LSE

LSE_DA301 ASSIGNMENT

ASH JOHNSON

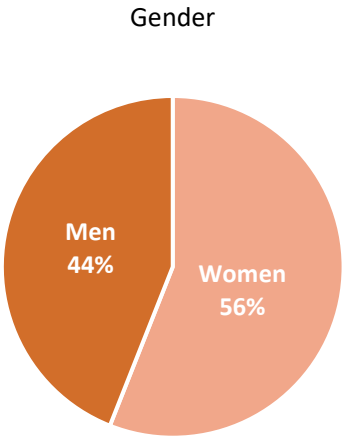
LSE DA301 ASSIGNMENT

OBJECTIVES

We have conducted exhaustive analysis for Turtle Games to find actionable information based on their sales and reviews data. During the course of this document we will cover the methodologies and tests done on the data along with conclusions on each test.

REVIEW ANALYSIS

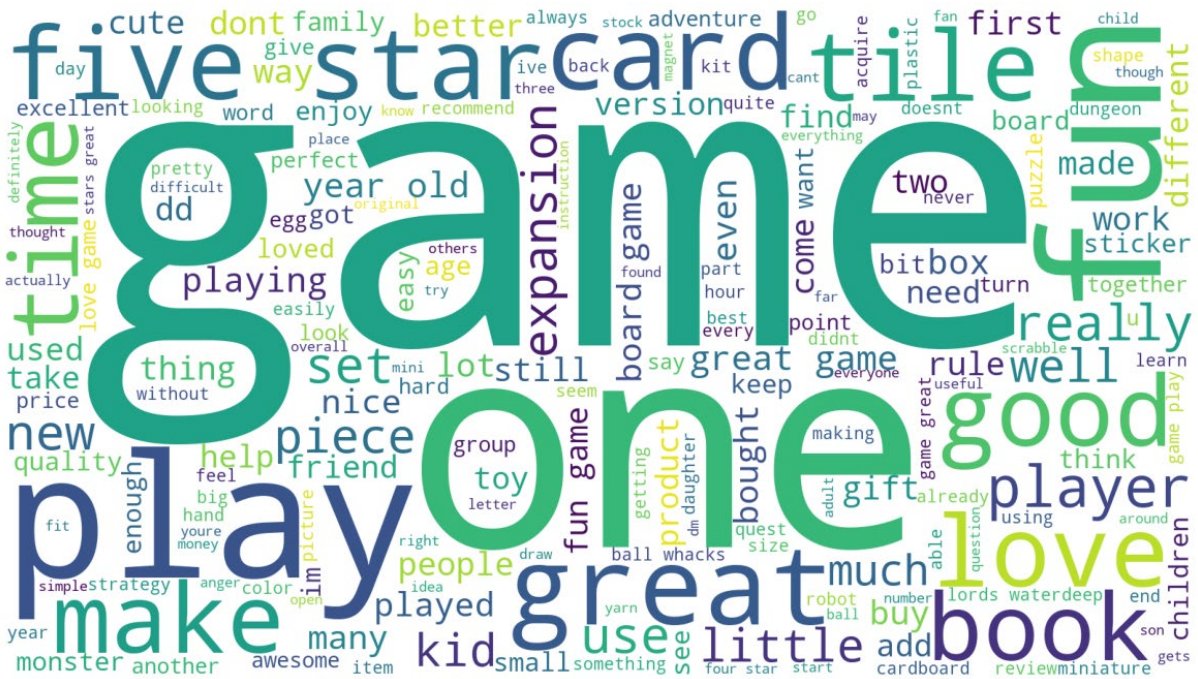
We looked at 2000 reviews by customers. Some of the key factors on the demographics of our users is as follows:



Education	Count
Basic	50
PhD	460
Diploma	190
Graduate	900
Postgraduate	400

As we can see from the above information – majority of our reviews were from women. In addition, our sample is very well educated with majority of them be graduates or higher.

We ran Natural Language Processing (NLP) tests on the data and came up with the following results on the reviews and summary reviews. We have 2000 records in our dataset. Our main aim is to do sentiment analysis on our data and come up with the top words in our dataset. We started by creating a word cloud of the most common words used and can see that the data is mostly positive with words like 'five star', 'great', 'love', 'good', 'fun', 'great game' and 'love' featuring in our data.



Our sentiment analysis shows that the mean polarity is .23 which is just above positive. However, to get a better idea of the sentiment of our data we will plot a histogram of the polarity. We can see that the data is mostly positive with a few negative reviews. There are a lot of neutral reviews in our data which brings down our mean polarity. Once we account for the neutral reviews our mean polarity, we can see very few negative reviews with most of them being positive. We can conclude that our data is mostly positive with a few negative reviews.

