Background

Turtle Games has a business objective of improving overall sales performance by utilising customer trends. With this broad business objective in mind, this analysis is attempting to help Turtle Games understand their customer base and revenue sources by region, using the social and sales data available. Turtle Games have raised the below questions specifically and we will answer them throughout this report:

- how customers accumulate loyalty points
- how groups within the customer base can be used to target specific market segments
- how social data (e.g. customer reviews) can be used to inform marketing campaigns
- the impact that each product has on sales
- how reliable the data is (e.g. normal distribution, skewness, or kurtosis)
- what the relationship(s) is/are (if any) between North American, European, and global sales?

Analytical approach, visualisation and insights

Question 1: How do loyalty points correlate with customers' characteristics?

Analytical Approach: to answer this question, I used Python to clean, wrangle the customer review data ('turtle_reviews'), visualised the relationships between 'loyalty_points', 'age', 'remuneration', 'spending_score', used linear regression models to check the statistical significance of the relationships. I utilised the relevant data and statistical and visualisation packages including numpy, pandas, pylab, sklearn, scipy, statsmodels, seaborn, matplot.

Visualisation and insights

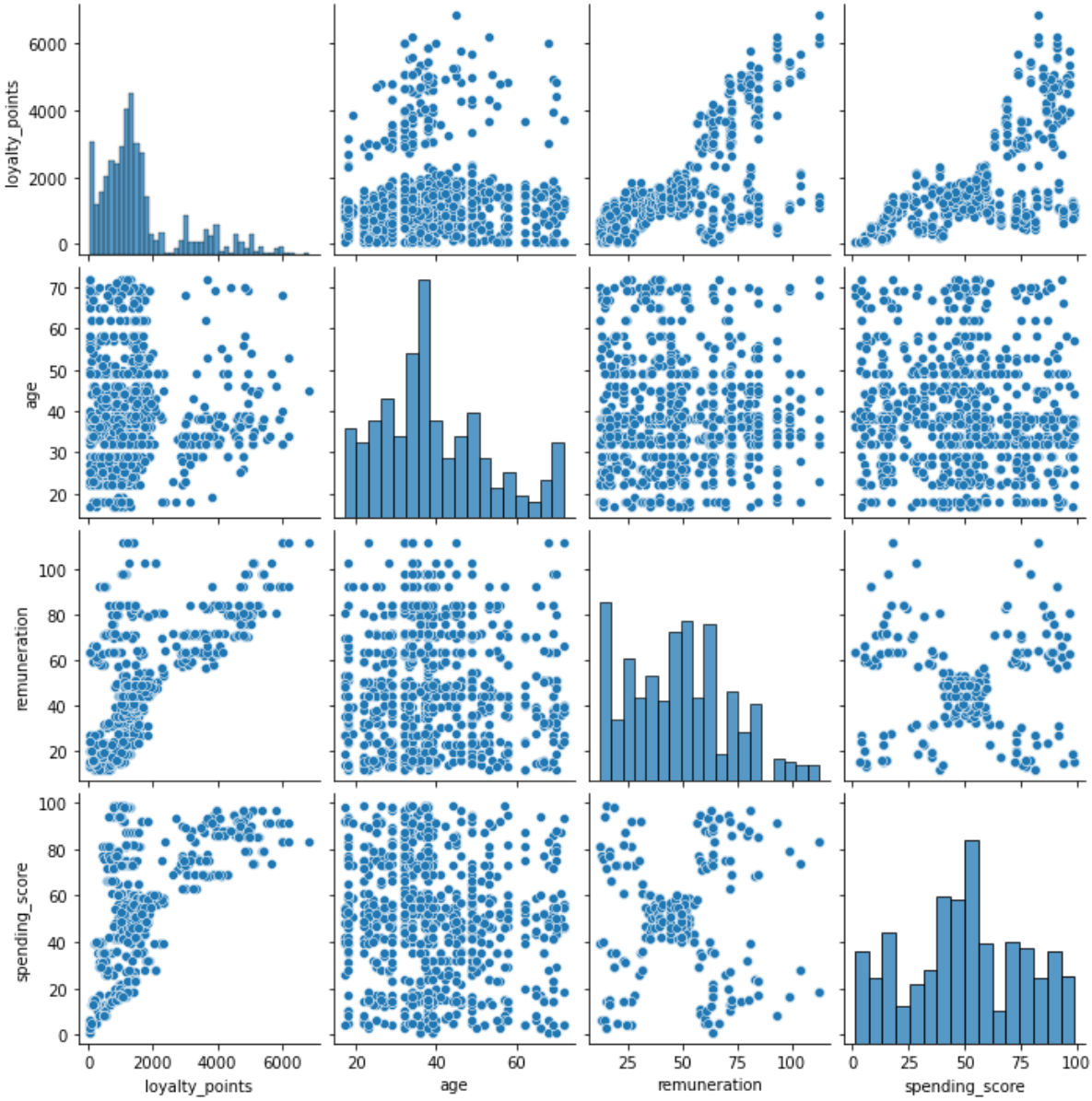
In plot 1 below I visualised the relationship between loyalty_points and age/remuneration/spending scores BEFORE running the regressions. This would already give us a lot of insights and help determine the right models to use, in my view.

Visual Observation 1: 'loyalty_points' seems to have a stronger correlation with 'remuneration','spending_score'. But the graphs suggest potential heteroskedasticity. Its relationship with 'age' looks weak.

Visual Observation 2: There doesn't appear to be linear relationship between the independent variables 'age','remuneration' and 'spending_score', which suggests multicollinearity might not be an issue.

Visual Observation 3: However, there is an obvious cluster in the middle looking at the plot of 'remuneration' vs 'spending_score'.

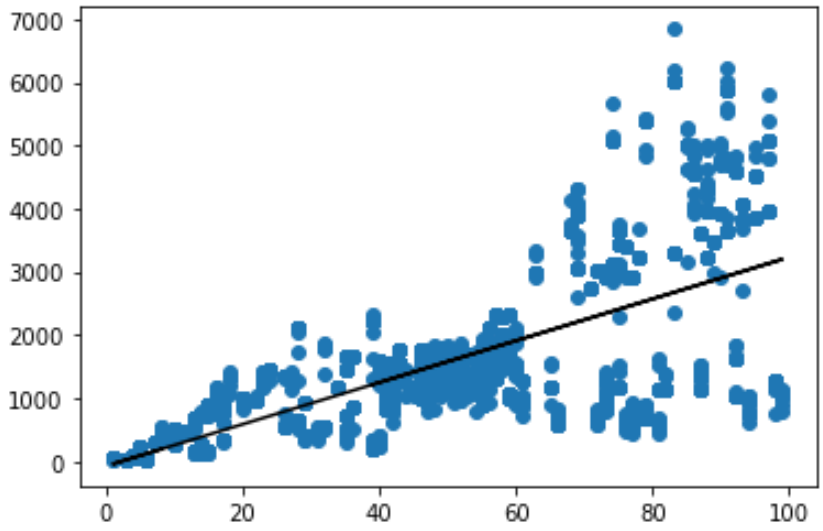
Plot 1. Pair-wise relationships of the variables



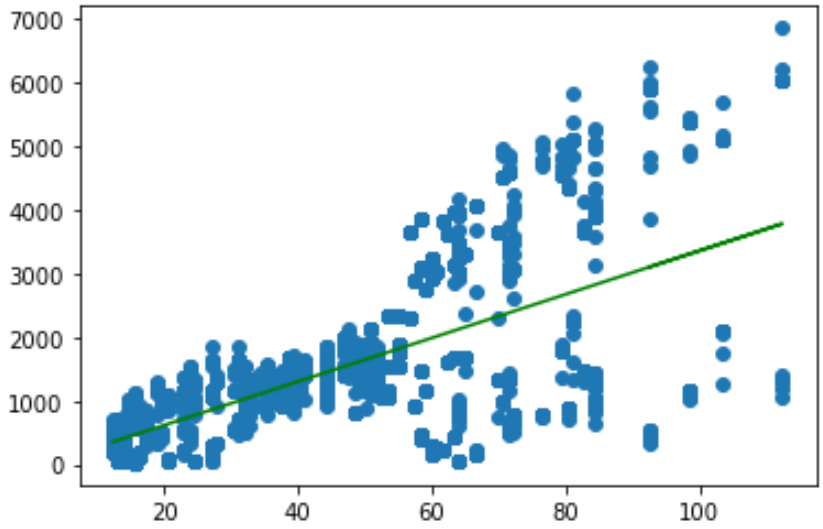
Simple Linear Regression Model

In the Python script, I have provided the linear regression model results and here is a summary of the visualisations (plots 2-4) and insights. Note in each model/plot, spending/remuneration/age are the x variables.

