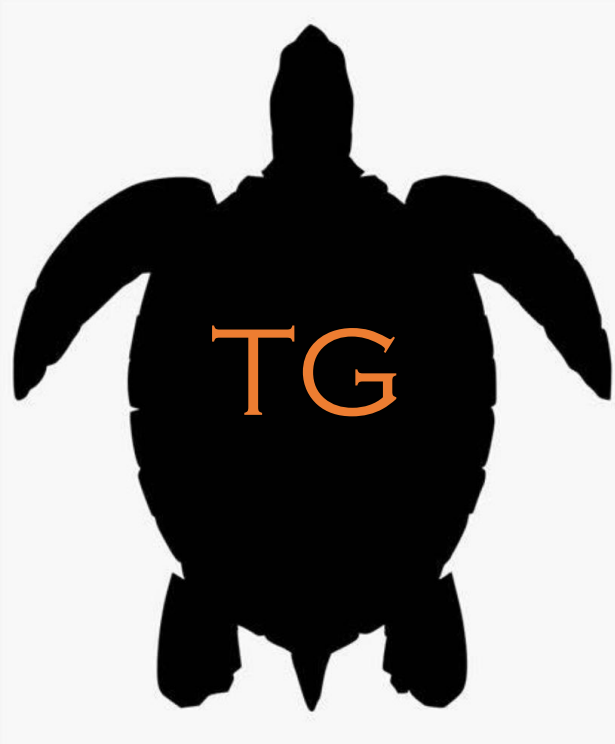


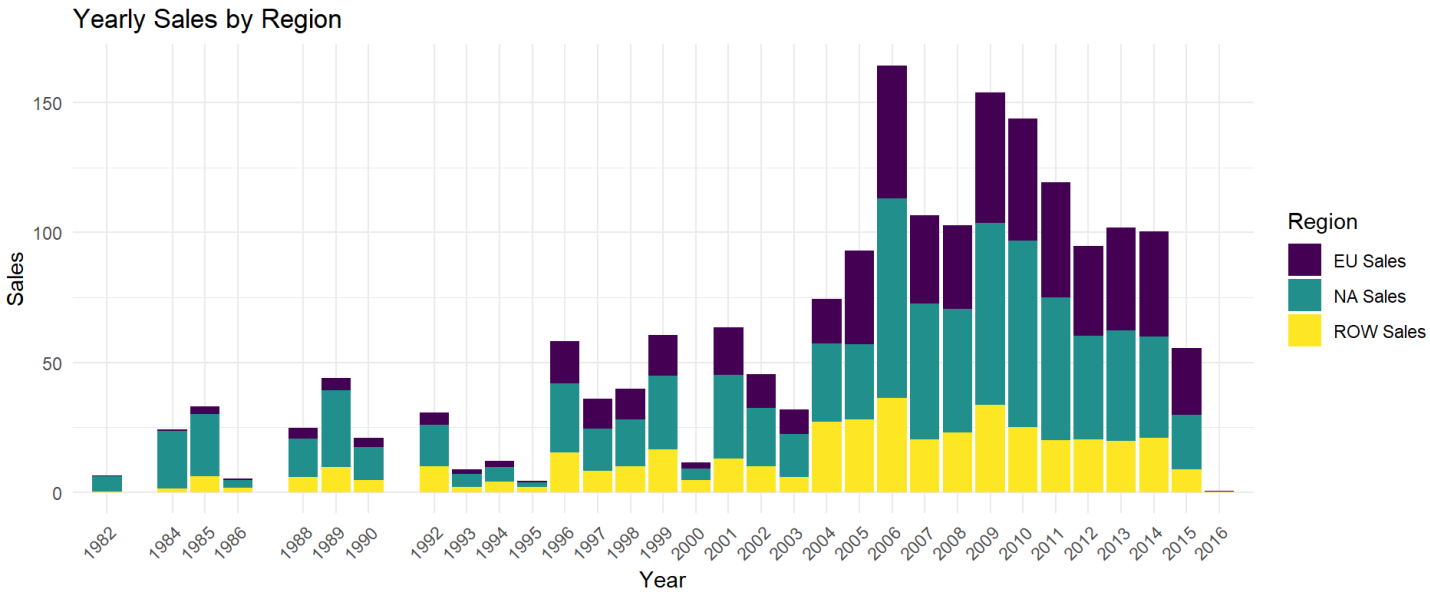
Data Analysis by Alasdair Bell

Contents

Background

Identify and Define the Problem

Sales at Turtle Games have decreased dramatically in recent times.



Considering this, TG are pursuing the objective of improving overall sales performance. We have been tasked with analysing customer trends to provide insights as to how this objective can be achieved. The analysis will be focused on addressing the following questions in an IDEAL framework:

1. How do customers engage with and accumulate loyalty points?
2. How can customers be segmented into groups, and which groups can be targeted by the marketing department?
3. How can text data (e.g. social data such as customer reviews) be used to inform marketing campaigns and make improvements to the business?
4. What trends and insights can be identified from the sales data?
5. Is the sales data suitable for modelling purposes? (Provide a statistical review and interpretation (e.g. normal distribution, skewness, or kurtosis) to justify the answer.)
6. Are there any possible relationships between North American, European, and global sales? (If yes, please provide a description.)

1. Loyalty points

Identify the Problem and Opportunities

TG does not have a full understanding of how the accumulation of loyalty points relates to customer variables such as spending, renumeration and age.

Understanding how loyalty points are accumulated presents several opportunities:

- Targeted marketing
- Improved loyalty program design.
- Increase customer satisfaction and retention.
- Product and service development.
- Foster data driven decision making.

Define Goals

To use linear regression to assess the relationship between customer variables and loyalty points.

Explore

The data was assessed in accordance with the following procedural approach which aligns with the accompanying Jupiter notebook.

Process	Action	Results	Observations
1.1 Load and Explore data			
Necessary packages imported.			
Data loaded.	CSV file read into new df 'reviews'.		
Missing values determined.	reviews.isna().sum()	No missing values detected.	
Data explored.	print(reviews.info())	11 columns, 2000 records.	
Descriptive statistics produced.	print(reviews.describe())		
1.2 Drop Columns	reviews_clean = reviews.drop(['language', 'platform'], axis=1)	language and platform columns removed.	
1.3 Rename Columns	reviews_final = reviews_clean.rename(columns={'remuneration (k£)': 'remuneration', 'spending_score (1-100)':'spending_score'})	Columns renamed to more readable format.	
1.4 Clean df saved to csv.	reviews_final.to_csv('turtle_reviews_final.csv', index=False)	Cleaned df saved to csv and reloaded as rf dataframe.	
1.5.1 Simple linear regression			
Independent and dependent variables defined.	Dependent variable - loyalty_points. Independent variables - spending_score. remuneration and age respectively.		
OLS test conducted.	E.g. model_sp = sm.OLS(y_sp, sm.add_constant(x_sp)).fit()		
Summary stats produced.	E.g. print(model_sp.summary())	See results section.	
Predicted values extracted.	E.g. print(model_sp.predict())		No requirement for params and standard errors as they are contained in summary stats.
Regression table generated using the estimated parameters.	E.g. y_pred_sp = -75.0527 + 33.0617 * x_sp		
Scatter plot with line of best fit produced.			
Predictions made for a given independent variable.	E.g. New_sp = 70 y_pred_sp = -75.0527 + (33.0617 * New_sp)		
Residuals plotted.	E.g. plt.scatter(x_sp, model_sp.predict()- y_sp, s=10) plt.plot(x_sp, y_sp - y_sp, color='black')		
qq plot produced.	E.g. residuals = model_sp.predict() - y_sp sm.qqplot(residuals, fit=True, line='45')		
Models evaluated.	Summary stats analysed for model predictive accuracy.	See Observations section.	
Boxplots produced for categorical variables.			

1.1 Linear Regression – Basic Assumptions

Assumption	Explanation
Linear Model	This assumption posits a linear relationship between independent and dependent variables, as non-linear relationships in real-world data would lead to inaccurate predictions with significant variance from actual observations in a linear regression model.
No multicollinearity	Multicollinearity occurs when predictor variables are correlated, leading to redundancy in the dataset as these variables contain similar information. This redundancy increases model complexity without contributing new information or patterns, making it advisable to avoid highly correlated features even in complex models.
Homoscedasticity of Residuals	Homoscedasticity refers to the condition where residuals from a linear regression model are uniformly spread, indicating a satisfactory model.
No Autocorrelation in residuals	Absence of autocorrelation, where residuals are independent of each other.
Number of observations Greater than the number of predictors	To enhance model performance, the quantity of training data should exceed that of test data. Specifically, in linear regression, the number of observations must be greater than the number of independent variables.
Unique Observations	Each observation in the dataset should be independent, implying that it is uniquely measured for each occurrence of the event causing the observation.
Normal distribution of Residuals	This assumption states that the residuals (errors) from the regression model should be normally distributed. This is particualrly important if the data is to be used in hypothesis testing.

A. Spending score vs Loyalty Points

Summary Statistics

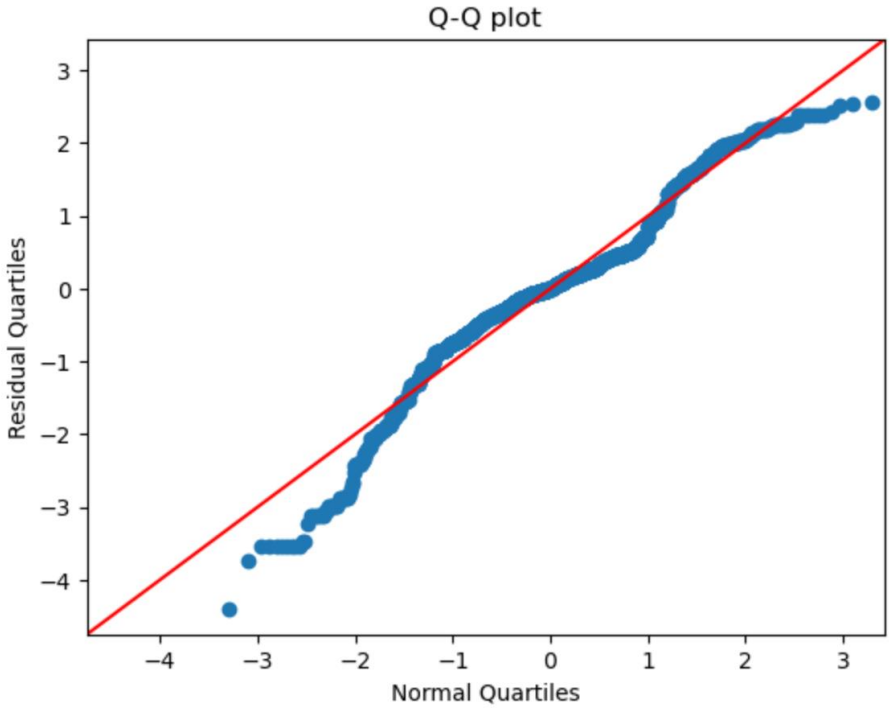
OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:	0.452
Model:	OLS	Adj. R-squared:	0.452
Method:	Least Squares	F-statistic:	1648.
Date:	Fri, 12 Jan 2024	Prob (F-statistic):	2.92e-263
Time:	14:01:26	Log-Likelihood:	-16550.
No. Observations:	2000	AIC:	3.310e+04
Df Residuals:	1998	BIC:	3.312e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-75.0527	45.931	-1.634	0.102	-165.129	15.024
spending	33.0617	0.814	40.595	0.000	31.464	34.659

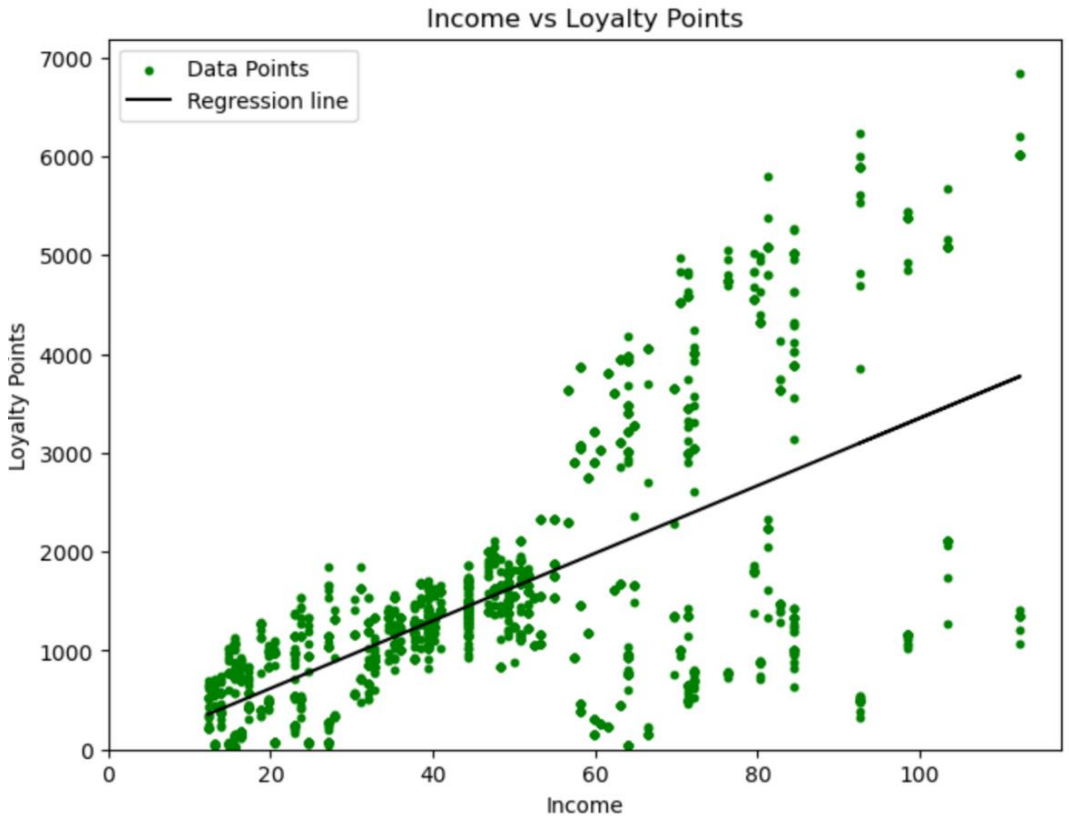
Omnibus:	126.554	Durbin-Watson:	1.191
Prob(Omnibus):	0.000	Jarque-Bera (JB):	260.528
Skew:	0.422	Prob(JB):	2.67e-57
Kurtosis:	4.554	Cond. No.	122.

A notable funnel pattern implies heteroscedasticity.



Deviation from the best fit line suggests that the residuals are not normally distributed.

B. Remuneration vs Loyalty Points



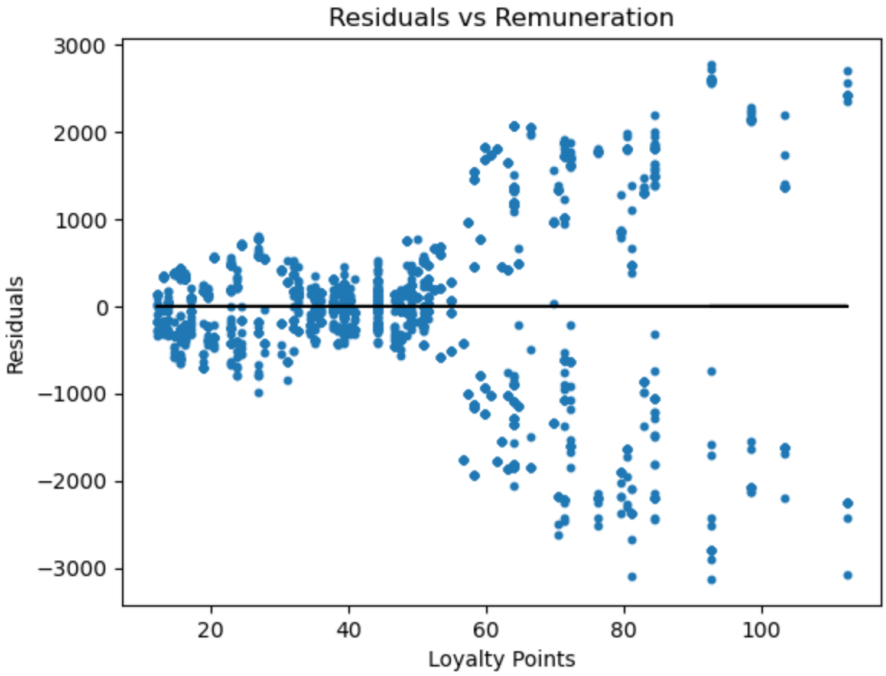
Summary Statistics

OLS Regression Results

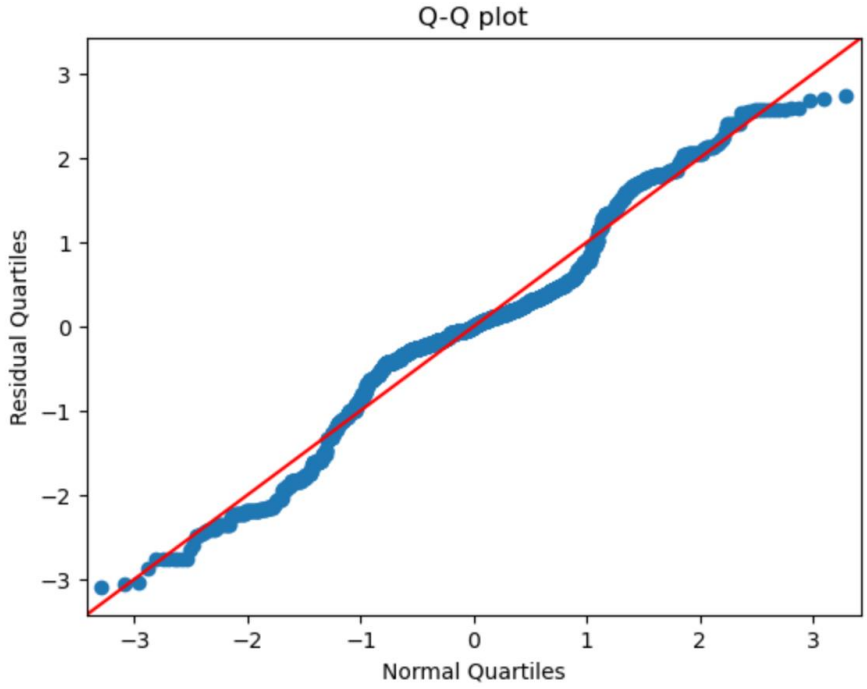
Dep. Variable:	loyalty_points	R-squared:	0.380
Model:	OLS	Adj. R-squared:	0.379
Method:	Least Squares	F-statistic:	1222.
Date:	Fri, 12 Jan 2024	Prob (F-statistic):	2.43e-209
Time:	14:17:28	Log-Likelihood:	-16674.
No. Observations:	2000	AIC:	3.335e+04
Df Residuals:	1998	BIC:	3.336e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-65.6865	52.171	-1.259	0.208	-168.001	36.628
income	34.1878	0.978	34.960	0.000	32.270	36.106

Omnibus:	21.285	Durbin-Watson:	3.622
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31.715
Skew:	0.089	Prob(JB):	1.30e-07
Kurtosis:	3.590	Cond. No.	123.

