Context of the business scenario (100)

Turtle Games sells to its global customer base both its own product range (including books, board games, video games, toys) and products manufactured by others. The objective of the analysis is to provide insights and predictions on how to boost sales performance using customer trends.

The data provided for the analysis are customers' reviews and sales volumes by region; a first observation on the datasets suggests customers' reviews are missing the customer's geographical area. Having this level of granularity would make it possible to conduct sentiment analysis on a regional level.

Global sales are made by three main geographical areas (North America, Europe, Rest of the world). Although North America seems the more significant market, a further level of granularity on the rest of the world (i.e. MENA and Asia) would allow identifying some region-specific trends to consider for market expansion.

In terms of product, the data available only seem to be related to video games, which I assume are the company's core business. There is no indication of sales performances for other products, especially for the company own's products, so I will limit providing insights on the available data.

Analytical approach (350)

My approach has used two different techniques for the two main areas of analysis:

1. Customer insights (Python)
    a. Analysis of the correlation between customer loyalty and other variables
        - Import and exploration of the data on a Pandas dataframe, understanding shape, size and structure of the data
        - Sense-check the database for data types and null values
        - Initial data wrangling (drop of the non-relevant columns and sensible re-naming)
        - Visualisation of the dataset with scatterplot (Matplotlib) to check for linearity of:
            - Spending vs Loyalty
            - Remuneration vs Loyalty
            - Age vs Loyalty
        - Fitting of regression with OLS models, generating regression tables to establish the lines of best fit (for each correlation)
        - From the result of the initial analysis, exploration of the correlation Remuneration vs Spending (as both are correlated to Loyalty):
            - Visualisation of the data with scatter plot and pairplot (Seaborn)
            - K-means clustering using Elbow and Silhouette methods
            - Evaluation of k values (4,5,6) and final fit with k=5
    b. NLP and sentiment analysis of customers' reviews
        - Import of the clean database (from initial wrangling)
        - Drop irrelevant columns (for this analysis I kept only the reviews and summary data)

- Removal of capitalisation (transforming all strings to lowercase with lambda function) and punctuation
- Identification and removal of duplicate entries (and re-indexing of database)
- Tokenisation (with loop function) of the two columns distinctively and creation/plot of the preliminary correspondent wordclouds
- Determination of the frequency distributions of the tokens for both columns
- Joining of the token lists and removal of alphanumeric values and stopwords (based on English language)
- Creation and plot of the final wordcloud
- Identification of the 15 more common words
- Visualisation of Polarity scores for both reviews and summary
- Sentiment analysis: top 20 positive and negative reviews and summary

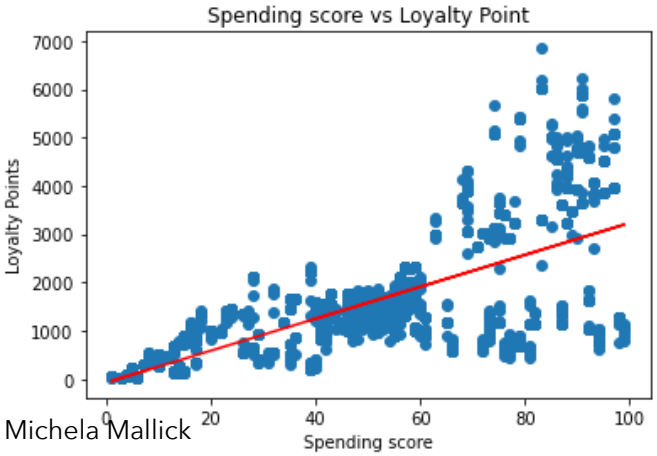
2. Sales insights (R)
   a. Imported data in R and subset the dataframe keeping only the sales columns, product and platform
   b. Created a column for the rest of the world sales (ROW = Global Sales – NA – EU) to observe the data between regions
   c. Visualisations of the dataset to gather insights:
      - Scatterplots capturing the correlation of sales between different regions
      - Histogram of product by platform
      - Boxplots of regional sales
   d. Top 10 product per region: insights
   e. Checks of the normality of the data set: Q-Q Plots, Shapiro-Wilk test , Skewness and Kurtosis
   f. Plot of the data to gather insights on correlation
   g. Linear and multiple linear regression models to provide recommendations