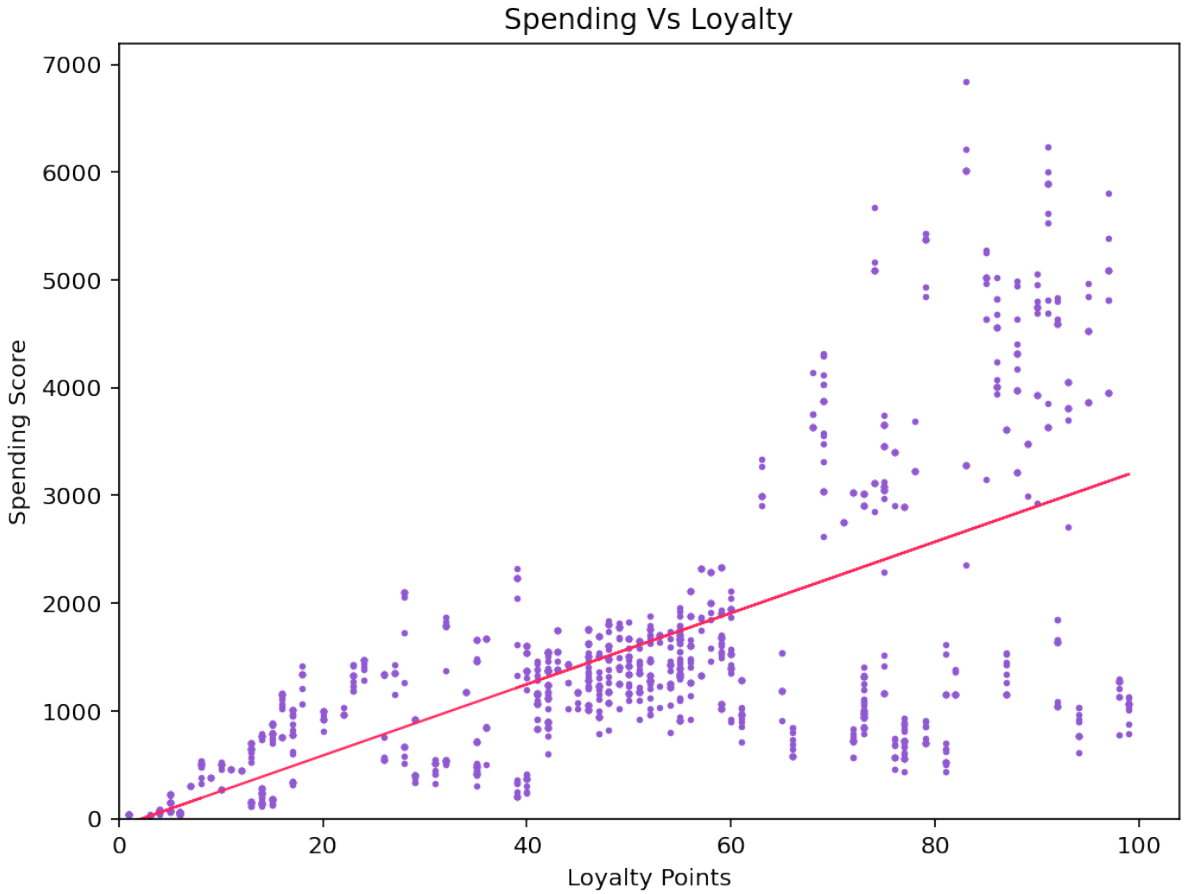
From our low amount of negative reviews we can see that a few people were disappointed and found the game difficult. However as most of our reviews were positive we should look at the positive sentiments. Some of the top 20 positive reviews have words like 'perfect', 'awesome gift' and 'awesome addition to my rpg gm system'. Some of the top 20 positive summaries are 'one of the best', 'excellent', 'awesome learning tool' and 'perfect gift'.

In conclusion we can conclude the game was a success.

CONSUMER ANALYSIS

In order to understand correlations between different features of our customers we plotted linear regression lines to compare the correlation between them. Some of the

