## 

# 

**TDS 2101**

**Introduction to Data Science Assignment**

**REPORT**

# 

**Industry : Entertainment**

**Tutorial Group : TT01**

**Group Members :**

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| Samuel Wong Ing Shiing (PL) | 1151103687 |
| Kavisha Kamini AP Ramah Moorthy | 1151305146 |
| Muhamad Aliff Mustaqim Bin Abd Rahman | 1151104153 |
| Wayile Jialade | 1151102347 |

# **Table of Content**

**Netflix** [**Part A**](#_j3vixotr5tn) **3**

1.[The Company and The Focus](#_9120kcnsnepa) 3

[2. Netflix’s Leverage on Data Analytics](#_gtdomn8rbww1) 3

[3. Benefits and Impacts of Data Analytics](#_co4c89na8wk3) 12

[4. Strengths and Weakness of Analytics Implementation](#_fpk303q56y68) 13

[5. Potential Future Business Cases or Questions](#_yjx7kz6zcl3) 14

**Netflix** [**Part B**](#_gtd589tglniy) **16**

1[.](#_gtdomn8rbww1) Context and Research Questions 16

2[.](#_co4c89na8wk3) Related Datasets and Its Content [1](#_co4c89na8wk3)6

3[.](#_fpk303q56y68) Datasets Examination and Data Preprocessing 17

4. Exploration of Descriptive Statistics of the Data 17

5. Statistics Modelling Technique for Prediction 23

6. Summary

**Miniclip** [**Part A**](#_j3vixotr5tn) **28**

1.[The Company and The Focus](#_9120kcnsnepa) 28

[2. Miniclip’s Leverage on Data Analytics](#_gtdomn8rbww1) 28

[3. Benefits and Impacts of Data Analytics](#_co4c89na8wk3) 29

[4. Strengths and Weakness of Analytics Implementation](#_fpk303q56y68) 30

5. [Potential Future Business Cases or Questions](#_yjx7kz6zcl3) 31

**Miniclip** [**Part B**](#_xfg3de1gjvcz) **33**

1[.](#_gtdomn8rbww1) Context and Research Questions 33

2[.](#_co4c89na8wk3) Related Datasets and Its Content 33

3[.](#_fpk303q56y68) Datasets Examination and Data Preprocessing 33

4. Exploration of Descriptive Statistics of the Data 34

5. Statistics Modelling Technique for Prediction 39

6. Summary 41

**Spotify** [**Part A**](#_j3vixotr5tn) **43**

1. [The Company and The Focus](#_9120kcnsnepa) 43

[2. Spotify’s Leverage on Data Analytics](#_gtdomn8rbww1) 44

[3. Benefits and Impacts of Data Analytics](#_co4c89na8wk3) 47

[4. Strengths and Weakness of Analytics Implementation](#_fpk303q56y68) 49

5. [Potential Future Business Cases or Questions](#_yjx7kz6zcl3) 50

# **Part A**

### **The Company and The Focus**

Company : Netflix

Focus : Becoming the best global entertainment distribution service

### **2. Netflix’s Leverage on Data Analytics**

* Data Collection :

Netflix, a company that started as a DVD rental company has grown into a multi-billion entertainment company that provides services such as media streaming and video-on demand. Netflix being a data-driven company has been using big data analytics to aid their important decisions and this strategy has proved itself effective and useful. As the first stage within the data science pipeline, data collection is essentially the prerequisite for any other data science process to work. For Netflix, the company has certainly no issue with data availability when they have over 100 million active streaming subscribers worldwide as of Q3 of 2017.

Netflix tracks a number of “events” from the interaction of their subscribers including the time when users pause, rewind or fast forward a content, what day, time and date users watch contents on Netflix’s platforms, where do users watch contents, the device used to watch contents, when users pause and leave contents, ratings users give, searches done by users, browsing and scrolling behaviours of users, data within the contents and so on. These “events” form an enormous datasets for Netflix and they were consisted by about 500 billion individual events or about 1.3 Petabytes of data per day, or in other words, 8 million individual events or about 24 Gigabytes of data every second. All of these data are collected from all of Netflix’s platforms such as the Netflix app on iOS devices, Android devices and the Netflix website.

* Data Preprocessing :

1. Data Cleaning

At Netflix, tonnes and tonnes of data are piling up every second at an astonishing rate. To keep the cost regulated and for the convenience of data analysis and data mining, Netflix implemented an automated process of data cleaning using a service under Amazon Web Service called Janitor Monkey.

Janitor Monkey is a service that runs in Amazon Web Service cloud servers that locates unwanted resources in datasets before cleaning them up. The way Janitor Monkey is used in Netflix is that Netflix has given it a set of rules as a guideline to determine whether a resource is redundant or unwanted and provide it with a schedule to run automatically in regard of Netflix’s enormous amount of incoming data everyday. After a resource is detected to be unwanted, Janitor Monkey will flag it and push a notification to the manager days before cleanup time.

At Netflix, four types of resources are managed by Janitor Monkey, they are EBS Volumes, EBS Volume Snapshots, Auto Scaling Groups and Instances. Each of them has their own rules to determine whether a resource is unwanted.

Janitor Monkey executes data cleaning at Netflix in 3 stages that are, “mark”, “notify” and “delete”. Janitor Monkey crawls datasets to find cleanup candidates and flag them based on the collection of rules provided to it. After that, it schedules a time to remove the resources flagged, the time duration can be configured within the rules. The “manager” of the flagged resource will too be notified as a layer of verification before the actual removal. The manger is also able to flag a resource to be excluded by Janitor Monkey using the REST interface.

Janitor Monkey events are also logged into Amazon SimpleDB in Netflix so that the records are always available to be checked on for activities done by Janitor Monkey.

1. Data Integration

Before 2016, Netflix has its billing infrastructure residing on Netflix Data Centre and Netflix since then has decided to migrate its billing data of different forms in its data centre (DC) to Amazon Web Service (AWS) cloud. This requires Netflix to integrate all kinds of data and locate them in another location, in this case on AWS cloud.

Netflix’s billing infrastructure includes open/paid billing periods, amount of credit on member’s account, management of payment status of members, charge request, payment date and so on. Other than that, the infrastructure also feeds into Netflix’s financial system on revenue and tax reporting. This requires Netflix to integrate different services such as member account service, payment processing, customer service, customer messaging, DVD website and shipping.

Considering most of these services’ data are interacting with Oracle, Netflix first operation is to disintegrate their giant Oracle based solution into a services based architecture before splitting them again into multiple data stores. Subscriber data was moved to Cassandra data store on AWS, whereas for payment processing integration which requires ACID integration, all relevant data was moved to MYSQL on AWS.

Netflix’s approach towards this operation mentioned above is to first ensure that only essential data is integrated into the cloud and purge all data that are useless to any function.

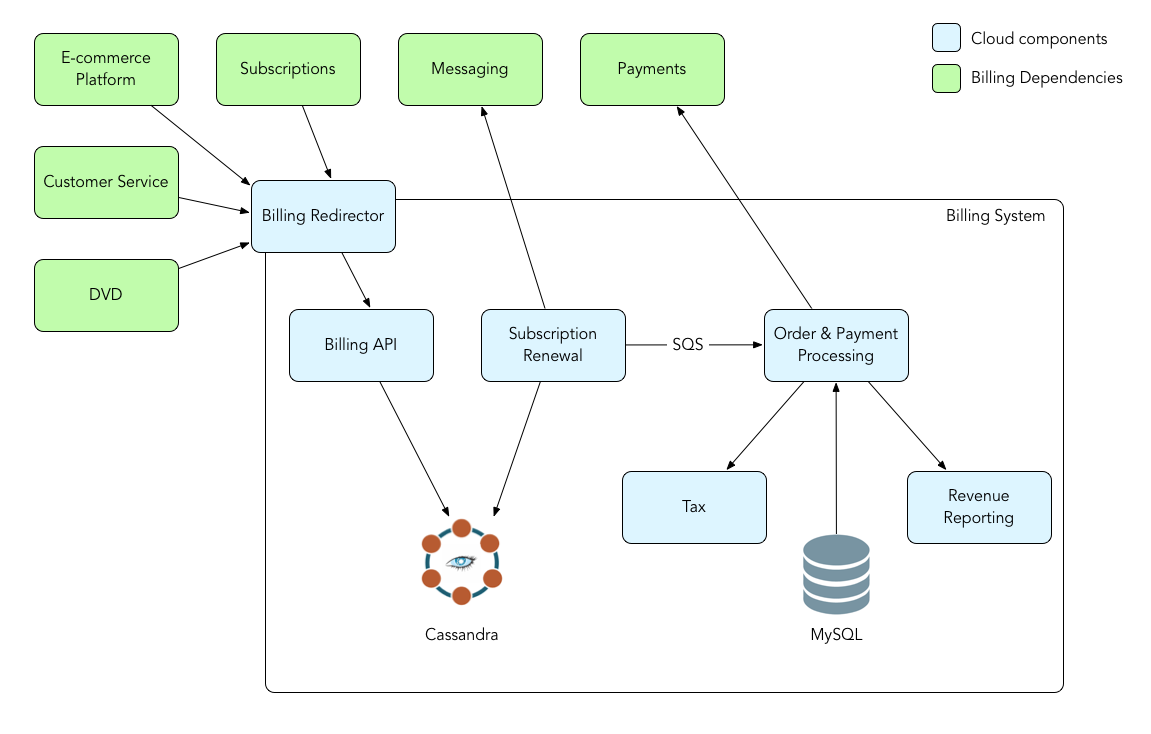


Diagram of Netflix Billing Infrastructure

1. Data Reduction

Netflix has designed a tool called “Aegisthus” to handle its map/reduce operation for their data. The program reads Cassandra SSTables and executes data conversion for the dimensional data they hold in Cassandra into data that Netflix would consume in their big data environment.

As Netflix has a very substantial utilization of Cassandra database on its AWS cloud, the flow of Cassandra data moving into Netflix big data environment includes reduction of Cassandra SSTables to json formatted data in S3 for the convenience of downstream batch data analysis as illustrated below.

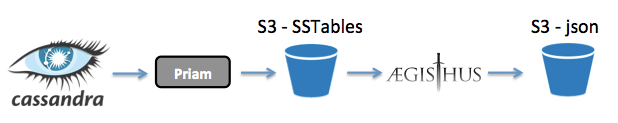


Diagram of Data Reduction Process from Cassandra to JSON

At the very first stage of the flow, “Priam” is used as a tool to backup tables from Cassandra in the form of SSTables in an incremental fashion onto physical disks. After that, Aegisthus will read relevant SSTables to create the wanted dataset, by removing duplicate records, it compacts the records into a single view of data. Finally, it serializes the data into a format that can be processed by their batch processing system. The format that works for Netflix is a line of json per record. The ultimate goal of this data reduction is to split up large SSTables which in turn will ramp up processing rate of Netflix’s data analysis.

1. Data Transformation



Netflix in the aim of solving process flow orchestration has built and use a “orchestration engine” called “Conductor” to fulfill the requirement of tracking and management of workflow, to visualize the workflow, to synchronously process all task when needed, to scale millions of concurrently running process flows, just to name a few. Within the operation of Conductor, data transformation is carried out in a regular basis based on JSONPath.

With JSONPath, data extraction from JSON data structure is possible with no extra scripting before undergoing data transformation process. JSONPath is implemented with templates such that a new object can be created based on the template given. A template property can extract a single property of interest from the source data before transforming them into newly created objects.

* Data Analysis :

With over 100 million streaming subscribers worldwide, data analysis is an important if not the most important approach for Netflix to understand its user base and to gain insights on what has happened to them as well as what is happening in real time. With results from data analysis, Netflix is able to make better decision or even adjustment of strategy that will benefit the business as well as its users.

The goal of data analysis is to understand characteristics of data and Netflix is achieving that by a number of methods. For example, to understand the characteristics of data relating to user engagement on Netflix, Netflix looks towards “Completion Rate”, which is the rate of user watching a content finishing it to the end, of a certain series. With the completion rate available, Netflix would then look into the cutoff point where a percentage users stop watching a series as well as the time gap of the cutoff period. This type of data analysis is able to help Netflix to understand what its users like and dislike, and from there, Netflix will be able to make more reasonable decisions in the future based on the data analysis such as producing series that includes more elements that are previously known to increase user engagement.

Netflix’s has shown huge success in their “recommendation algorithm” to suggest new movies that a user might be interested to watch, but before the algorithm can work, Netflix must first what each user wants and things or genres that they might be interested in. This is where data analysis is crucial to Netflix. In order for Netflix to understand each of its multimillion-size users, Netflix collects data on what a user normally or frequently watch on netflix based on the genres, producers, actors staring and other related factors, with this, Netflix is able to group or tag its users into different category for future application on prediction algorithm.

* Data Mining :

With the data mining goal of examining large databases to discover patterns or new information, Netflix has been very successful with its movie recommendation algorithm called “Cinematch”. This is an algorithm built by Netflix to suggest or recommend new movies or series to Netflix users based on data analytics done on each users. Movie recommendation is extremely crucial to Netflix as 75% of movies or series watched by users on Netflix are recommended by some sort of recommendation system on Netflix. For Netflix, user engagement is a key factor for its success as proved statistically that users tend to unsubscribe when they do not or failed find good movies or contents on Netflix.

