SHAQUIL BISWAS

DA301

Turtle Games

Analysis for improving overall sales performance by utilising customer trends

The analysis conducted for Turtle Games has been done in order to help reach business objectives by answering various questions which have been posed to help better understand the company and each of the components which attribute to its performance.

The main business objective Turtle Games wishes to satisfy is: '**Improving overall sales performance by utilising customer trends.'**

Smaller questions which Turtle Games wishes to address in solving the business objective include:

- How customers accumulate loyalty points
- How groups within the customer base can be used to target specific market segments
- How social data (e.g. customer reviews) can be used to inform marketing campaigns
- The impact that each product has on sales
- How reliable the data is (e.g. normal distribution, skewness, or kurtosis)
- What the relationship(s) is/are (if any) between North American, European, and global sales?

One of the main means of answering the questions includes use of analytical models such as predictive analytics which is a means of quantifying uncertainty by taking assumptions (what we know to be true) and using this data to forecast and predict future outcomes.

SHAQUIL BISWAS

DA301

Linear regression

In order to carry out much of the investigation Python was used to gain a better understanding of the customer base to decipher and place quantifiable metrics behind much of the social data at hand.

The data was initially cleaned, reimported and scrutinised to ensure visualisations were complete. Various packages and libraries were harnessed for running analysis such as regression, sentiment, clustering and visualisations. These packages included Sci-Kit Learn, Sci Py, Matplot Lib and Seaborn.

```
# Any missing values?
rnull = reviews.isnull().sum()
rna = reviews.isna().sum()

print('Null Check: \n')
print(rnull,'\n')
print('NA Check: \n')
print(rna)
```

After the data was imported from the CSV file into the Jupyter Notebook the data was cleaned by ensuring there were no null/na values.

```
Null Check:

gender                    0
age                       0
remuneration (k£)         0
spending_score (1-100)    0
loyalty_points            0
education                 0
language                  0
platform                  0
product                   0
review                    0
summary                   0
dtype: int64
```

```
NA Check:

gender                    0
age                       0
remuneration (k£)         0
spending_score (1-100)    0
loyalty_points            0
education                 0
language                  0
platform                  0
product                   0
review                    0
summary                   0
dtype: int64
```

The Data was then explored by checking headings, structure and descriptive statistics.

```
# Explore the data.
print('columns,rows:\n',reviews.shape)
```

```
columns,rows:
 (2000, 11)
```

```
# Descriptive statistics.
reviews.describe()
```

Further cleaning of the data included dropping unnecessary columns and renaming them, once the data was fully cleaned the data was exported as a CSV and reimported for the analysis to begin.

SHAQUIL BISWAS

DA301

In order to answer many of the questions simple linear regression analysis was conducted between dependent and independent variables. Simple tests for homoscedasticity were conducted to see whether there was linearity, upon determining this OLS regression analysis was conducted between

- spending vs loyalty
- renumeration vs loyalty
- age vs loyalty

For each variable the independent and dependent variable were identified and as such the analysis to check for the strength of the linearity (if it was present) and the measure of the explicable proportion of the total variation in the dependant variable accounted and allocated against the variation in the independent variables. At this stage the F Stat, correlation coefficients and standard errors were also observed.