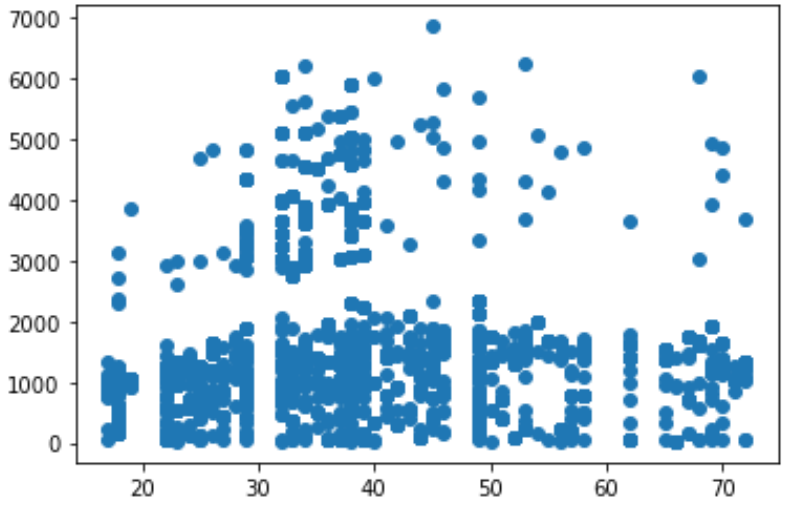
Plot 2. Spending and Loyalty Points plot



Plot 3. Remuneration and Loyalty Points



Plot 4. Age and Loyalty Points



Insights from the simple linear models and plots:

1. The model and coefficient are statistically significant. However, there are signs of heteroscedasticity from the plots.
2. In the age vs loyalty model, p-values of F-stat and x coef are both > 0.05. Age doesn't seem to be a good explanatory variable for loyalty in a linear way.
3. However, the models had low R-sq individually. I decided to try a multiple linear regression with both spending and remueration as independent variables.

Multiple linear regression

Insights from the multiple linear regression model:

1. The results look good with p-values of F-stat and both independent variables < 0.05, and R-sq at 82.7%.

2. However in the heteroscedasticity tests, both p-values are extremely low, suggesting heteroscedasticity or a flawed model.

3. This confirms what we observed in the plots early on - the variance of loyalty increases as spending or remuneration increases. Considering the above, some further treatments of the variables may be needed.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.827
Model:	OLS	**Adj. R-squared:**	0.827
Method:	Least Squares	**F-statistic:**	4770.
Date:	Fri, 03 Mar 2023	**Prob (F-statistic):**	0.00
Time:	21:45:49	**Log-Likelihood:**	-15398.
No. Observations:	2000	**AIC:**	3.080e+04
Df Residuals:	1997	**BIC:**	3.082e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1700.3051	35.740	-47.575	0.000	-1770.396	-1630.214
remuneration	33.9795	0.517	65.769	0.000	32.966	34.993
spending_score	32.8927	0.458	71.845	0.000	31.995	33.791

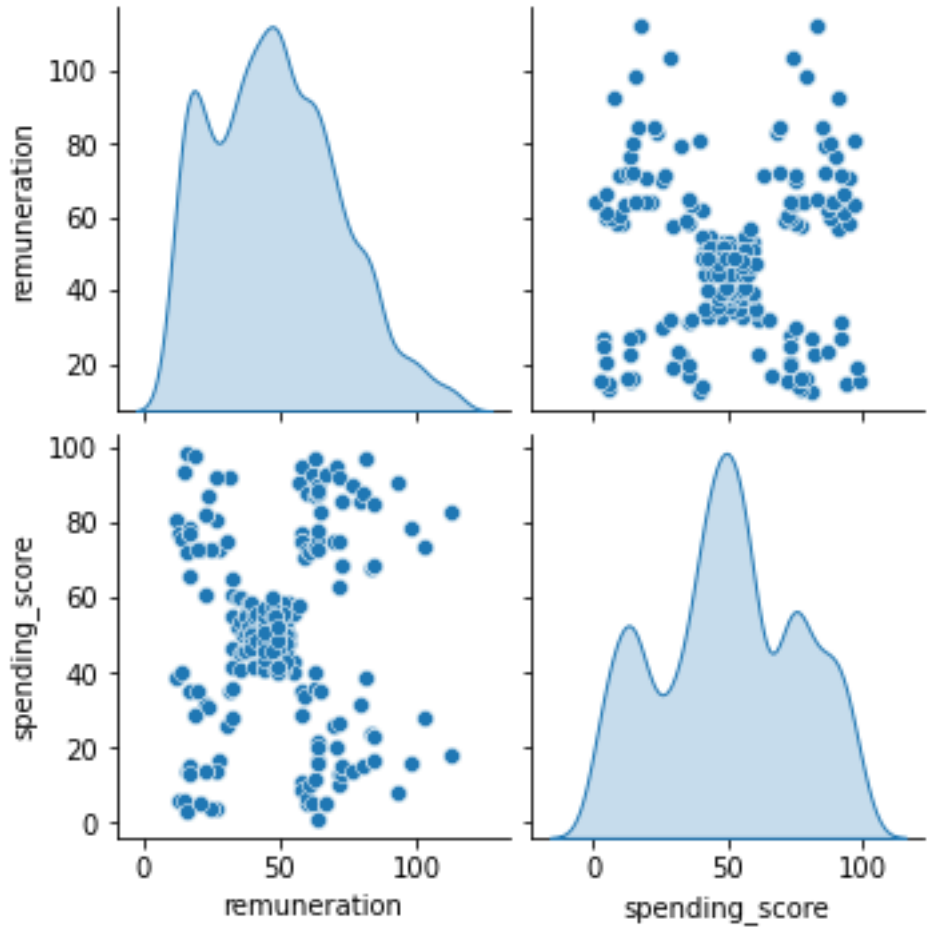
Omnibus:	4.723	**Durbin-Watson:**	3.477
Prob(Omnibus):	0.094	**Jarque-Bera (JB):**	4.650
Skew:	0.103	**Prob(JB):**	0.0978
Kurtosis:	3.115	**Cond. No.**	220.

Question 2: Customer groups and market segments

Analytical approach: to identify groups within the customer base that can be used to target specific market segments, I used k-means clustering to identify the optimal number of clusters and then apply and plot the data using the created segments. The two characteristics I used to create the groups are remuneration and spending score. To determine the optimal number of clusters, we used the Silhouette and Elbow methods. Finally, I fitted a final model using the selected number of clusters (k) and checked the predictive power of the model. These were all done in Python with the help of relevant libraries and packages such as 'kmeans', 'silhouette_score' from Sklearn.

