# 

# Netflix TV Shows and Movies Analysis

Intro to Data Science - IST 687  
Final Project  
Group 1

Kelsey Kirby, Blessy Thomas, Hannah VanTilburg, Patrick Walsh, Mark Stiles

## **Project Objective**

When selecting a dataset, we were looking for something we all had some domain knowledge of and a reasonably clear goal. We’re all familiar with online media and felt we understood the purpose of predicting the IMDB score. We found several different movie datasets but settled on this one because it had sufficient data to analyze, a good number of fields to compare, and a wide range of scores to predict. Ultimately we wanted to see if we could predict if a potential future movie would be critically successful and determine what factors lead to high user satisfaction.

<https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies>

## **IMDB (Internet Movie Database) and Netflix**

The dataset we selected relies on the IMDB scores from a list of Netflix movies and television shows. IMDB is a website that tracks movies, television shows, the cast and production crews. It allows users to vote on titles and then averages the scores.

Netflix is a streaming service that has a combination of rotating titles and originally produced content. The dataset we worked on was a subset of Netflix titles at an undetermined time that had associated IMDB scores.

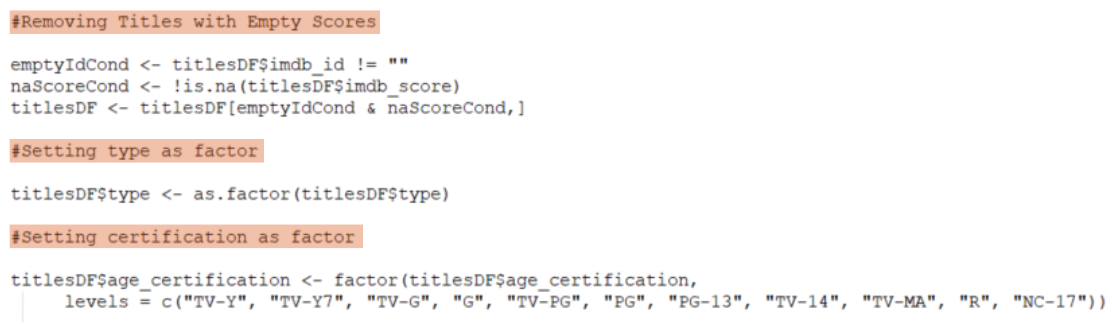
## **Data Processing**

This is a list of the variables our dataset provided:

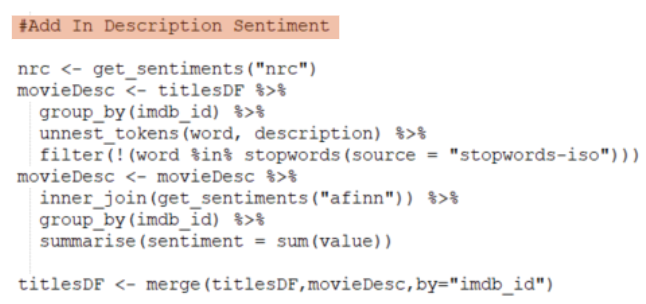
**Title Dataset**

| **Name** | **Description** |
| --- | --- |
| ID | Dataset unique id |
| Title | Movie or TV Show title |
| Type | Either Movie or TV Show |
| Description | Movie or TV Show Description |
| Release Year | The year it was initially released |
| Age Certification | The title rating (G, PG, R) |
| Runtime | The title length in minutes |
| Genres | The list of genres associated with the title |
| Production Countries | The list of countries where the title was produced |
| Seasons | How many seasons the TV Shows had |
| IMDB ID | IMDB’s tracking ID |
| IMDB Score | IMDB’s average score for the title |
| IMDB Votes | IMDB’s number of votes on the title |
| TMDB Popularity | The Movie Database popularity |
| TMDB Score | The Movie Database score |

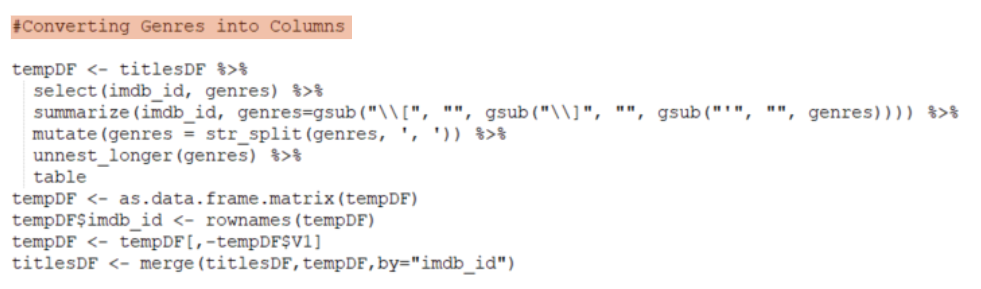
We cleaned up the values and developed additional variables over time with what we realized needed to be done and additionally what we thought may add more quality to our model after we had spent time with the variables and had ideas about additional information encoded into the variables. To begin with we removed titles with empty scores. It wouldn’t make sense to train on data where the score couldn’t be inferred through other means. We also then converted the title ‘Type’ and ‘Age Certification’ to factors.



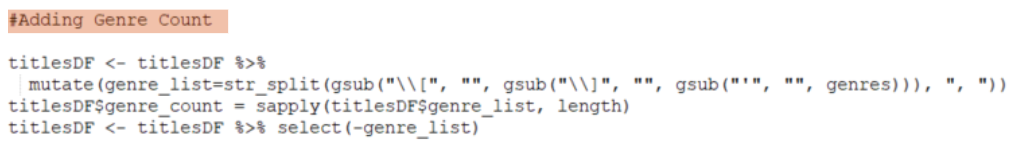
Next we used the description field to generate a sentiment value for the title. We thought there may be a correlation between sentiment and score.



We then converted the list of genres into a list of columns with a logical value. We thought this would be helpful in training the models later.

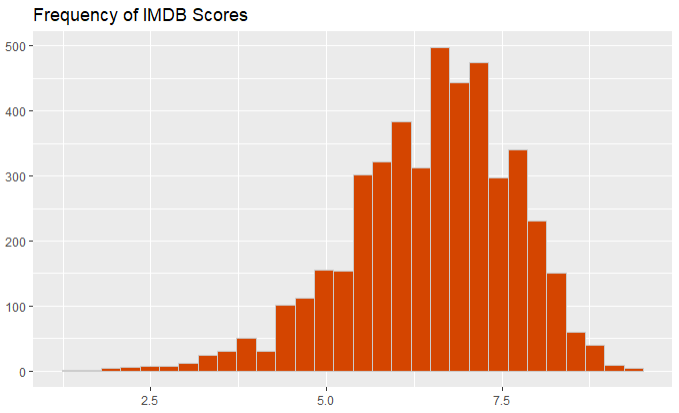


We also thought there would be a correlation between the number of genres attached to a title and the score. We added another variable for that to each title.

