**ASSIGNMENT REPORT, COURSE 3**

**PREDICTING FUTURE OUTCOMES**

**ALI FAZEL**

Using data from sales and reviews, Turtle Games wants to improve overall sales performance. This analysis supports with that by understanding customer trends, insights as well as dissecting sales data and determining relationships and correlations.

**Week 1: Make predictions with regression**

In order to better understand how customers, accumulate loyalty points three relationships have been explored within the data:

1. loyalty points vs spending score
2. loyalty points vs remuneration
3. loyalty points vs age

Using linear regression, it was determined that there was a correlation between spending score and loyalty points. The OLS regression results produced an R-squared value of 0.45 (or 45%) which suggests that 45% of the variability in the dependent variable (loyalty points) can be explained by the variability in the independent variable (spending score). Whilst the figure suggests there is still 55% of the variation unexplained, it helps to understand that the more a customer spends, the higher their loyalty points are likely to be.

When exploring loyalty points and remuneration, an R-squared value of 0.38 (38%) was calculated. The relationship between these two variables is not emphatic however there still exists a weak correlation. For both this and the former regression model, the p-value is less than 0.05 suggesting these are significant relationships.

We can confidently conclude that there is no relationship between age and loyalty points after this regression showed only 0.2% of the variability in loyalty points can be attributed to the variability in age. The probability of the t-value is also greater than 0.05 suggesting that this is not a significant relationship – it is non-existent.

Therefore, we can use age and remuneration, with caution, to predict the loyalty points a customer might have, however it would not be advisable to use age.

**Week 2: Making predictions with clustering**

In order to correctly advise the marketing department, remuneration and spending score were further explored with k-means clustering to identify segments of customers that can be targeted with specific marketing.

Using the Elbow and Silhouette methods, it was determined that the most appropriate number of clusters for this dataset was five. The elbow method looks for a change of slope from steep to shallow (an elbow) to determine the optimal number of clusters, this was evident at four or five clusters.

By employing the Silhouette method, we were able to measure similarity of the object to its assigned cluster, with respect to other clusters - this confirmed five clusters of customers was the most appropriate as it was at this point we were able to maximise the silhouette coefficient.

**Week 3: Analyse customer sentiment with reviews**

We applied natural language processing (NLP) to the customer reviews for Turtle Games in order to inform the marketing department on how to approach future campaigns. After reading the data and dropping the unnecessary columns, the pre-processing begun. This involved changing all the reviews and summaries to lower case using a *lambda* function, removing punctuation using the *str.replcace* function, and dropping duplicate entries.

With the data cleaned, we were able to move onto creating word clouds (*See Appendix 1).* With a cleaned list of tokens, the *Counter* class was employed alongside the *most\_common* function to both columns to identify the 15 most frequently used words. For the review column ‘game’, ‘one’, ‘play’, ‘fun’ and ‘great’ were the top five, whilst for the summary column it was ‘game’, ‘great’, ‘fun’, ‘good’ and ‘love’. Using a user defined function to generate the polarity score which forms part of the sentiment score alongside subjectivity, polarity was calculated for each review and summary and this was added to the data frame.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedA histogram was plotted indicating polarity for both review and summary. The review histogram indicated that the majority of feedback was positive as the majority of the data fell beyond the neutral mark of 0. This should give Turtle Games some confidence that on the whole customers are largely happy with their products. Whilst there is a significant amount of feedback that is neutral, negative sentiment is limited when compared to positive sentiment. The histogram for the summaries told a similar story, although the feedback was more varied with negative sentiment stretching across the entire scale all the way to -1. That said, the vast majority of the data was positive with a very large proportion of it also being neutral, at 0.

Finally, the 20 most positive and negative reviews and summaries were printed based on polarity as Turtle Games requested – more on this in the presentation. *(See Appendix 2)*.

**Week 4: Visualise data to gather insights**

The *qplot* function was used to plot various graphs to determine some relationships in the data as part of the exploratory data analysis process. In some cases, for example when plotting boxplots, it was necessary to specify the *geom* to tell R exactly what plot was wanted.

The scatter plot shows us that there is a positive correlation between European sales and North American sales. This would be helpful for the company as it informs them that similar games are popular in each region. Positive correlations between European and Global Sales as well as North American and Global Sales were expected as these figures contribute towards the Global Sales. The scatter plots are easy to read and so would be useful for stakeholders.

