Introduction

The growing quantity of data that companies and everyday life produces nowadays allows for great analysis and answers to long dated problems. The applications of data science are limitless. Going from smart cities efficiency to cancer identification, it applies to every industry of the current world. Predictive analytics is what the following paper is all about, we saw it in different field including stock price prediction for example but today what we aimed to save you some money! Have you ever gone out for a movie night and regretted instantly the hour and a half spent in the theatre? Or complained about the huge waste of the 15$ spent on the ticket? Well, not anymore! Our work will allow you to evaluate a movie and make sure it is worth watching even before it is out.

Data

Overview

Our dataset contains 5457 observations on movies in various languages from different countries, across different genres. It consists of 55 predictive variables, and these variables are a combination of continuous and categorical variables ranging from movie budget to total number of directors. We encountered variables that we never thought would be meaningful in including in a predictive model, such as the gender of the director or the number of production countries. As it will be explained in more detail throughout our analysis, some of these variables were important to include and some of them had no statistical significance nor predictive power.

As we can see from the summary of IMDB scores we have in Appendix1, the IMDB scores of movies in our dataset have a mean of 6.67 and Appendix 2 shows that it is distributed with a right skew because most of the movies have high IMDB scores that are between 6 and 8 as opposed to strict critics in the lower range.

Cleaning and Adjustments

a. Formatting

The first adjustment we had to make was to ensure all of our continuous variables were in the same format before we did any analysis. When we examined the dataset, we detected inconsistencies about the representation of numerical values. Star-meter variable was in the "comma separated number" and the rest was in "general" format. We converted all the numerical values to general format for simplicity and to avoid distortions in our findings while working with R. Next, we identified all the categorical variables and changed their format to "factor variables" by using *as.factor.*

b. Missing Entries: Budget

After making sure all variables are in the right format, we needed to decide how to address the problem of missing entries in the dataset. Half of the observations were missing the budget values, as some producers simply choose not to share the movie budget with the public. We came up with multiple possible ways to tackle this problem. First thing we thought of was eliminating the observations that did not have budget entries. We chose not to take that path, because we did not want to lose more than a thousand observations. A larger sample size means more reliable results, greater precision and higher predictive power, so we decided to keep these observations and assign budgets to the empty cells instead. We considered assigning the median or mean value of the existing budgets for missing values, but quickly realized that would not be a good representation of the reality. Imagine a very short movie with no famous actors and no serious production costs. Assigning 22 million dollars (median budget) to that movie's budget would hurt the accuracy of our model, and lead to misleading solutions. So, we decided to predict the missing budgets with a multiple regression model. First, we split our dataset into two: The movies with an assigned budget, and the ones without a budget. Then, we trained a model on the dataset with budget values. To evaluate the performance of our model, we split that same dataset with budget values into training and test groups to perform a cross validation test. The model ended up having an adjusted r-squared of 68% and an error term of 10 million dollars. Although a high adjusted r-squared does not automatically indicate a good model, it shows that our model captures a satisfactory amount of variability in our data. We are also aware that such a model is not perfect

due to the relatively high error component, but we thought it was worth the compromise, since it was still a better option than using median or mean values for missing values. Therefore, we predicted the missing budgets with our model and obtained our complete dataset.

c. Adding New Variables

Besides having information on the day of the month the movie was released, we thought including the day of the week (Monday, Tuesday, Wednesday, …) as a predictor could provide some interesting insights. As management students with busy schedules, we prefer seeing movies in the weekends, and our workload and stress level on a day definitely has an impact on how much we enjoy a movie. Thus, we anticipated a meaningful relationship there and wanted to analyze it. Furthermore, the same day in a month corresponds to a different day of the week every year. For example; while 28/07/2018 is a Saturday, 28/07/2016 corresponds to a Thursday. So, we used a function on Excel that converts the information on date (year, month, day) to the day of the week, found the day each movie way released, and added this information to our dataset.

Using this variable, we created another variable dividing months into 2 different categorical representing the beginning and end of the month. For example, the beginning of the month variable contained the first 14 day of a month.By doing this, we tried to analyze whether there is an effect on rating if a movie is released at the beginning or at the end of a given month.

Considering that the ratings of the movies that we are trying to predict are all in English, we decided to include a dummy variable for English language, instead of including all language levels. The number of English movies in the dataset is a lot higher than other languages, so we had enough sample size in both English and Non-English buckets to inform our predictions.

Finally, when we went over the categorical variables in the dataset, we realized that some of them had tens of and even hundreds of different levels. For example, the variable “main director name” had more than 3000 levels alone. Imagine creating a dummy variable for each and every producer out there! Plus, putting each level into the model would be both very inefficient and meaningless, since some of the categories only had a few occurrences. While Steven Spielberg directed 23 movies, Nikolaj Arcel only directed 1 movie. Consequently, predictions regarding the rating of a

movie that Arcel directed would be far less accurate, because its sample size is not large enough to train a decent model. So, we decided to create a new variable called “director experience” with 5 levels and grouped the directors based on the number of their occurrences in the dataset as we believed it represented their level of experience in the business . We tried to divide the occurrences uniformly across buckets, such that each would have close numbers of directors in them. We repeated this process for main production company, production country, producer name and editor name. The exact distribution for each level can be found in Appendix 3. This technique allowed us to make intuitive assumptions with occurrence and experience, and group categories with few occurrences together to form acceptable sample sizes.

d. Analysis and Interesting Findings

After making necessary adjustments on our data and preparing it for analysis, we proceeded with analyzing the variables in our dataset and scrutinized their relationships with IMDB scores. We started with running individual regressions with IMDB score and certain variables that we thought would be interesting to look at, and found some interesting results.

