DECEMBER 22, 2022

TURTLE GAMES REPORT
CO3_LSE_DA_301 ADVANCED ANALYTICS FOR ORGANISATIONAL IMPACT

TAMAS BALOG (STUDENT)
LSE

1. Introduction

The data includes information of customers and products of Turtle Games, a computer games retailer. The analysis process was highly iterative, every step added more information and allowed to consider new ideas and questions, for example:

Products

- older games have more accumulated sales than new games
- count of platforms for each product or publisher can be relevant
- is there a trend in products by platforms, genre or another variable

Customers

- sales and loyalty points are impacted by customer satisfaction and sentiment, which can be analysed based on reviews?
- why is Turtle Games not mentioned in reviews?
- what is the role of Turtle Games, how do customers purchase products?
- is there correlation between age and remuneration?
- two groups can impact sales, young people play a lot of games and older people with higher salaries can afford more games

Observations:

- product 107, one outlier with great impact
- does it make sense that the sorting of global sales of the company is fully in line with the global ranking of the product
- age starts from 17 only and release year is only until 2016
- missing information like unit price, profit margins, this is important because it can impact the business value of the analysis
- lack of time-series information on sales or customer acquisition and retention

2. Predicting loyalty points

Investigating the possible relationships between the loyalty points, age, remuneration, and spending scores. Different categorical and numerical variables were explored, originally by simple visualisations:

- remuneration should impact the spending of customers
- spending score might use similar inputs like loyalty points
- age could impact any other variable too
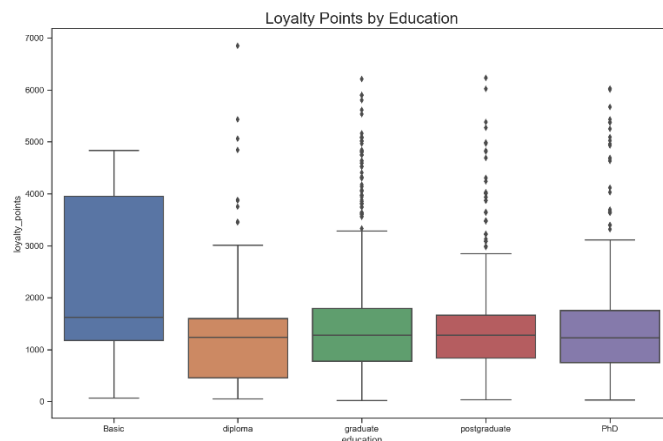
The distribution for the basic category has a clearly higher interquartile range of loyalty points in terms of level of education. On the one hand this can be related to the years you need to achieve a PhD. Perhaps, external comparisons from the industry could be used to look into this. On the other hand, it can indicate a group that is highly engaged with the products like younger customers. In contrast, there is no customer with basic education over 5000.

