LSE Predicting Future Outcomes

**Python / Jupyter Notebook**

1. **Linear regression model:**

Before discussing the approach to the analysis, it is important to note a significant level of 0.05 (95%) has been set.

Turtle Games wanted to find out how customers accumulate loyalty points.

Steps for preparing data for analysis included: loading the data, checking for null values, dropping unnecessary columns, and changing column names. This cleaned data can now be used to create linear regression models.

Linear or Multivariate Linear Regression models answer how customers accumulate loyalty points and enable predictions to be made.

**Linear / Ordinary Least Squared (OLS) models:**

Created three models:

* Spending vs loyalty
* Remuneration vs loyalty
* Age vs loyalty

Loyalty points remains the dependent (y) variable with the three other variables each running as independent (x) variables. Once the variables have been defined we can run the regression formula, fit the model and print the results.

See appendix 1. for the code snippet.

The key linear regression metric to use is R-squared, which measures the variation of the dependent variable explained by the independent variable(s).

The best R-squared comes from Spending Score vs Loyalty Points with an R-squared of 0.452. Meaning around 45.2% of the variation in loyalty points is explained by spending score.

See appendix 2 for OLS summary results.

Ideally, we would be looking for an R-squared higher than 0.452 and therefore will build MLR models.

Ran four MLR tests:

* Spending Score & Remuneration vs Loyalty Points
* Spending Score & Age v Loyalty Points
* Remuneration & Age vs Loyalty Points
* Spending Score, Remuneration & Age vs Loyalty Points

A similar process in running MLR to the OLS model – see appendix 3 for code.

Adding more variables to a linear regression will increase the R-squared but it will also increase the complexity therefore Adj. R-squared is the key metric to consider here.

Of the first three MLR models, Spending Score & Remuneration has the best R-squared (0.827). Finally, created an MLR with all three independent variables, Spending Score, Remuneration & Age, which produces an Adj. R-squared of 0.842.

See appendixes 4 & 5 for code and outputs.

When using MLR it is important to test for multicollinearity which is where there is a strong correlation between the variables. To do this we look for the VIF factor of our variables, the closer this is to 1 the better. Both models have good VIF factor results.

See appendix 6 for results.

Therefore, to determine how customers accumulate loyalty points recommend using the Spending Score, Remuneration & Age vs Loyalty Points model.

1. **Clustering model:**

Clustering helps identify trends in each group, in this case with groups of customers. The clustering model will be conducted using a new data frame with only remuneration and spending score.

The number of clusters (k) is determined by using the Elbow and Silhouette methods

The Elbow method is a decision rule used to determine k by calculating the within-cluster-sum for squared errors (WSS) for each possible value of k. We are looking for a balance between low SS-distance and the number of clusters.

The Silhouette method works out how similar a point is to its own cluster compared to other clusters. Silhouette method values range between -1 and +1. We want a high value where the line peaks whilst also striking a balance between the number of clusters.

See appendix 7 for the code and results of each method.

The results suggest that five clusters is the correct number, checked this by running the k-means model using five and six.

See appendix 8 for results

The output shows five clusters provide the best segmentation of customers. There are clear distinct groups of customers based on their remuneration and spending score.

Once each customer has been allocated a cluster, can concatenate this K-means cluster column to the reviews data frame. This means customer traits can be identified allowing Turtle Games to improve its sales performance.

Clusters 3 and 4 are groups of customers whose spending scores are low and therefore need to be targeted to improve sales. Cluster 4 is particularly interesting as it has the most customers educated as PhD level and more male than female customers.

Recommendation:

Turtle Games should target customers in cluster 4 by offering products which appeal to individuals at a higher education level and / or appeal more to male customers.

1. **Customer Sentiments**

Turtle Games wants to see how customer reviews can inform marketing campaigns.

Prepared a new data frame including only the review and summary columns.

To prepare each column for analysis, a lambda function is applied to convert words into lowercase with a space between. Afterwards, all the punctuation was removed from the columns and true duplicates were removed, of which there were 39.

See appendix 9 for all code.

Once a copy of the data frame has been created, the review and summary columns are tokenised separating each review & summary into individual words. New columns are created to store tokenised reviews & summaries.

The Final part of prepping the data for visualisations is removing stop words such as ‘the’ and ‘and’, leaving only the words relevant for our analysis.

Sentiment analysis takes three parts: word clouds, frequency distributions and polarity scores.

Word clouds show the words that appear most frequently. Frequency Distributions and the bar plots provide a numeric representation of top 15 most frequent words. Lastly, looking at polarity scores allows us to gauge the overall sentiment displayed in each review.

See appendix 10 displaying the results of each analysis.

Recommendation:

Use words from the frequency distribution which have a positive sentiment, such as ‘good’, ‘great’, and ‘love’, to inform the language of marketing campaigns. Identify the games that receive positive reviews and increase the marketing of said products.

**R and RStudio**

1. **Visualise data**

The purpose of this analysis is to look at the impact each product has on sales.

Created a subset, titled sales\_subset, of the original file (turtle\_sales) keeping the relevant columns. Deducted North American (NA) and European (EU) sales from Global to produce Rest of the World (ROW) sales column.