We predicted a very strong relationship between the IMDB rating and the star-meter of the actors, especially the main actor. Our analysis did not support such predictions, however. From the graph in Appendix 4 that shows the correlation between star-meter and IMDB scores, it is understandable that the relationship is not linear, but relying on graphs for drawing insights is not always enough. So, we decided to run a regression to see what p-value gives to us in terms of linearity. After running the proper regression, shown in Appendix 5, we found out that the star-meters for actors are statistically insignificant with very high p-values(0.392) and very low adjusted r-squares. This contradicts with the intuitive logic that movies with popular actors have higher IMDB ratings (Low star meter indicates high actor popularity). After getting these results we did some research about this variable. The star-meter is a ranking system that is updated weekly. Each time someone visits an actor’s IMDB page, it has a positive effect on that actor’s star-meter. For instance, this October the movie ‘Joker’ got released and it was a big hit as a result of successful marketing, positive reviews and international awards. This is the reason why Joaquin Phoenix has the lowest star-meter (highest popularity) in our data. It also explains the high variance of star-meter when plotted against IMDB rating. When an actor is very popular, the average interest of users over time probably remains constant. The star-meter of Leonardo

Dicaprio is consistently low, because he has been starring in a movie once in every two years since 2000. On the other hand, with less popular names it's more common to have sudden and sharp increases or decreases in the number of people who visit their page. If we think of a movie that was released 3-4 years ago which has a main actor that is not so popular, he/she might have had a high star-meter before but as this is a system that changes weekly this actor may not be so popular right now, which will not reflect the impact on the IMDB rating.

As gender equality becomes an increasingly important topic of discussion in many industries today, we thought it would be interesting to examine the effect of actors' gender on the IMDB ratings. We had information on the gender of the main actor, second main actor and the third main actor in our dataset. It was surprising to see that movies with female main actresses have lower IMDB scores than those with male actors, in all three categories! Although the difference in IMDB scores are not very high (around 0.1) we thought that it was worth mentioning, since we got similar results for all three main actor categories with significant p-values. So, the actor's gender seems to make a difference. In addition, the bar charts on the gender counts for main actors and directors (see Appendix 6) show some interesting findings. We were surprised to see that the majority of the main actors in our dataset is male, but we were truly shocked to see almost all of the directors are male. With further analysis we found that only 3% of the directors in our dataset are female. These results made us come to the conclusion that the movie industry still seems to be dominated by men. We decided to go even further with our analysis on the gender of the main actors by analyzing some interaction variables. The first one we observed is *main_actor1_is_female*main_actor2 _is_female*, from our initial finding on gender we were surprised to notice a significant positive coefficient of 0.156 when regressed on the rating. This might suggest that when a movie is tilted toward a specific gender the ratings increases. Another interesting finding results from the interaction term of genre_adventure*main_actor1_is_female, this term resulted in a significant negative coefficient of 0.5 when regressed in the model on the rating of the movie.

We were quite frustrated when we observed the relationship between the budget and rating. Thinking of all the great movies with extremely high budgets such as Avatar and Harry Potter, we were so sure that there would be a strong positive relationship between the two variables. As explained earlier, we even put in extra effort to predict the missing budgets in the dataset before we began our analysis. That is because we thought it was an important predictor in the sense that it would explain a good amount of variance in IMDB ratings. Yet, we were proven wrong about

both of these assumptions when we plotted the relationship between them and did a regression analysis (See appendix 7). There was no remarkable correlation between the budget of a movie and it's IMDB rating, with a very low adjusted r-squared. (Appendix 8)

An interesting relationship was found between the movie duration and IMDB rating. We anticipated a negative relationship between them, considering our bad experiences with awfully long and boring movies. However, when we plotted their relationship **(see Appendix 9)**, we saw that the IMDB rating increased as the duration increased. It is still important to point out that as the movies get too long (longer than 200 minutes) the visual evidence in the scatter plot shows a downward trend, indicating that the relationship is not linear. To take this analysis one step further in order to understand qualitatively this relationship we looked at the producer of the movie and if their skills were the source of capturing their audience for such a longer movie length**.** Therefore, looking at Appendix 9, showing the relationship of a movie duration with its rating for two producer skill level, a lot of high rated and long duration movies where produced by skilled persons. This lead us to create an interaction variable *producer_experience*movie_duration* which ended up being significant. Personally when I watch a movie knowing it was realized by Francis Ford Coppola I am a lot more attentive and kin to appreciate the movie, no matter the length.