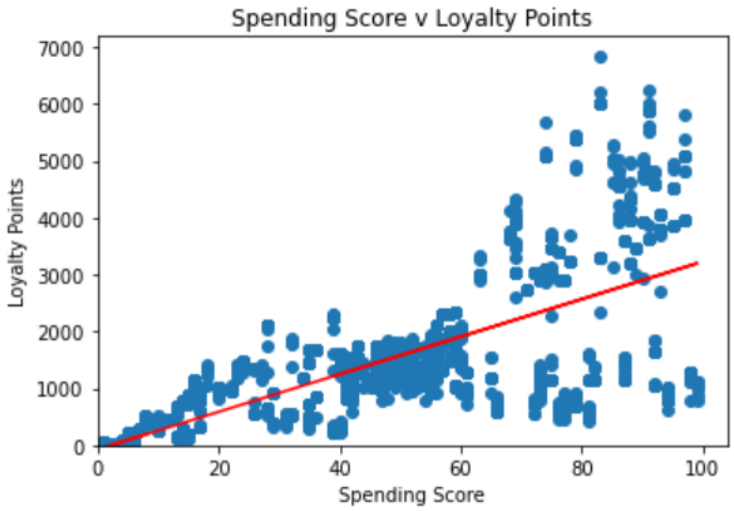
spending vs loyalty

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Sun, 15 Jan 2023	**Prob (F-statistic):**	2.92e-263
Time:	08:59:36	**Log-Likelihood:**	-16550.
No. Observations:	2000	**AIC:**	3.310e+04
Df Residuals:	1998	**BIC:**	3.312e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-75.0527	45.931	-1.634	0.102	-165.129	15.024
x	33.0617	0.814	40.595	0.000	31.464	34.659

Omnibus:	126.554	**Durbin-Watson:**	1.191
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	260.528
Skew:	0.422	**Prob(JB):**	2.67e-57
Kurtosis:	4.554	**Cond. No.**	122.



The R squared is a measure between independent variable (spending) and dependent variable (loyalty). The spending vs loyalty showed had the strongest R squared value and indicated a relationship was present between variables with R squared at 0.45, the R squared indicates the proportion of total variation in the dependent variable which can be explained or allocated by total variation in the independent variable.

As we understand the closer the R squared to 1 the better the fit, we still must recognise that over half of the variations in the dependent variables are inexplicable.

SHAQUIL BISWAS

DA301

The P value (F Stat) for spending v loyalty was very low at less than 0 indicating the probability of observed difference happening by chance is low it is also less than alpha 0.05 indicating P is significant and the relation between variables are present.

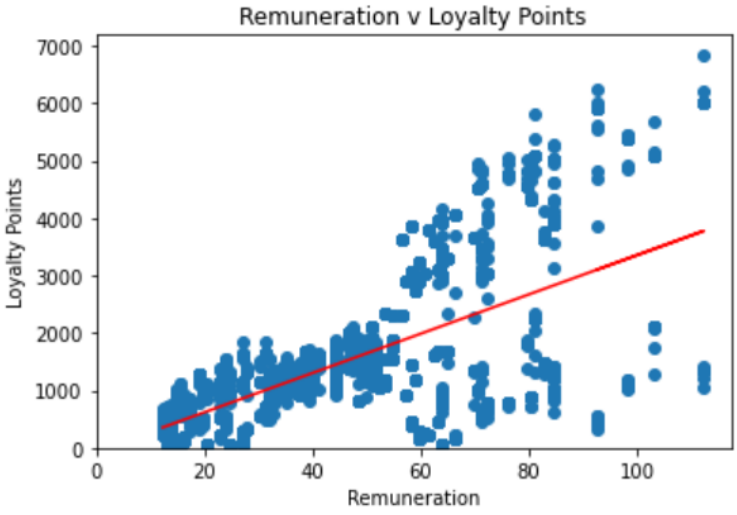
renumeration vs loyalty

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.380
Model:	OLS	**Adj. R-squared:**	0.379
Method:	Least Squares	**F-statistic:**	1222.
Date:	Sun, 15 Jan 2023	**Prob (F-statistic):**	2.43e-209
Time:	08:59:37	**Log-Likelihood:**	-16674.
No. Observations:	2000	**AIC:**	3.335e+04
Df Residuals:	1998	**BIC:**	3.336e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-65.6865	52.171	-1.259	0.208	-168.001	36.628
x	34.1878	0.978	34.960	0.000	32.270	36.106

Omnibus:	21.285	**Durbin-Watson:**	3.622
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	31.715
Skew:	0.089	**Prob(JB):**	1.30e-07
Kurtosis:	3.590	**Cond. No.**	123.



The strength of relationship between renumeration and loyalty is less strong compared to spending and loyalty. This therefore means even less of the variances in the dependent variable can be explained by the independent variable. It can also be seen from the graph that the homoscedasticity of the data is not as present as the data points seem to cone away from each other towards the latter stages. That being said the P value is still significant.

