DA301 Assignment Report
by Jena Shubotheesh

Word count: 1195

Background

This report aims to look at Turtle Games' provided data and provide some insights on their products. They are a game manufacturer and retailer on a global scale. We have been provided sales information as well as customer reviews to be manipulated and visualised. By analysing trends and patterns, we will consider potential future decisions that could be made by the business to improve sales in the future.

Analysis

Before being able to analyse the data, we need to clean it. Removing unnecessary columns and renaming complex names made it easier to code. Both Python and R have internal functions to drop columns or rows. Looking and omitting rows with incorrectly formatted data was important since future visualisations can be easily skewed by false data. For example, turtle_sales had an entry with 'Platform' having an incorrect value, so was removed.

Python

I required certain tools to perform our analysis. First was statsmodel, which let us create regression models from the dataframes. I used this to create a simple linear regression looking at loyalty points against spending, remuneration and age. By adding a line of best fit, we can then get an understanding of the correlation each variable has. This helps us answer Turtle Games' question on what influences loyalty points.

I then performed cluster analysis to help Turtle Games see different customer types that may exist. I used the sklearn library to do K-means analysis. This paired with seaborn let me create effective visualisations of clusters. To evaluate the best number of clusters to use, I used the elbow and silhouette method.

Finally, I analysed customer reviews using natural language processing (nltk). I needed to prepare the review data before analysis. This involved making the words lower case and removing punctuation. I would have also stemmed and lemmatised the data to simplify the words more but our wordclouds showed some good insights regardless. Removing stopwords was imperative since 'and', 'that' and 'the' were most popular prior. I was also able to create polarity histograms using the textblob function to see the positivity of reviews. This shows us how marketing campaigns have and can influence sales.

R

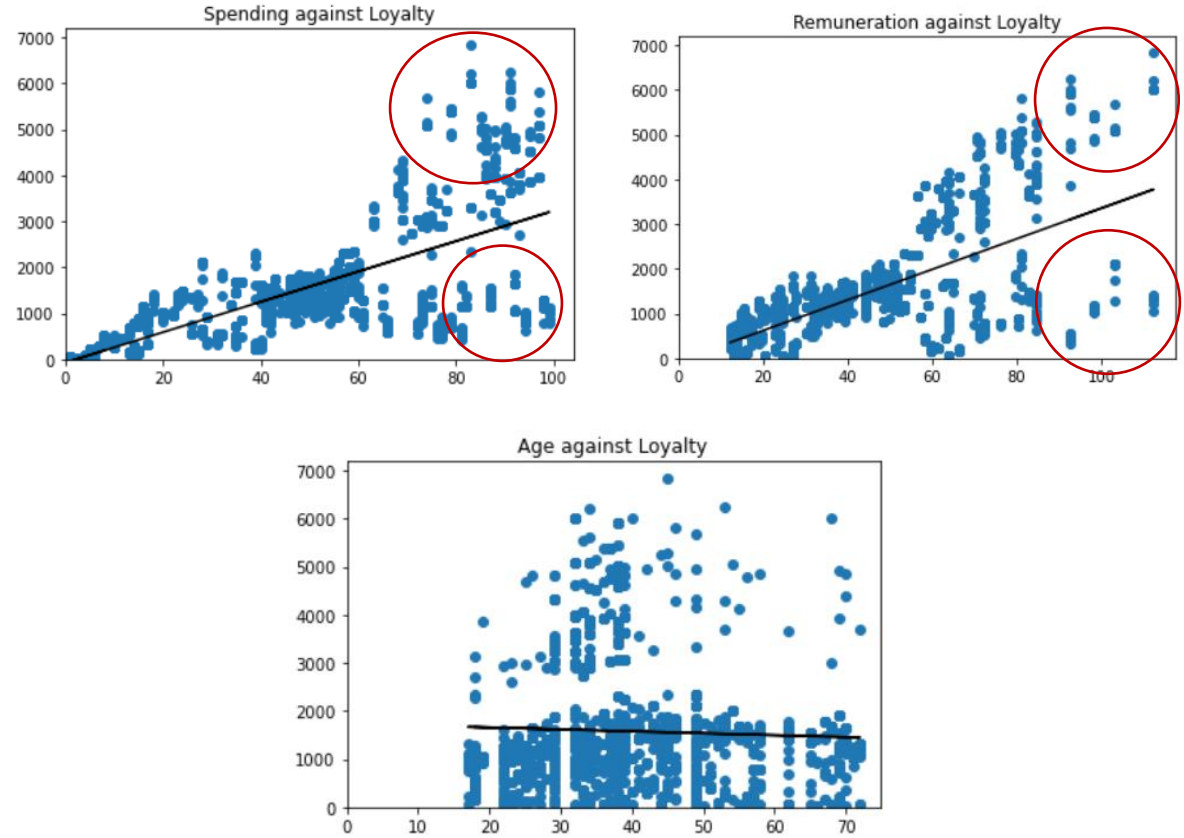
Since the sales team preferred R, we used this to look at sales data. Using tidyverse, I could plot data on sales for different regions, splitting it by platform. The three plots I considered were scatter, histograms and boxplots. Platforms were not of significant interest to Turtle Games, so I have omitted them from my insights.