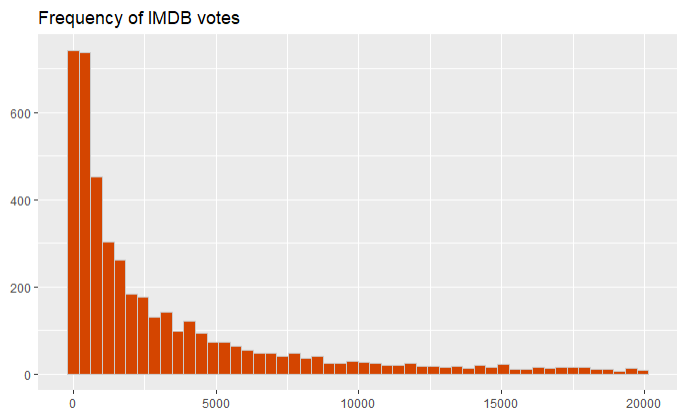


## **Exploratory Analysis**

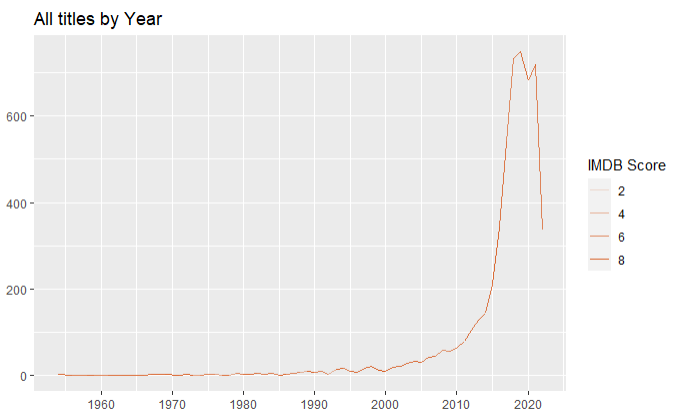
In doing our analysis, the first thing we looked at was the distribution of IMDB scores. We found that the mean IMDB score for all titles was 6.5, and the range went from 1.5 to 9.5. This is interesting because it means that most media was well-liked (above 5). There were samples of greatly liked and disliked titles. It would better predict a platykurtic distribution with more examples at each tail. Still, it does at least have some representation on the lower end.



We also wanted to know how the vote distribution may affect the outcome of scores and looked at the distribution of votes. We found the mean was 25,074 votes, ranging from 5 to 2,294,231. This is good because it means that most titles were scored by a sufficient number of people and would accurately reflect what people like.

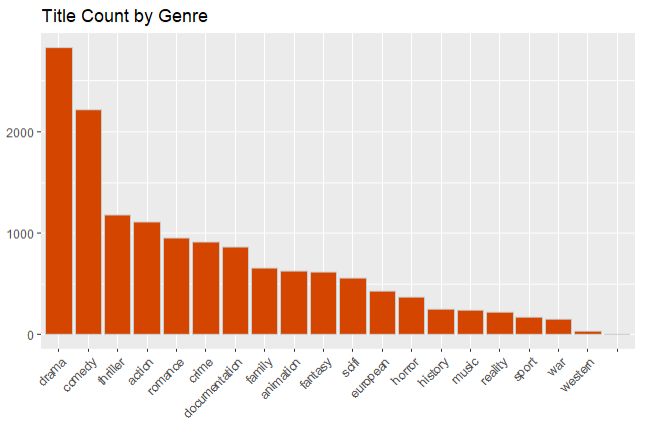


We then explored how the titles were distributed throughout the years. The dataset was heavily weighted toward newer movies. This was a challenge since we wanted a wider distribution throughout the years. Any predictor we create may be more accurate for newer titles but potentially inaccurate for older titles.