Spending Score vs Loyalty Points

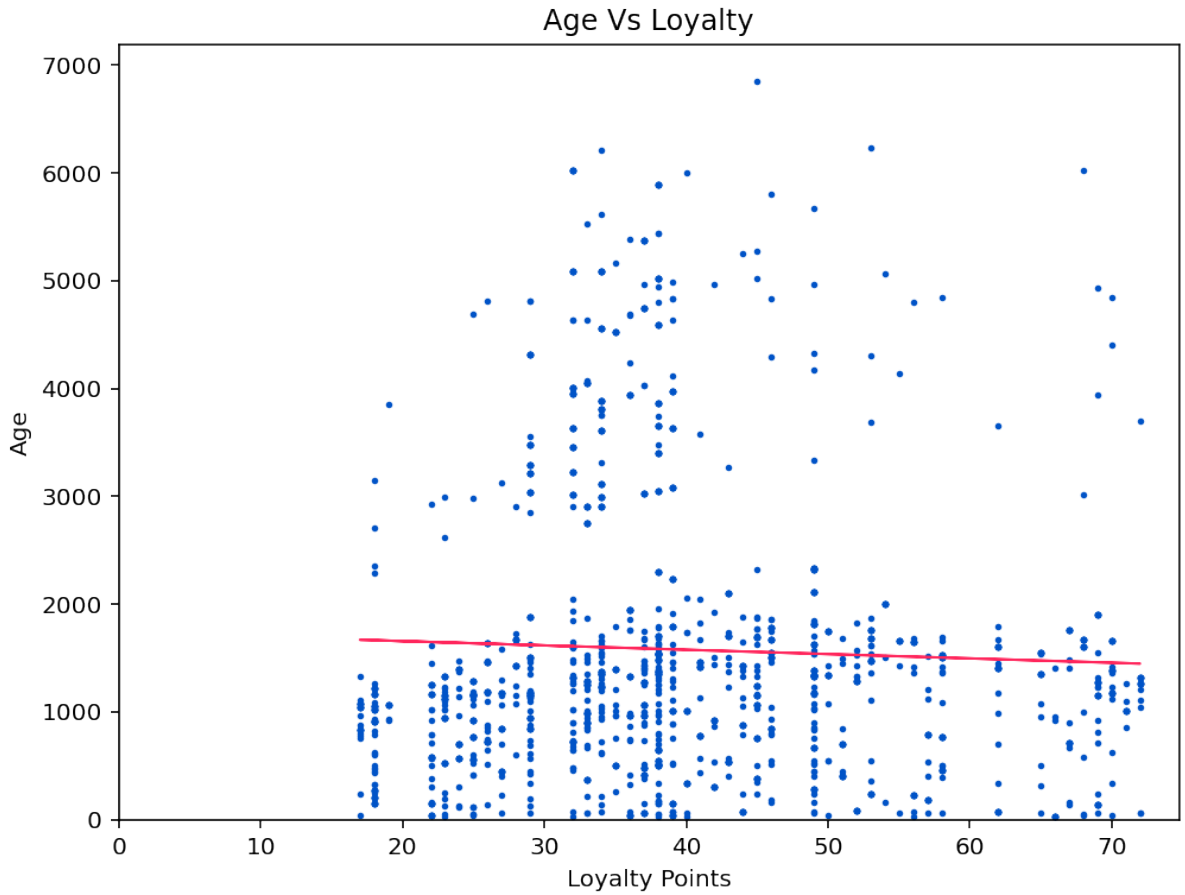


There is a weak correlation between loyalty points and spending score, our scatterplot has a megaphone shape showing high variance in the two meaning larger predicted values are also associated with large errors or residuals. In our regression model we have an R-squared 45.2% showing some corelation between the variables. Our T test has a value of 40.595 with a probability of T equal to 0 < 0.5 which would indicate that there is a corelation between these two variables. In addition, our F statistic is 1648 with a very low probability score. We can conclude that there is `weak correlation` between Spending and Loyalty Points

Remuneration vs Loyalty

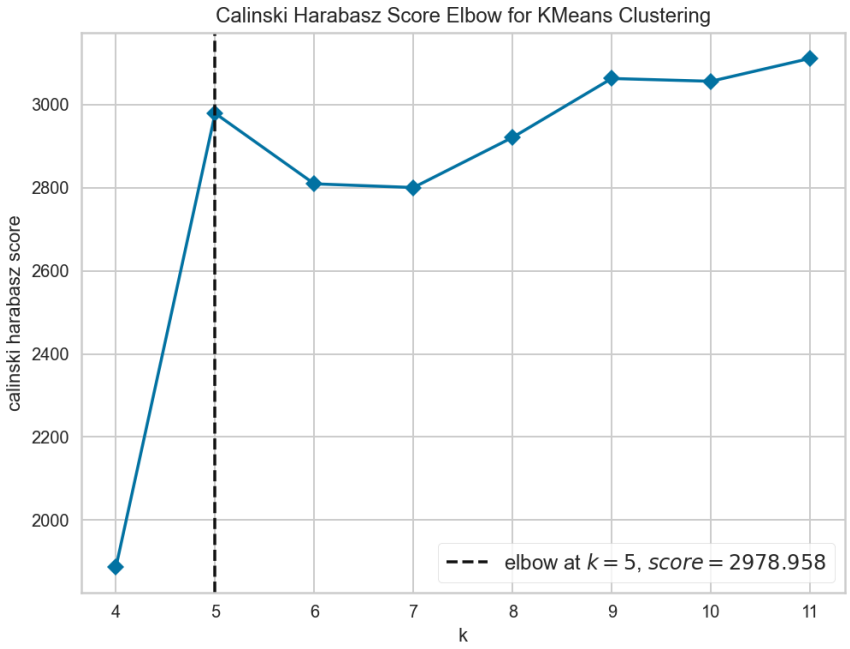
In our regression model we have an R-squared 38.0% showing some corelation between the variables. Our T test has a value of 0.978 with a probability of T equal to 0 < 0.5 which would indicate that there is very weak corelation between these two variables. In addition our F statistic is 1222 with a very low probability score. We can conclude that there is `very weak correlation` between Spending and Loyalty Points

