Turtle Games,
Data analysis

INDEX

CONTEXT

Turtle Games is a game manufacturer and retailer with a global customer base. The company manufactures and sells its own products, along with sourcing and selling products manufactured by other companies. Its product range includes books, board games, video games, and toys. The company collects data from sales as well as customer reviews.

This report summarises an analysis of the following data:

- **Reviews data** across products from their customers
- **Sales data,** related to the video games sold across North America, Europa and worldwide.

In this context, the data analysis is split between:

A. Review analysis, using Python
B. Sales analysis, using R

A. Reviews - data analysis

A.1. APPROACH

Using Python, through numpy, pandas, seaborn, matplotlib, statsmodels and sklearn, the following procedures were done to structure, clean and analyse the data:

Linear relationships

- **Import** the three datasets provided as DataFrames.
- Check for **missing data**: not founded.
- **First exploration of the data** through functions (dtypes, head, tail, shapes, describe)
- **Drop unnecessary columns** (language and platform) and rename other columns
- **Perform and plot single linear regression**, using loyalty points as dependent variable and spending/remuneration/age separately as independent variables
- **Perform multi linear regression (MLR)**, using loyalty points as dependent variable and spending and remuneration as independent variables. Additional, including age also dependent variable was studied but not included, as it is concluded a similar Adjusted R-squared and decided to keep the simplest
- **Train and test the model**
- **Check heterocedasticity using Breuschpagan**
- Finding evidence of heterocedasticity, it is **transformed the data using sqrt(loyalty) instead of loyalty**. Repeated again the previous steps using this new variable
- With sqrt(loyalty), verified the assumption of homocedasticity.

- **Mean Absolute Error and Mean Square Error** calculated to analyse if it as plausible values
- **Detecting multicollinearity** using VIF Factor

Clustering

- First exploration, using **scatterplot and pairplot,** of remuneration vs spending
- Estimate **Elbow and Silhoutte methods** to determine the best options for clusters' number
- **Evaluate k-means model at the different values of k** selected
- **Plot the model with the optimal K considered**

NLP

- **Change** reviews and summary **data to lowercase**
- **Remove punctuation**
- **Drop duplicates in both columns**
- **Tokenised and create wordclouds,** firstly without any changes and after **removing alphanumeric characters and stopwords**
- **Calculate the frequency distribution**
- **Identify the most common words**, separately for each variable and then identifying the **common top words in both cases (plotting it for better visualisation using barplot)**
- **Polarity and Sentiment Analysis, using TextBlob**, plotting histograms for better visualisation
- **Additional analysis using Vader**, for summary variable, plotting **using boxplot**

A.2. KEY INSIGHTS AND RECOMMENDATIONS

- **Strong MLR model that allows to predict loyalty points, using remuneration and scoring variables.**
  - Adjusted R-squared: 88,2%, representing the proportion of loyalty explained by independent variables (spending/remuneration).
  - An increase of 1 unit on spending and remuneration, increases 0.42 and 0.40 of sqrt(loyalty), respectively

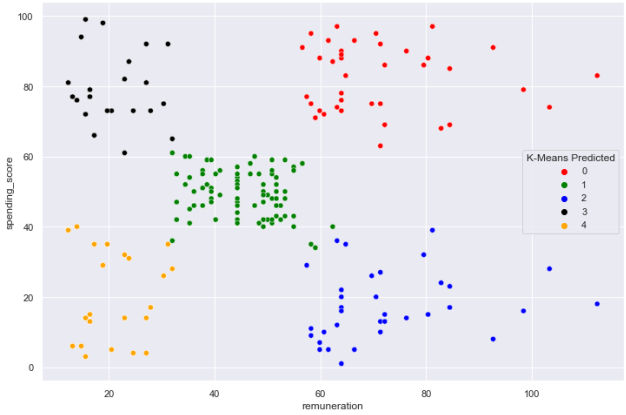
OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:	0.882
Model:	OLS	Adj. R-squared:	0.881
Method:	Least Squares	F-statistic:	5944.
Date:	Sun, 11 Sep 2022	Prob (F-statistic):	0.00
Time:	10:48:04	Log-Likelihood:	-4895.5
No. Observations:	1600	AIC:	9797.
Df Residuals:	1597	BIC:	9813.
Df Model:	2		
Covariance Type:	nonrobust		