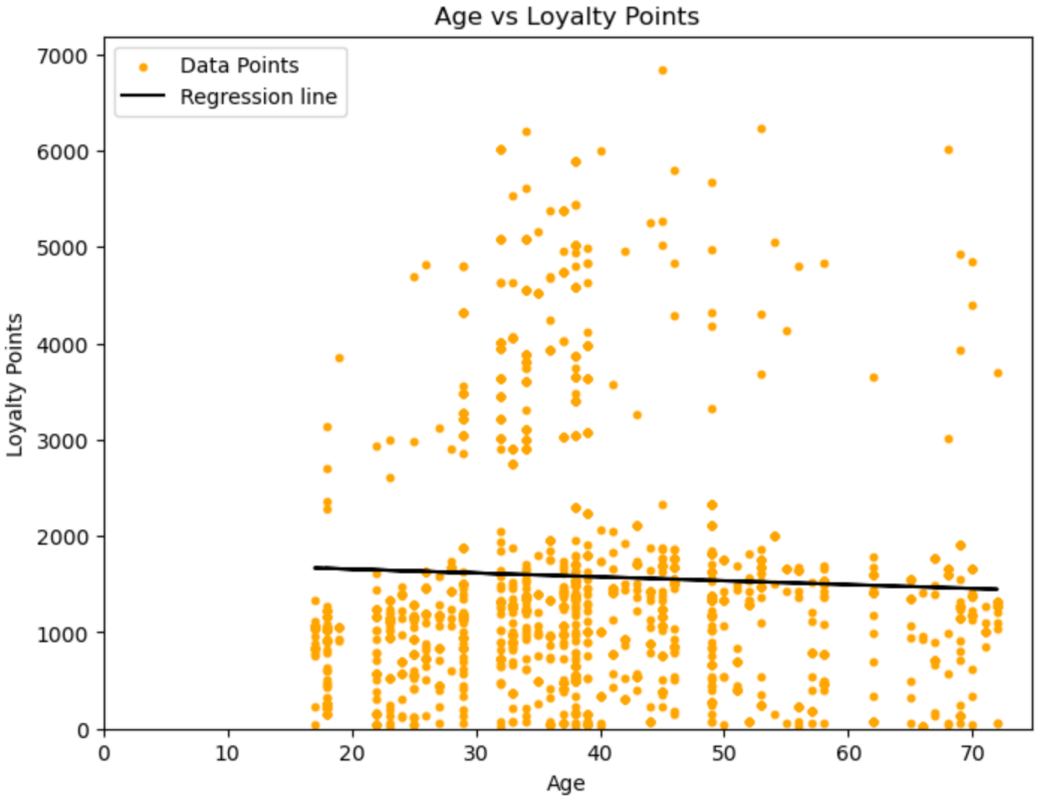


A notable funnel pattern implies heteroscedasticity.



Deviation from the best fit line suggests that the residuals are not normally distributed.

C. Age vs Loyalty Points



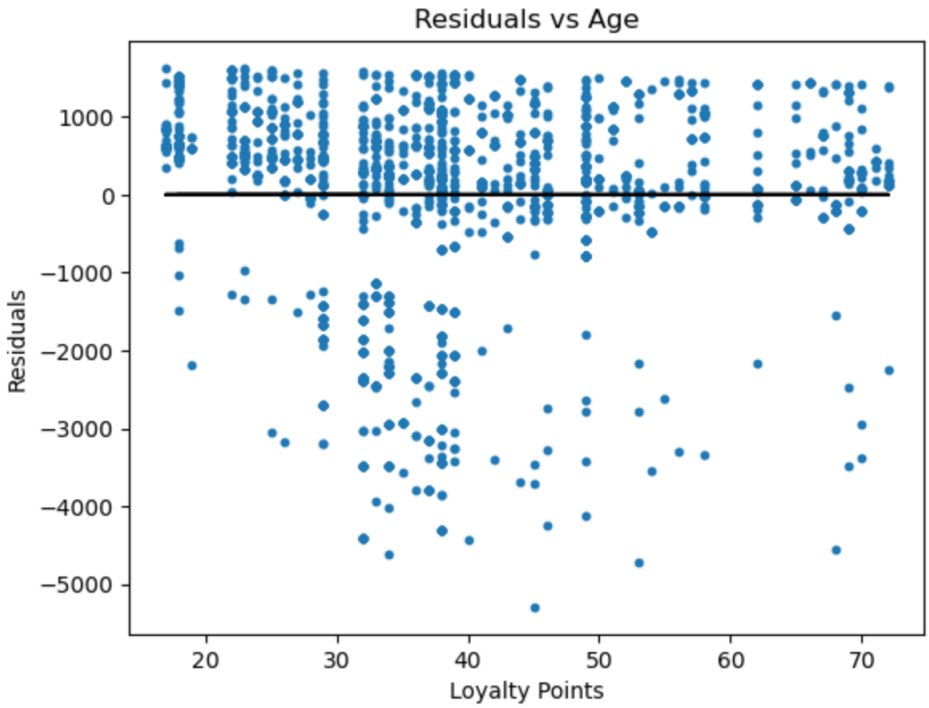
Summary Statistics

OLS Regression Results

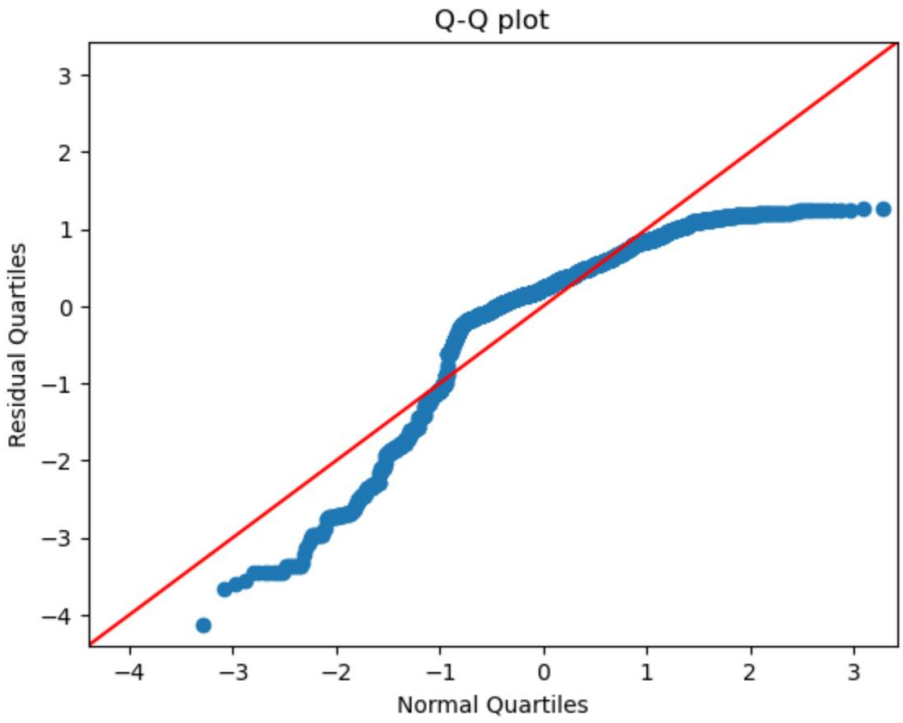
Dep. Variable:	loyalty_points	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	3.606
Date:	Fri, 12 Jan 2024	Prob (F-statistic):	0.0577
Time:	14:24:05	Log-Likelihood:	-17150.
No. Observations:	2000	AIC:	3.430e+04
Df Residuals:	1998	BIC:	3.431e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1736.5177	88.249	19.678	0.000	1563.449	1909.587
age	-4.0128	2.113	-1.899	0.058	-8.157	0.131

Omnibus:	481.477	Durbin-Watson:	2.277
Prob(Omnibus):	0.000	Jarque-Bera (JB):	937.734
Skew:	1.449	Prob(JB):	2.36e-204
Kurtosis:	4.688	Cond. No.	129.

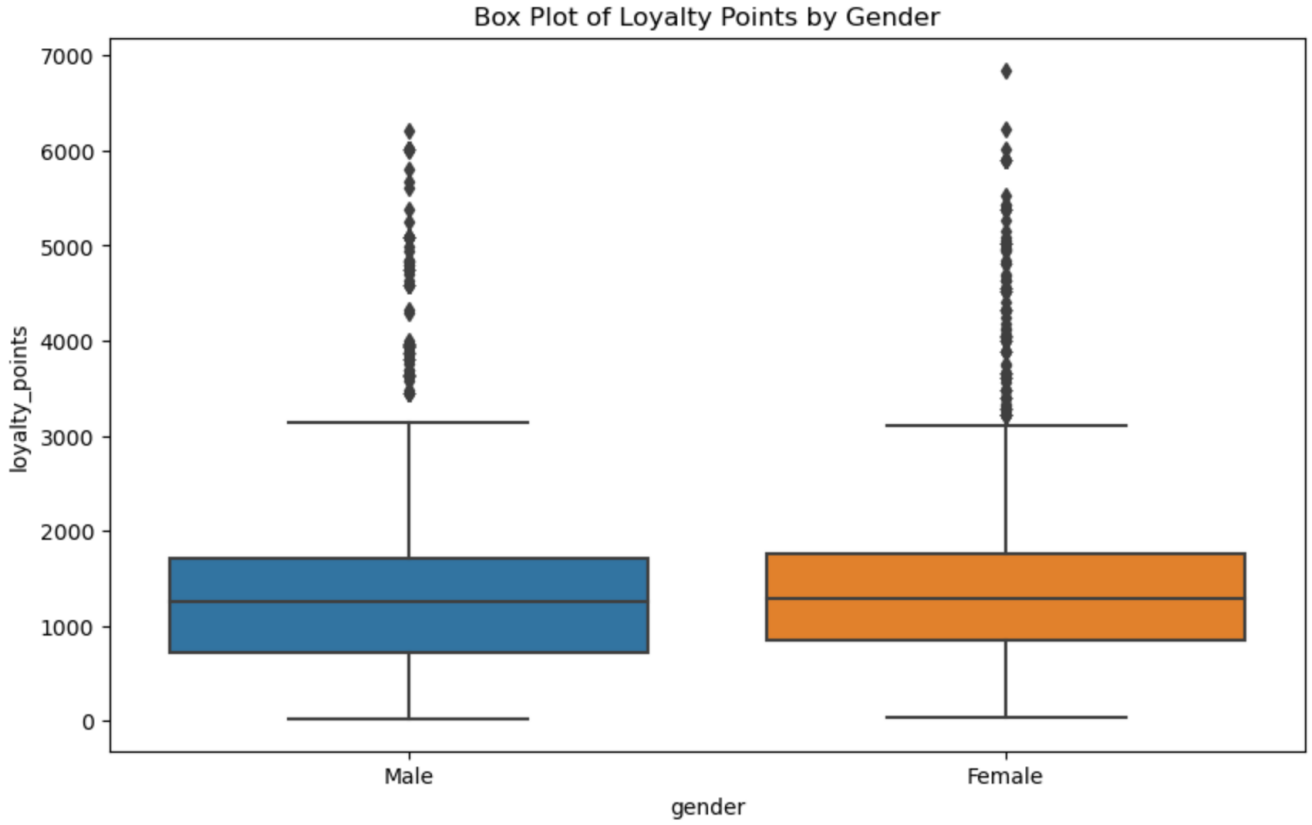


A clear pattern suggests that heteroscedasticity is present.



Deviation from the best fit line suggests that the residuals are not normally distributed.

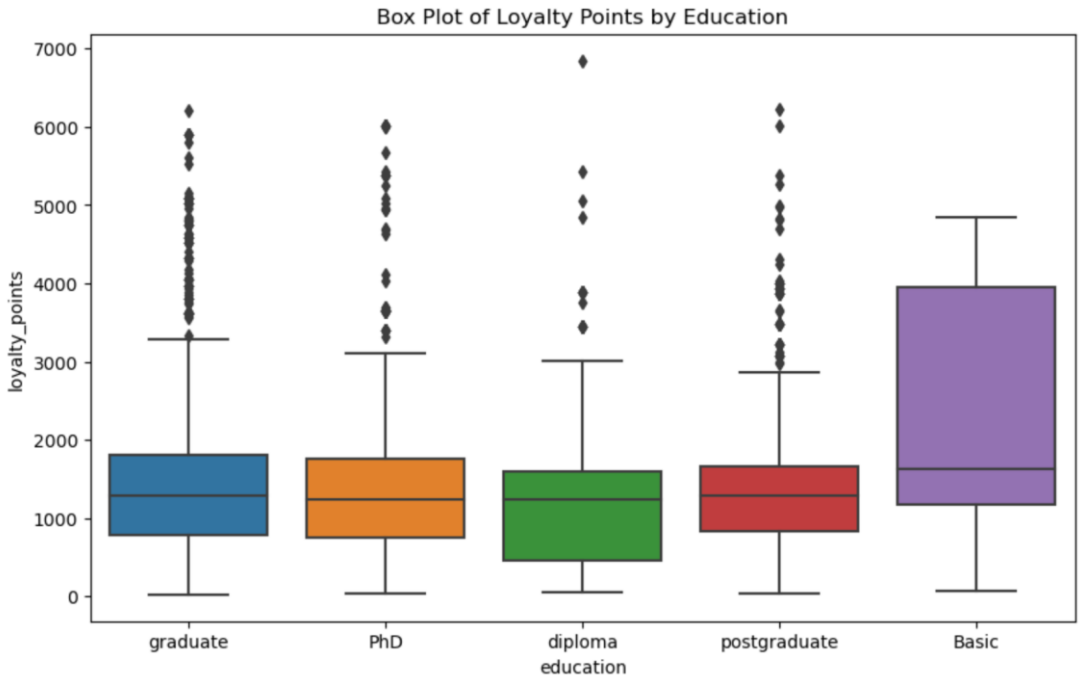
D. Gender vs Loyalty Points



gender	count	mean	std	min	25%	50%	75%	max
Female	1120.0	1601.166964	1251.215501	30.0	842.00	1281.0	1752.0	6847.0
Male	880.0	1548.587500	1323.008802	25.0	724.75	1248.0	1703.0	6208.0

Loyalty points are not dependent on gender.

E. Education vs Loyalty Points



education	count	mean	std	min	25%	50%	75%	max
Basic	50.0	2265.040000	1510.312347	66.0	1177.00	1622.0	3954.0	4837.0
PhD	460.0	1499.750000	1274.458561	30.0	752.00	1232.0	1756.0	6020.0
diploma	190.0	1336.021053	1162.759075	51.0	459.75	1239.0	1601.0	6847.0
graduate	900.0	1666.057778	1341.090733	25.0	780.00	1285.0	1800.0	6208.0
postgraduate	400.0	1499.077500	1136.152156	35.0	840.00	1281.0	1668.0	6232.0

It appears that those customers with a basic education are more likely to collect loyalty points.