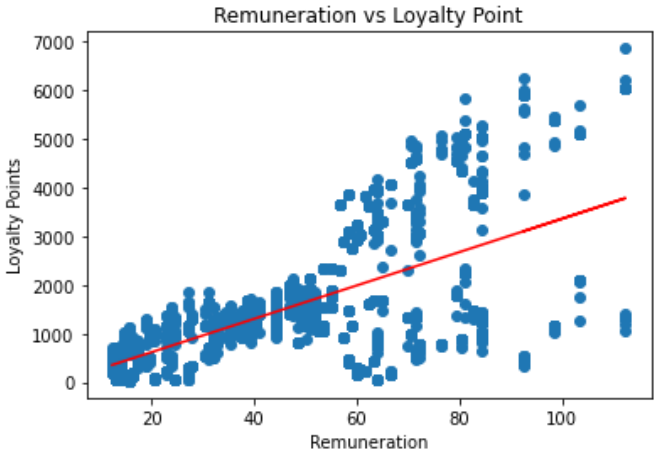
Visualisation and insights (350)

In terms of visualisations and insights, I'll follow the analysis structure;

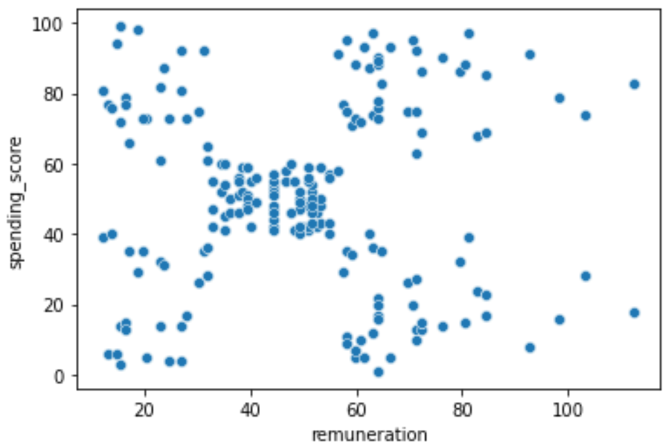
Customer insights:

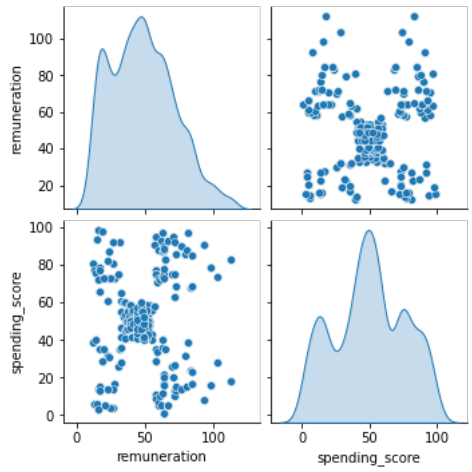
- Scatterplots visualisations of Spending score vs Loyalty Point and Remuneration vs Loyalty Points suggest a correlation:



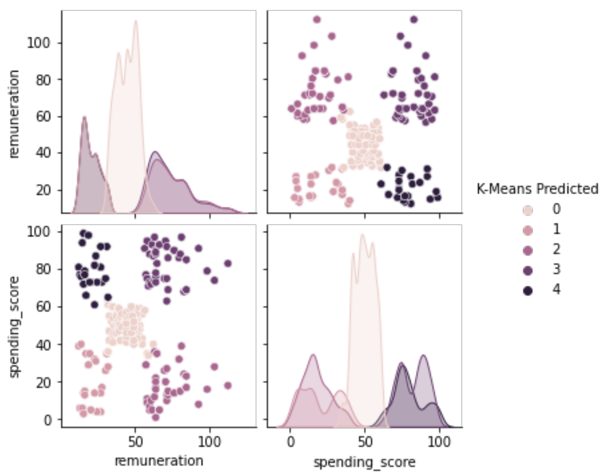


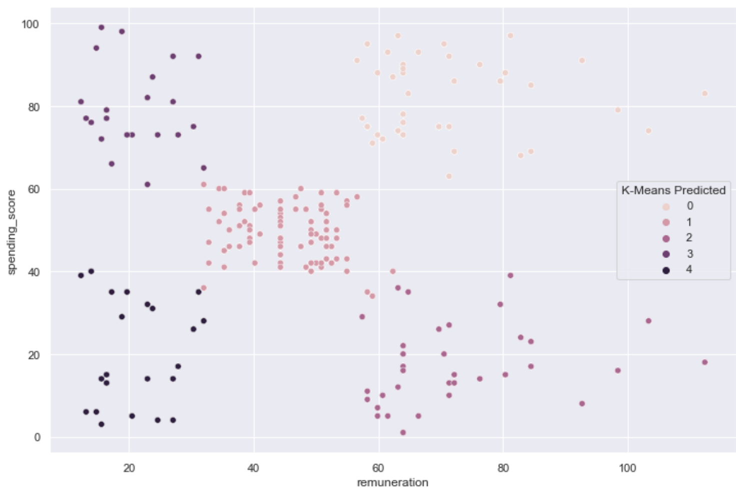
- This correlation prompted me to also evaluate the correlation between spending score and remuneration:



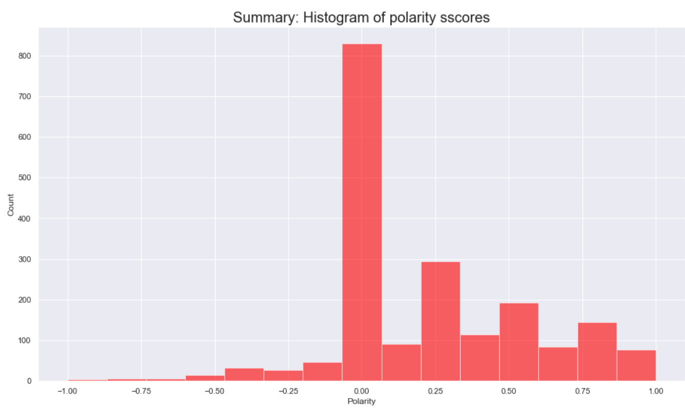


The scatterplot suggests that the client base could be divided into 5 distinct clusters. Aside from cluster 0, which is bigger in size, the other 4 clusters are similar in size. Further exploration could certainly provide valuable insights into effective customer segmentation.



