National College of Ireland

Postgraduate Diploma in Science in Data Analytics (PGDDA\_SEPOL)

Database and Analytics Programming Project

Instructor Name

Student Name

Date

## 3.2 Research Questions

To derive insights from the Netflix dataset used in the analysis, this kind of project will require research questions. The research questions will play a great role by acting as the target point of my study case. I will focus on the research questions to derive those useful insights which are necessary in improving Netflix company financially and its general performance so as to maintain its position on the market like its competitors.

Below are the research questions derived to ensure that the research goes successful and doing the data analysis by help of machine learning and deep learning techniques;

1. What are the most watch shows on Netflix based on the contribution of Artificial intelligence integration to its search engine?
2. How is the rating from customer’s increases customer turnover?
3. How does customer satisfaction impact rating and financial growth of the company?

## 4. Methodology

## 4.1 Information About Dataset

To perform my data analysis using artificial intelligence techniques, I had chosen to perform the analysis for the data collected about Netflix Company. In simple terms, Netflix is a streaming service that costs money to use. This also means that you can also download TV shows and movies to your iOS, Android, or Windows 10 device so that you can watch them without having to be online. The Subscribers to Netflix can watch TV shows and movies without commercials on an internet-connected device.

### 4.1.1 About Netflix Dataset

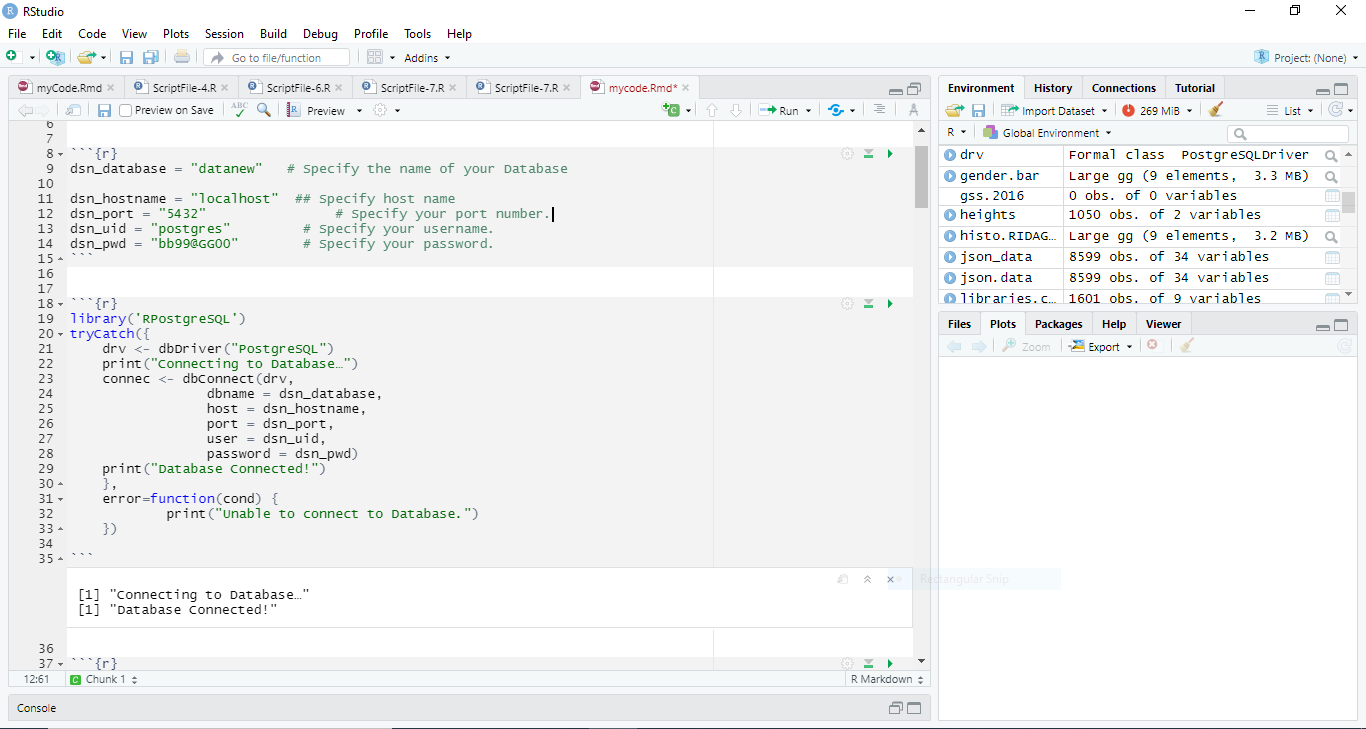
Netflix is one of the most popular ways to watch media and videos. They have more than 8000 movies and TV shows (the dataset observations available ) on their platform. As of mid-2021, they have more than 200 million subscribers around the world. This is a list of all the movies and TV shows that are available on Netflix. It includes information about the cast, directors, ratings, release date, length, rating, description, listed\_in, country and many more. These are the main features of the dataset that are used in analysis by the help of artificial intelligence skills to come up with the insights hidden in such dataset which is of more useful when implemented by Netflix company (Narayanan & Shmatikov, 2016).

