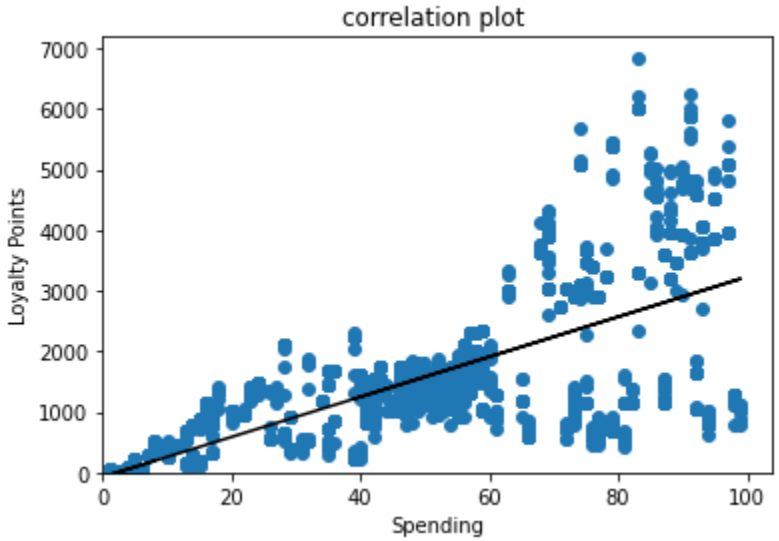
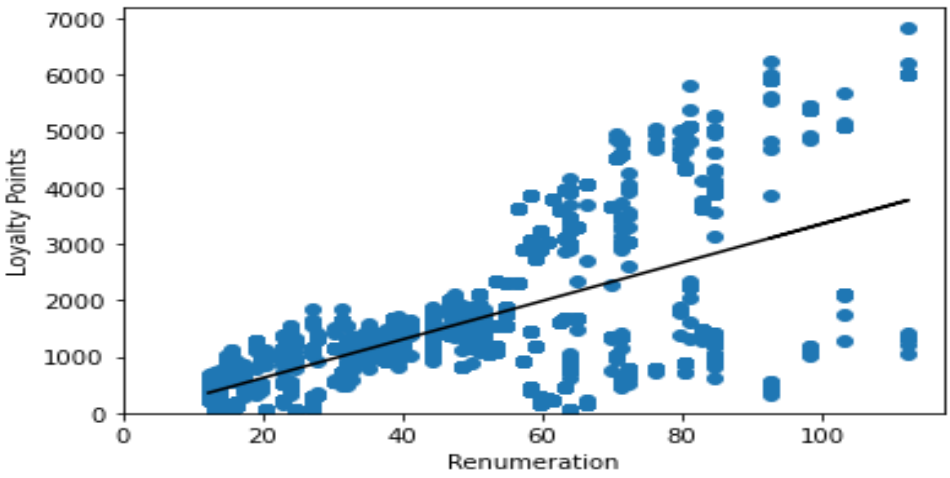
Making predictions with regression

We perform Linear regression.
First we implement OLS regression.
The first case is where the independant variable is the spending score
and the dependant variable is the loyalty points. The scatter plot implies
a potential positive linear relationship.The intercept value is -75.053
and the X coefficient value is 33.062. We create an object called y_pred
which is a linear combination of the intercept and X coefficient to make predictions.
Looking at the line of best fit it indicates a positive correlation albeit with some outliers.

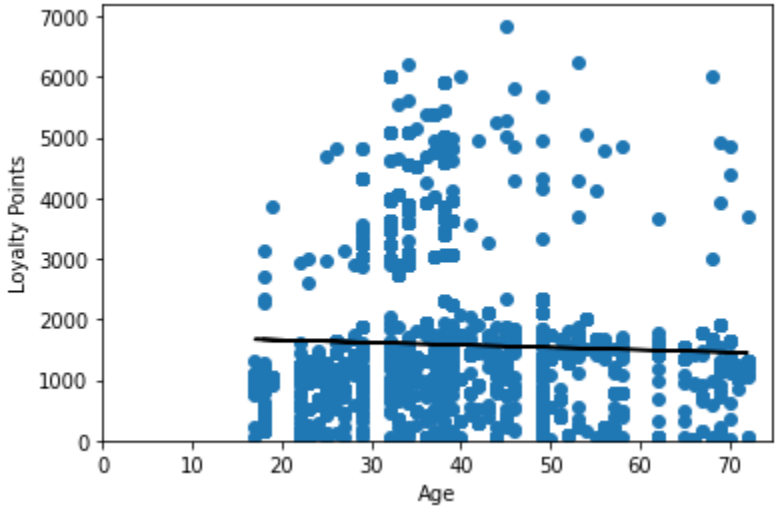
R-Squared value is only 0.452 showing the spending score explains 45.2 percent
Of the variance in the loyalty points. This is not terrible considering the simplicity of the model used.



We now look at remuneration as the independent variable and loyalty points as
The dependant variable. We run the ols test again to perform least squares regression.
The X coefficient value now is 34.19 and the Intercept value is -65.7.Looking at the plot with
A line of best fit added we see a clear linear positive trend/relationship between the variables.
The R-squared value is only 0.38 however which shows that only 38 percent of the variance in
The loyalty points explained by renumeration.This makes sense as this is a very simple model
based on only one parameter. Definite underfitting.



We now explore the relationship between age and loyalty using OLS regression.
The Intercept value is 1736.5177 and the X coefficient value (age) is -4.0128.
Looking at the plot there seems to be no linear relationship, however we have to consider that The scale of the loyalty points is much larger than age. In this case there seems to be no correlation regardless. The R-squared values is very loew at 0.01 showing age is a Terrible predictor for the loyalty points.