SHAQUIL BISWAS

DA301

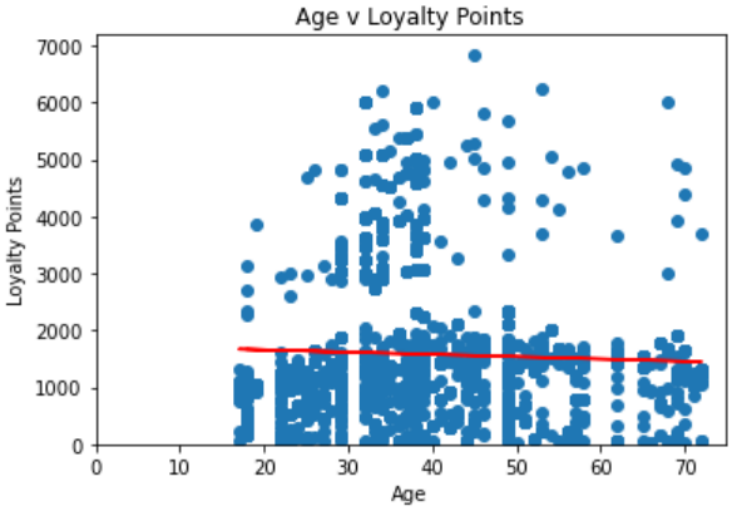
age vs loyalty

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.606
Date:	Sun, 15 Jan 2023	**Prob (F-statistic):**	0.0577
Time:	08:59:38	**Log-Likelihood:**	-17150.
No. Observations:	2000	**AIC:**	3.430e+04
Df Residuals:	1998	**BIC:**	3.431e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1736.5177	88.249	19.678	0.000	1563.449	1909.587
x	-4.0128	2.113	-1.899	0.058	-8.157	0.131

Omnibus:	481.477	**Durbin-Watson:**	2.277
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	937.734
Skew:	1.449	**Prob(JB):**	2.36e-204
Kurtosis:	4.688	**Cond. No.**	129.



The age vs loyalty regression analysis indicates practically 0 relation between the two variables, which means the model has no significance. Furthermore the P value is above 5 which is above alpha level and therefore not significant. Finally the graph shows no homoscedasticity and we can safely disregard the relation between these two variables.

Using the regression analysis data we have identified '**How customers accumulate loyalty points**'. This is done by

1. Spending
2. Renumeration.

That being said much of the relation is still no explainable and as such we must consider that we do not have conclusive knowledge in how customers truly accumulate points.

SHAQUIL BISWAS

DA301

Clustering with k-means

In order to understand the customer groups better clusters need to be determined to understand how different groups of people interact and behave, these would be directly answering some of the questions posed by Turtle Games such as how those in a customer base can be used to target specific market segments.

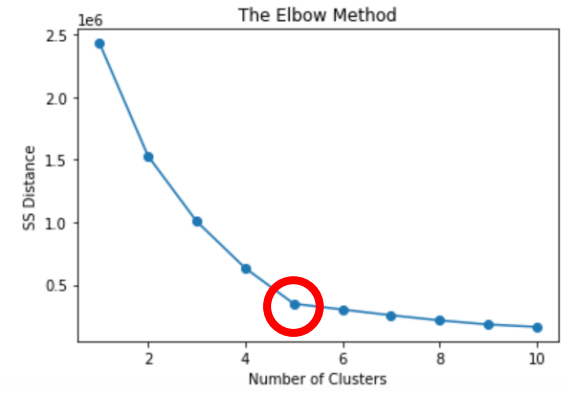
3. Elbow and silhoutte methods

```
# Determine the number of clusters: Elbow method.
ss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i,
                    init = 'k-means++',
                    max_iter = 500,
                    n_init = 10,
                    random_state = 0)
    kmeans.fit(x)
    ss.append(kmeans.inertia_)

# Plot the elbow method.
plt.plot(range(1, 11),
         ss,
         marker='o')

plt.title("The Elbow Method")
plt.xlabel("Number of Clusters")
plt.ylabel("SS Distance")

plt.show()
```



