St. Joseph’s Institute of Management (JIM)
A Jesuit Business School
St. Joseph’s College (Autonomous), Tiruchirappalli 620 002

GOVINDH J
(18PBA131)

Prediction of movie rating using machine learning techniques

Master Thesis
July 4, 2019

Faculty Guide: Dr. A. PAPPU RAJAN

ABSTRACT

Every week, a number of movies are released worldwide. There is a lot of data available on the internet related to the movies, because of the big quantity of data available, it is an interesting subject of data mining. In the latest trends, the area of generating predictive models using machine learning has grown in size. Prediction in data mining is to identify data points purely on the description of another related data value. The prediction in data mining is known as Numeric Prediction. Movie prediction is one the popular area in the data mining techniques. Since, movie prediction is very helpful for various members such as producers, movie cast people, and movie viewers.

The purpose of this study is to examine whether the movie rating can be predicted after learning the relationship between the rating and the movies various attributes from an IMDb movie dataset. This was done by building a backward stepwise regression model (machine learning algorithm) with metadata retrieved from the Internet Movie Database (IMDb) and Rotten Tomatoes (RT) websites.

There are around 10 regression models were built with backward stepwise elimination method that has higher p-value and trained in the movie dataset. The final model developed herein has a very strong predictive power and got an accuracy level of 80.7% with a confidence interval of 95%. This study managed to correctly predict the IMDb rating for a particular movie with an accuracy of about 80% with a confidence interval of 95%. This study also proved that there are strong association on certain specific attributes of a movie drive to success. The accuracy of the predictions can further be increased with a larger dataset with more attributes.

LIST OF FIGURES

CHAPTER I

1.1 INTRODUCTION

Movie is a motion picture or film produced for the purpose of entertainment that tells a story. Movies are greater source of entertainment and people are crazy about movies. Every year, the movie industry produces thousands movies of various genres like drama, action, adventure, romance, sci-fi, war, animation etc. Most of the time, people are confused to choose which movie to watch during their spare time. There are numerous online platforms like Internet Movie Database, Rotten Tomatoes, Metacritic which keep track of the movies along with the information such as actors, directors, budget, as well as user ratings and comments about the movie. A lot of research has been done on prediction of movies based on social media, user ratings. But less work has been done on prediction of movie ratings based on various movie attributes. This study explores the movie rating prediction with the help of various attributes of the movies such as genre, runtime, audience rating, critics ratings etc.

This chapter deals with the overall view of the entire project and the need for this study and also an explanation about the movie industry through which the researcher going to carry out this study. This chapter further deals with the importance of the study and scope of the study. Many concepts have also been used in this chapter which will also be explained.

1.2 Concepts and their definitions in the study

1.2.1 Big data

According to Andrea De Mauro (2016), “Big data is the information asset characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value”.

1.2.2 Data mining

Since the availability of data over the internet has increased, companies started to dig the data to gain valuable insights from those data.

Data mining is defined as the process of discovering patterns in data. The process must be atomic or semi atomic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic one. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner.

1.2.3 Predictive modelling

Predictive modelling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated.

Independent variable – is the variable that is changed or controlled in a scientific experiment to test the effects on the dependent variable.

Dependent variable – is the variable that depends on the independent variable. As the experimenter changes the independent variable the effect on the depended variable is observed and recorded.

1.2.4 Machine learning

It is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Further, machine learning algorithms can be broadly classified into supervised and unsupervised learning. Supervised machine learning algorithms can apply what has been learned in the past to new data using labelled examples to predict the future events. Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labelled.

This study uses the supervised learning approach since the dependent variable (predictor variable) is known.

1.2.5 IMDb

IMDb stands for Internet Movie Database is an online database of information related to films, television programs, home videos and video games, and internet streams, including cast, production, crew and personnel biographies, plot summaries, trivia, and user reviews and ratings. IMDb was launched on 17 October 1990 as an independent company in United Kingdom.

Most data in the database is provided by volunteer contributors. The site enables the registered users to submit reviews and ratings to the movies, programs etc. All registered users choose their own site name, and most operate anonymously. They have a profile page which shows how long a registered user has been a member.

1.2.6 Rotten Tomatoes

Rotten tomatoes is an American review aggregation website for film and television. The company was launched in August 1998 by three undergraduate students at the university of California, Berkeley.

Rotten tomatoes rating are based on reviews by television and movie critics as below,

1. Fresh
   - The fresh rating icon has a tomato symbol named tomato-fresh which means that the media item has at least 60% positive reviews.
2. Splat
   - The splat rating icon has a greenish splash symbol named tomato-splat which means the media item has less than 60% positive reviews.
3. Certified fresh
   - The certified fresh rating icon which have the resource names tomato-certified and tomato-certified-m. Tomatoes assigns a certified fresh rating to media items that meet the following criteria.
     a. A steady rating score of 75% or higher.
     b. At least five reviews from top critics.
     c. At least 80 reviews for films in wide release.
     d. At least 40 reviews for films in limited release.
     e. At least 20 reviews for an individual season of a television show. Only individual seasons are eligible for a rating.

1.3 Importance and scope of the study

According to a study, movie industry in the United States generate revenue up to 10 billion dollars. Each movie cost 100 million approx. but still there is a great deal of uncertainty that the movie will do business or not. Movie industry is a big business, which can give profits or loss up to several millions dollar. Given this background, it is decided to carry out the movie prediction problem. The study deals with the importance of movie rating prediction prior to their release which will be useful to the producer to take necessary actions. The focus of this study is to formulate a method of how to pre-process the data set and evaluate which attributes are the most relevant, by evaluating the correlation between the attributes using the machine learning technique.

1.4 Limitations of the study

The study is limited to minimal number of movies, this is because if the television programs are included, the data will become large which require more time to analyse the study. Due to time constraint researcher has restricted the study to the movies so that researcher can study the data and analyse it with more accuracy. Hence the researcher's study was limited to the following:

- Time constraint
- Limited to movies
- Data accuracy
- May not be applicable to television programs

1.5 Profile of film industry

The film industry or motion picture industry, includes the both technological and commercial institution of filmmaking. Film industry is a huge sector for investment but larger business sectors have more complexity and it is hard to return back the investment. Big investment comes with bigger risks as well as bigger profits. The CEO of Motion Picture Association of America (MPAA) J. Valenti mentioned that 'No one can tell you how a movie is going to do in the marketplace. Not until the film opens in darkened theatre and sparks fly up between the screen and the audience'.

The definition of a movie success is relative, some movies are called successful based on its worldwide gross income, and some movies may not shine in business part but can be called successful for good critic's review and popularity. There are many movies which did not produce good amount of profit during its release time but become famous after few years.

The worldwide theatrical market had a box office of US$38.6 billion in 2016. The top three continents/regions by box office gross were: Asia-Pacific with US$14.9 billion, the U.S. and Canada with US$11.4 billion, and Europe, the Middle East and North Africa with US$9.5 billion. United Kingdom holds the largest box office market as of 2016. As of 2011, the countries with the largest number of film productions were India, Nigeria, and the United States.

Besides being the largest producer of films in world, India also has the largest number of admissions. Indian film industry s multi-lingual and the largest in terms of ticket sales but 3rd largest in terms of revenue mainly due to having amongst the lowest ticket prices in the world. The largest film and most popular industry in India is the Hindi film industry mostly concentrated in Mumbai, and is commonly referred to as Bollywood.

Through the researcher’s internship in SmartMegh Solutions, the company gave an insight about Predictive Analytics. Along with the learning from SmartMegh Solutions, in this study the company has given a wide description about predictive analytics. The next chapter deals with articles which helped the researcher to do the project and the methods which to do the project.

The remaining portion of this report is organised as below:

- In chapter 2: Literature of review of previous research works.
- In chapter 3, Research methodology.
- In chapter 4, Data analysis and interpretation
- In chapter 5, Findings, recommendations and conclusion.

CHAPTER II

REVIEW OF LITERATURE

Introduction

Review of literature is a written work that compiles significant research published on a topic by accredited scholars and researchers, surveys scholarly articles, books, dissertations, conference proceedings, and other sources, examines contrasting perspectives, theoretical approaches, methodologies, findings, results, conclusions. Reviews critically, analyses, and synthesizes existing research on a topic; and, performs a thorough "re" view, "overview", or "look again" of past and current works on a subject, issue, or theory.

Literature reviews are important because they are usually a *required* step in a thesis proposal (Master's or PhD). The proposal will not be well-supported without a literature review. Also, literature reviews are important because they help to learn important authors and ideas in this field. This is useful for the coursework and the writing. Knowing key authors also helps to become acquainted with other researchers in this field.

In this chapter the researcher has explained about the need and importance of predictive analytics. This chapter also explains about the articles which helped the researcher to analyse various topics in predictive analytics like machine learning, regression model and backward stepwise elimination method. The chapter further deals with the already existing solutions for the same problem.

Adam Sadovsky, Xing Chen (2018), tried to find out the efficient machine learning model to predict how a user will rate movies he or she has not seen, given the ratings for movies he or she has seen. This study is based on the Netflix Prize challenge created by Netflix to individuals. The researcher has built various machine learning approaches such as matrix factorization, clustering, various types of linear and logistic regression, etc. One of the most successful prediction methods thus far has been logistic regression with L2 regularization, which achieves approximately 0.93 root mean squared error (RMSE) on the probe set. However, because L2 regularization imposes a Gaussian prior on feature weights, it tends to assign non-zero weight to all features for various reasons, L1 may outperform L2. Furthermore, because L1 regularization results in zero weight for a large majority of features, it can also help to construct more complex models (such as Markov networks with movie potentials).

The study achieved creating a successful model named L1– regularized logistic regression that worked surprisingly well for the Netflix movie rating prediction task.

Nikki Castle (2018), discussed in this article that Machine learning generates a lot of buzz because it's applicable across such a wide variety of use cases. That's because machine learning is actually a set of many different methods that are each uniquely suited to answering diverse questions about a business. To better understand machine learning algorithms, it’s helpful to separate them into groups based on how they work.

The author classifies the machine learning techniques into three categories namely supervised learning, unsupervised learning and semi-supervised learning. Also, the author explains about the two popular types of machine learning methods namely regression and classification. This article is helpful for my study as it briefly explains how to classify the machine learning techniques with the available data.

K. Meenakshi, G. Maragatham, Neha Agarwal and Ishitha Ghosh (2018), says that in the real-world prediction models and mechanisms can be used to predict the success of a movie. This study proposes a system based upon data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. An attempt is made to predict the past as well as the future of movie for the purpose of business certainty or simply a theoretical condition in which decision making towards the success of the movie is without risk, because the decision maker (movie makers and stakeholders) has all the information about the exact outcome of the decision, before he or she makes the decision to release the movie. The study concludes that it is difficult to apply data mining techniques to the data in the IMDb dataset. It requires proper cleaning and integration, and this consumed a large proportion of the time available for this analysis.

Pojan Shahrivar (2017), explained about the areas of creating predictive models using machine learning algorithms. The article states different approaches to apply for a varied predictive analytics problems. The article further explains that the regression analysis generates an equation to describe the relationship between one or more predictors and the response variable and to predict new observations. Linear regression usually uses the ordinary least squares estimation method which derives the equation by minimizing the sum of the squared residuals.

The article also describes about the types of regression analysis as follows,

- Simple linear regression
- Multiple linear regression

Simple linear regression examines the linear relationship between two continuous variables: one response (y) and one predictor (x). When the two variables are related, it is possible to predict a response value from a predictor value with better than chance accuracy.

Multiple linear regression examines the linear relationships between one continuous response and two or more predictors. If the number of predictors is large, then before fitting a regression model with all the predictors, you should use stepwise or best subsets model-selection techniques to screen out predictors not associated with the responses. This article concludes that the use of simple linear regression and multiple linear regression is based on the nature of variable available in the problem.

According to Muhammad Hassan Latif (2016), there is a large amount of data related to the movies is available over the internet, because of that much data available, it is an interesting data mining topic. The prediction of movies is complex problem. Every viewer, producer, director’s production houses all are curious about the movies that how it will perform in the theatre. The study proposes a classification algorithm model which will predict the movies popularity with so many attributes related to a movie and all in different dimensions. The study achieved best results through simple logistic and logistic regression at around 84%. The attributes that contributed the most to information are metascore and number of votes for each movie, Oscar awards won by the movies and the number of screens the movie is going to be screened.

As defined by Jeffrey Ericson, Jesse Grodman (2016) in “What makes a movie successful?”, the right combination of talent, chemistry and timing, all with some good luck. The researcher developed a model using support vector machine (SVM) in order to predict the success rate of the movie based on different attributes. The study suggests production team should consider the release time and season in order collect estimated revenue. Specifically, the finding suggests that there is likely a high box office revenue during the second half of the year.

Tavish Srivastava (2016), the article states that every predictive modelling problems need to be break down into 4 different parts on the basis of time as follows,

- Descriptive analysis of the data – 50% time
- Data treatment (missing value and outlier fixing) – 40% time
- Data modelling – 4% time
- Estimation of performance – 6% time

With advance machine learning tools coming in race, time taken to perform this task can be significantly reduced. The articles suggest that for the initial analysis, there is no need for the future engineering. Hence, the time might need to do descriptive analysis is restricted to know the missing values and big features which are directly visible. The article elaborately describes each part with a sample problem. This article is helpful for the proposed study as it says how important is to classify the predictive modelling problems on the basis of time.

As defined by Achal Augustine (2015) in “User rating prediction for movies” the user rating can be predicted after learning the relationship between the rating and movie’s special characteristics from a training set. The researcher also presents an algorithm for scoring individual members based on the movie rating and also a neural network framework for estimating various parameters for combining individual scores for prediction.

In a study by Nahid Quader, MD. Osman Gani (2015), discussed about making a prediction of society’s reaction to a new product in the sense of popularity and adaption rate become an emerging field of data analysis. The study proposes a decision support system for movie investment sector using machine learning technique. The system will predict an approximate success rate of a movie based on its profitability by analysing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic. The results of the research say that the movie’s box office profit can be predicted based on some features like cast, budget, members, movie release time, movie reviews and various types of movie ratings.

Through this chapter the study about IMDb movie rating prediction using machine learning technique were discussed and further this chapter also helped to gain tools required for analysing the predictive analytics. The fore coming chapter will deal about the Research methodology and tools required to analyse the given data.

Chapter III

RESEARCH METHODOLOGY

This chapter explains the study topic briefly. It says about the methodology of the study. The chapter explains the population of the study, sampling and methods of sampling. The chapter also deals with the survey method and sources of the data, tools for data collection and data analysis. The chapter describes the objective of the study, variables considered, and the hypotheses of the study. Briefly the chapter explained how the study has been selected and researched and carried out.

3.1 STATEMENT OF THE PROBLEM

As every movie contain various characteristics information about the movie, it might be interesting to know how audiences rate the success of a movie. Do audience responded differently to different type of movies, the length of the movie, the year or month it's released or more concerns over who casted in the movie. These information may provide some important insights about how popular the movie can be. A few features about the movie can make it more or less popular, and it's not only meant the content of drama, humour, horror, etc. A few biases can make people unintentionally vote for high or low scores in online ratings.

Popular movie is likely to attract bigger crowds which leading to higher ratings and hence generating more ticket sales which literally means better revenues to the production studio. In summary, every producers would wish to have better ratings for their movies as, to them this is the key driver to success. Online movie renters, such as Netflix, already do this kind of analysis to have the best return when deciding which movie, they are going to make available in their libraries.

This study mainly focuses on predicting the success of a movie based on certain characteristics of a movie. A good prediction model would be valuable to the producers as they will know the likelihood of success of the movie before it is released to the public.

3.2 STUDY DESIGN: EXPLORATORY

The researcher has chosen the Explorative design because of the nature of the study. The raw data provided for this analysis consists of 651 movies which released prior to 2016. The data source comes from both IMDb and Rotten Tomatoes websites in the URL www.imdb.com and www.rottentomatoes.com. The data set is comprised of 651 randomly sampled movies produced and released before 2016, each row in the dataset is a movie and each column are a characteristic of a movie. Alongside with these 651 movies, there are 32 observation columns (i.e. variables) detailing the title of the movie, movie genre, runtime, MPAA rating, studio production, IMDB rating, RT audience rating, Oscar nominations, Actor/Actress/Director names and many more.

3.3 OBJECTIVES

- To predict the viewers perception for the unreleased movie based on the prediction attributes.
- To Find the movie attributes which influences the IMDB movie rating.
- To develop a model which learns itself (Machine Learning) to improve the result accuracy.
- To compare the predicted result with IMDB movie rating.

Variables

The researcher has taken totally 32 variables. Using data reduction Out of 32 variables only 9 are character variables, 12 are factor variable, 10 are numerical variable, one is integer variable and among the present 10 numerical variable 6 are related to date.