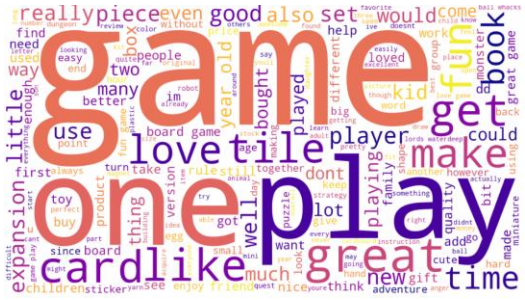
	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.0766	0.397	-10.270	0.000	-4.855	-3.298
spending_score	0.4258	0.005	84.999	0.000	0.416	0.436
remuneration	0.4024	0.006	71.086	0.000	0.391	0.414

Omnibus:	247.088	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	369.184
Skew:	-1.144	Prob(JB):	6.80e-81
Kurtosis:	3.554	Cond. No.	225.

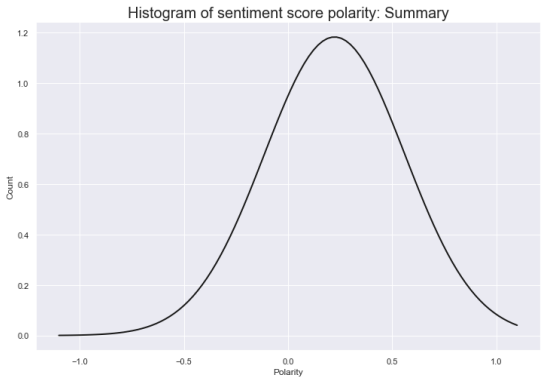
- **Five-clusters model, identifying patterns between remuneration and spending variables**



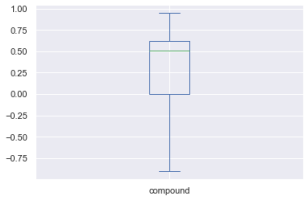
- **NLP – Most common words include the word which could represent that fun games are crucial for your customers, so you should potentiate this segment**



- **NLP – Positive results on sentiment analysis of reviews, recommending further analysis of negative reviews to improve**



summary	polarity_summary
the worst value ive ever seen	-1.000000
boring unless you are a craft person which i am	-1.000000
boring	-1.000000
before this i hated running any rpg campaign dealing with towns because it	-0.900000
another worthless dungeon masters screen from galeforce9	-0.800000
disappointed	-0.750000
disappointed	-0.750000
disappointed	-0.750000
disappointed	-0.750000
promotes anger instead of teaching calming methods	-0.700000
too bad this is not what i was expecting	-0.700000



B. Sales - data analysis

B.1. APPROACH

Using R, through tidyverse, ggplot2, plotly, moments and data.table libraries, the following procedures were done to structure, clean and analyse the data:

Overview

- **Import** the dataset provided
- **First exploration using DataExplorer, creating a html report**
- **Save new dataset, dropping unnecessary columns** (Platform, Ranking, Year, Genre and Publisher)
- Use **view and summary functions** to analyse the dataset
- Plotting, through **scatterplots, histograms, boxplots** to visualise the pattern of Sales variables
- Additionally, **calculate Other sales (Global-EU-NA) and plot with all Sales columns to view the dimension of it.** Then, drop Other_sales created, focusing the analysis without this disaggregation