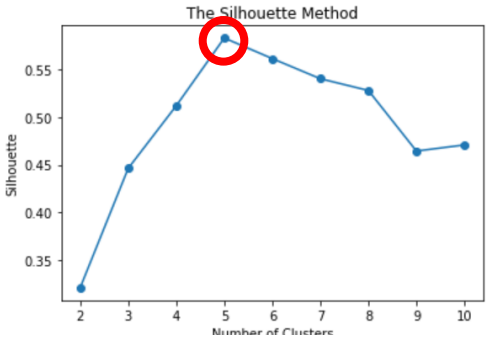

```
# Determine the number of clusters: Silhouette method.
sil = []
kmax = 10

for k in range(2, kmax+1):
    kmeans_s = KMeans(n_clusters = k).fit(x)
    labels = kmeans_s.labels_
    sil.append(silhouette_score(x,
                                labels,
                                metric = 'euclidean'))

# Plot the silhouette method.
plt.plot(range(2, kmax+1),
         sil,
         marker='o')

plt.title("The Silhouette Method")
plt.xlabel("Number of Clusters")
plt.ylabel("Silhouette")

plt.show()
```

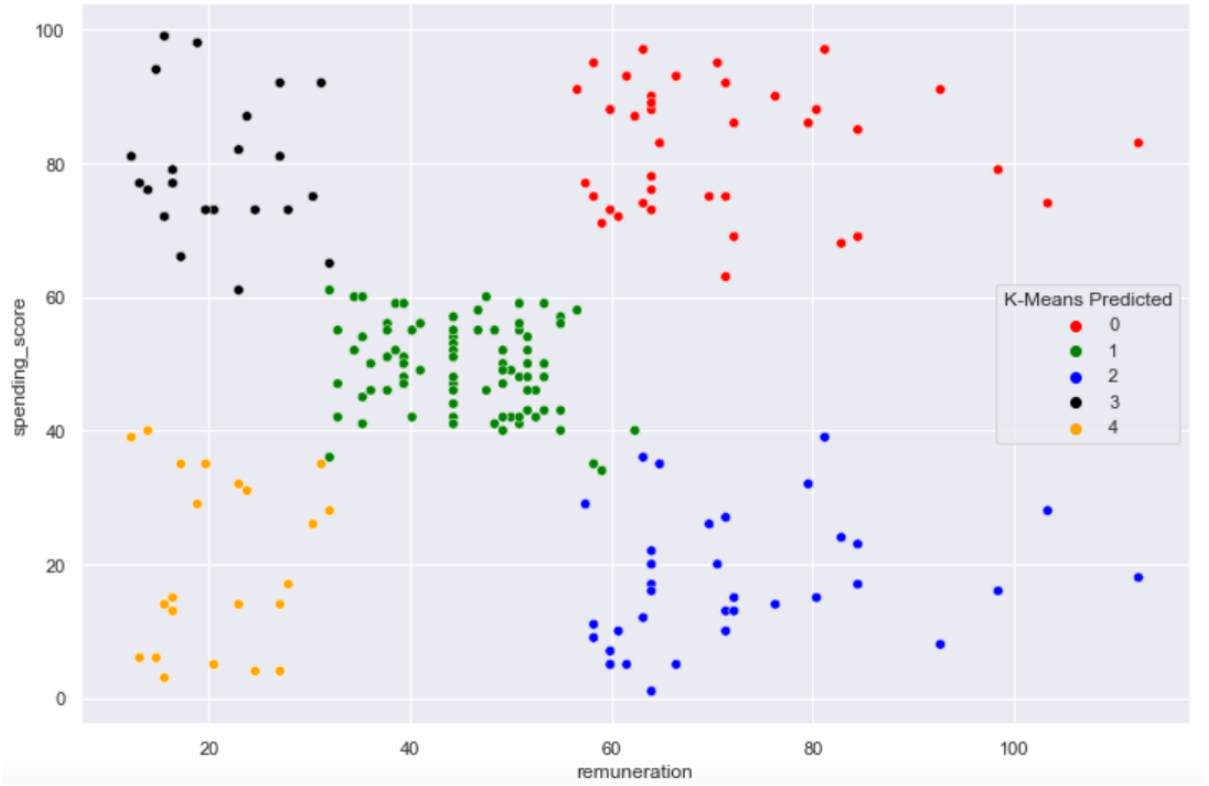


Using the elbow and silhouette methods as shown above we can identify the ideal number of clusters for use in k-means clustering, this can then be applied to a model where cluster groups can then be attributed. E.g. cluster 1 may have attributes of high spend and low renumeration, we may then identify those in this cluster are more likely to leave positive/ negative reviews.