During the implementation of the analysis on the Netflix dataset, several methods/techniques and steps were involved depending on their various use cases in the study. R programming is the language used during implementation.

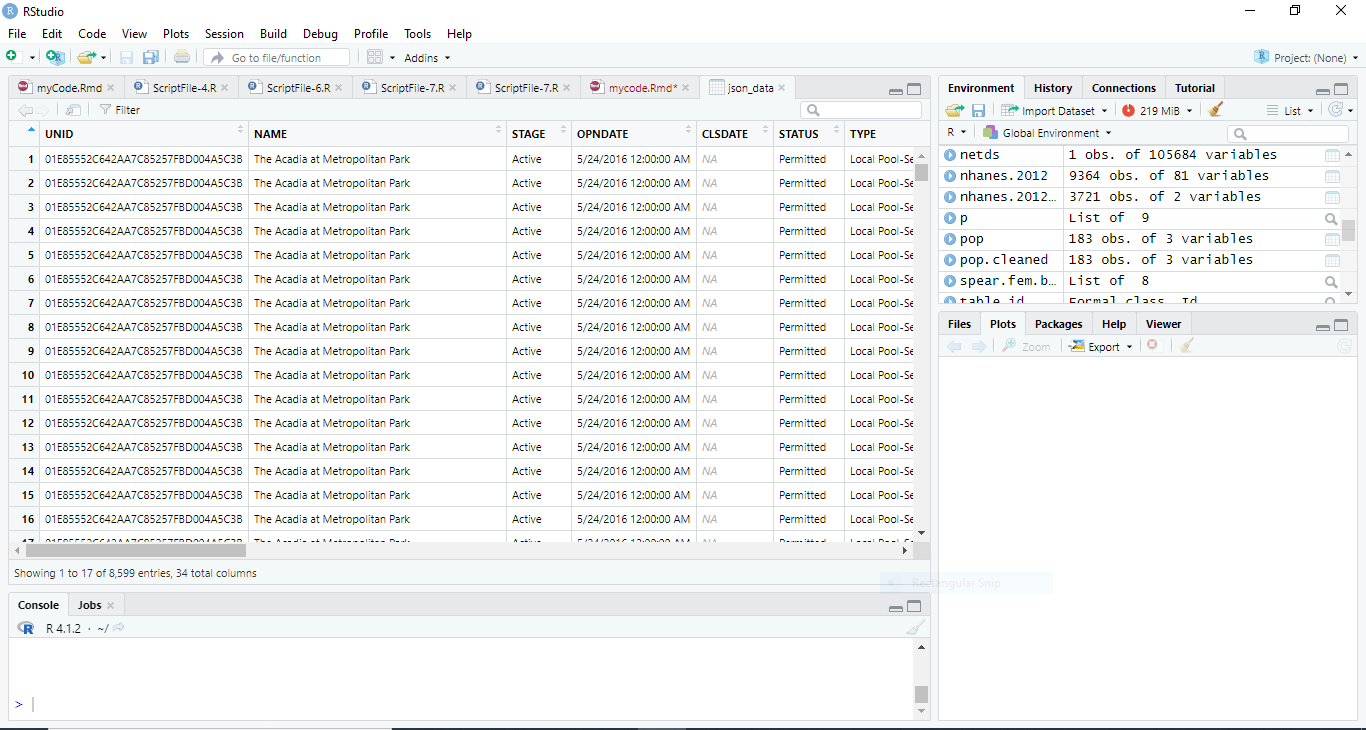
Firstly, I performed data reading. This involves loading dataset information into my working environment from postgreSQL database by help of various R packages(using fromJSON() function to read the JSON dataset file since the dataset was stored in a json format). Second part, I did dataset summarization. This was useful in understanding the dataset(the features available and their description before I perform analyses on the dataset. At this point, the data scientist is in a good position to explore the dataset fully and understand the observations in the whole dataset, its features, the different data types used in the dataset as well as the dominated data type. Then, there was need for data cleaning. This is the basic fundamental step that one needs to perform before analyzing the data to get insights.