Analyse variables relationship

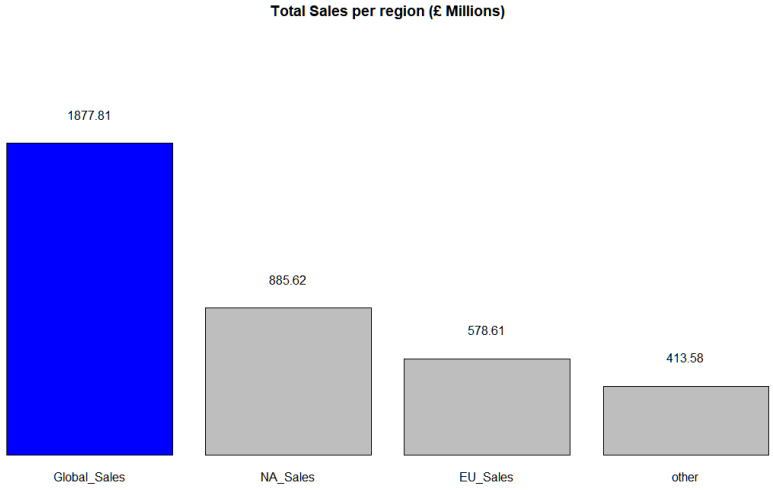
- **Aggregate the data by product**, using sum.
- **Apply view, summary and dim** to have a perception of the dataset
- Develop **scatterplots to analyse the relationship between Global, NA and EU Sales**.
- Perform boxplots mainly to check the **evidence of outliers**
- Additionally, **remove outliers** (values higher than 15 – limit selected throughout graphic analysis) and scatterplot EUvsNA **to zoom** the data analysis in this interval
- **QQ-Plots to analyse the normality** of each sales variables
- **Perform Shapiro-Wilk tests** to assess the normality.
- **Calculate Skewness and Kurtosis** for each sales variables
- Assess the **correlation** between Sales variables
- Plot **density functions** to a better visualisation of each variable.
- Additionally, plot the **density functions in the same graph** to compare between variables. In this case, melt function it is applied to plot jointly
- Plot, using barplot, **top 10 products in Global Sales to identify the most relevant**.

Analyse variables relationship

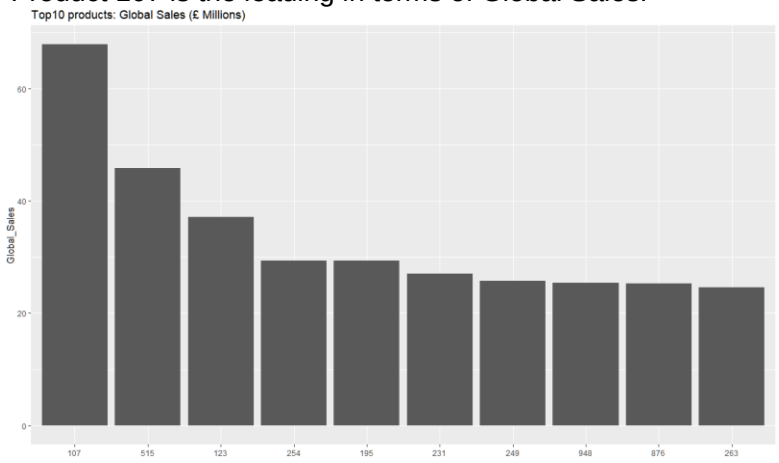
- Perform and plot **single linear regressions** between sales variables
- **Perform multi linear regression (MLR)**, using Global as dependent variable and EU/NA as independent variables.
- **Plot the residuals** to analyse if follows a normal distribution
- **Estimate predicted values** for the indicated values and compare with the observed values
- **Plot to compare observed values vs confidence interval estimated**
- Plot **dynamic scatterplots** for sales, differentiating points by publisher using colours, to detect if the higher values have a predominant one.
- Reestimate MLR, adding TRUE/FALSE variable of Publisher “Nintendo” as they represent all of the highest values. Test it versus the observed values.