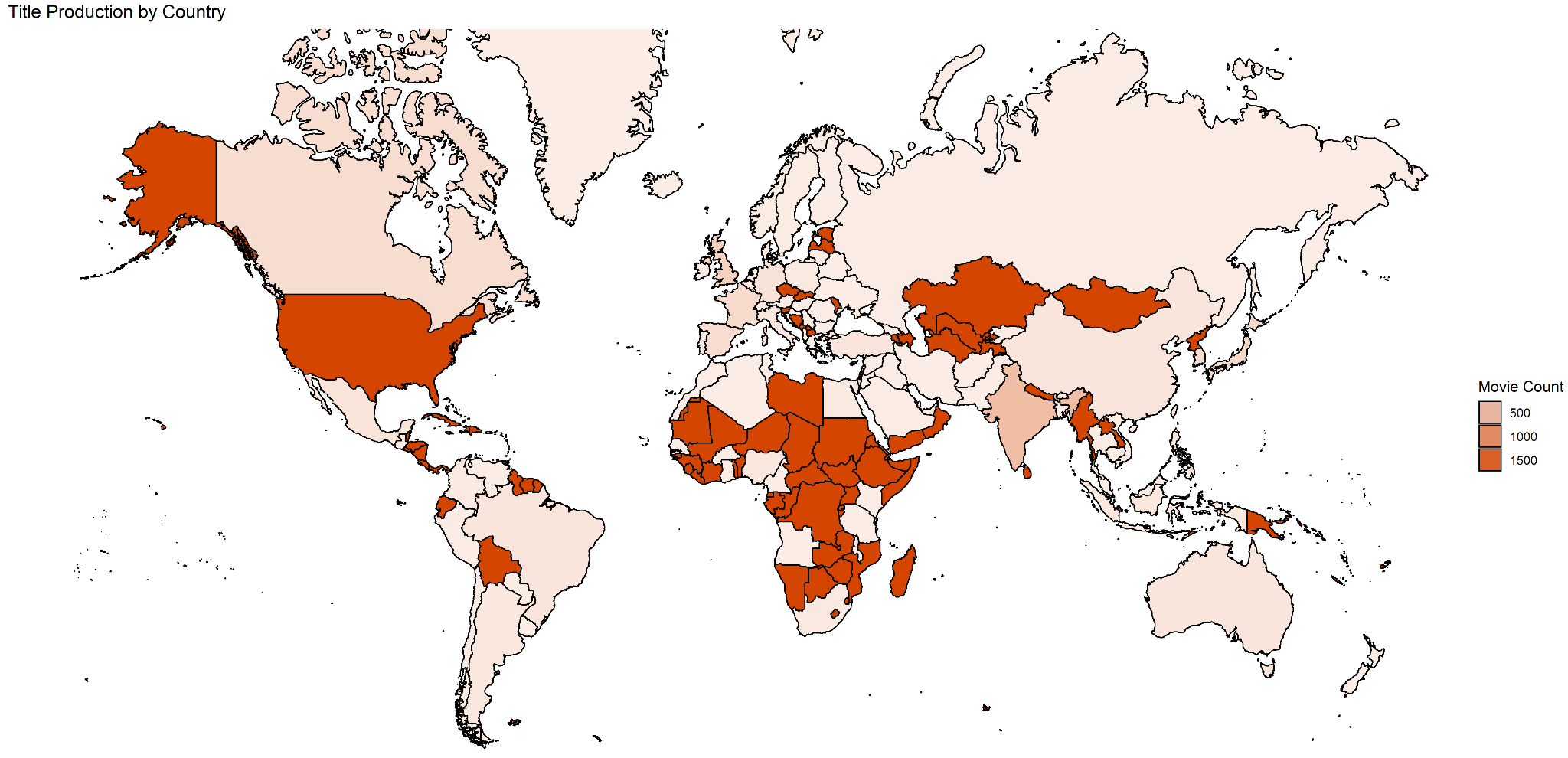


We then extracted the genres from each title to get a clearer view of how they were distributed. Titles could be assigned more than one genre, so we needed to break them out into their variables to sum them across all titles. We found that there wasn’t an equal amount of titles per genre. Drama and comedy held the highest volume. It’s unclear if this is an artifact of the specific selection of the dataset or an accurate representation of the broader view of all titles. Given that the dataset is not a true sampling of all titles created but rather a curated collection of titles by

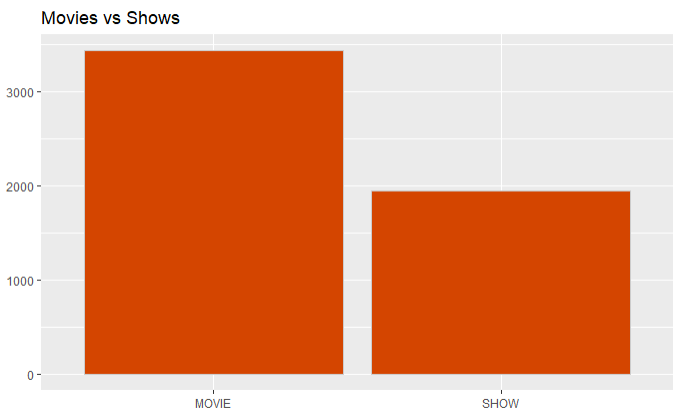
Netflix, it probably reflects more of what people would pay to watch.



Since we had the production country available, it made sense to see what the distribution looked like, which we thought was surprising. Countries in Africa and Asia produced more movies than countries in Europe and South/Central America. It was interesting to see but ultimately didn’t produce anything helpful in determining what a score might be.



One of the final statistics that we used was the title type. We wanted to see how the media distribution was, and it was surprising that most of the Netflix titles were movies by almost a 2-to-1 margin. We would have expected more TV shows given the rate of in-house production occurring with them, and also that binge-watching is easier to do with shows than movies. We still opted to dig in further with the TV shows because there would be more insight into television since it had more characteristics to analyze, such as seasons.

****

**Expected Findings**

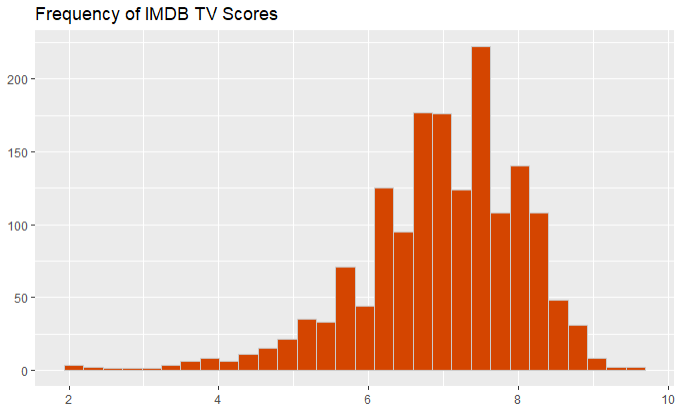
* We expected a heavier weight on newer titles, given the rate of new production and the financial benefit to Netflix over content licensing.
* We did expect there to be a large dip in the runtime of television shows during the 80s and 90s due to the growth of ad time during episodes, and it receded after streaming took over.
* We didn’t find all variables useful and focused on the Type, Description, Release Year, Age Certification, Runtime, Genres, Seasons, IMDB Votes and IMDB Score. The rest didn’t make *the final cut*.

**Unexpected Findings**

* Movies made up the majority of the titles instead of shows which we did not expect.
* We didn’t expect the average ratings to decrease over time. We expected new technology and higher budgets to improve movie quality.
* We didn’t expect that the two genres with the most titles were drama and comedy given the prolific amount of war movies produced each year.
* There were a lot of countries where titles were produced that we didn’t expect (African countries) and many countries where production didn’t occur much (European countries).
* The mean of the ratings for titles was around 6.5, so generally, movies were well-liked.
* The average titles had 25,000 votes which were more than we expected.
* Most titles were for adults, and fewer and fewer were for younger viewers.

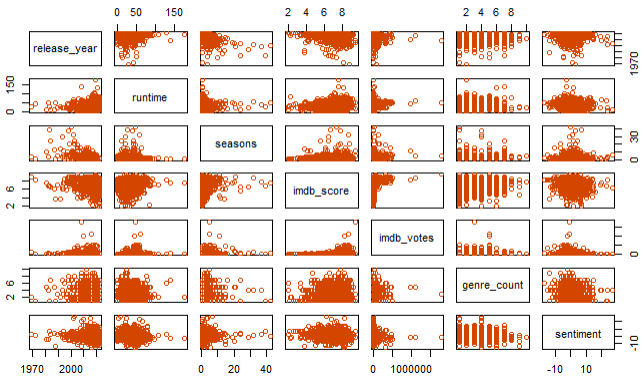
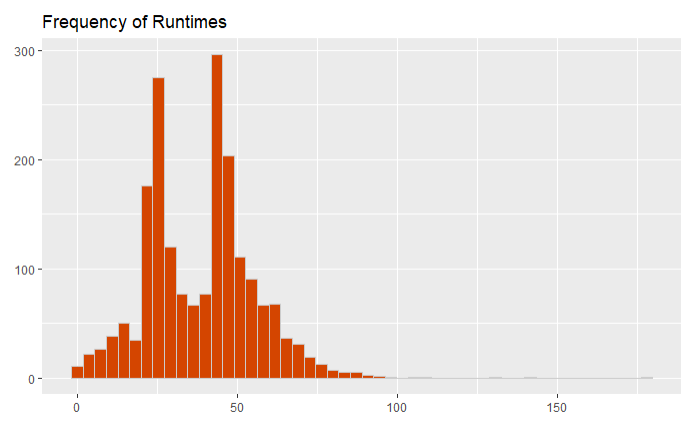
## **TV Show Analysis**

In our initial exploratory analysis, we found two groups of titles: tv shows and movies. We decided to focus on tv shows, given they had some additional variables to analyze. We found that the mean for only TV shows was 6.9, which was higher than all titles. The range of scores went from 2 to 9.5.