- Title
- Title type
- Genre
- Runtime
- Mpaa rating
- Studio
- Theatre release year
- Imdb num votes
- Imdb rating
- Critics score
- Critics rating
- Audience score
- Audience rating
- Best pic winner
- Best actor winner
- Best actress winner
- Best director winner
- Top 200 box
- Cast

- Imdb url
- Rt url

3.4 HYPOTHESIS

H_0 - There is no significant relationship between movie attributes and viewer's perception.

H_1 - There is significant relationship between movie attributes and viewer's perception.

3.5 SAMPLE DESIGN

The type of data which is used in the study is Secondary data. The data is collected from various websites. Websites are major source of data collection place where large amount are data are collected and converted in to information to identify the user rating for a movie. In this study various websites were taken for analysis. Hence the sample is of huge volume of quality data retrieved prior to 2016.

The sample data are congregated from:

a. Internet Movie Database: www.imdb.com

b. Rotten Tomatoes: www.rottentomatoes.com

3.7 DATA COLLECTION: METHODS AND TOOLS

Secondary data related to the study has been collected from the IMDb and RT websites. The data is then extracted from the Kaggle website where the variables related to the movie are filtered using elimination method. The secondary data for the study is collected using qualitative method, where the details are extracted from online database and movie ratings websites. The data collected by users were from the years prior to 2016 thereby giving the suitable solution for the above-mentioned problem.

3.8 METHODS AND TOOLS OF DATA ANALYSIS

The data from secondary sources has been arranged and the missing data were cleansed. The cleaned data has been analysed with various machine learning algorithms by the R studio using R programming language.

3.8.1CONCEPTUAL FRAMEWORK FOR RESEARCH WORK

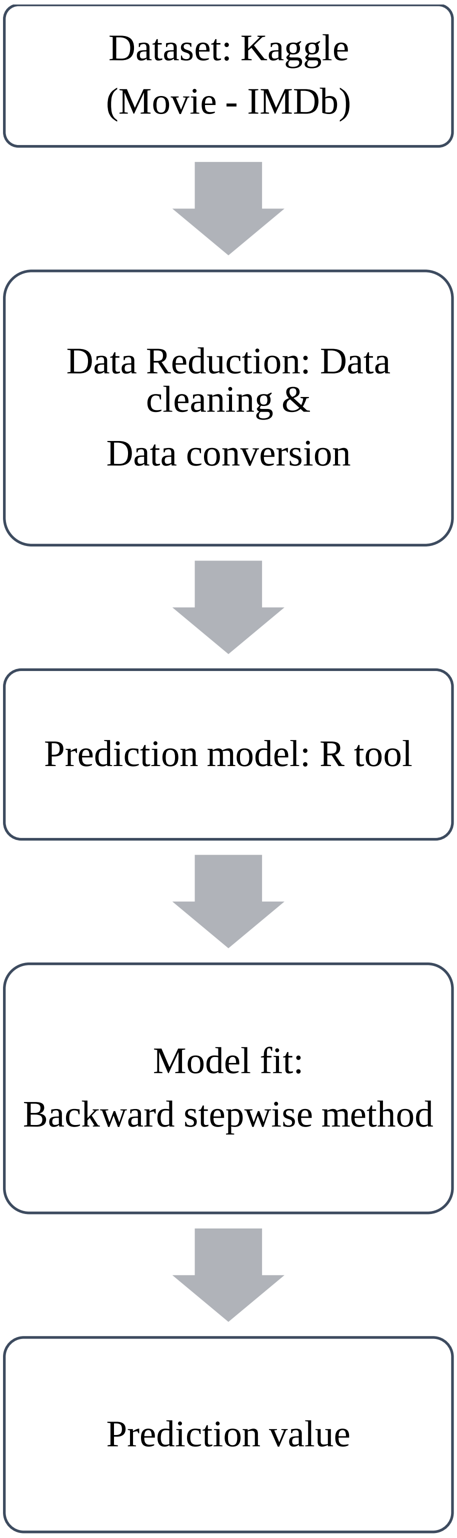


Figure 3.0

Conceptual frame work for predictive analysis

3.8.2 Predictive analysis

The secondary data taken from the websites are then taken for analysis in the R studio using R programming to predict the IMDb rating prediction for movies. The data is further taken for developing machine learning model using regression algorithm.

The chapter gives a clear idea about how the research has been done. This was very helpful for the researcher to formulate the method and the way to carry out the study. Through this chapter, the researcher has always known what to find and what to interpret through this chapter. The data analysis and data interpretation has been done in the following chapter.

CHAPTER IV

DATA ANALYSIS AND INTERPRETATION

4.1 Data setup

This research shows how to perform regression Analysis with R and how to use the final model to predict the rating of an unreleased movie. To perform this analysis, you need to install these R packages,

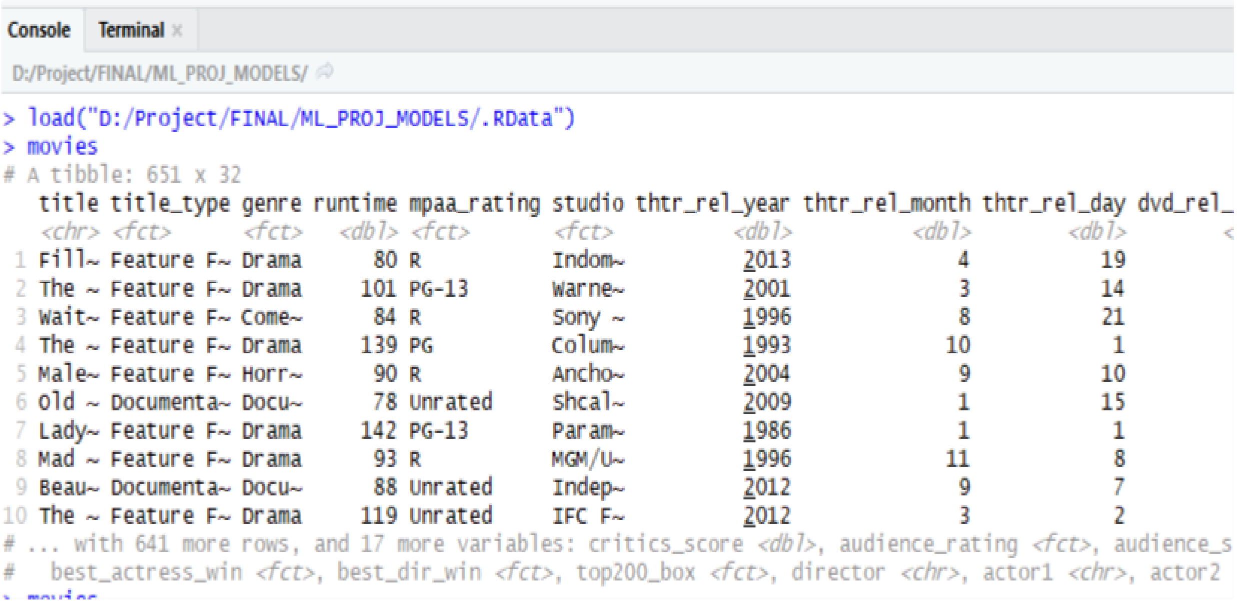
a. ggplot2
b. dplyr
c. GGally
d. corrplot
e. gridExtra
f. statsr
g. caret

This bit of code is to import the csv file:

Snippet 1:

load(“d:/Project/DATASET/movies.Rdata”)

Figure 4.1 - Output for snippet 1: Load data



```
D:/Project/FINAL/ML_PROJ_MODELS/
> load("D:/Project/FINAL/ML_PROJ_MODELS/.RData")
> movies
# A tibble: 651 x 32
   title title_type genre runtime mpaa_rating studio thtr_rel_year thtr_rel_month thtr_rel_day dvd_rel_
   <chr> <fct>      <fct>   <dbl> <fct>       <fct>          <dbl>          <dbl>        <dbl>        <
 1 Fill~ Feature F~ Drama      80 R           Indom~          2013              4           19
 2 The ~ Feature F~ Drama     101 PG-13       Warne~          2001              3           14
 3 Wait~ Feature F~ Come~      84 R           Sony ~          1996              8           21
 4 The ~ Feature F~ Drama     139 PG          Colum~          1993             10            1
 5 Male~ Feature F~ Horr~      90 R           Ancho~          2004              9           10
 6 Old ~ Documenta~ Docu~      78 Unrated     Shcal~          2009              1           15
 7 Lady~ Feature F~ Drama     142 PG-13       Param~          1986              1            1
 8 Mad ~ Feature F~ Drama      93 R           MGM/U~          1996             11            8
 9 Beau~ Documenta~ Docu~      88 Unrated     Indep~          2012              9            7
10 The ~ Feature F~ Drama     119 Unrated     IFC F~          2012              3            2
# ... with 641 more rows, and 17 more variables: critics_score <dbl>, audience_rating <fct>, audience_s
#   best_actress_win <fct>, best_dir_win <fct>, top200_box <fct>, director <chr>, actor1 <chr>, actor2
```

Output for snippet 1: Load data

Inference:

The above bit of code is used to load the IMDb movie dataset into the R studio.

4.2 Data cleansing

Before proceeding further, it is worth noting if there are any missing values in the dataset. This is an important step when working with any regression model.

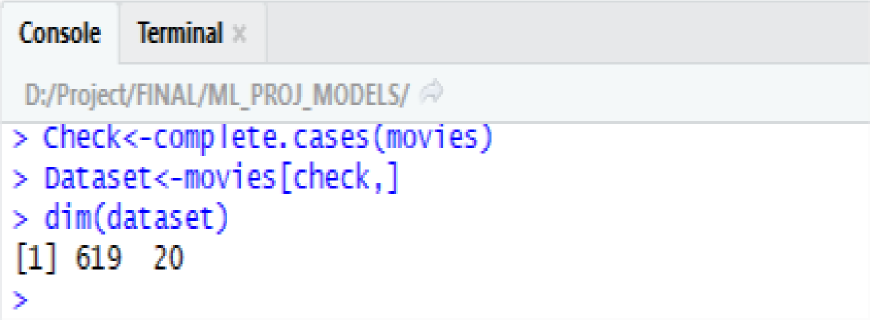
Snippet 2:

Check<-complete.cases(movies)

Dataset<-movies[check,]

Dim(dataset)

Figure 4.2 - Output for snippet 2: Dataset view



```
> Check<-complete.cases(movies)
> Dataset<-movies[check,]
> dim(dataset)
[1] 619  20
>
```

Output for snippet 2: Dataset view

Reduced version – movie dataset with NA removed

Snippet 3:

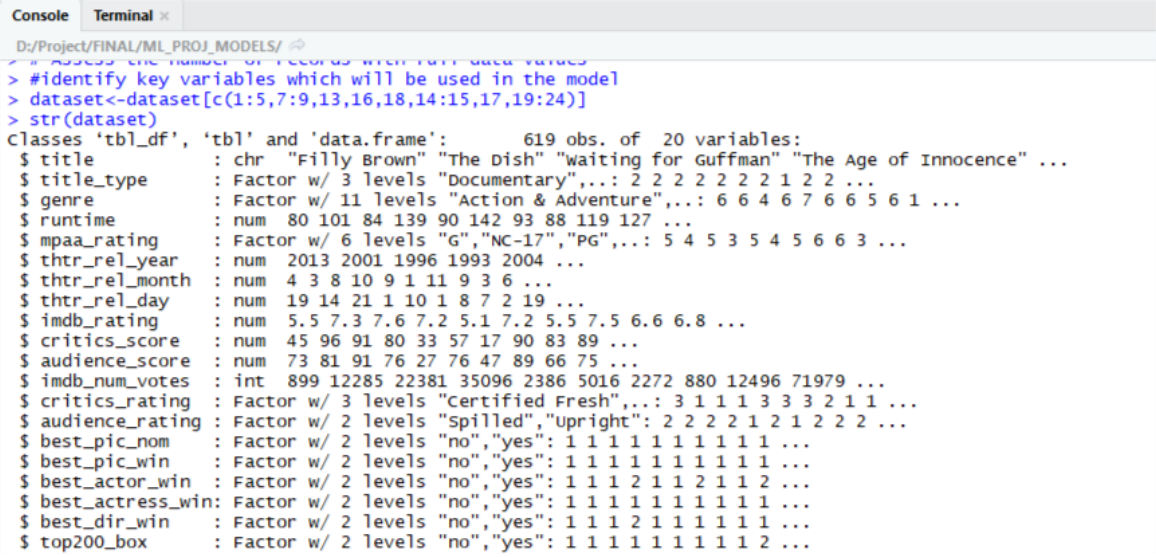
#identify key variables which will be used in the model

Dataset<- dataset[c(1:5,7:9,13,16,18,14:15,17,19:24)]

Str(dataset)

Summary(dataset)

Figure 4.3 - Output for snippet 3: Reduced version of dataset



```
> # Assess the number of records with full data values
> #identify key variables which will be used in the model
> dataset<-dataset[c(1:5,7:9,13,16,18,14:15,17,19:24)]
> str(dataset)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       619 obs. of  20 variables:
 $ title           : chr  "Filly Brown" "The Dish" "Waiting for Guffman" "The Age of Innocence" ...
 $ title_type      : Factor w/ 3 levels "Documentary",..: 2 2 2 2 2 2 2 1 2 2 ...
 $ genre           : Factor w/ 11 levels "Action & Adventure",..: 6 6 4 6 7 6 6 5 6 1 ...
 $ runtime         : num  80 101 84 139 90 142 93 88 119 127 ...
 $ mpaa_rating     : Factor w/ 6 levels "G","NC-17","PG",..: 5 4 5 3 5 4 5 6 6 3 ...
 $ thtr_rel_year   : num  2013 2001 1996 1993 2004 ...
 $ thtr_rel_month  : num  4 3 8 10 9 1 11 9 3 6 ...
 $ thtr_rel_day    : num  19 14 21 1 10 1 8 7 2 19 ...
 $ imdb_rating     : num  5.5 7.3 7.6 7.2 5.1 7.2 5.5 7.5 6.6 6.8 ...
 $ critics_score   : num  45 96 91 80 33 57 17 90 83 89 ...
 $ audience_score  : num  73 81 91 76 27 76 47 89 66 75 ...
 $ imdb_num_votes  : int  899 12285 22381 35096 2386 5016 2272 880 12496 71979 ...
 $ critics_rating  : Factor w/ 3 levels "Certified Fresh",..: 3 1 1 1 3 3 3 2 1 1 ...
 $ audience_rating : Factor w/ 2 levels "Spilled","Upright": 2 2 2 2 1 2 1 2 2 2 ...
 $ best_pic_nom    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ best_pic_win    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ best_actor_win  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 2 1 1 2 ...
 $ best_actress_win: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ best_dir_win    : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ top200_box      : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 2 ...
```

Output for snippet 3: Reduced version of dataset

4.3 DATA DISTRIBUTION USING R

The other aspect of the work is to assess the data distribution of each movie characteristics (e.g. by Title, by Genre, by Runtime, by MPAA Rating, by Theatre Release Year, Month or Day etc)

Snippet 4:

Data distribution -1:

Type <-ggplot(data = dataset, aes(x = title_type)) + geom_bar(fill="black") + xlab("Type") + ylab("Count") + theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))