Observations

Independent Variable	R-squared	F-stat and Prob (F-statistic)	Coefficients	Standard Errors and confidence intervals	Summary
Spending	Moderate. 45.2% of the variance in the loyalty points can be explained by spending. Other factors / variables are also affecting loyalty points.	High and extremely low respectively. The model is therefore statistically significant.	For each unit increase in x, y is expected to increase by 33.06 units, the coefficient is statistically significant with a p-value of almost 0.	Low and narrow respectively, suggesting that the estimate is precise.	Significant and positive relationship between spending and loyalty points. However, moderate R squared would suggest others factors in play.
Income	Moderate. 38% of the variance in the loyalty points can be explained by income. Other factors / variables are also affecting loyalty points.	High and extremely low respectively. The model is therefore statistically significant.	For each unit increase in x, y is expected to increase by 34.19 units, the coefficient is statistically significant with a p-value of almost 0.	Low and narrow respectively, suggesting that the estimate is precise.	Significant and positive relationship between income and loyalty points. However, moderate R squared would suggest others factors in play and high Durbin-Watson might require further investigation.
Age	Extremely Low. Age is not a good predictor of loyalty points.	Low and just above the confidence level of 0.05 respectively. The model is therefore **not** statistically significant. There is a weak relationship between age and loyalty points.	For each unit increase in x, y is expected to **decrease** by 4.013 units, the coefficient is **not** statistically significant with a p-value of 0.058.	Large and wide (also crosses 0) respectively, suggesting that the estimate is not precise.	There is a very limited relationship between age and loyalty points indicated by the statistically insignificant coefficient for age and a low R squared value. The residuals are not normally distributed as indicated by the Jarque-Bera, Kurtosis and Skew.

The simple linear regression R-squared values illustrate a moderate fit with the data.

Residuals

When the residuals in each model were assessed, heteroscedasticity appears to be present i.e. the variance of the residuals varies at different levels of the predictors. Furthermore, the Q-Q plots reveal a non-normal distribution of the residuals. Two of the basic assumptions for linear regression are not met.

Anticipate Outcomes and Improve the model.

1.3 Multiple linear Regression

A multiple linear regression analysis was conducted to assess whether combining independent variables could produce a more predictively accurate model.

Process	Action	Results	Observations
1.5.2 Multiple Linear Regression			
Independent and dependent variables defined.	Dependent variable - loyalty_points. Independent variables - spending_score, remuneration and age collectively.		
Initial multicollinearity check on entire dataframe to assess potential problematic variables conducted.	VIF analysis performed on full dataset.	Low multicollinearity between the independent variables. All retained.	
Model fitted using Scikit_learn.	mlr = linear_model.LinearRegression() mlr.fit(X, y)		
Predictions for x called.	mlr.predict(X)		
R-squared intercept and coefficients ascertained.	print("R-squared: ", mlr.score(X, y)) print("Intercept: ", mlr.intercept_) print("Coefficients:") list(zip(X, mlr.coef_))		
Predictions for new variable values performed.			
Data split into training and test sets.	x_train, x_test, y_train, y_test = sklearn.model_selection.train_test_split(X, y, test_size = 0.20 random_state = 42)		
Model produced using OLS for summary stats.	model = sm.OLS(y_train, sm.add_constant(x_train)).fit() y_pred = model.predict(sm.add_constant(x_test)) print_model = model.summary() print(print_model)		
Model trained on training data.	mlr = LinearRegression() mlr.fit(x_train, y_train)		
Predictions for independent variables (X) ascertained from training data.	y_pred_train = mlr.predict(x_train)		
Model tested on test data (Predictions for independent variables (X) ascertained from test data).	y_pred_test = mlr.predict(x_test)		
Model evaluated.	**Training data metrics** mae_train = metrics.mean_absolute_error(y_train, y_pred_train) mse_train = metrics.mean_squared_error(y_train, y_pred_train) r2_train = mlr.score(x_train, y_train) * 100 **Test data metrics** mae_test = metrics.mean_absolute_error(y_test, y_pred_test) mse_test = metrics.mean_squared_error(y_test, y_pred_test)	See results section.	
Final multicollinearity test conducted on training data.	VIF analysis performed on training dataset.	Low multicollinearity between the independent variables. All retained.	
Observations and insights considered.			See observations section

Results

Summary Statistics Comparison

Full data

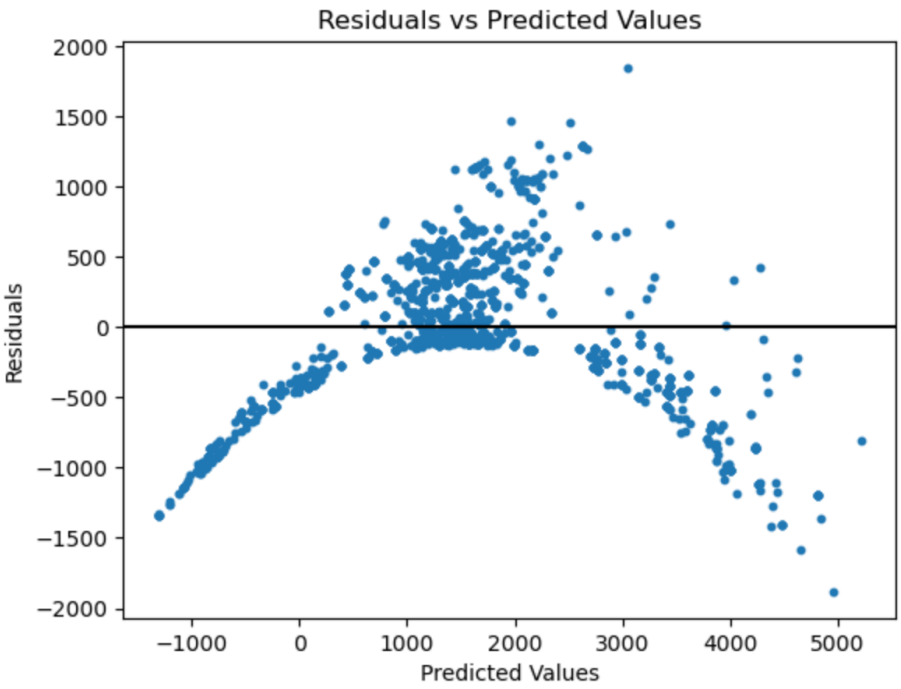
OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:				0.840
Model:	OLS	Adj. R-squared:				0.840
Method:	Least Squares	F-statistic:				3491.
Date:	Fri, 12 Jan 2024	Prob (F-statistic):				0.00
Time:	14:24:05	Log-Likelihood:				-15320.
No. Observations:	2000	AIC:				3.065e+04
Df Residuals:	1996	BIC:				3.067e+04
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P>|t|	[0.025	0.975]
const	-2203.0598	52.361	-42.075	0.000	-2305.747	-2100.372
spending	34.1832	0.452	75.638	0.000	33.297	35.070
income	34.0084	0.497	68.427	0.000	33.034	34.983
age	11.0607	0.869	12.730	0.000	9.357	12.765
Omnibus:	22.644	Durbin-Watson:				3.453
Prob(Omnibus):	0.000	Jarque-Bera (JB):				24.110
Skew:	0.227	Prob(JB):				5.82e-06
Kurtosis:	3.290	Cond. No.				377.

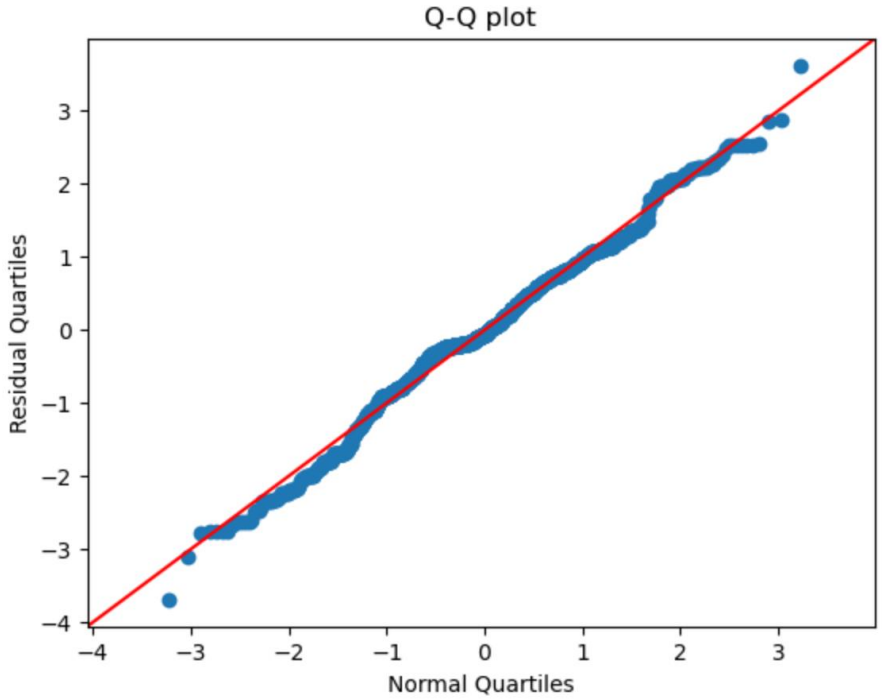
Training Set

OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:				0.842
Model:	OLS	Adj. R-squared:				0.842
Method:	Least Squares	F-statistic:				2846.
Date:	Fri, 12 Jan 2024	Prob (F-statistic):				0.00
Time:	14:24:05	Log-Likelihood:				-12246.
No. Observations:	1600	AIC:				2.450e+04
Df Residuals:	1596	BIC:				2.452e+04
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P>|t|	[0.025	0.975]
const	-2197.0105	58.134	-37.792	0.000	-2311.037	-2082.984
spending	33.9681	0.505	67.253	0.000	32.977	34.959
income	34.2457	0.552	62.004	0.000	33.162	35.329
age	11.0137	0.974	11.313	0.000	9.104	12.923
Omnibus:	14.722	Durbin-Watson:				2.050
Prob(Omnibus):	0.001	Jarque-Bera (JB):				15.856
Skew:	0.189	Prob(JB):				0.000360
Kurtosis:	3.308	Cond. No.				377.



A clear pattern suggests that heteroscedasticity is present.



Residuals are close to being normally distributed.

Observations

Multilinear Regression - Independent Variables: spending, income & age

	R-squared	F-stat and Prob (F-statistic)	Coefficients	Standard Errors and confidence intervals	Summary
Full data vs training data	Both models show similar R-squared values (84%, 84.2%). Strong and consistent explanatory power across the full and training data set.	High and extremely low respectively. The model is therefore statistically significant and consistent.	Similar across both models. Low p-values on both suggesting statistical significance.	Low and narrow respectively, suggesting that both models' coefficients are reliable and precise.	Most metrics are consistent suggesting the model is reliable and robust. The Durbin-Watson value is significantly different and may require further investigation.

Evaluation of Model Performance

	Metric	Training Data	Test Data
0	Mean Absolute Error	393.811746	402.235031
1	Mean Squared Error	260165.530533	277188.702332
2	R-squared (%)	84.248918	82.907234

The MAE values are relatively close between the training and test sets. This suggests that the model has a consistent performance in terms of average error per prediction, both on the data it was trained on and on unseen (test) data.

The MSE values are also relatively close, though slightly higher for the test set. This means that the model's overall error (considering the square of the errors) is again consistent across both datasets.

The R-squared values are quite high for both datasets, indicating that the model explains a significant portion of the variance in the dependent variable (loyalty points). The fact that these values are close (within about 1.34 percentage points) suggests good generalisability of the model.

To conclude, for different variable inputs of spending score, remuneration and age, the model shows good predictive power for loyalty points and generalises well to unseen (test) data.

Possible further investigation could be conducted using a validation dataset.

Residuals

When the residuals in MLR model were assessed, heteroscedasticity appears to be present i.e. the variance of the residuals varies at different levels of the predictors.

Multicollinearity

	VIF Factor	features
0	20.8	const
1	1.1	spending_score
2	1.0	remuneration
3	1.1	age

All variables show low multicollinearity.

Removal of Age as a dependent variable

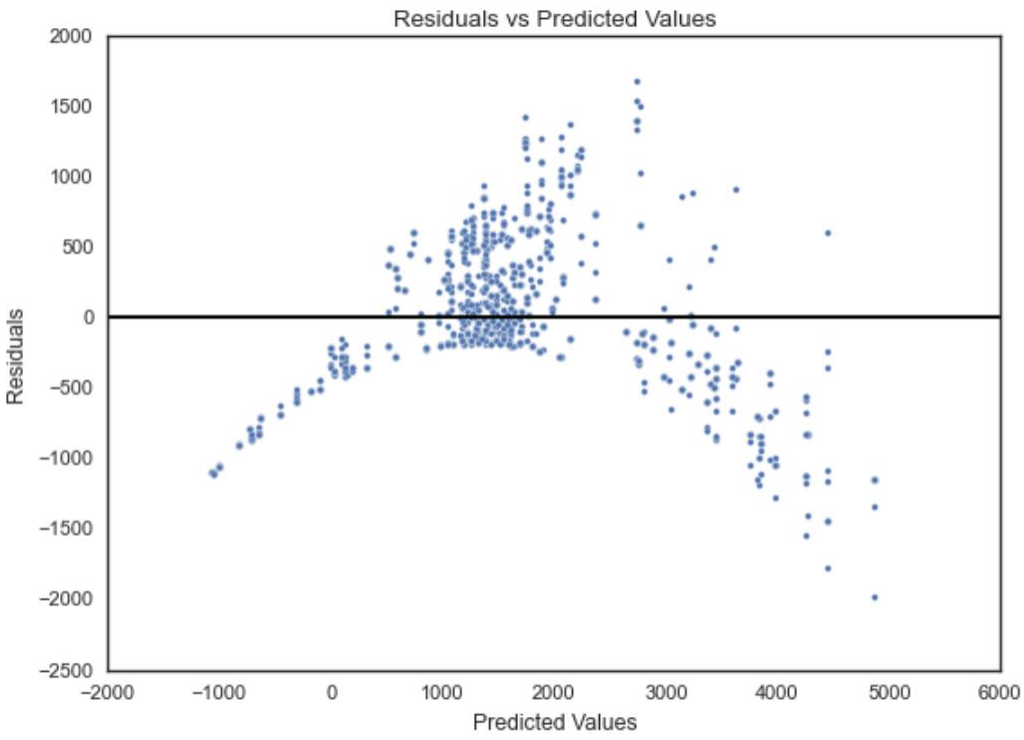
A second model was created removing age as an independent variable and the following summary statistics produced:

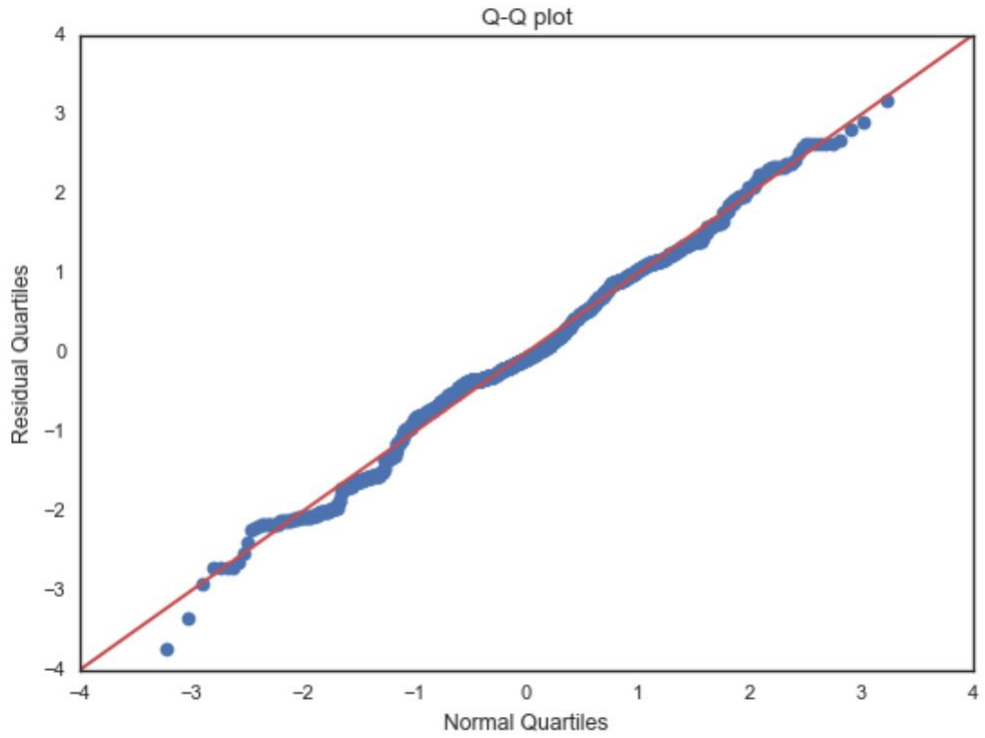
OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:	0.830
Model:	OLS	Adj. R-squared:	0.830
Method:	Least Squares	F-statistic:	3895.
Date:	Tue, 27 Feb 2024	Prob (F-statistic):	0.00
Time:	08:53:02	Log-Likelihood:	-12307.
No. Observations:	1600	AIC:	2.462e+04
Df Residuals:	1597	BIC:	2.464e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1700.3237	39.588	-42.950	0.000	-1777.974	-1622.674
spending_score	32.6439	0.510	63.947	0.000	31.643	33.645
remuneration	34.3346	0.574	59.838	0.000	33.209	35.460

Omnibus:	2.977	Durbin-Watson:	2.034
Prob(Omnibus):	0.226	Jarque-Bera (JB):	2.923
Skew:	0.075	Prob(JB):	0.232
Kurtosis:	3.147	Cond. No.	220.