Following this, we can look at how each product has influenced sales. I used the groupby function to sum sales by product id. Using this, I planned to create similar visualisations as with platforms. However, an immediate issue was that the column was read as a continuous variable, so the plots were wrong. After changing it to a factor in R, I realised there were too many

individual products to plot effectively. I therefore decided to look at the top and bottom 10 selling products. I did this by sorting the columns then subsetting the first 10 rows.

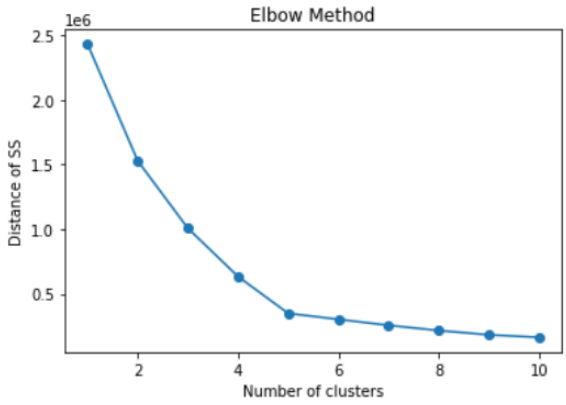
Finally, I used regression to see relationships between sales from different regions and predict potential global sales using this. I used the original dataframe without grouping by products as different platforms sales may impact predictions. I used the 'lm' function on R to do a regression on EU and NA separately then together as a multiple linear regression. I then used the predictmodel function to calculate potential global sales based off EU and NA sale inputs.

