## Part I

## 

## Domain situation

The film industry can be defined as all the companies, [studios](https://www.collinsdictionary.com/dictionary/english/studio), people etc involved in making [commercial](https://www.collinsdictionary.com/dictionary/english/commercial) films collectively. According to the 2018 IBISWorld report, the film industry generated $136 billion of revenue in 2018, with countries such as the United States, Canada, China, Japan, the United Kingdom and India leading most of the revenue and production. The top movie genres within the United States, the largest film industry, are adventure, action, and drama in order of box office revenue from 1995 to 2019. This is reflected by top grossing films such as Avengers:Endgame, Avatar, and Titanic, all of which belong to at least of the above genres. Movies, television shows, and movies are common leisure activities with data from 2000 to 2018 showing that 70% of Austrlialians reported going to the cinemas at least once, with the average frequency being around 7 times a year (Screen Australia, 2018). In recent years, however, there has been a sharp rise in online streaming sites such as Netflix and Amazon, that have changed the way consumers interact with traditional television and film. Dispute PwC predicting a [“terminal decline](https://www.theguardian.com/media/2010/nov/29/dvd-industry-sales-slump-blu-ray) for DVD and blu ray sales,” cinema revenue is forecasted to remain largely unaffected as those who use streaming services also contribute more to box office sales (Sweney, 2017).

Questions we are looking to explore

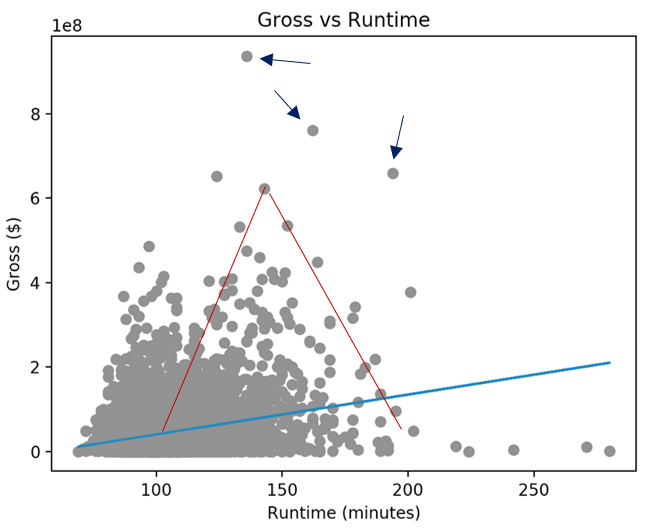
The film industry loves numbers, especially when the numbers relate to money; despite all the artistic prowess of a film, the gross revenue is the single quantitative measure of a movie’s success. We wanted to investigate the relationship of gross revenue in relation to the various aspects highlighted in our dataset. We intended to find out whether aspects such as budget, runtime and viewer score lead to higher box-office figures. This investigation is important for our stakeholders in knowing if there is a predefined recipe that directors could follow and achieve big gross revenues.

‘gross’ in the dataset is the total revenue generated by the film, and is represented in USD ($)

How does the runtime of the movie affect the gross revenues generated by it?

‘runtime’ in the data is the total duration of the film in minutes. We hypothesised that the relationship between runtime and gross revenue would replicate a parabolic shape, whereby we expected there to be ‘sweet-spot’ for the runtime which would yield the highest revenue. We expected the viewer to perceive the runtime of the movie as an indicator of the type of movie they are going to see; movies that are usually shorter in length are usually factual, whereas longer movies are usually fictional. It is important to remember; these are simply our expected findings for the runtime and gross revenue relationship.

In order to test our hypothesis, we denoted the runtime as the independent variable and the gross revenue as the dependent variable, and plotted them on x-axis and y-axis respectively. We decided to use a scatter plot to present our findings because we were representing two quantitative variables.

The initial observation from the graph seem to falsify the hypothesised trend, there does not seem to be a parabolic shape to the graph. However, if we were to remove the clutter of movies in the <150 mins range, patterns begin to appear. Although there isn’t a balanced parabolic shape in the diagram, there is a distinct rise in the gross numbers for movies within the 100 – 200 min range. The top 3 highest grossing movies, namely Star Wars: The Force Awakens, Avatar and Titanic run 136, 162, 194 mins respectively. Perhaps this is the sweet spot for the duration of a movie, in order to maximise revenues? 

However, a possible confounding factor is that this range is far too big to be considered a guaranteed revenue maximiser; most movies usually fall within this range in regards to duration. The line of best fit of this graph is marked by the blue line, and the correlation coefficient of the graph is 0.063. The low coefficient suggests a very-weak correlation between runtime and gross revenue.