From the OLS regressions we see that the spending score is the best predictor of the Loyalty points and age is the worst.
Let us implement a more complex Multiple linear regression model.
We now use it to predict values.
To evaluate the accuracy of the predictions and thus the model we need to check the R-squared value. Here it is 0.83 which means the independent variables explain 83 percent of The variance in the spending score. The intercept is -1700.3 and the coefficients are

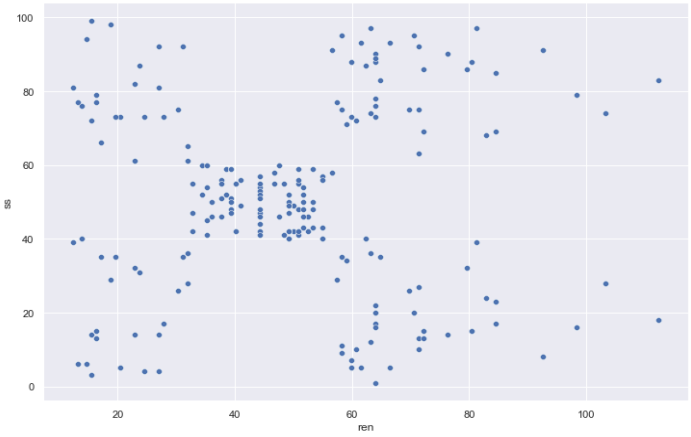
33.98 for remuneration and 32.9 for spending score.

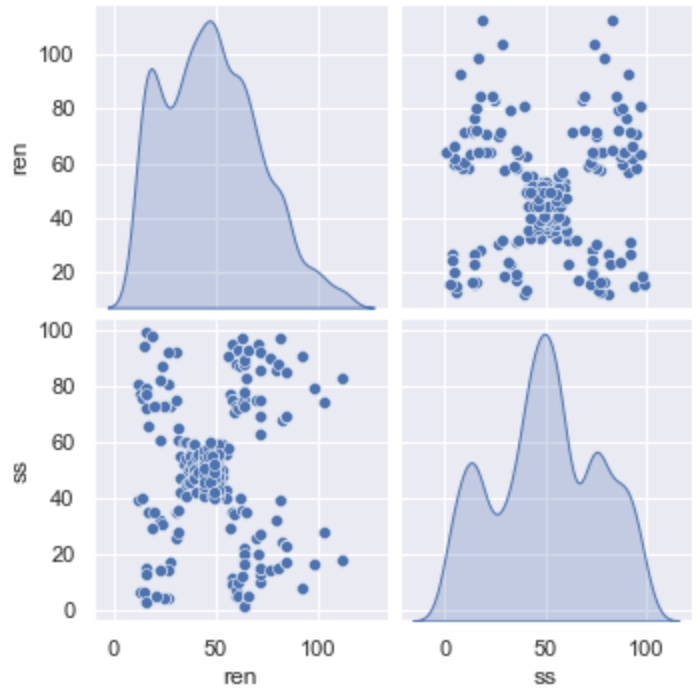
We also implement a different method by creating a training and test split.
The training set is used to train the model and the test data is used to evaluate the accuracy of
The trained model on new data, i.e how well it generalizes.
The R-squared score is 0.8151 on the test set. We check for multicollinearity
Using the variance inflation factor and conclude there is non-significant correlation
Between the independent variables. Our model is also has significant p-value less than
0.05 in the breuch pagan test for heteroscedasticity, thus we reject the null hypothesis
Of homoscedasticity and accept the alternative hypothesis of heteroscedascity.
The significant F-test value also indicates our model is a better fit for the data than a
Model with no independent variables.
Our Mean Absolute Error is 425.36. Which means on average each predicted value is
Within 425.36 of the true value.

Overall our model is decent given the amount of data provided and considering all the assumpotions of Multiple Linear Regression have been met.

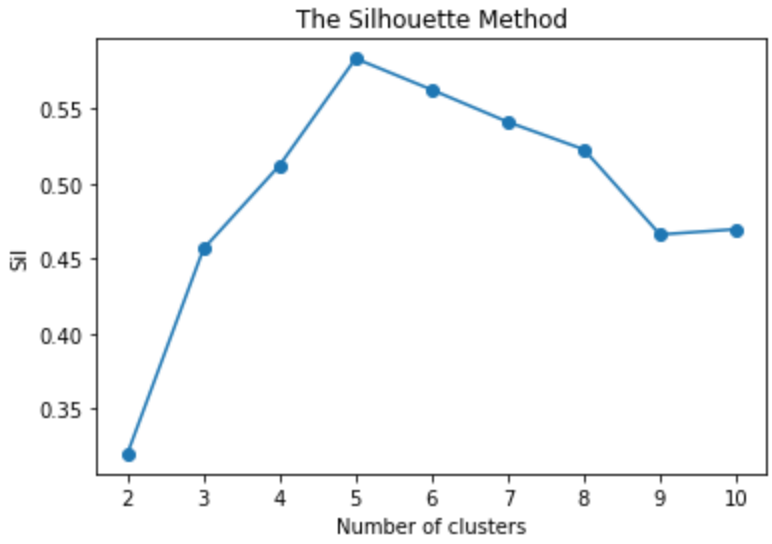
Making predictions with K-means clustering

Looking at the scatterplot of renumeration against spending score it seems
Like there are 5 distinct clusters, which hints at remuneration and spending score being
Good predictors for education level as it has 5 levels and the only other categorical variable is gender which would only have two levels. We need to explore further to confirm this.