Removing age as an independent variable has no material effect on residuals.

The model was tested, and the following metrics ascertained:

	Metric	Training Data	Test Data
0	Mean Absolute Error	411.564600	429.663620
1	Mean Squared Error	281026.485091	300944.091783
2	R-squared (%)	82.985943	81.442364

Evaluation of Model Performance

- The model incorporating Age outperforms, evidenced by its lower MAE and MSE.
- It also shows higher R-squared and adjusted R-squared values, indicating a stronger ability to explain the variance in loyalty points. Despite Age's low correlation in simple linear regression, its inclusion, especially when combined with spending score and remuneration in multiple linear regression, significantly enhances the model's predictive accuracy and fit to the data.

How do customers engage with and accumulate loyalty points?

Look Back and Learn

Spending Behavior: A crucial driver of loyalty points accumulation. Customers who frequently engage with the business and have higher spending scores accumulate more points, emphasizing the importance of encouraging customer spending to boost loyalty.

Financial Capacity: Remuneration, serving as a proxy for financial capacity, shows that customers with higher earnings tend to accumulate more loyalty points. This might reflect both a greater capacity for spending and possibly a tendency to engage in behaviors or purchase products that are rewarded by the loyalty program.

Age Factor: While not as strong a predictor as spending score and remuneration when considered alone, the positive impact of age on loyalty points suggests that customer engagement could also be influenced by the length of the relationship with the business or by life stage-related spending habits.

Model Preference: . Both models are effective. However, the model including age has a slightly higher R-squared, indicating a better fit for predicting loyalty points accumulation

Conclusion: Customers engage with and accumulate loyalty points primarily through their spending behaviors and financial capacity, with age also playing a role.

Enhancing customer engagement strategies, therefore, could focus on maximizing spending opportunities and tailoring loyalty programs to cater to the spending capacity and lifestyle of different customer segments.

2. Customer Segmentation

Identify the Problem and Opportunities

TG hasn't optimally segmented their customers based on shared characteristics.

Customer segmentation can provide the following opportunities:

- Targeted marketing
- Improved pricing strategies

Define Goals

To employ K means clustering to segment customers into clusters, where each cluster represents a collection of data points that are similar to each other more than to those in other clusters. To determine the optimal number of clusters.

Explore

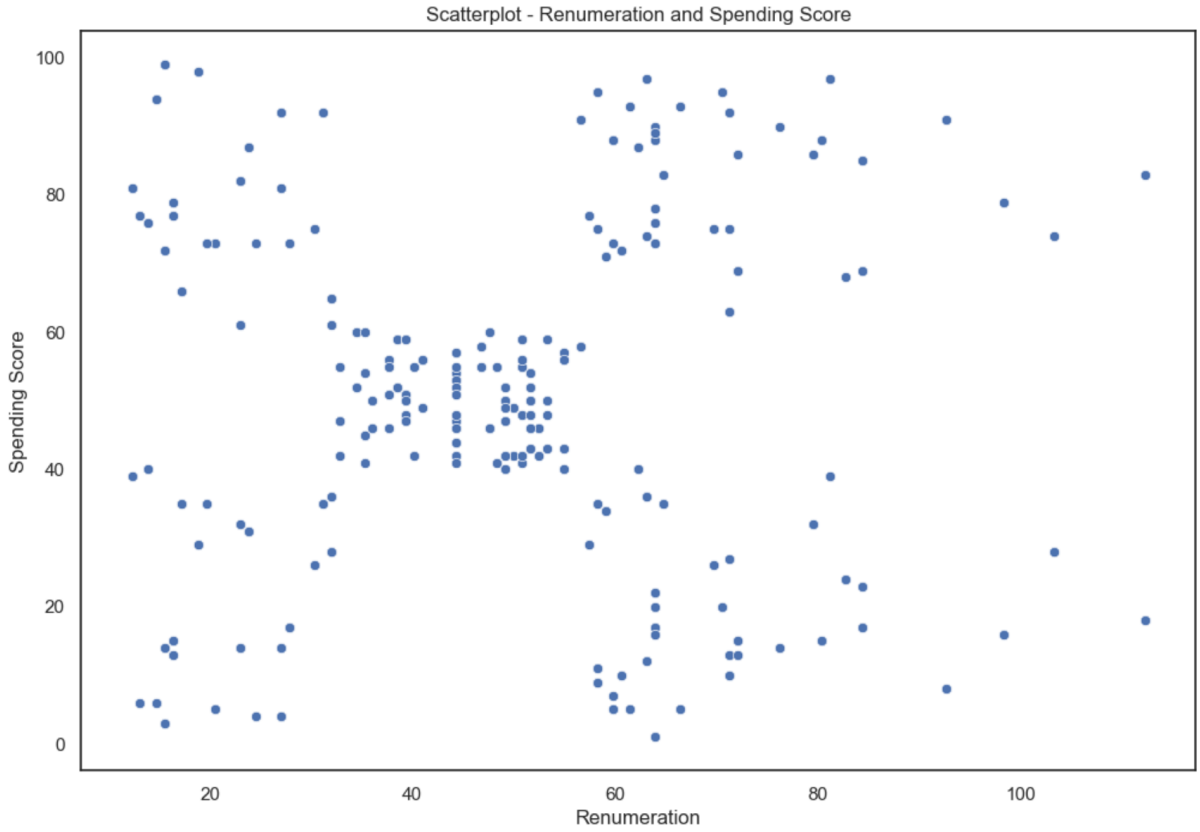
The following process was used to explore the data.

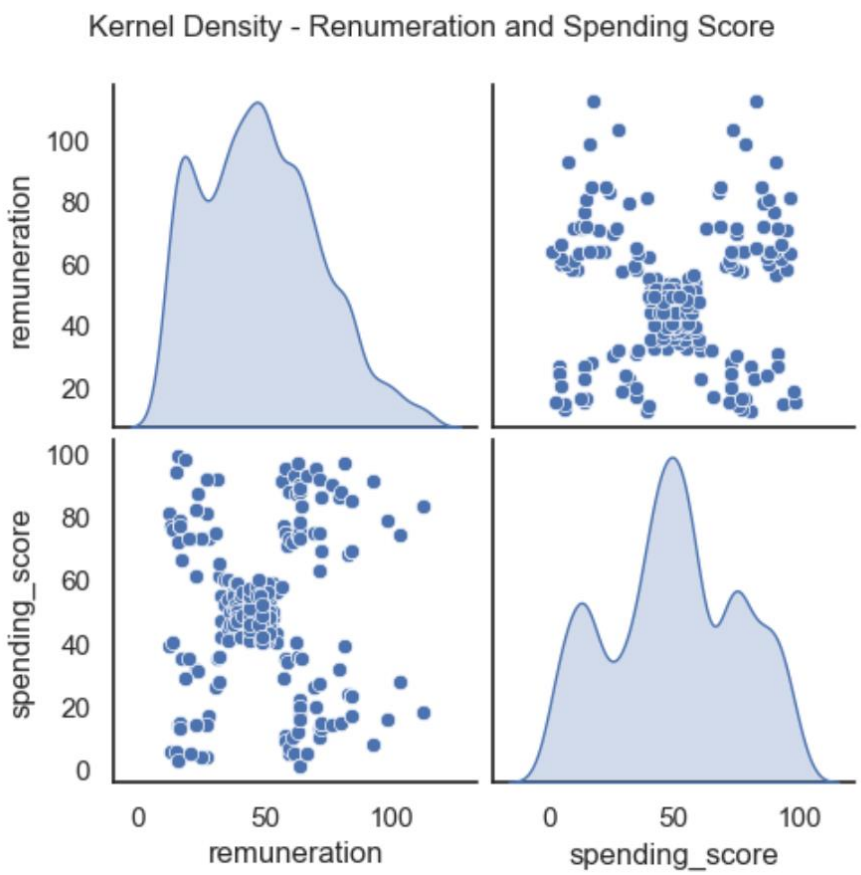
Process	Action	Results	Observations
2.1 Load and Explore data			
Necessary packages imported.			
Data loaded.	New df, rf2 with remuneration and spending_score		
Data explored.	print(rf2.info())	2 columns, 2000 records.	
Descriptive statistics produced.	print(rf2.describe())		
2.2 Plot			
Scatterplot	sns.scatterplot(x='remuneration', y='spending_score',data=rf2)	See results.	
Pair plot	x = rf2[['remuneration', 'spending_score']] sns.pairplot(rf2, vars=x, diag_kind='kde')	See results.	
2.3 Elbow and Silhouette Methods			
Elbow method	Sum of squares empty list (ss) created. List ss populated with inertia values using a for loop with number of clusters ranging from 1 - 10. Elbow chart plotted.	Fig 2.3a	Obvious elbow at 5 clusters
Silhoutette method	Empty list (sil) created to store silhouette scores. Max clusters 10. For loop iterates over different cluster sizes. For each iteration a k-means fitted to data (x) with k number of clusters. Silhouette score calculated and appended in sil list. Silhouette score plotted - max values illustrate how similar an object is to its own cluster when compared to other clusters.	Fig 2.3b	Obvious peak at 5 clusters
2.4 Evaluate k-means at different values of k.	4, 6 and finally 5 clusters were evaluated.		
Kernel Density plots		See results.	
Value count K-means predicted	x['K-Means Predicted'].value_counts()		
View K-means predicted df.	print(x.head())		
2.5 Fit Final Model and Justify	5 Clusters used. K means predicted counted and df printed.	See results.	
2.6 Plot and Interpret Clusters	Clusters plotted and interpreted		
2.7 Observations and Insights			See observations and insights.

2.1 Descriptive Statistics

	remuneration	spending_score
count	2000.000000	2000.000000
mean	48.079060	50.000000
std	23.123984	26.094702
min	12.300000	1.000000
25%	30.340000	32.000000
50%	47.150000	50.000000
75%	63.960000	73.000000
max	112.340000	99.000000

2.2 Scatter and Pair Plot



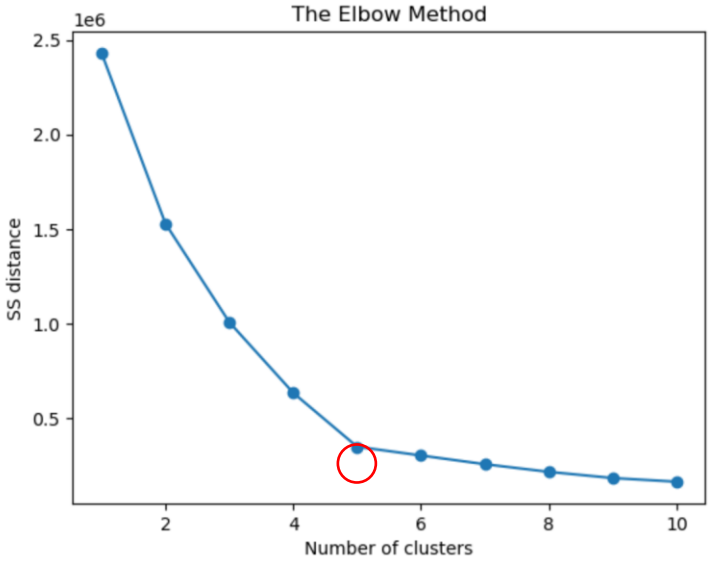

Kernel Density - Renumeration and Spending Score
remuneration
spending_score
100
80
60
40
20
0
0
50
100
remuneration
spending_score

2.3 Optimum k-value

Elbow Method

Fig 2.3a

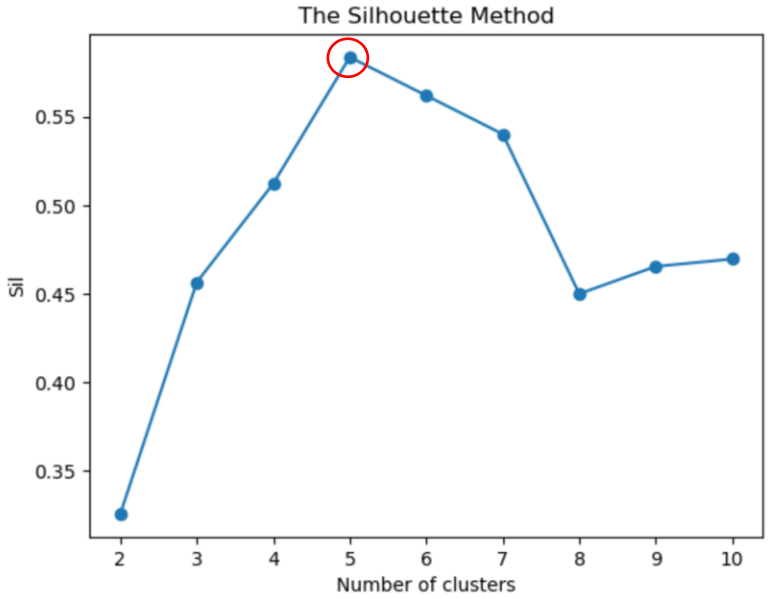
The elbow method indicates a clear 'elbow' at 5 clusters.



Silhouette Method

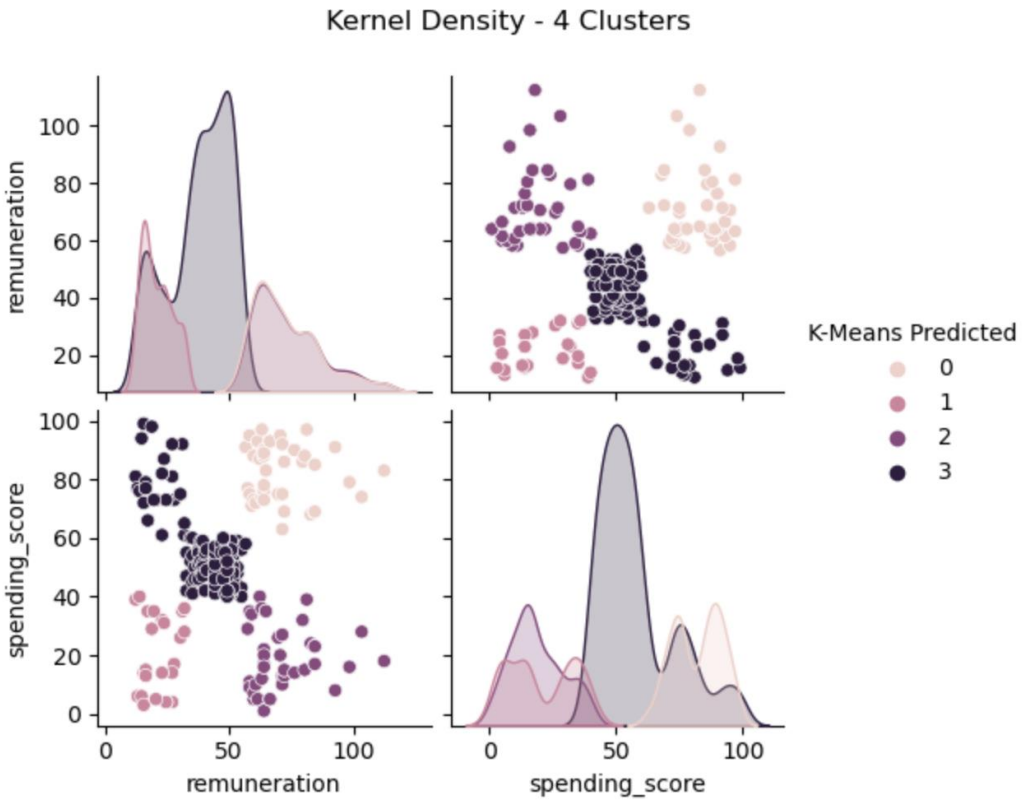
Fig 2.3b

The silhouette method indicates a clear peak at 5 clusters. The peak value indicates that an object within the cluster is well matched to its own cluster and poorly matched to neighbouring clusters.