Data Mining being the centre process of the algorithm, “Cinematch”, the algorithm finds patterns in terms of “matches” between contents instead of matching users to movies. There are a number of steps on how the algorithm matched contents, simplified as follows:

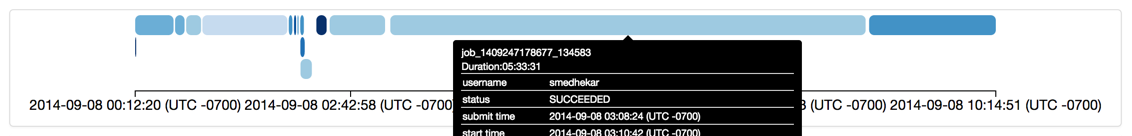
1. To search a match for a movie, the algorithm searches the Cinematch database for the people who have rated the exact movie.
2. Extracts users from the group that have also rated another movie.
3. Computes the statistical probability of the people who liked the first movie will also like the second one.
4. Iterate the process to generate a pattern of correlations between subscribers’ ratings of different contents.

* Data Visualization :

After having a peek on the scale of data analytics done by Netflix on a regular basis, it is obvious that Netflix has a huge number of data processing platforms used internally with different purposes. In fact, to acquire information on a specific run, piles and piles of logs, records, tools and data need to be extracted and this process is normally a nerve-wrecking and time consuming one.

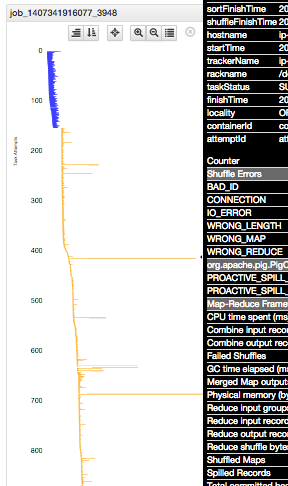
In the name of easing the process of different internal platform users to explore and have a better insight of their job performance, Netflix has built a job search and visualization tool named “Inviso”. This tool has only one goal that is to help big data users to better understand their job performance. The way Inviso achieve effective job search is by indexing all jobs across all clusters into ElasticSearch and at the same time provides an user interface to do basic querying. The search result is displayed in a concise table containing the job, its ID, running duration, completion status, clustered runned on, and other historical data archived.

To further visualize job performance, Inviso is able to generate a swimlane diagram as shown in the diagram below to visualize different stages of a job executed in a parallel fashion. This approach provides big data users to have a better view on the whole process of a job and to deduce a focus within the process to improve on.



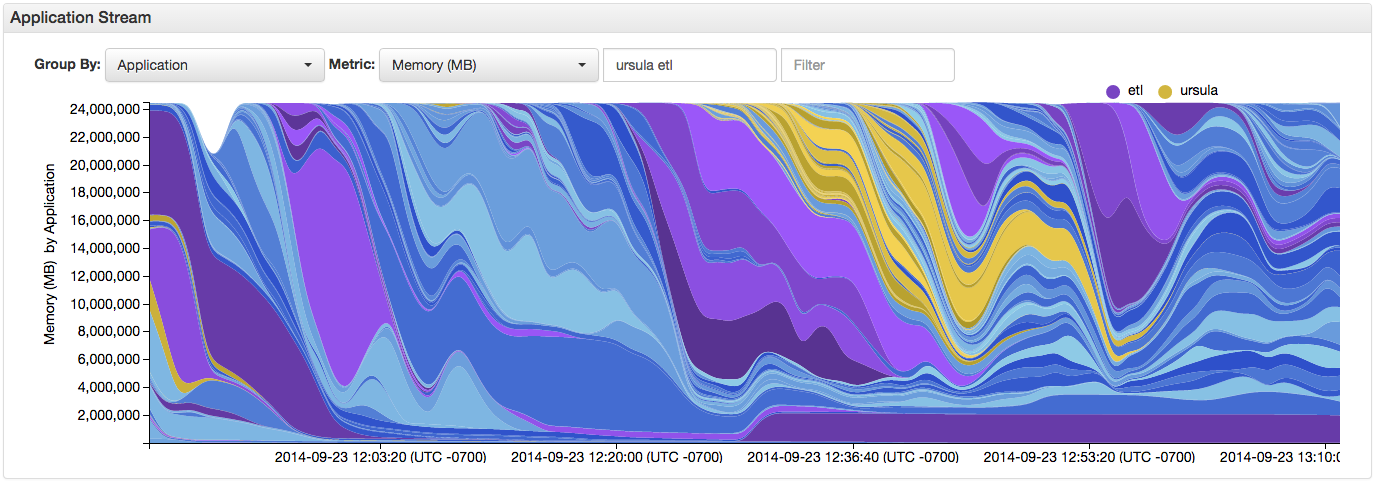
Swimlane Diagram of a Job

With Inviso, a workflow diagram as shown in the diagram below, can also be produced for a better view on the individual execution of every task attempt of every job. As the these task are listed in a time order, it is easier to spot potential issues within individual task of a job such as bad resource allocation, slow attempts, data skews, just to name a few.



Workflow Diagram on individual task of a job

The performance of a job is entirely dependant on the resource allocated to it and Netflix has thousands and thousands of jobs running on a regular basis such as production jobs, ad-hoc queries, reporting, benchmarking and so on, taking on relatively limited resources. Hence, for Netflix to better monitor their resource allocation for these everyday jobs, Inviso is designed to index all jobs and their current metrics on the clusters that they are on, into ElasticSearch. With this, reconstruction of the state of any cluster on any time span can be done to evaluate the effects of a specific job execution on a cluster’s performance. In the diagram below is the visualization done by Inviso on the capacity and backlog of a cluster based on predefined metrics.



Performance Visualisation of a Cluster

### **3. Benefits and Impacts of Data Analytics**

In 2011, Netflix has made a big decision to outbid big television channels in the US such as HBO and AMC to acquire the rights for a US version of the series, “House of Cards”. This costs Netflix a whopping 100 million USD for a series of 2 seasons, each with 13 episodes and with the price tag of each episode of 4 - 6 million USD. This is one of the biggest investments that Netflix has made at the time and the decision was made based on data analytics, more precisely “prediction algorithm” designed by Netflix.

Before Netflix has set their eyes on the series, “House of Cards”, through past data analytics, Netflix found out that a lot of their subscribers watched the movie, “The Social Network” directed by David Fincher with high completion rate. Other than that, the british version of “House of Cards” has achieved high acceptance rate. Netflix also realized that subscribers that watched the british version of “House of Cards” also watched other Kevin Spacey films as well as films directed by David Fincher. All of the three facts mentioned above has taken up a good portion of Netflix’s entire subscriber base. In fact, in an interview with Steve Swasey, VP of Corporate Communication at Netflix, he publicly stated that they have a very high degree of confidence in “House of Cards” based on director, producer and the stars. He said that they think they do not need to spend millions to get their subscribers to tune into this as Netflix knows that there is a big portion of their subscribers that will be interested to watch it based on their data analytics. Swasey adds that they are able to look into consumer data to look for the appeal for the director, stars and producer.

After successfully acquiring the rights for streaming US version of “House of Cards”, Netflix was not done applying their knowledge learnt from data analytics. Through categorizing their subscribers based on their historical interest on genres of content, Netflix actually made a number of different versions of trailers for the series, “House of Cards” with the goal of a better catch of interests of different types of subscribers. Netflix plays different versions of the trailer to its subscribers based on their historical interest with the aim of a better gear towards different audiences.

Years after the first stream of the US version of the series, “House of Cards”, metrics are available to inspect the trajectory of one of the biggest investment done by Netflix. According to Netflix, “House of Cards” in the US has attracted 2 million new US subscribers in the first quarter of the 2013, a 7% increase over the previous quarter and that does not include the 1 million new subscribers from countries outside of the US. In accordance to The Atlantic Wire, these 3 million new subscribers alone has paid the entire cost of the series.

### **4. Strengths and Weakness of Analytics Implementation**

One of the main big data analysis tools used by Netflix at a regular basis is the “Netflix Recommendation Algorithm” as introduced before in this report. Most of the recommendation algorithm works by capturing patterns from the past and using them to predict something. There is a weakness with this type of algorithm that is algorithms of this sort might have a hard time keeping up the the scale of factoring attributes of a product that it is trying to recommend or predict. In other words, it is very unlikely for a algorithm to consider each and every attributes in a field or product that might have effect on future trends of that particular field or product as there might be an infinite amount of them considering these factors are different for every individual. The same logic can be applied to Netflix’s recommendation algorithm too.

Other than that, there is also a problem of “unpredictable attributes”. For Netflix, there might be an issue with eccentric movies, the type of movies that often generates very two-sided opinions among its audiences. These type of movies are harder to recommend on as a certain group or type of audience might have very high similarity on types of movies favoured but might both hate or love these eccentric movies at the same time.

Despite all the weaknesses, the recommendation algorithm implemented by Netflix has a huge advantage for Netflix itself and it is the algorithm works great when the database that it based its prediction on is huge. This types of algorithm tend to be able to capture more underlying trends or attributes when it feeds on a variety of huge related database. In other words, the algorithm can have a more comprehensive consideration of factoring attributes when dealing with databases that Netflix can provide in term of size. This algorithm will not work nearly as well if the datasets that it feeds on is relatively small.

The way that Netflix’s recommendation algorithm is implemented, it takes both content filtering and collaborative filtering into account. For instance, Person A and B are categorized into a subgroup based on types of audiences, if only collaborative filtering is implemented, for a movie to be recommended to Person A or B, some other member of that subgroup would have to be watching that movie and gave a good rating before, but if only content filtering is implemented, only one person from the subgroup can be recommended to by the algorithm at a time. Hence, by combining both implementations into the algorithm, efficiency, accuracy and effectiveness can be improved by a huge margin.

### **5. Potential Future Business Cases or Questions**

With Netflix extraordinary growth of number of subscribers in the recent years, around half of the subscriber base of Netflix is now international subscribers outside of the United States. One of Netflix’s approach to attract more subscribers is by licensing contents to stream and producing original series. But majority of these content are tilted towards the liking or tastes of US subscribers. With international subscribers taking half of Netflix’s subscriber base for the first time in Netflix’s history, Netflix should think about creating more contents that would appeal to the unique interests of viewers in countries with diverse taste. This might involve analysing interest or taste of subscribers from different countries on video contents. This can potentially be an effective approach towards attracting more international subscribers as well as reinforcing Netflix’s influence in the entertainment industry around the globe instead of just the industry within US.

**Reference :**

[Philip Fisher-Ogden](http://twitter.com/philip_pfo), [Matt Zimmer](http://twitter.com/zimmermatt), [James Kojo](http://www.linkedin.com/in/jameskojo), and [Jinhua Li](http://www.linkedin.com/pub/jinhua-li/44/307/8). (2015). Netflix’s Viewing Data : How we know where you are in the House of Cards. Retrieved from : *https://medium.com/netflix-techblog/netflixs-viewing-data-how-we-know-where-you-are-in-house-of-cards-608dd61077da*

[Steven Wu](https://twitter.com/stevenzwu), [Allen Wang](https://twitter.com/allenxwang), [Monal Daxini](https://twitter.com/monaldax), [Manas Alekar](https://twitter.com/ManasAlekar), [Zhenzhong Xu](https://twitter.com/ZhenzhongXu), [Jigish Patel](https://twitter.com/jigishjpatel), [Nagarjun Guraja](https://twitter.com/gvnagarjun), [Jonathan Bond](https://twitter.com/ma4jpb), [Matt Zimmer](https://twitter.com/zimmermatt) and [Peter Bakas](https://twitter.com/peter_bakas). (2016). Evolution of the Netflix Data Pipeline. Retrieved from:

*https://medium.com/netflix-techblog/evolution-of-the-netflix-data-pipeline-da246ca36905*

Zach Bulygo. How Netflix Uses Analytics to select movies, create content and make multimillion decisions. Retrieved from :

*https://blog.kissmetrics.com/how-netflix-uses-analytics/*

Charles Smith and Jeff Magnusson. (2013). Aegisthus Is Now Part of NetflixOSS. Retrieved from :

*https://medium.com/netflix-techblog/aegisthus-is-now-part-of-netflixoss-a6364b254973*

[Stevan Vlaovic](https://www.linkedin.com/in/stevanvlaovic), [Rahul Pilani](https://www.linkedin.com/in/rahul-pilani-7910b82), [Subir Parulekar](https://www.linkedin.com/in/subir-parulekar-19ab403) and [Sangeeta Handa](https://www.linkedin.com/in/sangeetahanda). (2016). Netflix Billing Migration to AWS. Retrieved from :

*https://medium.com/netflix-techblog/netflix-billing-migration-to-aws-451fba085a4*

Michael Fu and Cory Bennett. (2013). Janitor Monkey - Keeping the Cloud Tidy and Clean. Retrieved from :

*https://medium.com/netflix-techblog/janitor-monkey-keeping-the-cloud-tidy-and-clean-d517ad74d648*

Viren Baraiya. (2017). Netflix Conductor - Inversion of Control for Workflow. Retrieved from : *https://medium.com/netflix-techblog/netflix-conductor-inversion-of-control-for-workflows-c495621bc1c5*

Xavier Amatriain. (2012). Netflix Recommendations: Beyond the Five Stars (Part 1). Retrieved from : *https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429*

Daniel C. Weeks. (2014). Inviso: Visualizing Hadoop Performance. Retrieved from : *https://medium.com/netflix-techblog/inviso-visualizing-hadoop-performance-f834175c6df8*

# **Part B**

1. **Context and Research Questions**

The research questions are :

1. Does higher historical rating of a particular anime content has effects on the ratings given by a particular individual?
2. Does a particular type of anime content has effects on the ratings given by a particular individual?
3. Does higher number of total episodes of a anime content has effects on the ratings given by a particular individual?
4. Based on the results of the analysis on the descriptive questions above, if anime contents with historical ratings of 7 and above are suggested by the system to the users, will the users be able to discover more anime contents that are within their liking?
5. **Related Datasets and its Contents**

A related dataset that has been found is the dataset of recommendation data from theanimelist.net. The dataset includes information on data of user preference from 73,516 users on 12,294 anime. Each of the users is able to add anime to their completed list and give it a rating and this data set is a compilation of those ratings.