_

Interestingly, from running a regression of the variable day_of_week (Monday, Tuesday ,Wednesday…) we realized that the most popular release day was Friday(Appendix 10). Friday also had the greatest significance as its p-value is the smallest (day_of_week 5 in **Appendix 10**). Its effect on the dependent variable is also the largest as the b5 is equal to -.33, therefore, contributing to a decrease of -.33 on the rating of a movie. In fact, from **Appendix 11**, showing the boxplot of each day distribution we see that Friday has the lowest mean and most of the outlier below the first quartile. There really seems to be negative pressure on the rating when a movie is released on Fridays. Our belief is that customers have strong expectations about a movie when going on a Friday as they finish a long week of work and really hope for a decent movie. When not satisfied, they do not hesitate writing terrible reviews and thus are more sensitive, explaining why there is so much more outlier below the 1st quartile on Fridays. The same thing applies for great movies because customers are more sensitive on Fridays they tend to exaggerate their rating when the movie is great which explains why there is more outliers above the 4th quartile on Friday than any other day of the week although less frequently than very bad reviews

Model Building and Calculations

We wanted to start building our model by looking at variables which did not have a linear relationship with IMDB rating. After testing for linearity of each variable, we found that *duration*, *number of actors*, *budget* and *number of producers* are not linear. Then, we looked at the best fit lines of these variables when plotted against IMDB ratings and concluded that we needed to use polynomials for them rather than splines based on their respective results. Next, we ran K-fold tests to find the most suitable polynomial degree for each variable. Specifically, K-fold tests told us which polynomial degree would yield the lowest MSE for each selected variable, so that we could use the findings as a benchmark while assigning the degrees for them in the final model. However, it is important to point out that K-fold test falls short when finding the best degree, because it does not show the significance of an improvement in MSE when the polynomial degree is increased. So, we ran an ANOVA test following the K-fold test, because we did not want to decrease our degrees of freedom for an insignificant improvement in our model. The results indicated that the "best degrees" are degree 4 for the budget, degree of 2 for the number of actors, degree 5 for number of producers and degree 6 for the duration.

The next step was to initiate building our model. We ran a multiple regression starting with variables that had the highest correlation with IMDB rating, regardless of the degree of their relationship. (See Appendix 12 for the Correlation Matrix). The first model we built to get a sense of where we stand included only duration with a polynomial degree of 6; which gave us an MSE of 0.9. This was a good start, but as data scientists, we were not satisfied, and we knew we could do better.

Using that first model as a benchmark, we added variables one after another, starting from the continuous variables that had the strangest correlation with IMDB rating. If adding the variable decreased the MSE, we kept it in the model, otherwise we took it out and added the one with the next highest correlation. Once we chose all the continuous variables that improved our model, we moved on to categorical variables, and repeated the trial and error process of fine-tuning.

Once all variables were tested, we reached the lowest MSE of 0.73. This means that for any movie, given that we have information on the variables that are in our model, we can predict its IMDB rating within a range of +/- 0.83.

However, as dedicated students of Juan Serpa, we were still not satisfied with this result. We wanted to incorporate the effect of interactions mentioned earlier in the Data section and other logical interactions. We used our intuition to come up with interesting interactions and continued our trial and error approach; and managed to decrease the MSE to 0.63.

RESULTS

As a result of all of our analysis and trials, we reached the lowest MSE at a level of 0.637134 using K fold cross validation test at k designated at 74 (from rule of thumb: square root of number of observations). Below, you can see all the variables included in the final model with their respective coefficient (**See Appendix 13)**. After finding the best model, we tried to detect possible issues with our model that might affect the results. Our tests indicated heteroskedasticity in the model as the p value of the non-constant variance test was very low. After correcting heteroskedasticity, we moved on to identifying outliers and removed them to obtain a more accurate model. These corrective measures are required in such models where there are thousands of observations, because a skew in our data due to outliers would reduce the quality and integrity of the model. After removing outliers; we ran our model one more time. Following these two corrections the final MSE obtained was 0.63.

Some interesting results were encountered when we ran the summary of our regression model. Budget, along with duration and the genre comedy had the largest effect on the rating. Another interesting finding was the relationship between action movies and the rating. Before doing the analysis, we thought action movies would have a high tendency of getting better ratings than other genres; however the results show otherwise (coefficient estimate of -2.65). We did some research to find the reason behind such relationship, and saw that there are lots of action movies with bad production setup resulting from bad technology usage or just boring scenarios lacking creativity. Perhaps the successful action movies such as Fast and Furious set the expectations too high, and it gets harder to impress the audience with mediocre stundts and animations. Horror movies also have a seem to decrease the IMDB rating, with a coefficient of -0.5; which is understandable as they are often scary and not appealing to the general public. As per the director experience, only the expert level had a positive coefficient which means that the director experience makes a difference only when they are an expert.

Overall, we are pleased with our model's predictive power. With an MSE of 0.63, we are in the range where we can predict the ratings of upcoming movies within an error margin of 0.8.