2.4 Evaluate k-means at different values of k.

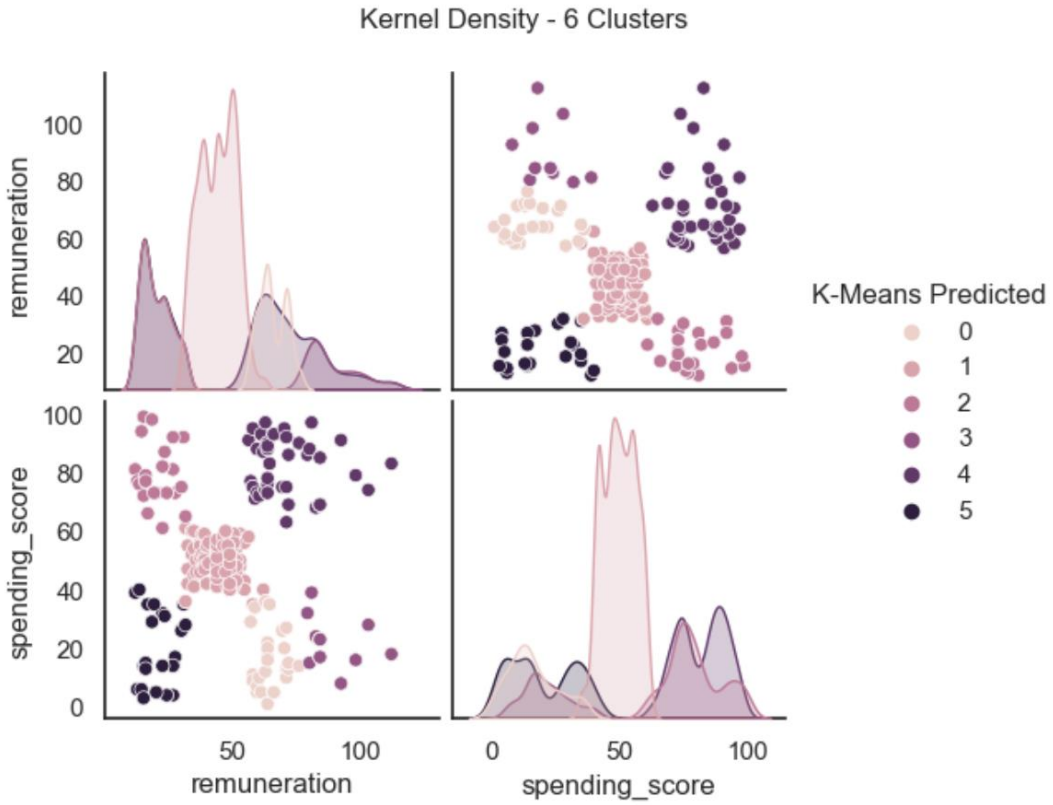
K = 4



K-Means Predicted	
3	1013
0	356
2	351
1	280

Cluster 3 is too large and encapsulates 2 distinct clusters in the central region and the high spending / low remuneration region.

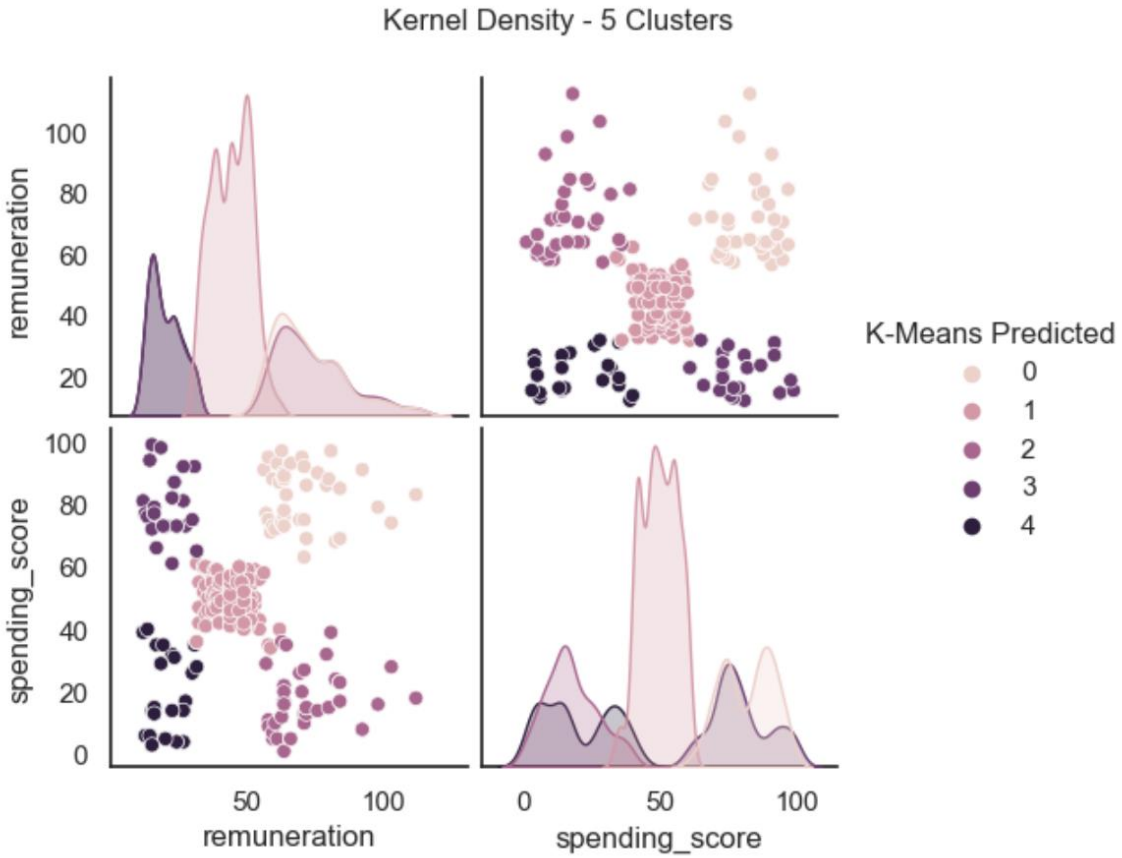
K = 6



K-Means Predicted	
1	767
4	356
5	271
2	269
0	214
3	123

Cluster 3 is considerably smaller and should be incorporated within cluster 0.

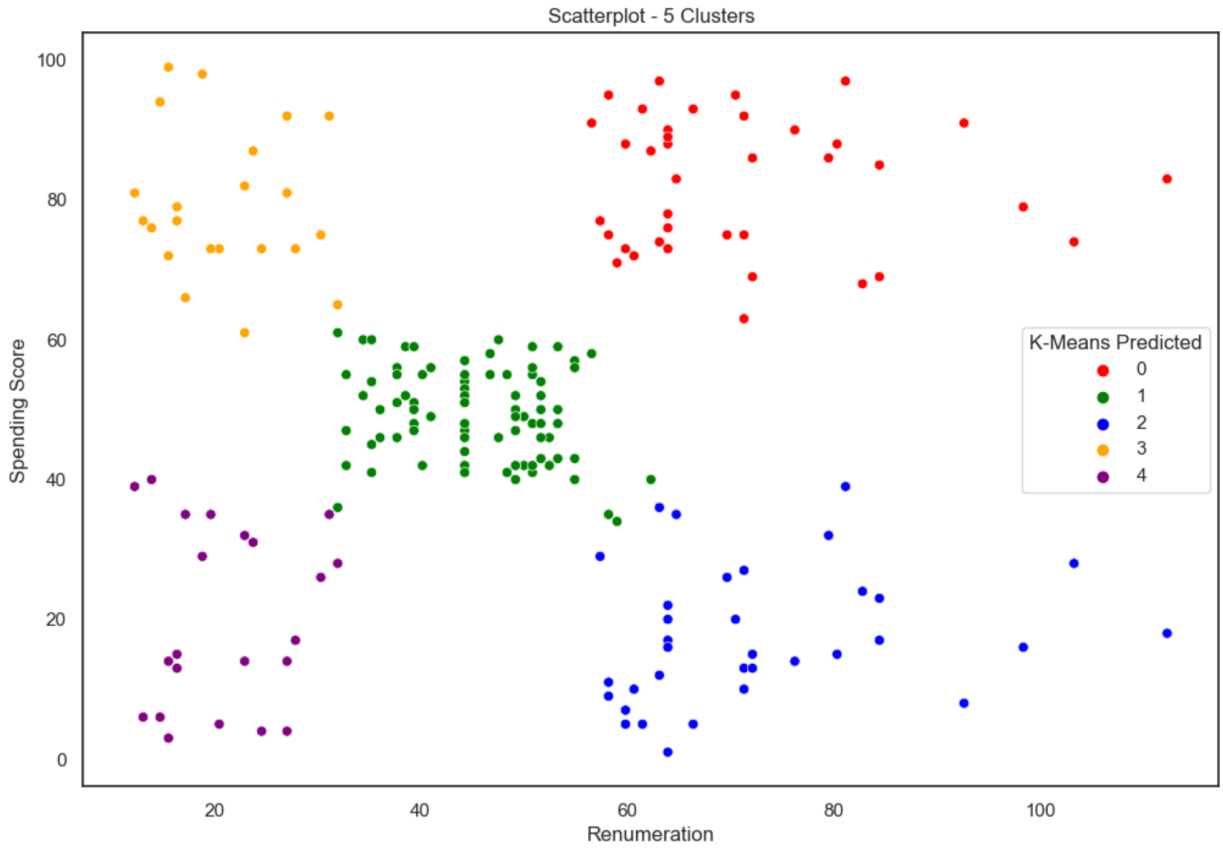
2.5 Fit Final Model with k = 5



	K-Means Predicted
1	774
0	356
2	330
4	271
3	269

Considering the results of the elbow and silhouette methods, 5 clusters appear optimal.

2.6 Plot and Interpret Clusters



How can customers be segmented into groups, and which groups can be targeted by the marketing department?

The customers can be categorised optimally into the following 5 segments.

Cluster	Customer Type	Value
1	Core	774
0	High Income, High Spend	356
2	High income, Low spend	330
4	Low income, low spend	271
3	Low income, high spend	269

The core group in the centre represents the largest cluster of customers with a 774 value count. The high-income groups are evenly distributed between high and low spend as are the low-income groups which represent the smallest clusters within Turtle's customer group.

Observations and Insights

Specific actions could be orchestrated to target each individual customer category with regards to marketing strategies, product customisation, customer relationship management, growth potential and revenue maximisation.

Core

The largest cluster, this group represents the most common combination of income and spending patterns among Turtle's customers. With moderate income and spending levels, but the most numerous, understanding the characteristics of this group is key since they form a significant portion of Turtle's customer base.

High Income, High Spend

The second largest customer category. Potential Opportunities:

- Marketing: Targeted with premium products and exclusive offers.
- Product Customisation: Focus on more luxury products.
- Customer relationship Management: personalised services and exclusivity.

High Income, Low Spend

- Perhaps low frequency purchasers of high value products.
- Identifying high-income, low-spending customers can be an opportunity to introduce products or services that might appeal to this segment, thereby potentially increasing their spending.
- Growth focus category. Potential revenue maximization opportunities.

Low Income, Low Spend

- Possible inducement to spend more through targeted marketing of budget friendly options and promotional offers.
- Receptive to loyalty programs and cost-saving tips.

Low Income, High Spend

- These customers may be more receptive to budget-friendly deals, promotional offers.
- Customise for more economical options.
- Receptive to loyalty programs and cost-saving tips.
- Potential credit risks associated with high spenders in relation to income.

Look back and learn.

Conclusion

Customer segmentation can help improve sales through:

- **Targeted marketing:** By understanding the distinct needs, preferences, and behaviors of each segment, businesses can tailor their marketing messages more effectively.
- **Improved Product Development:** Segmentation allows companies to identify specific needs and gaps in the market that they can address with new or improved products.
- **Enhanced Customer Experience:** With customer segmentation, companies can provide a more personalized experience to their customers, whether through customized communications, tailored product recommendations, or bespoke services.
- **Efficient Allocation of Resources:** By focusing on the most profitable segments or those with the highest growth potential, businesses can allocate their resources more efficiently.
- **Competitive Advantage:** Segmentation helps businesses understand their customers better than their competitors, offering insights into customer needs that may not be apparent on the surface.
- **Pricing Strategies:** Different customer segments may have different sensitivities to pricing. Through segmentation, businesses can adopt varied pricing strategies that match the purchasing power and perceived value for each segment, optimizing revenue opportunities across different parts of the market.
- **Cross-Selling and Up-Selling Opportunities**

3. Customer Sentiment Analysis

Identify the Problem and Opportunities

TG doesn't have a full understanding of customer sentiment towards their products, services and brand.

Sentiment Analysis can provide the following opportunities:

- Identify high value customers with negative experiences.
- Identify potential growth customers.
- Differentiate products based on positive and negative reviews.

Define Goals

To employ NLP to understand sentiment in the review data and identify the opportunities above.

Explore

Sentiment analysis was conducted using the following process:

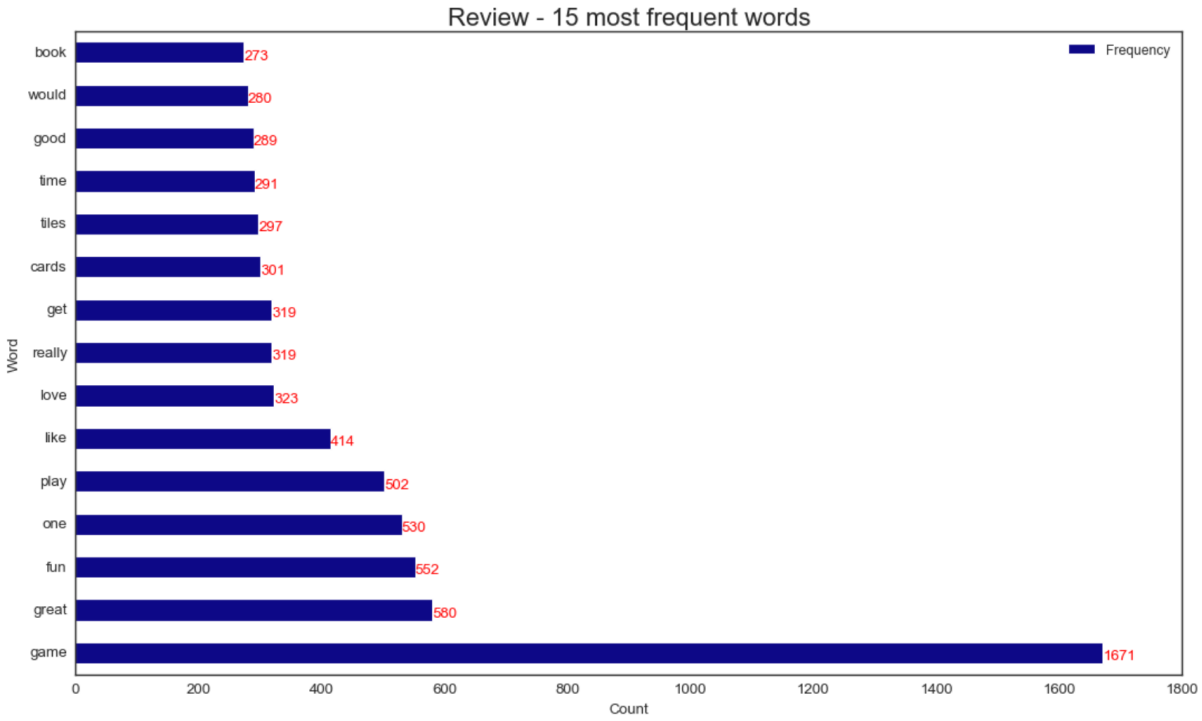
Process	Action	Results	Observations
3.1 Load and Explore Data	New dataframe df3 created. Only review and summary columns retained. df3 = rf[['review', 'summary']]		
Explore the data	df3.info(), df3.isna().sum()	2000 non null records, no null values. Data types - objects	
3.2 Prepare data for NLP			
Change to lower case and join elements with a space.	df3['review'] = df3['review'].apply(lambda x: " ".join(x.lower() for x in x.split()))	All uppercase letters replaced in review and summary columns	
Replace punctuation in each of the columns respectively (review and summary)	df3['review'] = df3['review'].str.replace('[^\w\s]','')	Punctuation marks removed.	
Drop duplicates in both columns	df3 = df3.drop_duplicates(subset=['review', 'summary']) df3.reset_index(inplace=True)	39 duplicated removed.	Duplictaes were removed only if they occurred in both columns simultaneously. There could be numerous duplicates such as "5 stars" that form part of different records.
Tokensise and remove non alphanumeric and stopwords	New df4 created		
Tokenisation applied using word_tokenise	df4['review_tokens'] = df4['review'].apply(word_tokenize) df4['summary_tokens'] = df4['summary'].apply(word_tokenize)		
Remove non alphanumeric items	E.g. review_tokens = [word for word in all_review_tokens if word.isalnum()]		
Create a set of English stopwords.	english_stopwords = set(stopwords.words('english'))		
Create a filtered list of tokens without stopwords	E.g. review_tokens2 = [x for x in review_tokens if x.lower() not in english_stopwords]		
Define an empty string variable. Add each filtered token word to the string.	review_tokens2_string = ' ' for value in review_tokens: review_tokens2_string = review_tokens2_string + value + ' '		
3.3 Frequenct Distribution and Word Clouds			
Frequency distribution created Word clouds created	E.g. fdist_review = FreqDist(review_tokens2) wordcloud = WordCloud(width = 1600, height = 900, background_color ='white', colormap='plasma', min_font_size = 10).generate(review_tokens2_string)		
15 most common words determined and plotted.	counts_review = pd.DataFrame(Counter(review_tokens2).most_common(15), columns=['Word', 'Frequency']).set_index('Word')		

3.4 Review polarity and sentiment
Function to extract polarity
 	
Polarity determined
 	
Subjectivity determined
 	
Histograms of Polarity and Subjectivity produced
	def generate_polarity(comment):
	  '''Extract polarity score (-1 to +1) for each comment'''
	  return TextBlob(comment).sentiment[0]
	E.g. df4['review_polarity'] = df4['review'].apply(generate_polarity)
	def generate_subjectivity(comment):
	  return TextBlob(comment).sentiment[1]
	df4['review_subjectivity'] = df4['review'].apply(generate_subjectivity)		

3.5 Identify top positive and negative reviews and Summaries	E.g review_negative_sentiment = df4.nsmallest(20, 'review_polarity')		
3.6 Using Sentiment Analysis to improve overall sales Identify the relationship between Summary Polarity and Spending Score Identify dissatisfied high spenders and satisfied low spenders Identify Loyal Customers giving bad reviews Identify Products with the best and worst reviews	Scatterplot of summary polarity and spending score Define thresholds for high and low spenders eg 25% upper and lower quartile. Identify summary polarity less than -0.5 for high spenders and above 0.5 for low spenders. Define thresholds for loyal customers eg top 50% . Identify summary polarity less than -0.5 for high spenders. Group by product and mean summary_polarity: product_summaries = df4.groupby('product')['summary_polarity'].mean().reset_index(). Sort by highest and lowest.		