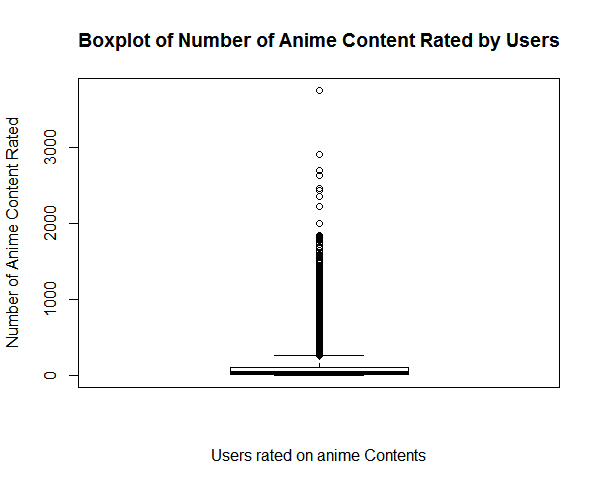
The dataset has been taken from the dataset website, Kaggle.com. The information that has been included within the dataset includes anime ID, name of anime, genre, type of content, number of episodes, average historical ratings given, members, user ID and rating of a specific user towards a specific anime content.

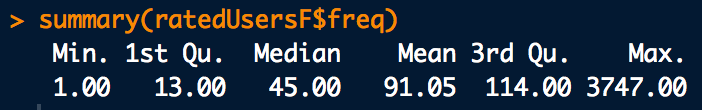
1. **Datasets Examination and Data Pre-Processing Activities**

A data pre-processing activity has been carried out on the dataset chosen that is to remove unwanted or unnecessary data from the dataset or also known as data cleaning. This is due to the fact that the dataset has include all historical behaviour of a specific user including the anime content that they watched but did not assign a rating to. Situations like this may happened for a huge variety of reasons other than disliking the content. Hence, for consistency reason, records of users watching a content but did not rate it after, have been removed from the dataset. The removal process is done through RStudio by importing the original dataset of the user ratings and extracting records with valid ratings only before saving it as a new updated dataset.

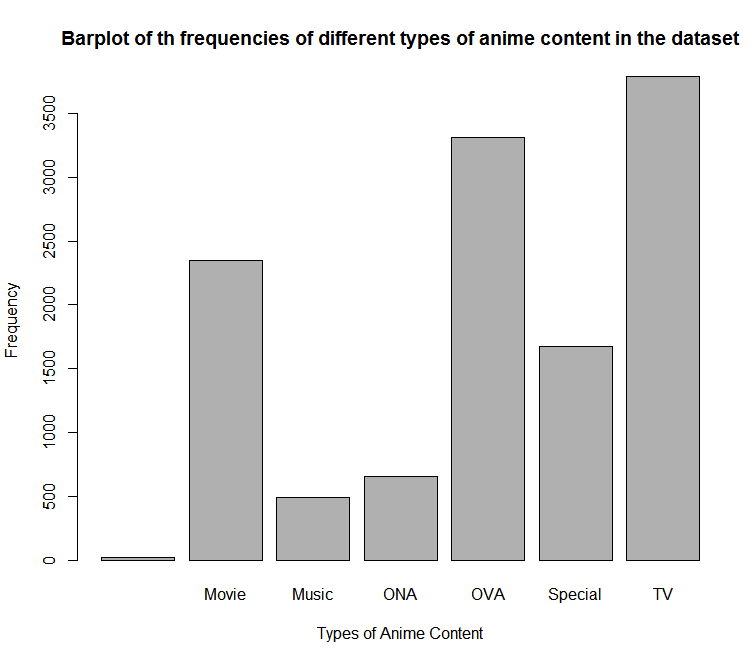
1. **Exploration of Descriptive Statistics of the Data**

After detailed analysis of the datasets, the team were able to come out with some descriptive statistics about the datasets. First and foremost, the team decided to first get a view on the numbers of anime content rated by the users in the dataset. A box plot is generated on the numbers of anime content rated by a single user, through the box plot, it is obvious that there is a lot of outliers beyond the 4th quartile and the mean of the number of anime content rated per person is quite low at 91.05 when the data has a maximum of 3747. In other words, the average number of anime content rated by a user in the dataset is 91.05 with a minimum number of 1 and maximum number of 3747. Conclusively, the variation of the number of anime content rated by a user in the dataset is huge.

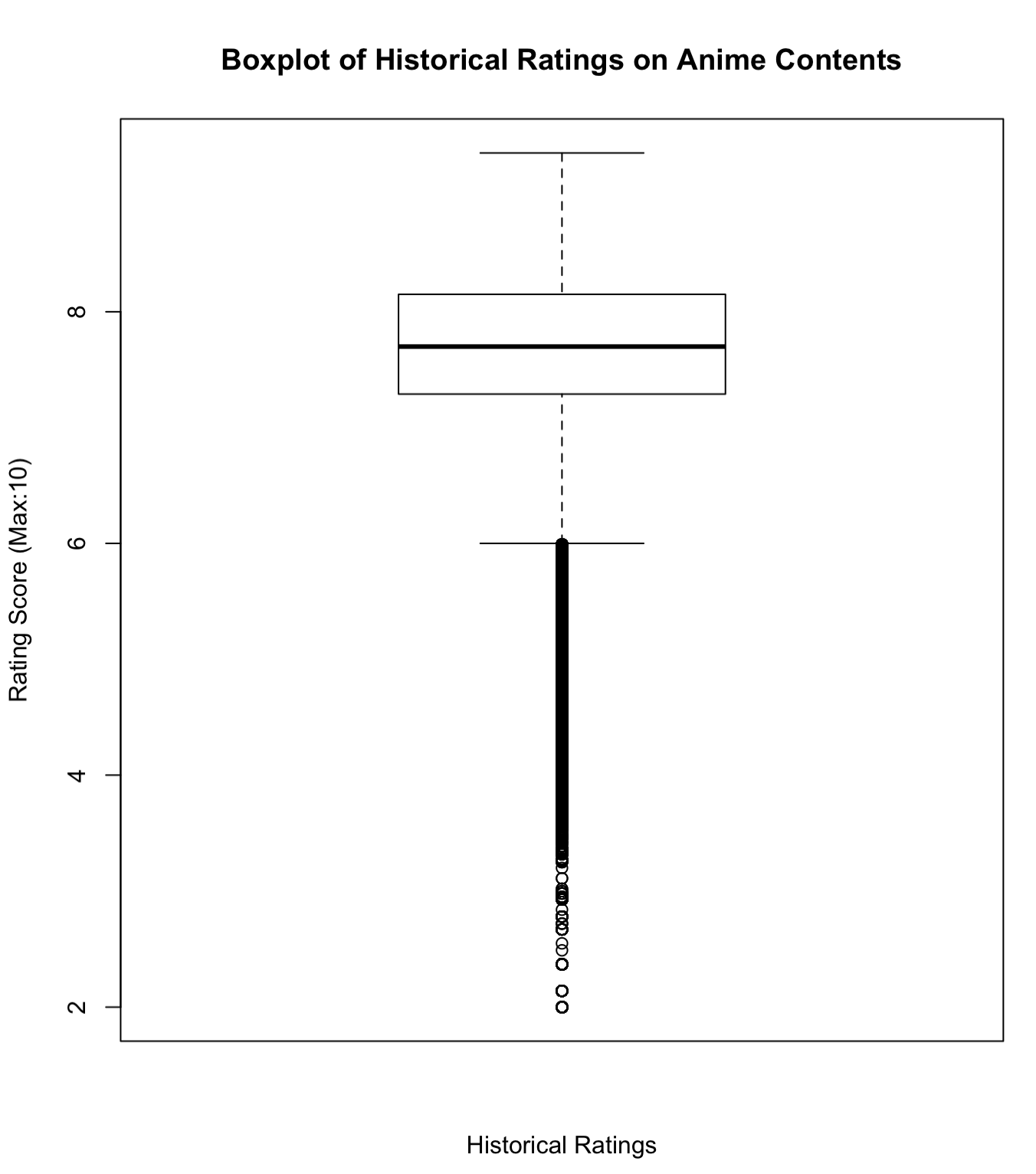


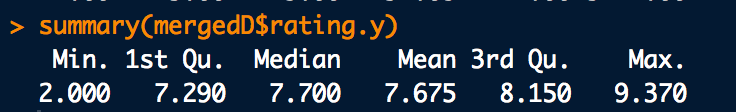


Other than that, the team decided to understand the proportion of each type of anime content in the dataset. A barplot is generated based on the number of anime content for each of the 6 types of anime content. It turns out that anime content of type TV has the highest number, followed by OVA, Movie, Special, ONA and Music.

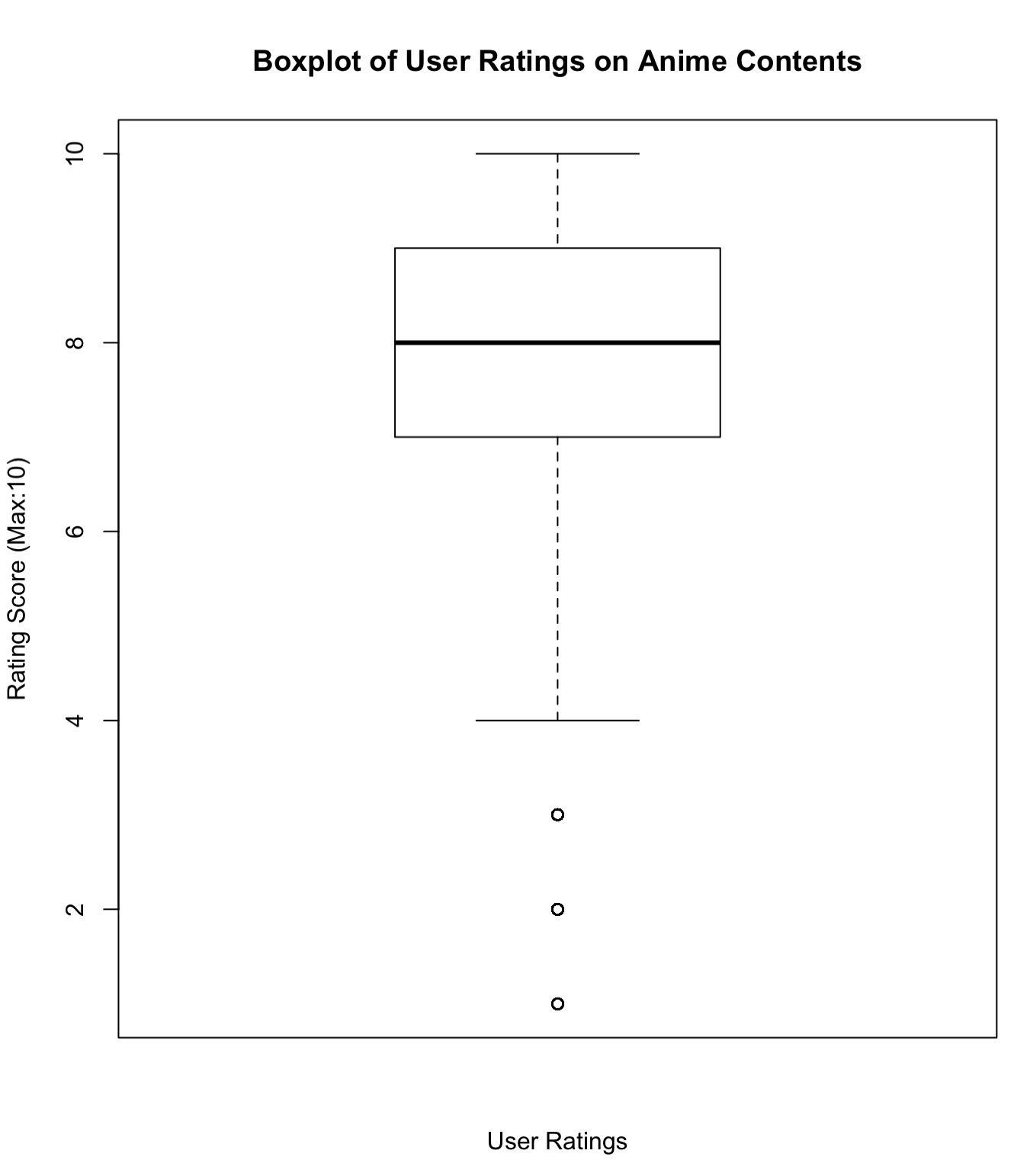


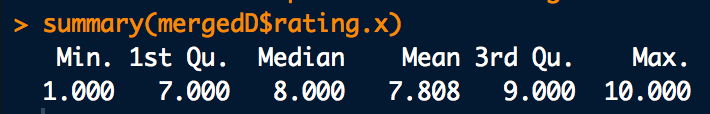
The team was also able to understand the statistics on the historical rating of the anime content in the dataset. A box plot is generated based on the historical rating of the anime content in the dataset and it turns out that the average rating is at around 7.675 out of 10 with a minimum of 2 and maximum of 9.37.