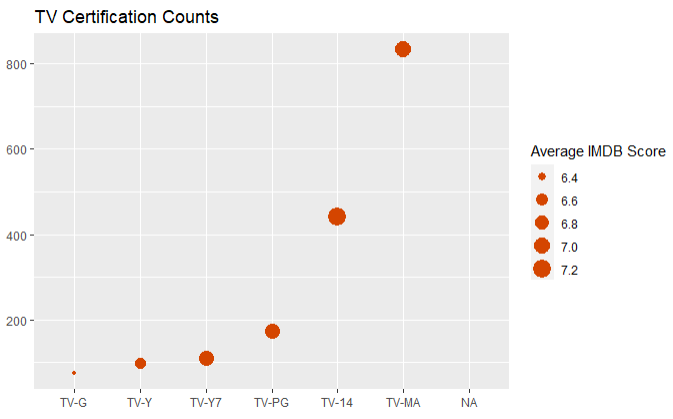


One interesting pattern we did find was in the frequency of runtimes. There were two peaks due to the unique property of television shows. Due to commercials, broadcast television shows had 30 or 60-minute ranges equal to 25 or 45-minute episodes. Newer streaming-only shows have provided more spread, but it is still apparent from the distribution.

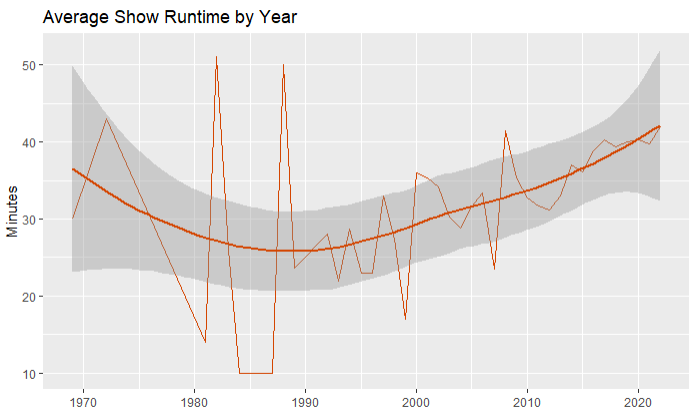
In looking for interesting patterns in the remaining variables, we used the pairs function to find potential highlights. There weren’t any obvious patterns, but it did help us eliminate a few options.



When analyzing the television certifications, we weren’t expecting to see such a curve between youth-rated and adult-rated shows. Still, Netflix must be focused on providing most of the content to paying adults. However, we expected more content to be aimed at children and the low scores for children’s content could be attributed to parents not liking it and a lack of children voting on IMDB titles.



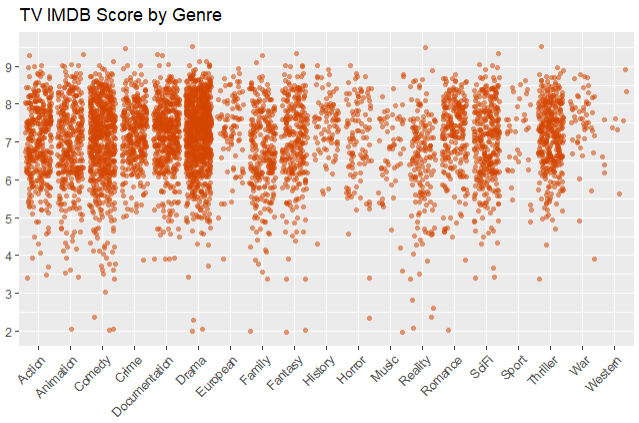
When taking a look at how the average show runtime changed over the years we weren’t surprised at what we saw. There’s a clear dip between the 80s and 90s due to what we attributed to a heavier reliance on commercial breaks. As they became more lucrative, television shows dedicated more time to them for each episode. The episode length skewed up as the internet streaming services began producing content. Many shows may release fewer episodes but with longer runtimes.



We also wanted to look at the frequency of words in television show descriptions. It contains shows of all genres but it was interesting to see what the most frequent words were. It didn’t surprise us that “Life”, “Love”, “Family” and “Friends” were at the top as a lot of television dramas revolve around families and relationships.



One visual that we developed gave a sense of how densely the votes were packed for each genre and in a sense how valid the score was. Drama has a very high density; from that, you can infer that there’s a lot of confidence in the scores it receives. On the other hand, Sport and Western do not have nearly as many votes, making the overall score seem less valuable because there was so much less data involved.

****

**Expected Findings**

* Within the shows there were two distinct groups: 25 min long episodes and 45 min long episodes

**Unexpected Findings**

* There were a lot of shows that fell outside of the 25/45 minute span
* Most content was aimed at adults. We expected a more even spread

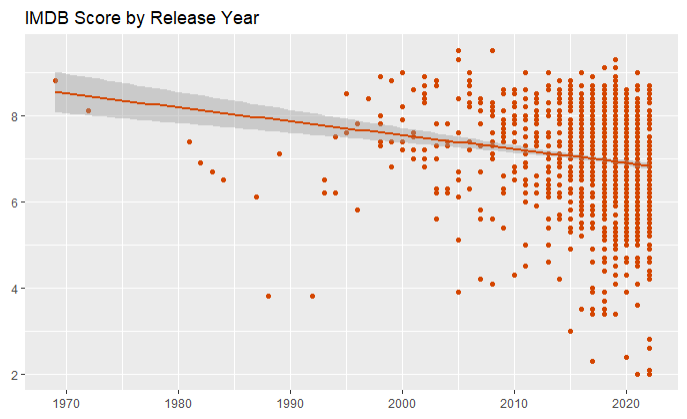
**Interesting Insights**

* Most content was targeted at adults and very little towards children.
* Netflix contains a lot of shows that are generally liked, but only a few have great content.
* A significant number of shows fit outside a 25 or 45-minute runtime length.