Visualisation and insights

Plot 5. Visualising remuneration vs spending score

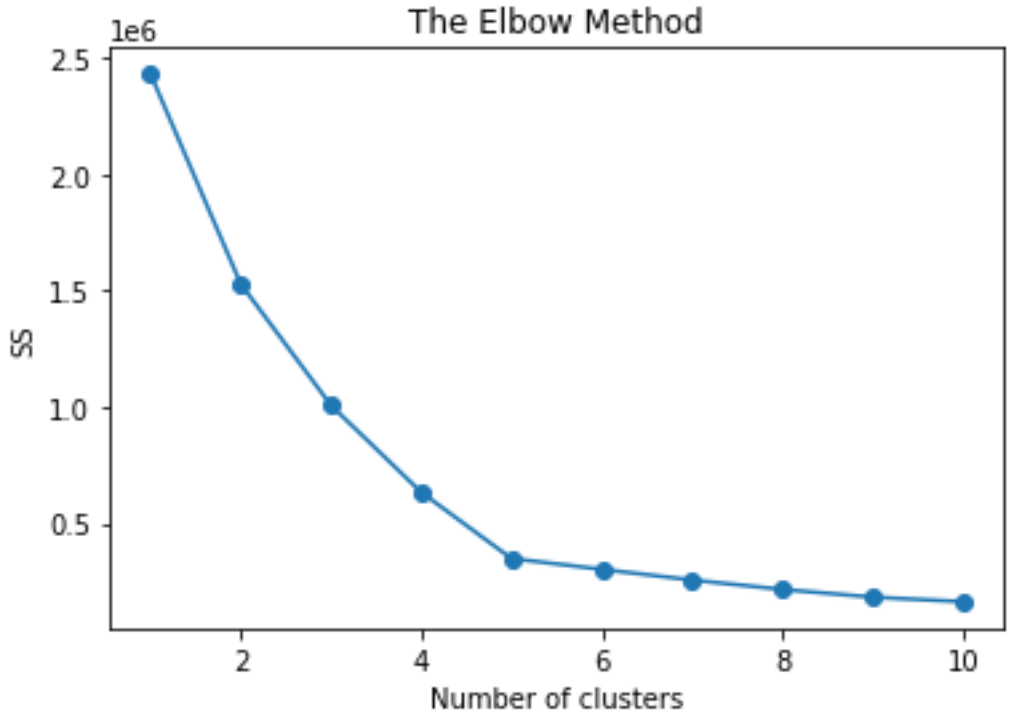


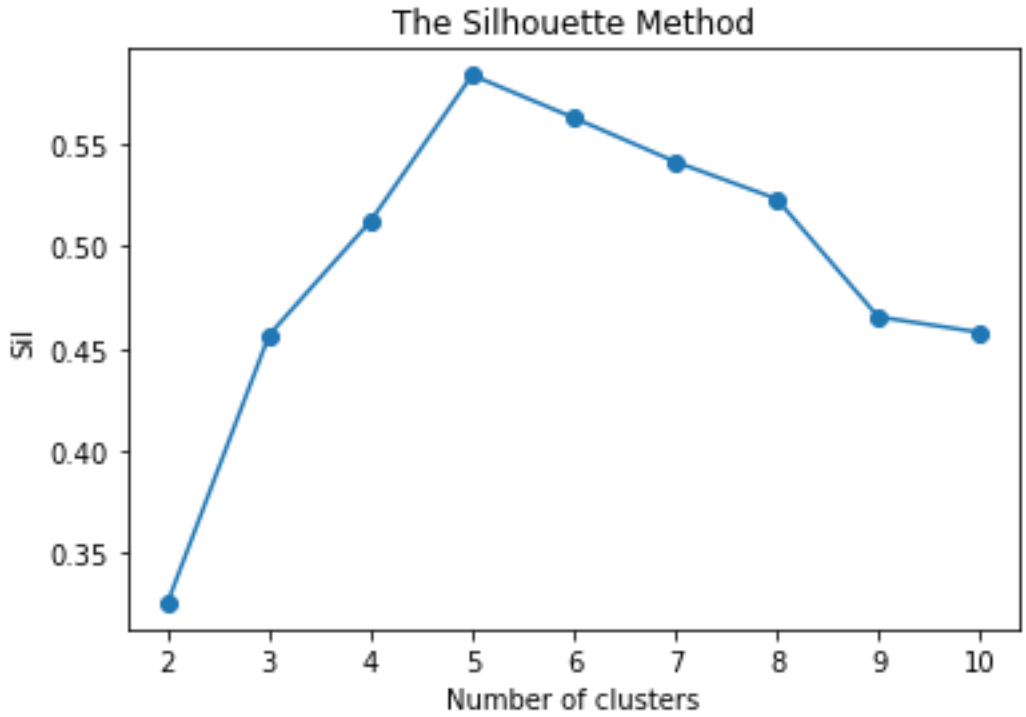
Modelling

Deciding the optimal number of clusters (k)

Both methods (see below) indicate 5 would be the optimal number of clusters.

Plot 6. Methods of deciding the optimal k value

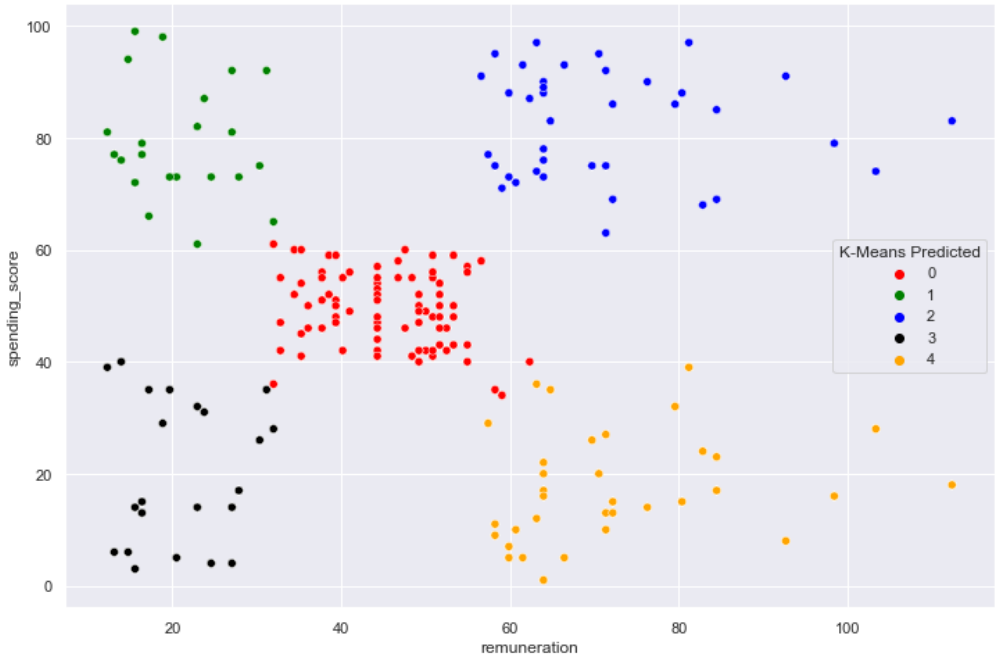




Evaluating k-means model at different values of k

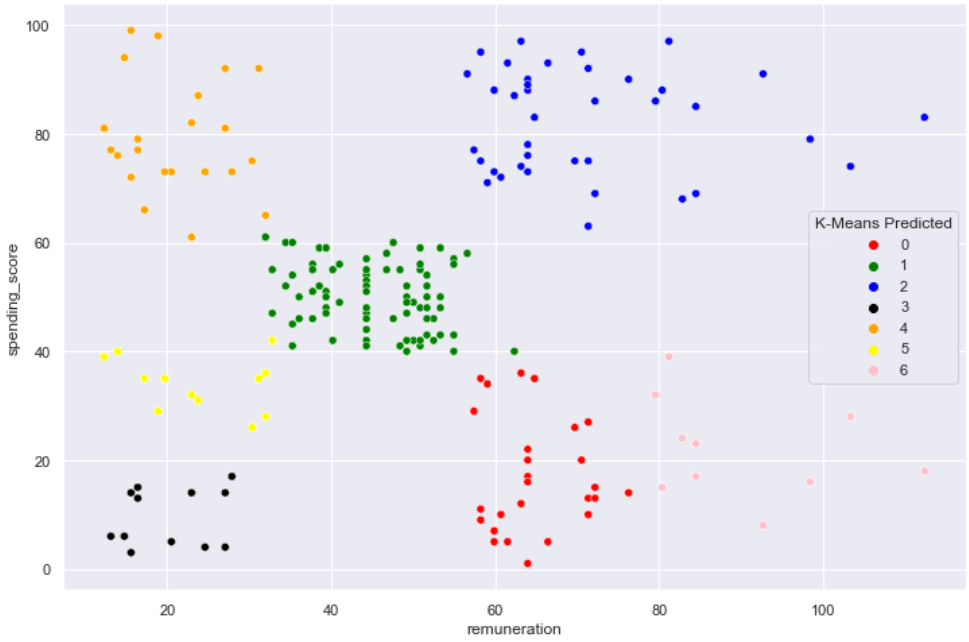
I used 5 clusters to model the groups and here is the plot of the predicted groups.

Plot 7. Plot of predicted groups using k=5



As a comparison, I also modelled using k=7 i.e. 7 clusters, and plotted the predicted groups. The reason I chose 7 was because class 2 and 4 above seem to have obvious outliners compared to other classes.

Plot 8. K=7



Looking at the results, I would choose 5 clusters to fit the final model. 7 clusters reduce observation size of class 0/3/5/6, and we run a higher risk of over-fitting.

In conclusion, the elbow and the silhouette methods both suggest k=5 seems to be the optimal. The number of predicted values per class indicates a better distribution for k=5 than k=7.

The clusters suggest we can group the customers into five groups based on the spending and pay patterns. Therefore, I would recommend grouping the customers based on their remuneration and spending score into 5 groups:

1. low pay and low spending

2. low pay and high spending

3. Medium pay and medium spending

4. High pay and high spending

5. High pay and low spending

We can analyse each group separately to see if there are clear patterns in other characteristics within groups and create targeted marketing strategy.

Question 3: Customer reviews and marketing campaign implications

Analytical approach: The marketing department wanted to identify the 15 most common words used in online product reviews and a list of the top 20 positive and negative reviews received from the website. I applied NLP on the data set using Python to produce the lists. I prepared the data (changed case, removed punctuations, dropped duplicates etc), tokenised the words, removed alphanumeric characters and stopwords, reviewed the frequency of words and their polarity and sentiment. The additional libraries needed were wordcloud, nltk, textblob and scipy.