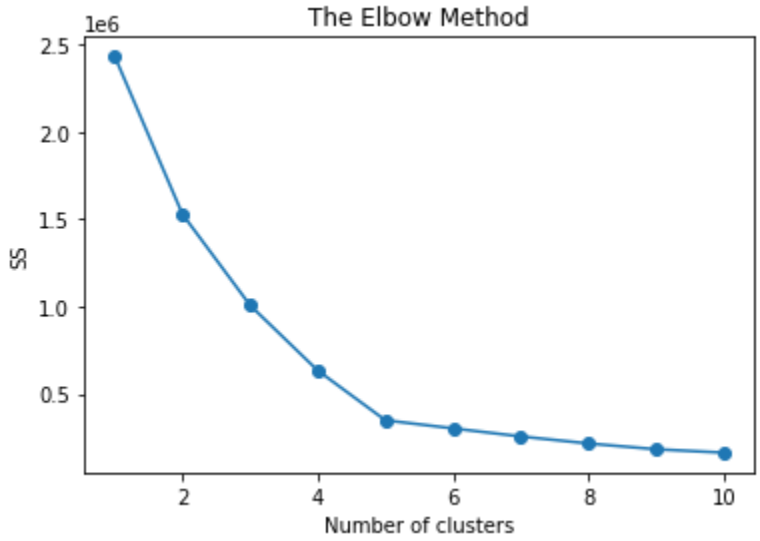




We have to determine the value of ‘k’ in the k-means clustering algorithm thus we implement both the elbow method and the silhouette method to find an optimum value of K.

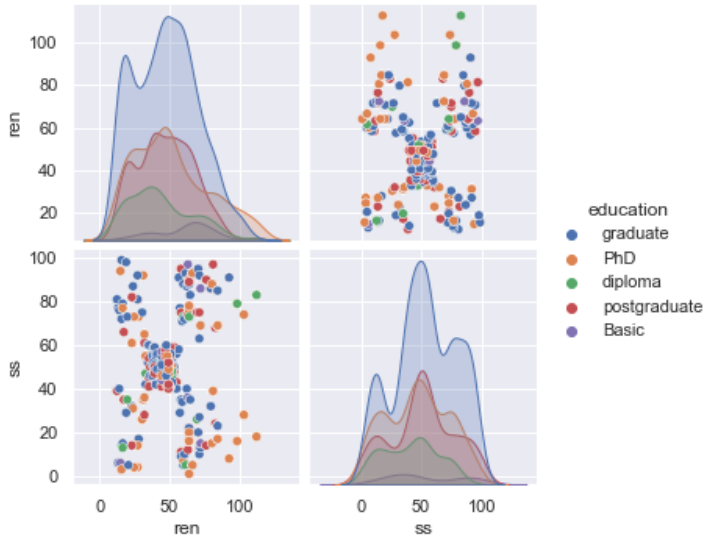


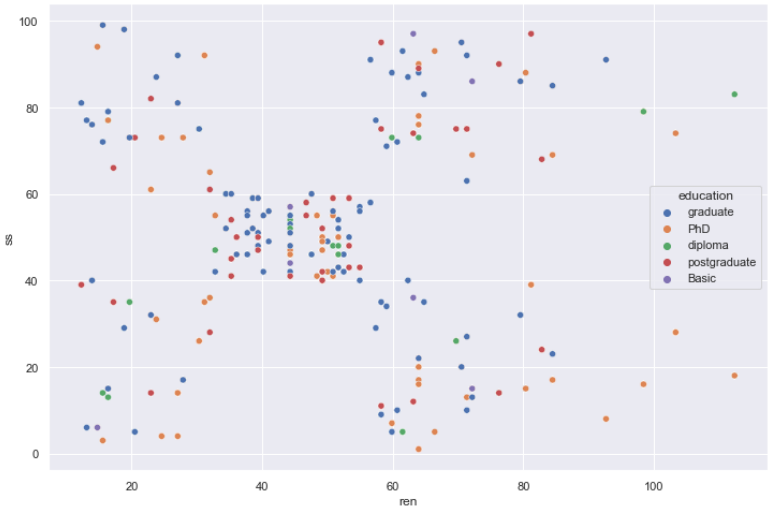
We



Using our knowledge of the silhouette and elbow method k=5 seems to be the ideal\
fit.

We plot the same pairplot and scatterplot but now with the color representing the different education levels:





We see that although there seems to be 5 distinct clusters the distribution plots indicate There is a lot of overlapping in the groups and thus they are more similar and harder to Distinguish from each other. We see some patterns such as PHD students tending to have slightly higher remunerations, graduate students have the largest frequency and thus should be the largest cluster. We also see a high density of points at a spending score of 50.
Let us run our K-mean algorithm and observe the results.

We set k=5 due to us already making a reasonable assumption of what categorical variable seems to be a viable prediction as well due to the conclusions from the elbow and silhouette methods.

Plot of our clustering output:

Overall, We can make some useful predictions using Renumeration and spending score regarding the education level of the customers. We can accurately predict if a customer is a graduate as it is clearly the largest group. We can predict a customer as being at diploma or basic level however we cannot accurately say which of the two. We can predict a customer at being at PHD and postgraduate level but once again cannot accurately say which of the two. This indicates that PHD and postgraduate students have similar patterns of remuneration and spending scores and basic and diploma level students also have similar patterns of remuneration and spending scores.

Analyse customer sentiments with reviews.