As shown in the Elbow method above the line plateaus after the 5th cluster, this indicates that 5 clusters is the optimum number. Similarly the silhouette method shows that 5 clusters is ideal as the 5th cluster has the greatest peak.

SHAQUIL BISWAS

DA301



- The cluster model as shown above indicates that the main cluster group is cluster 1 which has strongest renumeration density at around 45 and spending score between 40 and 60.
- Interestingly cluster 0 has very high spending score and a very high renumeration which could be an indicator that this particular group may be a benefical target of marketing campaigns.

Using this cluster data we have answered the question '**How groups within the customer base can be used to target specific market segments**'. We can recognise that;

1. There are evidently different groups within the customer base
2. Each group has different attributes which are shared
3. These groups can be targeted specifically based on these attributes

SHAQUIL BISWAS

DA301

Natural Language Processing

The data was then reimported and manipulated in a way so that there were two columns for the customer reviews. The text in the reviews had been changed to lowercase, punctuation removed and duplicates removed.

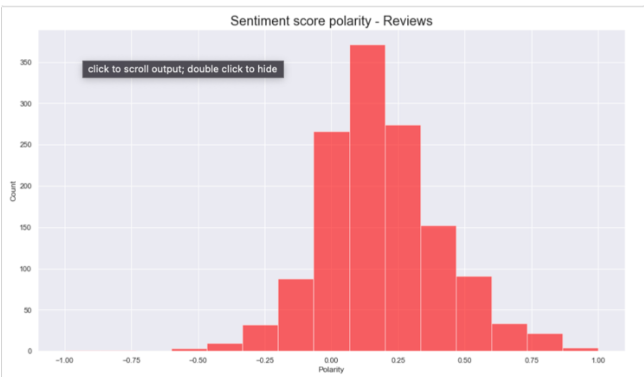
```
# Review: Change all to lower case and join with a space.
df3['review'] = df3['review'].apply(lambda x: " ".join(x.lower() for x in x.split()))
```

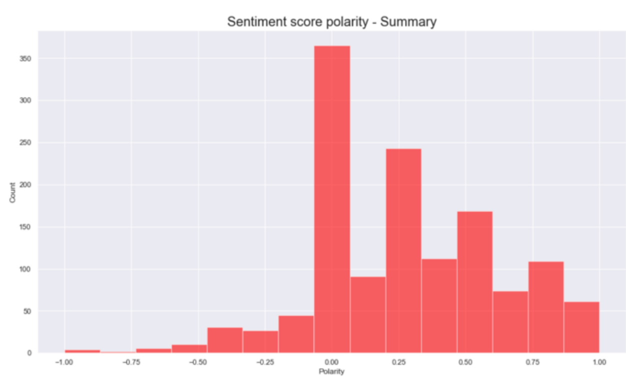
```
# Summary: Change all to lower case and join with a space.
df3['summary'] = df3['summary'].apply(lambda x: " ".join(x.lower() for x in x.split()))
```

The words were then tokenised and had alphanumeric and stopwords removed. The result of all of this was transformed data which could be passed through a wordcloud generator to highlight the most frequently used words.

As shown in the images above (left review, right summary) the most frequent words do vary between each dataframe. The overall sentiment shows a positive one with words such as fun, great and love appearing.

A test for polarity was conducted to show the most extreme sentiments on both the positive and negative side as determined by the sentiment analyser package.