Visualisations and insights

Plot 9. Wordclouds – Summary column

Plot 10. Wordclouds – Review column

Table 1 & 2. The top 15 common words in summaries and reviews

Summary, Word	Frequency
stars	427
five	342
game	319
great	295
fun	218
love	93
good	92
four	58
like	54
expansion	52
kids	50
cute	45
book	43
one	38
awesome	36

Review, Word	Frequency
game	1671
great	580
fun	552
one	530
play	502
like	414
love	323
really	319
get	319
cards	301
tiles	297
time	291
good	289
would	280
book	273

Table 3 & 4. The top 20 negative and positive reviews.

	neg	neu	pos	compound	polarity
difficult	1.000	0.000	0.000	-0.3612	-0.500000
crappy cardboard ghost original hard believe shame hasbro disgusting	0.723	0.157	0.120	-0.8885	-0.305556
incomplete kit disappointing	0.615	0.385	0.000	-0.4939	-0.600000
found directions difficult	0.556	0.444	0.000	-0.3612	-0.500000
rather hard year old alone	0.531	0.469	0.000	-0.3400	-0.095833
yearold granddaughter frustrated discouraged attempting craft definitely young child difficulty understanding directions disappointed	0.520	0.359	0.121	-0.8360	-0.450000
got product damaged condition	0.492	0.508	0.000	-0.4404	0.000000
uno questions anger okay way discuss anger gets repetitive students start get bored half round	0.491	0.427	0.081	-0.8625	-0.288095
doesnt love puppies great instructions pictures fun	0.477	0.286	0.236	-0.5207	0.533333
bought thinking would really fun disappointed really messy isnt nearly easy seems also glue useless year old instructions difficult	0.455	0.427	0.118	-0.8513	-0.159524
im sorry find product boring frank juvenile	0.419	0.581	0.000	-0.3818	-0.583333
hard put together	0.412	0.588	0.000	-0.1027	-0.291667
hard complicated make	0.412	0.588	0.000	-0.1027	-0.395833
flimsy get pay	0.412	0.588	0.000	-0.1027	0.000000
fun way kids talk anger identify okay feel angry	0.407	0.291	0.302	-0.4215	-0.100000
review previous screens completely unnecessary nearly useless skip definition waste money	0.394	0.606	0.000	-0.7063	-0.316667
im ot intended using aggressive kiddos used even frustrated repetitive asking questions different format	0.389	0.611	0.000	-0.7184	-0.316667
cute idea horrible execution want child tears book seven year old got frustrated whole thing	0.381	0.433	0.186	-0.6705	-0.180000
game hard frustrating first fun get hang	0.371	0.345	0.284	0.0000	-0.108333
smaller thought kind disappointed	0.365	0.235	0.400	0.0772	-0.050000

	neg	neu	pos	compound	polarity
liked	0.0	0.0	1.0	0.4215	0.600000
great	0.0	0.0	1.0	0.6249	0.800000
satisfied thanks	0.0	0.0	1.0	0.6908	0.350000
awesome gift	0.0	0.0	1.0	0.7906	1.000000
satisfied	0.0	0.0	1.0	0.4215	0.500000
super fun	0.0	0.0	1.0	0.8020	0.316667
love	0.0	0.0	1.0	0.6369	0.500000
nice	0.0	0.0	1.0	0.4215	0.600000
cool	0.0	0.0	1.0	0.3182	0.350000
fun	0.0	0.0	1.0	0.5106	0.300000
good	0.0	0.0	1.0	0.4404	0.700000
fine	0.0	0.0	1.0	0.2023	0.416667
fun enjoyable	0.0	0.0	1.0	0.7351	0.400000
loved loved loved	0.0	0.0	1.0	0.9136	0.700000
fun entertaining	0.0	0.0	1.0	0.7351	0.400000
super cute	0.0	0.0	1.0	0.7845	0.416667
cute	0.0	0.0	1.0	0.4588	0.500000
inspiring creativity	0.0	0.0	1.0	0.6597	0.500000
love helpful	0.0	0.0	1.0	0.7906	0.500000
ok	0.0	0.0	1.0	0.2960	0.500000

Table 5 & 6. The top 20 negative and positive summaries

	neg	neu	pos	compound	polarity
frustrating	1.000	0.000	0.00	-0.4404	-0.400000
disappointed	1.000	0.000	0.00	-0.4767	-0.750000
meh	1.000	0.000	0.00	-0.0772	0.000000
boring	1.000	0.000	0.00	-0.3182	-1.000000
disappointing	1.000	0.000	0.00	-0.4939	-0.600000
worn	1.000	0.000	0.00	-0.2960	0.000000
defective poor qc	0.857	0.143	0.00	-0.7184	-0.400000
sided die	0.796	0.204	0.00	-0.5994	0.000000
bad expecting	0.778	0.222	0.00	-0.5423	-0.700000
uno angry	0.767	0.233	0.00	-0.5106	-0.500000
mad dragon	0.762	0.238	0.00	-0.4939	-0.625000
weak game	0.744	0.256	0.00	-0.4404	-0.387500
damaged product	0.744	0.256	0.00	-0.4404	0.000000
sadly cheap	0.737	0.263	0.00	-0.4215	0.400000
bad set limited applicability	0.730	0.270	0.00	-0.6597	-0.385714
crappy cardboard ghost original hard believe shame hasbro disgusting	0.723	0.157	0.12	-0.8885	-0.305556
students fight	0.722	0.278	0.00	-0.3818	0.000000
faulty product	0.697	0.303	0.00	-0.3182	0.000000
small boring	0.697	0.303	0.00	-0.3182	-0.625000
money trap	0.697	0.303	0.00	-0.3182	0.000000

	neg	neu	pos	compound	polarity
fun useful	0.0	0.0	1.0	0.7351	0.30
good	0.0	0.0	1.0	0.4404	0.70
nice	0.0	0.0	1.0	0.4215	0.60
fun cute	0.0	0.0	1.0	0.7430	0.40
ok great	0.0	0.0	1.0	0.7430	0.65
perfect gift	0.0	0.0	1.0	0.7650	1.00
great	0.0	0.0	1.0	0.6249	0.80
ok best	0.0	0.0	1.0	0.7506	0.75
fun	0.0	0.0	1.0	0.5106	0.30
great gift	0.0	0.0	1.0	0.7906	0.80
great helper	0.0	0.0	1.0	0.7579	0.80
cute	0.0	0.0	1.0	0.4588	0.50
ok ok	0.0	0.0	1.0	0.5267	0.50
happy gift	0.0	0.0	1.0	0.7650	0.80
love play friends	0.0	0.0	1.0	0.8658	0.50
satisfied	0.0	0.0	1.0	0.4215	0.50
good fun	0.0	0.0	1.0	0.7351	0.50
fun fun fun	0.0	0.0	1.0	0.8720	0.30
engaging	0.0	0.0	1.0	0.3400	0.40
fantastic	0.0	0.0	1.0	0.5574	0.40

To summarise, the wordclouds/most common words and the histograms show that consumer sentiment towards the company's products is overall positive.

Reading the top 20 negative reviews and summaries, there are complaints about the product instructions being complicated/difficult, children getting frustrated or incomplete/faulty products being delivered. This shows that the company should look into their product design - perhaps it needs to be more age appropriate - as well as quality control.

The top positive reviews and summaries show words such as 'good gift', 'cute','fun' - this shows that a lot of customers are buying the products as presents and perhaps the company can offer customers gift packaging to enhance this point. In marketing compaigns they could emphasise the 'fun' aspect even more.

A wordcloud of the most common negative/positive words would also have been useful in this analysis.

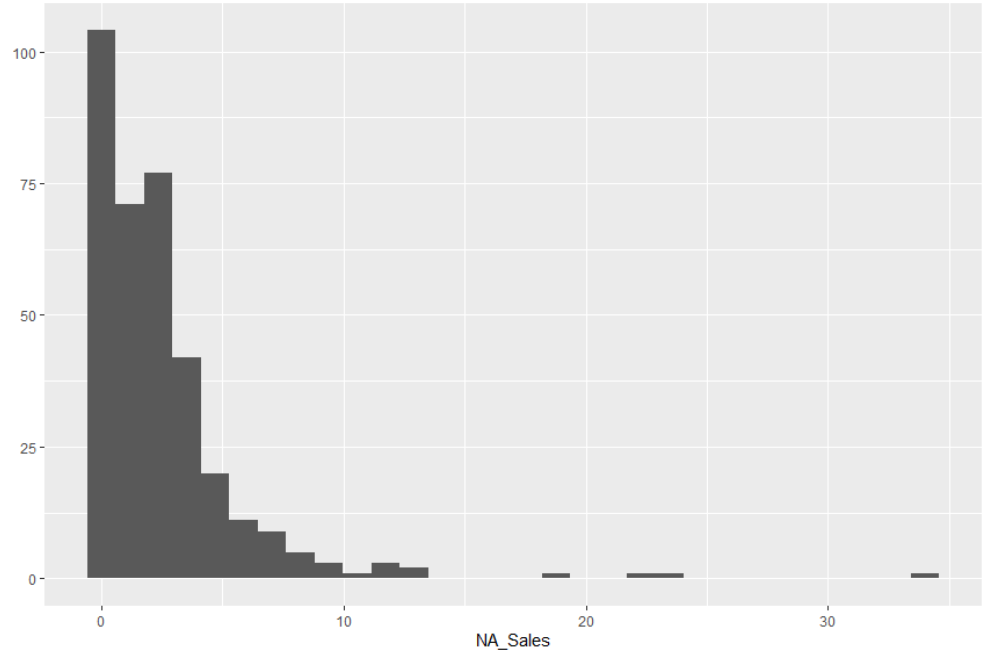
Question 4: Sales per product and data reliability