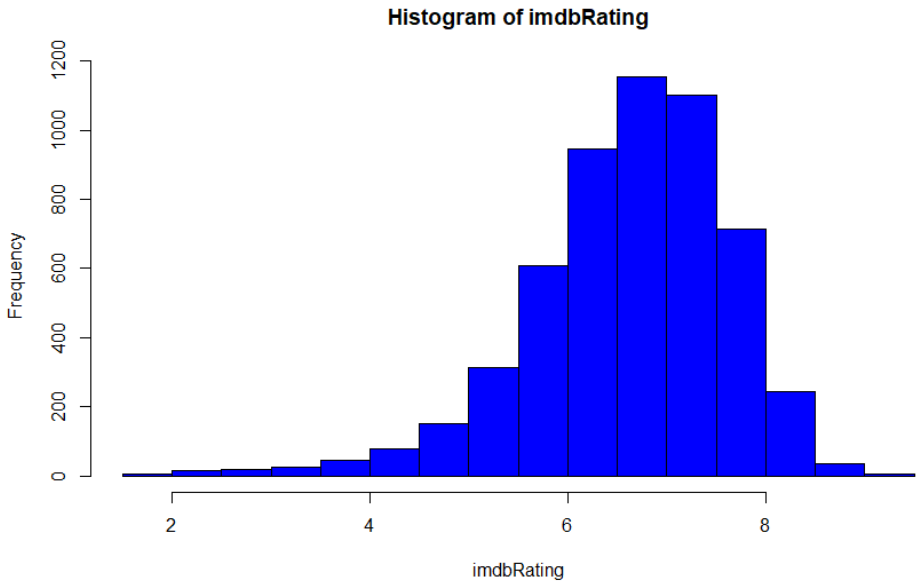
After these processes, we have the ability to calculate the ratings for the upcoming movies. For the predicted IMDB ratings, see(Appendix 14), the mean of our 12 predictions were 6.8390/10, which is close to initial data's mean, which is 6.66/10. We find this encouraging because our small size sample with has a great connection in terms of mean with the 5456 observations. Small sample (movies that are coming out) has a median of 6.82, and our dataset has a median of 6.8. We believed that using median over range would be a better choice in comparing the accuracy of the small sample data in relation to big data. Note that some of the movies that we had to predict did not disclose their budget publicly. Thus, we estimated those undisclosed values with the budget regression model we developed.

However, there are some caveats that has to be addressed. One of them is the budget forecast. Although we believe that the way we tackled the issue of non-available budgets is a great way that helped us to progress forward in our model as well as analysis, it still had a lot of room for improvement in terms of predictive power.Further steps could include fine tuning this specific predictive model. Another factor that has to be considered is the given data. The movies that are included in the data goes from 1933 to 2014, with no detailed description of movies that come out later. As the movies we're analyzing will come out in 2019, there is a possibility of error due to interpolation. People's preferences could have changed during these 5 years, and it won't be accounted for in our model. Moreover, we can think of other variables that can be interesting to analyze including unstructured data such as the number of tweets about a movie occurring before its release for example or the number of times it appears on live TV as part of interviews on tv shows to name a few.

APPENDIX

Summary of imdbRating					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.6	6.1	6.8	6.667	7.4	9.3

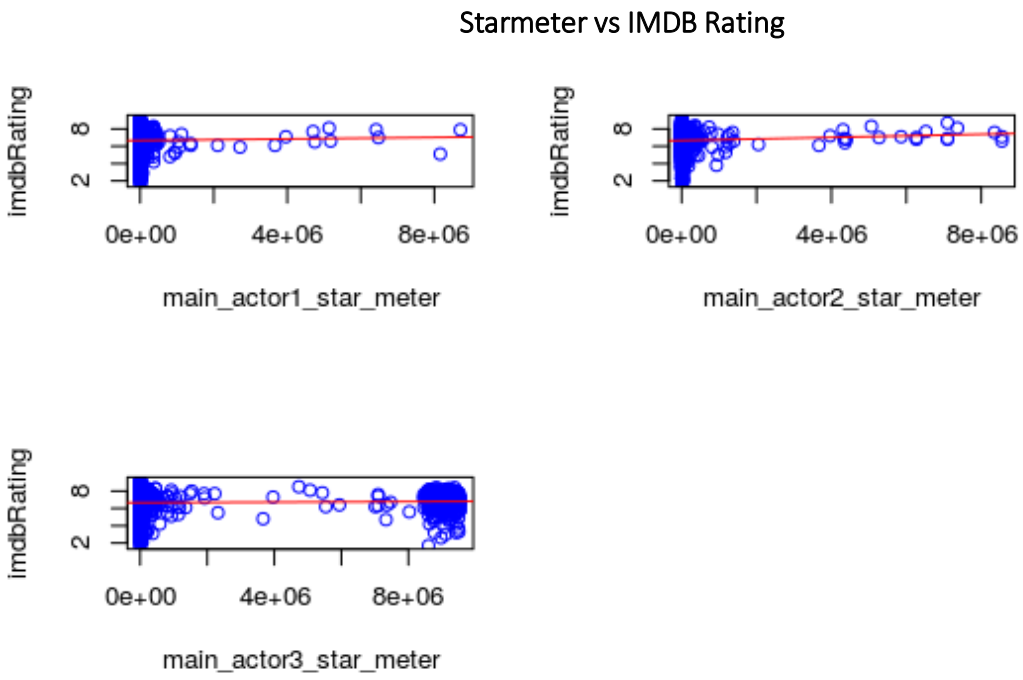
Appendix 1: Summary of imdRating of the movies on our dataset