Visualised the data using scatterplots, histograms and boxplots. Ran several visualisations at this stage to gain insights into the impact of products and platforms.

Scatterplot insights:

* Can see that for each sales column (Global, NA, EU and ROW) a product on the Wii platform has significantly higher sales than any other product.
* In NA, products on the NES platform perform disproportionately better compared to the EU and ROW. The NES is a very old platform so many of these are likely to be historic sales.
* Similarly, PS2 products have higher sales in ROW compared to NA and EU. The PS2 isn’t as old as the NES and was a major gaming platform, which could indicate there is still a market for these products in ROW.

Histogram insights:

* Across all four sales columns, the data had a right skew. Products that sold most frequently are at the lower end of each scale.

Boxplot insights:

* Supports insights of previous visualisations. The first boxplots confirm the skewness of the data with the Q1 value much closer to the min value in all four columns.
* Also enables us to see the outliers much clearer of which there are quite a few for each sales column.
* Interestingly, for ROW we can see four products manufactured by Nintendo have large spreads of data, with much higher Q3 and max values than other platforms. Suggests a bias towards Nintendo products in the ROW.

1. **Clean and manipulate data**

This analysis provides us with a gauge of the reliability of the data.

Aggregating sales\_subset data frame groups products by their unique code. The Data frame now has 175 values compared to 352.

Then used Shapiro test, Q-Q plots, and skewness and kurtosis tests to judge the reliability of the data.

Q-Q norm plots help to provide an idea of the skewness and distribution of the data. If the data points are closer to the fitted qqline the data is normally distributed.

In the case of the Q-Qnorm plots for each of the sales columns, the points either sit on or above the qqline. Suggesting our data is positively skewed.

See appendix …..

Used the Shapiro-Wilk test on each column to corroborate the results. Shapiro test gives us a statistical representation of whether the data is skewed. If the p-value from the test is below 0.05 we can reject the null hypothesis that the data is normally distributed. For all three columns the p-values are below 0.05 confirming the skew analysed from the Q-Q plots.

See appendix ….

Lastly, skewness and kurtosis tests were run on each column as a final reliability test of the data.

A result of 0 from the skewness test would represent data which is perfectly symmetric in its distribution. Given the Q-Q plots and Shapiro test, we would be expecting positive values indicating positive skew. All three of our columns produce positive values from the test again confirming the positive skew.

The kurtosis test helps us identify if our data has heavy or light tails. The benchmark for this test is a normal distribution which has a kurtosis value of 3; below 3 indicates light tails and above 3 indicates heavy tails. All three of our columns produce kurtosis scores significantly higher than 3 indicating heavy tails.

Insight:

Is skewness or kurtosis bad? Ideally, we want a skewness value between -3 and +3. Only the European Sales column has a value in this range. Too much skewness can raise concerns about the reliability of the data and could cause when applying statistical models. To improve the scores, we could remove outliers like the products on the Wii platform identified earlier. However, decided against this to avoid overfitting the data for good values on each test.

1. **Predict sales with regression**

Like with models in Python built both linear and multivariate linear regression models, to determine what relationship is there between North American, European and global sales.

Linear Regression Models:

Created three models plotting each sales column against each other:

* sum\_global~sum\_NA
* sum\_global~sum\_EU
* sum\_NA~sum\_EU

Created each model using the lm() function.

Once each model was created could then use summary() to view the statistics and determine the strength of the model.

See appendix ….

R-squared results:

* sum\_global~sum\_NA: 0.8395
* sum\_global~sum\_EU: 0.7201
* sum\_NA~sum\_EU: 0.3856

The first two linear regressions indicate a strong relationship between the two variables, particularly the sum\_global~sum\_NA regression which shows that 83.95% of the variation in global sales is explained by NA sales.

To improve on this R-squared, created an mlr model using the original data, plotting Global\_Sales~NA\_Sales+EU\_Sales. Produces a high R-squared of 0.9687 / 96.87% of the variation in global sales explained by NA and EU sales.

This model can then be applied against new values to make predictions. For instance, if a new product had NA sales and EU sales of 34.02 and 23.80 our model would predict this product has global sales of 71.47. When checking the original data set, we can see the product with ranking number 1 has this exact number of NA and EU sales and has global sales of 67.85. The difference between the two values is only 3.62. Shows our model is good at placing accurate predictions.

Insight:

NA sales and EU sales do not have much of a relationship to each other. However, both have a strong relationship to global sales and creating an mlr with both against global sales produces a model with strong predictive qualities.

**Overall Recommendations:**

* Use mlr model of spending score, remuneration and age to predict loyalty points.
* Turtle Games should look to target customers in cluster 4 who have a high remuneration but low spending score, which could help increase sales performance.
* Utilise sentiment analysis to form the language for marketing campaigns. Also, analyse the polarity score of reviews to see if any products consistently receiving positive reviews are not being marketed properly.
* Use mlr model of Global\_Sales~NA\_Sales+EU\_Sales to make accurate predictions of product sales.