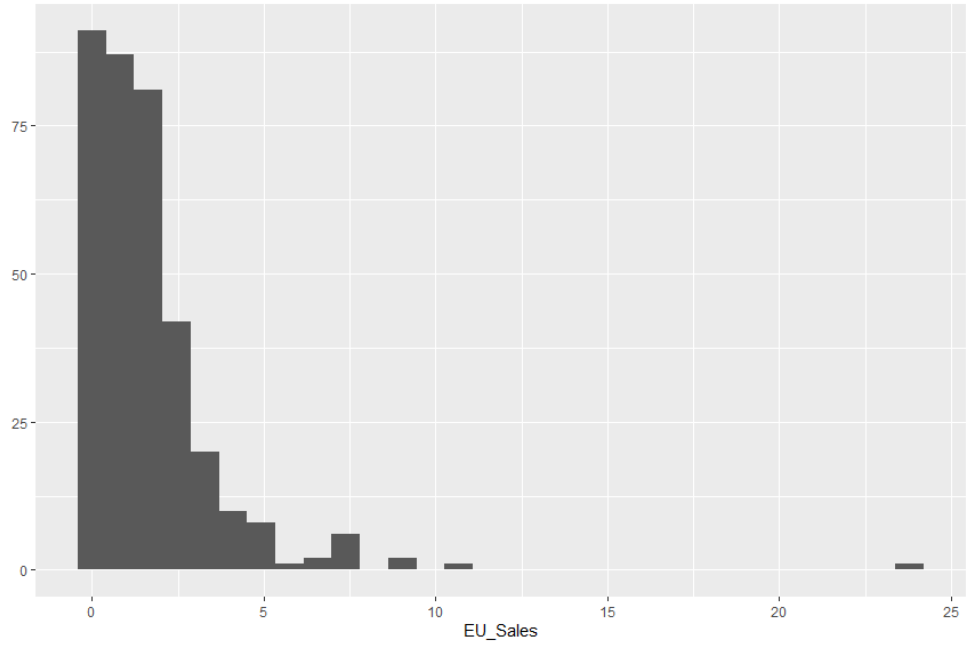
Analytical approach: The sales department of Turtle games would like to understand the sales per product and they prefer R to Python. I used R to explore, prepare, manipulate and did quick plots of the 'turtle_sales' data set to obtain some early insights.

Further I investigated the normality of the data set based on plots, and statistical tests (Skewness, Kurtosis, and a Shapiro-Wilk test). I then looked at the sum of sales per product to see what insights we can get from there.