Therefore, given the findings from the graph, the hypothesis must be false, there must be other factors that lead to high gross-revenues.

Another interesting observation made from the findings is that movies that are longer than 200 mins in duration have a significant impact on the box-office performance, with all the movies in this range performing significantly worse than any other movies.

## Origin of data

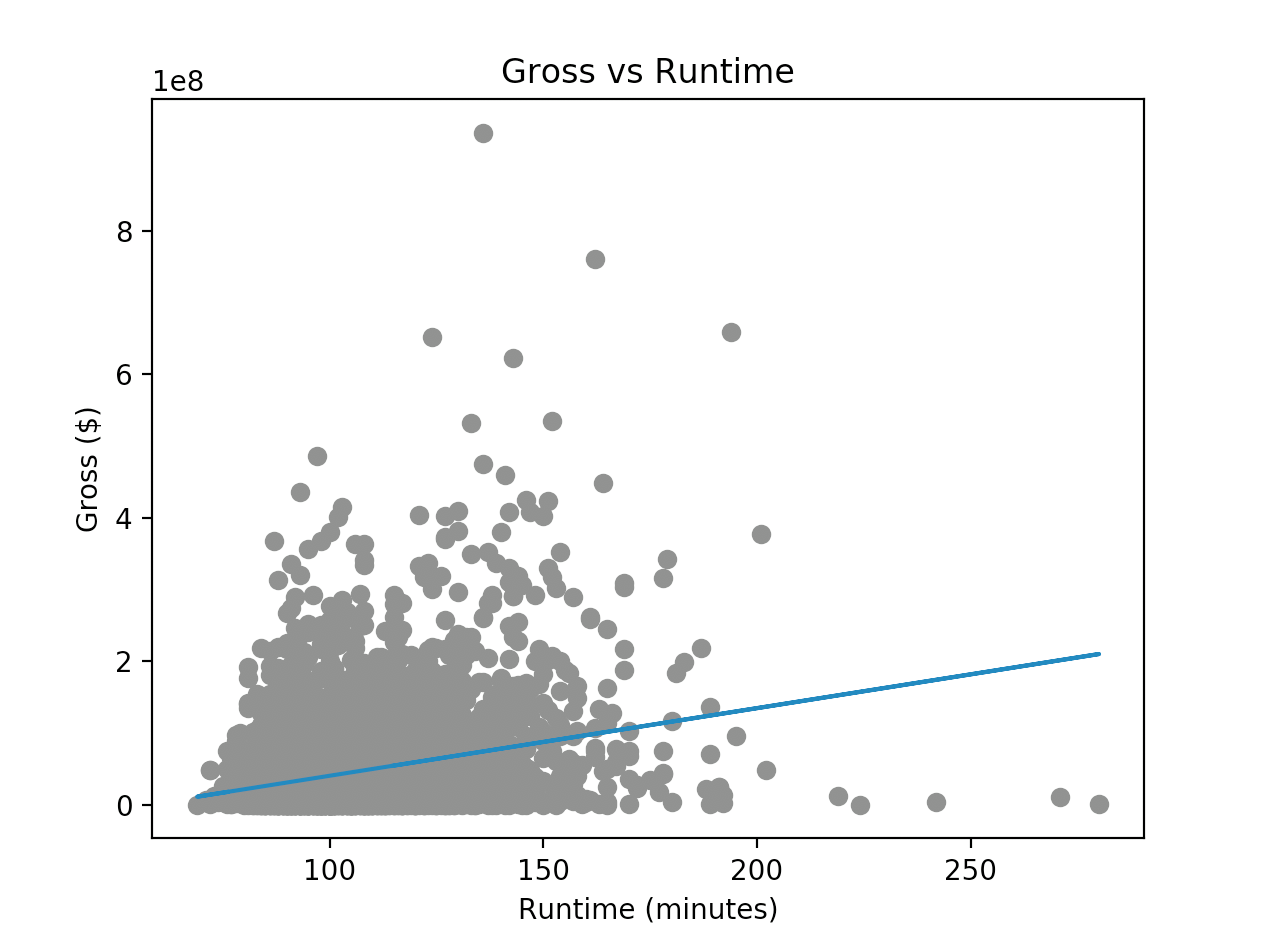
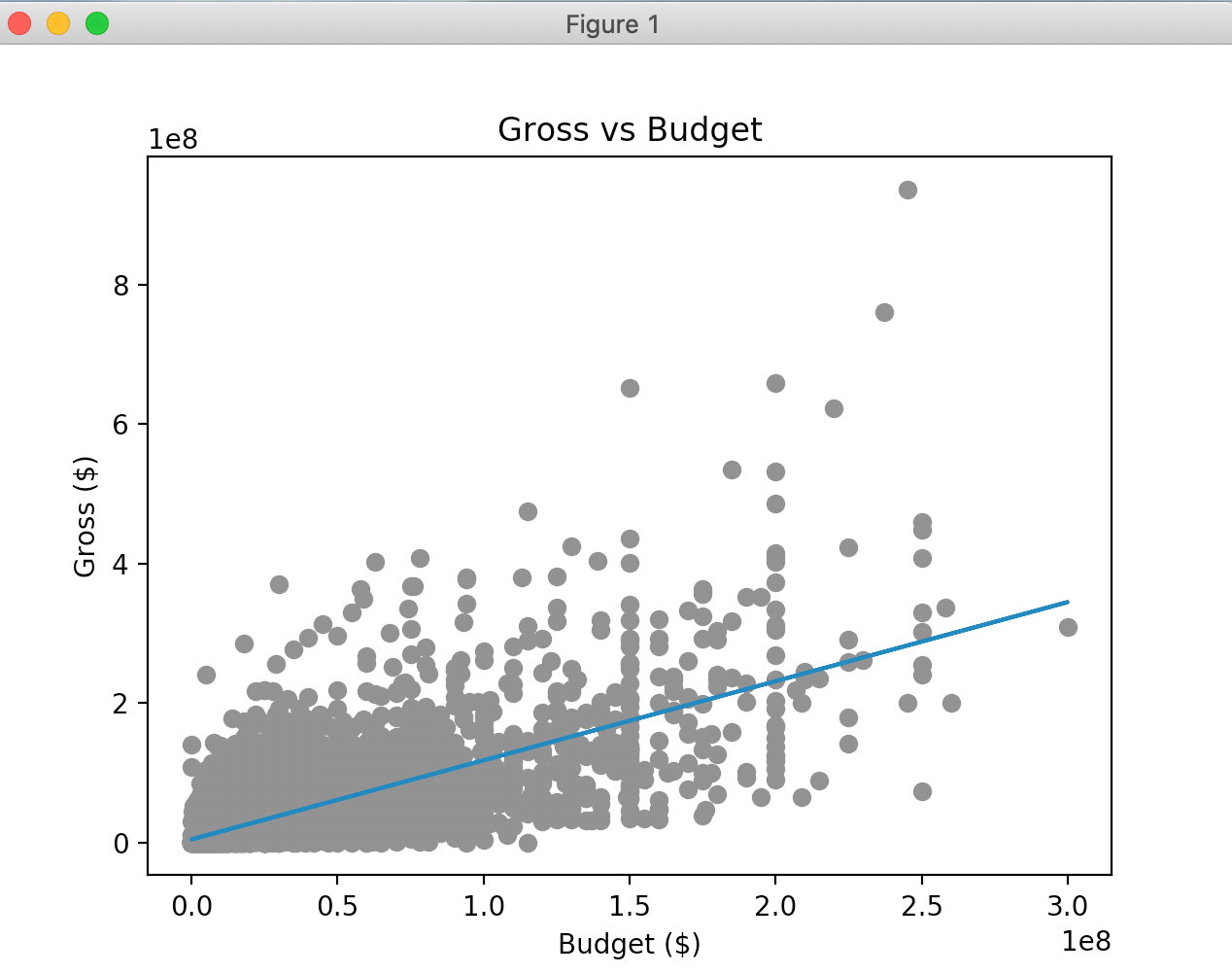
To answer our questions, we use a dataset obtained from the IMDb website, a public online database, containing information in regards to movies, TV, and celebrity’s content. The raw dataset contains 6821 movies released between 1990 (inclusive) and 2016 (inclusive), consisting of 15 columns where 6 were quantitative discrete data (such as the budget and earning of the movie), and the remaining 9 were qualitative nominal data (such as names of the director, production company and lead actor/actress) (Grijalva n.d.). We decided to filter out movies with no input budget in stage 1, leaving us to work with the remaining 4639 movies.