We use Natural language processing to analyze the sentiment of the customer reviews of the Website Turtle Games.

Below is the initial ‘Reviews’ WordCloud:

We see the word game is the most frequent as shown by its largest size. Play,great,fun and love are also frequent hinting at positive sentiment in the reviews .

Below is the WordCloud for the ‘Summary’ columns:

We see the words five and stars are quite frequent which could indicate a ‘5-star’ rating indicating positive sentiment. Other frequent terms are great,love,good And fun. With initial insights of these wordclouds we tend to see an overall positive sentiment.

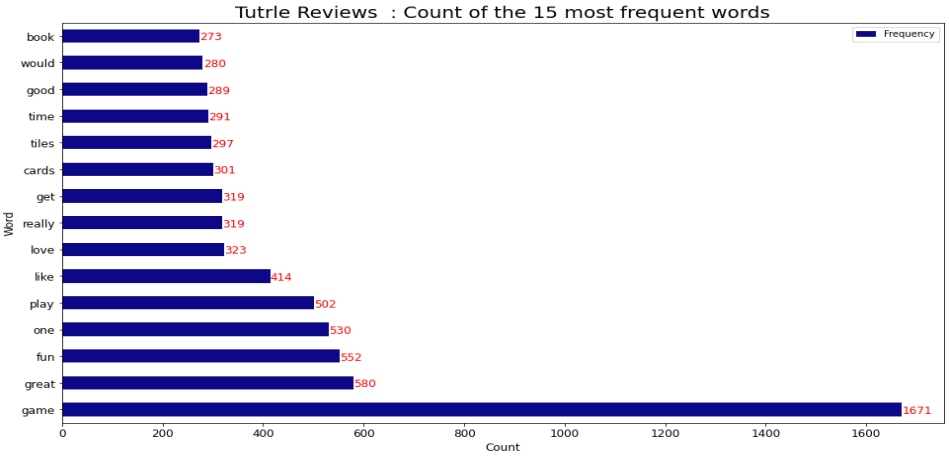
We now remove the stopwords and alphanumeric characters as these do not contribute To sentiment.

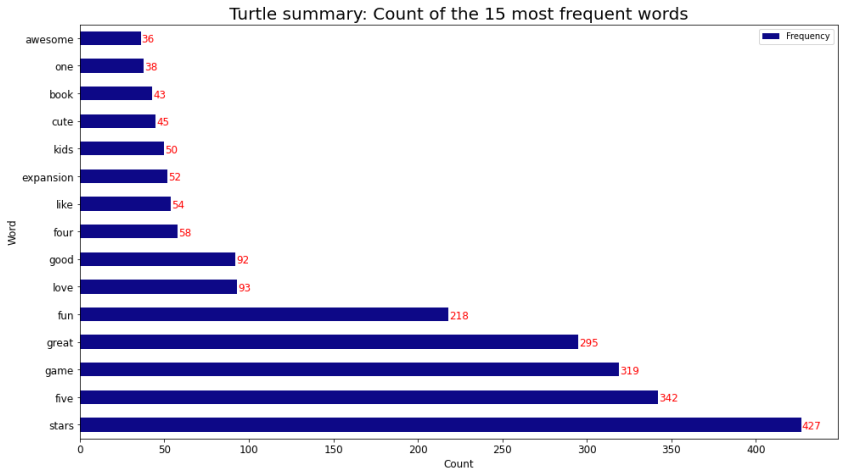
We now visualize the new wordclouds without the stopwords and alphanumeric characters:

Summary:

Review:

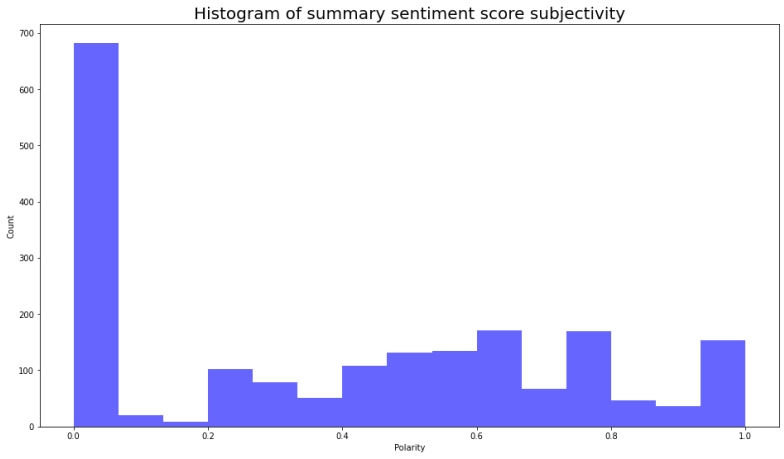
Top 15 words for both columns visualization:



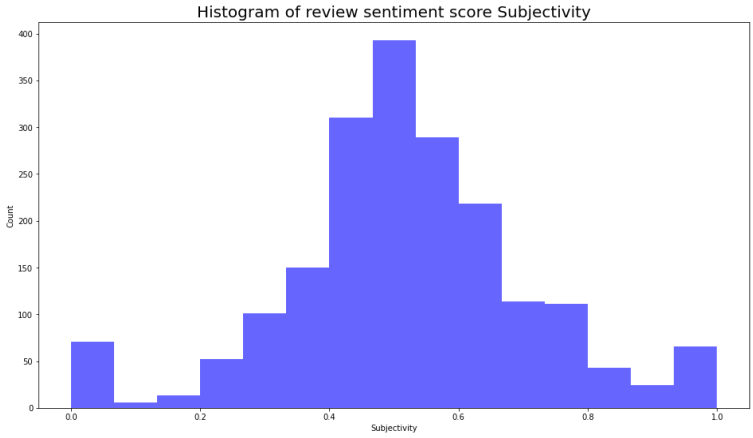


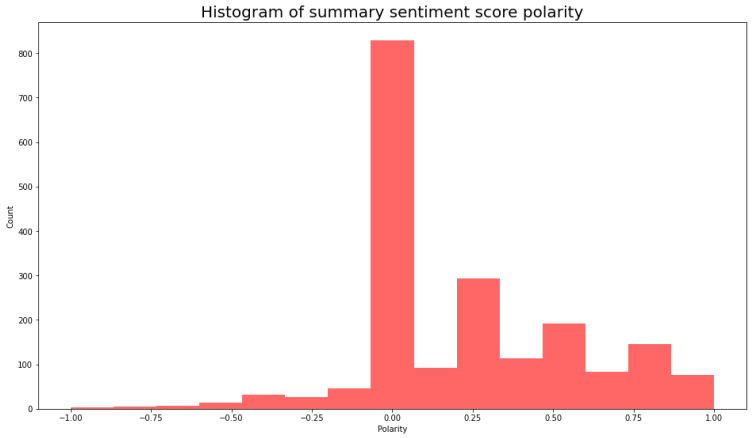
Finally we will review the polarity and sentiment of the texts.
The polarity score tries to determine a score for the text that determines if it is
Negative or positive. The higher the polarity score the more positve the text is.
The subjectivity score compares the personal opinion in the text
To the actual factual information in the text.The higher the subjectivity score the higher the amount of personal opinion in the text.

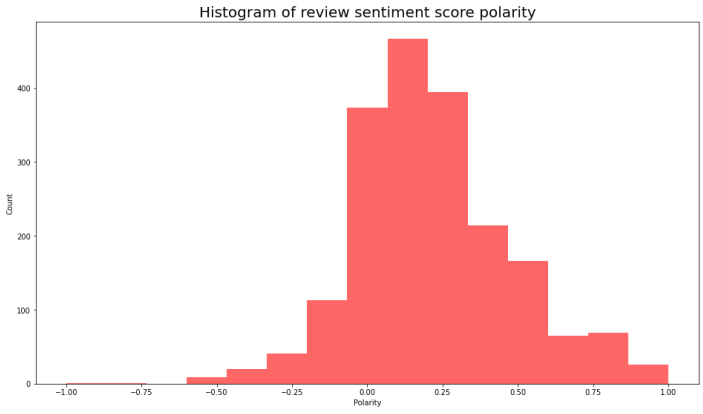
Histograms for the polarity/subjectivity of the two columns:



:







The example negative review polarity scores also correspond to the actual text.

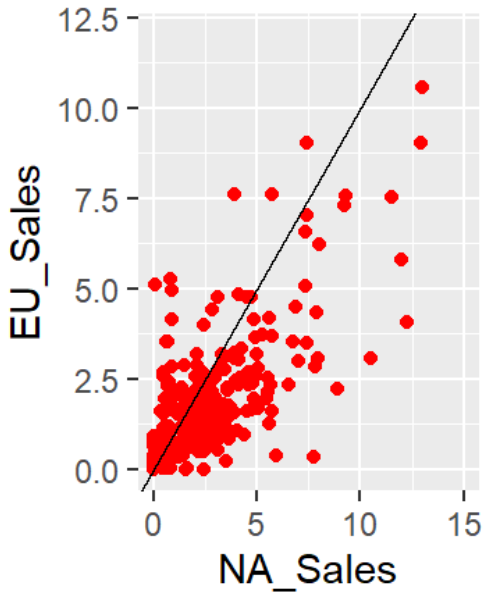
The positive polarity scores of the reviews also seems to correspond to the text.
Overall it seems that our model has done a good job in classifying text as positive,negative or neutral and it seems we have mostly positive feedback from the customers where the reviews seem To be normally distributed in terms of subjectivity which indicates a good mix of personal opinion and fact which is good when describing video games.

The summary sentiment subjectivity scores indicates that in conclusion infact most of the reviews may be based on fact than personal opinion. This again is good that the video games would generalize to wider market better as factually they seem to be good.

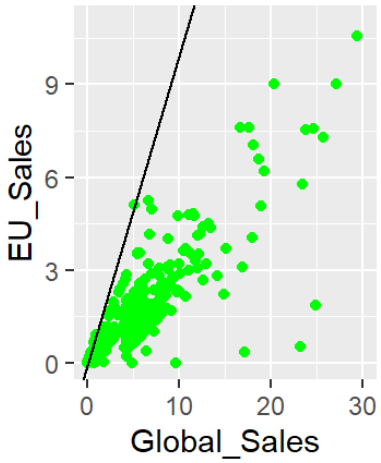
Visualise data to gather insights.

We use the very efficient programming language R to import and view the turtle sales data.
We create scatter plots to view the correlation between sales with
The sales in various regions.