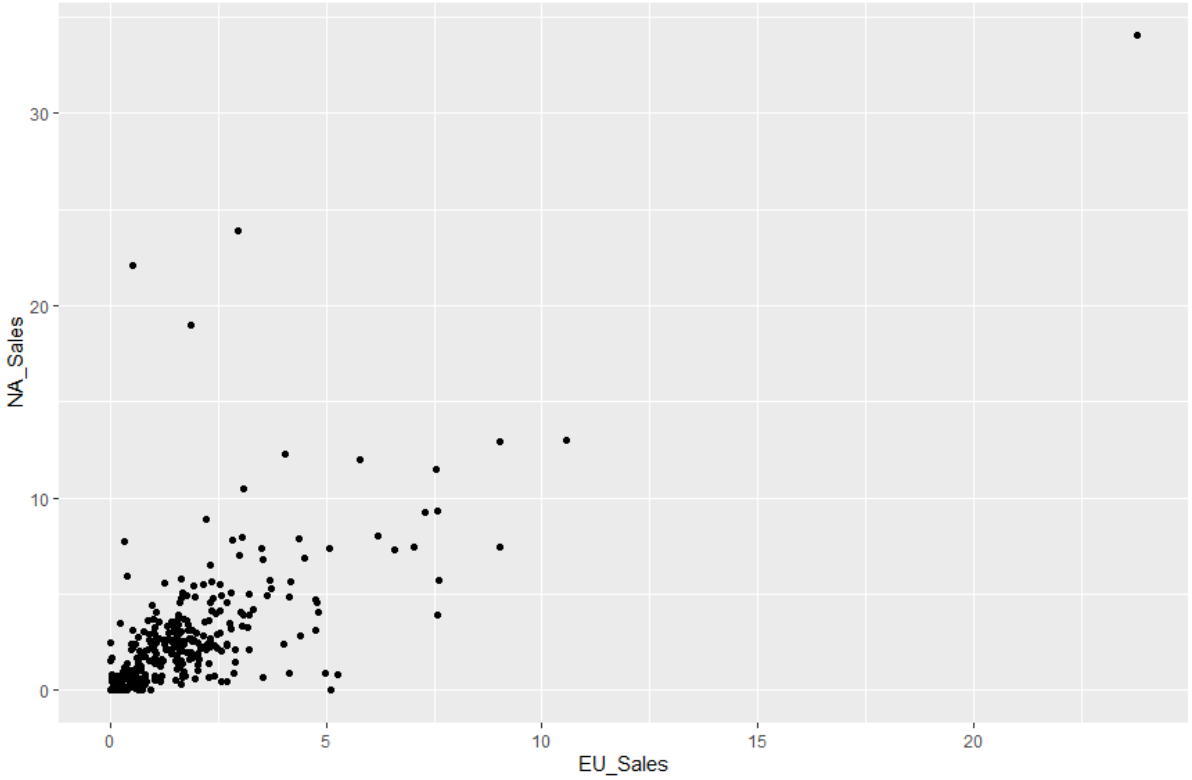
Visualisations and insights

1. Looking at the histograms, they sell on average more games in North America than in Europe – plot 11 & 12:

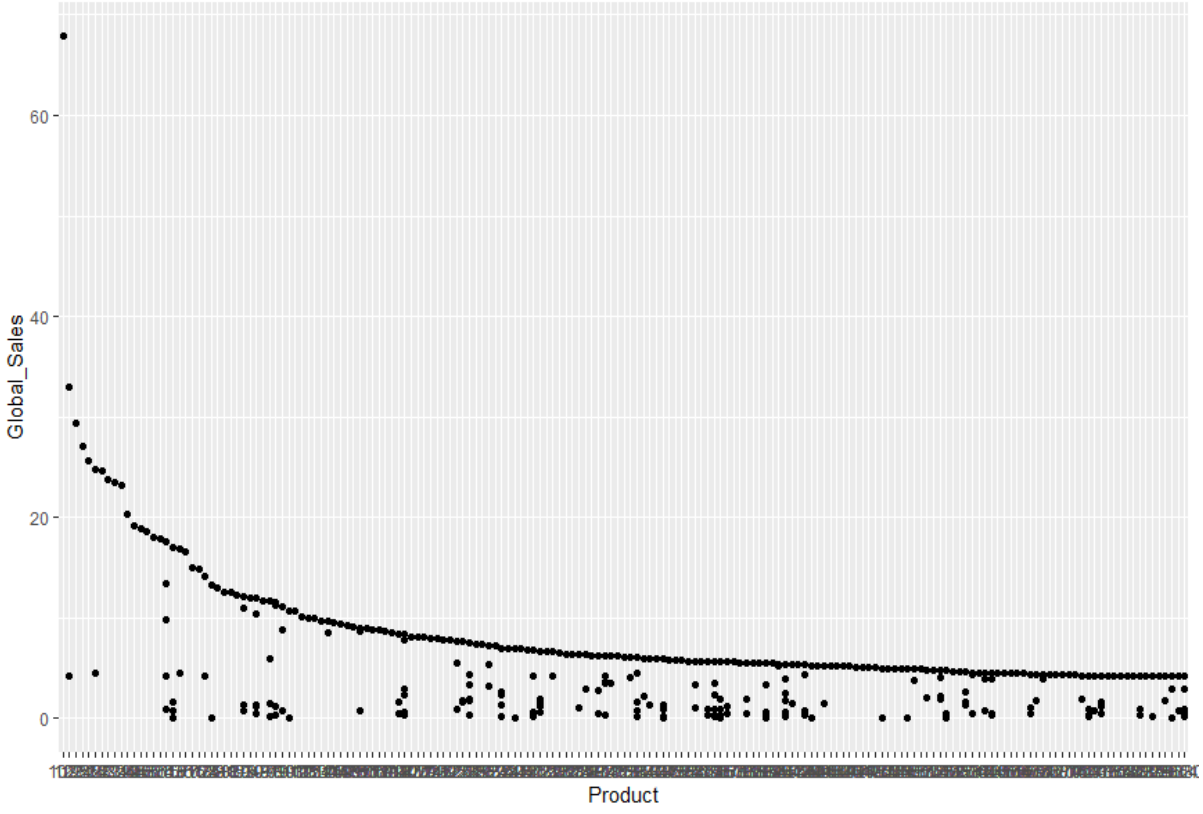




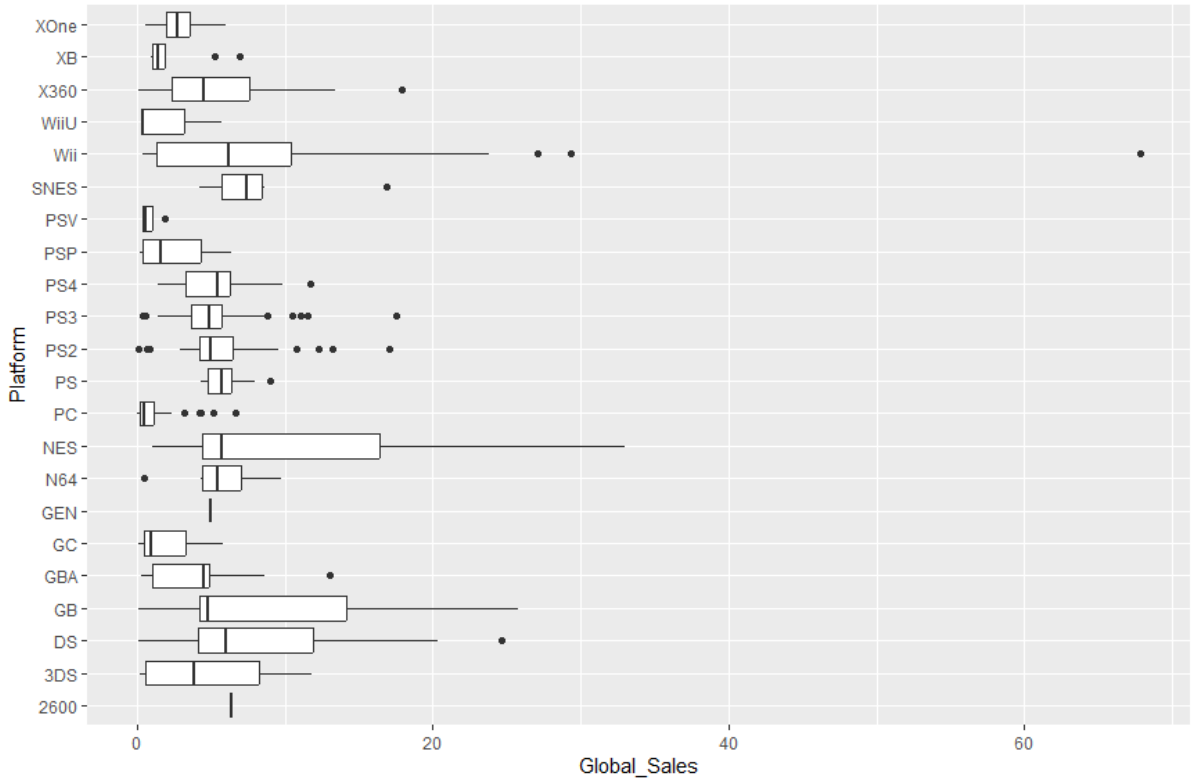
2. North America sales and EU sales data look positively correlated but there are few outliners – plot 13:



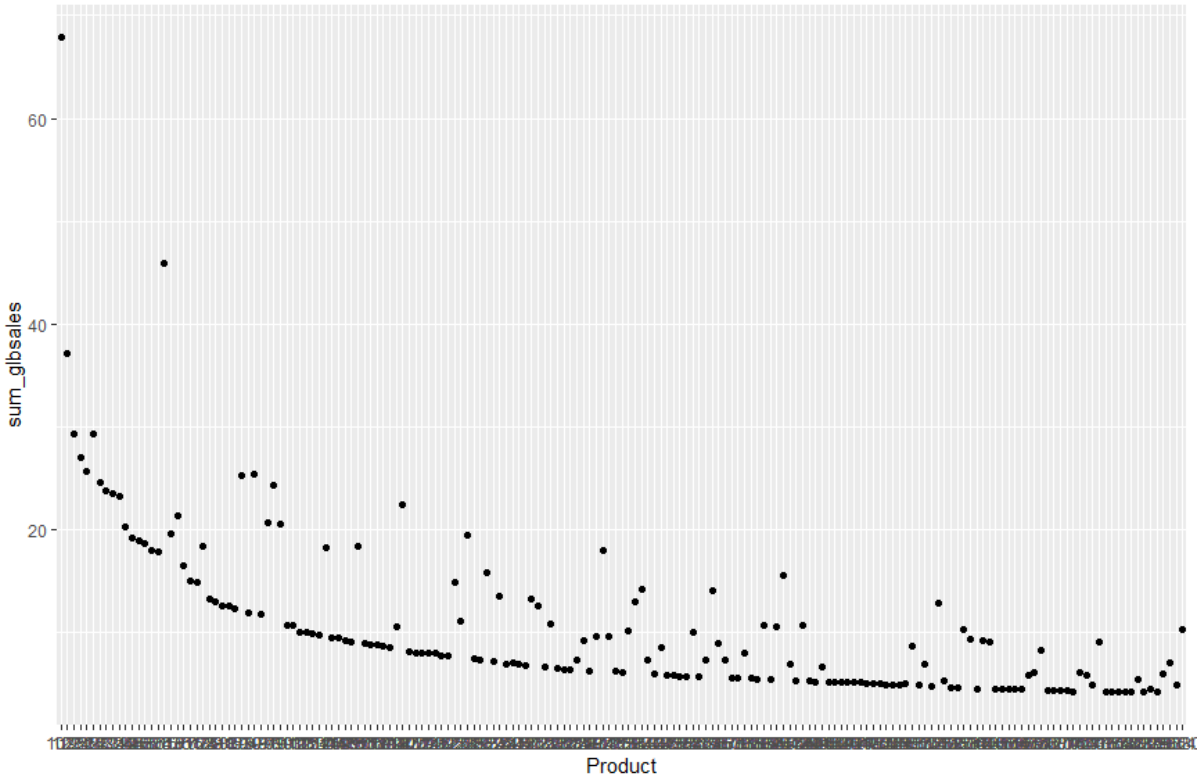
3. Global sales and Product code shows that smaller product codes have higher sales - is this a coincident and should we look into this further? – Plot 14



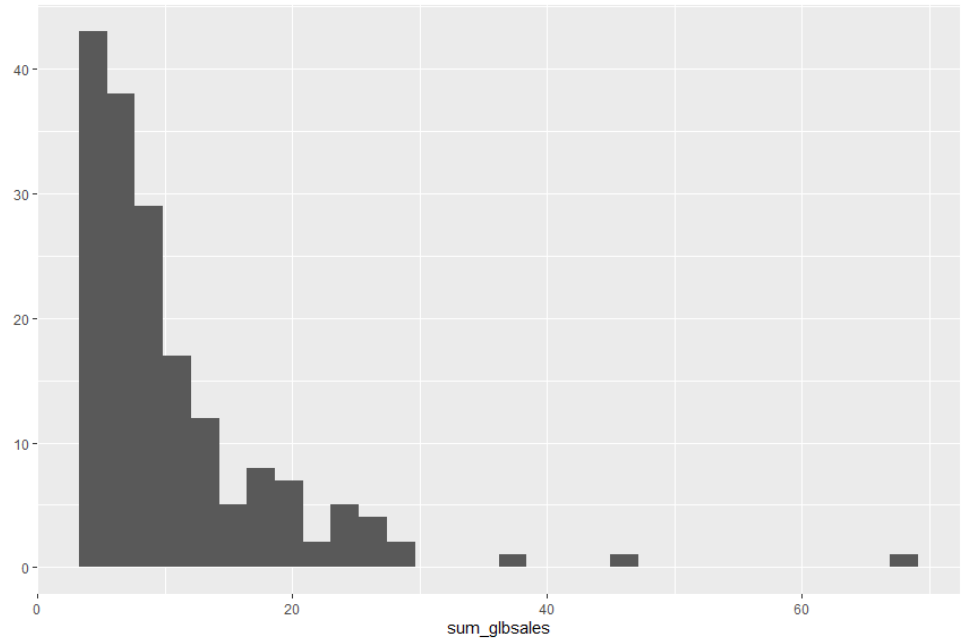
4. The platform that sold more games on average are Wii and SNES but the range is very wide – plot 15:



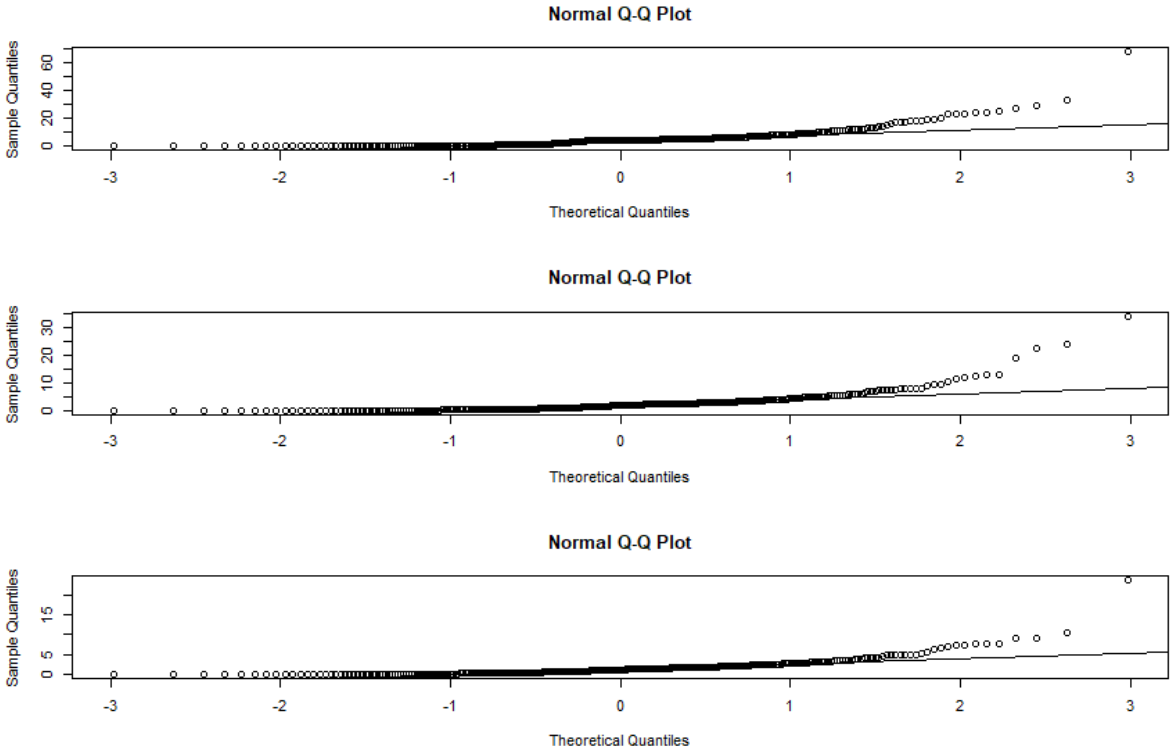
5. After trying a few different plots, I found that due to the large number of products, visualising total sales per product using boxplots were not very useful. Scatterplots of the sum of sales per product revealed that, similar to what we observed earlier, the smaller the product id the larger the total sales. This could be that the smaller product ids were older products that had been on the market for longer. Therefore, we may want to subset the data into groups of product ids, or find more information about the number of years they had been on sale for. Plot 16:



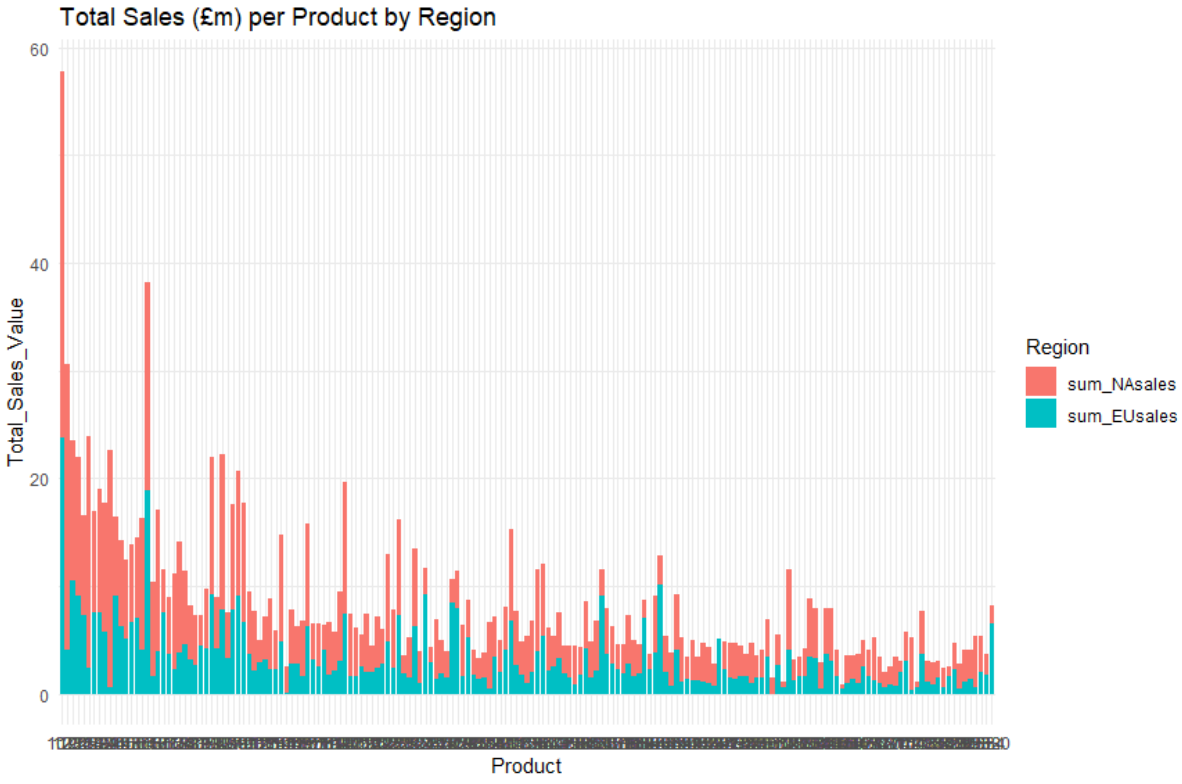
6. Most of the games sold between 0-15million pounds globally - plot 17:



7. All the normality tests suggest that the sales data are not normally distributed. This is supported by the histogram plots earlier. They are very skewed with fat tails suggesting significant outliners.
   Plot 18: Q-Q plots of Global sales, NA sales, EU sales.



8. They sell more games in North America (NA) than in Europe (EU).
   Plot 19



9. Global sales and NA_Sales are highly positively correlated. While NA sales and EU sales are also positively correlated, the correlation is less strong – see correlation table below:

	Global_Sales	NA_Sales	EU_Sales
Global_Sales	1.00	0.93	0.88
NA_Sales	0.93	1.00	0.71
EU_Sales	0.88	0.71	1.00

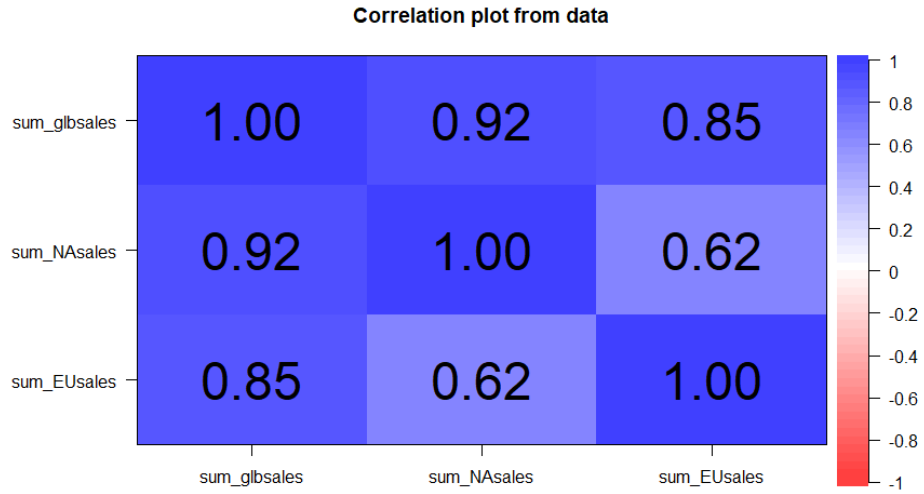
Question 5: Relationships between regional (North American and European) and global sales

Analytical approach: I investigated the potential relationship(s) in the sales data utilising simple and multiple linear regression models.

