**LSE Predicting Future Outcomes**

**Python / Jupyter Notebook**

1. **Linear Regression Models:**

* A significance level of 0.05 has been set=

Turtle Games wants to find out how customers accumulate loyalty points:

* After loading, prepared data by checking for null values, dropping unnecessary columns and changing remuneration and spending score column names

Built linear and multivariate linear regression models to see what variables explain customers' loyalty points.

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 other three variables as independent (x) variables
* Run the regression formula, fit the model and print the results

See appendix 1. for code snippet

* Key metric is R-squared which measures the variation of the dependent variable explained by the independent variable
* Highest R-squared was produced by Spending Score vs Loyalty Points with a result of 0.452.
  + This means around 45.2% of the variation in loyalty points is explained by spending score

See appendix 2 for OLS summary results

Ideally, want an R-squared higher than 0.452. Therefore, built four MLR models.

MLR models:

* Spending Score & Remuneration vs Loyalty Points
* Spending Score & Age v Loyalty Points
* Remuneration & Age vs Loyalty Points
* Spending Score, Remuneration & Age vs Loyalty Points
* Adding more variables to a linear regression will increase the R-squared
* However, it will also increase the complexity
* Therefore, Adj. R-squared is the metric to consider here
* Of the models with two independent variables, Spending Score & Remuneration has the best Adj. R-Squared – 0.827
* MLR with all three independent variables produces an Adj. R-squared of 0.842

See appendix 3 for code snippets

See appendix 4 for MLR results

* When using MLR it is important to test for multicollinearity which is where there is a strong correlation between the variables
* To do this look for the VIF factor of the independent variables. The closer to VIF is to 1 the better.
* Both models have low VIF Factor results

See appendix 5 for results

Recommendation:

* 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 from a data set, in this case, Turtle Games’ customers. The clustering model will be conducted using a subset data frame of only remuneration and spending score.

* Number of clusters (k) determined by using the Elbow and Silhouette method

Elbow method

* Decision rule used to calculate k by calculating the within-cluster-sum for squared errors (WSS) for each possible value of k
* Looking for a balance between a low SS-distance and the number of clusters

Silhouette method

* Works out how similar a point is to its own cluster compared to other clusters
* Values range between -1 and +1
* Want a high value where the line peaks whilst also striking a balance between the number of clusters

See appendix 6 for code snippers and results of each method

* Results of each method suggest that five clusters is the correct number of clusters
* Double-checked running k-means model using five and six clusters

See appendix 7 for results

* Output shows five clusters segments the customers the best
* Provides clear and distinct groups of customers based on their remuneration and spending score
* Once each customer has been allocated a cluster, can concatenate the K-means column to the reviews data frame
* Allows customer traits to be identified meaning Turtle Games target customer to help improve its sales performance
* Clusters 3 and 4 are groups of customers where the spending score is low and therefore have the potential to improve sales performance
* Cluster 4 is particularly interesting as it has the most customers educated as PhD level and more males than females

Recommendation:

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

1. **Customer Sentiments**

Turtle Games want to see how customer reviews can inform marketing campaigns

* Created a subset data frame including only the review and summary columns
* Prepared each column for analysis by applying a lambda function to convert words into lowercase with a space between
* Removed all punctuation and true duplicates, of which there were 39

See appendix 8 for code.