Then, it can be also expected that there is a relationship between age and remuneration, but the initial visualisation of data is not enough to determine this. Further grouping and filtering should allow to find more details. This level of detail can be seen in the comparison of remuneration and loyalty points, sorted by level of education in a different way as on the following graph:
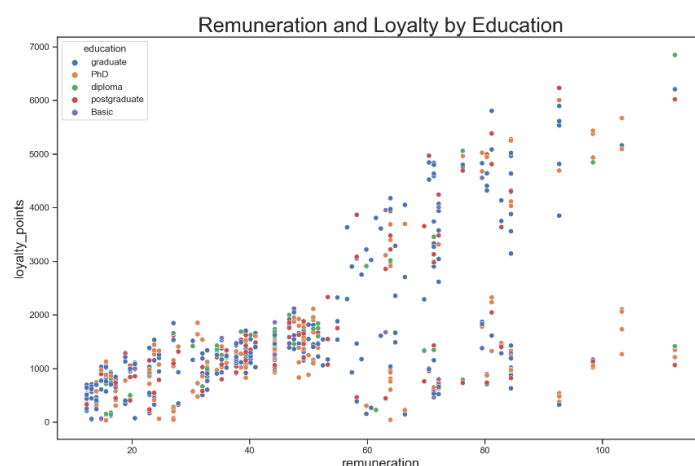
Further depth of analysis is possible with categorical variables, but the initial analysis focused on exploring relationship of loyalty points with the numerical variables. There does not seem to be linear relationship between loyalty points and age. There is approximately linear relationship with the two other variables, remuneration and spending score. This relationship becomes less predictable at the higher values, heteroscedasticity impacts regression analysis with single or multiple variables. Therefore, a nonlinear transformation was used and a model with two variables seems to be a good predictor of the loyalty points. This prediction is not directed it is derived from the square root value of loyalty points which is predicted by the model. The typical range of tests were done to evaluate the model. These proved that the assumptions of the linear regression were met, and metrics do not show unacceptable errors.
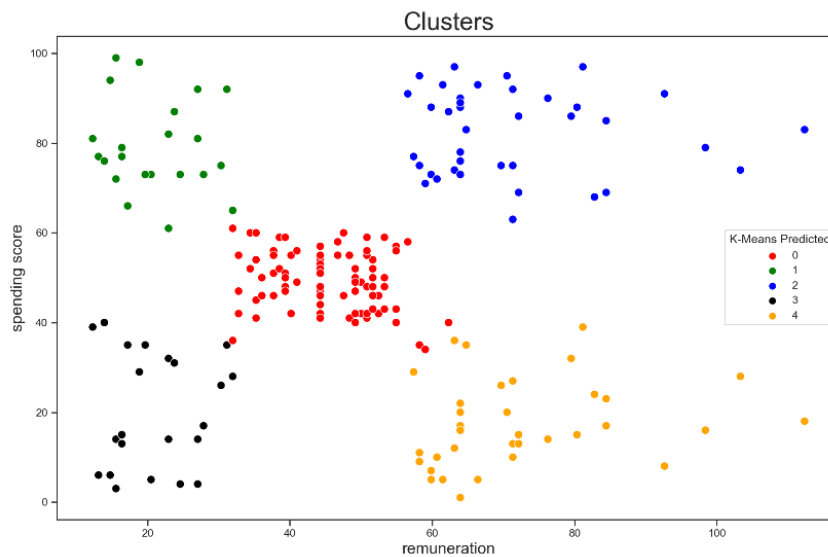
Pairplot, overview of numeric variables:
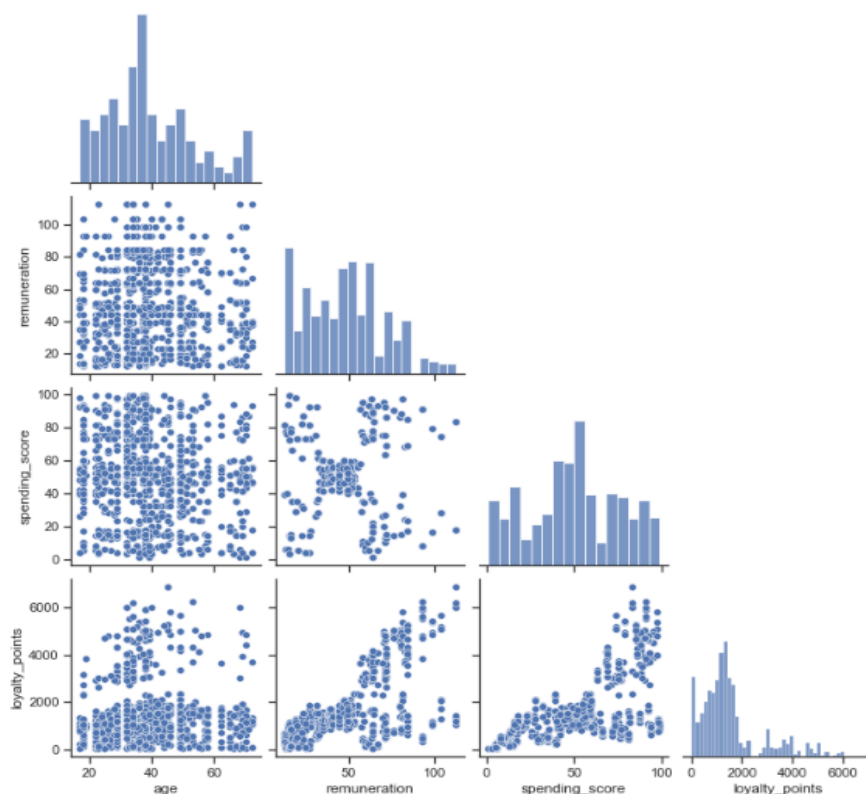
The pairplot shows