Appendix 2: A histogram showing the distribution of imdRating

	LEVELS					
VARIABLES	beginner	novice	intermediate	advanced	expert	N/A
director_experience	≤1	≤3	≤5	≤10	>10	N/A
producer experience						
editor experience						
	unpopular		popular		very popular	
production company experience	≤9		≤100		>100	
production country popularity						

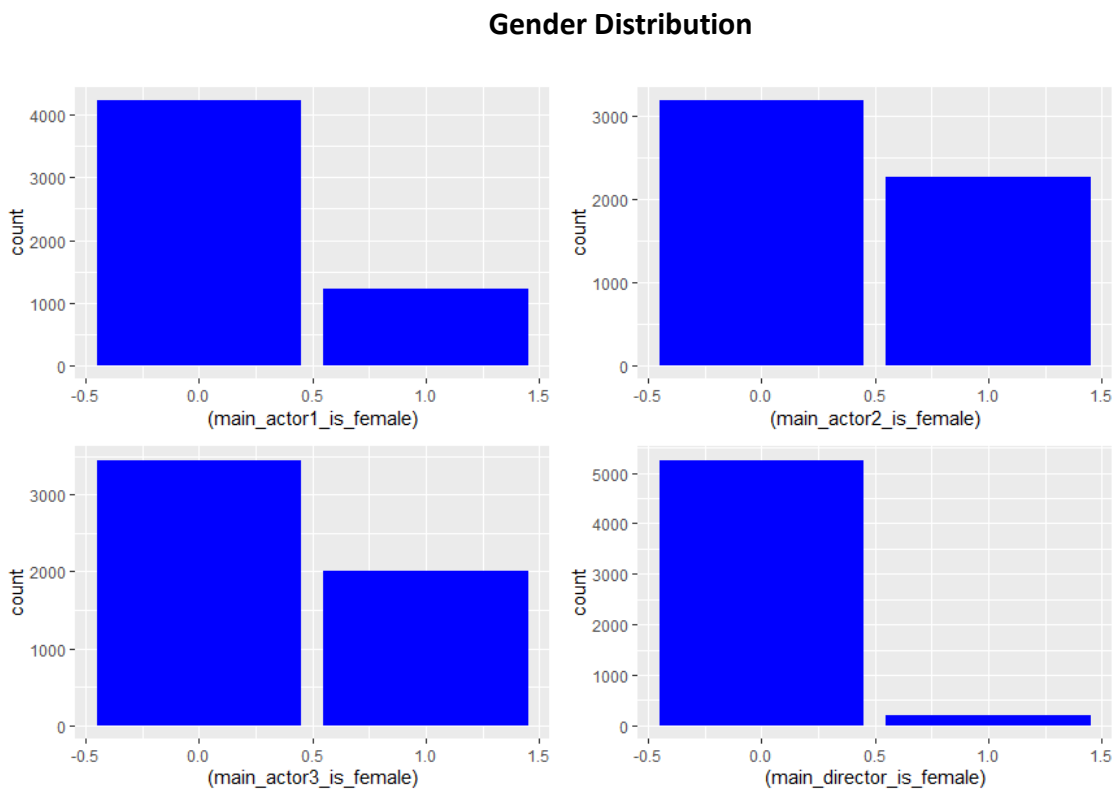
Appendix 3: A table showing the exact distribution for each level of variables that we created



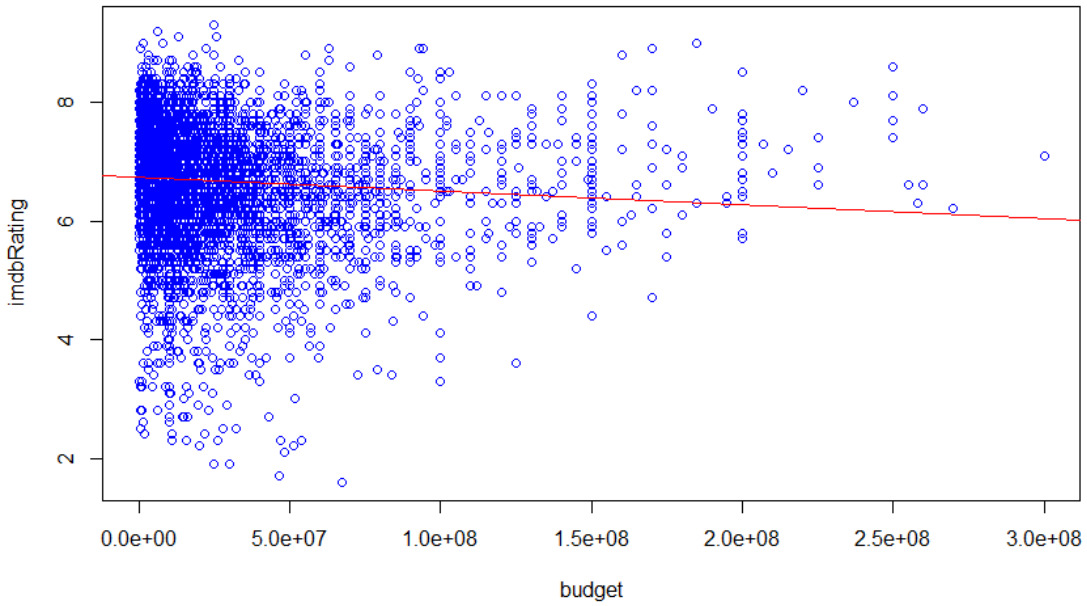
Appendix 4: A plot showing the correlation between star-meter of main actors and IMDB score