According to (Dasu & Johnson, 2017), data cleaning takes 80% of a data scientist during the data analysis period on any given dataset. This is because, for accuracy to be found in the prediction made during data analysis, we need data that is well normalized and standardized. This can be achieved by performing data cleaning which removes outliers in the dataset, the NAN values in the observations and features of the dataset. In addition, it involves changing the data types of the feature to their respective formats. For example, the target variable should always be categorical therefore in many cases, you will find the target variable in a character format, this needs to be changed to factor data type to allow better analysis of the dataset. By implementing all this, the analysis done will be accurate just as expected.

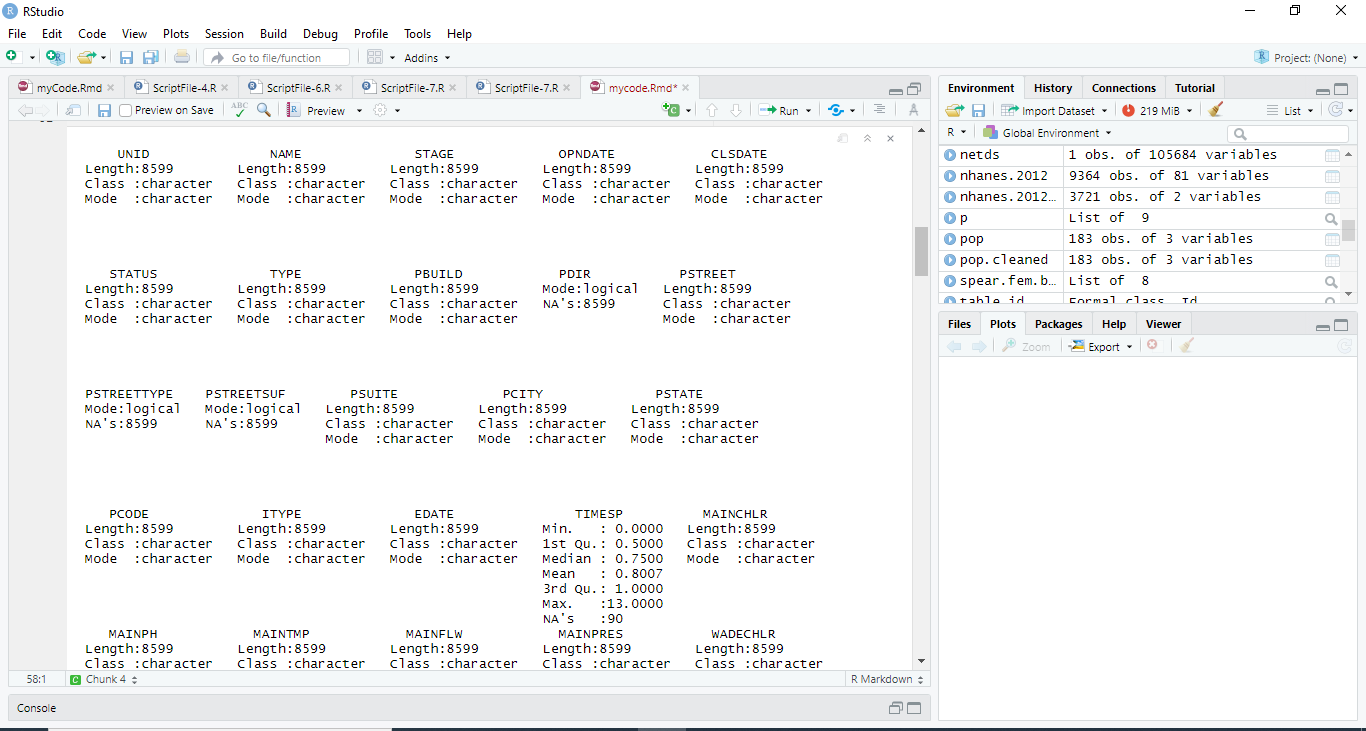
I involved the use of various techniques. In this project decision tree machine learning algorithm is the preferred one. I have used Decision tree algorithm to perform some of the classifications. However, I have several visualizations in the data analysis to indicate some insights that Netflix company needs. These visualizations have been achieved by the help of ggplot2 library, a library in R for performing all types of graphics. In simple terms, ggplot2 is the mother of visualization in R programming language (Taylor & Cihon, 2018).