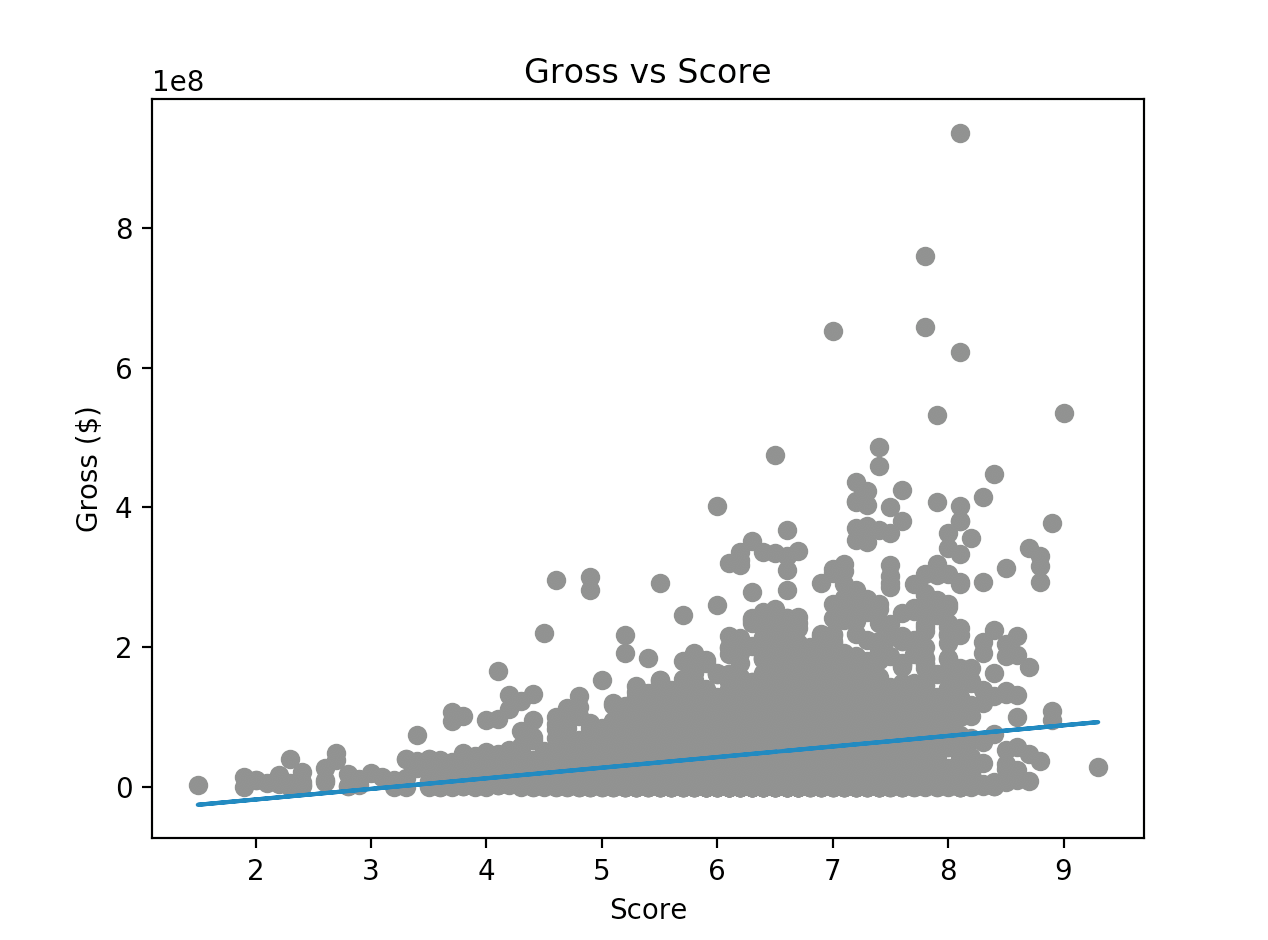
According to the IMDb conditions of use, any users all allowed copy, store and manipulate the information on this site only for non-commercial purpose (IMDb Conditions of Use n.d.). For this assignment, the use of the dataset for academic purposes falls within the given rights and does not violate licencing rules. Additionally, the dataset contains only public information standards for most movies. Hence, there is no risk of a breach in personal privacy due to storage, manipulation, and use of the dataset.