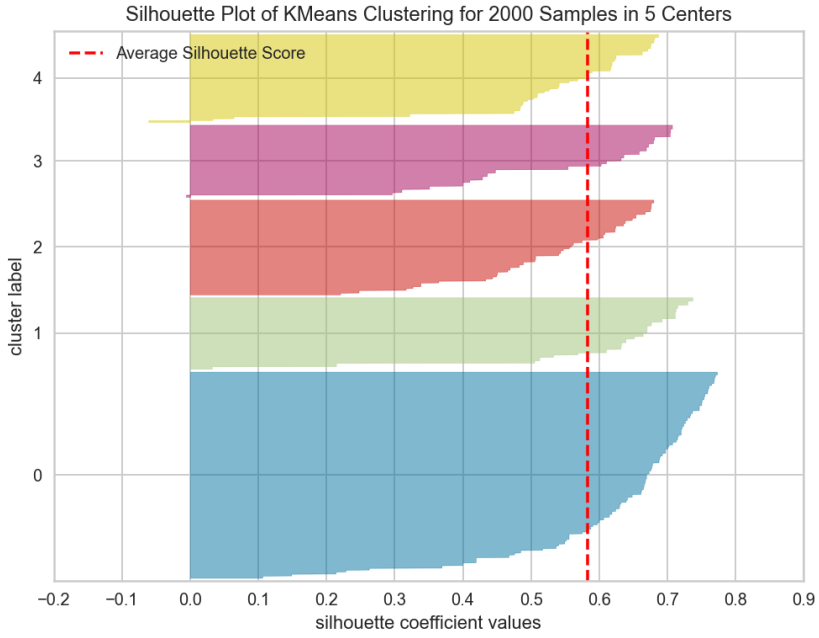
Age vs Loyalty



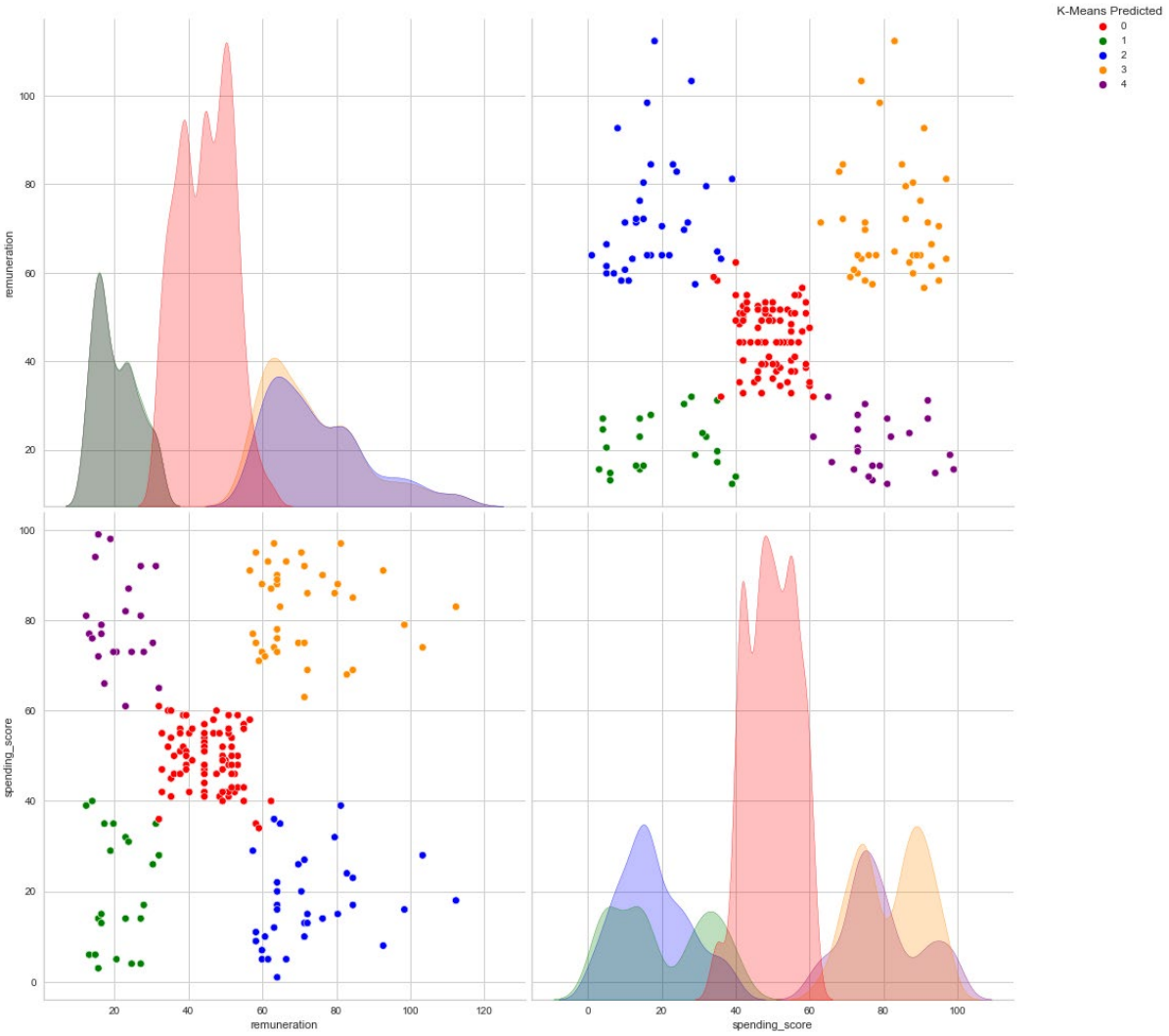
In our regression model we have an R-squared 0.2% showing no corelation between the variables. Our T test has a value of -1.899 with a probability of T equal to 0 .058< 0.5 which would indicate that there is no corelation between these two variables. In addition, our F statistic is very low at 3.606 with a very low probability score close to zero. We can conclude that there is `no correlation` between Spending and Loyalty Points