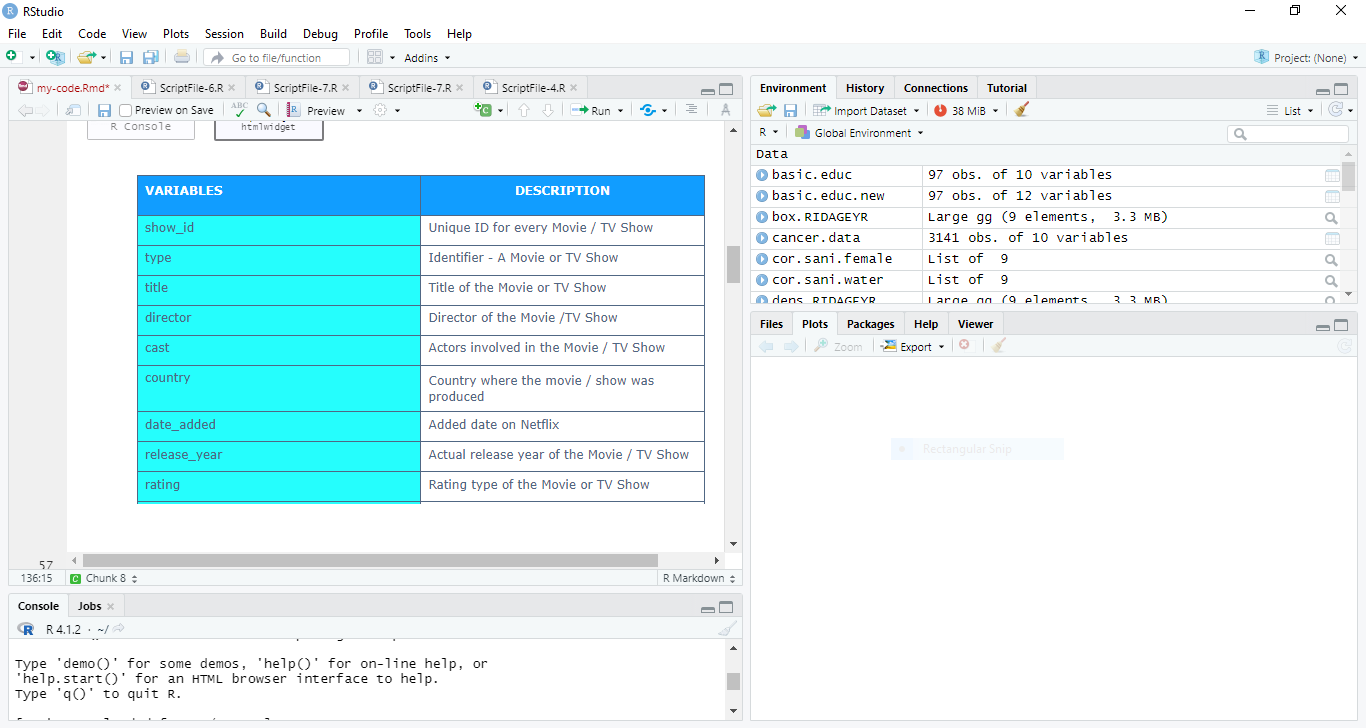
*Fig 1.0. Shows the connection of PostgreSQL in Rstudio*