- skewness of loyalty points
- that they also have heteroscedastic problem with the two meaningful variables
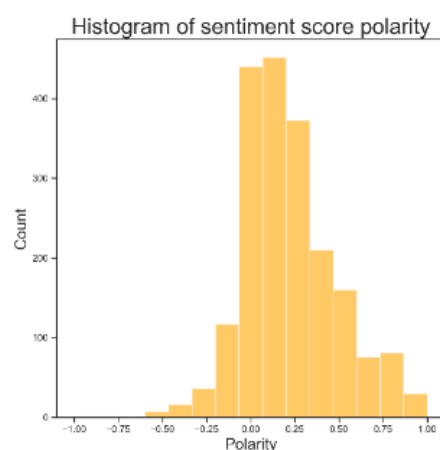
3. Clustering

K-means clustering, a simple and popular unsupervised machine learning algorithm was used to group customer data together based on remuneration and spending score. The five clusters are clearly visible, they are not overlapping, and the five clusters work well with the data. Both the elbow and the silhouette method confirmed the number of clusters and models with different number of clusters were less convincing. The skewness of both variables highlights a group of customers with both high earnings and loyalty points. Analysing this group further can be useful, why this group is different from those with high renumeration and lower loyalty points. Is this related to age and other demographics or product categories? Other clusters can be important and targeted marketing can help the high earning customers with low loyalty points spend more in the future.

### 4. Social data from 2000 reviews on Turtle Games products
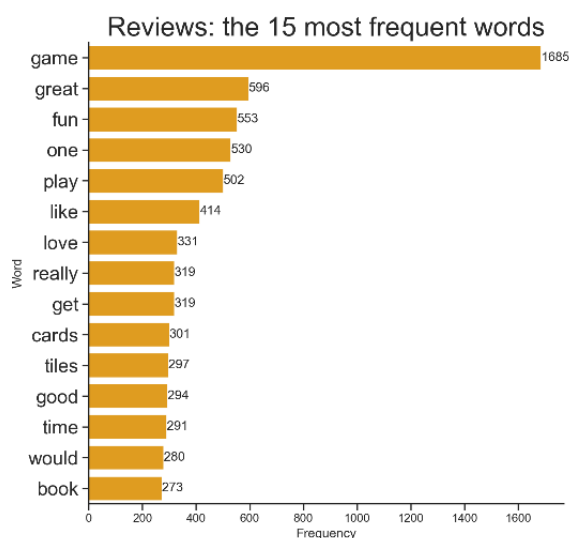
A few basic Natural Language processing tools were used to demonstrate how ML can help analyse customer reviews in Python. I also added a word search code snippet and spent some time exploring if any of the product platforms or publishers were mentioned but I could find no match.