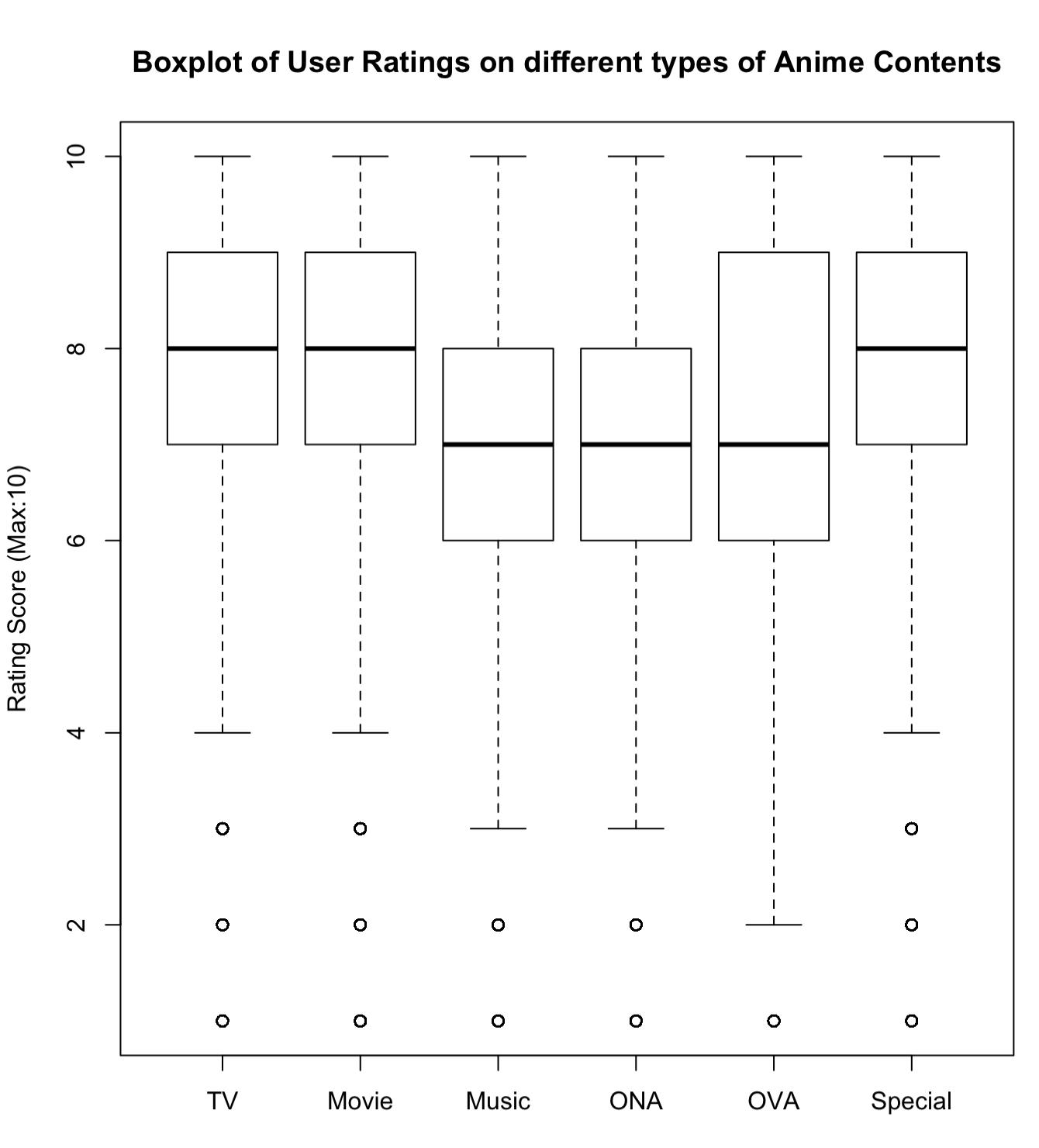
In contrast to the analysis above, the team was interested in the user rating of the anime content in the dataset. A similar box plot is generated based on the user rating of the anime content in the dataset and it turns out that the average user rating on the anime content in the dataset is 7.808 out of 10, quite similar to the average historical rating which is 7.675. The minimum is 1 and the maximum is 10.

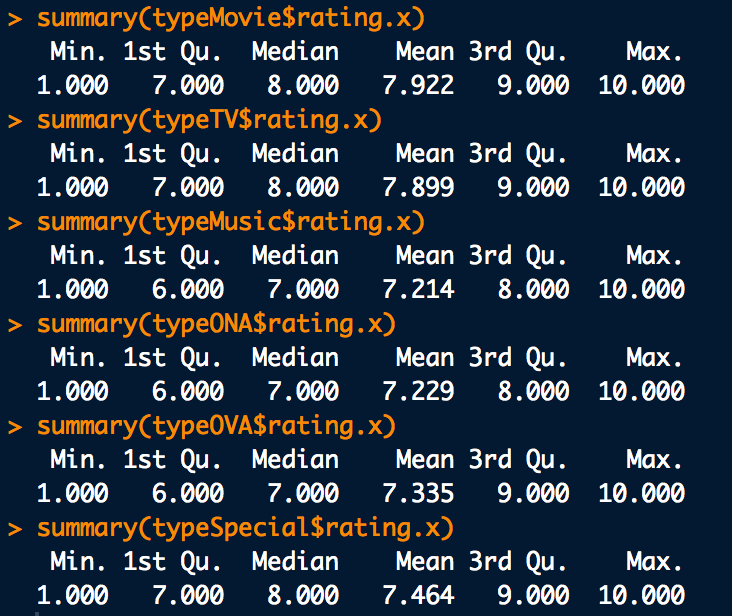


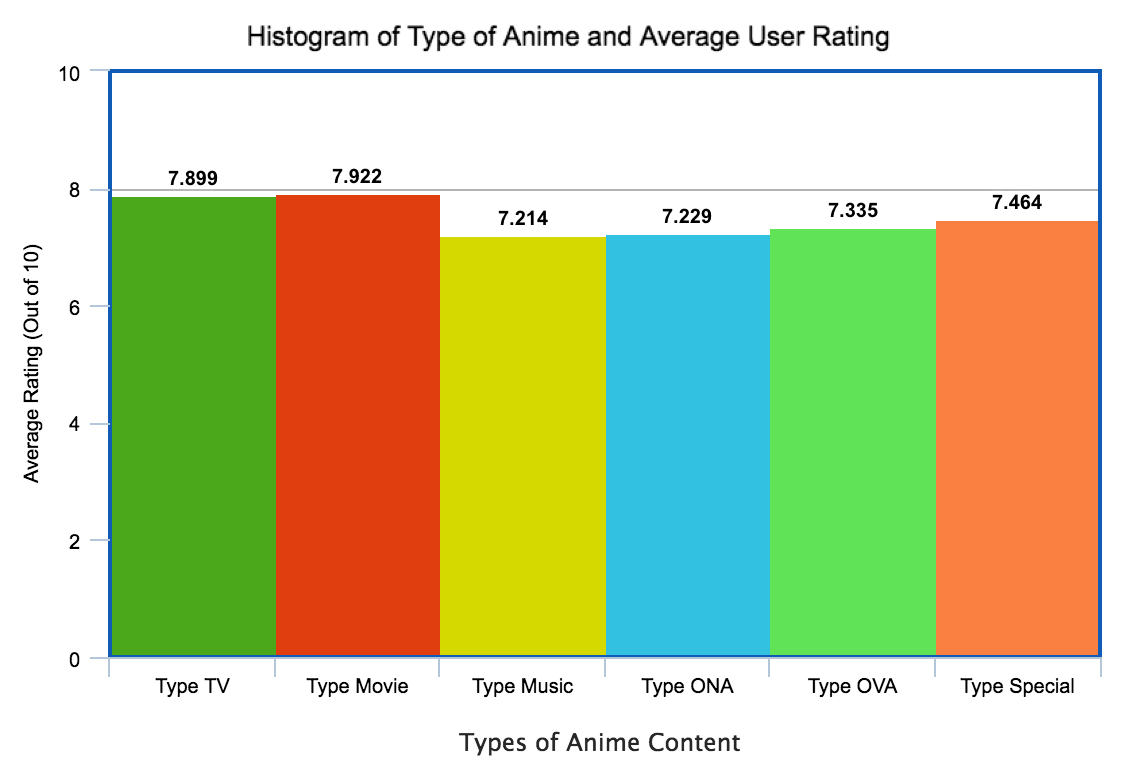


Based on both of the box plots on historical rating and user rating above, there seems to be a certain degree of correlation between the both as the average or mean scores of both ratings are interestingly close. A more detailed and in depth analysis on the correlation will be carried out and described later in the report.

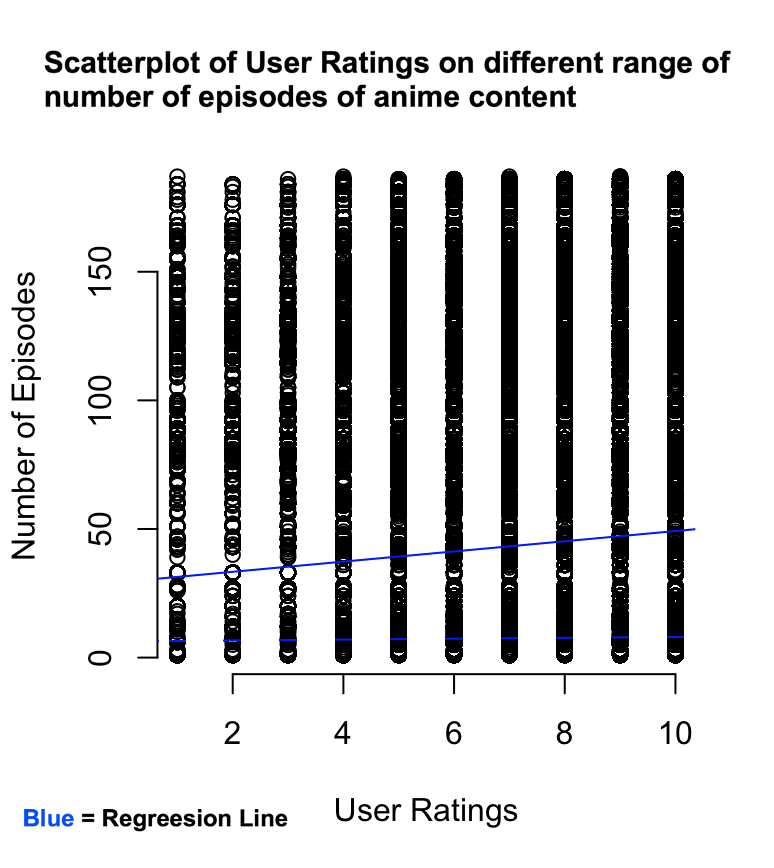
The team also looked into the user ratings given to each type of anime content. A box plot is generated based on the user rating on each type of anime content in the dataset. It turns out that the average user rating on anime contents of type TV and Movie are 7.899 and 7.922 respectively, which are slightly higher than that of type Music, Special, OVA and ONA at 7.214, 7.464, 7.335 and 7.229 respectively with type Music being the lowest.





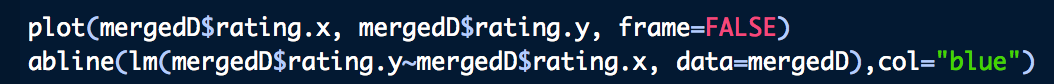


The team took the opportunity look into the user rating on different range of number of episodes of the anime content in the dataset. Due to the fact that the number of user rating contained in the dataset is too huge, the team decided to understand the correlation between number of episodes of an anime content and the user ratings given through linear regression analysis. It turns out that the regression line is implying a very slight correlation such that the ratings are the best when the number of episodes is closing to 50.