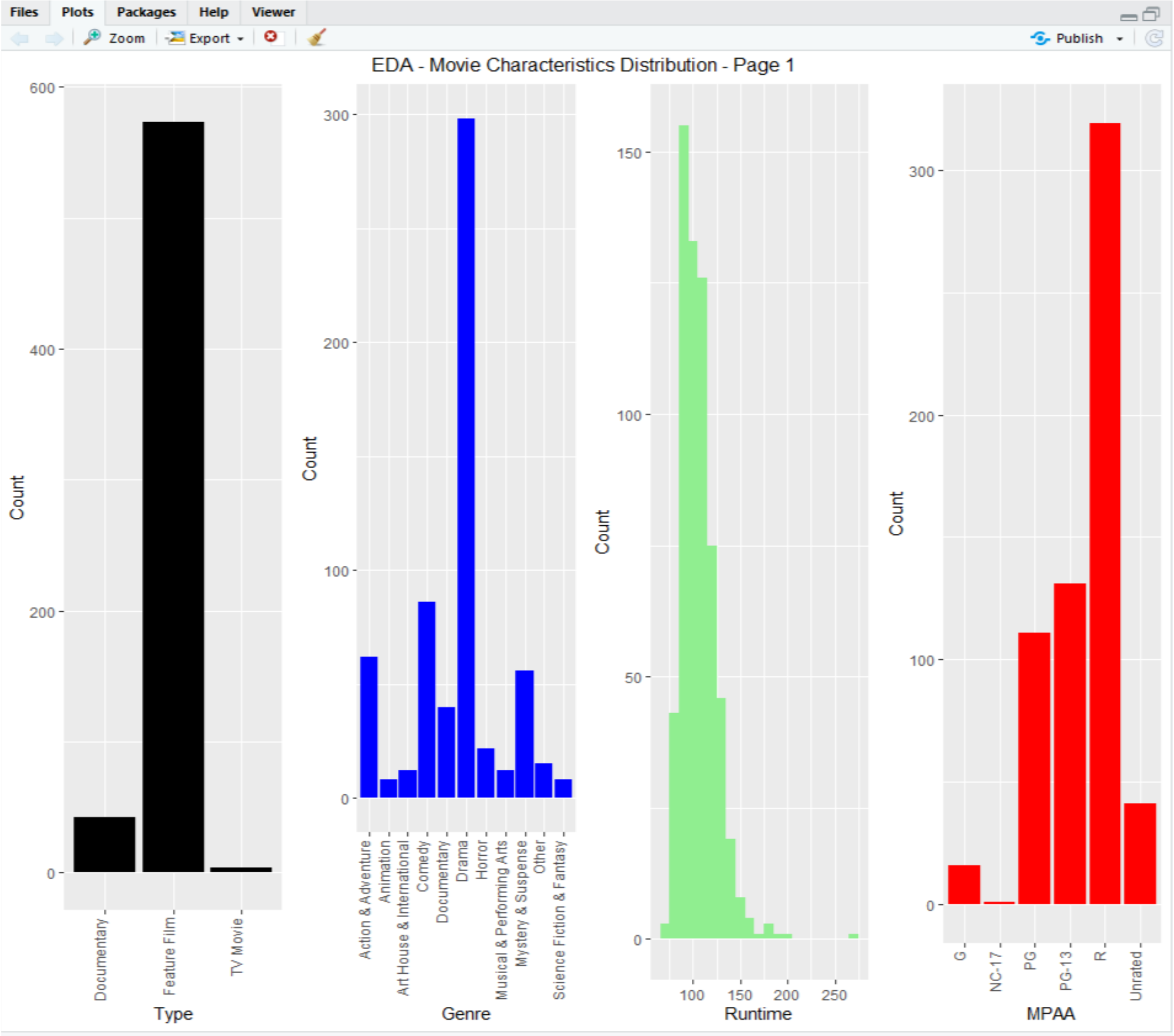
Genre <-ggplot(data = dataset, aes(x = genre)) + geom_bar(fill="blue") + xlab("Genre") + ylab("Count") + theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))

RunTime <- ggplot(data = dataset, aes(x=runtime)) + geom_histogram(binwidth=10,fill="light green") + xlab("Runtime") + ylab("Count")

MPAA <-ggplot(data = dataset, aes(x = mpaa_rating)) + geom_bar(fill="red") + xlab("MPAA") + ylab("Count") + theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))

grid.arrange(Type,Genre,RunTime,MPAA,nrow=1,top="EDA - Movie Characteristics Distribution - Page 1")

Figure 4.4 - Output for snippet 4: Data distribution 1



Output for snippet 4: Data distribution 1

Data distribution -2:

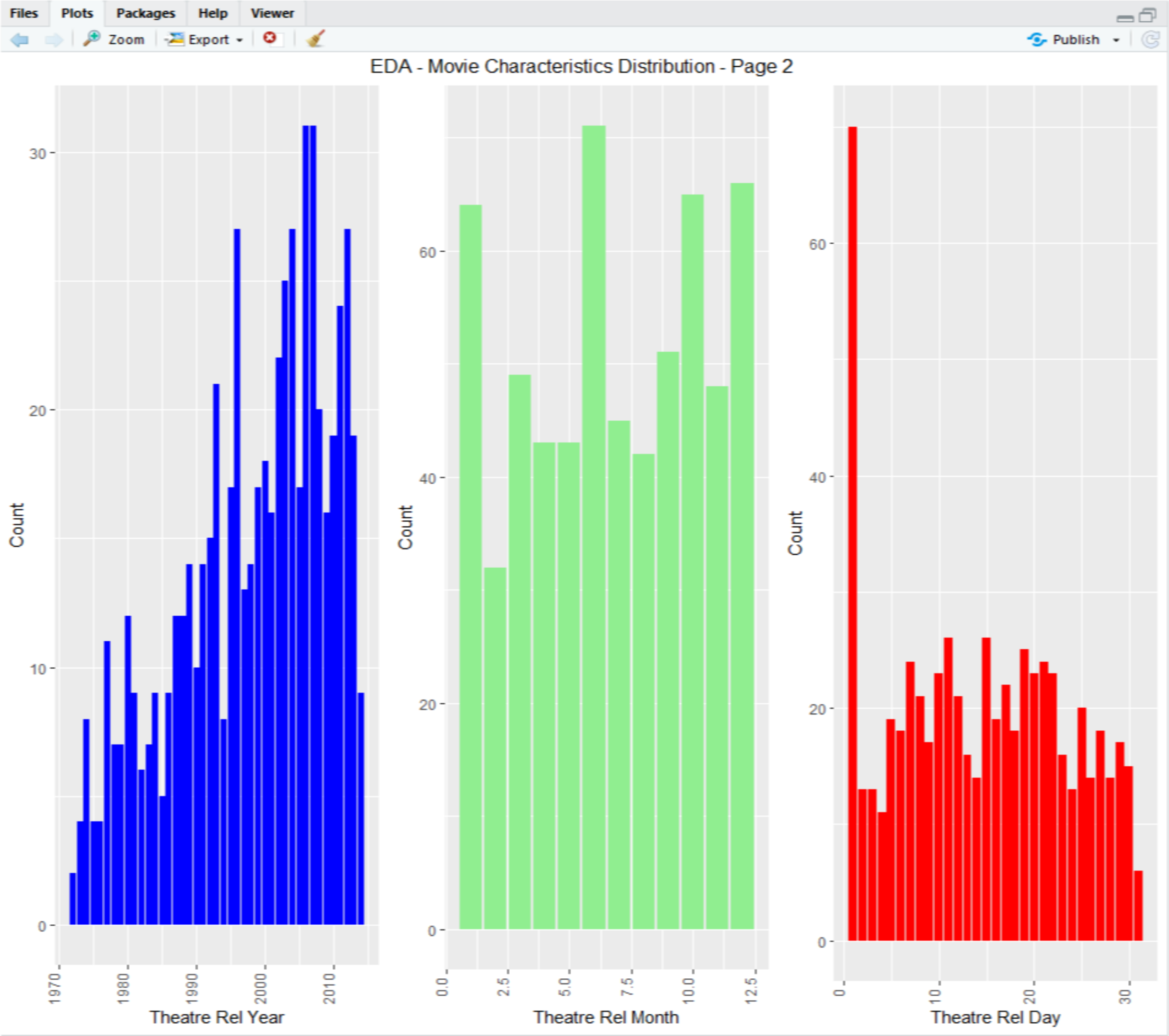
Thtr_Yr <-ggplot(data = dataset, aes(x = thtr_rel_year)) + geom_bar(fill="blue") + xlab("Theatre Rel Year") + ylab("Count") + theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))

Thtr_Month <-ggplot(data = dataset, aes(x = thtr_rel_month)) + geom_bar(fill="light green") + xlab("Theatre Rel Month") + ylab("Count") + theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))

Thtr_Day <-ggplot(data = dataset, aes(x = thtr_rel_day)) + geom_bar(fill="red") + xlab("Theatre Rel Day") + ylab("Count") + theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0))

grid.arrange(Thtr_Yr,Thtr_Month,Thtr_Day, nrow=1,top="EDA - Movie Characteristics Distribution - Page 2")

Figure 4.5 - Output for snippet 4: Data distribution 2



Output for snippet 4 (Data distribution 1)

Inference:

The above charts are to assess the data distribution of each movie characteristics (eg. by Title, by Genre, by RunTime, by MPAA Rating, by Theatre Release Year, Month or Day etc)

4.4 Modelling

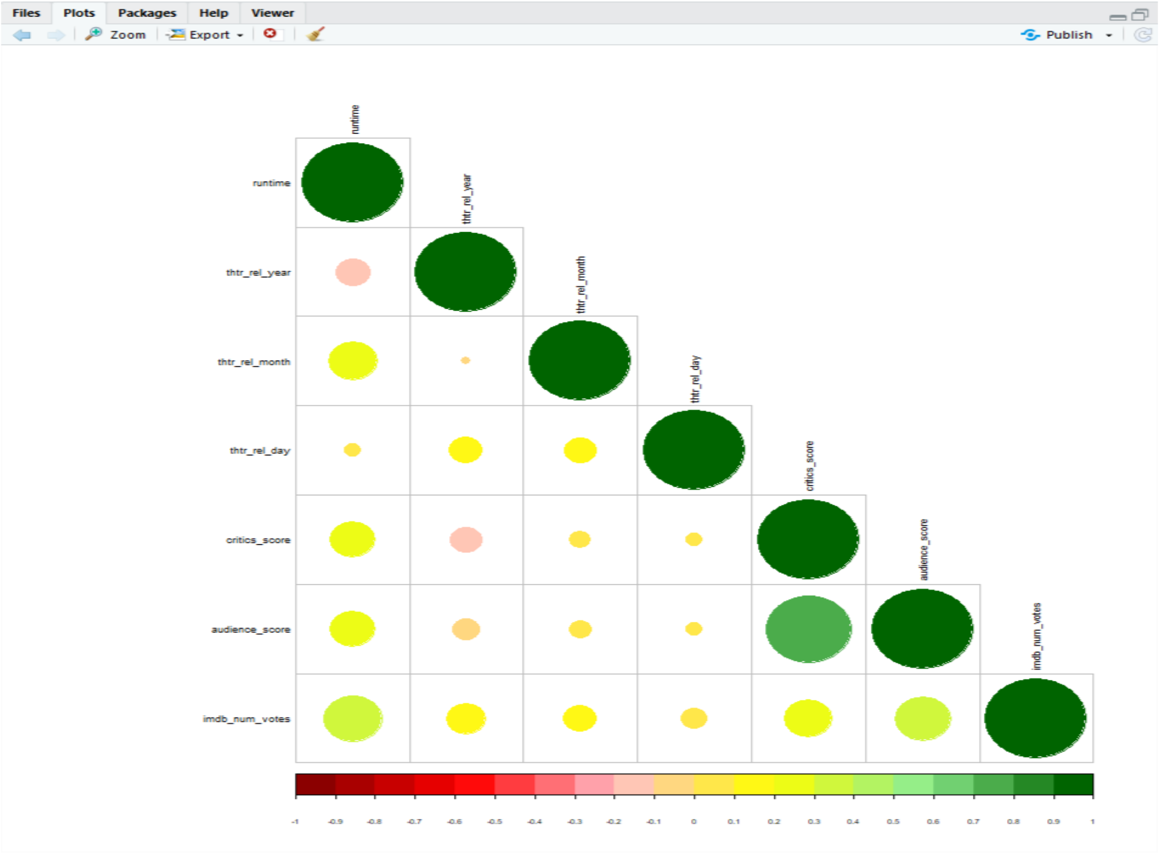
Check collinearity:

After explored on the data, the next thing is to check if there is any collinearity within the numerical explanatory variables. To do this, the researcher have created a sub-dataset of all the numeric variables e.g. runtime, thtr_rel_year, thtr_rel_month, thtr_rel_day, imdb_num_votes, critics_score, audience_score. This should ease us to do a correlation matrix to understand each of these explanatory variables.

Snippet 5:

```
num_expl_var <- dataset[c(4,6:8,10:12)]

#identify those numerical explanatory variables

corr<- cor(num_expl_var)

cex.before <- par("cex")

par(cex = 0.55)

col<- colorRampPalette(c("dark red","red","pink", "yellow","light green", "dark green"))

+(20)

corrplot(corr, method="circle", type="lower", col=col, sig.level = 0.01, tl.col="black")
```

Figure 4.6 - Output for snippet 5: Collinearity



output for snippet 5: Collinearity

Inference:

In the correlogram chart above, it seems Critics score and Audience score are highly correlated (i.e. Collinear). Literally it means if the two variables are added as explanatory variables, this will distort and complicate the model and hence redundant. As a result, one of these variables has to be removed from the model. In this case, the researcher has chosen to remove critics score from my regression model.

4.5 MULTIPLE REGRESSION MODEL - USING BACKWARD STEPWISE REGRESSION APPROACH

Model -1

Firstly, the researcher will create a full model and then work backward by eliminating variables which has the highest p-value. This process will take a while before we reach our final model. This approach is called backward stepwise regression.

Snippet 6:

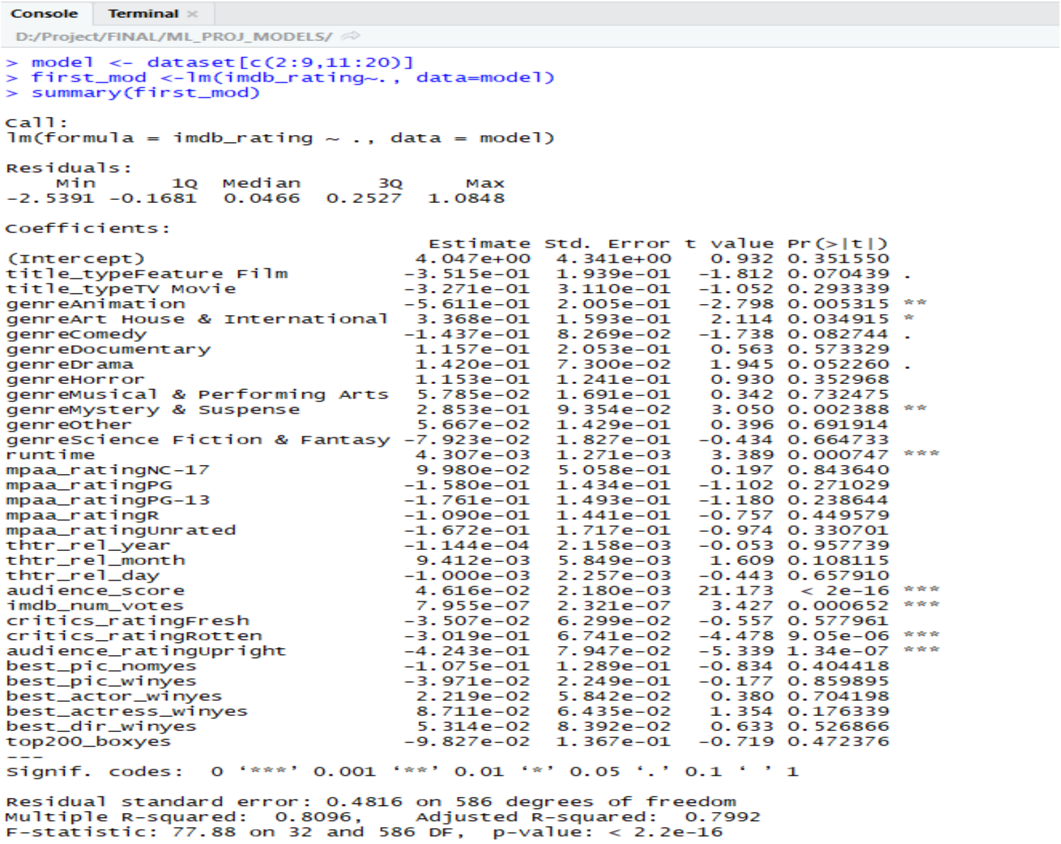
model <- dataset[c(2:9,11:20)]

first_mod <-lm(imdb_rating~., data=model)

summary(first_mod)

anova(first_mod)