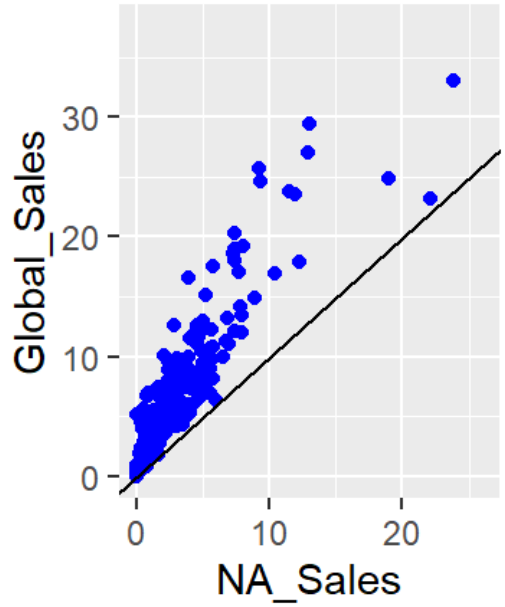
European sales vs North American sales:



European Sales vs Global Sales:

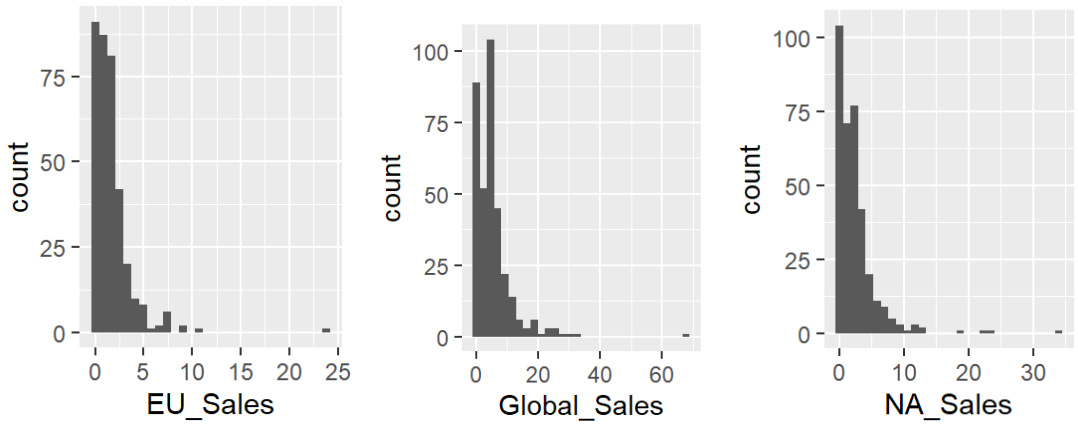


Global Sales vs North American Sales:



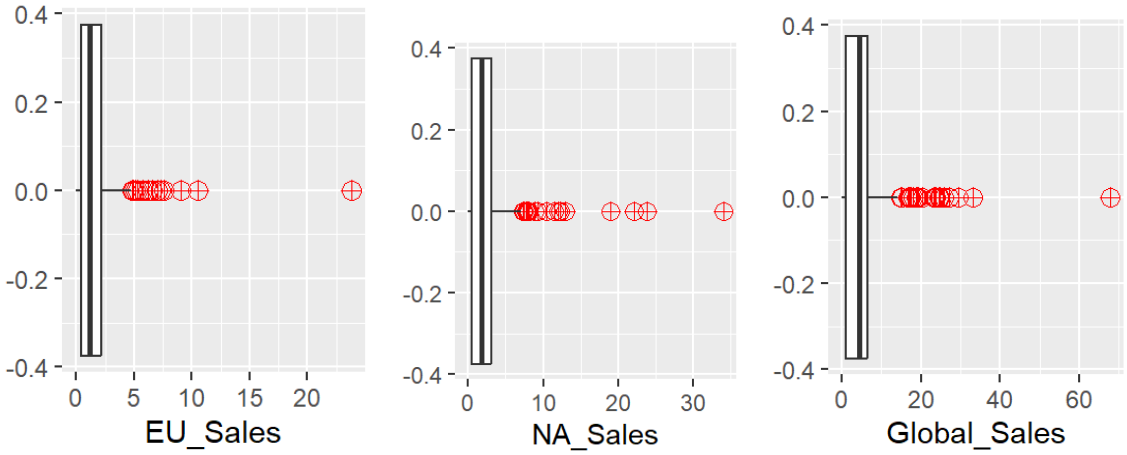
There seems to be positive correlation between all variables according to the line of Best fit.

We create histograms in order to visualize the distribution of the different sales:

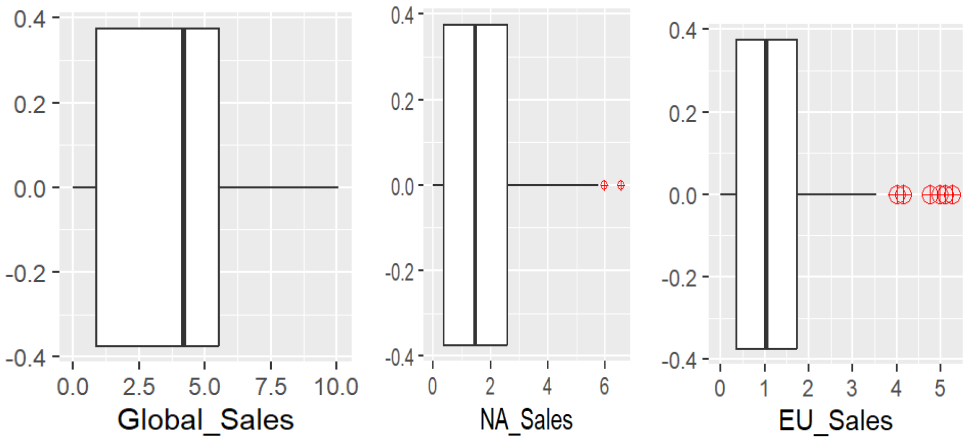


The all seem to be heavily right skewed.