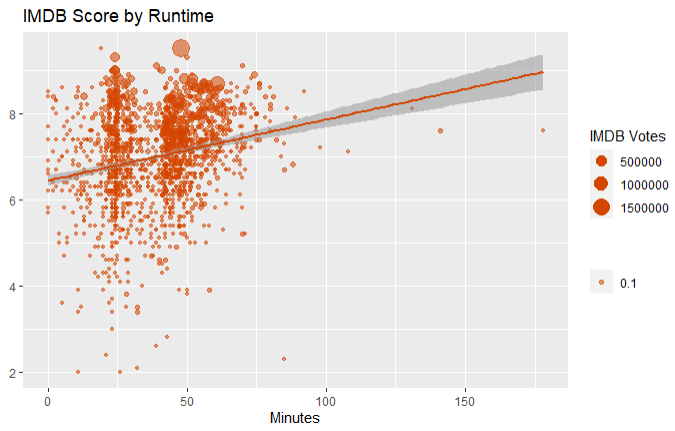
## **Data Analysis**

As we began focusing on predicting the IMDB scores, we first focused on variables that could pair with them.

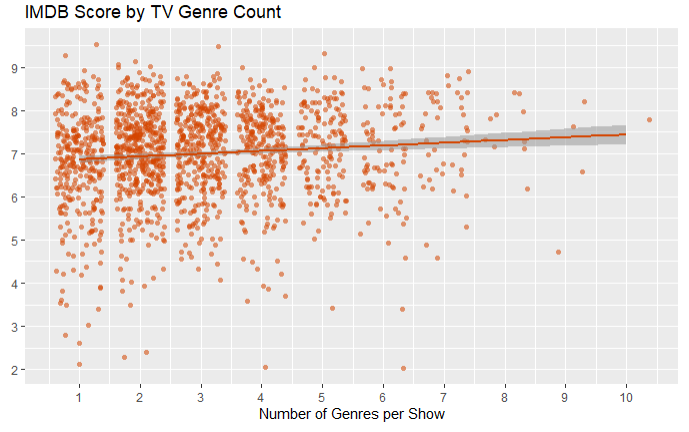
Release year was the first variable we tested, and there was a reasonable correlation. We expected a positive correlation; hence it was surprising that the average score decreased over time. This could be due to the lower volume of titles in the earlier years but even running the regression scoped to titles produced after 2015 still had a negative correlation.



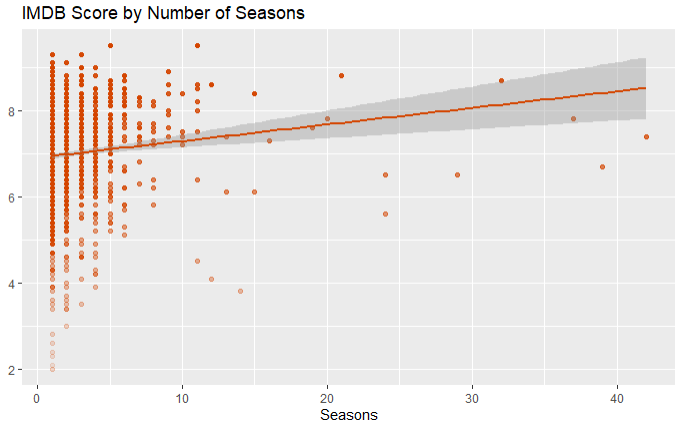
We didn’t expect there to be a correlation between score and runtime. We expected the score to be more related to the content, so it was surprising to find a high correlation. Even eliminating outliers of high and low runtimes produced a similar linear slope.



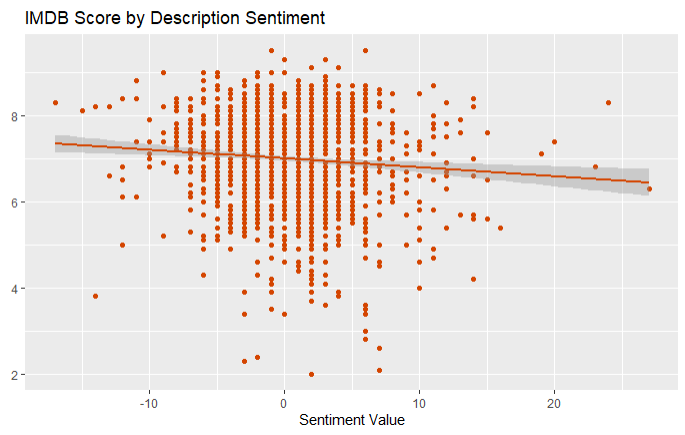
We did expect to find a significant relationship between the genre count and the score, but there wasn’t one. We reasoned that shows that tried to tackle too many topics would suffer from focus, but that wasn’t true.



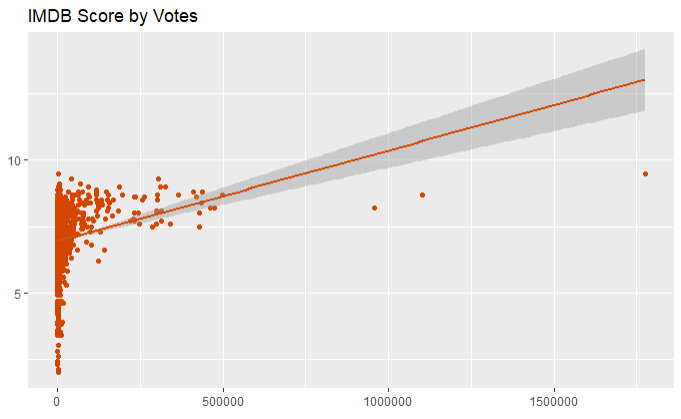
We expected that there would be a positive correlation between IMDB scores and the number of seasons that a show has. If a show continued to get renewed, it was assumed to be well-liked. We weren’t wrong, but we expected a higher positive effect than there was.