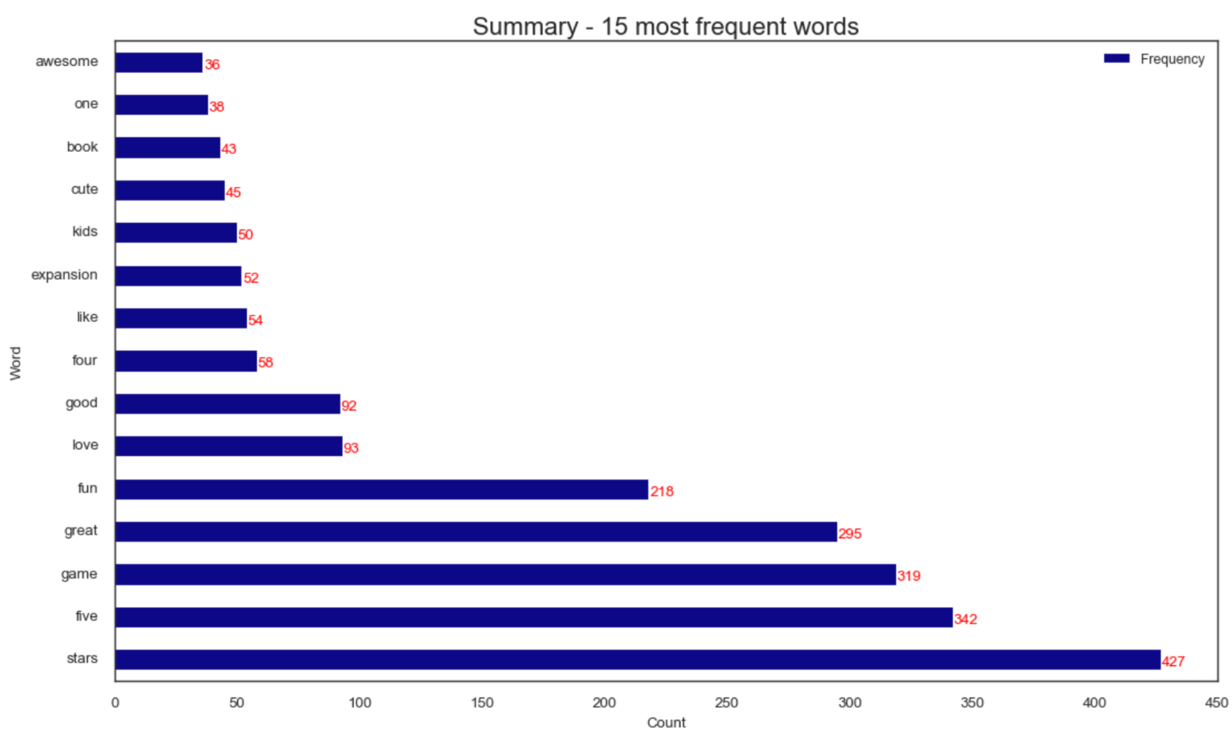
3.1 Word clouds and Frequency Distributions

Review

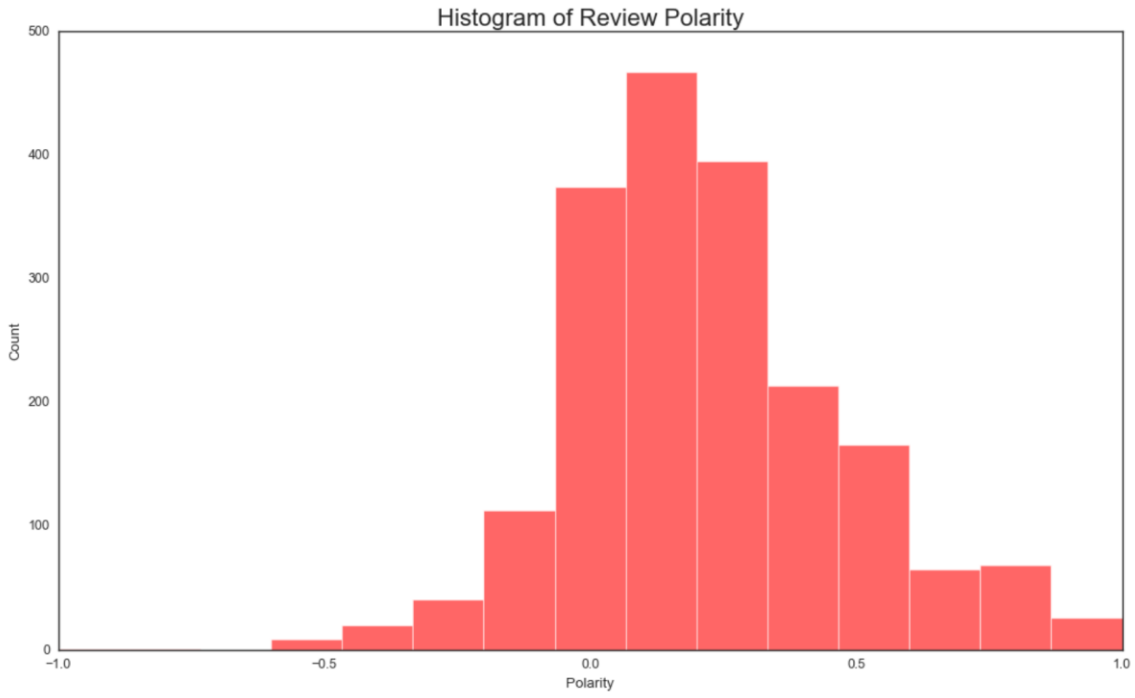


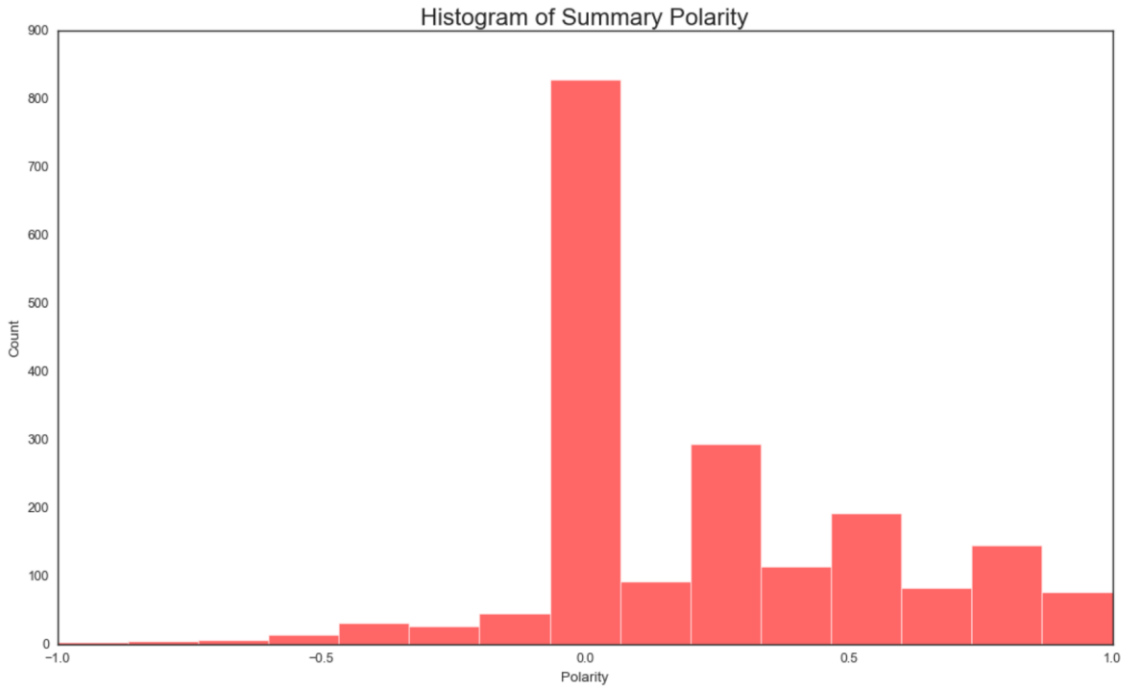
Summary


five stars
fun
game
great
stars
five
love
good
four stars
great game
stars great
expansion
kid
excellent
book
cute
fun game
awesome
really
product
quality
one


Summary - 15 most frequent words
Frequency
Word
Count
awesome 36
one 38
book 43
cute 45
kids 50
expansion 52
like 54
four 58
good 92
love 93
fun 218
great 295
game 319
five 342
stars 427
0
50
100
150
200
250
300
350
400
450

3.2 Polarity


Histogram of Review Polarity
Count
Polarity
500
400
300
200
100
0
−1.0
−0.5
0.0
0.5
1.0


Histogram of Summary Polarity
Count
Polarity
900
800
700
600
500
400
300
200
100
0
−1.0
−0.5
0.0
0.5
1.0

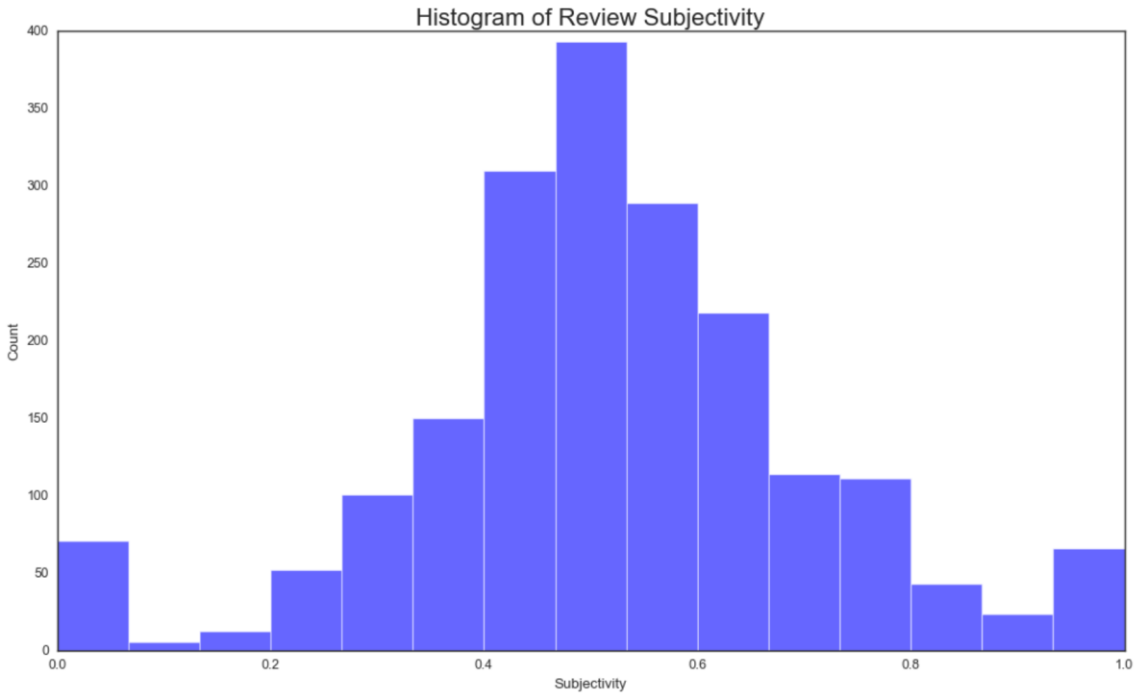
Polarity Statistics

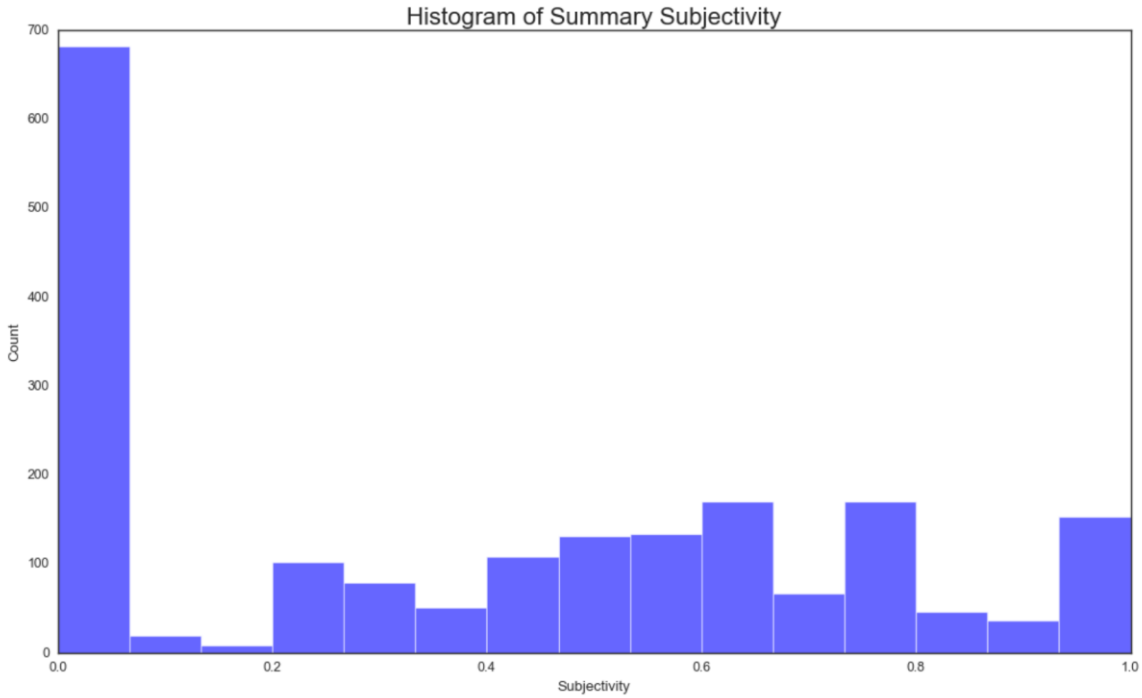
	review_polarity	summary_polarity
count	1961.000000	1961.000000
mean	0.213170	0.223678
std	0.260360	0.337507
min	-1.000000	-1.000000
25%	0.045833	0.000000
50%	0.177222	0.100000
75%	0.351562	0.475000
max	1.000000	1.000000

Overall Sentiment

- Broad spectrum of sentiment
- Overall positive average sentiment.
- Summaries are in the main more positive.
- Summaries tend to exhibit a wider range of sentiment compared to reviews, as indicated by the higher standard deviation, and more positively skewed upper quartile, despite having a slightly more positive mean sentiment. This could be due to the summarization process itself, where there's an attempt to encapsulate the essence of the review, leading to more pronounced sentiments.
- The presence of extreme values (both -1 and 1) highlights the diversity in perceptions and experiences among the respondents.

3.3 Subjectivity





Subjectivity Statistics

	review_subjectivity	summary_subjectivity
count	1961.000000	1961.000000
mean	0.516741	0.385615
std	0.192895	0.340746
min	0.000000	0.000000
25%	0.423637	0.000000
50%	0.508333	0.400000
75%	0.604286	0.650000
max	1.000000	1.000000

Overall Subjectivity

- Reviews are more subjective than summaries, which aligns with the expectation that reviews would contain more personal opinions.
- Summaries show a wide range of subjectivity, indicated by a higher standard deviation, with a notable portion being completely objective. This suggests that while some summaries may aim to objectively present information, others incorporate subjective views.
- The presence of completely objective (score of 0) and completely subjective (score of 1) entries in both reviews and summaries highlights the diversity of content in terms of subjectivity/objectivity.