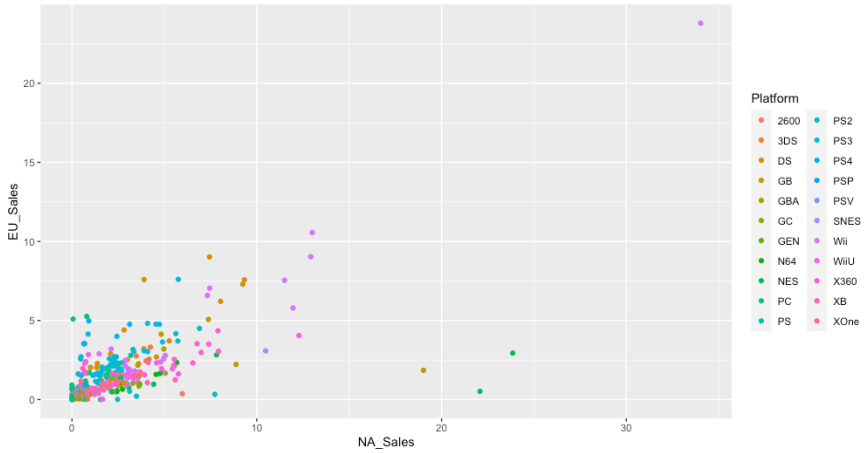


- In terms of NLP, the analysis

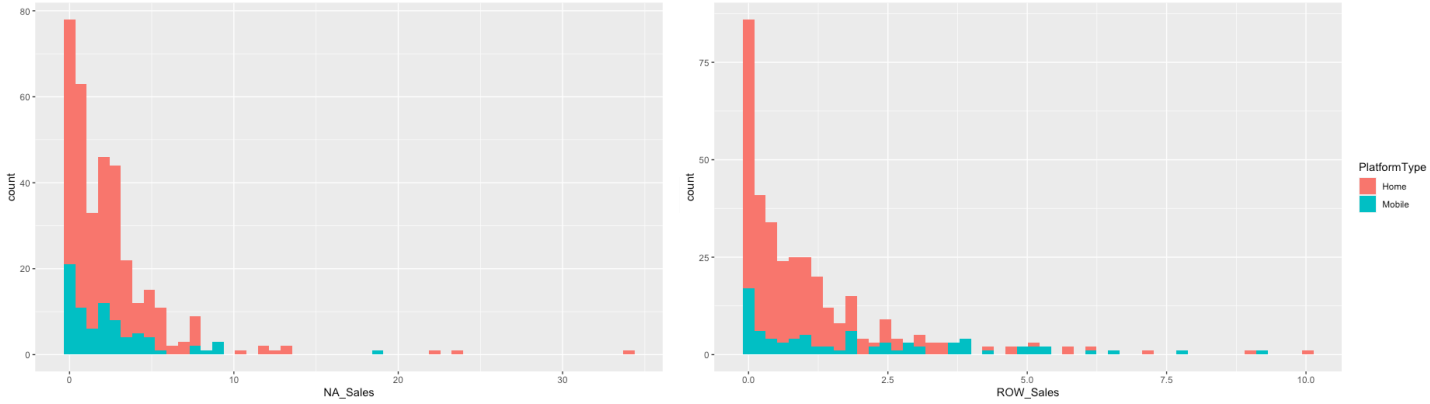


Predicting future outcomes

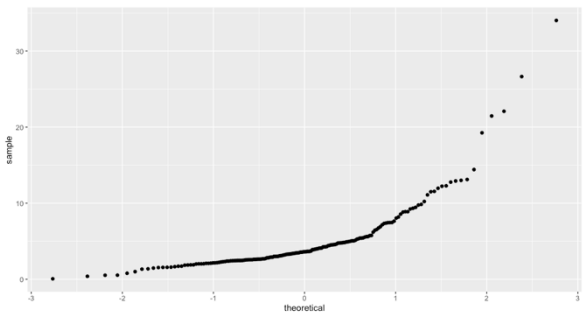
- I discovered that sales between regions were correlated, implying that products that succeeded in one region would be likely to succeed in others. In particular, North America has a strong correlation with Global Sales since it is the biggest regional market for Turtle Games. From the color fill of the scatterplot (Platform) we can also note another interesting insight: the correlations appear stronger for products within the same platform



- Exploring further the Platform element of the analysis, it is interesting to notice that in the Rest of the World data there is a discrepancy between console and mobile devices, where the latter seem to account for a higher number of sales. This aspect could be potentially explored further to focus marketing and developing resources in this region.



- As we can see from the above histograms, the sales data is also very highly right skewed and we confirmed this when aggregating sales by product across platforms for each region by conducting a number of statistical tests




```
        Shapiro-Wilk normality test

data:  sales_sub_product$NA_Sales
W = 0.69813, p-value < 2.2e-16
```

```
> # Skewness and Kurtosis.
> skewness(sales_sub_product[2:length(sales_sub_product)])
    NA_Sales     EU_Sales    ROW_Sales Global_Sales
    3.048198     2.886029     1.625362     3.066769

> kurtosis(sales_sub_product[2:length(sales_sub_product)])
    NA_Sales     EU_Sales    ROW_Sales Global_Sales
   15.602601    16.225544     6.078014    17.790717
```