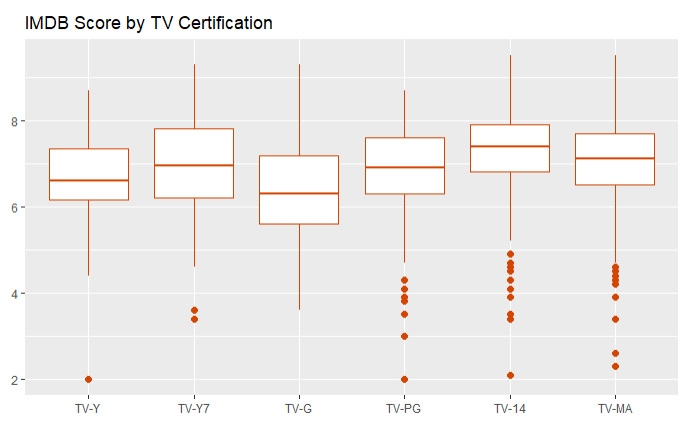
We weren’t sure if the description sentiment would affect the score. However, we were surprised to find a slightly negative effect. It could be that many shows dealing with real-life problems tend to be scored somewhat higher.



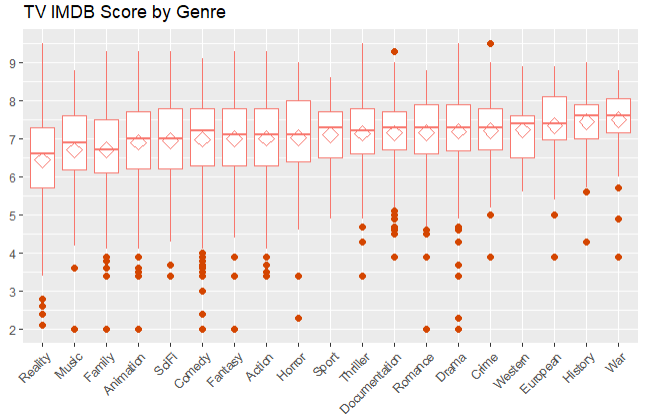
We expected a positive correlation between votes and scores because many people want to vote for an extremely popular title. We weren’t wrong; even removing extreme outliers resulted in a similar positive slope.



We were uncertain if there would be any relationship between the ratings and the score, and we found that there wasn’t. We converted the ratings to factors so that any relationship could be compared.



Given the distribution of genres, we expected to see a correlation with the score, but we were surprised that there wasn’t a hierarchy of genre scores. On the contrary, they all fell within a small range. This may be an artifact of Netflix’s method of selecting titles, but we expected more stratification.

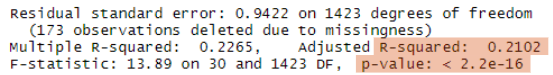


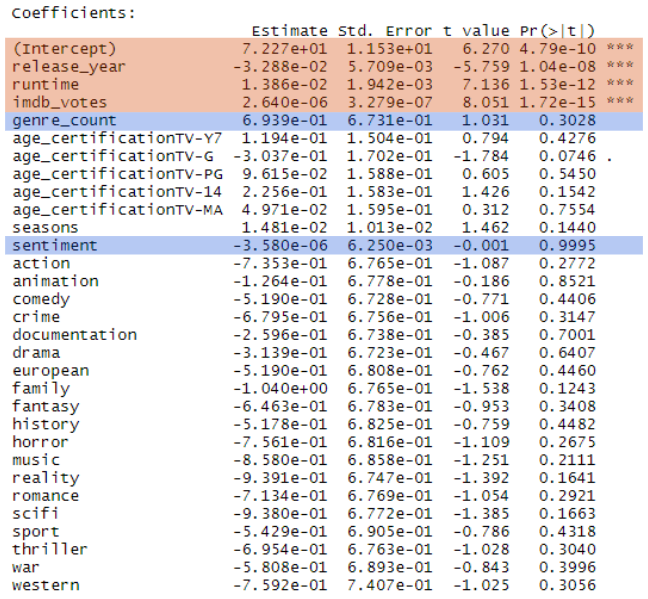
## **Predictive Modeling**

Now that we had identified the variables, we could predict the IMDB score. We tried four different models and the results varied. In the best cases, all could reach between .22 and .27 for an R-squared value but it was dependent on how the test/training split happened. In order to create some stability and reproducibility for the report we set a seed value and built the models on equal footing. This makes it a bit contrived but more consistent. Since we were doing a regression and not a classification we weren’t able to use a confusion matrix for results. Instead we measured the R-Squared value on the test data and used that as a final measure of success. We embedded this value into the prediction graphic.

### **Linear Model**

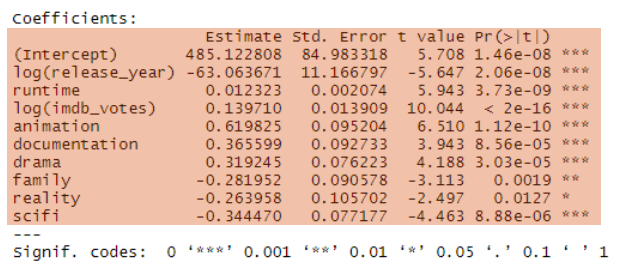
Our first model was a Linear Regression model. The initial results didn’t provide the best use of all the variables, and our two custom variables (genre\_count and sentiment) weren’t significant, so we removed them.

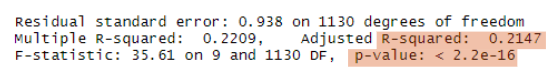




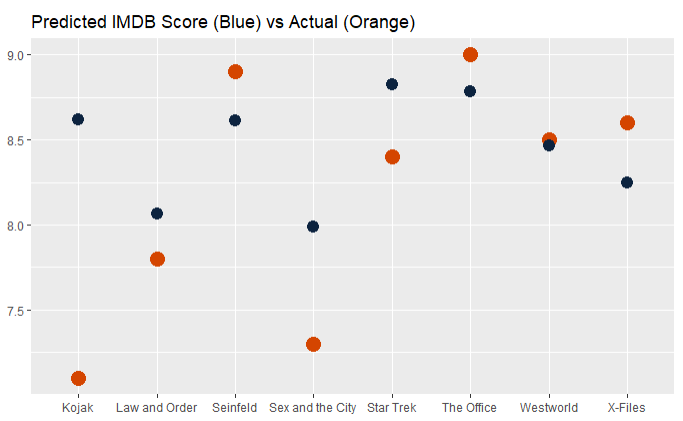
The top three variables (release\_year, runtime, and imdb\_votes) were taken ,and individual variables were added. We kept only the ones that improved the R-squared result and were also significant. We ended up with the following model that included several genres but not all. We also used the log of release\_year and imdb\_votes since they were not linear. In order to improve the quality of the values we wanted to bend them closer to a linear value to help reduce the error values. We ended up with an R-Squared value of 0.2147 which wasn’t terrible but also not great.