When plotting the histograms, it became clear that across the board, sales are skewed to the right with the vast majority of global sales falling into bins in the £0-10 million. Though the data did range going up to a maximum bin of £65-70 million. The company might be particularly interested in the data in this bin to understand what is driving sales.

The box plots compare the various platforms and the distribution of sales on a global, North American and European level. Though, these offer little value to the client and will not be included in the final presentation of insights. This is because firstly, boxplots are technical and may not be the most appropriate visualisation for a non-technical audience looking to gain value from insights instantly. Secondly, these boxplots are very busy, as there are several platforms which need to be compared. It may be an option to group together similar products (ie: PS, PS2, PS3, PS4 could all be grouped together so we can get a more high-level view of how this brand is performing overall in comparison to all the Wii products). This will make the boxplots a bit easier to understand.

**Week 5: Clean and manipulate data**

The impact on sales per product ID was determined by creating a new dataset which grouped the sales data across all regions by product and displayed the sum of NA\_Sales, EU\_Sales and Global\_Sales. After some exploration of the new grouped dataset, it was used to plot scatter graphs, histograms and boxplots, using *ggplot*. As already determined in week 4, the scatter plots provide more value than the histograms and boxplots *(See Appendix 3).*

With this in mind, the scatterplots were enhanced with a line of best fit and these would be presented to Turtle Games. They show that North American sales and European sales strong positive correlation of 0.706, suggesting that if a game is popular in Europe, it is likely to also be popular in North America. We saw even stronger correlations with North American sales and global sales (0.935) as well as European sales and global sales (0.878). The trend lines will allow Turtle Games to make rough estimations of sales in Europe or North America for a certain product based on the sales in one region.

Separately, a data frame was created which grouped sales across all regions by platform using the *groupby* and *summarise* functions. This data was then used to generate a bar plot on *ggplot* which allows Turtle Games to understand what platforms are most popular globally – the bar plot shows Wii as most popular with over £300 million in sales, X360 in second place with just over £250 million and PS3 with over £200 million.

Chart, histogram

Description automatically generated

Similar plots were generated for Europe and North America sales, however, to add more value, a new bar plot showing both regions sale figures were displayed on the same plot. (*See Appendix 4 for the technical process*). The graph shows that whilst Wii and X360 are the most popular, both platforms are far more revenue generating in North America than in Europe. Also noteworthy is that there are only two platforms which generate greater revenues in Europe than in North America: PS3 and PS4. Depending on business objectives, this information might allow the marketing strategy to either focus more on pushing games on these platforms or widening appeal and trying to create more demand for games on other platforms in Europe in order to match demand in North America.

Chart, bar chart

Description automatically generated

*(See Appendix 5)*

**Week 6: Predict sales with regression**

The correlations in the dataset were initially explored using the *cor* function. Three simple linear regression models were created exploring the relationships between:

1. NA\_Sales and EU\_Sales (Multiple R-Squared: 0.386)
2. NA\_Sales and Global\_Sales (Multiple R-Squared: 0.840)
3. Chart, scatter chart

   Description automatically generatedEU\_Sales and Global\_Sales (Multiple R-Squared: 0.720)

This confirms that 38.6% of the variation in European sales can be explained by the variation in North American sales. Much stronger positive correlations can be seen in the following models: 84% of the variation in global sales can be explained by variation in North American sales and 72% of the variation in global sales can be explained by variation in European sales.

*(See Appendix 6)*

The *abline* function was used to add a line of best fit on each model which would allow Turtle Games to predict the y variable if the x-variable was provided. However, a multiple linear regression model would be more appropriate as there will be more than just one variable that impacts global sales.

The multiple linear regression model took two x-variables: EU\_Sales and NA\_Sales which would together be used to determine Global\_Sales (y-variable). Two models were created – one with the aggregated data set and one with the raw data. As advised however, the aggregated dataset was used to predict global sales. The prediction data was made into a data frame and then inserted as the ‘*newdata’* parameter in the *predict* function.

Text

Description automatically generated with medium confidence

The predicted global sales are noticeably close to the observed values. The model’s R-squared value of 0.9668 informs us that 96.68% of the variation in global sales can be explained by NA\_Sales and EU\_Sales. The Adjusted R-squared of 0.9664 is also high suggesting we can confidently trust the model to make accurate enough predictions to inform stakeholders as seen in the examples above. Although these are not exactly accurate, Turtle Games will be able to confidently utilise this model to roughly predict global sales.