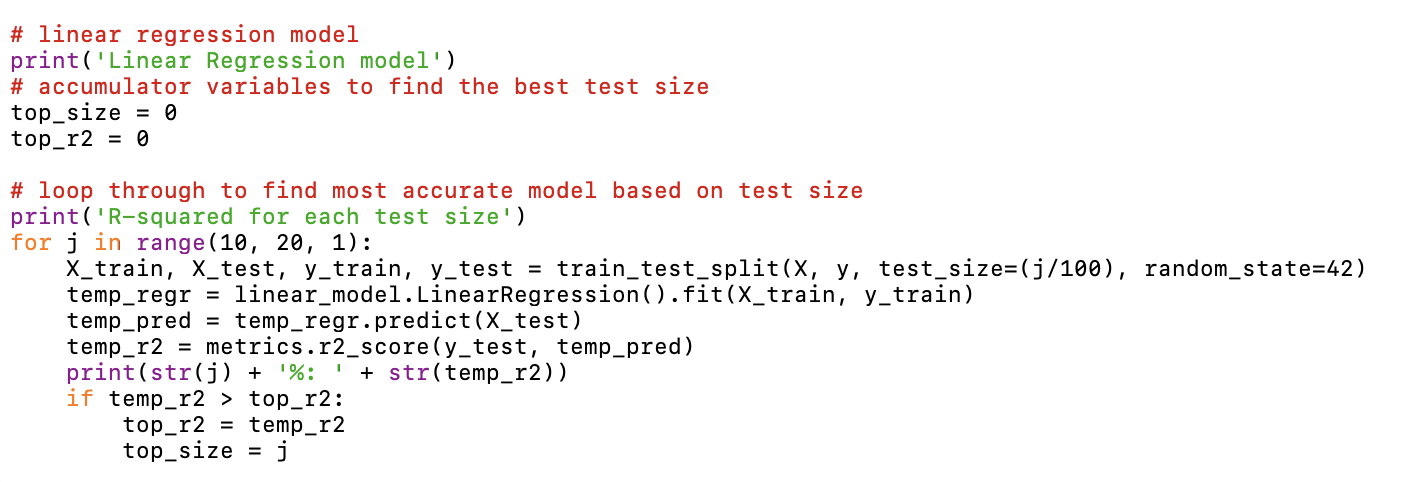
Part III

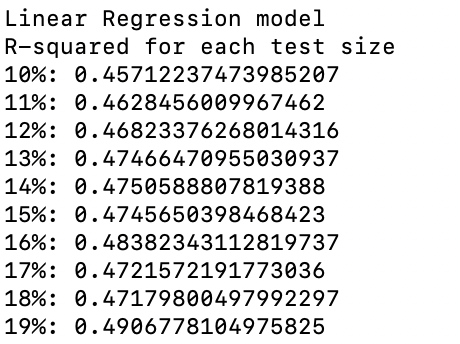
As stated above, we wanted to explore the effect of three attributes, budget, runtime, and score, on the gross of movies in US dollars. Because we have clean data about the movies industry, we used supervised machine learning to create a predictive model. More specifically, we used two machine learning regression algorithms, linear regression and k-nearest neighbors (k-NN), as our output variable is numerical.

We started with the linear regression algorithm. Typically, this model works best when data has a linear relationship, but it is hard to understand the relationship between four variables in a graph. We used pyplot from the library matplotlib to graph the pairwise relationships and the library sci-kit learn to analyze their linearity.

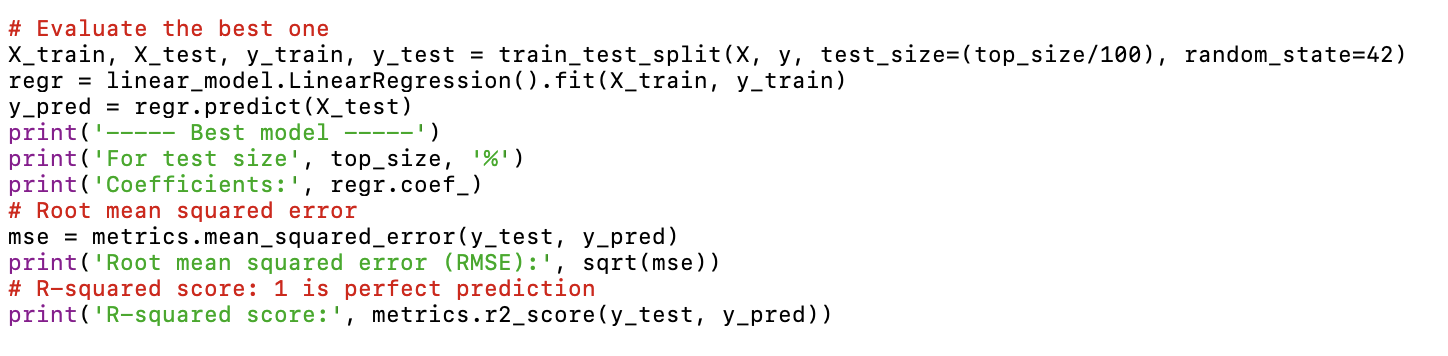


Overall, we found that the relationships were not necessarily linear, but we were curious to see how well a multiple linear regression model would predict the gross. We imported linear\_model and metrics from scikit learn, train\_test\_split from model selection in scikit learn, and sqrt from math, then proceeded with linear regression model calculations. To determine the best amount of data to allocate for testing, we created linear regression models for test sizes 10% through 20% using a for loop that added 1% each time. Accumulator variables recorded the test size for the highest R-squared value (1 is the best). 