SHAQUIL BISWAS

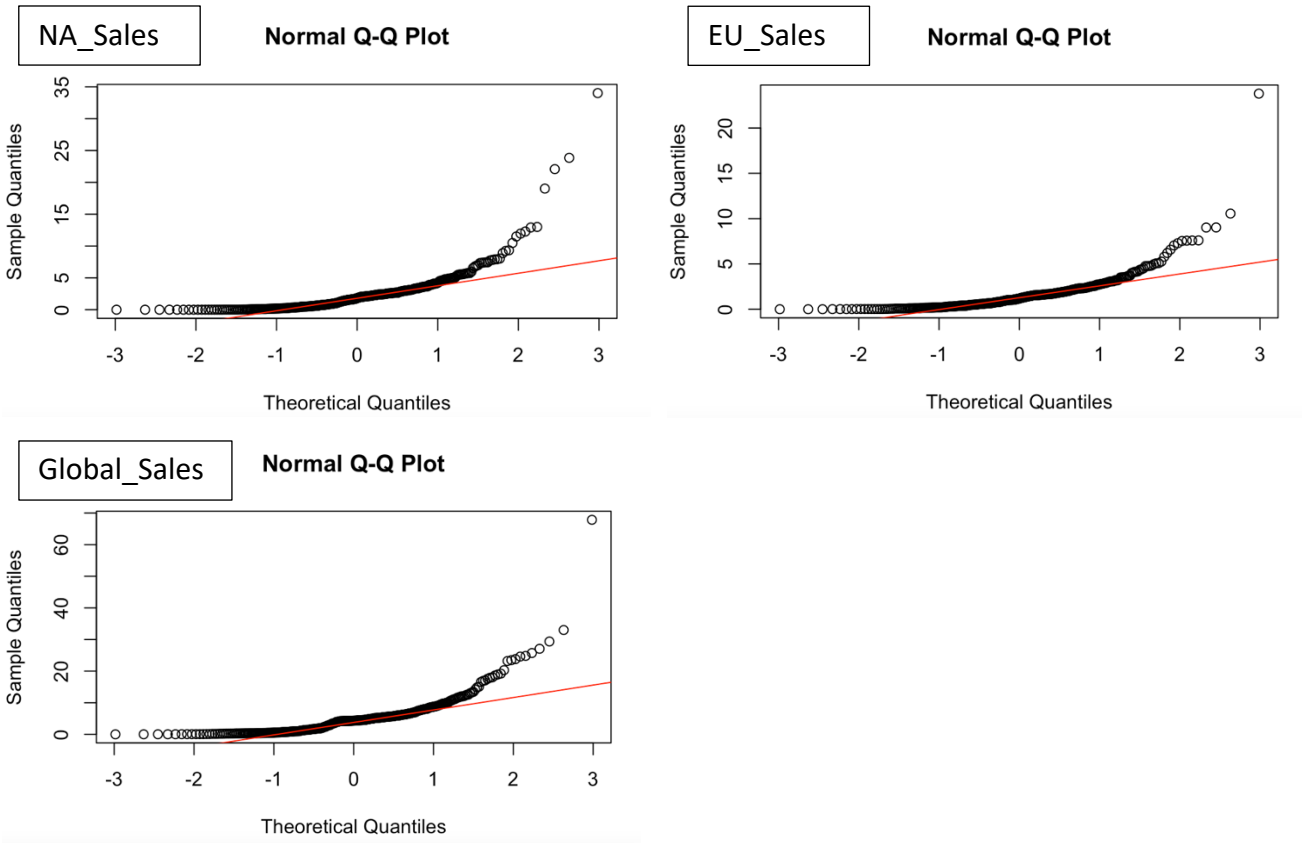
DA301

Although there are negative sentiments the overarching positive reviews considerably outweigh the negative and as such the volume of the positive sentiments as extracted by the sentiment analysis is evidently present throughout wordclouds. It is also evident as per the histograms above which show the skew towards the positive polarities which are the positive words.

The use of Natural Language Processing has allowed for us to see '**How social data (e.g. customer reviews) can be used to inform marketing campaigns**'. We can identify that certain elements of the products can evoke a stronger positive response. By recognising which products have the most affinity to positive reviews we can push those products more with marketing campaigns.