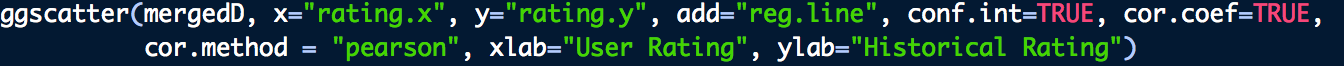


1. **Statistical Modelling Technique for Prediction**

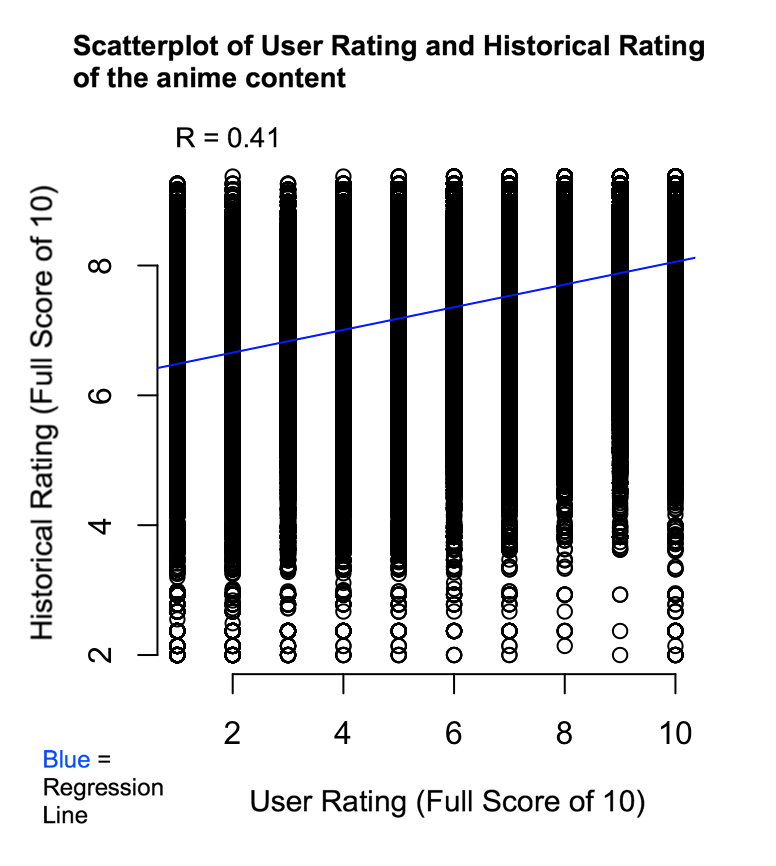
The linear regression model is fitted using the lm() function, in this case, to predict user rating based on historical rating of an anime content as shown in the diagram below.



A regression line and the correlation coefficient are generated in the model using the correlation method of “pearson” as shown in the diagram below.



A scatter plot is generated together with the regression line and correlation coefficient value calculated as shown in the diagram below.



In order to answer the predictive research question, assuming that user’s liking towards an anime content is directly proportional to the ratings they assign, it is logical to assume that when users are recommended with anime contents with historical ratings of 7 and above, the users have higher probability to give them higher ratings which implies that they have higher or greater likings towards the contents. Hence, if anime contents with historical ratings of 7 and above are suggested by the system to the users, statistically the users will be able to find more contents that are within their likings.

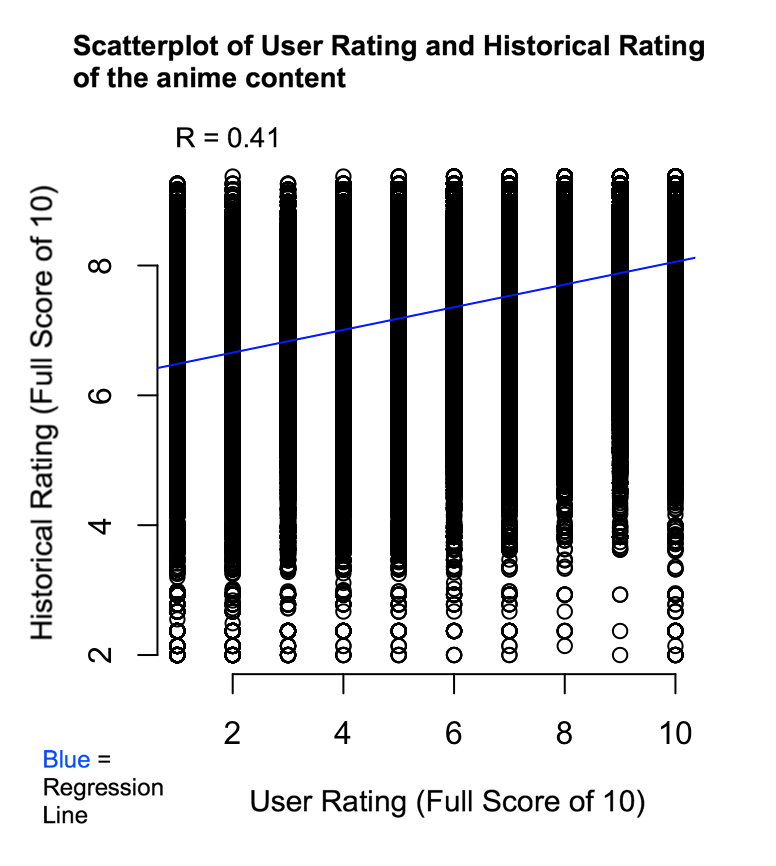
1. **Summary**

The data analysis has helped answer all the research questions as described below:

First research question :

* Does higher historical rating of a particular anime content has effects on the ratings given by a particular individual?

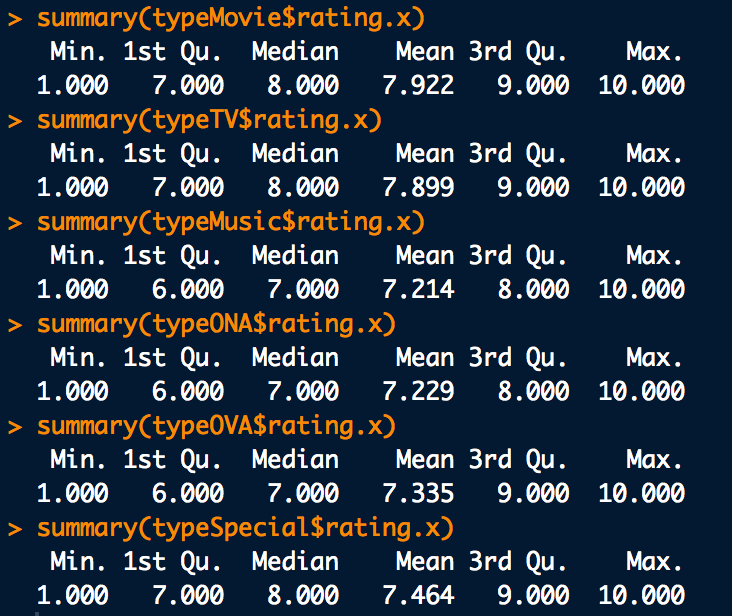
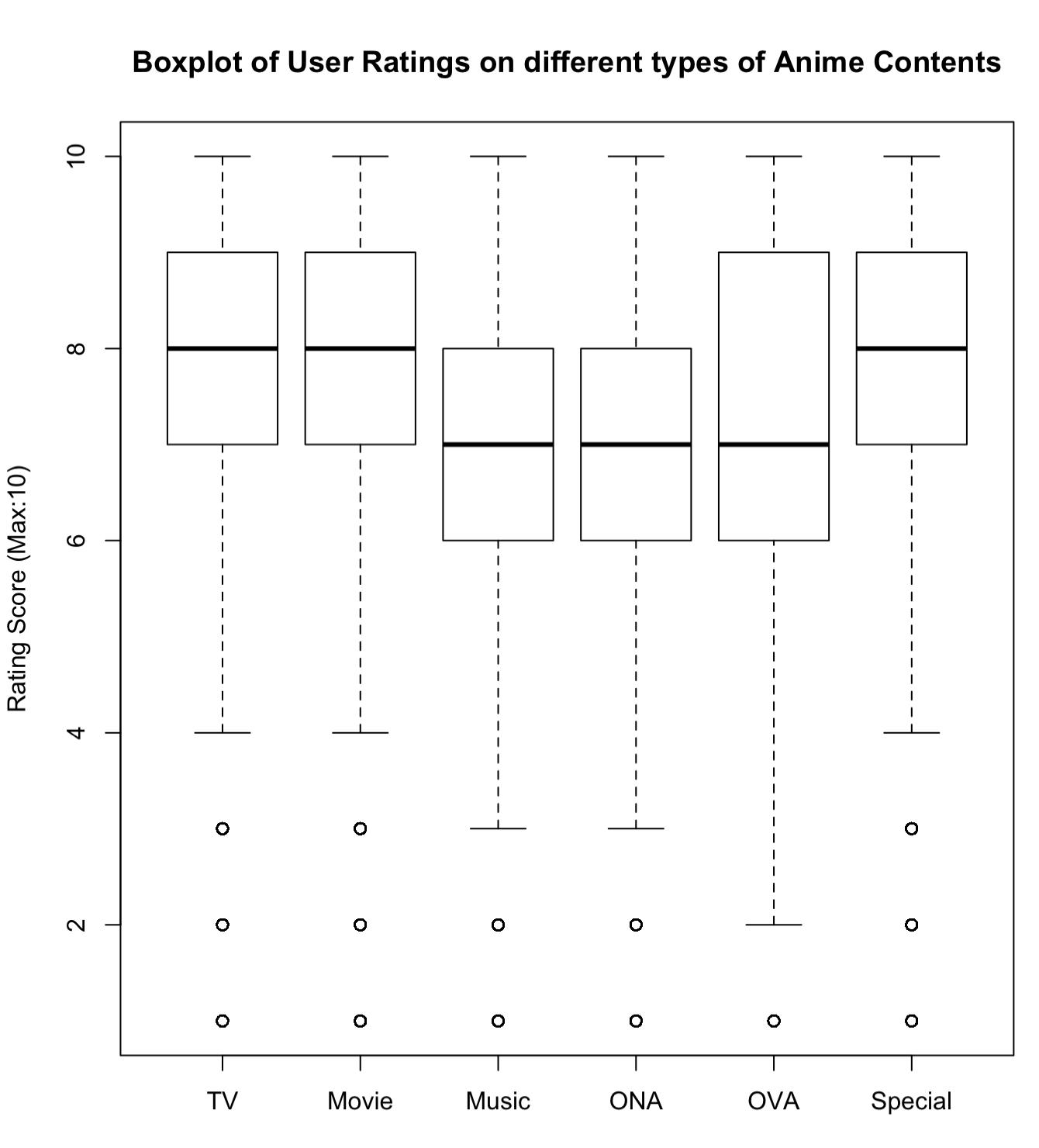
By implementing linear regression statistical modelling technique, a scatter plot with a regression line and the value of correlation coefficient has been generated as shown in the diagram below. Through the scatter plot, it is obvious that there is a correlation coefficient value of 0.41 which means user ratings do have relation with the historical ratings such that higher historical rating can lead to higher user rating.

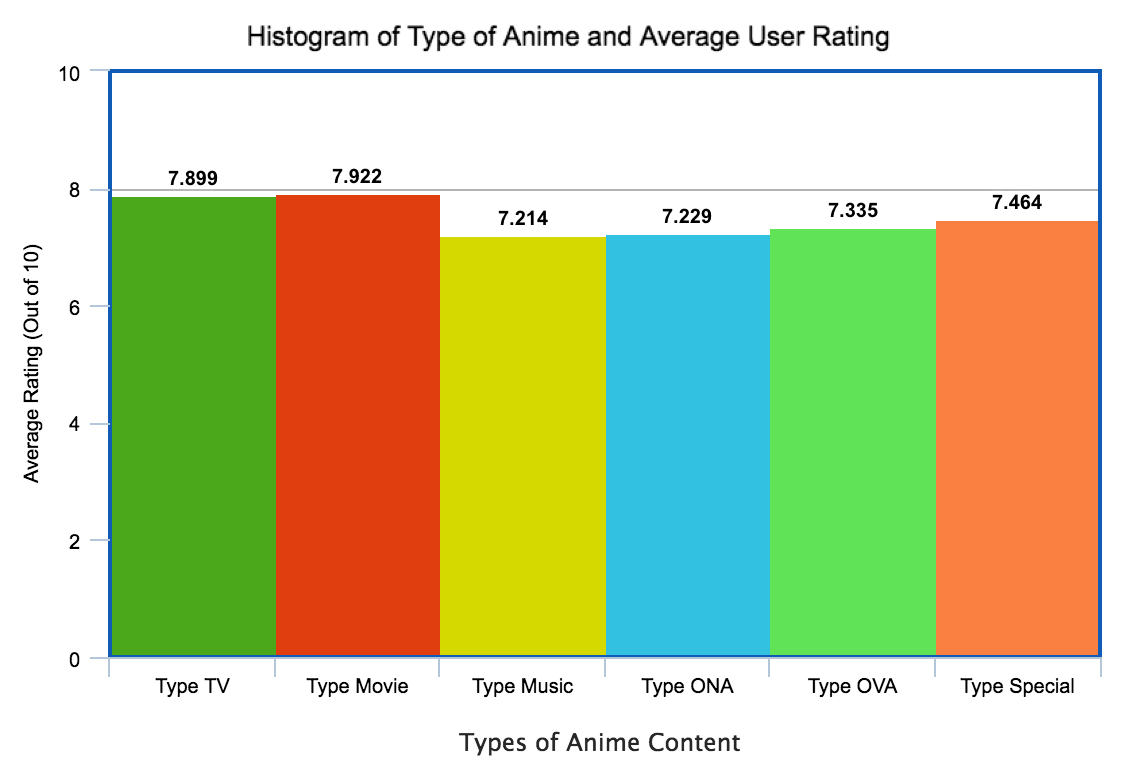


Second research question :

* Does a particular type of anime content has effects on the ratings given by a particular individual?