CLUSTERING OUR CUSTOMERS



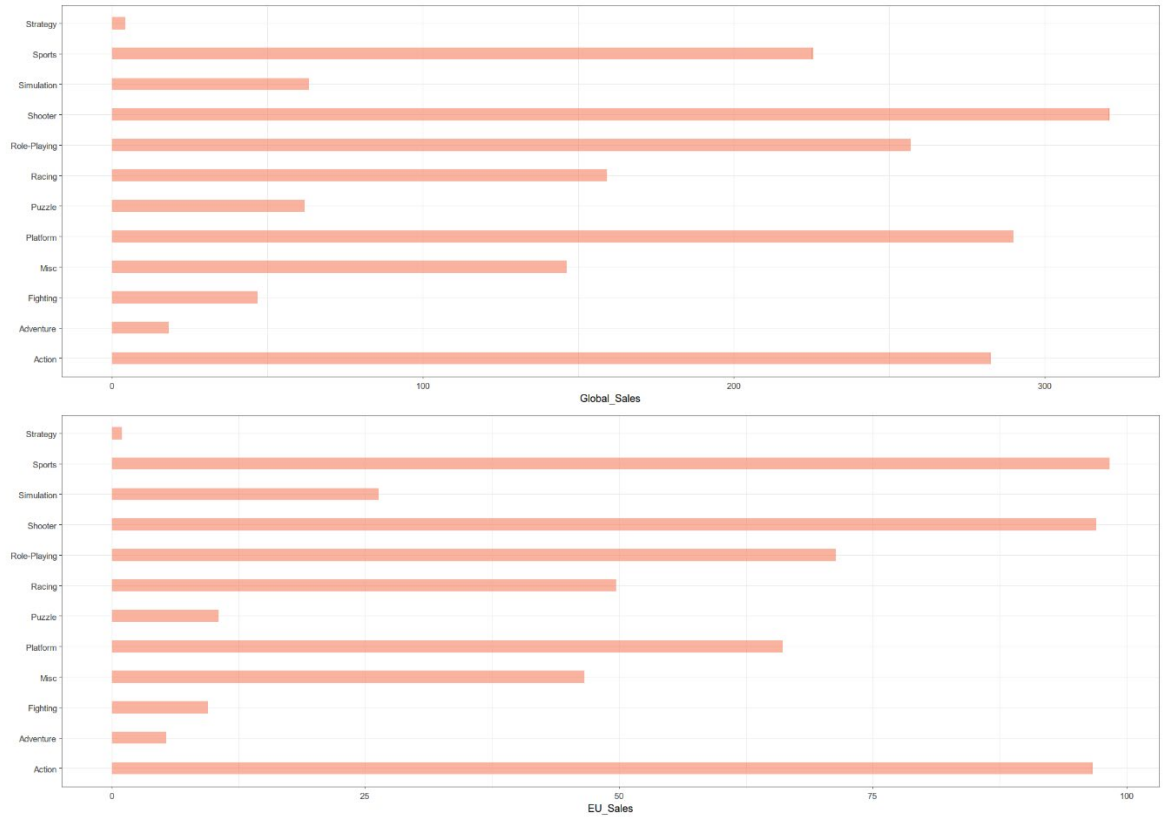


We have examined our data to see the clusters for renumeration and spending score. Our dataset contains 2000 records and used the Calinski Harabasz Score Elbow to decide how many clusters best fit our data. We concluded the optimal elbow was at 5 with a score of 2978.958. This can also be seen in our silhouette plot of Kmeans clustering with 5 centres. Our first cluster has the maximum data points of 774 values with cluster 5 having the least amount of values at 269.

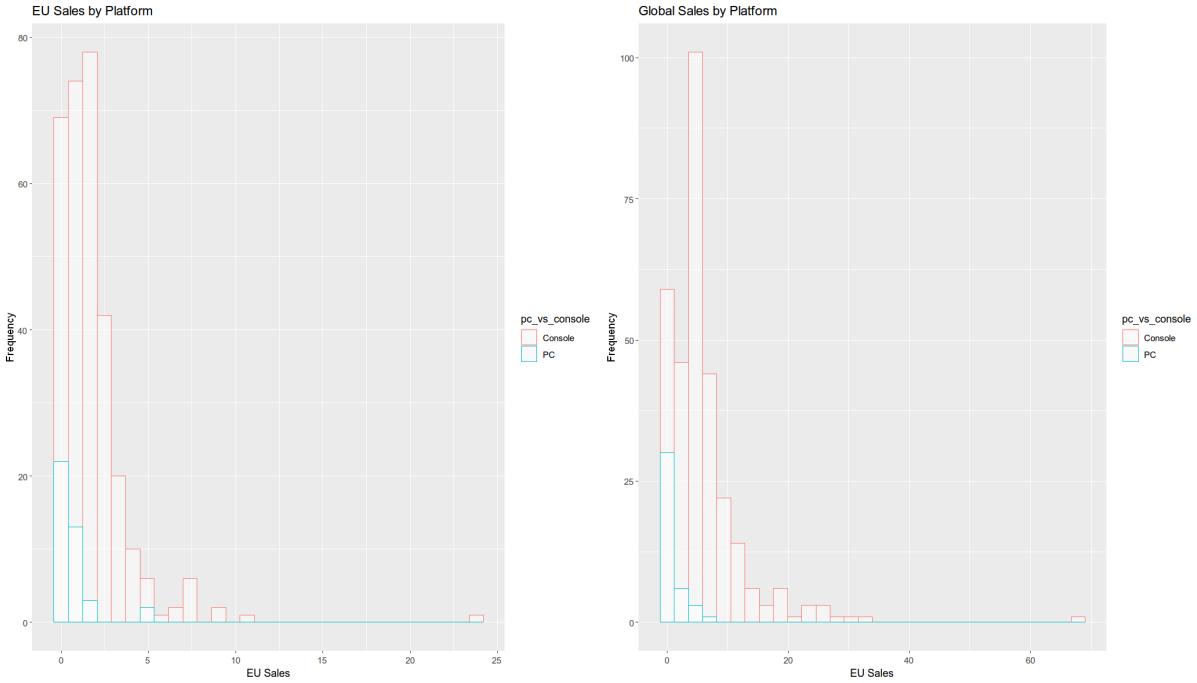


Upon visual inspection we can clearly see that cluster 3 and 4 are the most interesting to us as they have the highest spending score. We can also see that cluster 2 and 3 have the highest remuneration with cluster 1 and 4 having the lowest remuneration. It would be good to target clusters 3 and 4 with marketing campaigns to increase their loyalty points. We should run a separate campaign to increase the spending score for cluster 2 as they have high income and could lead to higher turnover if we can convince them to spend more.