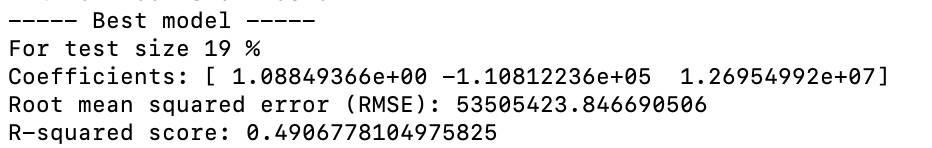
The code above gave us the following results:



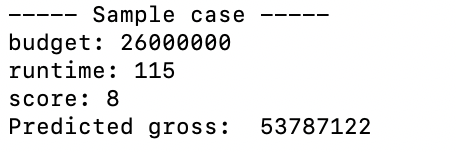
We then used the best test size for our real linear regression and calculated the coefficients of the multiple linear regression (there are 3 for 3 input variables) along with the root mean squared error and R-squared values. This is where we used metrics from scikit learn.



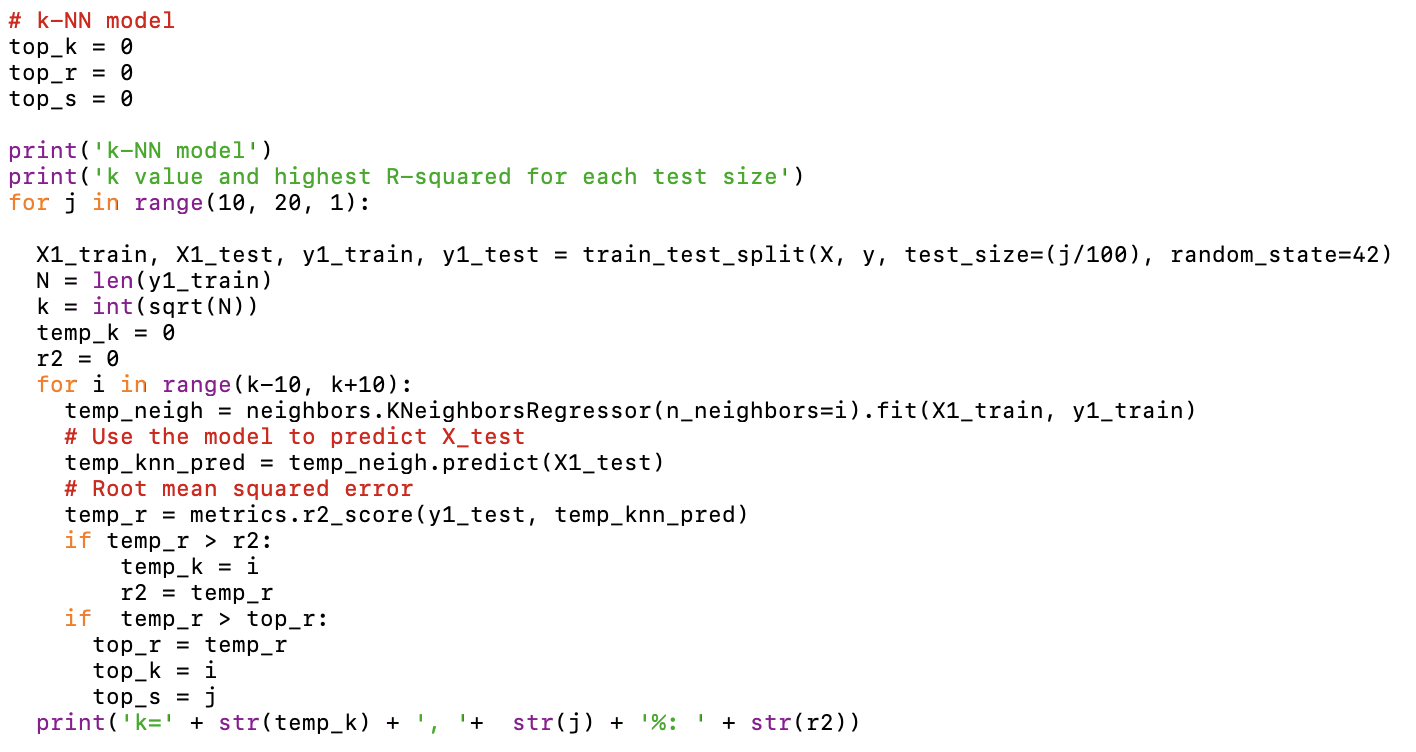
The code gave the following results:



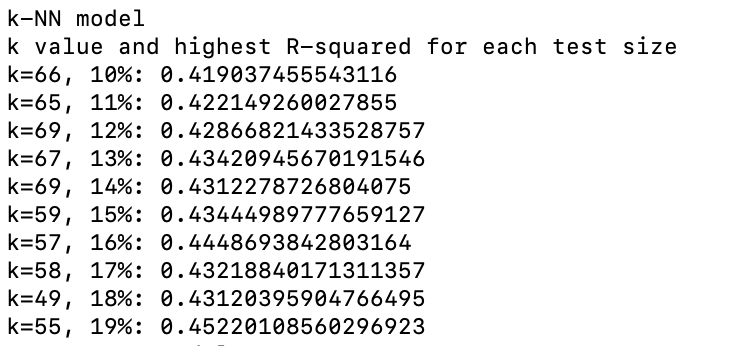
To be a good model, we would want a small root mean squared error and an R-squared score close to 1. Our error calculations revealed that our data does not have a strong linear relationship, like we predicted. Even so, we tested our model with a sample case and found the following results:



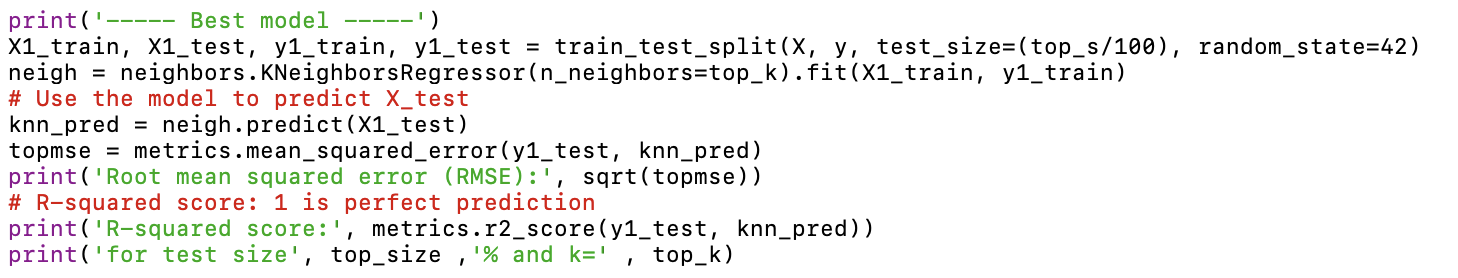
We also tried a different regression algorithm, k-NN, because it does not require linearity. For this, we had to import neighbors from the scikit learn library, and we used metrics, train\_test\_split, and sqrt that we already had from our linear regression. Just like our linear regression, we used a for loop to determine the best amount to allocate for testing, but we nested a for loop inside of that to determine the best k, which is the amount of nearest neighbors. The k chosen for k-NN should be around the square root of the amount of data, so we made predictive models using the range k-10 and k+10 iterating by one. We used accumulator variables to record the highest R-squared value and the k and test size that created the model for that highest R-squared value.



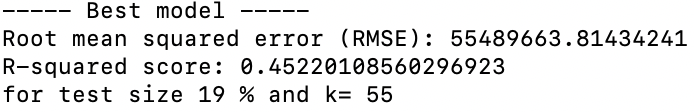
This code displayed the k value that calculated the highest R-squared value for each test size.



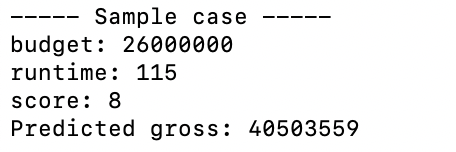
We then used the k value and test size that had the highest overall R-squared value to create the real model and make error calculations.