By finding out the average or mean rating of all users in the dataset on each type of anime content in the dataset as shown in the diagram below, anime content of types TV and Movies have slightly higher average ratings compared to other types of anime content. Hence, the conclusion is that the type of anime content does have a very slight influence or correlation on the ratings given by the users.

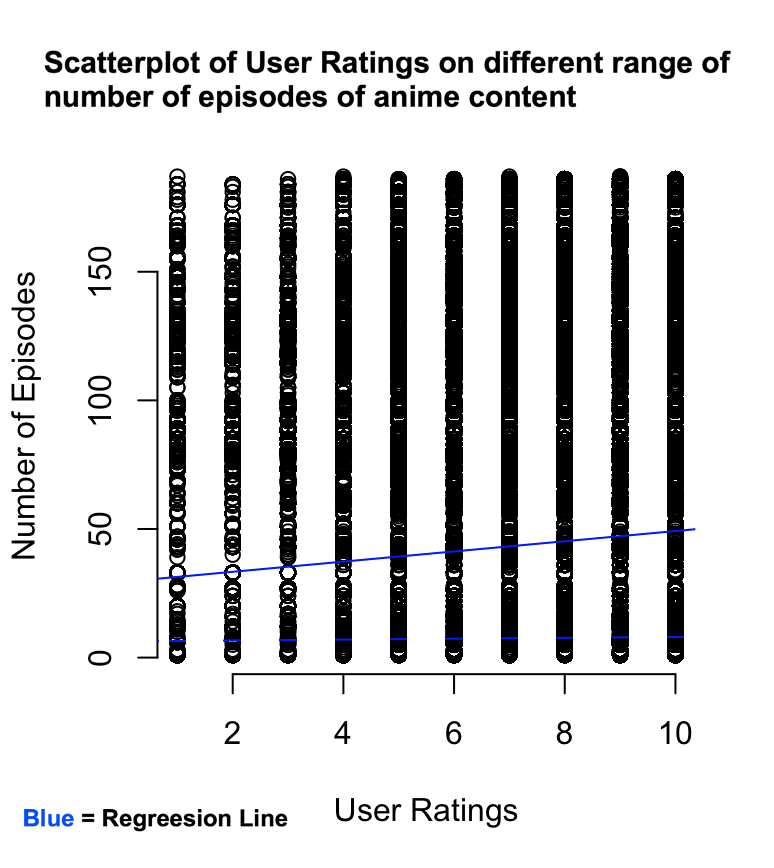




Third research question :

* Does higher number of total episodes of a anime content has effects on the ratings given by a particular individual?

By implementing linear regression statistical modelling technique, a scatter plot with a regression line has been generated and the regression line has shown that there is a slight correlation between the number of episodes of an anime content and the ratings given by the users. It shows that the ratings given by users are peaked when the number of episodes is closing to 50. Hence, higher number of total episodes of a anime content indeed result in higher ratings given by a particular individual but only when the number of episodes is below 50.



Forth research question (Predictive) :

* Based on the results of the analysis on the descriptive questions above, if anime contents with historical ratings of 7 and above are suggested by the system to the users, will the users be able to discover more anime contents that are within their liking?

According to the answer to the first research question, it is safe to say that when users are recommended with anime contents with historical ratings of 7 and above, the users have higher probability to give them higher ratings which implies that they have higher or greater likings towards the contents. Hence, if anime contents with historical ratings of 7 and above are suggested by the system to the users, statistically the users will be able to find more contents that are within their likings.

Restrictions and Challenges :

In the process of analysing the dataset using RStudio, the volume and size of the datasets have caused some slight efficiency problems related due the hardware limitations. For example, a single scatter plot generation for the variables, User Ratings and Historical Ratings of all anime contents in the dataset has over 6,337,239 individual plots and the generation has taken RStudio over 15 minutes to process.



Further Insights to be generated :

For further, more in-depth insights on the dataset, multiple factoring variables can be analysed by combining type of anime content, genre of anime content, historical rating of anime content and members of anime content, to understand the correlation of the combination of variables towards user rating on the anime contents.

# Part A

**The Company and Focus**

Company: Miniclip

Focus: MIniclip focuses on building better games and also provide the best experience possible for its players by understanding the gamers themselves as in what games they like or what actions they take in the game.

**Questions Answered**

- How to give the best service/experience to the users?

- What are the steps taken to build better games for the gamers?

- How to understand the users to provide them the best experience possible?

**2.MIniclip’s Leverage on Data Analytics**

Miniclip started on data collection and analysis by understanding the data and metrics they wanted to track , and also determine how to consume the data with the company team and its key stakeholders from every department.

Next, Miniclip wanted a tool in place that made customer data easily accessible to everyone at Miniclip.The company chose BIME which is a cloud business intelligence (BI) pioneer that has partnered with Google to provide robust big data analytics experience.

Miniclip used Amazon Web Services(AWS) which is a secure cloud services platform, offering compute power, database storage, content delivery and other functionality to help businesses scale and grow.This AWS was used by Miniclip to create useful data products.One such product is UA LTV which is an interactive data product that allows business users to analyse predicted LTV(Lifetime Value) across all possible cohort combinations.This data product was coded in R with Shiny and interacts with Amazon S3, Redshift and Amazon EC2 instances in AWS.

Miniclip used Amazon DynamoDB one of the AWS for a custom game data service that allows third-party games to store almost any type of data in this service without requiring development support which was a perfect fit for DynamoDB's non-schema-based data storage.Miniclip also used DynamoDB for simple relational schemas, which allowed the company to store and retrieve the data in a fast and highly available manner.

**3.Benefits and Impacts of Data Analytics**

The benefits were,the product team were able to better assess the economy of each game by having better knowledge of the performance of their games and their efforts in customer acquisition and engagement by using BIME.In BIME, the data team has created custom KPIs(performance indicator) that are core to Miniclip’s business, like revenue per user, daily active user, average tenure, and customer LTV by cohort. This has helped the product team at Miniclip to monitor the KPI dashboards everyday in which game performance and experience provided to the users could be improved.

Other than that,the Amazon DynamoDB used enabled Miniclip to empower their business by being able to analyze consumer needs, preferences and buying behaviors faster and in return launch better products in the future as the company was able to store and retrieve the data in a fast and highly available manner.

**Different Analytic Implementation to Business**

-Miniclip had used Amazon S3 and Amazon DynamoDB from the many services provided by Amazon Web Services(AWS).

Amazon Simple Storage Service (Amazon S3) is object storage with a simple web service

interface to store and retrieve any amount of data from anywhere on the web.Amazon S3 can be used as primary storage for cloud-native applications such as a bulk repository, or "data lake," for analytics.It's simple to move large volumes of data into or out of Amazon S3 with Amazon's cloud data migration options. Once data is stored in Amazon S3, it can be automatically tiered into lower cost, longer-term cloud storage classes like Amazon S3 Standard, Infrequent Access and Amazon Glacier for archiving.

Meanwhile, Amazon DynamoDB is a fast and flexible NoSQL database service for all applications that need consistent, single-digit millisecond latency at any scale.It is a fully managed database and

supports both document and key-value data models. Its flexible data model and reliable

performance make it a great fit for mobile, web, gaming, ad-tech, Internet of Things (IoT), and

many other applications.

This shows that Amazon S3 provides data storage service while Amazon DynamoDB provides NoSQL database service both important for game analytics and product creation at MiniClip.

BIME is an easy yet powerful service that connects to and analyzes data in any organization.

It is a simple-to-use BI app based on cloud computing innovations and data visualization.

With BIME, you can use the internet as your data warehouse, produce compelling visualizations, and share insights.Meanwhile with AWS, a cloud service from Amazon,provides various services/building blocks which are designed to work with each other, and result in applications which are sophisticated and highly scalable. Both BIME and AWS provides useful cloud services to its users such as the Amazon DynamoDB and Amazon S3 where BIME focuses more on customer analytics while AWS enables focus on data products by Miniclip.

**4.Strength and Weaknesses of Analytics Implementation**

**Strengths**

- BIME is able to create advanced visualizations including all the classics plus maps, funnels, bullet charts, relational analysis, and hundreds of customization points.This enabled visualization of relationships and patterns between operational and business activities which easily permitted the company to discover the problems in the business such as performance of the games.

-BIME has connectors to Big Data sources such as BigQuery, HANA and Redshift which enabled the company in getting a 360-degree view of their business as the many Big Data sources in BIME helps in providing different services in data development in the company.

-Amazon DynamoDB is flexible data model and provides reliable performance which makes it a great fit for mobile and web gaming.Amazon DynamoDB is designed to deliver consistent,fast performance at any scale for all applications. Average service-side latencies are typically single-digit milliseconds. As your data volumes grow and application performance demands increase, Amazon DynamoDB uses automatic partitioning and SSD technologies to meet your throughout requirements and deliver low latencies at any scale.

-Amazon S3 provides durable infrastructure to store important data and is

designed for durability of 99.999999999% of objects. Data is redundantly stored

across multiple facilities and multiple devices in each facility.Also by using Amazon S3, there is unlimited data storage and data can be accessed whenever needed.

**Weakness**

- BIME does not work in all available platforms such as Android app and Iphone app.

- The business intelligence solution at BIME doesn’t offer on-premise deployment option which can be a problem for some users.

-Hosting is provided only for static websites in Amazon S3

-If running a website with heavy-duty activities the size of the files you store will increase the cost will be more.

-Unit tests are slow and expensive with Amazon DynamoDB

-Amazon DynamoDB lacks a sleek query language. This makes performing ad-hoc queries much harder.

**5.Potential Future Business Cases/Questions**

**-What are the skills the company should focus on to compete in the market?**

The company team should focus on sharpening their skills and also learn new skills to keep up with the market in order to provide content/services to customers which would help to pull in customers.

**-What will be the most effective marketing and promotional strategy?**

The company should have an insight on strategic marketing and do research before promoting their company products/services to make sure they make maximum revenue out of their products and services.this ensures the constant growth in business of the company and promotes the company well.

-**What are actions/activities carried out by the customers when they are using the gaming website?**

**Reference**

Stephanie Zou(2016). How gaming company Miniclip is getting to know their players. Retrieved from : *https://www.zendesk.com/blog/miniclip/*

Stephanie Zou(2016). Analyst Hero of the Month: Paul Bugryniec @ Miniclip.Retrieved from :*<https://blog.bimeanalytics.com/english/category/miniclip-centralizes-its-data-strategy-to-better-understand-its-gamers>*

AWS Case Study: Miniclip.Retrieved from: *<https://aws.amazon.com/solutions/case-studies/miniclip/>*

About:What We Do.Retrieved From*:<https://corporate.miniclip.com/about/>*

Ricardo Vladimiro(2016).Who wants to be a Data Scientist at Miniclip?!Retrieved From:

*<https://ongamesndata.wordpress.com/tag/data-science/>*

# Part B

**Context and Research Questions**

The context to be examined will be customer behavior towards a game.

A.What are the games most liked(played the most) by the customers?

B. For which games are the purchase rate higher?

C.Which games are played the longest by the users?

D. Do the users play or purchase games more?

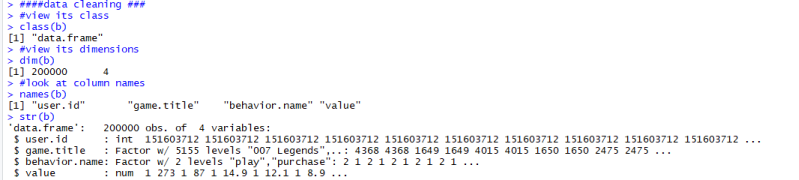
E.Does the no of hours a game is played affect the the number of purchase for that particular game?

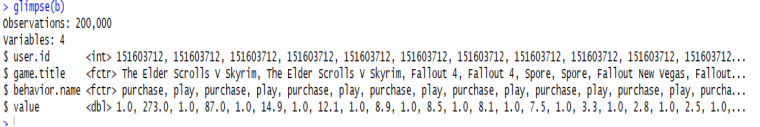
**Related Datasets and Its Contents**

The dataset used is a gaming analysis dataset which is generated from Steam, the world's one of the most popular PC Gaming hub.The dataset is gotten from<https://www.kaggle.com/datsets.> This dataset is a list of user behaviors, with columns: user-id, game-title, behavior-name, value. The behaviors included are 'purchase' and 'play'. The value indicates the degree to which the behavior was performed - in the case of 'purchase' the value is always 1, and in the case of 'play' the value represents the number of hours the user has played the game.