Further to this, positive reviews themselves can act as marketing campaigns as they are viewed by other customers who can recognise the positive sentiments to the products.

EDA using R



Through use of Q-Q plots, skewness and kurtosis, and Shapiro-Wilk testing there is an evident right skew. NA sales are slightly higher than global and EU sales. The Shapiro-Wilk tests provided a result which were all less than the alpha value 0.05 which indicates significance.

SHAQUIL BISWAS

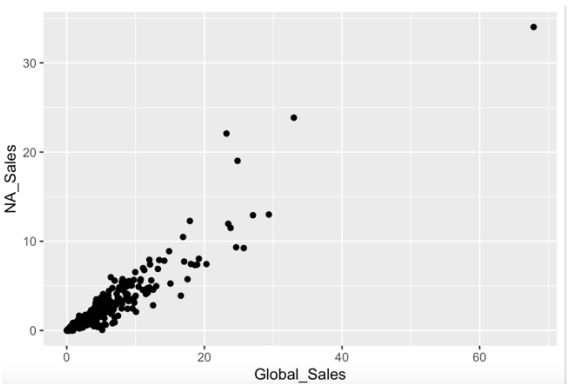
DA301

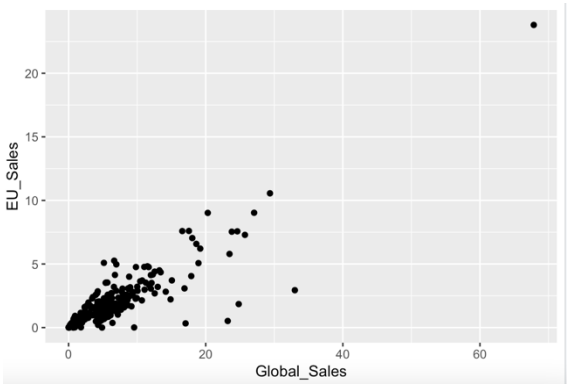
Here we can see some of how '**each product has and its impact on sales'.**

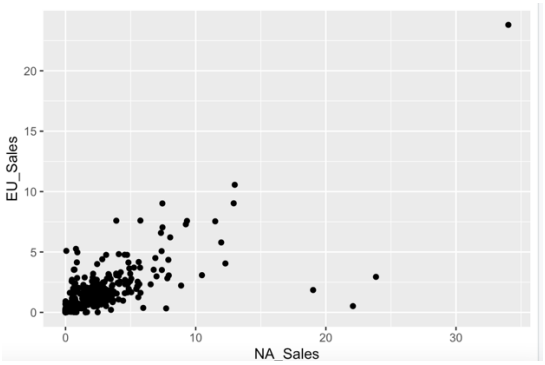
	Platform	NA_Sales	EU_Sales	Global_Sales
1	X360	153.39	76.01	253.81
2	Wii	149.52	104.99	312.56
3	PS3	77.78	88.52	211.61
4	DS	72.56	65.60	205.02
5	GB	68.66	28.18	133.97

The Xbox 360 was the most popular individual platform in NA although the Wii was the most popular globally.

The PlayStation as a complete gaming platform including PS1, PS2, PS3, PS4 and PS5 successful brand in terms of sales, followed by Microsofts Xbox







The strongest correlation is evident between NA sales and Global Sales, the least strong correlation is evident between NA and EU Sales.

SHAQUIL BISWAS

DA301

Final Comments

There may be a link between loyalty points and spending score, if this information is coupled with the cluster data we can identify particular groups of people that are spending more and possibly alter the loyalty points scheme as incentives.

The clusters show that the customer are segmented into groups, these individuals can have targeted campaigns, for example those spending less can possibly be targeted with cheaper products.

The cluster data can also be linked to the reviews they are leaving, certain groups may be more inclined to leave reviews and given on the whole the reviews are positive it may be beneficial to look at extrapolating this so there is a push for more reviews.

It would be interesting to look at other social media data such as Facebook, Twitter and Instagram to see an idea of a less contained environment where people are sharing their views.