	Dependent variable:		
	imdbRating		
	(1)	(2)	(3)
main_actor1_star_meter	0.3921		
main_actor2_star_meter		0.0134*	
main_actor3_star_meter			0.0163*
Constant	6.666***	6.664***	6.659***
	(0.014)	(0.014)	(0.014)
Observations	5,456	5,456	5,456
R^2	0.0001	0.001	0.001
Adjusted R^2	-0.00005	0.001	0.001
Residual Std. Error (df = 5454)	1.005	1.004	1.004
F Statistic (df = 1; 5454)	0.732	6.116**	5.774**
Note:	*p<0.1; **p<0.05; ***p<0.01		

Appendix 5: Regression on star-meter of main actors



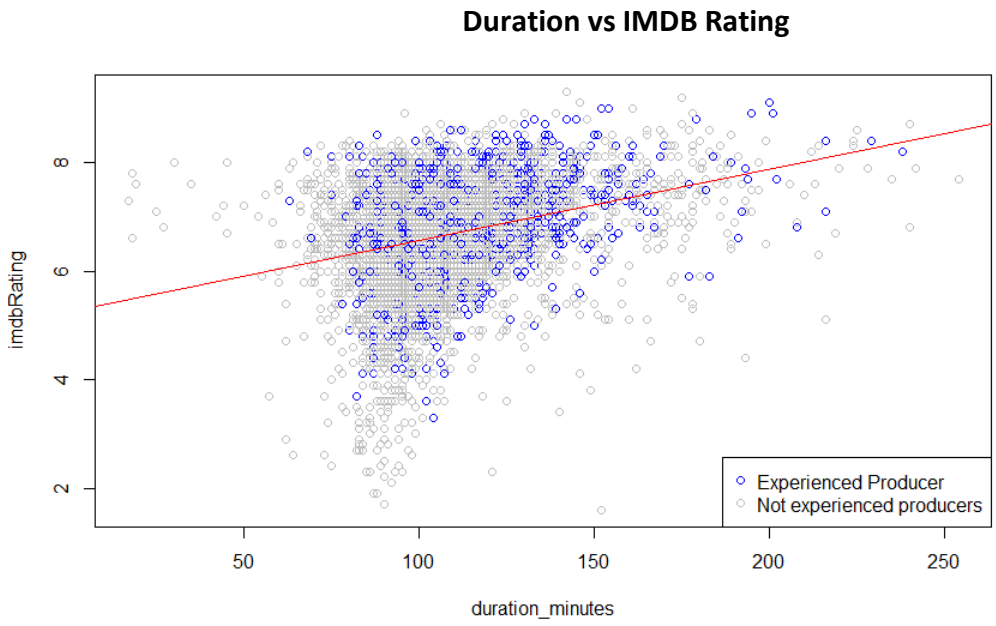
Appendix 6: A bar chart showing gender distribution of the main actors and directors



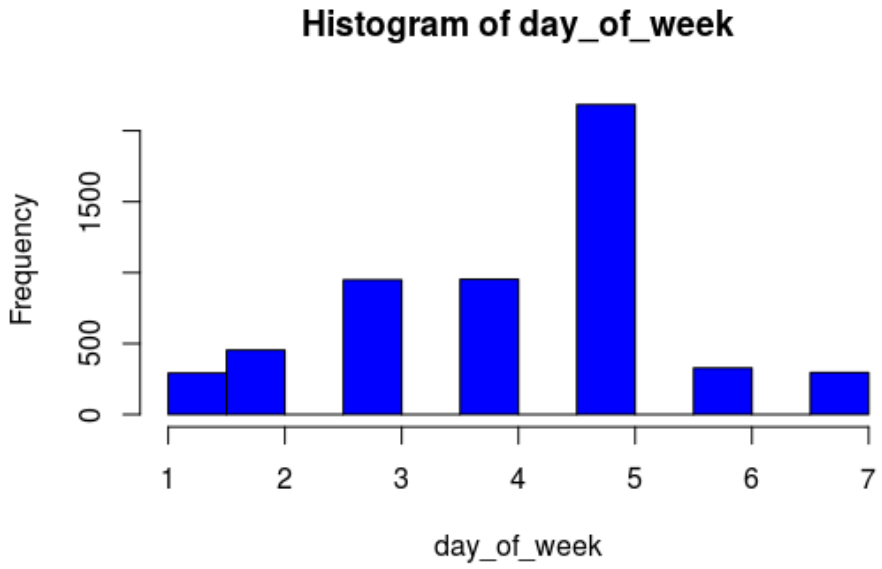
Appendix 7: A plot showing the lack of correlation between movie budget and IMDB score