**Datasets Examination and Data Preprocessing Activities**

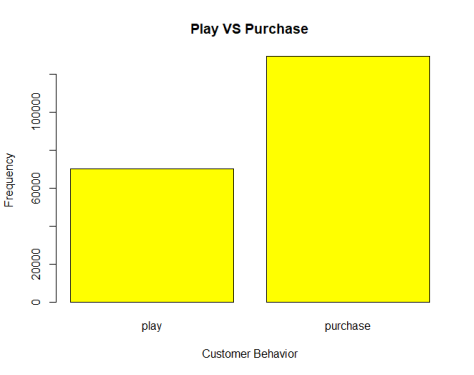
The data preprocessing activities done are first is data cleaning. The dataset was checked for unwanted data and update the dataset if any error is detected.Examples of data checking done are checking the class,dimensions,column names,view the top and bottom of the data,summary of the data.





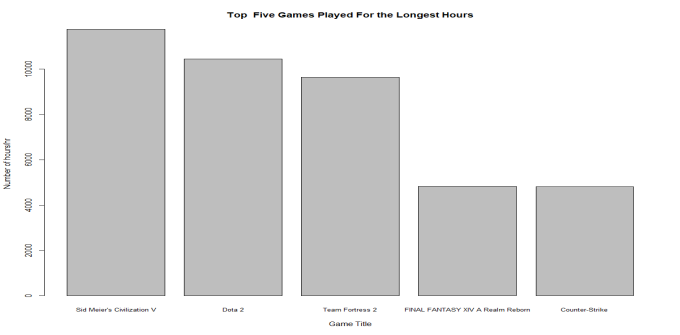
**Exploration of Descriptive Statistics of the Data**

I)The number of purchase is found to be higher for games than the number of times the games are played.

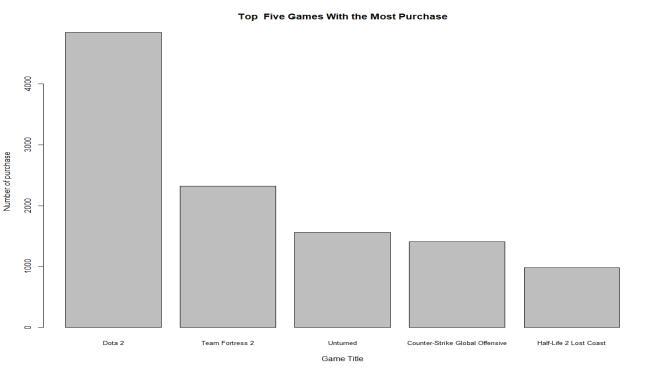


ii)Below is the five games which were played the longest hours by the users with the highest number of hours of 11754.00 for the Sid Meier’s Civilization V game.

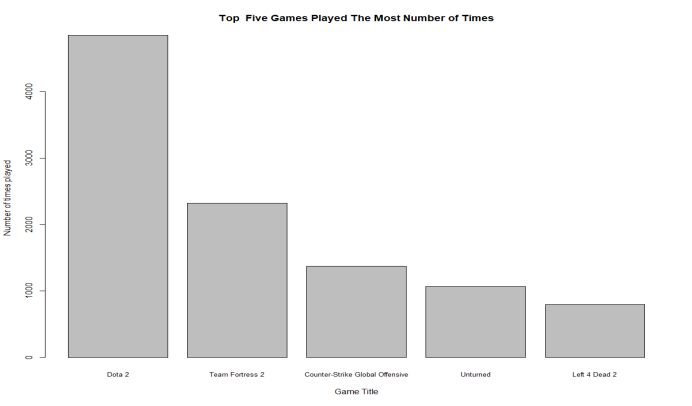




ii)Below is the five games with the most purchase by the users with the highest number of purchase for Dota 2.



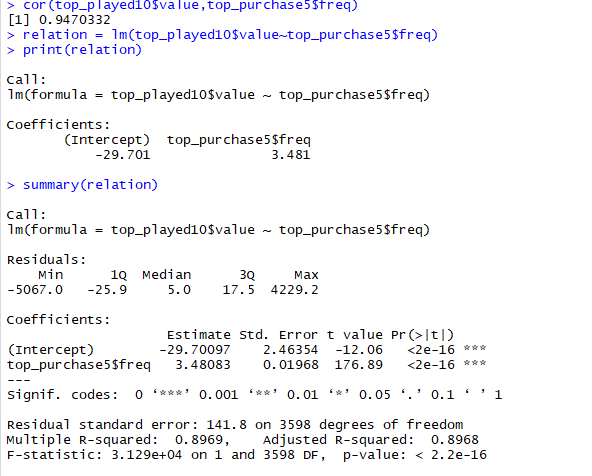
Iii)Below is the five games which is played the most by the users with the highest number of times played is for Dota 2.

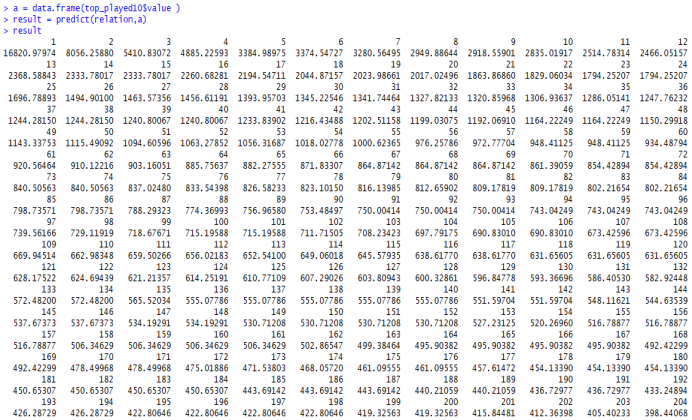


From the data obtained we can say that , Dota 2 is most purchased and most played game in Steam and the second game played the longest hour.The data shows that the most popular game among the users is Dota2 based on the aspects studied(no.of purchase,no.of times played,how long played).

**Statistics Modelling Technique for Prediction**

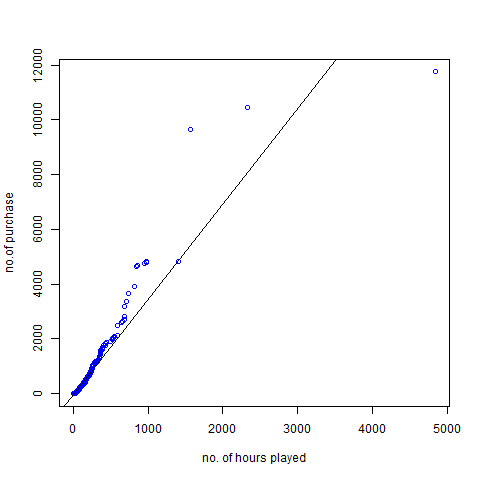
Linear regression modelling technique was used study whether the number of hours used to play a certain game affects the purchase rate of the games.

.



**Summary**

Below is the linear regression visualization.The graph shows a linear increase in the number of purchases for games as the no of hours played increases.



By knowing the number of hours a game is played affects the number of purchases per game, more emphasis should be given to create a better and updated version of the game to keep the interest intact and the users are compelled to buy the game.

Challenges:

RStudio easily consumes memory and hence some R commands take a longer time to load, which is time consuming and data takes a longer time to process.

Way forward for further insights to be generated is to have a list of expectation of the data you have before going ahead with data exploration.Also learn to tell a story with data as more understanding of the data is obtained especially with more visualizations.

**PART A**

1. **The Company and Focus**

Company : Spotify

Spotify is a digital music, podcast, and music streaming service that gives user access to millions of songs and other content from artists all over the world. It was officially launched on 7 October 2008 at Stockholm, Sweden by Daniel Ek and Martin Lorentzon. It has 140 million user (70 million paying) around the world.

**Focus of the Company**

To provide a best and broader service for music streaming and entertainment beside regenerate the lost value by converting music from poorly monetized format (piracy) to a paid streaming format.

**Translate the focuses into Questions that were answered:**

a. Can Spotify survive with their current business model and how?

b. Are Spotify harming piracy more than they are harming sales?

c. How Spotify distributing revenues fairly to artists and songwriters?

d. What are Spotify ideas on how to increase people to use their platform for music streaming?

e. What are the steps taken by Spotify to keep update with user preferences or choices given there are thousands of new songs created every day?

2. **Spotify’s Leverage on Data Analytics**

**A. Data Collection**

Basically, Spotify company uses Identity-Driven Information Gathering Technique to collect data from user. Spotify uses this technique by following or keep tracking with user when they click the ‘Play’ button on certain playlists, on what day and what time for Spotify enhancing their user personal experience. The company puts a requirement that every user, either paying it or not paying, must sign up first to use the service. From sign-up session only, Spotify collected an enormous amount of data on what people are listening into, where and in what context. From the data collected from users, Spotify knows a lot about our personal information, such as age, gender, and location.

A recent privacy policy changes shows that Spotify will collect information store inside user’s mobile device, which are contacts, photos, media files and GPS (the location of the user). This new announcement received many outrage from users around the world but as explained by Spotify CEO, Daniel Ek in his blog post, the reent changes in policy regarding collecting data is for upgrading personal experience for each user. That is how Spotify collect data, which is by keep tracking on what user click on application and by through user registration to create an account.

**B. Data Preprocessing**

At Spotify, they use Pythonflow in data preprocessing pipelines because it automatically caches computationally expensive operations, any part of the computational graph can be easily evaluated for debugging purposes, and lastly it allows Spotify’s programmer to distribute data preprocessing across multiple machines. When Pythonflow evaluates an operation, it will check whether the current context provides a value for the operation and return it immediately if possible. If the current context does not provide a value for the operation, Pythonflow will evaluate the dependencies of the operation, evaluate the operation of interest, store the computed value in the context, and return the value.

Spotify also use Luigi, a recently open-sourced Python framework that simplifies batch data processing, helping them to build complex pipelines of batch jobs, handle dependency resolution and create visualizations to help manage multiple workflows besides from Pythonflow. It also comes with Hadoop support build in. Luigi provides an infrastructure that powers several Spotify features including recommendations, top lists, A/B test analysis, external reports, internal dashboards and many more. The objective of Luigi is to address all the plumbing typically associated with long-running batch processes. Failures will happen if programmer wants to chain many tasks and automate them. These tasks can be anything, but are typically long running things like Hadoop jobs, dumping data to or from databases, running machine learning algorithm, or anything else. Conceptually, Luigi is similar to GNU Make where programmer has certain tasks and these tasks in turn may have dependencies on other tasks.

For collaborative filtering algorithm, they use Spark. They use it because Spotify made audio recommendation system as an important part in their product. Apache Spark is a fast and general-purpose cluster computing system. Spark provides data engineers and data scientists with a powerful, unified engine that is simple, highly scalable, and effective integrable with other tools. Spark can help data scientists solve highly complex machine learning problems involving graph computation, streaming, and real-time interactive query processing in an interactive way and at a much greater scale. This already explains why Spotify choose Apache Spark as one of their analytics tools for data preprocessing.

**C. Data Analysis**

Spotify analyses the data to recognize trends, discover bugs, and analyse the effect of an event on a user and the entire ecosystem. Spotify uses three tools as their analytics tool to analyse data. The three tools are Dashboards, Data Warehouse, and Luigi. Dashboards provides an interface similar to Google Analytics and allows users to create their own custom screens containing data they are interested in Spotify’s pipeline. As example, Spotify have dashboards that show user growth in particular regions, or user engagements. While Data Warehouse is a more complex system that allows user to access Spotify data-set directly. From this tool, user can query data, create map or reduce jobs using Hive, and even create mini data pipelines. Luigi is for more operations. It is made to talk to any storage systems, run machine learning algorithms and even provide daily reports.

The infrastructure of data analysis implemented in Spotify is a set of daemons that constantly parse the syslog on production machine. It always adding most of Spotify recurring data into their data analysis pipelines. Matching data is compressed and periodically synced to HDFS. Within 24 hours, typically data is available in their Data Warehouse and Dashboards. However, in some cases data is available within a few hours or even instantly through tools like Storm.

**D. Data Mining**

Spotify leverage on data mining techniques and solutions by allowing user to take data as personal experiences. Spotify user can take the data from Spotify to know more details about what genre they like, what songs they hear the most and many more. Hence, Spotify created and launched Spotify Web API. This is for user to extract data from Spotify. Commonly known as a RESTful API (Representation State Transfer), today the API has 40 distinct endpoints through which developers could retrieve and manage Spotify over the Internet. The Web API uses the same HTTP protocol as other Web in Internet.

Spotify Web API lets users to fetch data from the Spotify music catalogue and manage user’s playlists and saved music. Based on simple REST principles, Spotify Web API endpoints return metadata in JSON format about artists, albums, and tracks directly from the Spotify catalogue. The API also provides access to user-related data such as playlists and music saved in a “Your Music” library, subject to user’s authorization.

**E. Data Visualization**

For data visualization, Spotify is using two tools, which are Carlo and EchoNest. Carlo is a platform that visualizes data for easier navigation. Spotify exported their internally collected data into Carlo to create a visualization that update the distinctive city playlists twice a month, each involving the analysis of approximately 20 billion listener and track relationships. The entire process does not involve the need for back-end developers or other specialists. Pop-up windows will display the name of the city and its corresponding top track lists if user hovering over the map offered by Carlo.