Figure 4.7 - Output for snippet 6: Model 1



```
> model <- dataset[c(2:9,11:20)]
> first_mod <-lm(imdb_rating~., data=model)
> summary(first_mod)

Call:
lm(formula = imdb_rating ~ ., data = model)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5391 -0.1681  0.0466  0.2527  1.0848 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     4.047e+00  4.341e+00   0.932 0.351550    
title_typeFeature Film         -3.515e-01  1.939e-01  -1.812 0.070439 .  
title_typeTV Movie             -3.271e-01  3.110e-01  -1.052 0.293339    
genreAnimation                 -5.611e-01  2.005e-01  -2.798 0.005315 ** 
genreArt House & International  3.368e-01  1.593e-01   2.114 0.034915 *  
genreComedy                    -1.437e-01  8.269e-02  -1.738 0.082744 .  
genreDocumentary                1.157e-01  2.053e-01   0.563 0.573329    
genreDrama                      1.420e-01  7.300e-02   1.945 0.052260 .  
genreHorror                     1.153e-01  1.241e-01   0.930 0.352968    
genreMusical & Performing Arts  5.785e-02  1.691e-01   0.342 0.732475    
genreMystery & Suspense         2.853e-01  9.354e-02   3.050 0.002388 ** 
genreOther                      5.667e-02  1.429e-01   0.396 0.691914    
genreScience Fiction & Fantasy -7.923e-02  1.827e-01  -0.434 0.664733    
runtime                         4.307e-03  1.271e-03   3.389 0.000747 ***
mpaa_ratingNC-17                9.980e-02  5.058e-01   0.197 0.843640    
mpaa_ratingPG                  -1.580e-01  1.434e-01  -1.102 0.271029    
mpaa_ratingPG-13               -1.761e-01  1.493e-01  -1.180 0.238644    
mpaa_ratingR                   -1.090e-01  1.441e-01  -0.757 0.449579    
mpaa_ratingUnrated             -1.672e-01  1.717e-01  -0.974 0.330701    
thtr_rel_year                  -1.144e-04  2.158e-03  -0.053 0.957739    
thtr_rel_month                  9.412e-03  5.849e-03   1.609 0.108115    
thtr_rel_day                   -1.000e-03  2.257e-03  -0.443 0.657910    
audience_score                  4.616e-02  2.180e-03  21.173  < 2e-16 ***
imdb_num_votes                  7.955e-07  2.321e-07   3.427 0.000652 ***
critics_ratingFresh            -3.507e-02  6.299e-02  -0.557 0.577961    
critics_ratingRotten           -3.019e-01  6.741e-02  -4.478 9.05e-06 ***
audience_ratingUpright         -4.243e-01  7.947e-02  -5.339 1.34e-07 ***
best_pic_nomyes                -1.075e-01  1.289e-01  -0.834 0.404418    
best_pic_winyes                -3.971e-02  2.249e-01  -0.177 0.859895    
best_actor_winyes               2.219e-02  5.842e-02   0.380 0.704198    
best_actress_winyes             8.711e-02  6.435e-02   1.354 0.176339    
best_dir_winyes                 5.314e-02  8.392e-02   0.633 0.526866    
top200_boxyes                  -9.827e-02  1.367e-01  -0.719 0.472376    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4816 on 586 degrees of freedom
Multiple R-squared:  0.8096,	Adjusted R-squared:  0.7992 
F-statistic: 77.88 on 32 and 586 DF,  p-value: < 2.2e-16
```

Output for snippet 6: Model 1

This probably a good start as the adj R2 of 0.7992 is relatively strong. But the next procedure is to remove those insignificant variables one-by-one, based on the highest p-value first. In this specific model, best_pic_win has the highest p-value and hence will be removed accordingly. This is an iterative process and may take some time to get to the final model.

Model-2:

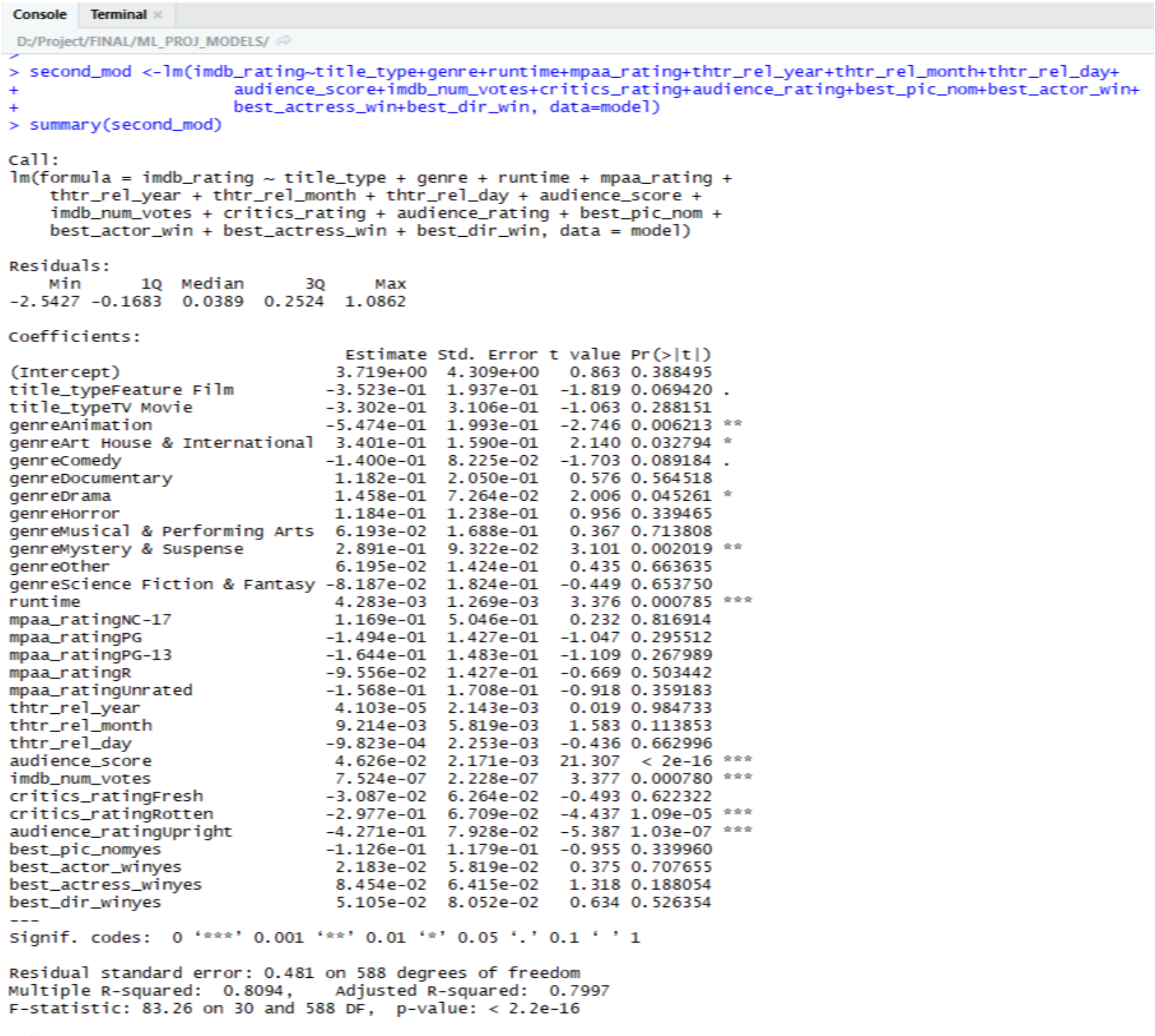
Snippet 7:

second_mod<- lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+thtr_rel_day+ audience_score+imdb_num_votes+critics_rating+audience_rating+best_pic_nom+best_actor_win+ best_actress_win+best_dir_win, data=model)

summary(second_mod)

anova(second_mod)

Figure 4.8 - Output for snippet 7: Model 2



```
> second_mod <-lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+thtr_rel_day+
+                    audience_score+imdb_num_votes+critics_rating+audience_rating+best_pic_nom+best_actor_win+
+                    best_actress_win+best_dir_win, data=model)
> summary(second_mod)

Call:
lm(formula = imdb_rating ~ title_type + genre + runtime + mpaa_rating +
    thtr_rel_year + thtr_rel_month + thtr_rel_day + audience_score +
    imdb_num_votes + critics_rating + audience_rating + best_pic_nom +
    best_actor_win + best_actress_win + best_dir_win, data = model)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5427 -0.1683  0.0389  0.2524  1.0862 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    3.719e+00  4.309e+00   0.863 0.388495    
title_typeFeature Film        -3.523e-01  1.937e-01  -1.819 0.069420 .  
title_typeTV Movie            -3.302e-01  3.106e-01  -1.063 0.288151    
genreAnimation                -5.474e-01  1.993e-01  -2.746 0.006213 ** 
genreArt House & International 3.401e-01  1.590e-01   2.140 0.032794 *  
genreComedy                   -1.400e-01  8.225e-02  -1.703 0.089184 .  
genreDocumentary               1.182e-01  2.050e-01   0.576 0.564518    
genreDrama                     1.458e-01  7.264e-02   2.006 0.045261 *  
genreHorror                    1.184e-01  1.238e-01   0.956 0.339465    
genreMusical & Performing Arts 6.193e-02  1.688e-01   0.367 0.713808    
genreMystery & Suspense        2.891e-01  9.322e-02   3.101 0.002019 ** 
genreOther                     6.195e-02  1.424e-01   0.435 0.663635    
genreScience Fiction & Fantasy -8.187e-02  1.824e-01  -0.449 0.653750    
runtime                        4.283e-03  1.269e-03   3.376 0.000785 ***
mpaa_ratingNC-17               1.169e-01  5.046e-01   0.232 0.816914    
mpaa_ratingPG                 -1.494e-01  1.427e-01  -1.047 0.295512    
mpaa_ratingPG-13              -1.644e-01  1.483e-01  -1.109 0.267989    
mpaa_ratingR                  -9.556e-02  1.427e-01  -0.669 0.503442    
mpaa_ratingUnrated            -1.568e-01  1.708e-01  -0.918 0.359183    
thtr_rel_year                  4.103e-05  2.143e-03   0.019 0.984733    
thtr_rel_month                 9.214e-03  5.819e-03   1.583 0.113853    
thtr_rel_day                  -9.823e-04  2.253e-03  -0.436 0.662996    
audience_score                 4.626e-02  2.171e-03  21.307  < 2e-16 ***
imdb_num_votes                 7.524e-07  2.228e-07   3.377 0.000780 ***
critics_ratingFresh           -3.087e-02  6.264e-02  -0.493 0.622322    
critics_ratingRotten          -2.977e-01  6.709e-02  -4.437 1.09e-05 ***
audience_ratingUpright        -4.271e-01  7.928e-02  -5.387 1.03e-07 ***
best_pic_nomyes               -1.126e-01  1.179e-01  -0.955 0.339960    
best_actor_winyes              2.183e-02  5.819e-02   0.375 0.707655    
best_actress_winyes            8.454e-02  6.415e-02   1.318 0.188054    
best_dir_winyes                5.105e-02  8.052e-02   0.634 0.526354    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.481 on 588 degrees of freedom
Multiple R-squared:  0.8094,	Adjusted R-squared:  0.7997 
F-statistic: 83.26 on 30 and 588 DF,  p-value: < 2.2e-16
```

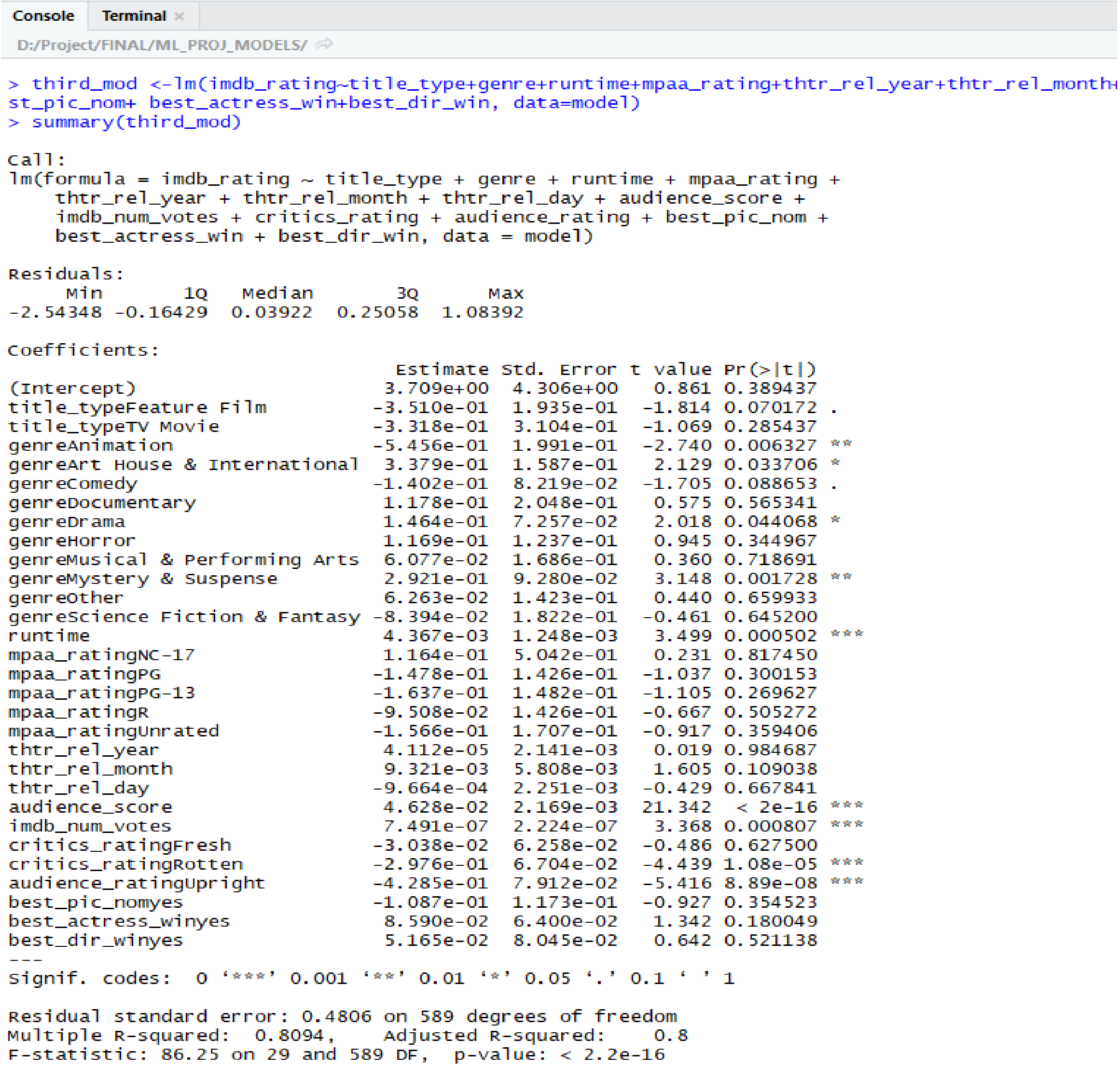
Output for snippet 7: Model 2

Model-3:

Snippet 8:

```
third_mod <-lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+thtr_rel_day+audience_score+imdb_num_votes+critics_rating+audience_rating+best_pic_nom+ best_actress_win+best_dir_win, data=model)

summary(third_mod)

anova(third_mod)
```

Figure 4.9 - Output for snippet 8: Model 3



```
> third_mod <-lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+
st_pic_nom+ best_actress_win+best_dir_win, data=model)
> summary(third_mod)

Call:
lm(formula = imdb_rating ~ title_type + genre + runtime + mpaa_rating + 
    thtr_rel_year + thtr_rel_month + thtr_rel_day + audience_score + 
    imdb_num_votes + critics_rating + audience_rating + best_pic_nom + 
    best_actress_win + best_dir_win, data = model)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.54348 -0.16429  0.03922  0.25058  1.08392 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     3.709e+00  4.306e+00   0.861 0.389437    
title_typeFeature Film         -3.510e-01  1.935e-01  -1.814 0.070172 .  
title_typeTV Movie             -3.318e-01  3.104e-01  -1.069 0.285437    
genreAnimation                 -5.456e-01  1.991e-01  -2.740 0.006327 ** 
genreArt House & International  3.379e-01  1.587e-01   2.129 0.033706 *  
genreComedy                    -1.402e-01  8.219e-02  -1.705 0.088653 .  
genreDocumentary                1.178e-01  2.048e-01   0.575 0.565341    
genreDrama                      1.464e-01  7.257e-02   2.018 0.044068 *  
genreHorror                     1.169e-01  1.237e-01   0.945 0.344967    
genreMusical & Performing Arts  6.077e-02  1.686e-01   0.360 0.718691    
genreMystery & Suspense         2.921e-01  9.280e-02   3.148 0.001728 ** 
genreOther                      6.263e-02  1.423e-01   0.440 0.659933    
genreScience Fiction & Fantasy -8.394e-02  1.822e-01  -0.461 0.645200    
runtime                         4.367e-03  1.248e-03   3.499 0.000502 ***
mpaa_ratingNC-17                1.164e-01  5.042e-01   0.231 0.817450    
mpaa_ratingPG                  -1.478e-01  1.426e-01  -1.037 0.300153    
mpaa_ratingPG-13               -1.637e-01  1.482e-01  -1.105 0.269627    
mpaa_ratingR                   -9.508e-02  1.426e-01  -0.667 0.505272    
mpaa_ratingUnrated             -1.566e-01  1.707e-01  -0.917 0.359406    
thtr_rel_year                   4.112e-05  2.141e-03   0.019 0.984687    
thtr_rel_month                  9.321e-03  5.808e-03   1.605 0.109038    
thtr_rel_day                   -9.664e-04  2.251e-03  -0.429 0.667841    
audience_score                  4.628e-02  2.169e-03  21.342  < 2e-16 ***
imdb_num_votes                  7.491e-07  2.224e-07   3.368 0.000807 ***
critics_ratingFresh            -3.038e-02  6.258e-02  -0.486 0.627500    
critics_ratingRotten           -2.976e-01  6.704e-02  -4.439 1.08e-05 ***
audience_ratingUpright         -4.285e-01  7.912e-02  -5.416 8.89e-08 ***
best_pic_nomyes                -1.087e-01  1.173e-01  -0.927 0.354523    
best_actress_winyes             8.590e-02  6.400e-02   1.342 0.180049    
best_dir_winyes                 5.165e-02  8.045e-02   0.642 0.521138    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4806 on 589 degrees of freedom
Multiple R-squared:  0.8094,	Adjusted R-squared:    0.8 
F-statistic: 86.25 on 29 and 589 DF,  p-value: < 2.2e-16
```