* Copy the subset data frame
* Tokenised the columns separating each review & summary into individual words
* New columns created to store tokenised reviews & summaries
* Removed stop words such as ‘the’ and ‘and’, leaving only words relevant to the analysis

Sentiment analysis takes on three elements:

* Word clouds
  + Visualisation of which words feature regularly
* Frequency distributions
  + Provides numeric representation of top 15 more frequent words
* Polarity scores
  + Allows us to gauge overall sentiment displayed in each review

See appendix 9 for graphs

Recommendation:

* Words from frequency distributions which have a positive sentiment such as ‘good, great and love’ can inform the language of marketing campaigns
* Identify the games that receive positive reviews and increase marketing of said products

**R and R Studio**

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 turtle\_sales data keeping the relevant columns
* Deducted North American (NA) and European sales (EU) from global sales to produce Rest of the World (ROW) sales column
* Visualised the data using scatterplots, histograms and boxplots to gain insights into the impact of products and platforms

Scatterplot insights:

* A product on the Wii platform has significantly higher sales than any other product for all sales columns
* In NA, products on the NES platform perform disproportionately better compared to EU and ROW. NES is a very old platform, so these are likely historic sales.
* PS2 products have higher sales in ROW compared to NA and EU. PS2 isn’t as old as the NES and was a major gaming platform so could indicate there is still a market for these products in ROW

Histogram insights:

* All four sales columns, the data has a right skew.
* Products that sold most frequently are at the lower end of the scale.

Boxplots insights:

* Supports insights from previous visualisations
* First four boxplots confirm the skewness of the data with Q1 much closer to the min value for all sales columns
* Can see the outliers much clearer of which there are several for each sales column
* Interestingly, ROW has four products manufactured by Nintendo which have much higher Q3 and max values than other platforms. Suggests a bias towards Nintendo products in the ROW

1. **Clean and manipulate data**

The analysis provides 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.

Afterwards, used Q-Qnorm plots, Shapiro tests and skewness & kurtosis tests to judge the reliability of the data.

Q-Qnorm

* Helps 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
* For each sales column Q-Qnorm plots the points either sit on or above the qqline
* Suggests the data is positively skewed

See appendix 10 for plots

Shapiro Tests

* 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
* All three columns have p-values below 0.05 confirming the skew from Q-Qnorm plots

See appendix 11 for results

Skewness and kurtosis

* Run on each column as a final test of reliability
* Result of 0 from the skewness test represents data that is perfectly symmetric in its distribution
  + All four columns produce positive values, above 0, confirming the positive skew
* Kurtosis test helps identify if the data has heavy or light tails
  + Benchmark for the test is a normal distribution which has a kurtosis value of 3
  + Below 3 indicates light tails and above 3 indicated heavy tails
  + All three sales columns’ product kurtosis scores are significantly higher than +3 indicating heavy tails

See appendix 12 for results

Insight:

* Is skewness or kurtosis bad?
* Ideally, 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 cause issues when applying statistical models
* Could remove outliers, like Wii products, to improve scores from tests. However, decided against this to avoid overfitting the data for good values.

1. **Predict sales with regression**

Like with Pyton built both linear and multivariate linear regression models, to determine what relationship exists 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. After each model was created can then see a summary of the results and determine strength of the model.

See appendix 13 for results

R-squared results:

* sum\_global~sum\_NA: 0.8395
* sum\_global~sum\_EU: 0.7201
* sum\_NA~sum\_EU: 0.3856
* First two linear regressions indicate strong relationships between NA and EU columns to Global Sales
* Particularly the sum\_global~sum\_NA regression which shows that 83.95% of variation in global sales is explained by NA sales

To improve on this R-squared, created an MLR model from the original data, plotting Global\_Sales~NA\_Sales+EU\_Sales.

See appendix 14 for results

* Produces a high R-squared of 0.9687
* 96.87% of the variation in global sales is explained by NA and EU sales

This model can then be applied against new values to make prediction:

* For instance, if a new product has NA sales and EU sales of 34.02 and 23.80 the model would predict this product has global sales of 71.47
* A product in the turtle\_sales data, with ranking number 1, has exactly these NA and EU sales and global sales of 67.85
* Difference of 3.62 between the predicted and actual
* Shows the model is very good at producing accurate predictions

Insight:

* When combined in an MLR, NA and EU sales have a strong relationship with global sales allowing us to make good predictions
* Turtle Games should use this model to help predict sales of new products

**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.