3.4 Identifying Positive and Negative Reviews

The Top 20 positive and negative reviews for review and summary were determined (miscategorized comments are highlighted):

Positive Reviews

	review	review_polarity	review_subjectivity
7	came in perfect condition	1.000000	1.000000
164	awesome book	1.000000	1.000000
193	awesome gift	1.000000	1.000000
489	excellent activity for teaching selfmanagement skills	1.000000	1.000000
517	perfect just what i ordered	1.000000	1.000000
583	wonderful product	1.000000	1.000000
601	delightful product	1.000000	1.000000
613	wonderful for my grandson to learn the resurrection story	1.000000	1.000000
782	perfect	1.000000	1.000000
923	awesome	1.000000	1.000000
1119	awesome set	1.000000	1.000000
1150	best set buy 2 if you have the means	1.000000	0.300000
1159	awesome addition to my rpg gm system	1.000000	1.000000
1282	its awesome	1.000000	1.000000
1380	one of the best board games i played in along time	1.000000	0.300000
1523	my daughter loves her stickers awesome seller thank you	1.000000	1.000000
1580	this was perfect to go with the 7 bean bags i just wish they were not separate orders	1.000000	1.000000
1684	awesome toy	1.000000	1.000000
1689	it is the best thing to play with and also mind blowing in some ways	1.000000	0.300000
1695	excellent toy to simulate thought	1.000000	1.000000

Negative Reviews

	review	review_polarity	review_subjectivity
207	booo unles you are patient know how to measure i didnt have the patience neither did my daughter boring unless you are a craft person which i am not	-1.000000	1.000000
181	incomplete kit very disappointing	-0.780000	0.910000
1773	im sorry i just find this product to be boring and to be frank juvenile	-0.583333	0.750000
362	one of my staff will be using this game soon so i dont know how well it works as yet but after looking at the cards i believe it will be helpful in getting a conversation started regarding anger and what to do to control it	-0.550000	0.300000
116	i bought this as a christmas gift for my grandson its a sticker book so how can i go wrong with this gift	-0.500000	0.900000
226	this was a gift for my daughter i found it difficult to use	-0.500000	1.000000
229	i found the directions difficult	-0.500000	1.000000
289	instructions are complicated to follow	-0.500000	1.000000
300	difficult	-0.500000	1.000000
1501	expensive for what you get	-0.500000	0.700000
173	i sent this product to my granddaughter the pompom maker comes in two parts and is supposed to snap together to create the pompoms however both parts were the same making it unusable if you cant make the pompoms the kit is useless since this was sent as a gift i do not have it to return very disappointed	-0.491667	0.433333
345	my 8 yearold granddaughter and i were very frustrated and discouraged attempting this craft it is definitely not for a young child i too had difficulty understanding the directions we were very disappointed	-0.446250	0.533750
531	i purchased this on the recommendation of two therapists working with my adopted children the children found it boring and put it down half way through	-0.440741	0.485185
305	very hard complicated to make these	-0.439583	0.852083
421	kids i work with like this game	-0.400000	0.400000
430	this game although it appears to be like uno and have an easier play method it was still too time consuming and wordy for my children with learning disabilities	-0.400000	0.400000
490	my son loves playing this game it was recommended by a counselor at school that works with him	-0.400000	0.400000
795	this game is a blast	-0.400000	0.400000
798	i bought this for my son he loves this game	-0.400000	0.400000
814	was a gift for my son he loves the game	-0.400000	0.400000

Positive Summaries

	summary	summary_polarity	summary_subjectivity
6	best gm screen ever	1.000000	0.300000
28	wonderful designs	1.000000	1.000000
32	perfect	1.000000	1.000000
80	theyre the perfect size to keep in the car or a diaper	1.000000	1.000000
133	perfect for preschooler	1.000000	1.000000
139	awesome sticker activity for the price	1.000000	1.000000
160	awesome book	1.000000	1.000000
162	he was very happy with his gift	1.000000	1.000000
186	awesome	1.000000	1.000000
209	awesome and welldesigned for 9 year olds	1.000000	1.000000
412	perfect	1.000000	1.000000
468	excellent	1.000000	1.000000
536	excellent	1.000000	1.000000
541	excellent therapy tool	1.000000	1.000000
572	the pigeon is the perfect addition to a school library	1.000000	1.000000
591	best easter teaching tool	1.000000	0.300000
639	wonderful	1.000000	1.000000
643	all f the mudpuppy toys are wonderful	1.000000	1.000000
649	awesome puzzle	1.000000	1.000000
654	not the best quality	1.000000	0.300000

Negative Summaries

	summary	summary_polarity	summary_subjectivity
21	the worst value ive ever seen	-1.000000	1.000000
207	boring unless you are a craft person which i am	-1.000000	1.000000
819	boring	-1.000000	1.000000
1148	before this i hated running any rpg campaign dealing with towns because it	-0.900000	0.700000
1	another worthless dungeon masters screen from galeforce9	-0.800000	0.900000
143	disappointed	-0.750000	0.750000
623	disappointed	-0.750000	0.750000
785	disappointed	-0.750000	0.750000
1591	disappointed	-0.750000	0.750000
361	promotes anger instead of teaching calming methods	-0.700000	0.200000
875	too bad this is not what i was expecting	-0.700000	0.666667
880	bad qualityall made of paper	-0.700000	0.666667
177	at age 31 i found these very difficult to make	-0.650000	1.000000
100	small and boring	-0.625000	0.700000
511	mad dragon	-0.625000	1.000000
797	disappointing	-0.600000	0.700000
1001	disappointing	-0.600000	0.700000
1099	disappointing	-0.600000	0.700000
1773	disappointing	-0.600000	0.700000
991	then you will find this board game to be dumb and boring	-0.591667	0.633333

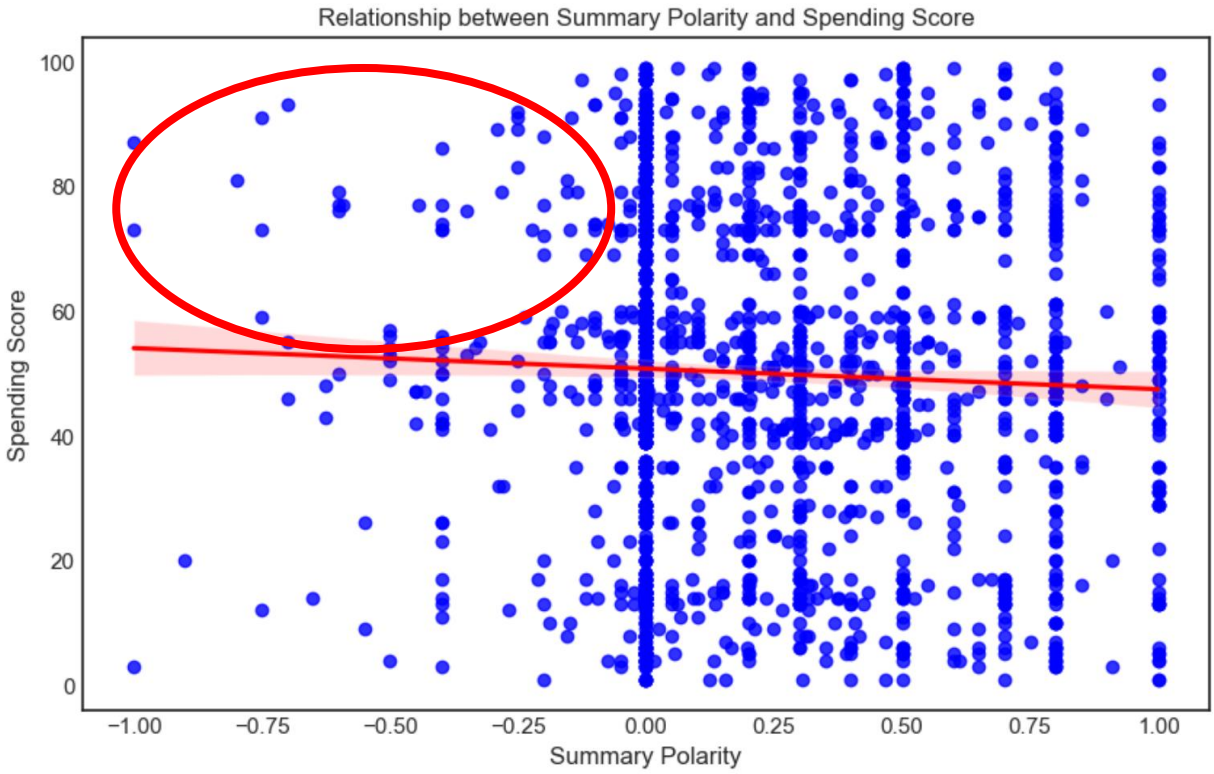
We have established that summaries tend to be less subjective and reflect the general sentiment more concisely and contain fewer erroneous categorisations. Summary polarity will be used in the following analysis.

Anticipate

3.5 Using Sentiment Analysis to improve Overall Sales

1.0 Identify High Spenders that are Dissatisfied with their Purchases.

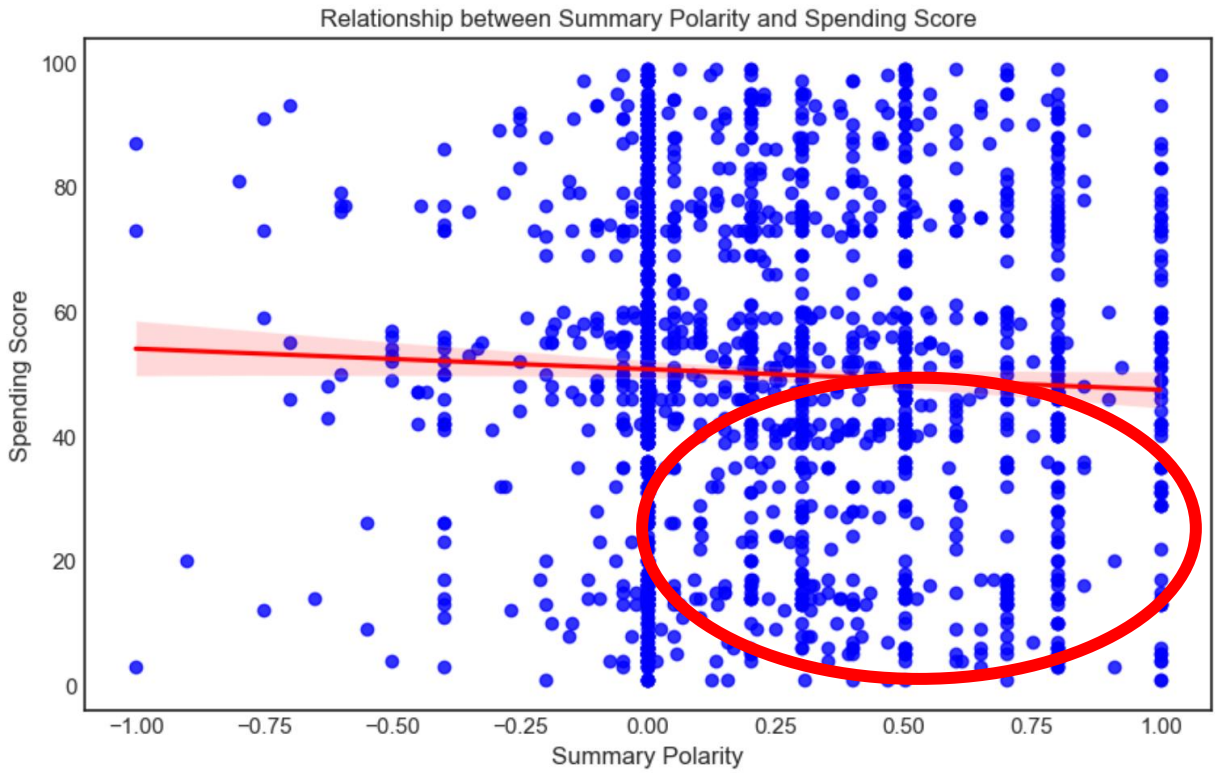
High-spending customers at Turtle Games, who are dissatisfied with their purchases, are at risk of turning to competitors. It's critical to quickly identify and engage these customers with targeted discounts and marketing efforts to retain their business



	summary	summary_polarity
21	the worst value ive ever seen	-1.000000
819	boring	-1.000000
1	another worthless dungeon masters screen from ...	-0.800000
623	disappointed	-0.750000
785	disappointed	-0.750000
1591	disappointed	-0.750000
361	promotes anger instead of teaching calming met...	-0.700000
875	too bad this is not what i was expecting	-0.700000
880	bad qualityall made of paper	-0.700000
100	small and boring	-0.625000
511	mad dragon	-0.625000
797	disappointing	-0.600000
1001	disappointing	-0.600000
1099	disappointing	-0.600000
1773	disappointing	-0.600000
991	then you will find this board game to be dumb ...	-0.591667

2.2 Identify Low Spenders that are Satisfied.

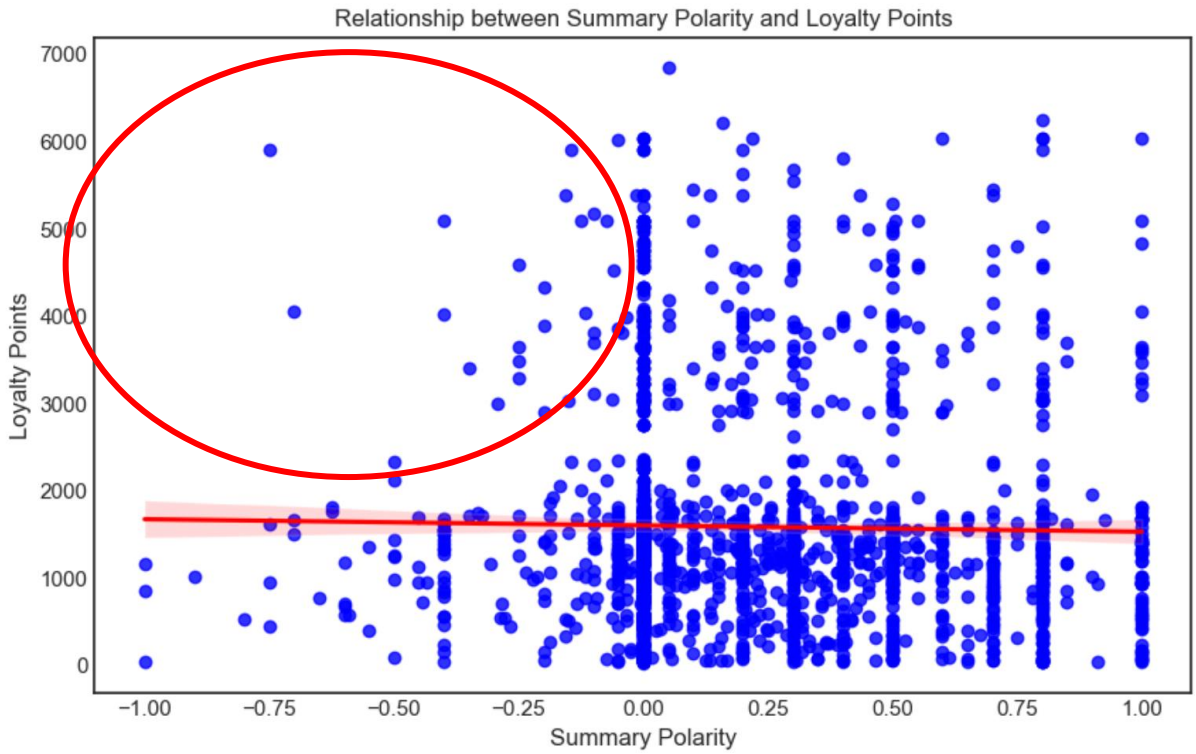
The following customers represent growth opportunities. There is potential to capitalise on the goodwill created by their recent purchases and increase their spending.



	summary	summary_polarity
6	best gm screen ever	1.000
1485	wonderful gift	1.000
1613	excellent teaching tool	1.000
968	the best among the dd boardgames	1.000
1611	perfect	1.000
...	...	...
726	really nice puzzle	0.600
1663	he liked it	0.600
1481	wow im impressed	0.550
163	not for beginners however great for the comic ...	0.525
1215	great first time set for creating a dungeon	0.525

3.0 Identify Loyal Customers giving Bad reviews

Turtle Games does not want to lose loyal and valued customers.



	summary	loyalty_points	summary_polarity
785	disappointed	5895	-0.750
1591	disappointed	1606	-0.750
361	promotes anger instead of teaching calming met...	4052	-0.700
875	too bad this is not what i was expecting	1501	-0.700
880	bad qualityall made of paper	1658	-0.700
100	small and boring	1809	-0.625
511	mad dragon	1752	-0.625
362	anger control game	1344	-0.550

4.0 Identify Products with the Best and Worst Reviews / Summaries.

- Grouped by product and aggregated by mean summary polarity.
- Sorted by best and worst scoring average summary polarity.

Worst 10 Products

	product	summary_polarity
0	11056	-0.030000
1	1501	-0.019375
2	466	0.006250
3	3525	0.008333
4	9507	0.015833
5	4047	0.018333
6	3967	0.029545
7	2285	0.031250
8	123	0.035000
9	7143	0.053333

Best 10 Products

	product	summary_polarity
0	10995	0.550000
1	3885	0.475833
2	7373	0.475000
3	4415	0.470000
4	1473	0.418000
5	9560	0.414000
6	263	0.411111
7	978	0.403704
8	3878	0.395625
9	979	0.392361

Look back and learn.