Output for snippet 8: Model 3

Model-4:

Snippet 9:

fourth_mod <- lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+thtr_rel_day+

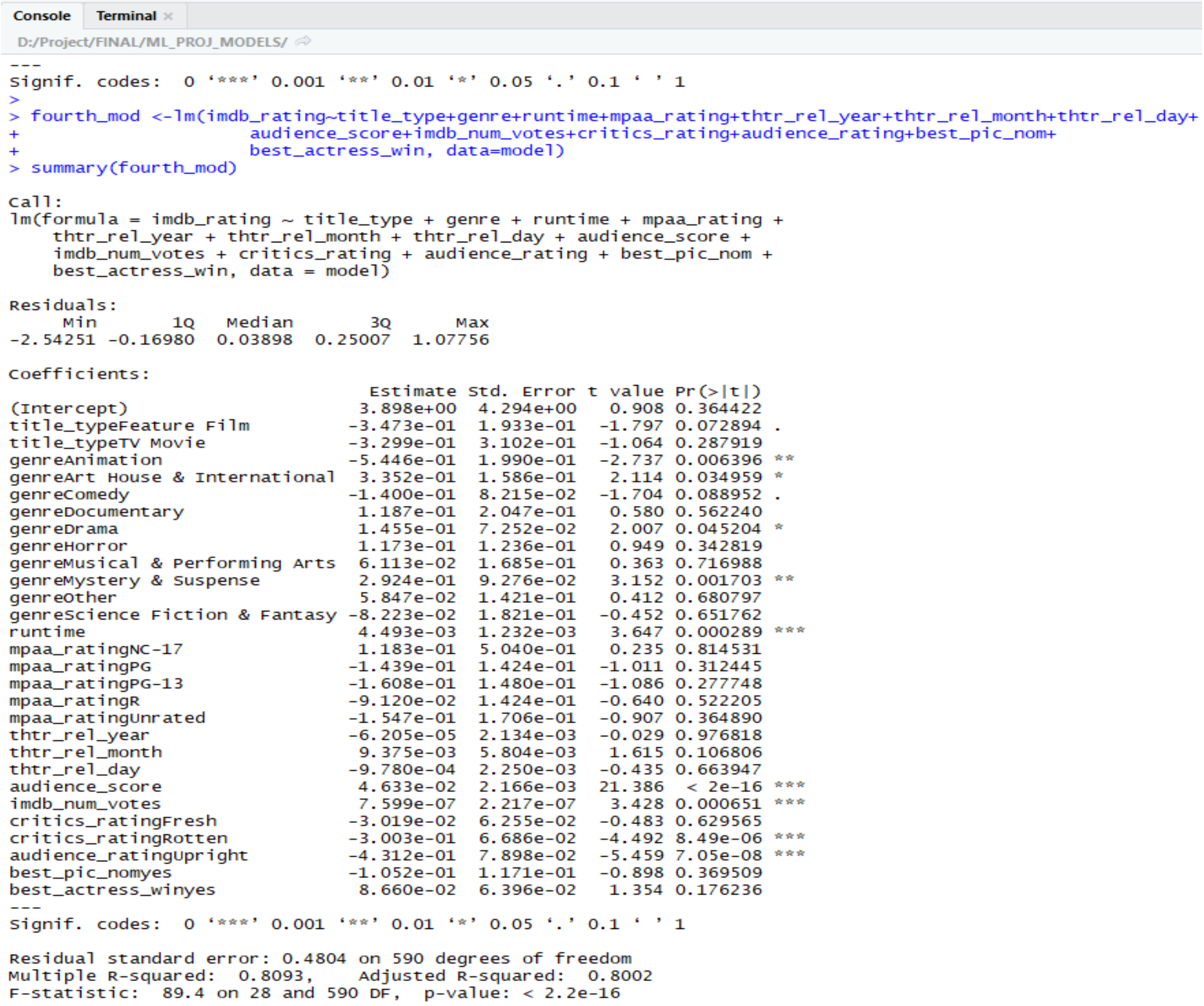
audience_score+imdb_num_votes+critics_rating+audience_rating+best_pic_nom+

best_actress_win, data=model)

summary(fourth_mod)

anova(fourth_mod)

Figure 4.10 - Output for snippet 9: Model 4



```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> fourth_mod <-lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+thtr_rel_day+
+                  audience_score+imdb_num_votes+critics_rating+audience_rating+best_pic_nom+
+                  best_actress_win, data=model)
> summary(fourth_mod)

Call:
lm(formula = imdb_rating ~ title_type + genre + runtime + mpaa_rating + 
    thtr_rel_year + thtr_rel_month + thtr_rel_day + audience_score + 
    imdb_num_votes + critics_rating + audience_rating + best_pic_nom + 
    best_actress_win, data = model)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.54251 -0.16980  0.03898  0.25007  1.07756 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    3.898e+00  4.294e+00   0.908 0.364422    
title_typeFeature Film        -3.473e-01  1.933e-01  -1.797 0.072894 .  
title_typeTV Movie            -3.299e-01  3.102e-01  -1.064 0.287919    
genreAnimation                -5.446e-01  1.990e-01  -2.737 0.006396 ** 
genreArt House & International 3.352e-01  1.586e-01   2.114 0.034959 *  
genreComedy                   -1.400e-01  8.215e-02  -1.704 0.088952 .  
genreDocumentary               1.187e-01  2.047e-01   0.580 0.562240    
genreDrama                     1.455e-01  7.252e-02   2.007 0.045204 *  
genreHorror                    1.173e-01  1.236e-01   0.949 0.342819    
genreMusical & Performing Arts 6.113e-02  1.685e-01   0.363 0.716988    
genreMystery & Suspense        2.924e-01  9.276e-02   3.152 0.001703 ** 
genreOther                     5.847e-02  1.421e-01   0.412 0.680797    
genreScience Fiction & Fantasy -8.223e-02  1.821e-01  -0.452 0.651762    
runtime                        4.493e-03  1.232e-03   3.647 0.000289 ***
mpaa_ratingNC-17               1.183e-01  5.040e-01   0.235 0.814531    
mpaa_ratingPG                 -1.439e-01  1.424e-01  -1.011 0.312445    
mpaa_ratingPG-13              -1.608e-01  1.480e-01  -1.086 0.277748    
mpaa_ratingR                  -9.120e-02  1.424e-01  -0.640 0.522205    
mpaa_ratingUnrated            -1.547e-01  1.706e-01  -0.907 0.364890    
thtr_rel_year                 -6.205e-05  2.134e-03  -0.029 0.976818    
thtr_rel_month                 9.375e-03  5.804e-03   1.615 0.106806    
thtr_rel_day                  -9.780e-04  2.250e-03  -0.435 0.663947    
audience_score                 4.633e-02  2.166e-03  21.386  < 2e-16 ***
imdb_num_votes                 7.599e-07  2.217e-07   3.428 0.000651 ***
critics_ratingFresh           -3.019e-02  6.255e-02  -0.483 0.629565    
critics_ratingRotten          -3.003e-01  6.686e-02  -4.492 8.49e-06 ***
audience_ratingUpright        -4.312e-01  7.898e-02  -5.459 7.05e-08 ***
best_pic_nomyes               -1.052e-01  1.171e-01  -0.898 0.369509    
best_actress_winyes            8.660e-02  6.396e-02   1.354 0.176236    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4804 on 590 degrees of freedom
Multiple R-squared:  0.8093,	Adjusted R-squared:  0.8002 
F-statistic:  89.4 on 28 and 590 DF,  p-value: < 2.2e-16
```

Output for snippet 9: Model 4

Model-5:

Snippet 10:

fifth_mod <- lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+thtr_rel_day+

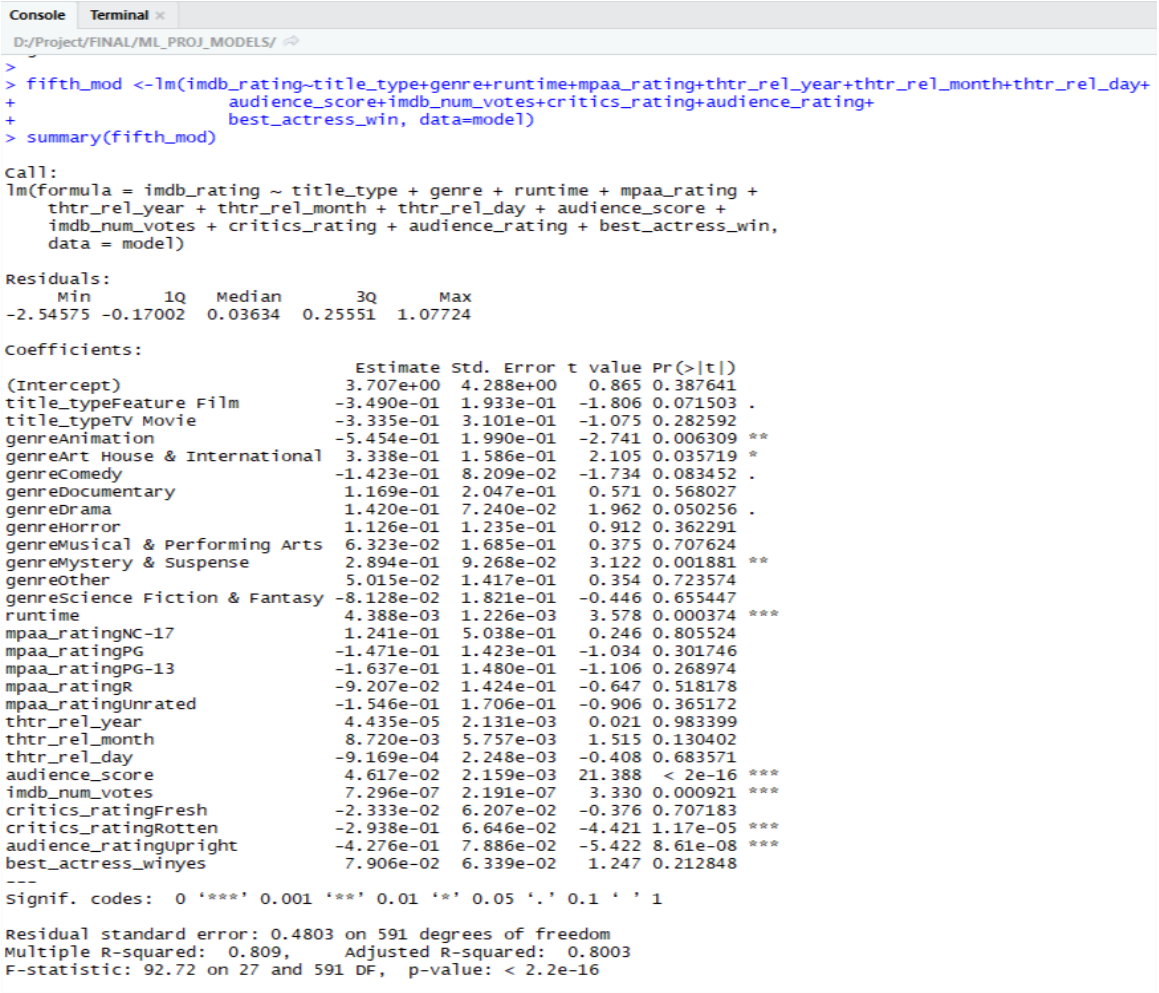
audience_score+imdb_num_votes+critics_rating+audience_rating+

best_actress_win, data=model)

summary(fifth_mod)

anova(fifth_mod)

Figure 4.11 - Output for snippet 10: Model 5



```
>
> fifth_mod <-lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+thtr_rel_day+
+                    audience_score+imdb_num_votes+critics_rating+audience_rating+
+                    best_actress_win, data=model)
> summary(fifth_mod)

Call:
lm(formula = imdb_rating ~ title_type + genre + runtime + mpaa_rating + 
    thtr_rel_year + thtr_rel_month + thtr_rel_day + audience_score + 
    imdb_num_votes + critics_rating + audience_rating + best_actress_win, 
    data = model)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.54575 -0.17002  0.03634  0.25551  1.07724 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    3.707e+00  4.288e+00   0.865 0.387641    
title_typeFeature Film        -3.490e-01  1.933e-01  -1.806 0.071503 .  
title_typeTV Movie            -3.335e-01  3.101e-01  -1.075 0.282592    
genreAnimation                -5.454e-01  1.990e-01  -2.741 0.006309 ** 
genreArt House & International  3.338e-01  1.586e-01   2.105 0.035719 *  
genreComedy                   -1.423e-01  8.209e-02  -1.734 0.083452 .  
genreDocumentary               1.169e-01  2.047e-01   0.571 0.568027    
genreDrama                     1.420e-01  7.240e-02   1.962 0.050256 .  
genreHorror                    1.126e-01  1.235e-01   0.912 0.362291    
genreMusical & Performing Arts  6.323e-02  1.685e-01   0.375 0.707624    
genreMystery & Suspense        2.894e-01  9.268e-02   3.122 0.001881 ** 
genreOther                     5.015e-02  1.417e-01   0.354 0.723574    
genreScience Fiction & Fantasy -8.128e-02  1.821e-01  -0.446 0.655447    
runtime                        4.388e-03  1.226e-03   3.578 0.000374 ***
mpaa_ratingNC-17               1.241e-01  5.038e-01   0.246 0.805524    
mpaa_ratingPG                 -1.471e-01  1.423e-01  -1.034 0.301746    
mpaa_ratingPG-13              -1.637e-01  1.480e-01  -1.106 0.268974    
mpaa_ratingR                  -9.207e-02  1.424e-01  -0.647 0.518178    
mpaa_ratingUnrated            -1.546e-01  1.706e-01  -0.906 0.365172    
thtr_rel_year                  4.435e-05  2.131e-03   0.021 0.983399    
thtr_rel_month                 8.720e-03  5.757e-03   1.515 0.130402    
thtr_rel_day                  -9.169e-04  2.248e-03  -0.408 0.683571    
audience_score                 4.617e-02  2.159e-03  21.388  < 2e-16 ***
imdb_num_votes                 7.296e-07  2.191e-07   3.330 0.000921 ***
critics_ratingFresh           -2.333e-02  6.207e-02  -0.376 0.707183    
critics_ratingRotten          -2.938e-01  6.646e-02  -4.421 1.17e-05 ***
audience_ratingUpright        -4.276e-01  7.886e-02  -5.422 8.61e-08 ***
best_actress_winyes            7.906e-02  6.339e-02   1.247 0.212848    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4803 on 591 degrees of freedom
Multiple R-squared:  0.809,	Adjusted R-squared:  0.8003 
F-statistic: 92.72 on 27 and 591 DF,  p-value: < 2.2e-16
```

Output for snippet 5: Model 5

Model-6:

Snippet 11:

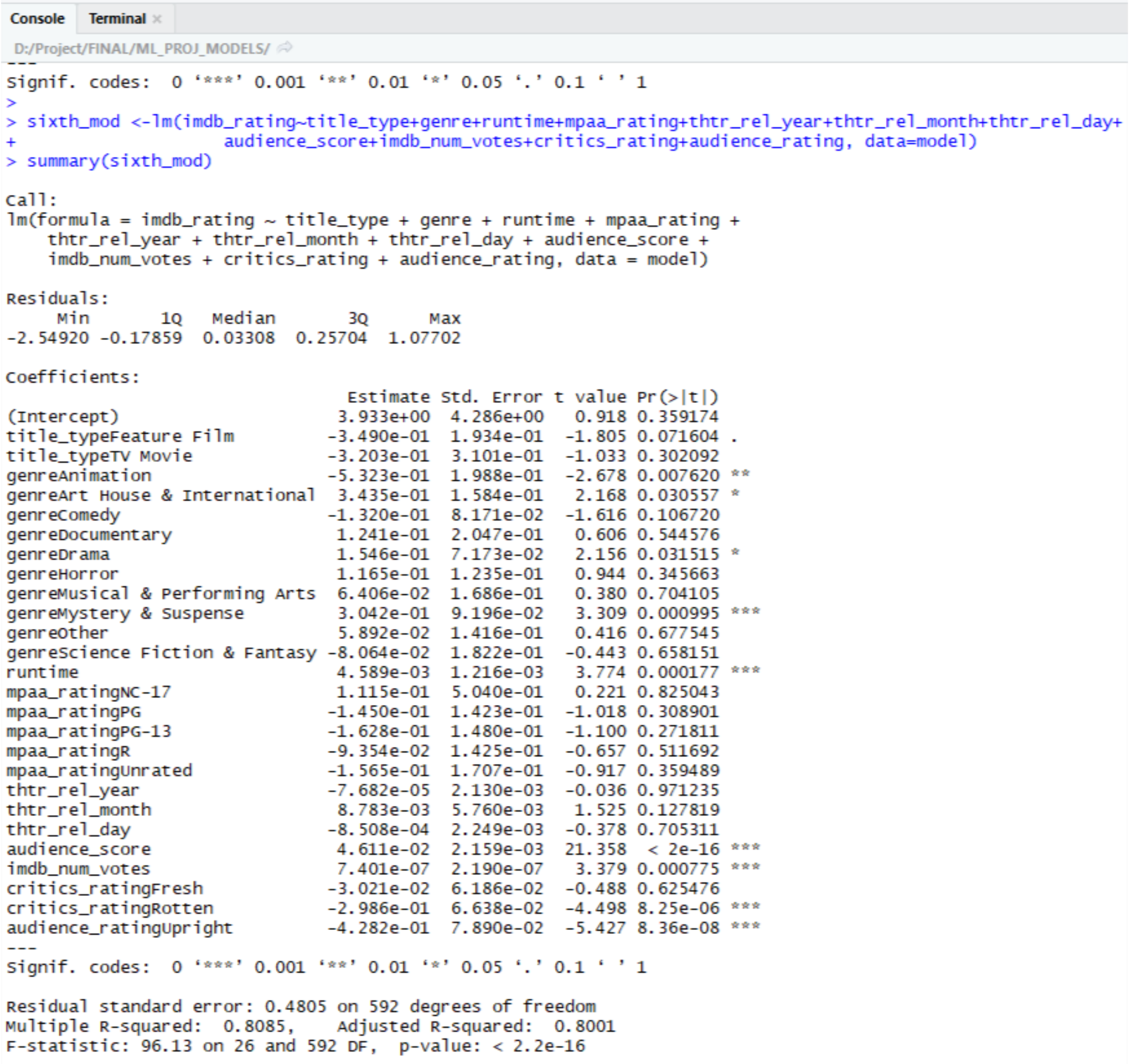
sixth_mod <- lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+thtr_rel_day+

audience_score+imdb_num_votes+critics_rating+audience_rating, data=model)

summary(sixth_mod)

anova(sixth_mod)

Figure 4.12 - Output for snippet 11: Model 6



```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> sixth_mod <-lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+thtr_rel_day+
+                     audience_score+imdb_num_votes+critics_rating+audience_rating, data=model)
> summary(sixth_mod)

Call:
lm(formula = imdb_rating ~ title_type + genre + runtime + mpaa_rating + 
    thtr_rel_year + thtr_rel_month + thtr_rel_day + audience_score + 
    imdb_num_votes + critics_rating + audience_rating, data = model)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.54920 -0.17859  0.03308  0.25704  1.07702 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    3.933e+00  4.286e+00   0.918 0.359174    
title_typeFeature Film        -3.490e-01  1.934e-01  -1.805 0.071604 .  
title_typeTV Movie            -3.203e-01  3.101e-01  -1.033 0.302092    
genreAnimation                -5.323e-01  1.988e-01  -2.678 0.007620 ** 
genreArt House & International  3.435e-01  1.584e-01   2.168 0.030557 *  
genreComedy                   -1.320e-01  8.171e-02  -1.616 0.106720    
genreDocumentary               1.241e-01  2.047e-01   0.606 0.544576    
genreDrama                     1.546e-01  7.173e-02   2.156 0.031515 *  
genreHorror                    1.165e-01  1.235e-01   0.944 0.345663    
genreMusical & Performing Arts  6.406e-02  1.686e-01   0.380 0.704105    
genreMystery & Suspense        3.042e-01  9.196e-02   3.309 0.000995 ***
genreOther                     5.892e-02  1.416e-01   0.416 0.677545    
genreScience Fiction & Fantasy -8.064e-02  1.822e-01  -0.443 0.658151    
runtime                        4.589e-03  1.216e-03   3.774 0.000177 ***
mpaa_ratingNC-17               1.115e-01  5.040e-01   0.221 0.825043    
mpaa_ratingPG                 -1.450e-01  1.423e-01  -1.018 0.308901    
mpaa_ratingPG-13              -1.628e-01  1.480e-01  -1.100 0.271811    
mpaa_ratingR                  -9.354e-02  1.425e-01  -0.657 0.511692    
mpaa_ratingUnrated            -1.565e-01  1.707e-01  -0.917 0.359489    
thtr_rel_year                 -7.682e-05  2.130e-03  -0.036 0.971235    
thtr_rel_month                 8.783e-03  5.760e-03   1.525 0.127819    
thtr_rel_day                  -8.508e-04  2.249e-03  -0.378 0.705311    
audience_score                 4.611e-02  2.159e-03  21.358  < 2e-16 ***
imdb_num_votes                 7.401e-07  2.190e-07   3.379 0.000775 ***
critics_ratingFresh           -3.021e-02  6.186e-02  -0.488 0.625476    
critics_ratingRotten          -2.986e-01  6.638e-02  -4.498 8.25e-06 ***
audience_ratingUpright        -4.282e-01  7.890e-02  -5.427 8.36e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4805 on 592 degrees of freedom
Multiple R-squared:  0.8085,	Adjusted R-squared:  0.8001 
F-statistic: 96.13 on 26 and 592 DF,  p-value: < 2.2e-16
```

Output for snippet 11: Model 6

Model-7:

Snippet 12:

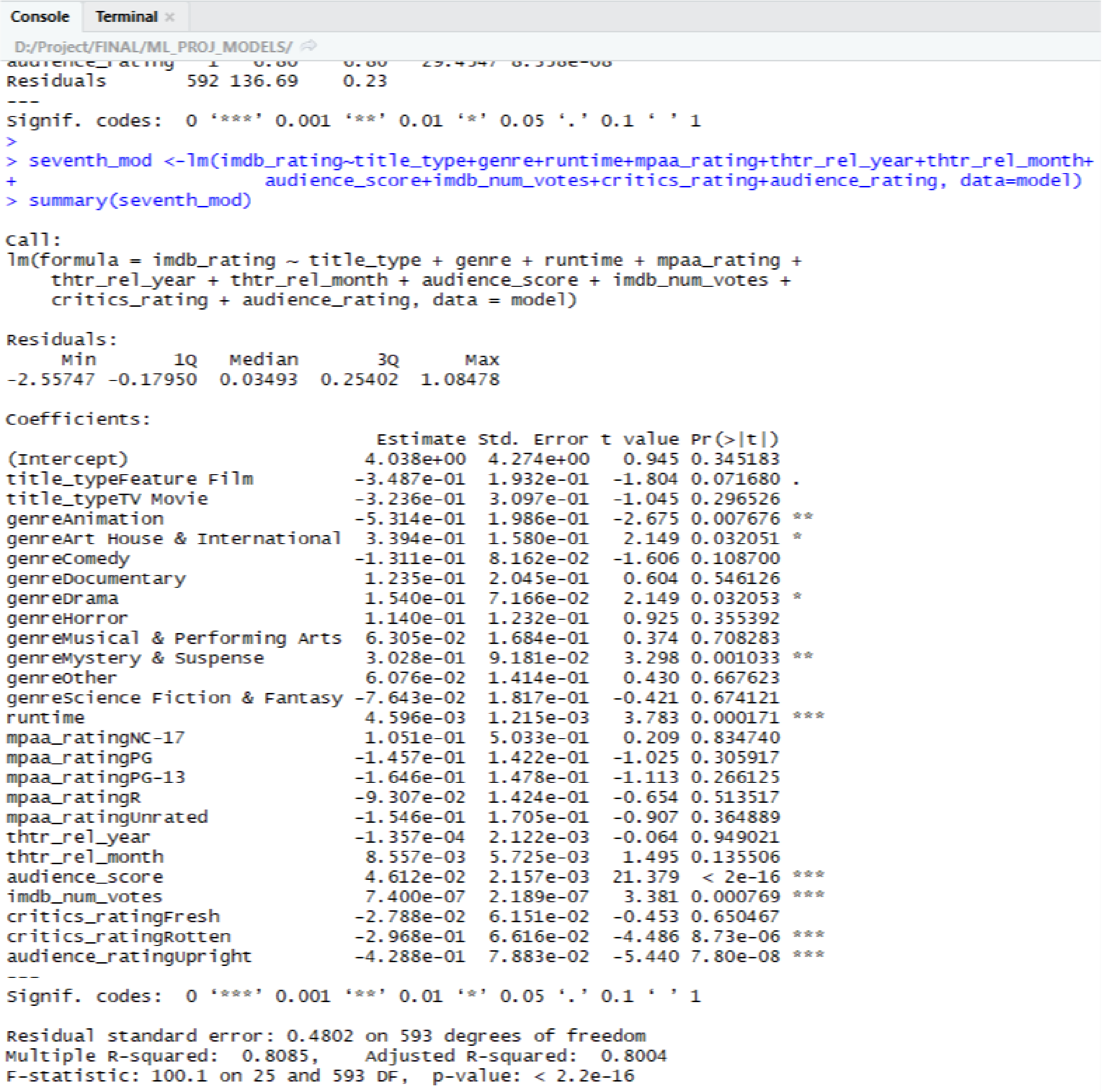
seventh_mod <- lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+

audience_score+imdb_num_votes+critics_rating+audience_rating, data=model)

summary(seventh_mod)

anova(seventh_mod)

Figure 4.13 - Output for snippet 12: Model 7



```
audience_rating   1   6.80     6.80   29.4347 8.358e-08
Residuals       592 136.69     0.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> seventh_mod <-lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+thtr_rel_month+
+                    audience_score+imdb_num_votes+critics_rating+audience_rating, data=model)
> summary(seventh_mod)

Call:
lm(formula = imdb_rating ~ title_type + genre + runtime + mpaa_rating + 
    thtr_rel_year + thtr_rel_month + audience_score + imdb_num_votes + 
    critics_rating + audience_rating, data = model)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.55747 -0.17950  0.03493  0.25402  1.08478 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     4.038e+00  4.274e+00   0.945 0.345183    
title_typeFeature Film         -3.487e-01  1.932e-01  -1.804 0.071680 .  
title_typeTV Movie             -3.236e-01  3.097e-01  -1.045 0.296526    
genreAnimation                 -5.314e-01  1.986e-01  -2.675 0.007676 ** 
genreArt House & International  3.394e-01  1.580e-01   2.149 0.032051 *  
genreComedy                    -1.311e-01  8.162e-02  -1.606 0.108700    
genreDocumentary                1.235e-01  2.045e-01   0.604 0.546126    
genreDrama                      1.540e-01  7.166e-02   2.149 0.032053 *  
genreHorror                     1.140e-01  1.232e-01   0.925 0.355392    
genreMusical & Performing Arts  6.305e-02  1.684e-01   0.374 0.708283    
genreMystery & Suspense         3.028e-01  9.181e-02   3.298 0.001033 ** 
genreOther                      6.076e-02  1.414e-01   0.430 0.667623    
genreScience Fiction & Fantasy -7.643e-02  1.817e-01  -0.421 0.674121    
runtime                         4.596e-03  1.215e-03   3.783 0.000171 ***
mpaa_ratingNC-17                1.051e-01  5.033e-01   0.209 0.834740    
mpaa_ratingPG                  -1.457e-01  1.422e-01  -1.025 0.305917    
mpaa_ratingPG-13               -1.646e-01  1.478e-01  -1.113 0.266125    
mpaa_ratingR                   -9.307e-02  1.424e-01  -0.654 0.513517    
mpaa_ratingUnrated             -1.546e-01  1.705e-01  -0.907 0.364889    
thtr_rel_year                  -1.357e-04  2.122e-03  -0.064 0.949021    
thtr_rel_month                  8.557e-03  5.725e-03   1.495 0.135506    
audience_score                  4.612e-02  2.157e-03  21.379  < 2e-16 ***
imdb_num_votes                  7.400e-07  2.189e-07   3.381 0.000769 ***
critics_ratingFresh            -2.788e-02  6.151e-02  -0.453 0.650467    
critics_ratingRotten           -2.968e-01  6.616e-02  -4.486 8.73e-06 ***
audience_ratingUpright         -4.288e-01  7.883e-02  -5.440 7.80e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4802 on 593 degrees of freedom
Multiple R-squared:  0.8085,	Adjusted R-squared:  0.8004 
F-statistic: 100.1 on 25 and 593 DF,  p-value: < 2.2e-16
```

Output for snippet 12: Model 7

Model-8:

Snippet 13:

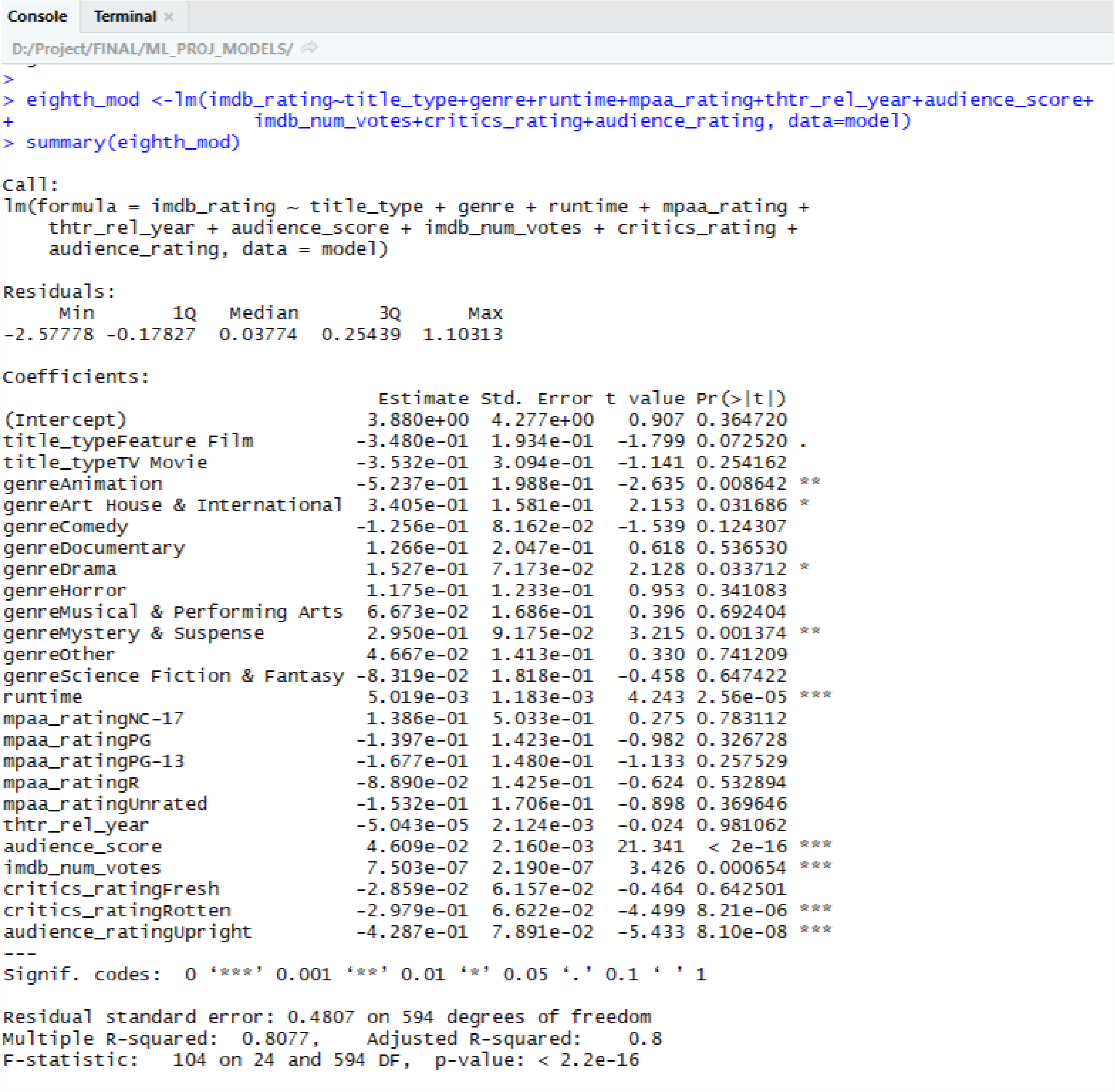
eighth_mod <- lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+audience_score+

imdb_num_votes+critics_rating+audience_rating, data=model)

summary(eighth_mod)

anova(eighth_mod)