Surprisingly, the brand 'Turtle Games` is not mentioned by customers in any text. Monitoring this could be important to identify positive and negative sentiment. The estimated sentiment score was added to each comment for the analysed words in the review and the summary columns combined. This originally shows a strong positive customer sentiment. Looking into the word clouds or comments with the different sentiment score allows to understand the potential and limitations of this quick and simple way of analysis. It is also important that most of the comments are related to the games themselves not Turtle Games itself.
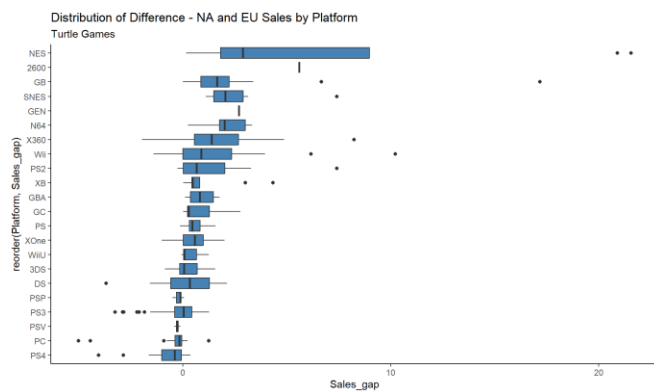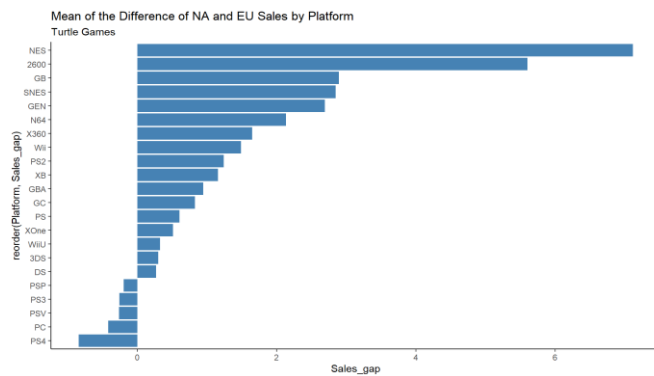
5. Impact of products on sales

5.1. Comparison of sales can be distorted by accumulation of earlier releases, review releases from the last few years accordingly:
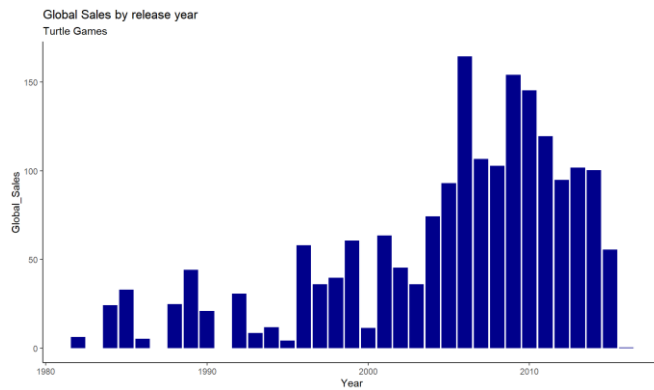

5.2. The simple sales_gap metrics helps to understand different regions and customer preferences by categorical variables like platform:

5.3. Sales figures are aggregated for products and grouped by categories. Overall, does the data show that people just buy the best games wherever they find it? The impact of Turtle Games is key to access how this data can benefit the business?
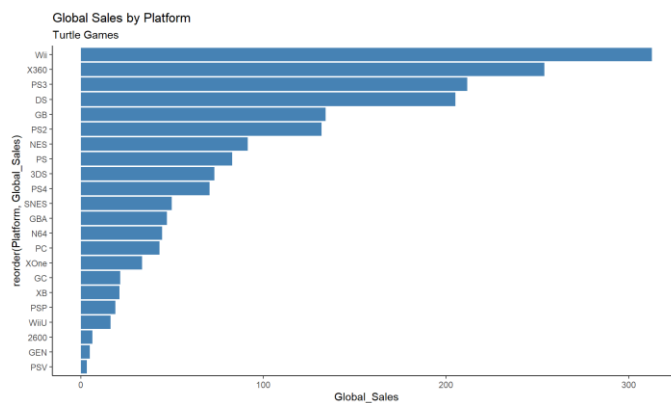
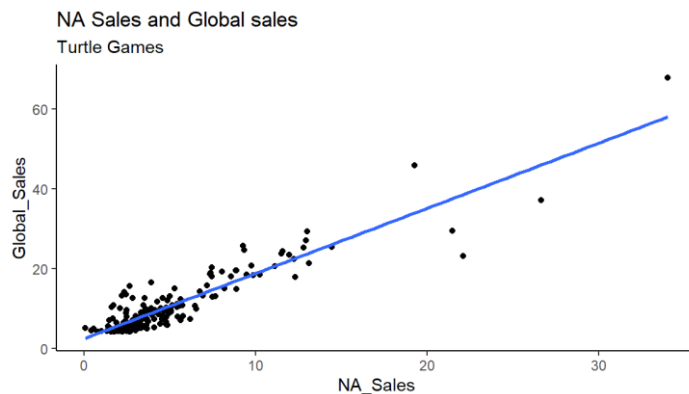5.4. There is heavy exposure to Nintendo sales:
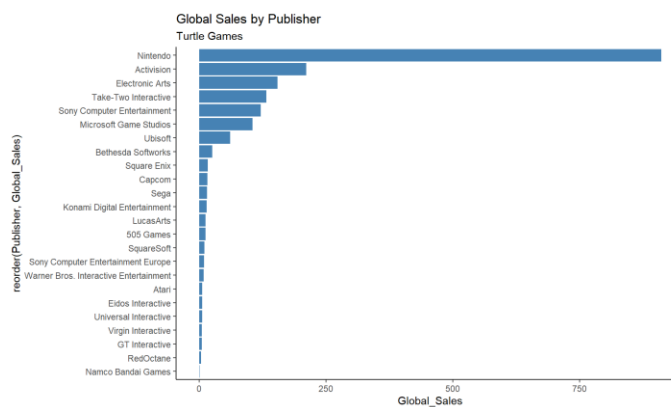
5.5. Pairplot shows strong correlation between regions but also skewed distribution and outliers:
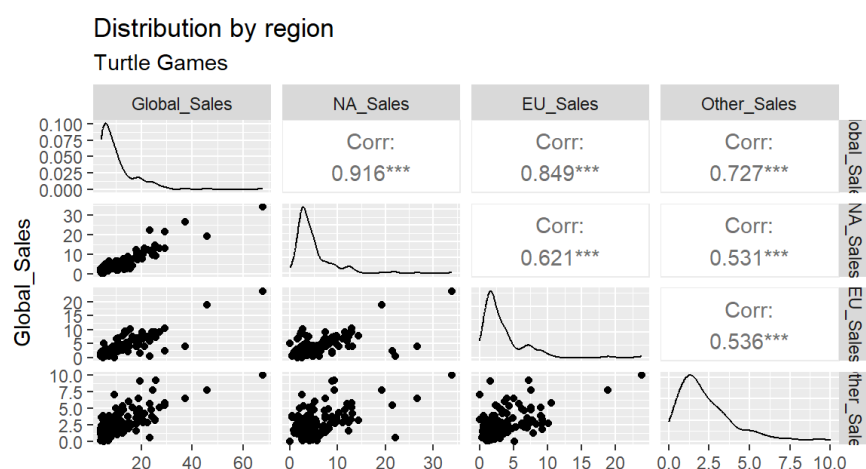
5.6. The detailed charts show linearity and that removing outliers might change line of best fit.

NA Sales and Global sales
Turtle Games

6. About the data

Overall, there were no issues with the tidiness of the data, some minor inconsistencies only that were fixed or do not distort the analysis. On the other hand, neutralising the impact of the outliers should improve the quality of the analysis. Detailed observations have been added to both assignment files.

7. Recommendations

First of all, monitoring social media can add significant value allowing to understand and react to changes in customer behaviour. On the other hand, advanced models of NLP can be costly and the potential business impact in relation to the company profile must be explored with stakeholders first. Further analysis should be directed towards the ways Turtle Game creates value with its services. Based on that, advanced ML algorithms and predictive analysis can support both sales and marketing and propagate brand identity.

Additionally, it must be checked what secures future Nintendo sales: 17 of top 20 games