SALES DATA ALAYSIS

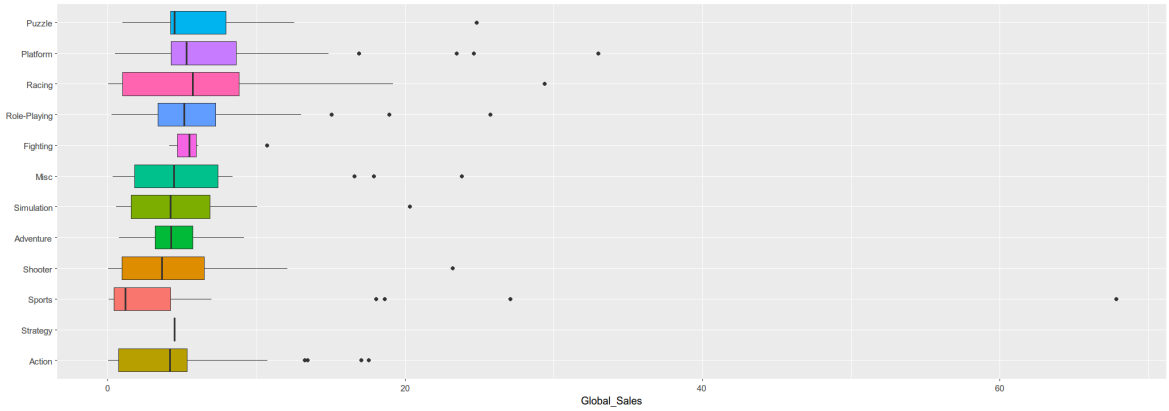


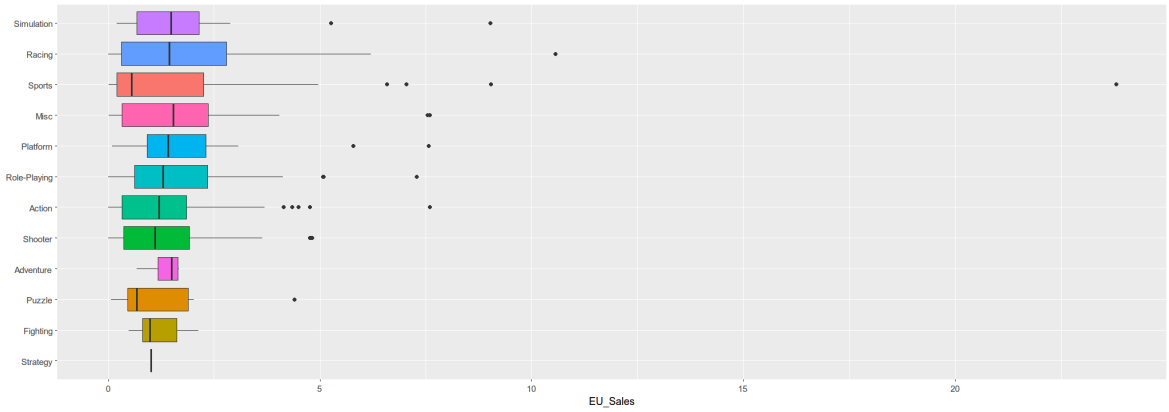
Our sales data contains 352 games, 10 platforms and 12 genres. The release date from games ranges from 1980 to 2016. Global mean sales are 5.335 million, while the EU mean sales are 1.644 million. Globally Sports, shooter and action games are the most popular. In Europe Shooters are the most popular, followed by Platform and Action. The least popular games are Strategy and Adventure. There are some outliers in sports sales with very high sales numbers that have very high sales numbers. The median Racing game sells better than any other game genre.

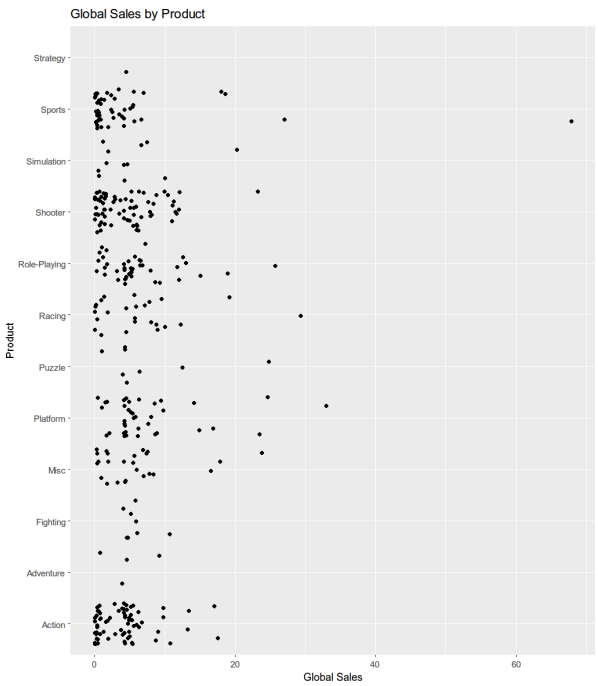


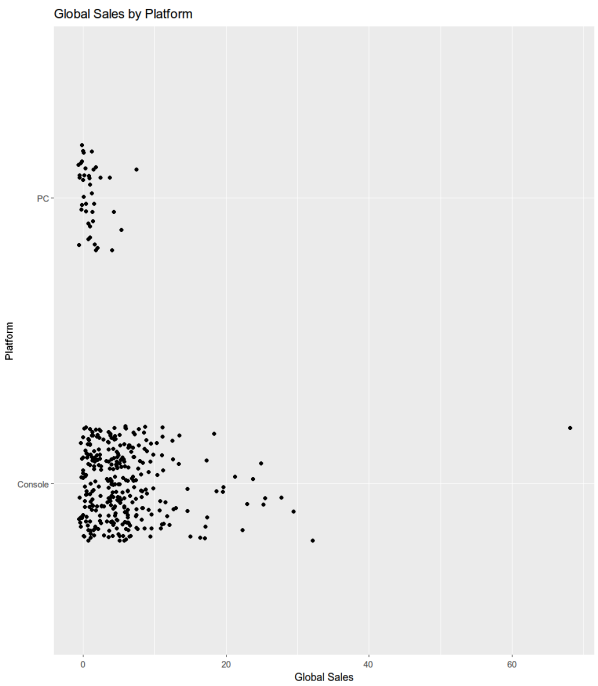
We can see that the data in all our sales figures is not normally distributed. With our Shapiro-Wilk showing p values of 2.2e-16 on all our sales data. We can also see that the data is leptokurtic with a very high positive skewness. This means shows that our data is not normally distributed and is very skewed. There is a game from Nintendo that has outsold all other games in our dataset. The sales of this game were 67.85 million. Nintendo also has the highest sales for any publisher in our list. Wii is the most popular platform followed by Xbox 360.