We now create boxplots:



We see a lot of outliers. We now remove or minimize these outliers for better analysis. Below are the resulting boxplots:



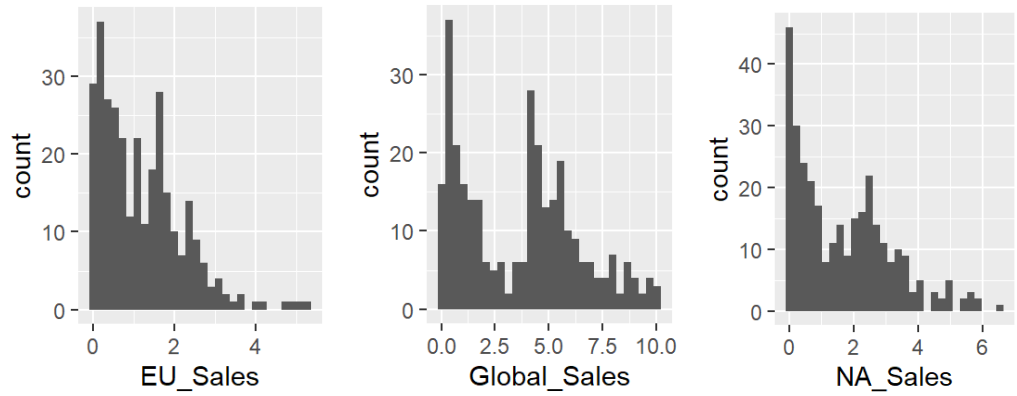
We have now greatly reduced the outliers.

Clean and manipulate data.

We use R to determine the max,min and mean of the different sales.
NA and EU sales minimum are both 0 while Global sales is 0.1.
NA sales maximum is 6.54, EU sales is 5.26 and Global sales is 10.06.
The mean of NA sales is 1.66717, EU sales is 1.185627, Global sales is 3.654.

It appears that scatterplots are the best for comparison of the sales variables.

The histograms of the final dataset with outliers removed is below:



In the above plots we have transformed the variables with the transformation that makes it as close to a normal distribution as possible using the transformTurkey package.
The right skew still exists therefore there is no transformation that can make the variables completely normal.

The correlation of the NA sales with Global sales is 0.86 which implies a strong positive correlation.

The correlation of the EU sales with the Global sales is 0.78.

The initial insights suggest that EU and NA sales could potentially be good predictors for Global sales.

We run t-tests for the sake of completeness to confirm that the mean sales in each region is different.
The significant scores for all of them verifies this.

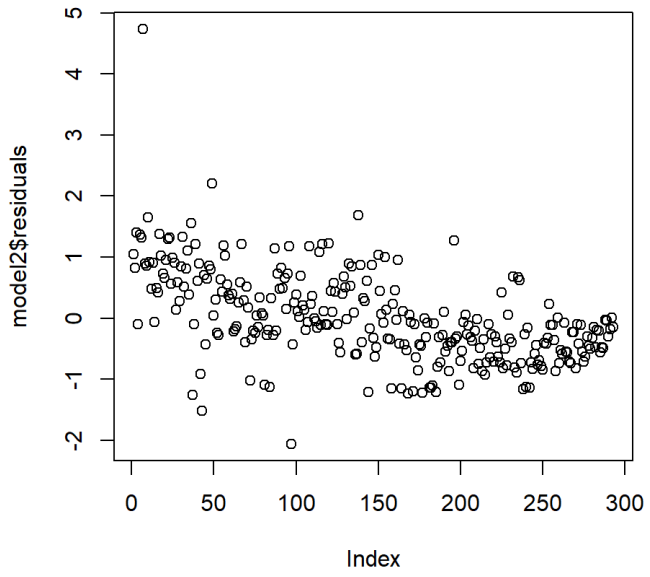
Predict sales with regression.

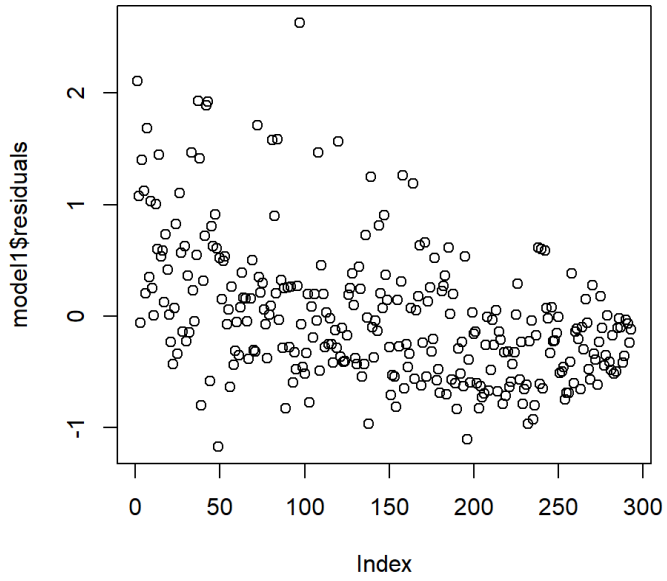
We will use the dataset with no outliers and data as close to normal as possible to perform regression analysis in R.

We create a subset with only the numerical sales columns.
We output the correlation matrix in our R script.
We create two simple linear regression models.
In both models we are predicting Global sales.
In model 1 we use NA sales as the independent variable.
In model 2 we use EU sales as the independant variable.

The summary of both models show that NA and EU variables are significant and thus good predictors.

Below are the residual plots for each model





There seems to be no significant residual autocorrelation.