Figure 4.14 - Output for snippet 13: Model 8



```
>
> eighth_mod <-lm(imdb_rating~title_type+genre+runtime+mpaa_rating+thtr_rel_year+audience_score+
+                      imdb_num_votes+critics_rating+audience_rating, data=model)
> summary(eighth_mod)

Call:
lm(formula = imdb_rating ~ title_type + genre + runtime + mpaa_rating + 
    thtr_rel_year + audience_score + imdb_num_votes + critics_rating + 
    audience_rating, data = model)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.57778 -0.17827  0.03774  0.25439  1.10313 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    3.880e+00  4.277e+00   0.907 0.364720    
title_typeFeature Film        -3.480e-01  1.934e-01  -1.799 0.072520 .  
title_typeTV Movie            -3.532e-01  3.094e-01  -1.141 0.254162    
genreAnimation                -5.237e-01  1.988e-01  -2.635 0.008642 ** 
genreArt House & International  3.405e-01  1.581e-01   2.153 0.031686 *  
genreComedy                   -1.256e-01  8.162e-02  -1.539 0.124307    
genreDocumentary               1.266e-01  2.047e-01   0.618 0.536530    
genreDrama                     1.527e-01  7.173e-02   2.128 0.033712 *  
genreHorror                    1.175e-01  1.233e-01   0.953 0.341083    
genreMusical & Performing Arts  6.673e-02  1.686e-01   0.396 0.692404    
genreMystery & Suspense        2.950e-01  9.175e-02   3.215 0.001374 ** 
genreOther                     4.667e-02  1.413e-01   0.330 0.741209    
genreScience Fiction & Fantasy -8.319e-02  1.818e-01  -0.458 0.647422    
runtime                        5.019e-03  1.183e-03   4.243 2.56e-05 ***
mpaa_ratingNC-17               1.386e-01  5.033e-01   0.275 0.783112    
mpaa_ratingPG                 -1.397e-01  1.423e-01  -0.982 0.326728    
mpaa_ratingPG-13              -1.677e-01  1.480e-01  -1.133 0.257529    
mpaa_ratingR                  -8.890e-02  1.425e-01  -0.624 0.532894    
mpaa_ratingUnrated            -1.532e-01  1.706e-01  -0.898 0.369646    
thtr_rel_year                 -5.043e-05  2.124e-03  -0.024 0.981062    
audience_score                 4.609e-02  2.160e-03  21.341  < 2e-16 ***
imdb_num_votes                 7.503e-07  2.190e-07   3.426 0.000654 ***
critics_ratingFresh           -2.859e-02  6.157e-02  -0.464 0.642501    
critics_ratingRotten          -2.979e-01  6.622e-02  -4.499 8.21e-06 ***
audience_ratingUpright        -4.287e-01  7.891e-02  -5.433 8.10e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4807 on 594 degrees of freedom
Multiple R-squared:  0.8077,	Adjusted R-squared:    0.8 
F-statistic:   104 on 24 and 594 DF,  p-value: < 2.2e-16
```

Output for snippet 13: Model 8

Model-9:

Snippet 14:

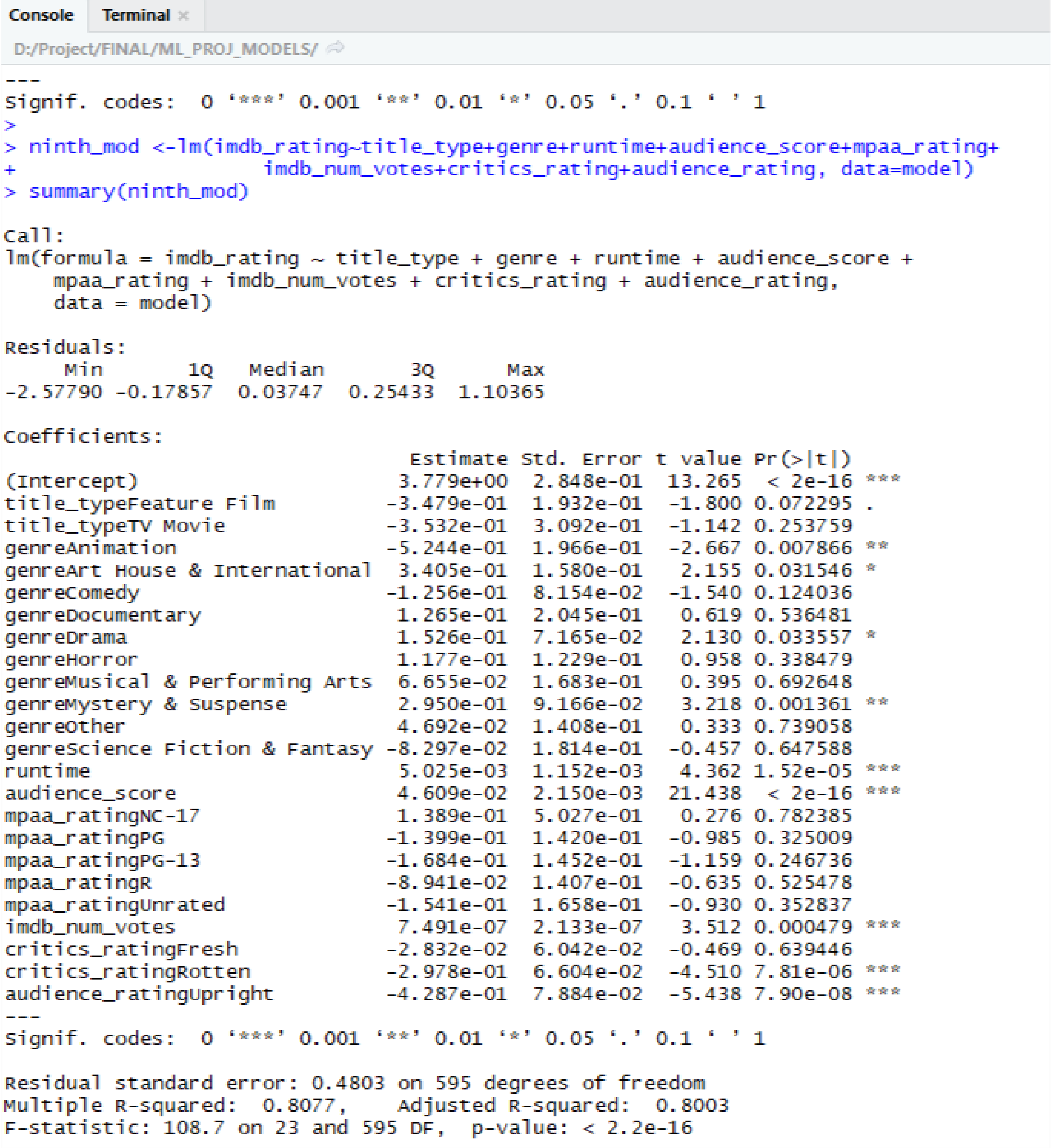
ninth_mod <-lm(imdb_rating~title_type+genre+runtime+audience_score+mpaa_rating+

imdb_num_votes+critics_rating+audience_rating, data=model)

summary(ninth_mod)

anova(ninth_mod)

Figure 4.15 - Output for snippet 14: Model 9



```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ninth_mod <-lm(imdb_rating~title_type+genre+runtime+audience_score+mpaa_rating+
+                     imdb_num_votes+critics_rating+audience_rating, data=model)
> summary(ninth_mod)

Call:
lm(formula = imdb_rating ~ title_type + genre + runtime + audience_score + 
    mpaa_rating + imdb_num_votes + critics_rating + audience_rating, 
    data = model)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.57790 -0.17857  0.03747  0.25433  1.10365 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     3.779e+00  2.848e-01  13.265  < 2e-16 ***
title_typeFeature Film         -3.479e-01  1.932e-01  -1.800 0.072295 .  
title_typeTV Movie             -3.532e-01  3.092e-01  -1.142 0.253759    
genreAnimation                 -5.244e-01  1.966e-01  -2.667 0.007866 ** 
genreArt House & International  3.405e-01  1.580e-01   2.155 0.031546 *  
genreComedy                    -1.256e-01  8.154e-02  -1.540 0.124036    
genreDocumentary                1.265e-01  2.045e-01   0.619 0.536481    
genreDrama                      1.526e-01  7.165e-02   2.130 0.033557 *  
genreHorror                     1.177e-01  1.229e-01   0.958 0.338479    
genreMusical & Performing Arts  6.655e-02  1.683e-01   0.395 0.692648    
genreMystery & Suspense         2.950e-01  9.166e-02   3.218 0.001361 ** 
genreOther                      4.692e-02  1.408e-01   0.333 0.739058    
genreScience Fiction & Fantasy -8.297e-02  1.814e-01  -0.457 0.647588    
runtime                         5.025e-03  1.152e-03   4.362 1.52e-05 ***
audience_score                  4.609e-02  2.150e-03  21.438  < 2e-16 ***
mpaa_ratingNC-17                1.389e-01  5.027e-01   0.276 0.782385    
mpaa_ratingPG                  -1.399e-01  1.420e-01  -0.985 0.325009    
mpaa_ratingPG-13               -1.684e-01  1.452e-01  -1.159 0.246736    
mpaa_ratingR                   -8.941e-02  1.407e-01  -0.635 0.525478    
mpaa_ratingUnrated             -1.541e-01  1.658e-01  -0.930 0.352837    
imdb_num_votes                  7.491e-07  2.133e-07   3.512 0.000479 ***
critics_ratingFresh            -2.832e-02  6.042e-02  -0.469 0.639446    
critics_ratingRotten           -2.978e-01  6.604e-02  -4.510 7.81e-06 ***
audience_ratingUpright         -4.287e-01  7.884e-02  -5.438 7.90e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4803 on 595 degrees of freedom
Multiple R-squared:  0.8077,	Adjusted R-squared:  0.8003 
F-statistic: 108.7 on 23 and 595 DF,  p-value: < 2.2e-16
```

Output for snippet 14: Model 9

Model-10:

Snippet 15:

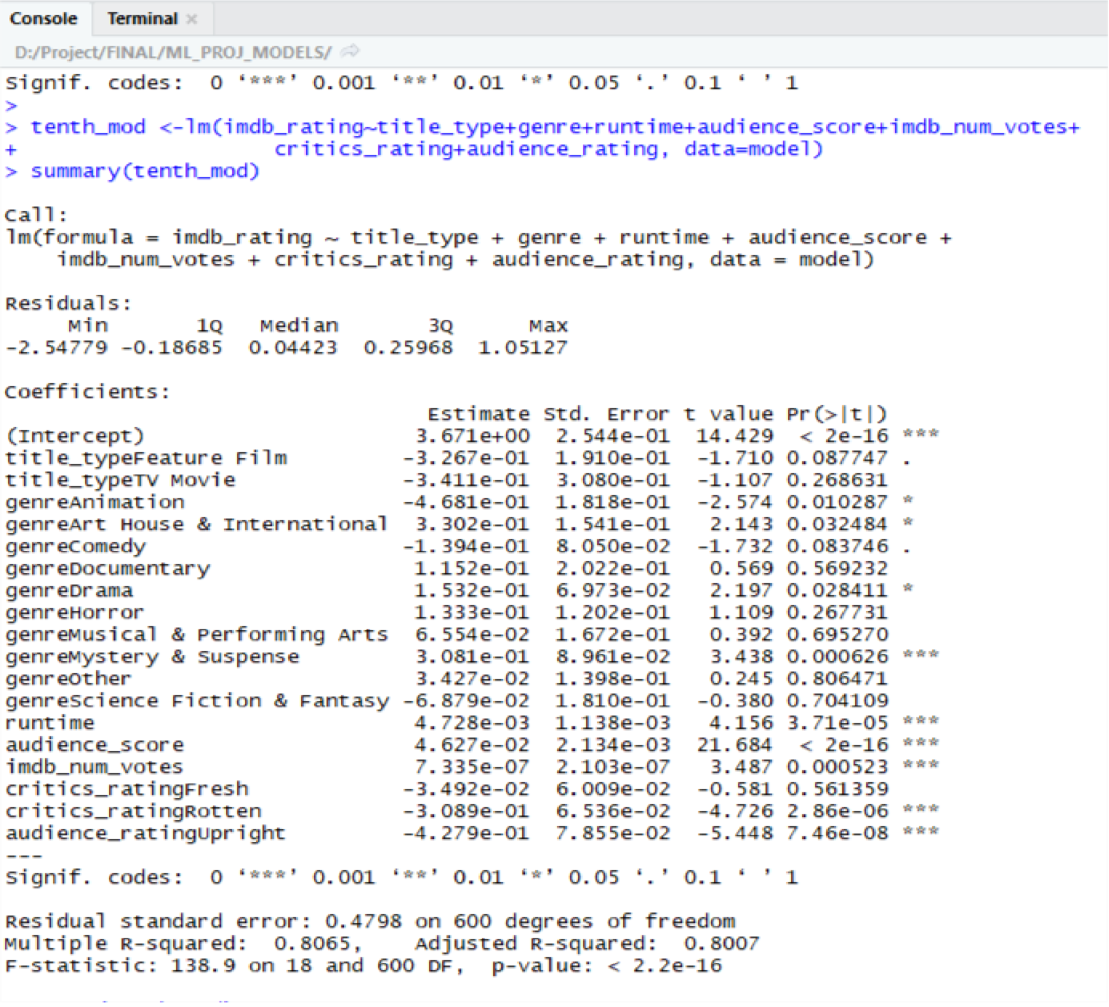
tenth_mod <-lm(imdb_rating~title_type+genre+runtime+audience_score+imdb_num_votes+

critics_rating+audience_rating, data=model)

summary(tenth_mod)

anova(tenth_mod)

Figure 4.16 - Output for snippet 15: Model 10



```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> tenth_mod <-lm(imdb_rating~title_type+genre+runtime+audience_score+imdb_num_votes+
+                     critics_rating+audience_rating, data=model)
> summary(tenth_mod)

Call:
lm(formula = imdb_rating ~ title_type + genre + runtime + audience_score +
    imdb_num_votes + critics_rating + audience_rating, data = model)

Residuals:
     Min       1Q   Median       3Q      Max
-2.54779 -0.18685  0.04423  0.25968  1.05127

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.671e+00  2.544e-01  14.429  < 2e-16 ***
title_typeFeature Film        -3.267e-01  1.910e-01  -1.710 0.087747 .
title_typeTV Movie            -3.411e-01  3.080e-01  -1.107 0.268631
genreAnimation                -4.681e-01  1.818e-01  -2.574 0.010287 *
genreArt House & International 3.302e-01  1.541e-01   2.143 0.032484 *
genreComedy                   -1.394e-01  8.050e-02  -1.732 0.083746 .
genreDocumentary               1.152e-01  2.022e-01   0.569 0.569232
genreDrama                     1.532e-01  6.973e-02   2.197 0.028411 *
genreHorror                    1.333e-01  1.202e-01   1.109 0.267731
genreMusical & Performing Arts 6.554e-02  1.672e-01   0.392 0.695270
genreMystery & Suspense        3.081e-01  8.961e-02   3.438 0.000626 ***
genreOther                     3.427e-02  1.398e-01   0.245 0.806471
genreScience Fiction & Fantasy -6.879e-02  1.810e-01  -0.380 0.704109
runtime                        4.728e-03  1.138e-03   4.156 3.71e-05 ***
audience_score                 4.627e-02  2.134e-03  21.684  < 2e-16 ***
imdb_num_votes                 7.335e-07  2.103e-07   3.487 0.000523 ***
critics_ratingFresh           -3.492e-02  6.009e-02  -0.581 0.561359
critics_ratingRotten          -3.089e-01  6.536e-02  -4.726 2.86e-06 ***
audience_ratingUpright        -4.279e-01  7.855e-02  -5.448 7.46e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4798 on 600 degrees of freedom
Multiple R-squared:  0.8065,    Adjusted R-squared:  0.8007
F-statistic: 138.9 on 18 and 600 DF,  p-value: < 2.2e-16
```

Output for snippet 15:Model 10

The last model has the highest adj R2 of 0.8007 and all predictors are statistically significant.

