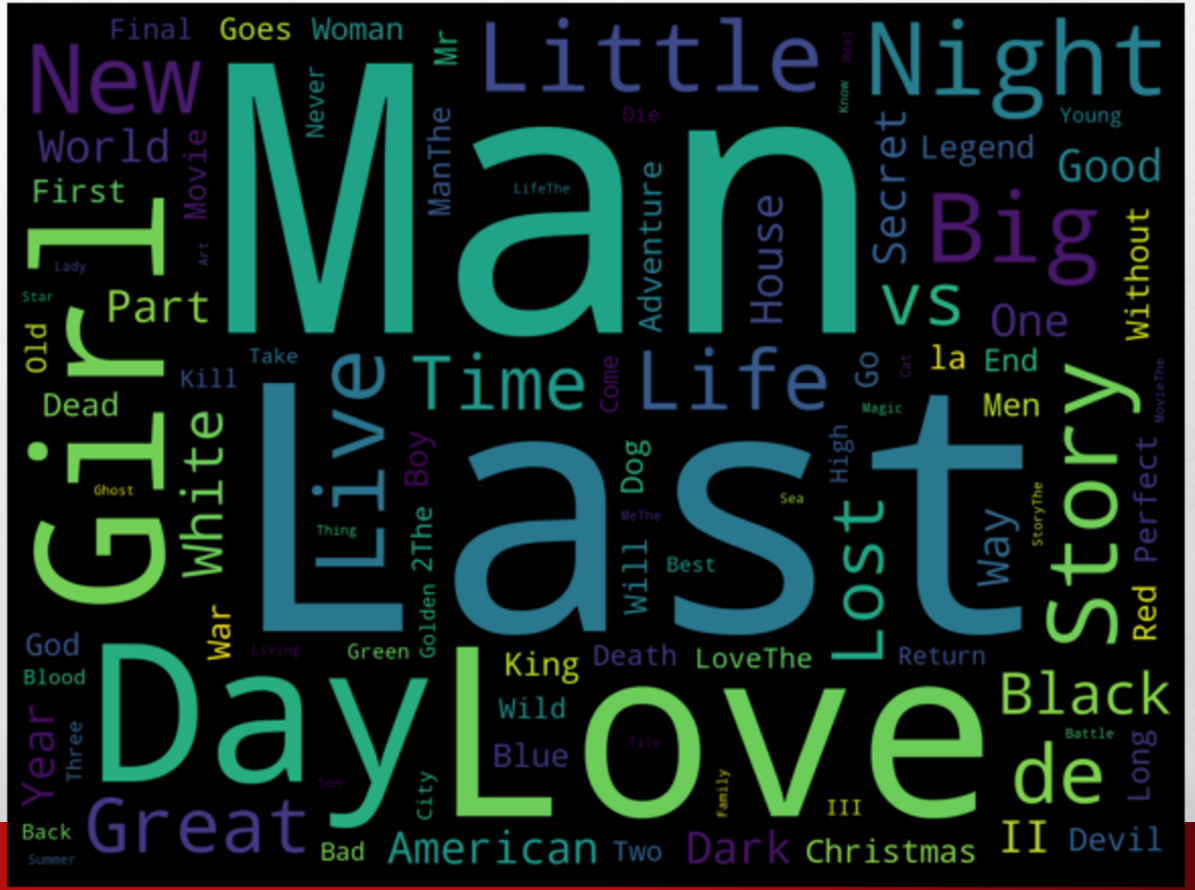




TITLE

- Using the titles of the most popular movies, we established the title of our movie as:

THE MAN WHO LOVED HIS LITTLE GIRL



KEY TAKEAWAYS

- Data preparation was key for this project since the dataset was so large
- Identifying a genre helped to focus which movie area we would work in and develop models
- Even with large datasets, we rarely had models that fit well- only once we limited scope we could identify models with good fit and minimal error
  - Unable to design practically predictive models for revenue using our data
- Although the dataset was not informative from a revenue perspective, it provided a lot of details to help design the movie- release date, movie title, language etc.

CONCLUSION

- Based on our data we will create a cartoon that tells the story the antics of a rugged Dad from Montana and his daughter who leaves the snow for sunny California
- The Animation will be released in the summer to an English speaking audience with a runtime of 100-160 minutes
- We will monitor early signs of the vote count to extrapolate potential sales and monitor markets for expansion.