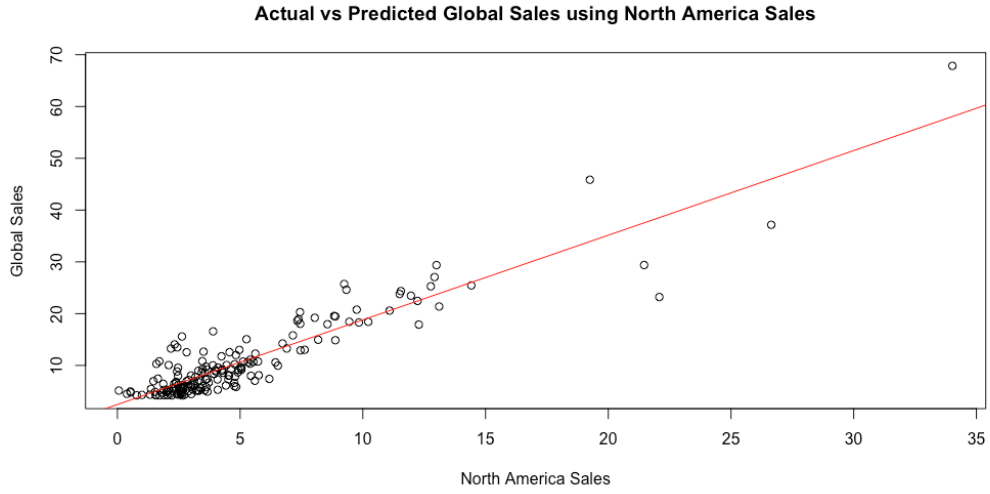
Predicting future outcomes

These tests results also revealed that the North America and EU region had sales even more highly right skewed than the rest of the world, indicating that commercial success in these regions has been carried by a few of the most popular products

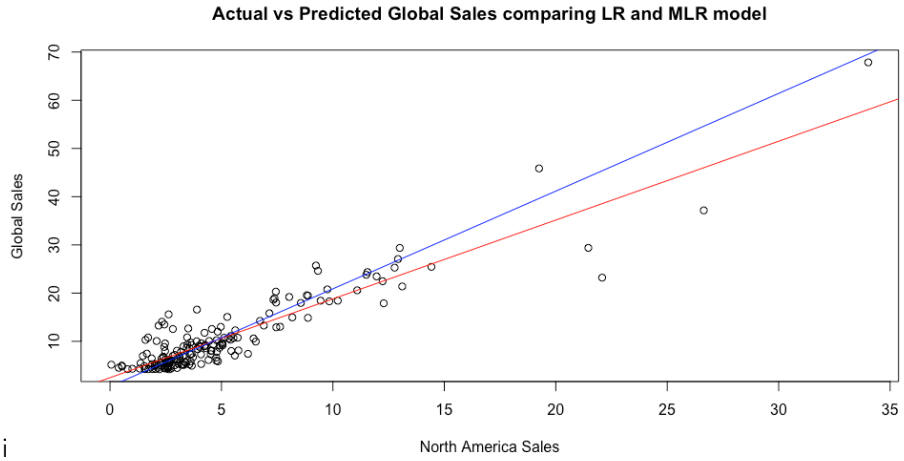
- To follow up on this finding, we explored the sales of the top 10 products per region also finding that three product seem to in the top 10 in all regions

	Product	NA_Sales	EU_Sales	ROW_Sales	Global_Sales	Percentage
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	107	34.0	23.8	10.0	67.8	3.61
2	195	13	10.6	5.81	29.4	1.56
3	515	19.2	18.9	7.73	45.9	2.44

- As North America was not only the biggest region but also the one with the strongest correlation with Global Sales, the linear regression model fit using North America sales to predict Global Sales achieved an adjusted R-Squared of 84%; this implies that we can explain the variance of Global Sales predominantly using just North America sales.



- I was able to improve significantly (R squared from 84% to 97%) the model that predicted Global Sales by making it into a multi-linear regression model (MLR) by incorporating EU Sales as a factor



Blue line represents the predictions from the MLR model while
Red line represents the predictions from the original LR model
As we can see, the blue line corrects a downward bias from the LR model (underestimate Global Sales)

Patterns and predictions

As a summary of the previous insights and recommendations:

- There is a correlation between loyalty points and both spending scores and remuneration. Using the correlation between spending score and remuneration, it is possible to cluster the customer base in 5 segments. Further explorations are required to better understand the characteristics of the clusters and target them accordingly
- The NLP analysis shows that the polarity scores are predominantly neutral with a right skew into the positive; in order to have better insights also related to the sales data, it would be useful to have these data with a regional granularity in order to inform marketing of regional peculiarities
- Product sales are very unevenly distributed between products reflecting a right-skewed distribution of product sales; a few products contribute to a disproportionately high number of sales.
- There is a differentiated distribution of sales by platform type (comparing ROW vs NA and EU). This also could suggest that a further analysis considering genre and product type by region could provide significant insights.
- There is a partial correlation between product sales in each of the regions. Overall Global sales are highly correlated with NA Sales. As a point of further exploration, we can test if the correlations of sales between regions become stronger when we focus on particular platforms.