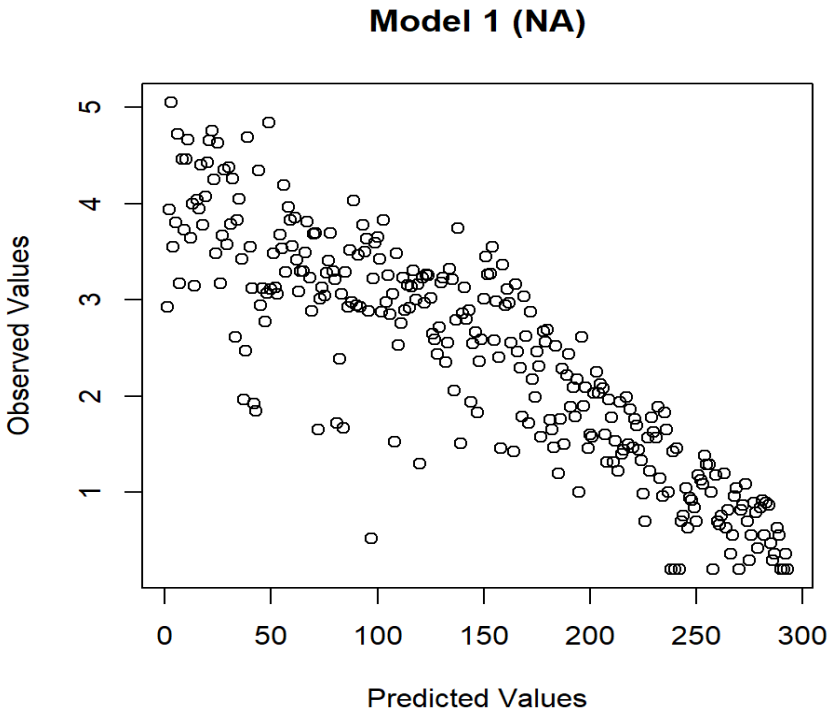
We create a data frame for the sales forecast using the given NA sales and EU sales values in order to make predictions.

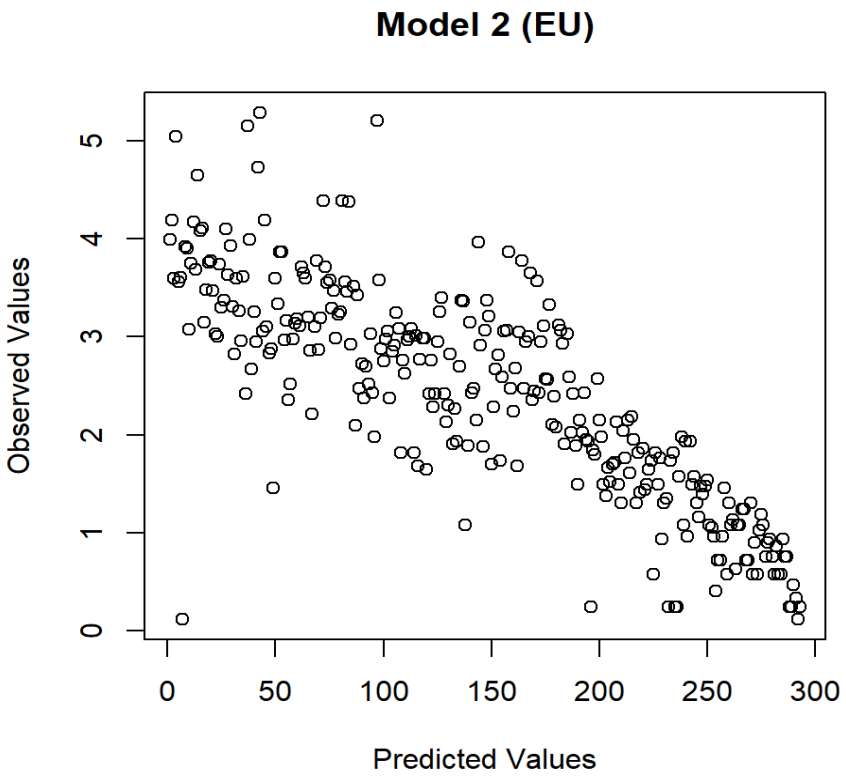
Finnally we will perform a multiple linear regression in R.

We create a model with the X vector containing the NA and EU sales models.

We see both the NA and EU sales variables are highly significant.
However the problem is that there exists multicolinearity between the independant variables.
Therefore, it is suggested to use simple linear regression models to predict global sales.

Below are the plots for the observed values vs the predicted values for both simple linear regression models, they seem to have a good fit to the 45 degree line of best fit indicating a good R squared score.


Model 1 (NA)
Observed Values
1
2
3
4
5
0
50
100
150
200
250
300
Predicted Values



Final Recommendations to the business:

- Do not use age as a predictor for Loyalty points.
- Use remuneration and spending score in a multiple linear regression model to predict loyalty points rather than using either renumeration or spending score in individual simple regression models.
- Renumeration and spending score can be used to predict education level, however we cannot distuingish efficiently between PHD and postgraduate students or basic and diploma students.
- Target graduate students as they form the majority of our customer base.
- We seem to have mostly positive feedback from customers regarding our games but try and we have too many neutral ratings, try and improve overall game quality to shift the neutral ratings to positive ones.
- EU,NA and global sales are all correlated with each other.
- EU and NA sales are good predictors of global sales.
- Fit simple linear regression models using either NA sales or EU sales as independant variables to predict global sales.
- Do not fit a multiple linear regression model as NA and EU sales are highly correlated.
- Sales data is not normal thus perform inference with caution.