Visualisation and insights

Correlation matrix

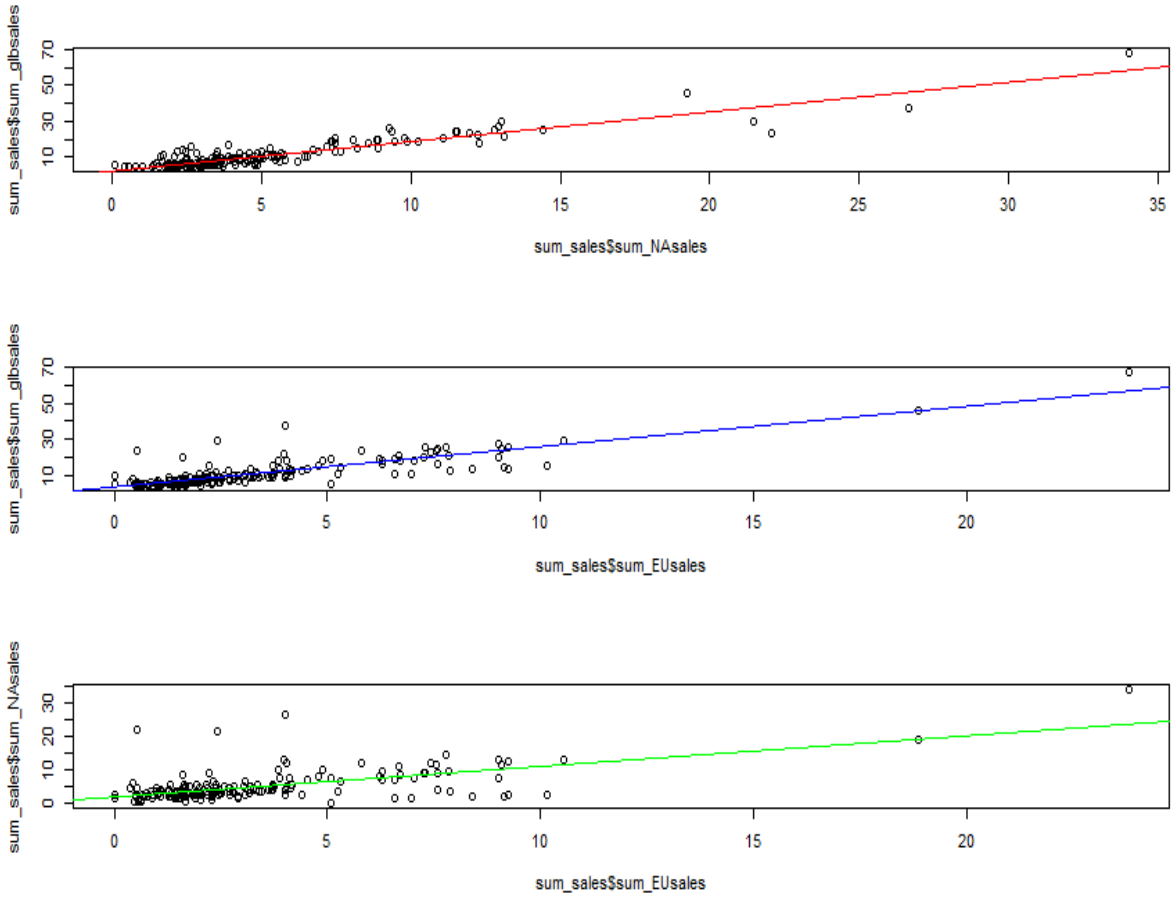
The correlation matrix suggests that global sales data has strong positive correlations with North American (NA) and European (EU) sales data, while the relationship between NA and EU sales data are somewhat weaker.



Simple linear regression model

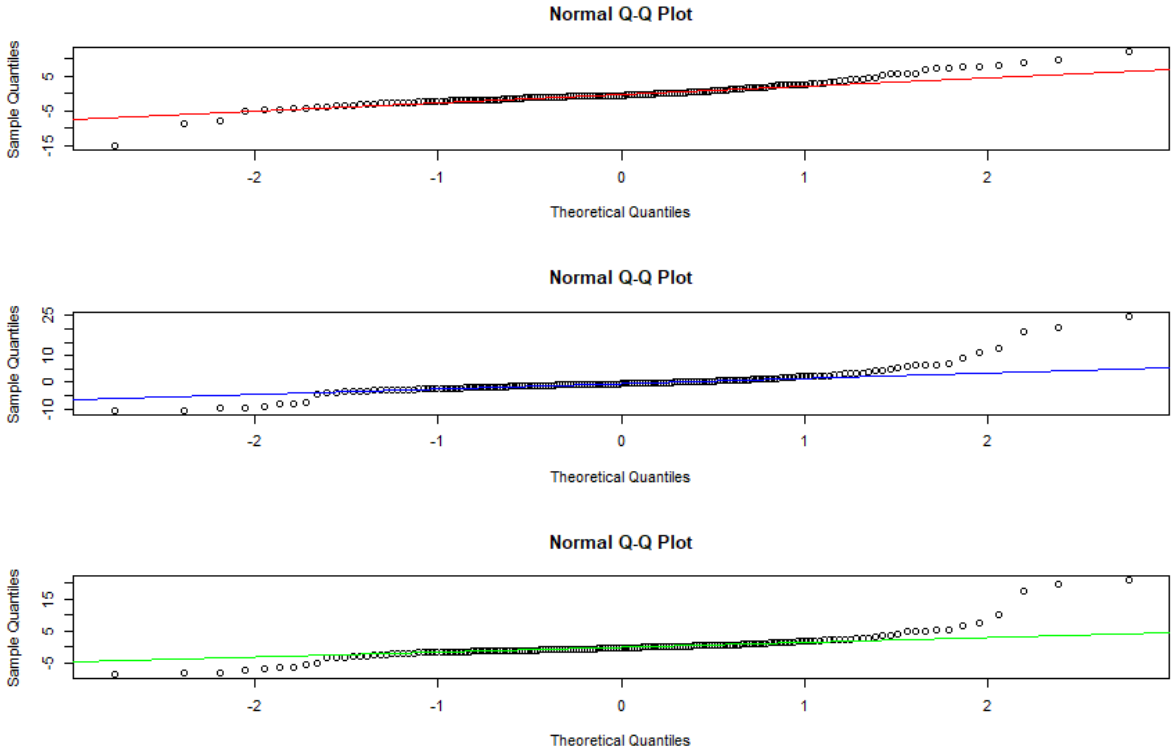
I regressed the sum of global sales per product on the North American (NA, model 1), European sales data (EU, model 2) respectively, and also regressed the North American data on the European sales data (model 3). The models all have statistically significant x coefficients - i.e. the sales columns have positive correlations. R-sq of model 3 is the lowest at 0.38, while model 1 and 2 have higher R-sq at 0.84 and 0.72 respectively. This suggests that EU sales is a poor explanatory variable for NA sales. The plots below also illustrate that the relationship between NA and EU sales is not well explained by a linear model.

Plot 20. Visualising the linear regressions



The residual plots below suggest that all three models produce residuals that are not normally distributed. There are too many extreme positive and negative residuals. This means one of the key assumptions of linear regression is not satisfied, and there may be problem with the model's reliability.

Plot 21. Checking normality of residual data



Multiple linear regression model

If we run a multiple linear regression model with NA sales and EU sales as the two X variables and Global sales as the dependent Y variable, the various statistical tests suggest that it's a good model with high R-sq (0.97) and significant x coefficients.

```
Call:
lm(formula = sum_glbsales ~ sum_NAsales + sum_EUsales, data = sum_sales3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4156 -1.0112 -0.3344  0.6516  6.6163

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04242    0.17736   5.877 2.11e-08 ***
sum_NAsales  1.13040    0.03162  35.745  < 2e-16 ***
sum_EUsales  1.19992    0.04672  25.682  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.49 on 172 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9664
F-statistic:  2504 on 2 and 172 DF,  p-value: < 2.2e-16
```

Comparing the predicted global sales values vs. the observed global sales values, for given NA and EU sales values, the multiple linear regression model seems to predict higher values than the observed. However, the predicted and observed global sales figures are reasonably close. The model seems to do a good job in explaining global sales.

```
> ModelvsObs
  sum_NAsales sum_EUsales glbsales_predicted Product sum_glbsales
1       34.02       23.80          68.056548     107        67.85
2        3.93        1.56           7.356754    <NA>           NA
3        2.73        0.65           4.908353    6815         4.32
4        2.26        0.97           4.761039    <NA>           NA
5       22.08        0.52          26.625558     326        23.21
```

Patterns and recommendations

- There is some evidence that a customer's loyalty points increase as their remuneration and spending increase. However, the linear models do not seem to be the appropriate models given problems of heteroscedasticity. Further treatment of the data or an alternative model is needed to confirm the relationship.
- There are obvious clusters when we visualise two key characteristics of a customer – spending score and remuneration. Using k-means clustering, I identified five clusters – i.e. five customer groups - which can be used by the marketing team to target specific market segments.
- With the help of NLP techniques, I understand from the reviews data that consumer sentiment towards the company's products is overall positive. The top 20 negative reviews and summaries show that the company should look into their product design - perhaps it needs to be more age appropriate - as well as quality control. The top positive reviews show that a lot of customers are buying the products as presents and perhaps the company can offer customers gift packaging to enhance this point. In marketing campaigns they could emphasise the 'fun' aspect even more. A wordcloud of the most common negative/positive words would also have been useful in this analysis.
- The sales data show that the company sells more games in North America than in Europe. Global sales data has strong positive correlations with North American (NA) and European (EU) sales data, while the relationship between NA and EU sales data are somewhat weaker. To predict global sales, the multiple linear regression model using both North American and European sales data as x-variables does a better job than the simple linear regression models. A recommendation here would be for Turtle Games to review their sales strategy on a regional basis. Data on how many products are sold relative to each market's population and income per capita would also be helpful – the higher sales in North America may simply be attributable to the higher population and/or higher income level for example.