*Fig 1.1. Shows the View of the json dataset in Rstudio in tabular format*

**

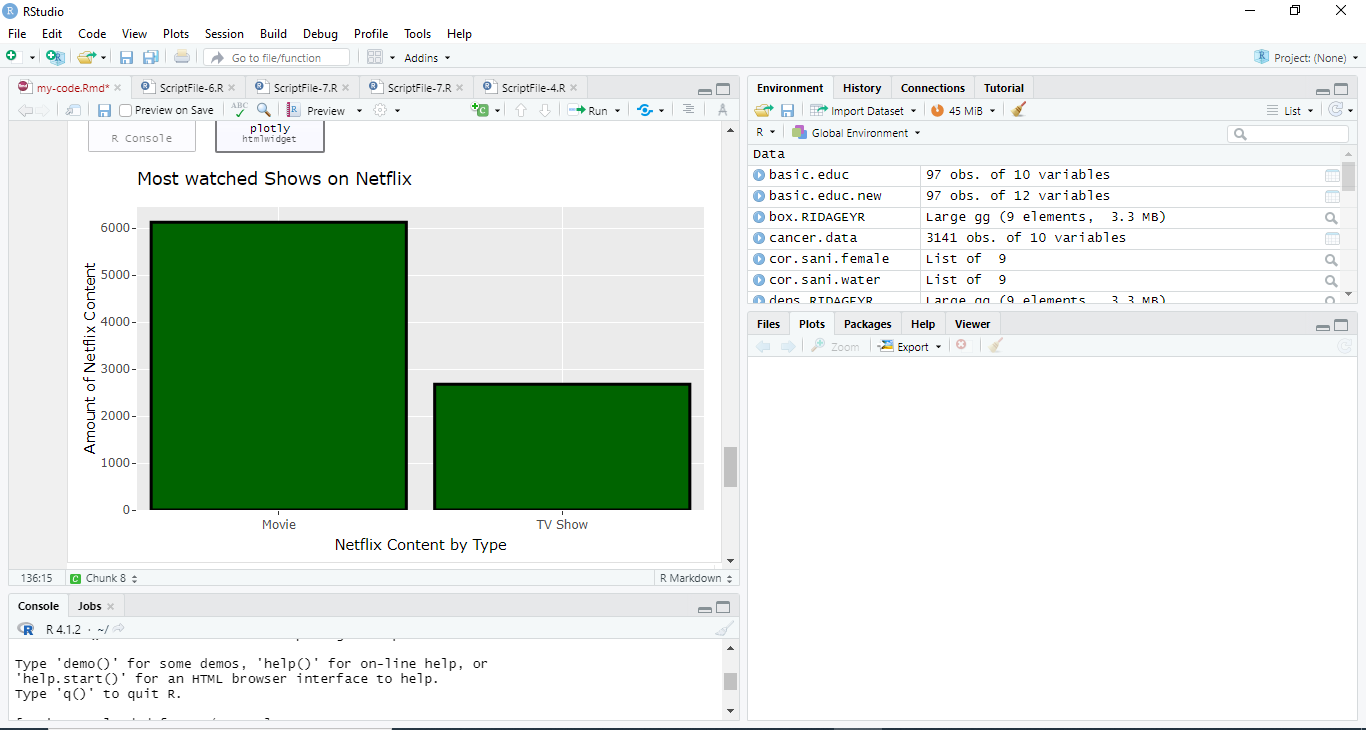
*. Fig 1.2. Shows the View of the json dataset in Rstudio in tabular format*