Final Model:

Snippet 16:

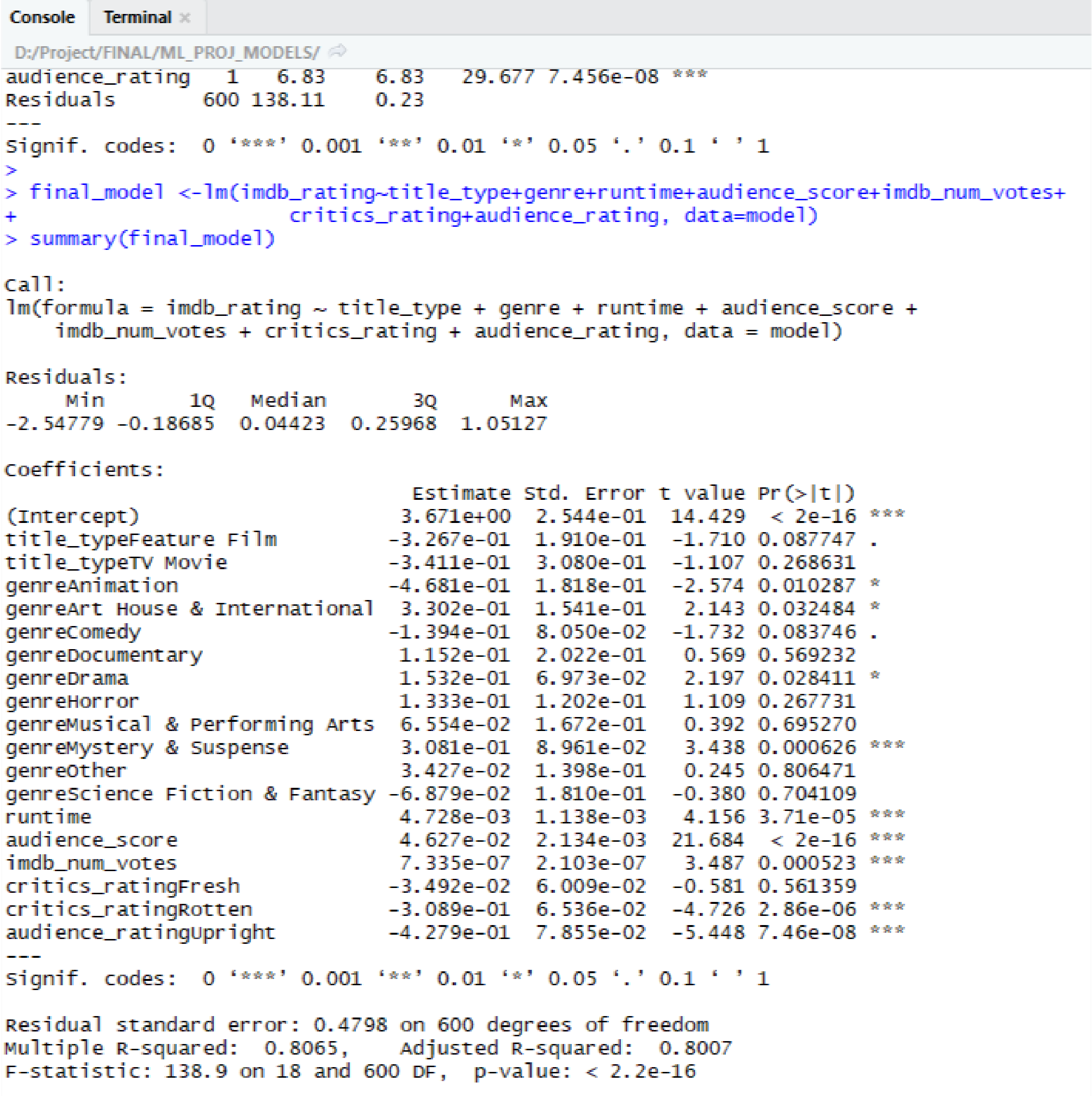
final_model <-lm(imdb_rating~title_type+genre+runtime+audience_score+imdb_num_votes+

critics_rating+audience_rating, data=model)

summary(final_model)

anova(final_model)

Figure 4.17 - Output for snippet 16: Final model



```
audience_rating   1   6.83    6.83   29.677 7.456e-08 ***
Residuals       600 138.11    0.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> final_model <-lm(imdb_rating~title_type+genre+runtime+audience_score+imdb_num_votes+
+                        critics_rating+audience_rating, data=model)
> summary(final_model)

Call:
lm(formula = imdb_rating ~ title_type + genre + runtime + audience_score + 
    imdb_num_votes + critics_rating + audience_rating, data = model)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.54779 -0.18685  0.04423  0.25968  1.05127 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     3.671e+00  2.544e-01  14.429  < 2e-16 ***
title_typeFeature Film         -3.267e-01  1.910e-01  -1.710 0.087747 .  
title_typeTV Movie             -3.411e-01  3.080e-01  -1.107 0.268631    
genreAnimation                 -4.681e-01  1.818e-01  -2.574 0.010287 *  
genreArt House & International  3.302e-01  1.541e-01   2.143 0.032484 *  
genreComedy                    -1.394e-01  8.050e-02  -1.732 0.083746 .  
genreDocumentary                1.152e-01  2.022e-01   0.569 0.569232    
genreDrama                      1.532e-01  6.973e-02   2.197 0.028411 *  
genreHorror                     1.333e-01  1.202e-01   1.109 0.267731    
genreMusical & Performing Arts  6.554e-02  1.672e-01   0.392 0.695270    
genreMystery & Suspense         3.081e-01  8.961e-02   3.438 0.000626 ***
genreOther                      3.427e-02  1.398e-01   0.245 0.806471    
genreScience Fiction & Fantasy -6.879e-02  1.810e-01  -0.380 0.704109    
runtime                         4.728e-03  1.138e-03   4.156 3.71e-05 ***
audience_score                  4.627e-02  2.134e-03  21.684  < 2e-16 ***
imdb_num_votes                  7.335e-07  2.103e-07   3.487 0.000523 ***
critics_ratingFresh            -3.492e-02  6.009e-02  -0.581 0.561359    
critics_ratingRotten           -3.089e-01  6.536e-02  -4.726 2.86e-06 ***
audience_ratingUpright         -4.279e-01  7.855e-02  -5.448 7.46e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4798 on 600 degrees of freedom
Multiple R-squared:  0.8065,	Adjusted R-squared:  0.8007 
F-statistic: 138.9 on 18 and 600 DF,  p-value: < 2.2e-16
```

Output for snippet 16: Final model

Inference:

After going through the iterative process above, the researcher has finally reached a model that fits a parsimonious concept. This last model has the highest predictive power (i.e.adj R2 of 0.8007) and all predictors are statistically significant (i.e. close to zero p-value). In summary, we can reject our null hypothesis that all the coefficients are indifferent and accept our alternative hypothesis that one of these coefficients are not equal.

4.6 Prediction

Now that the final model has been developed, this part is to assess its predictive capability. I have selected two movies which are not in the sample but prior to 2016 release. We will use the model developed to predict the IMDB rating. Two movies are chosen from both IMDB & Rotten Tomatoes websites.

Movie 1: Armageddon (1998)

Actual IMDb Rating: 6.6

http://www.imdb.com/title/tt0120591/?ref_=nv_sr_2

https://www.rottentomatoes.com/m/armageddon

Movie #2: The Smurfs (2011)

Actual IMDb Rating: 5.5

http://www.imdb.com/title/tt0472181/?ref_=nv_sr_3

https://www.rottentomatoes.com/m/the_smurfs

Characteristics	**Armageddon**	**The Smurfs**
Thtr_rel_year	1998	2011
Title_type	Feature Film	Feature Film
Genre	Action & Adventure	Animation
Runtime	151	103

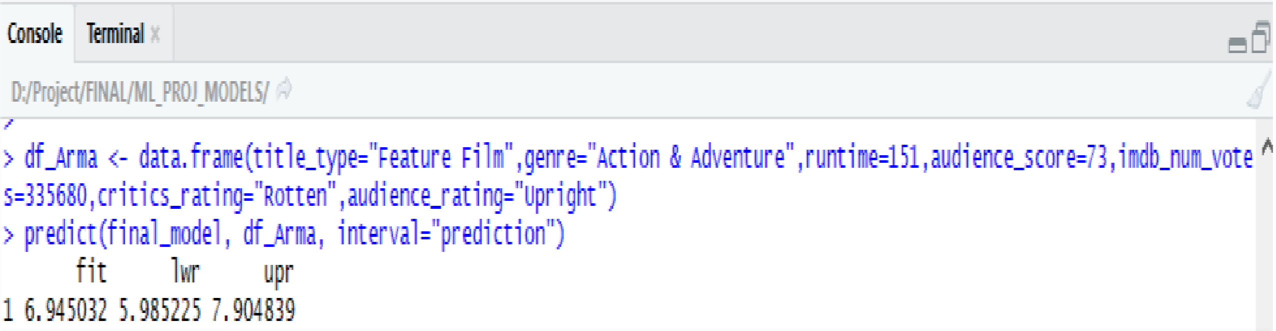
Characteristics	Armageddon	The Smurfs
Audience Score	73	44
IMDB Num Votes	335,680	69,572
Critics Rating	Rotten	Rotten
Audience Rating	Upright	Spilled

Snippet 17: (movie 1)

df_Arma <- data.frame(title_type="Feature Film",genre="Action & Adventure",runtime=151,audience_score=73,imdb_num_votes=335680,critics_rating="Rotten",audience_rating="Upright")

predict(final_model, df_Arma, interval="prediction")

Output for snippet 17:



```
> df_Arma <- data.frame(title_type="Feature Film",genre="Action & Adventure",runtime=151,audience_score=73,imdb_num_vote
s=335680,critics_rating="Rotten",audience_rating="Upright")
> predict(final_model, df_Arma, interval="prediction")
       fit      lwr      upr
1 6.945032 5.985225 7.904839
```

Figure 4.18 – output for snippet 17

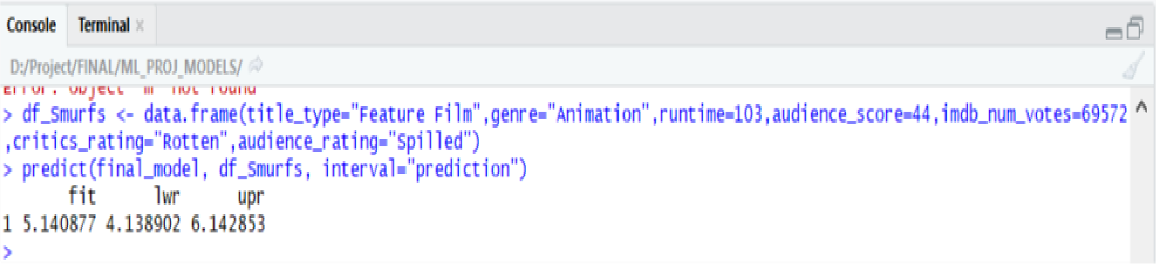
Inference:

The predicted value is 6.9 which is very close to the actual IMDb rating of 6.6. In fact, with 95% confidence interval, the actual IMDb_rating for Armageddon has a lower bound of 5.98 and a higher bound of 7.9.

Snippet 18: (movie 2)

df_Smurfs <- data.frame(title_type="Feature Film",genre="Animation",runtime=103,audience_score=44,imdb_num_votes=69572,critics_rating="Rotten",audience_rating="Spilled") *predict(final_model, df_Smurfs, interval="prediction")*

Output for snippet 18:



```
Error: object 'm' not found
> df_Smurfs <- data.frame(title_type="Feature Film",genre="Animation",runtime=103,audience_score=44,imdb_num_votes=69572
,critics_rating="Rotten",audience_rating="Spilled")
> predict(final_model, df_Smurfs, interval="prediction")
       fit      lwr      upr
1 5.140877 4.138902 6.142853
>
```

Figure 4.19 – output for snippet 18

Inference:

The predicted value is 5.1 which is very close to the actual IMDb rating of 5.5. In fact, with 95% confidence interval, the actual IMDb_rating for The Smurfs has a lower bound of 4.14 and a higher bound of 6.14.

The chapter has explained about the analysis done for the data collected. The interpretations were done by the researcher through the experiences during the internship. All the tests for analysis the hypothesis have been done by R Studio. The interpretation for all the analysis has been used to derive Findings, recommendations, and conclusion of the study which has been dealt in the following chapter.

CHAPTER V

FINDINGS, RECOMMENDATIONS AND CONCLUSION

The chapter concludes the study. It describes all the findings, recommendations and suggestions for the study. The researcher has analysed the data and found some inferences according to the objective of the study which were taken from the previous chapter. Suggestions and recommendation were proposed to improve the movie popularity as well as success of the film.

5.1 FINDINGS

The following are the major findings arrived from the analysis,

- From the data analysis, it has been found that the various movie attribute plays a vital role with respect to perception of viewer's towards movie rating.
- From the regression model, it has been found that there are strong association on certain specific characteristics of a movie to drive success.
- From the reduced version in data cleaning part **(Snippet 3),** it has been found that the following movie attributes are only related to movie rating,
    1. Title
    2. Title type
    3. Genre
    4. Runtime
    5. Mpaa rating
    6. Theatre release year
    7. Theatre release month
    8. Theatre release day
    9. IMDb rating
    10. Critics score
    11. Audience score
    12. Critics rating
    13. Audience rating
    14. IMDb number of votes
    15. Best picture nominee
    16. Best picture win
    17. Best actor win
    18. Best actress win

19. Best director win
20. Top 200 box

- With the help of backward stepwise elimination method, the final model **(snippet 16)** achieves an accuracy level of 80% at confidence interval of 95%.
- From the collinearity checking **(snippet 5),** it has been found that there is a linear relationship between IMDb and audience score.
- From the prediction part **(Snippet 17 & 18)**, it has been proved that the actual movie rating and predicted rating are very close to each other. Hence, the model has very strong prediction power.
- When analyzing the various movie attributes with IMDb movie rating, it has been found that the budget of the movie can also be added so that the revenue can also be predicted.

5.2 RECOMMENDATIONS

- Since the attributes which drives the movie to success is identified, this provides an interesting insight to the producers before they decide what movie to be launched and the likelihood of its success in public eyes.
- Movie industries can use the similar methods when producing movies that are more likely to be liked by the target audience.
- The production team of the movie can use such type of model to make decision related to the release of the movie, since the release date plays a vital role in user movie rating for unreleased movies.
- Various stakeholders such as actors, actress, producers, directors etc. can use these predictions to make more informed decisions. They can make the decision before the movie release in order make the movie success and also generate more revenues.
- The attributes that contributed the most to movie rating are number of votes for each movie, runtime of the movie and the genre of the movie in particularly the action & adventurous movies are more successful.

5.3 SUGGESTIONS FOR FURTHER STUDY

- The researcher has selected a limited number of samples for the analysis due to time constraint. Hence it would be suggested to have a large sample size for developing a predictive model in future.
- Likewise, it might also be of interest to evaluate whether or not the same is applicable to box office predictions, to further widen the scope of the study.

- A movie success does not only depend on features related to movies. Number of audiences plays very important role for a movie to become successful. Because the whole point is about audiences, the whole industry will make no sense if there is no audience to watch a movie. Hence, in the future work consider the various attributes such as number of tickets sold during a specific year.

5.4 CONCLUSION

The proposed research aims to predict the IMDb rating for movies with the help of various movie attributes. The researcher has used machine learning approach for the experimentation. Machine learning have powerful predictive algorithms for regression models. The research aims to improve previous researches. Performing data mining on IMDb dataset is a hard task because of so many attributes related to a movie and all in different dimensions with lots of noisy data and missing fields. After performing regression analysis, the researcher has found out that results are achieved far better with backward stepwise elimination method and linear regression at around 80% of accuracy at a confident level of 95%. When trying to predict the IMDb rating of a movie a highest success rate of 80% was achieved, this result show that the method used in this study are difficult use for an accurate prediction.

Through the study the researcher was able to analyse the problem and find possible solution using R Studio which is a statistical tool cum programming language.

References

Augustine, A. (2015). *User rating prediction for movies.* Texas: University of Texas at austin.

Grodman, J. E. (2015). *A predictor for movie success.* California: Standford Univeristy.

Irizarry, R. A. (2018, December). *Introduction to R: Explore the Data Frame.* Retrieved from Edx: https://www.edx.org

Irizarry, R. A. (2019). *Introduction to Data Science.* Texas, United States of America: HarvardX . Retrieved from https://rafalab.github.io/dsbook/

Meenakshi, K. (2018). A Data mining Technique for Analyzing and predicting the success of Movie. *Journal of Physics: Conference Series*, 4-10. Retrieved from https://doi.org/10.1088/1742-6596/1000/1/012100

Quader, N. (2015). *A machine learning approch to predict box office success.* School of Engineering & Computer Science , Department of Computer Science & Engineering. Mohakhali, Bangladesh: BRAC University.