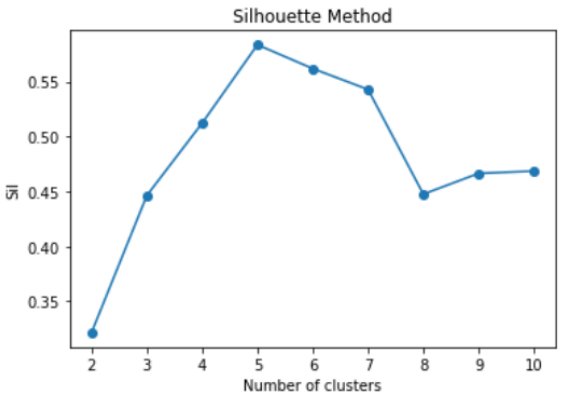
Insights



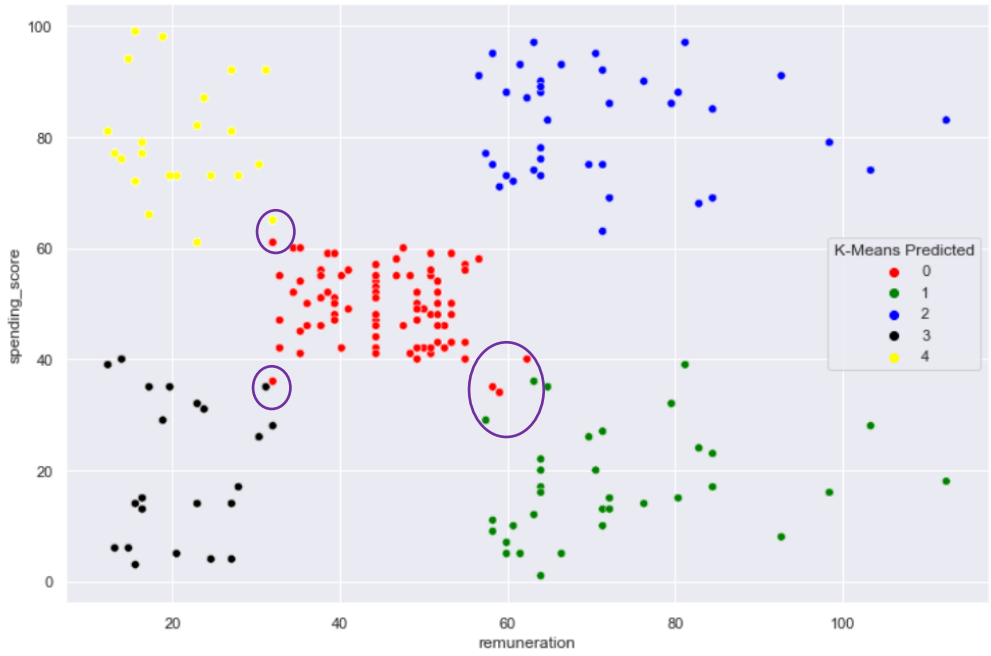
Here we see loyalty points mapped against different variables. The lines of best fit for spending and remuneration suggest there is some positive correlation between them and loyalty score. The LM p values for all 3 are well below 0.05, suggesting heteroscedasticity, which contradicts a key assumption required for this regression. To explore further I would try to add other variables or transform the data.

Simple linear regression is inherently a limited method. There is a risk of the data being underfitted given only 2 variables are considered and this is seen through the low R^2 values for all 3. We assume there is no multicollinearity. Furthermore, outliers are highly influential and impact the accuracy of the model. We can see clusters of potential outliers circled in red.