**

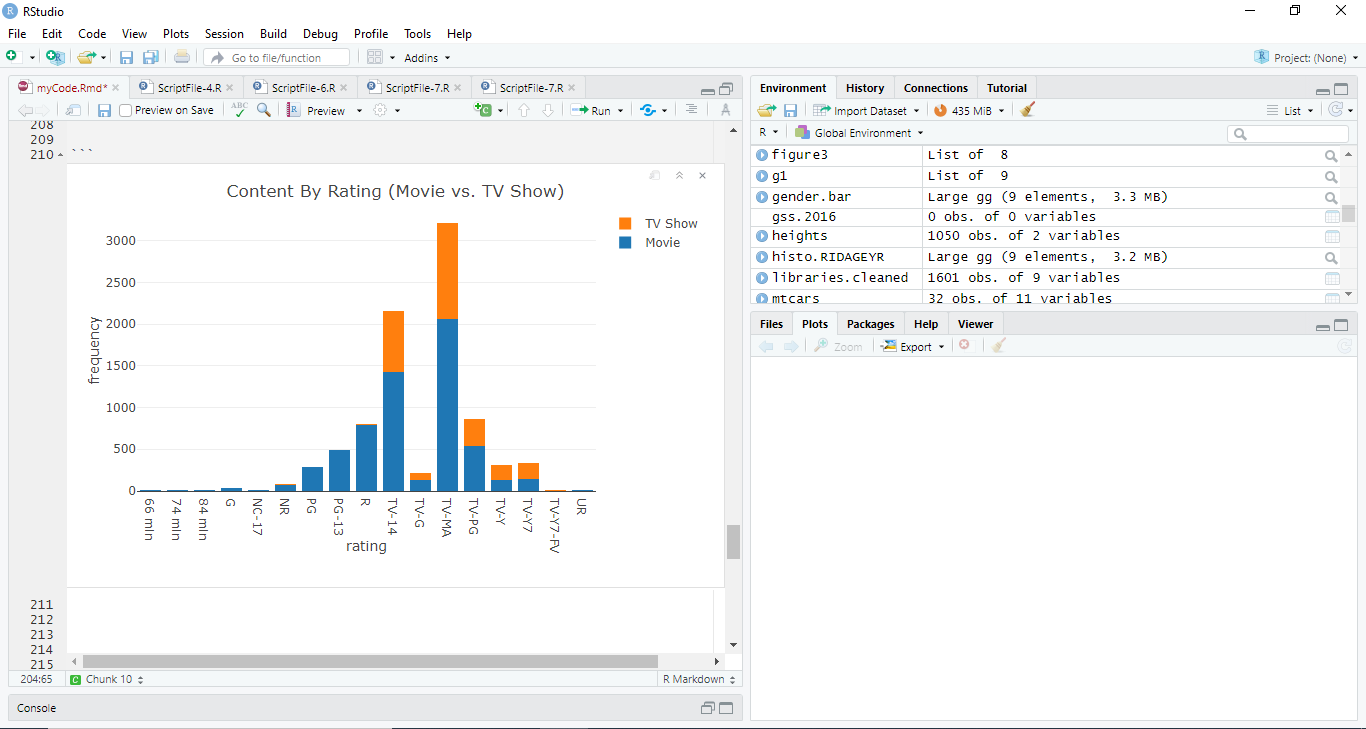
*Fig 2. Shows the features of the dataset and its description in Rstudio.*

## 5. Results

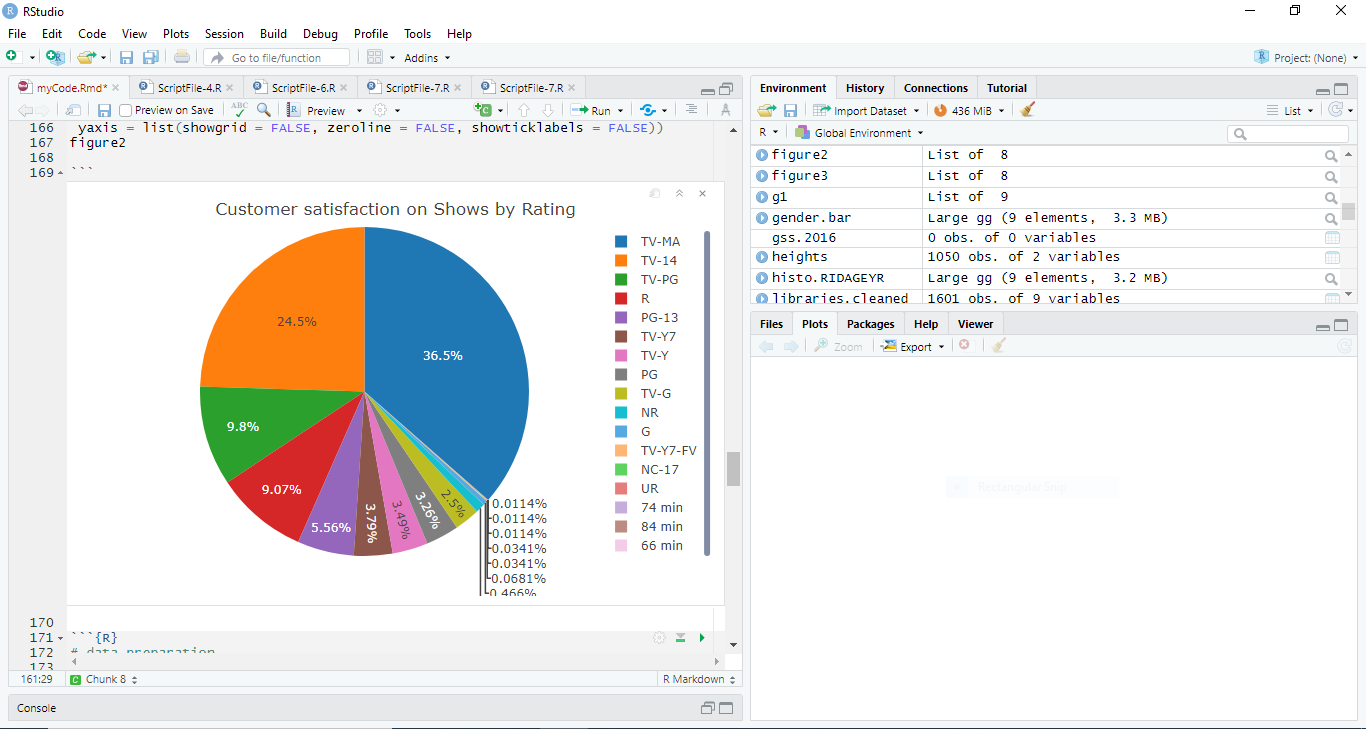
The following are the results generated after the data analysis by the help of artificial intelligence techniques has been used;