When we look at the correlation between the sales data we can see that there is a positive correlation between all the sales data. This means that as one sales figure increases the other sales figures will also increase. The highest correlation is between North American and Global Sales with a correlation of 93%.


Puzzle
Platform
Racing
Role-Playing
Fighting
Misc
Simulation
Adventure
Shooter
Sports
Strategy
Action
0
20
40
60
Global_Sales

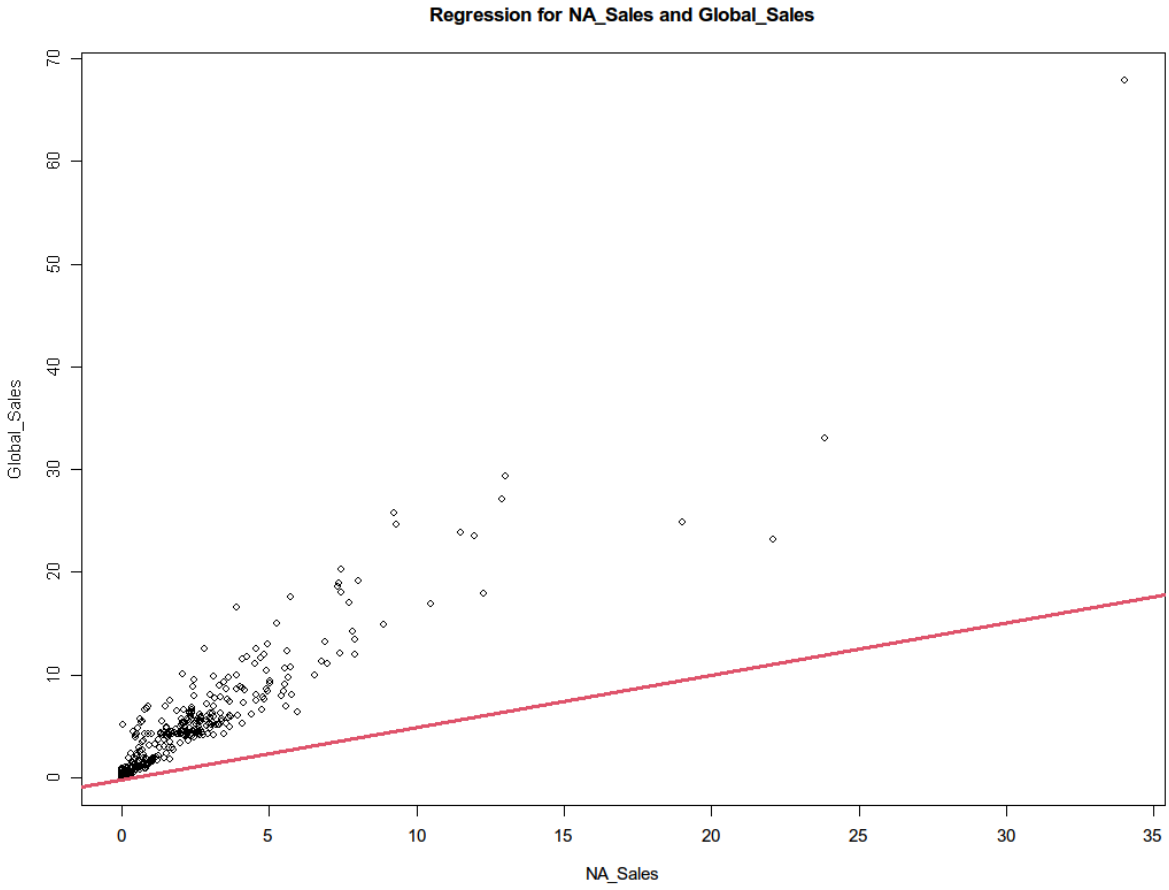

Simulation
Racing
Sports
Misc
Platform
Role-Playing
Action
Shooter
Adventure
Puzzle
Fighting
Strategy
0
5
10
15
20
EU_Sales


Global Sales by Product
Strategy
Sports
Simulation
Shooter
Role-Playing
Racing
Puzzle
Platform
Misc
Fighting
Adventure
Action
Product
0
20
40
60
Global Sales