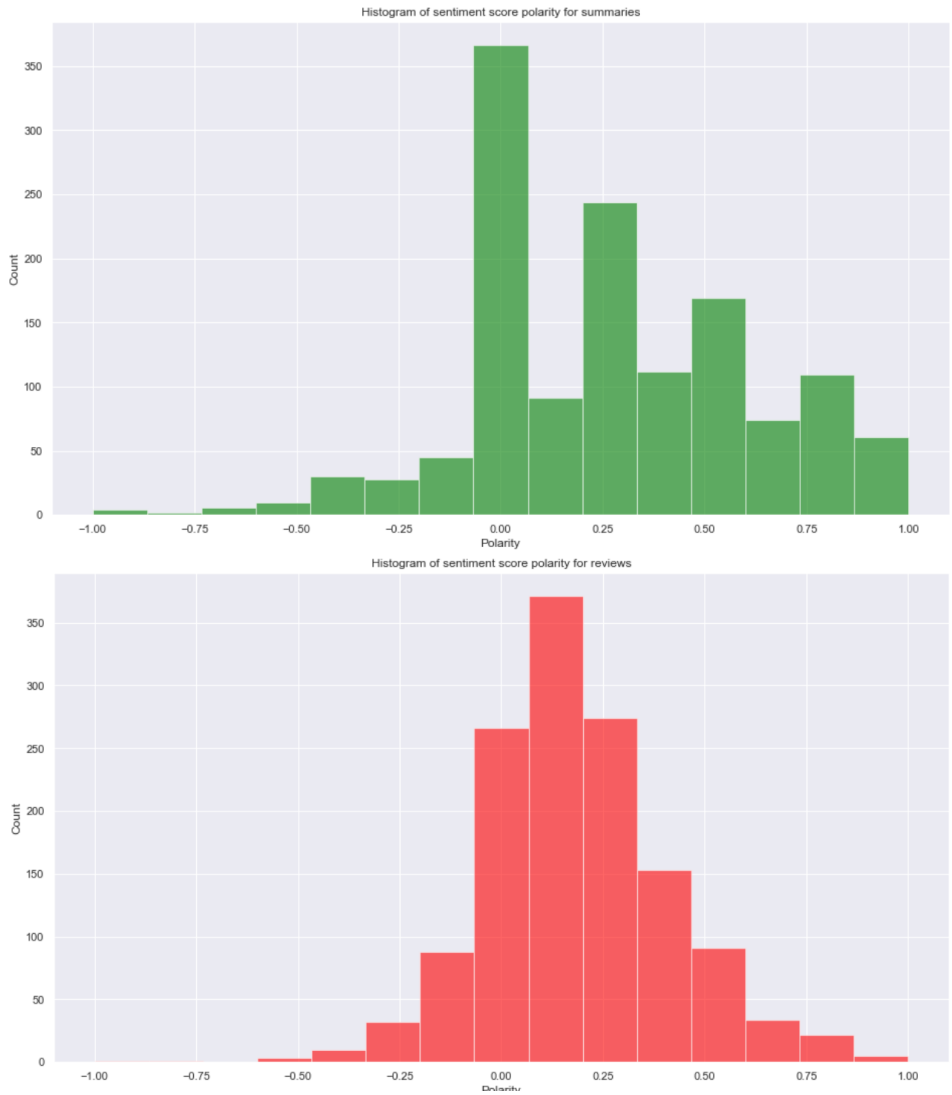


The elbow and silhouette method produced this. Both suggested 5 is the optimal number. Note this method is ineffective if the clusters aren't naturally distinct. To evaluate this, I compared pairplots with 2, 5 and 7 clusters and found 5 to be best. Looking closer, we see some close points between clusters (purple circles) but its not enough to justify more clusters. Turtle Games can try and market for these 5 groups separately to increase sales.



The polarity histograms show review sentiment better than the wordcloud since bars are easier to read. Negative polarity goes from -1 to -0.3, while the positive polarity goes from 1 to 0.8. Sentiment analysis may be inaccurate since certain terms can be misinterpreted. We still see a

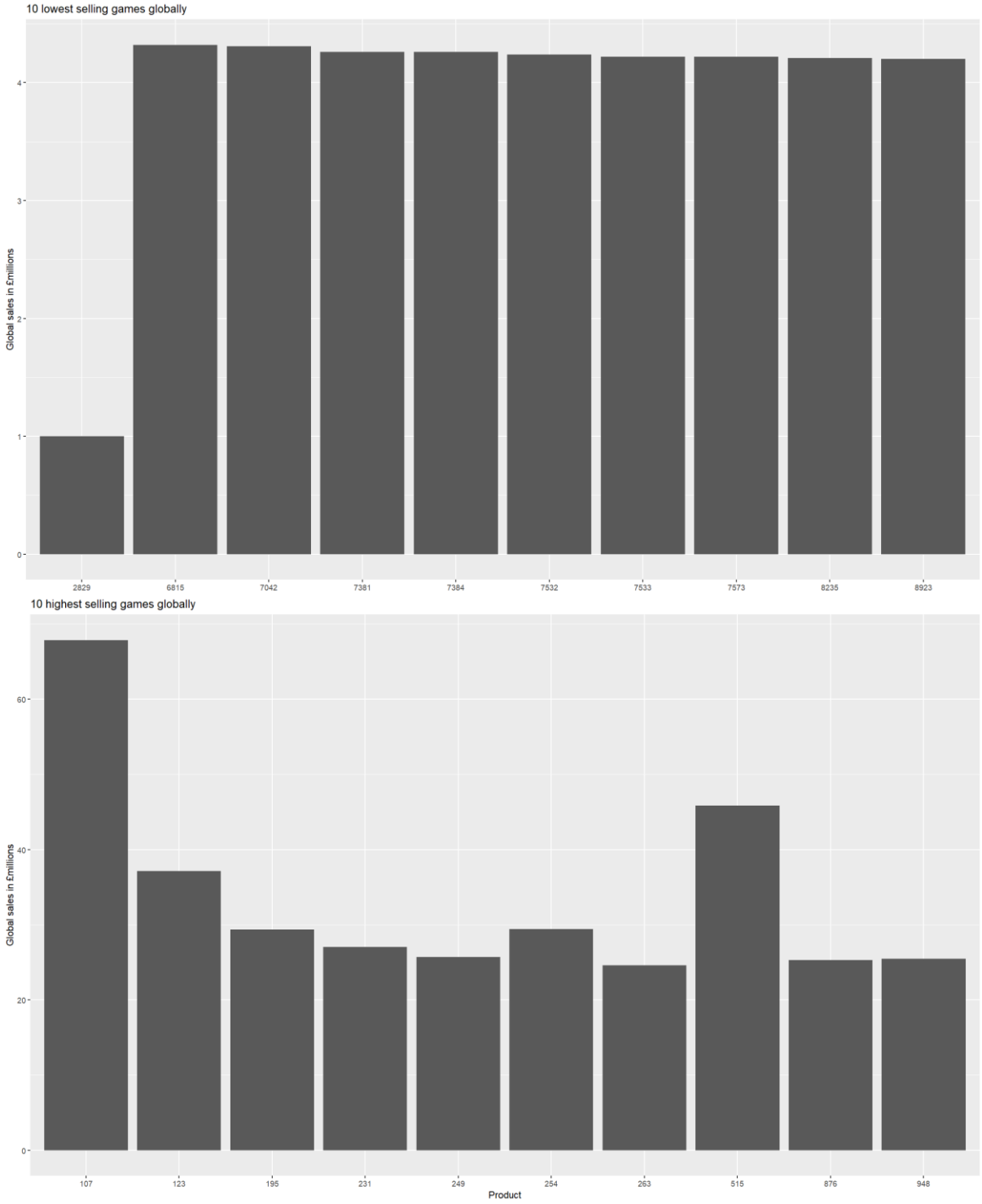
common trend through all visualisations. This signals that marketing has been effective for these products and should continue.