Observations	5,456
R^2	0.006
Adjusted R^2	0.005
Residual Std. Error	1.002 (df = 5454)
F Statistic	31.019^{***} (df = 1; 5454)
Note:	*p<0.1; **p<0.05; ***p<0.01

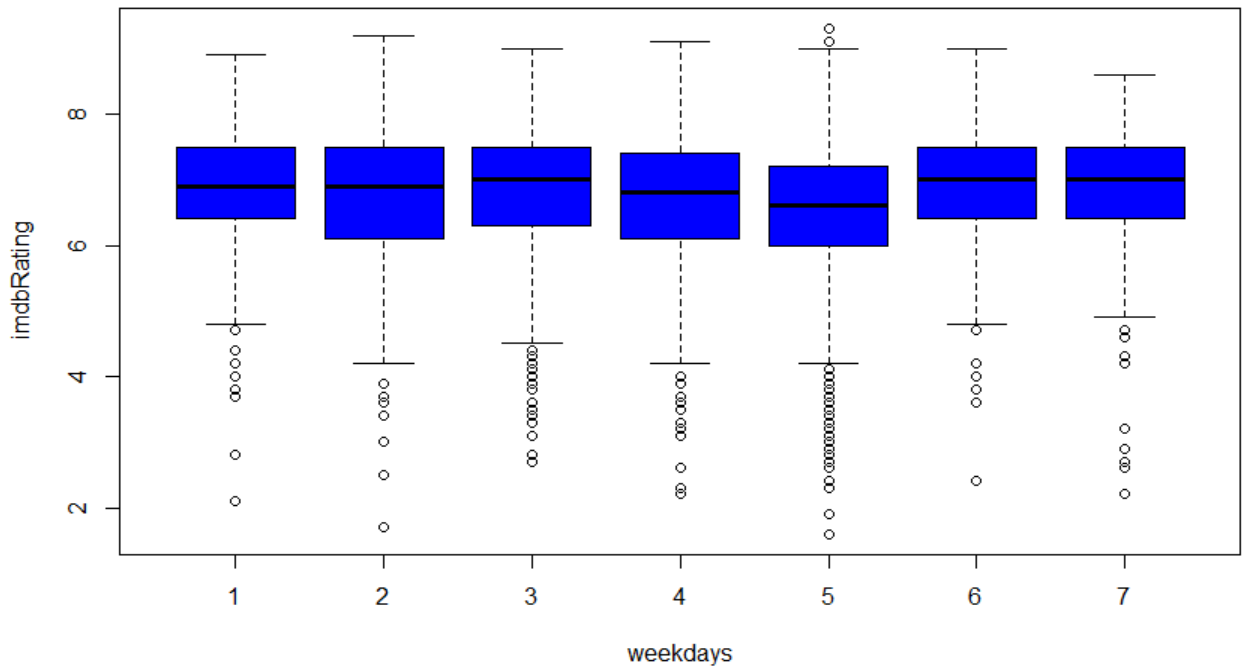
Appendix 8: Regression on budget showing it has a low adjusted R-squared



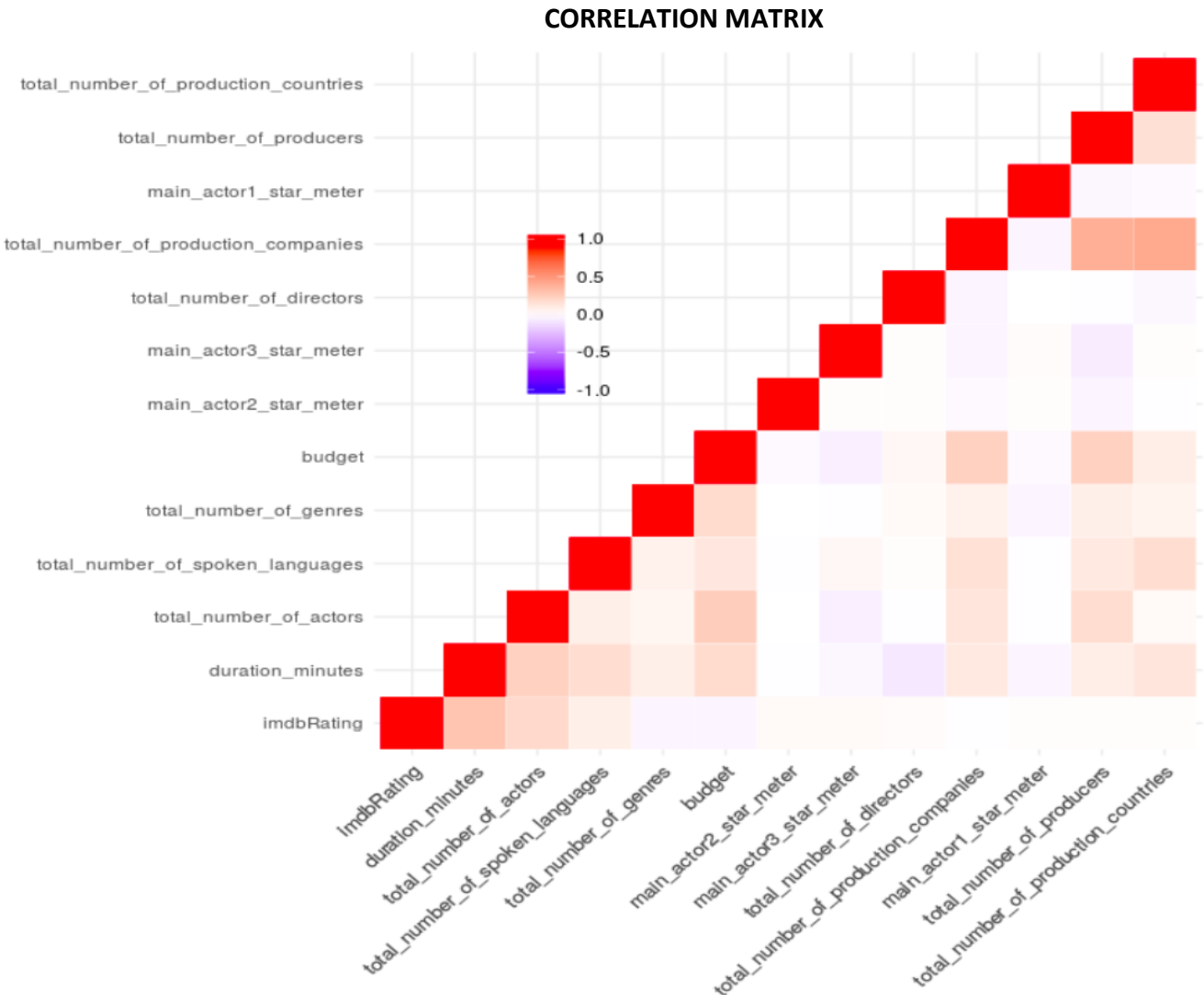
Appendix 9: A plot showing the correlation between duration and IMDB rating with an emphasis on producer experience