In March 2014, Spotify acquired EchoNest for their data visualization process. EchoNest is a music intelligence company that does things like determine what recommendations to make to listeners for automatic streaming radio services. EchoNest provides broad and deep data on millions of artists and songs, making it easy for Spotify programmers to create an awesome listening experiences for user. This tool can build world-class music apps that take advantage of all the capabilities and deep data provided by API.

**3.Benefits and Impacts of Data Analytics**

There are many benefits and impacts of analytics implementation implied in Spotify’s business. First, they got a very positive impact on user engagement and help more users to come back to using Spotify application more often. They implemented a system that holds a set of data consisting user emails for them to tell the users if their friends joined or if new songs were added to a playlist they subscribed to. Spotify programmers built a fairly simple system that had the ability to deliver a lot of emails and also provided a way for more people to create email templates. This system uses many big data analytics implementation such as data preprocessing and data mining. The system made Spotify get positive impact on user engagement between subscribers and company by getting good feedback from users and they were able to tracking the email if they had any effect on user’s listening habits, user’s account status and so on. Spotify know how to take advantage of using big data and it gave a powerful effect on the company business.

Spotify also seen the number of its active users grow from 15 million to 60 million since data analytics is being implemented. Spotify really mean it in doing big data business. They investing heavily by acquired the Echo Nest a few years back ago to improve its music recommendations system. From there, they created ‘Discover Weekly’ playlists for user of Spotify application. ‘Discover Weekly’ playlists got rave reviews from users all around the world and become the greatest upgrades did by Spotify company. This music recommendations system is using many data collection for analysis. They have to collect many data from each user to discover their personal favourite and then recommend to the user. This is why many people like ‘Discover Weekly’ playlist from Spotify and unintentionally increasing their user base.

Spotify also acquired Preact, a startup that acquire and retain subscribers by using analytics and machine learning. Preact are now working exclusively in helping Spotify production team to predict user behaviour around subscription sign-ups and upgrades to gain a much more new subscribers.

The impact of analytics implementation to Spotify is Spotify is considered as a symbol of resurgence of music industry along with Big Data technology. Each industry has a problem with technology in this era. For music industry, music has many times failed to keep up with the rapid pace of big data technology. This is due to many consumers turned to illegal downloading (piracy) for the sake of free pricing. Then, come Spotify with the idea of music streaming. Streaming offered solution with the potential to overcome illegal downloading. From recommendation engines to choosing the perfect individual playlist, data is redefining the dynamics of the music industry and the relationship between music and its listeners, in more creative ways than ever.

The other impacts created from the emergence of Spotify is many use Spotify’s data in decision making and providing forecasting information and business analytics. For example, in 2013, the database used streaming data to predict the winner of The Grammys Award. They made this possible by breaking down the user’s listening habit by taking into account the song and the album that was being streamed to determine the popularity of the music. As it turned out, 4 out of 6 of their predictions is true. Spotify would not have turned out the way it did without of the implication of big data analytics.

**Analytics implementation of Spotify**

**MLlib & PySpark**: Spotify using MLlib and PySpark that offer a collaborative filtering algorithm to implement audio recommendation system. These implementations also offer pattern detection analysis involving user interest. Basically, the system plays a critical role in providing personalized recommendations. There are two main approaches Spotify programmers use for building audio recommendation systems. The first one is content filtering, which uses known information about songs particularly and users to make recommendations. With this approach, they create profiles based on songs and users (demographic information and personal information). Spotify programmer choose a song based on hundreds of characterization while a user provides information about his/ her music preferences. Recommendations are made based on pairing these two sources. The other one is collaborative filtering. Spotify prefers this approach. This technique uses previous users’ input or behaviour to make the future recommendations.

**Luigi**: Luigi is a Python module that helps programmer to build a complex pipeline of batch jobs. It handles dependency resolution, workflow management, and data visualization. The purpose of Luigi is to address all the plumbing typically associated with long-running batch processes. One of the processes that related to Spotify is graph dependency. Using Luigi’s visualiser, people can get a nice visual overview of the dependency graph of the data workflow.

**Google Cloud Platform**: Spotify chooses Google Cloud Platform to power the implementation of data infrastructure . The company split their data business in Google Cloud into two streams, a services track and a data track. Spotify runs their products on a multitude of tiny microservices, several of which are now being moved from on premise data centers into Google’s cloud using Cloud Storage, Compute Engine and other products. With Compute Engine, teams can rely on consistent performance from ultra high IOPS SSD and local SSD storage capabilities. Spotify implements Google Cloud Datastore and Google Cloud Bigtable for their data storage. These two storage services lets Spotify engineers work on complex back end logic, instead of focusing on how to store the data and maintain databases. Spotify is also deploying Google’s Cloud Networking services to transfer data.

Hence, the above are some of the analytics implementation in Spotify business to achieve their main objectives.

**4.Strength and Weaknesses of Analytics Implementation**

**A. MLlib & PySpark**

Strengths:

1. The integration is simple, and migration of data provided by MLlib can be done easily by Spotify engineers.

2. The accuracy of scaling to big data is significantly improved in Spotify analytics implementation. Even though the Spotify datasets and model size are increasing, Spotify still can gaining the accuracy of the data.

Weaknesses:

1. This tool consumes a lot of memory and issues around memory consumption are not handled in a user friendly.

2. MLlib & PySpark would take large resources in processing the data in Spotify.

**B. Luigi**

Strengths:

1. Luigi comes with Hadoop built in which also used by Spotify in analytics implementation.

2. Luigi provides an infrastructure that powers several Spotify features including recommendations, top lists, A/B test analysis, external reports, internal dashboards, and many more.

Weaknesses:

1. Relatively small number of tasks, requires writing subclasses for most of Spotify requirements.

2. Possibly not suitable for streaming data, have performance concern for Spotify.

**C. Google Cloud Platform**

Strengths:

1. Spotify is one of the best music streaming services because of it improved performance in flat average response times, minimal errors and no spikes.

2. Google Cloud Platform has live migration of virtual machines. This is a very important functionality for Spotify as there are not many cloud providers have this function. Hence, there are no noticeable degradation in performance when they are live migrating VMs between host machines.

Weaknesses:

1. Spotify has to be fully dependent on Internet connection to use Google Cloud Platform because it is an internet based. No cloud provider, even the very best, would claim to be immuned from service outage.

2. Spotify has to be aware of the risk in security and privacy in using Google Cloud, especially when it comes to managing sensitive data such as personal information.

**5.Potential Future Business Cases/Questions**

1) **What are the steps will Spotify company take to going forward parallel with the rapid change of Big Data technology?**

As our Big Data technology is changing really fast, how Spotify planning their works to survive in music streaming industry and as well compete with other big names in industry such as Rdio, Apple iTunes and many more. Are they will take or acquire other company that offers better analytics tools to produce more better product?

2) **Does Spotify have to increase their variation of types of datasets for customer behavioural analysis in enhancing user engagement?**

By accessing data about consumer from multiple sources that Spotify already offered, such as demographic and personal information, Spotify is been great in creating personalized playlists for every subscribers around the world. Spotify has been used data from user profiles and user playlists to provide recommendations for each user. By combining data from million users, Spotify is able to make recommendations even if a particular user does not have an extensive history with the site. However, it is not enough when customer demand more than that. Are Spotify needs to variate their methods in collecting data (such as collect their personal profiles) to create better recommendation system? Is it a must for them to add more types of data such as include users’ friends behavioural in create ‘Discover Weekly’ for users?

3) **Does Spotify need to alternate their way or objectives in music industry by create algorithms for predicting the next big things in music industry?**

In this technology era, the music industry lays great emphasis on predicting the future.

This happens at all levels, from deciding what the individual user of streaming service wants next on their playlist, to discovering hidden gem in music. Recently, Big Data proved that it has the ability to do just that. Hence, does Spotify have to use this information regarding Big Data and use it as their advantages? Does they wants to create a new system to predict much bigger things such as estimating what is the next best songs that will go viral? Or can this song win bigger award in next year? This could be pretty helpful for Spotify in predicting the future and make users more intrigue than ever.

**References:**

Bernard Marr. (2017). The Amazing Ways Spotify Uses Big Data, AI And Machine Learning To Drive Business Success. Retrieved from:

[*https://www.forbes.com/sites/bernardmarr/2017/10/30/the-amazing-ways-spotify-uses-big-data-ai-and-machine-learning-to-drive-business-success/#43f4b6914bd2*](https://www.forbes.com/sites/bernardmarr/2017/10/30/the-amazing-ways-spotify-uses-big-data-ai-and-machine-learning-to-drive-business-success/#43f4b6914bd2)

Michael J. Miller. (2011). Spotify’s Goal: To Have All the World’s Music. Retrieved from:

[*https://forwardthinking.pcmag.com/show-reports/282252-spotify-s-goal-to-have-all-the-world-s-music*](https://forwardthinking.pcmag.com/show-reports/282252-spotify-s-goal-to-have-all-the-world-s-music)

Neville Li. (2017). Big Data Processing at Spotify: The Road to Scio (Part 1). Retrieved from:

[*https://labs.spotify.com/2017/10/16/big-data-processing-at-spotify-the-road-to-scio-part-1/*](https://labs.spotify.com/2017/10/16/big-data-processing-at-spotify-the-road-to-scio-part-1/)

Hayley Helms. (2016). How Spotify Uses Data-Capturing to Connect With Consumers. Retrieved from:

[*https://www.adventuresportsnetwork.com/transworld-business/how-spotify-captures-its-data/*](https://www.adventuresportsnetwork.com/transworld-business/how-spotify-captures-its-data/)

**PART B**  
  
In Part A, we had explained and broader our knowledge on the various entertainment business cases and insights that can be derived within entertainment industry. Picking up from the Questions we identified in Part A, we need to:   
  
1. **Set the context that we would like to examine i.e. the Question(s) we would like to answer**.  
  
Description questions:  
- Given the full listening history for 1 Million users, what are the most important attributes of a song for a music recommendation system?  
- By analysing song characteristics, how many songs that many people like and how many do not in a set of songs?  
  
ONE Predictive Question:  
- Do song characteristics and attributes can lead to higher revenue growth and steep increase in service subscription in years to come?

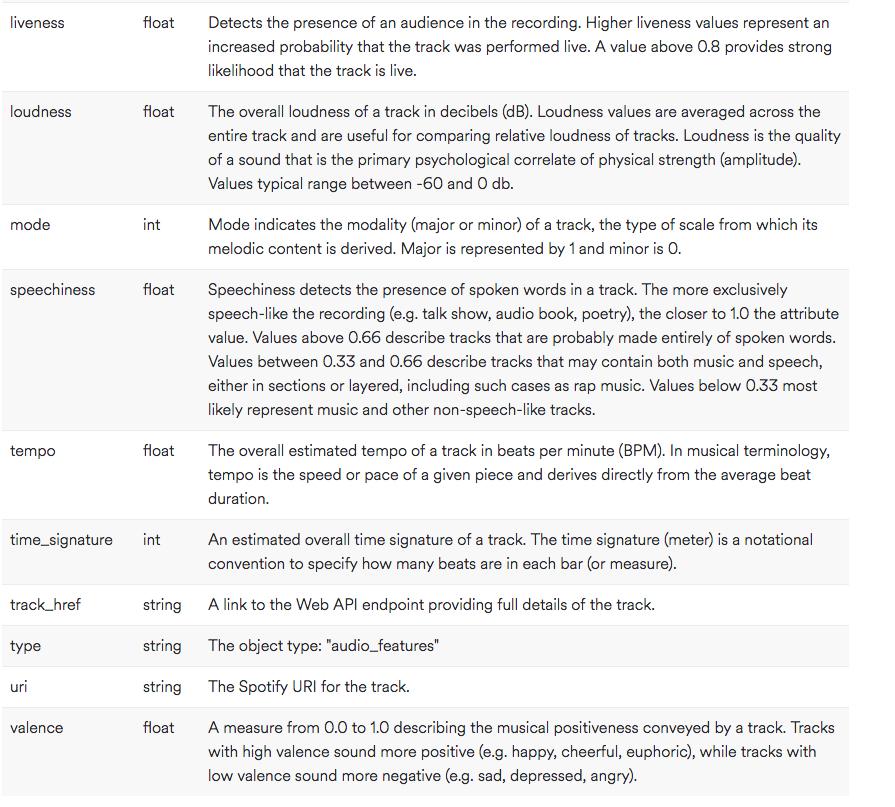
2. **Identify related dataset(s) that will be useful and describe the contents**.  
  
 The other dataset are consist of two subsets which are Spotify’s Worldwide Daily Song Ranking and Spotify Song Attributes datasets. Both we found at the same place, Kaggle.com. Daily Ranking Song contains information about the daily ranking of the 200 most listened songs in 53 countries from 2017 and 2018 by Spotify users. The dataset has ranking position (or index basically based on popularity), track name, artist, streams, URL, date and region. While for Spotify Song Attribute dataset, it consists of song name, artist name, label of the song, and the others are 13 track attributes: acousticness, danceability, duration\_ms, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time\_signature, and valence.  
  


Figure 1.1 Audio Features Object

3. **Examine the quality of the data and perform the necessary data pre-processing activities**.  
  
 For data pre-processing activity to check the quality of the data, there is no necessary action to be taken because the data is given by Spotify itself. Hence, the quality if the data is good.