The skewness and kurtosis for global data is extremely high, implying heavy tails on our data. We can infer that most products have low sum of sales with a few products dominating with high sales.

```
> skewness(data2_sums$sum_Global)
[1] 3.037823
> kurtosis(data2_sums$sum_Global)
[1] 17.6016
```

To look at this further, I decided to plot product. Looking at the top and bottom 10 selling products, we get the following. We see 107 selling the most and 2829 selling the least. Turtle Games may decide to discontinue or increase manufacturing based off this.



Finally, I did a multiple linear regression to look at the impact of EU and NA sales on global sales.

```
Call:
lm(formula = Global_Sales ~ EU_Sales + NA_Sales, data = data4)

Coefficients:
(Intercept)     EU_Sales     NA_Sales
     0.2254       1.3373       1.1584
```

EU has a slightly greater influence on global sales than NA. Our regression model had an R^2 value of 0.969 suggesting a high goodness of fit. We can predict using our model the global sales values based on certain regional sales.

Using the provided values in the model yielded us the following:

Actual value	Predicted value
67.85	71.463427
6.04	6.864220
4.32	4.257140
3.53	4.140621
23.21	26.498910

We see the predicted values are all slightly over the actual value meaning the model is overestimating. Looking at the upper and lower bounds shows that only 1 of the actual values lie in the boundaries. This suggests that the model is quite inaccurate. Perhaps a transformation for some of the variables may yield a more accurate prediction in future tests.

	fit	lwr	upr
1	71.463427	70.157637	72.769217
98	6.864220	6.725836	7.002603
175	4.257140	4.110077	4.404204
210	4.140621	4.014602	4.266641
10	26.498910	25.473937	27.523884

We can tell Turtle Games that there is a positive correlation between the different regions with sales and a slight dominance with EU. However, with my current model the prediction will not be accurate enough to give them insights.

Conclusion

In retrospect, I feel we have some strong insights from Turtle Games' data. We have seen the positive correlation between spending, remuneration and loyalty points. We also have found 5 logical customer groups that can be targeted individually by the company. Review analysis showed us that their products are favoured and enjoyed most, and we have looked closer at what products are most and least demanded. Finally, while my regression model may not be best at predicted, we have seen a positive correlation between