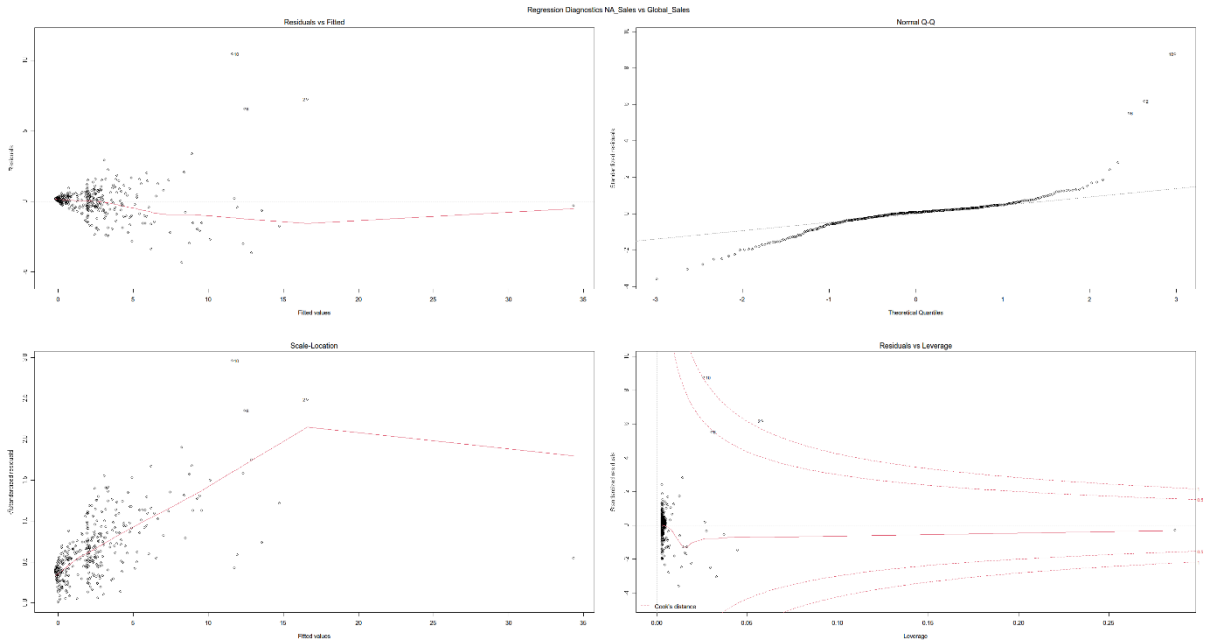

Global Sales by Platform
PC
Console
Platform
0
20
40
60
Global Sales

BUILDING MODELS FOR PREDICTION

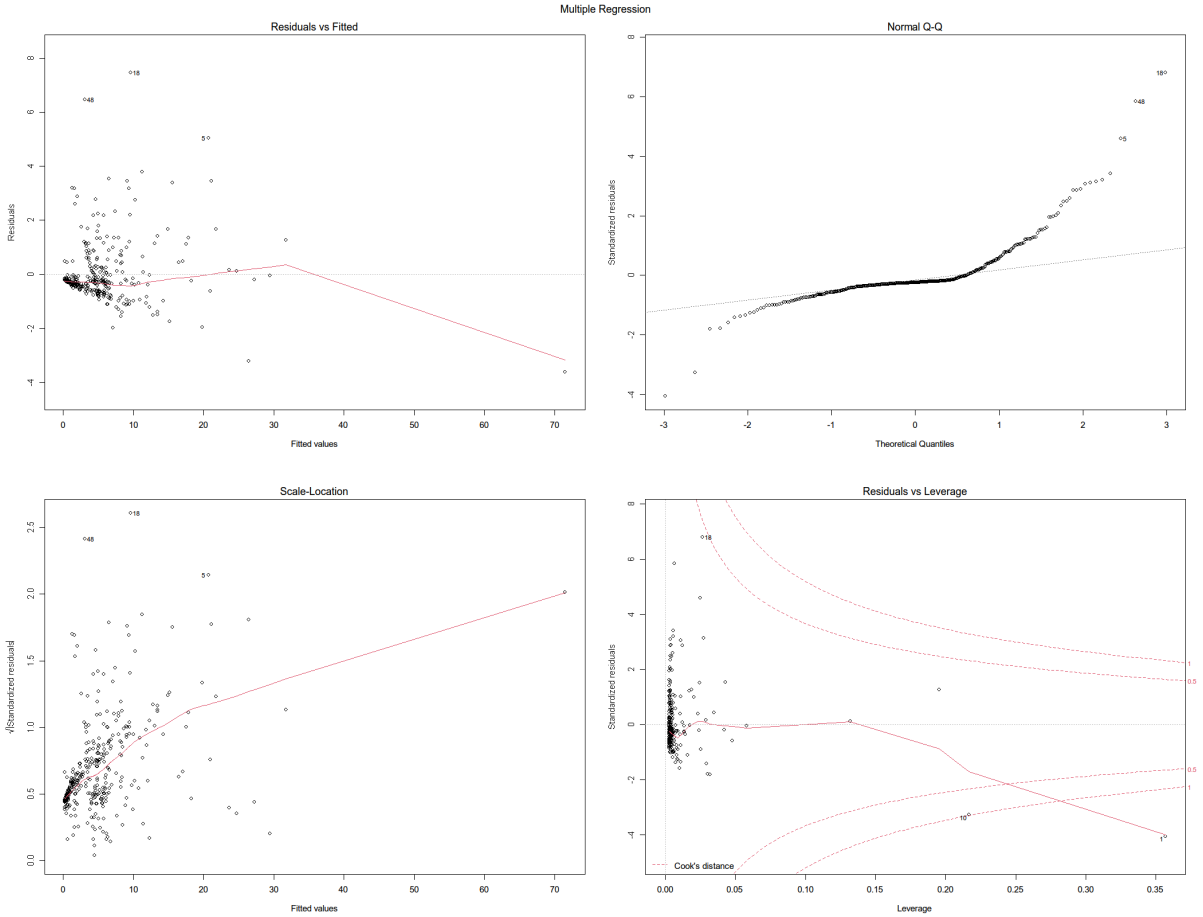
Looking at our Pearson's correlation we can see that the highest correlation in our data is between North America Sales and Global Sales with a index of 93.49% followed by EU and Global Sales of 87.75% with the least being EU and North American Sales of 70.55%.



We started with a Linear Regression model of North American and Global Sales. Although there is a very high positive correlation we can see some errors in our regression diagnostics.



Firstly, our Residual vs Fitted shows a megaphone shape which would indicate variances increase as the values go up. This would mean larger values are associated with large errors or residuals. In addition, our Normal Q-Q plot shows minimal errors between the -1 and 1 quantiles. The errors increase considerably when we are not within this zone.



Using multi linear regression we can see that product number has a negative correlation and our final model was made without product as a correlation vector. The regular vs fitted values changes dramatically for very high value showing that linearity is not met as we look at high sales figures.