**APPENDIX:**

**Appendix 1 (Week 3) – Preparing data for word clouds**

All the reviews and summaries were placed into their own strings using a for loop which allowed us to create the word clouds. Next, the columns were tokenised using the *apply.(word\_tokenize)* function. We then created a list of tokens for both the review and summary columns, this allowed us to calculate the frequency distribution using the FreqDist function. From the list of tokens, we used a for loop to remove all tokens that were not letters or numbers as well as all stop words – to do this we used for loops with embedded if statements. This list was then transformed to a string so that we could visualise word clouds without stop words.

**Appendix 2 (Week 3) – A note of caution with polarity scores.**

It is advised that they exercise caution with some of the polarity scores as the algorithm has assigned a polarity which does not correlate to some pieces of feedback. For example, “not the best quality” was assigned a polarity of 1.0, “I really like this game it helps kids recognize anger and talk about difficult emotions”, which is clearly a positive piece of feedback was assigned a polarity of -0.35. Whilst these cases of misaligned polarity scores are rare, it is worth bearing in mind that the algorithm may sometimes wrongly classify the feedback.

**Appendix 3 (Week 5) – Justifying choice of plots**

Scatter plots were chosen over histograms and boxplots as they show clear correlations of the different sales regions and can be used by Turtle Games to understand popularity of products in both regions. Whilst the histograms show a right skew for all sales data there is not much value gained from a commercial point of view from simply seeing the distribution of the data. Similarly, the boxplots do not help the stakeholders to understand correlations and relationships like the scatterplots do. Boxplots simply show distribution – which may be useful in the initial stages of exploratory data analysis, but not when feeding back insights and findings to the stakeholder.

**Appendix 4 (Week 5) – Technical process of melting data to plot bar plots**

A data set which summed EU and NA sale regions by platform was melted using the melt function so that both North American and European sales could be plotted on the same bar plot side by side. The reshape2 package was needed to restructure the data set. This allows Turtle Games to directly compare the popularity of a platform in each region which can support with marketing and commercial strategies. The melting put the variable (NA\_Sales or EU\_Sales) into a single column which meant this could be added as a ‘fill’ parameter in ggplot providing 2 bars for each platform.

**Appendix 5 (Week 5) – Normality of data**

When working with the descriptive statistics on the sales data to determine normality, qqplots were first plotted. They showed that for North American, European and Global sales, the data most certainly was not normally distributed. The qqline function confirmed this – it showed that the data moves significantly away from the normal distribution from 1-3 standard deviations above the mean of the normal in all sales data sets. This means the data is more extreme than the normal and so we can expect to have heavy tails.

The Shapiro-Wilk test was then applied to all rows of sales data, producing an extremely low p-value of 2.2e-16 for Global and NA sales and 2.987e-16 for EU Sales. Since the null hypothesis is that the data is normally distributed, a low p-value means we can confidently reject the null hypothesis and confirm what we have already seen on the qqplots: the data is not normally distributed.

Skewness and Kurtosis was then calculated for each column of sales data. As expected, the skewness was significantly above 0 for each column, 3.05, 2.89, 3.07 for NA\_Sales, EU\_Sales and Global\_Sales, respectively. This confirms a positive skew (data skewed to the right) in all cases. The Kurtosis test is a measure of how heavy or light the tail is. Comparing the outputs for the sales data to the Kurtosis for a normal distribution of 3, we see the figures are significantly higher:

|  |  |
| --- | --- |
| **Sales:** | **Kurtosis value:** |
| North American Sales | 15.6026 |
| European Sales | 16.22554 |
| Global Sales | 17.79072 |

This tells us that that the tails are extremely heavy in all cases. The extreme data points are where sales are significantly higher and so further away from the mean than expected. A heavy tail distribution is when we observe extreme values which are the main reason for the variance in the data.

**Appendix 6 (Week 6) – Determining whether to use logs**

To ensure the best possible fit, the logs of EU\_Sales were taken and a new variation of the first model was created exploring the relationship between the logs of EU\_Sales and NA\_Sales. In order to do this, we had to remove the ‘-inf’ value in the logsEU\_Sales column and change this to a NA value. This model produced a worse fit with a multiple R-Squared value of 0.195 and so a decision was made not to use the logs.