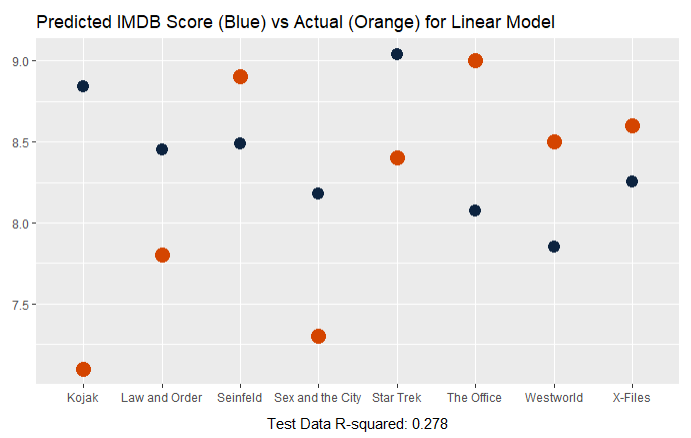
****

****

****

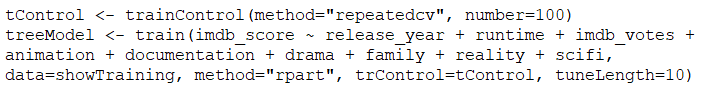
**Predictions**

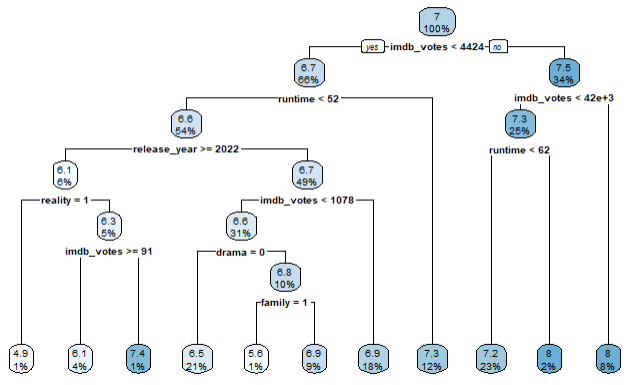
We tried to predict a few different shows that were randomly selected and not part of Netflix’s catalog. It was more accurate than we expected it to be, but a few titles were still missed by a wide margin. Overall, it appeared to be a decent model despite how low the R-squared value was. The R-squared value for the test data, which we included in the graph, was higher than from the training data. It wasn’t a lot higher but gave us a level of confidence that it was a stable value



### **Regression Tree**

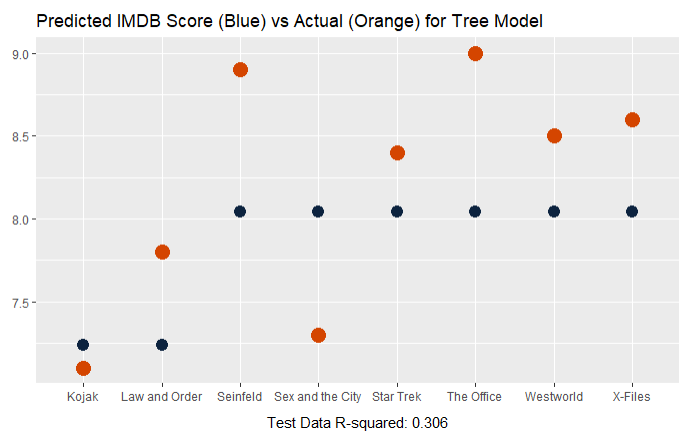
The second model we attempted to use was the Regression Tree model. The quality varied widely depending on how the testing and training data split. We didn’t feel good representing the best model given that it felt hard to reproduce. So we opted to show the most common model we were able to reproduce which had an R-Squared value of 0.306. When it did perform well we found that setting the number of repeats to 100 and a tune length of 10 worked well. We used the same set of variables as we did with the linear model but didn’t use the log of any variable since it didn’t work well when we tried.

****



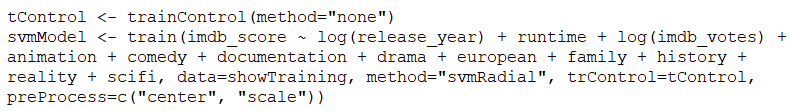
**Predictions**

The predictions were not very precise but did tend to get close enough on average to beat most of the rest of the models. This was pretty unexpected given the fairly simple logic it developed.



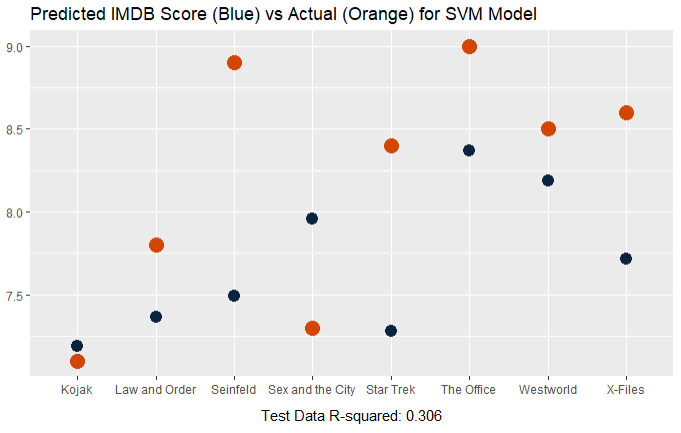
### **Support Vector Machine**

The third model we produced was a Support Vector Machine model and we applied a similar formula from the Linear model but with a few more genres that helped improve the score. This did outperform the Linear model but not by a lot. It did perform as well as the Regression Tree model.