Appendix 10: A histogram showing movie releases distribution across the days of the week



Appendix 11: Boxplot of each day distribution



Appendix 12: Correlation matrix of all quantitative variables

Final Regression Results

	Dependent variable:
	imdbRating
poly(duration_minutes, 6)1	14.775***
	(0.926)
poly(duration_minutes, 6)2	-1.083
	(0.840)
poly(duration_minutes, 6)3	-4.678***
	(0.818)
poly(duration_minutes, 6)4	3.031***
	(0.808)
poly(duration_minutes, 6)5	1.627**
	(0.805)
poly(duration_minutes, 6)6	-2.425***
	(0.809)
poly(total_number_of_actors, 2)1	10.599***
	(0.849)
poly(total_number_of_actors, 2)2	-6.632***
	(0.807)
poly(budget, 4)1	-8.802***
	(1.059)
poly(budget, 4)2	9.910***
	(0.849)
poly(budget, 4)3	-6.445***
	(0.815)
poly(budget, 4)4	2.927***
	(0.802)
day_of_week2	-0.018
	(0.060)
day_of_week3	0.007
	(0.054)

Appendix 13: A summary of our final model's results (part1)

day_of_week4	-0.004
	(0.054)
day_of_week5	-0.107**
	(0.050)
day_of_week6	0.001
	(0.064)
day_of_week7	0.051
	(0.066)
director_experiencebeginner	-0.420***
	(0.037)
director_experienceexpert	-0.022
	(0.048)
director_experienceintermediate	-0.103***
	(0.039)
director_experiencenovice	-0.224***
	(0.035)
genre_action	-2.813
	(3.106)
genre_drama	0.227***
	(0.027)
genre_horror	-0.555***
	(0.041)
genre_documentary	0.907***
	(0.111)
main_spoken_language_is_english1	0.181***
	(0.033)
release_year	-0.004***
	(0.001)
genre_comedy	12.896***
	(2.196)

Appendix 13: A summary of our final model's results (part2)

main_actor3_is_female	-0.026
	(0.026)
main_actor1_is_female	-0.138***
	(0.033)
main_actor2_star_meter	0.00000
	(0.00000)
genre_documentary:main_spoken_language_is_english1	0.239
	(0.344)
genre_action:release_year	0.001
	(0.002)
release_year:genre_comedy	-0.007***
	(0.001)
main_actor3_is_female:main_actor1_is_female	0.073
	(0.054)
release_year:main_actor2_star_meter	-0.000
	(0.000)
Constant	15.476***
	(1.472)
Observations	5,446
Log Likelihood	-6,459.186
Akaike Inf. Crit.	12,994.370
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix 13: A summary of our final model's results (part3)

Movie Title	Predicted IMDB Score
1) Mickey and the Bear	6.367383
2) Noelle	6.784421
3) Atlantics	6.708934
4) Charlie's Angels	5.513276
5) Le Mans' 66	7.286282
6) The Good Liar	7.331983
7) The Report	7.399469
8) Waves	7.632502
9) 21 Bridges	6.685861
10) Beautiful Day in the Neighbourhood	6.886171
11) Dark Waters	7.277594
12) Frozen II	6.13746

Appendix 14: Our predicted IMDB scores for the movies coming out this season