B.2. KEY INSIGHTS AND RECOMMENDATIONS

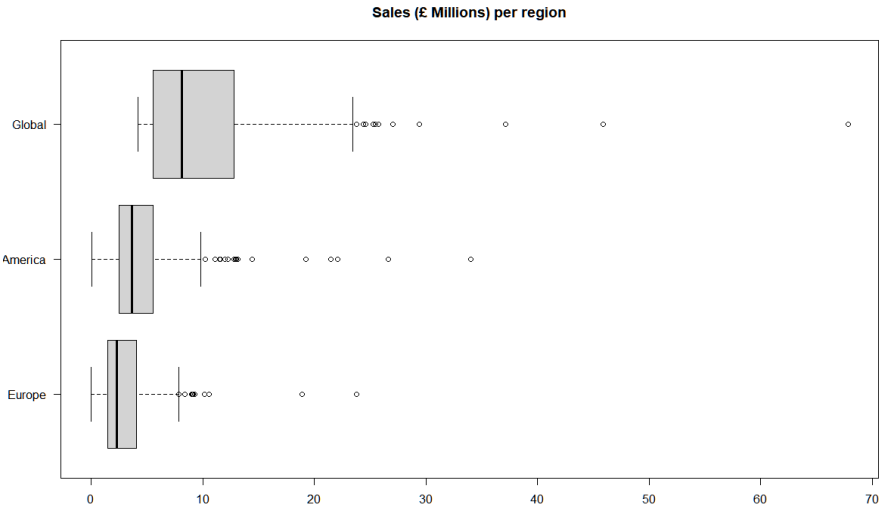
- Europe and North America represent almost 80% of the Global Sales. For further analysis, it could be useful explore and detail the others regions included “Other”



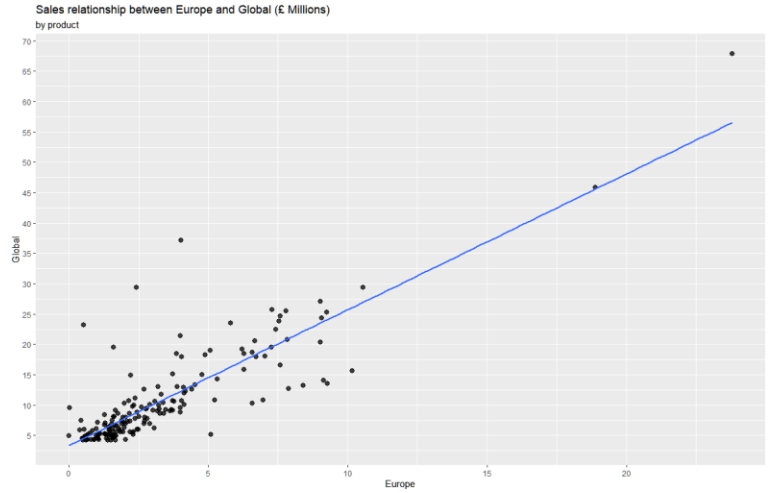
- Product 107 is the leading in terms of Global Sales.

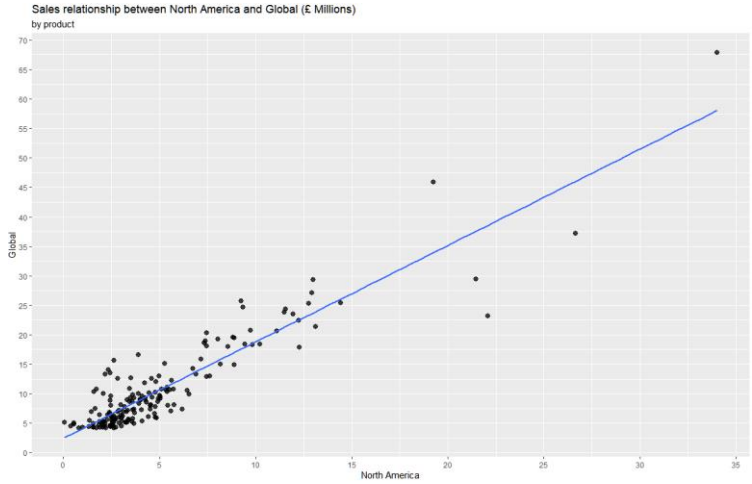


- All Sales variables have some highest values as outliers. Recommendation: perform a zoom analysis of these cases.