****

The code resulted in the following:



This means that with 19% of input data allocated for testing and 55 nearest neighbors, there is the least amount of error. However, the root mean squared error is quite large and the R-squared score is not close to 1 so the data does not have a strong relationship. We tested our predictive k-NN model with the same sample case as the linear regression and found:



Thus, the predictive models are pretty close. It is interesting that the two techniques we used both have similar errors. We expected that k-NN has stronger results because the data is not linear, but linear regression had less error. However, they can both be used to show that there is not a strong correlation between budget, runtime, and score and gross. Furthermore, the loops showed that for both of the algorithms allocating 19% for testing gave the best results. This would make sense because there is more data for the predictive model to test and fit, but it does lead to the possibility that the models used are overfitted.

**References**

IBISWorld (2018 August). [Global Movie Production & Distribution Industry: Industry Market Research Report](https://www.ibisworld.com/industry-trends/global-industry-reports/other-community-social-personal-service-activities/movie-production-distribution.html). Retrieved from <https://www.ibisworld.com/global/market-research-reports/global-movie-production-distribution-industry/>

Clark, T., Lynch, J. (2018) The 10 highest-grossing movies of all time, including 'Avengers: Endgame' Retrieved from

<https://www.businessinsider.com/highest-grossing-movies-all-time-worldwide-box-office-2018-4#D8VMLO3swidjyWkQ.99>

*IMDb Conditions of Use* n.d., Retrieced from <https://www.imdb.com/conditions>

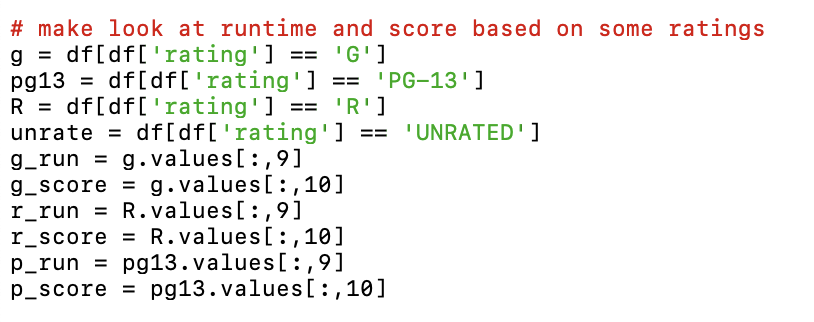
Grijalva, D n.d., *Movies Industry,* electronic dataset, kaggle.com, Retrieved from <https://www.kaggle.com/danielgrijalvas/movies>

Screen Australia. (2018). *Cinema audience attendance patterns.* Retrieved from <https://www.screenaustralia.gov.au/fact-finders/cinema/audiences/attendance-patterns>

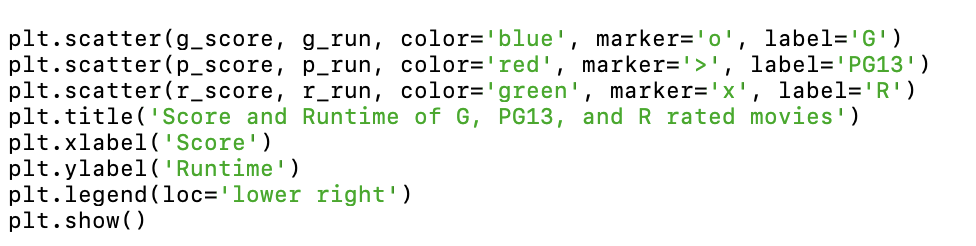
Sweney, M. (2017). Netflix and Amazon 'will overtake UK cinema box office spending by 2020’. Retrieved from <https://www.theguardian.com/media/2017/jun/14/netflix-amazon-uk-cinema-box-office-film-dvd-blu-ray-pwc>

How does rating play a role in score or runtime of movies?

Ratings of a movie determines the audience. For example, anyone can see a G rated movie, but an R rated movie is restricted to adults. It would make sense for them to have different scores and even runtimes, as it is more likely for an adult to vote for the score and an adult can stay focused in a longer movie. To analyze this, we looked specifically at three ratings: G, PG13, and R. After reading the clean csv file, we used pandas functions to slice the data.



Next, we used pyplot from the matplotlib library to plot the runtimes and scores on a scatter plot. Different colors and markers are included to represent the different ratings and improve readability.



When we run this, we get the graph below. Looking at it, we can see that the longest movie is rated R, and the next ones are rated PG13. The shortest movies are rated G, which is as expected. The highest rated movie is also rated R, but the lowest rated movies are PG13 and R. Overall, though, the majority of the movies are grouped between a score of 4.5 and 8 and runtime of 80 to 150 minutes, so the outliers tell us the most about the relationships.