**

*Fig 2. Shows the most watched shows on Netflix in Rstudio.*

**

*Fig 2. Shows the customer rating on the shows on Netflix in Rstudio.*

**

*Fig 2. Shows the customer satisfaction rating on the shows on Netflix in Rstudio.*

# 5.1 Evaluation

## 5.2 Evaluation of the Results Found

In summary, my research study by the help of artificial intelligence techniques focuses on finding out some of the most watched shows on Netflix. As well the effect of rating on customer rating on the shows that they stream via Netflix. Also, the study is after finding out the impact of customer satisfaction on the financial growth of the company.

According to the results found in the results section above, Movies shows are the most watched shows on Netflix with more than twice times that of Tv shows. This tells the Netflix management that most customers like movies than the Tv shows presented via Netflix. In this case, Netflix top management have the role to find out what they have done to the movies show that has not been implemented in the Tv show to improve the number of customers on streaming Tv shows.

In addition, the most rated show is the most watched show. Therefore, there is high turnover of customer to those shows they have highly rated. In my study, Movie shows have been found to be the most rated shows, while at that point, I the results also confirm that, that’s the show with high turnover of customers. Netflix top management should implement the necessary attractive thing to the Tv shows that they have done to the Movie show to increase rating and customer turnover on Tv shows as well.

Lastly, its obvious that customer satisfaction will increase the financial growth of any company. Just from our results, customers display clearly how they are satisfied with Movie shows. Netflix makes good amount from movie show compared to Tv shows. This is because once customers are satisfied they maintain the flow of streaming just as indicated by results.

## 4.2 Limitations

Despite performing a successful analysis on my dataset for Netflix company, I was limited in a way. The study only allowed me to used historical dataset for Netflix company. This means whatever the analysis am performing was based on the historical dataset of Netflix company. This is not bad in any way; however, It would have been better if I use real time data. This is where, I perform analysis on the already working, present data that Netflix company keeps generating every time (Amatriain, 2013).

## 4.2 Potential direction

If I would have been given more time, I would connect my PostgreSQL database with real time data. With real time data, the analysis performed provides predictions of the future based on the current status not the historical one. Therefore, as time changes , data changes as well thus, the more accurate the predictions done becomes.

# 5. Conclusion And Future Work

In conclusion, with the help of artificial intelligence techniques (the use of decision trees in performing the classifications) using the Netflix dataset, Netflix as a company is aware of where to adjust and why to adjust by using the results provided in our analysis done. The top management will not face difficult in finding out which show is not performing well and also which show majority of customers preferer in their streaming platform.

In simple terms, Netflix top management can deduct and make right decisions based on the analysis as a way of improving the company financial growth. Thus, all companies especially those that derives large datasets every day should employ artificial intelligence techniques to be able to track their company progress, something which they cannot make it on their own.

# Bibliography

Amatriain, X., 2013. *Big & personal: data and models behind netflix recommendations..* s.l.:s.n.

Dasu, T. & Johnson, T., 2017. *Exploratory data mining and data cleaning..* s.l.:John Wiley & Sons..

Narayanan, A. & Shmatikov, V., 2016. *How to break anonymity of the netflix prize dataset..* s.l.: arXiv preprint cs/0610105..

Taylor, J. & Cihon, C., 2018. *Statistical techniques for data analysis..* s.l.:Chapman and Hall/CRC..