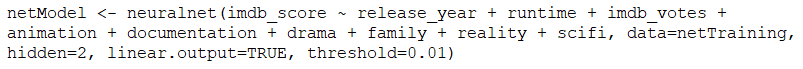
**Predictions**

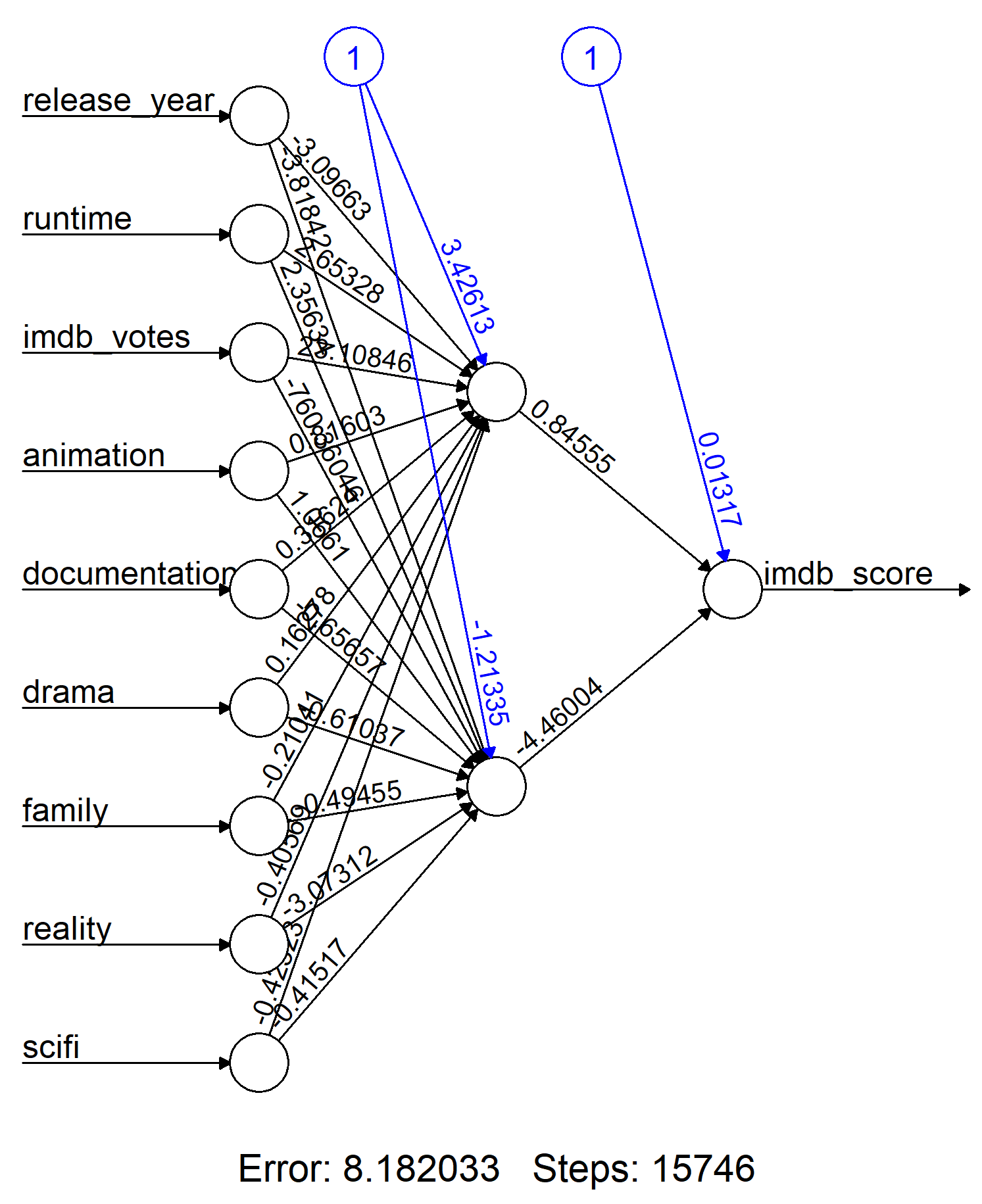
The predictions were a little more nuanced and seemed to have a better ability to generally predict values. It was obviously one of the best models we developed but I would expect it to beat the Regression Tree performance more generally.



### **Neural Network**

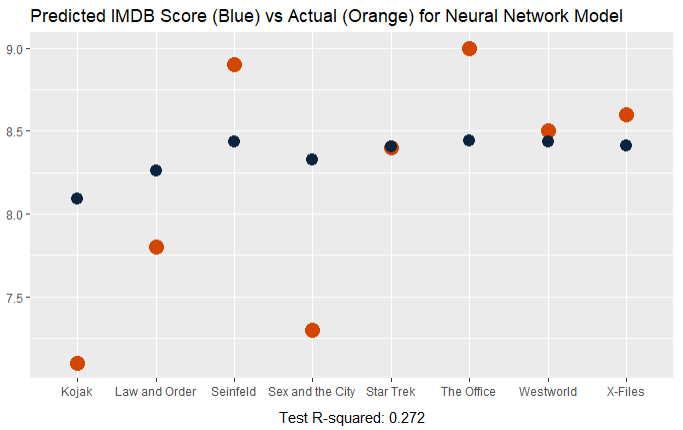
For the final model we developed a Neural Network. We started with a longer list of variables and ended up with the same relevant variables from the Linear regression. We didn’t end up using the log values because it didn’t work well but we were able to get a result similar to the rest of the models. We tried a lot of combinations of hidden nodes (12, 6, 4/2, 2/1, 3 and 2) but the best seemed to be two hidden nodes. We were surprised that it was a much simpler structure but it could be that there wasn’t enough valuable variable data to train with that a more complex model didn’t improve it any.





**Predictions**

The predictions were able to be somewhat nuanced unlike the Regression Tree. It was within a similar range of quality as the other models so it seems like there’s a lot of evidence that the value it was able to reach was about all the data could predict.



## **Results**

Here’s a table of the results of our models

| **Model** | **Training R-Squared** |
| --- | --- |
| Linear Model | 0.278 |
| Regression Tree Model | 0.306 |
| Support Vector Machine | 0.306 |
| Neural Network | 0.272 |

The best models were the Regression Tree and Support Vector Machine which was ok but not great. There are many reasons why we struggled to get a better model value. A wide number of years were used, but most of the titles were heavily weighted to one end. The years also produced significantly different shows with different qualities. Some genres didn't exist for the entire time covered. The people who voted also may not have voted for all of the titles but only the subset they watch. There are also external data points that may help provide insight into what drove results like cultural relevance. For example, if a show came out at a time when it dealt with an ongoing social issue, it might have made more of an impact in viewers' minds. If there's an oversaturation of a specific type of show, there may be diminishing returns on additional shows. This is also a subset of the total number of shows that Netflix thinks are a good mix for their audience. It may not be a representative sample of all shows, meaning a bias is built in. Ultimately, it is a complex problem, so there will never be a perfect model.

If we had more time or data available there were some improvements we would like to have made. The list of data that we wish we had is as follows:

* Number of views
* Percent of people who finished all episodes
* Average completion rate for each title
* The amount and type of referral to starting a show (we think you’ll like, homepage banner)
* Production budget
* Director IMDB Score
* Actor IMDB Scores

After spending this much time thinking about the problem, our team all agrees that with the right combination of information, we believe we could produce a reasonably well-performing model that could predict an IMDB score. Although it is very challenging to predict anything with a high level of certainty but even more with something that is so deeply personal as people’s viewing interest at any given place in time.