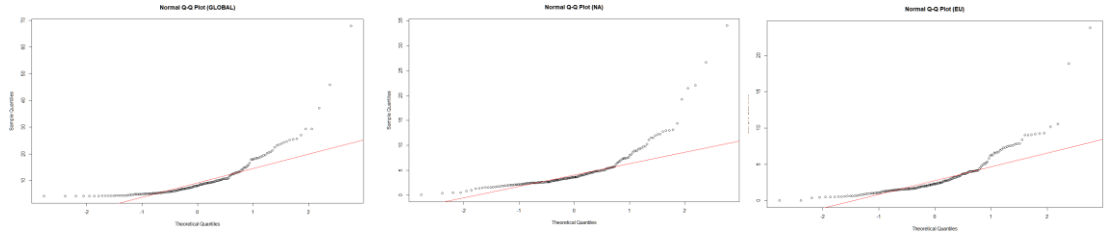


- Strong correlation between Global Sales vs NA Sales and Global Sales vs EU Sales

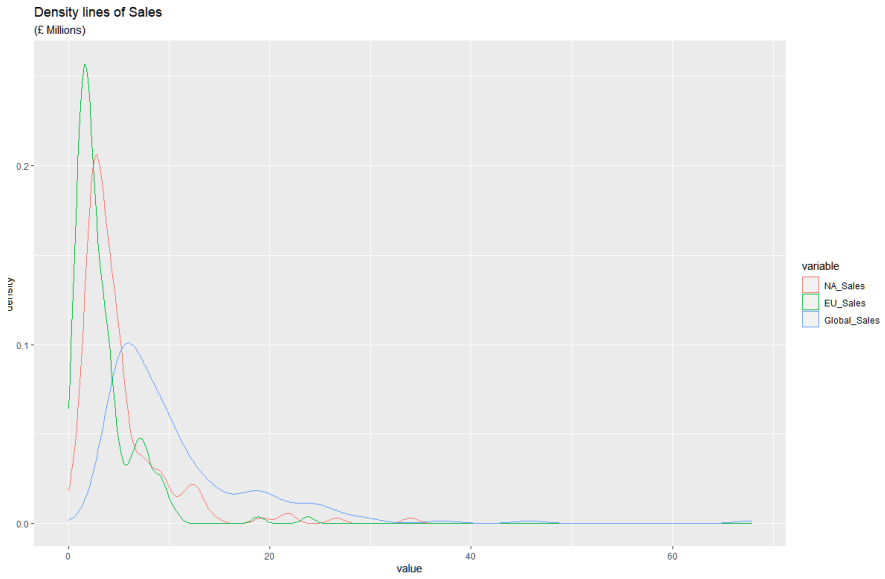




- QQ-plots show lack of normality in all variables, in particular values higher than 1 and lower than -1.



- **Right skewed and very high values of kurtosis**



- **MLR model with a very strong R-squared. Although does not follow a normal distribution, limiting our conclusions and estimations.** Recommendation: explore another alternative model could be an add-on.

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

Residuals:
    Min      1Q  Median      3Q     Max
-3.6186 -0.4234 -0.2692  0.0796  7.4639

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22175    0.07760   2.858  0.00453 **
NA_Sales     1.15543    0.02456  47.047  < 2e-16 ***
EU_Sales     1.34197    0.04134  32.466  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.112 on 349 degrees of freedom
Multiple R-squared:  0.9687,    Adjusted R-squared:  0.9685
F-statistic:  5398 on 2 and 349 DF,  p-value: < 2.2e-16
```