How can text data be used to inform marketing campaigns and make improvements to the business?

Optimizing Customer Segments for Increased Revenue

By pinpointing high spenders with dissatisfaction and low spenders with positive feedback, businesses can strategically address the concerns of key revenue contributors while leveraging insights from satisfied customers to elevate spending levels. Enhancing the experiences of dissatisfied segments helps mitigate churn risks and boosts customer lifetime value. For satisfied low spenders, implementing targeted promotions or loyalty incentives can encourage increased spending.

Revitalizing Customer Loyalty

Focusing on loyal customers who express dissatisfaction is essential for uncovering deep-rooted issues and maintaining a robust brand reputation. Personalized engagement and resolving specific grievances can rejuvenate their trust and loyalty, ensuring long-term retention.

Product Strategy Through Customer Feedback

Analyzing product feedback through summary polarities offers clear insights into customer preferences, guiding inventory, development, and marketing decisions. Improving or innovating based on this feedback can elevate customer satisfaction and drive sales, while adjusted marketing strategies can spotlight top-performing products.

Driving Sales through Strategic Initiatives

- **Product Development**: Utilize feedback for product enhancement and innovation, directly influencing customer satisfaction and sales.
- **Targeted Marketing**: Deploy segmented marketing strategies to resonate with distinct customer groups, increasing engagement and expenditure.
- **Customer Service Excellence**: Addressing loyal customers' concerns promptly and effectively can convert them into brand advocates, impacting loyalty and spending.
- **Pricing Strategies**: Use customer feedback to inform pricing approaches, optimizing revenue through premium pricing for well-received products or discounts on less popular items.

In essence, leveraging customer feedback and spending behaviors enables informed strategic decisions that not only refine product offerings and customer satisfaction but also foster sales growth. Catering to the nuances of high-value and potential growth customer segments can solidify a loyal base and attract new patrons, securing long-term business prosperity.

4. Sales

Identify the Problem and Opportunities

TG does not fully understand trends and patterns in the sales data, and the relationship between regional sales and global sales.

Opportunities:

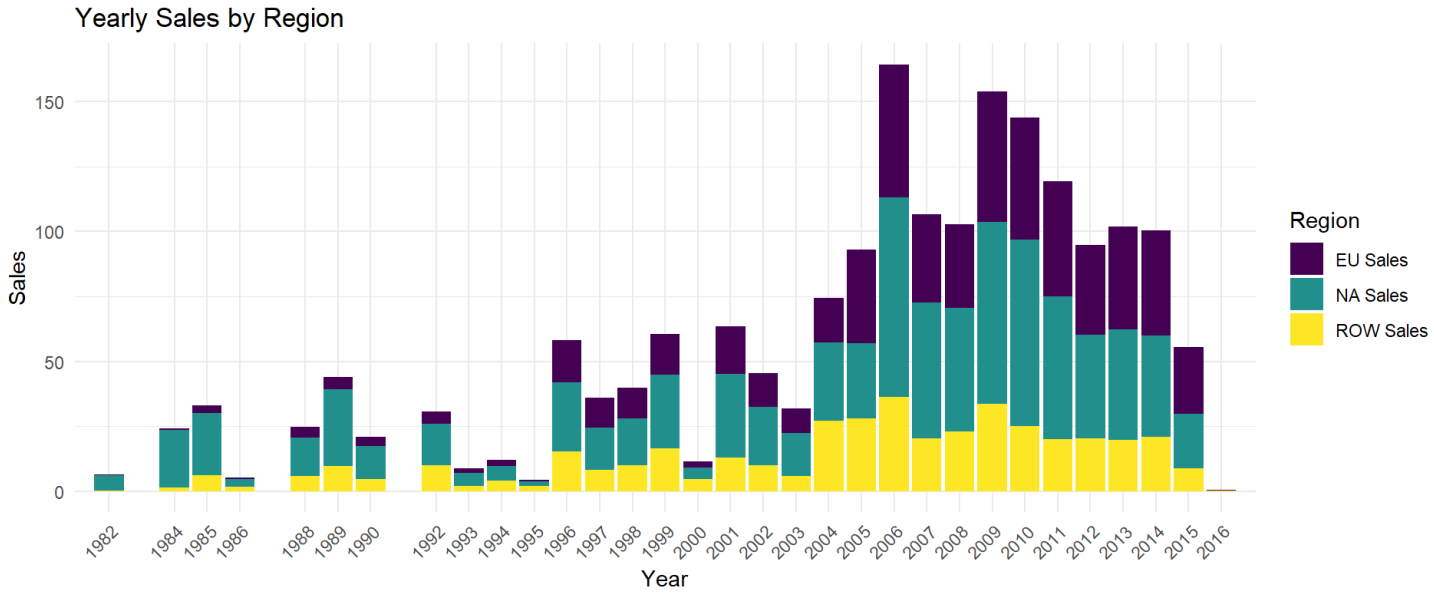
- Gain an overall picture of sales trends.
- Assess the data's suitability for modelling purposes.
- Analyse how regional and global sales are interrelated.

Define Goals

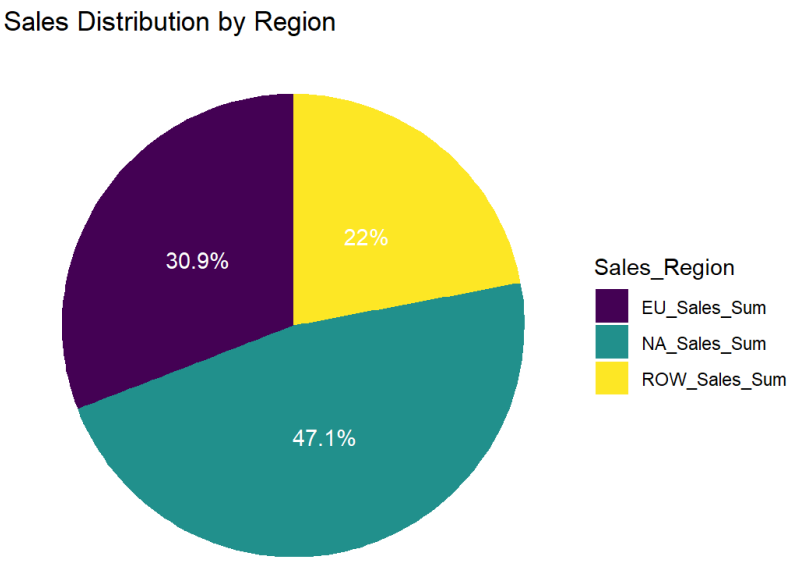
- To leverage R's plotting capabilities to create an overview of sales trends.
- To assess the data's suitability for modelling by comparing the data's characteristics to the prerequisites necessary for regression analysis.
- To apply linear regression analysis to regional and global sales data and ascertain their relationship and the effectiveness of the models created.

Explore

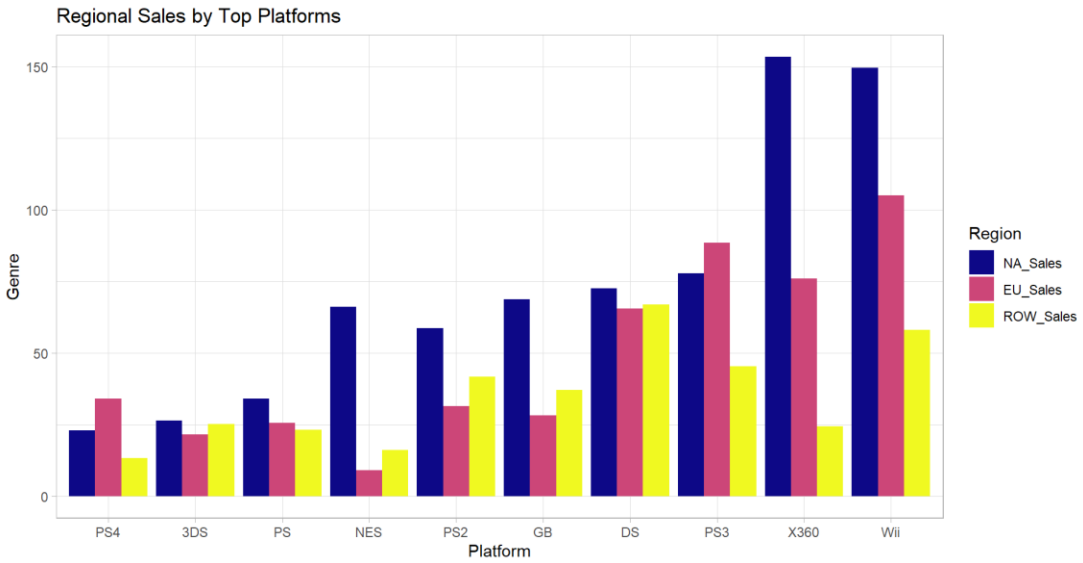
4.1 What trends and insights can be identified from the sales data?



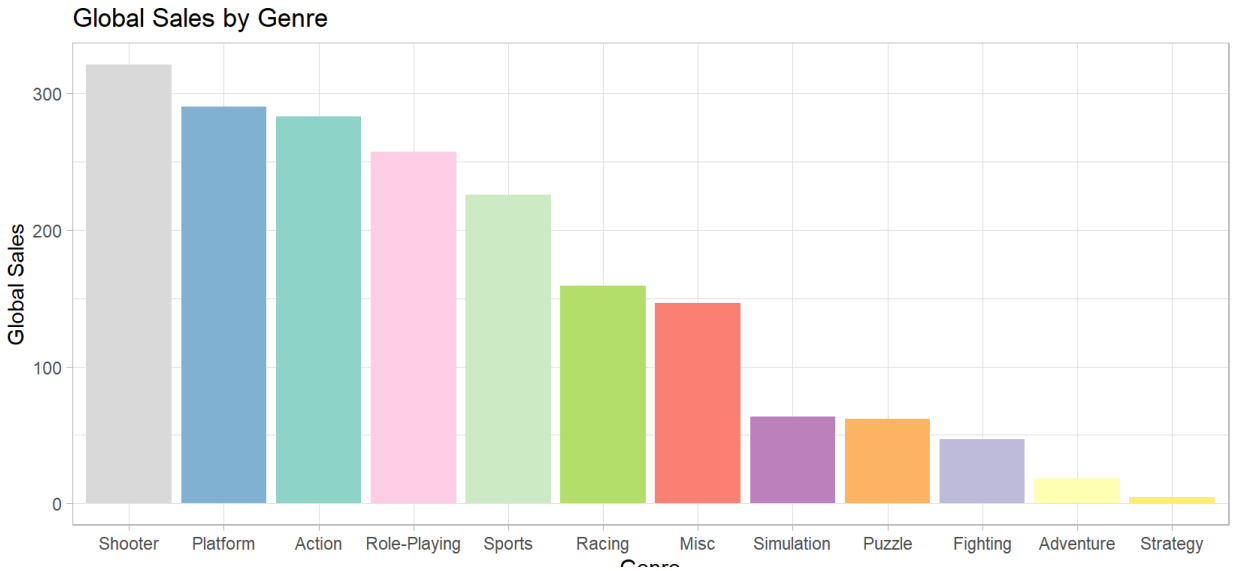
Sales peaked in 2006 and have dramatically fallen since 2009.



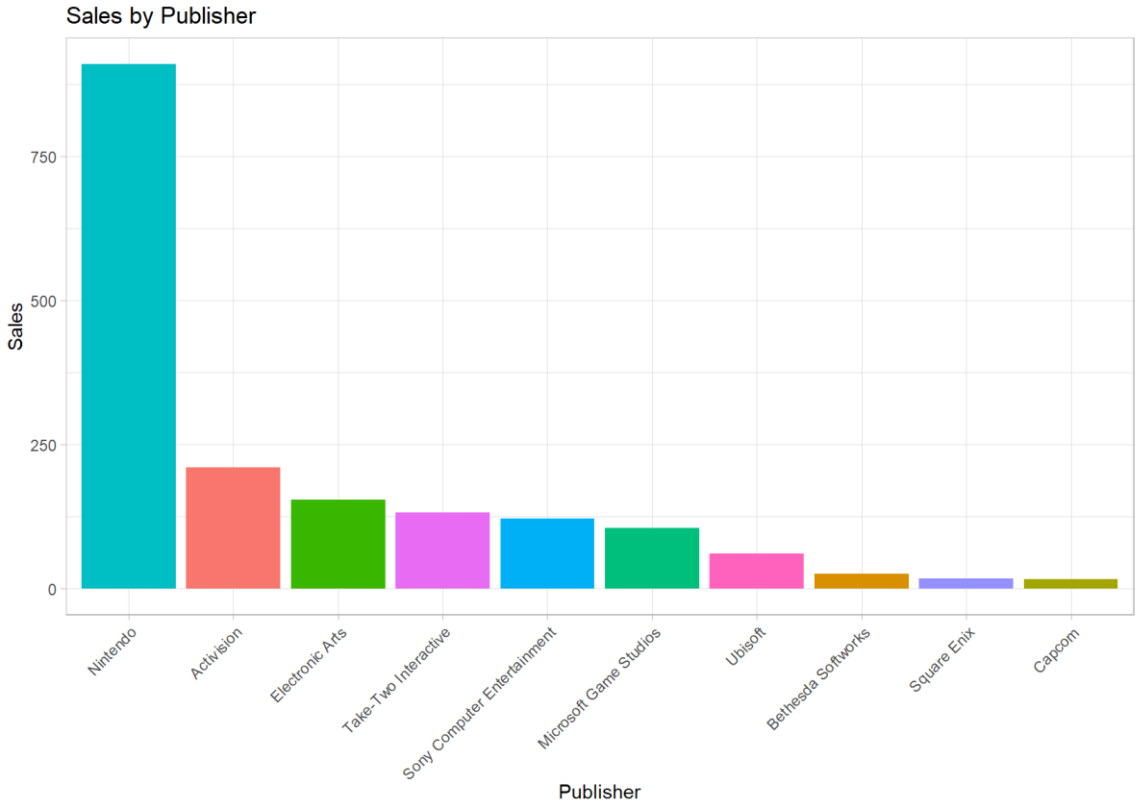
- North America accrues the most sales



- The highest grossing platforms: X Box 360 in North America, the Wii in Europe and the DS in the rest of the world.

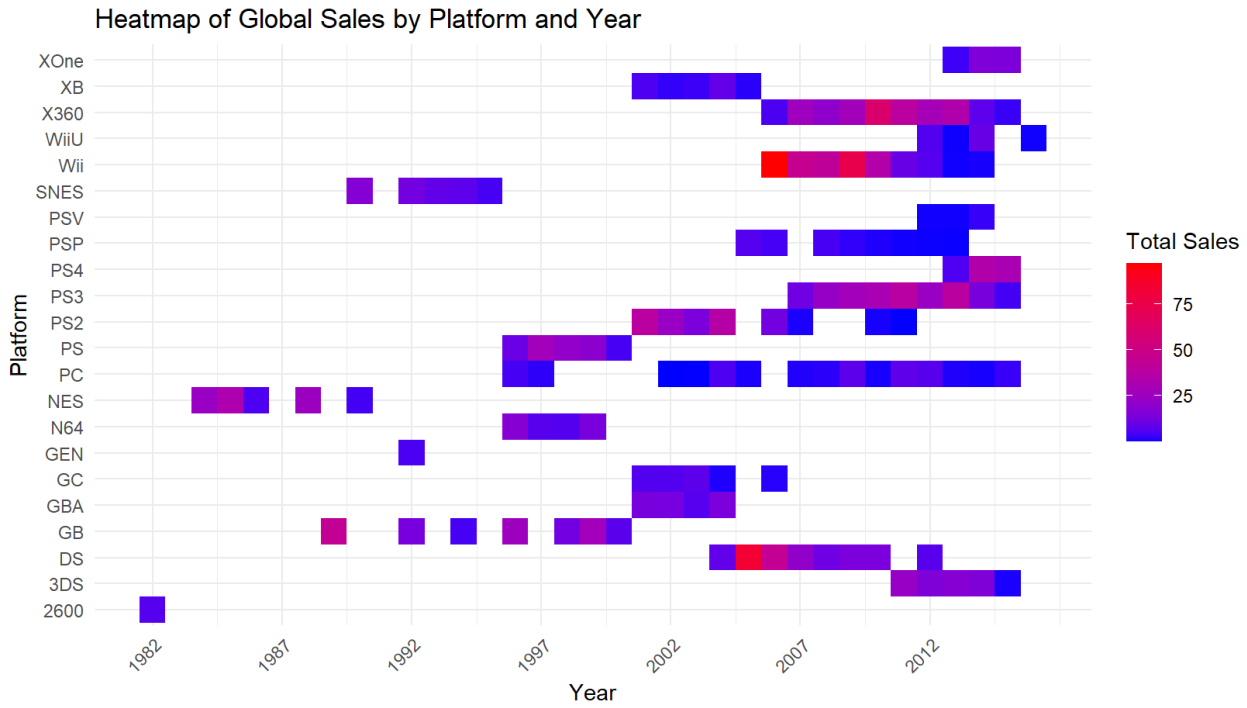


- Shooter games are the most popular globally.

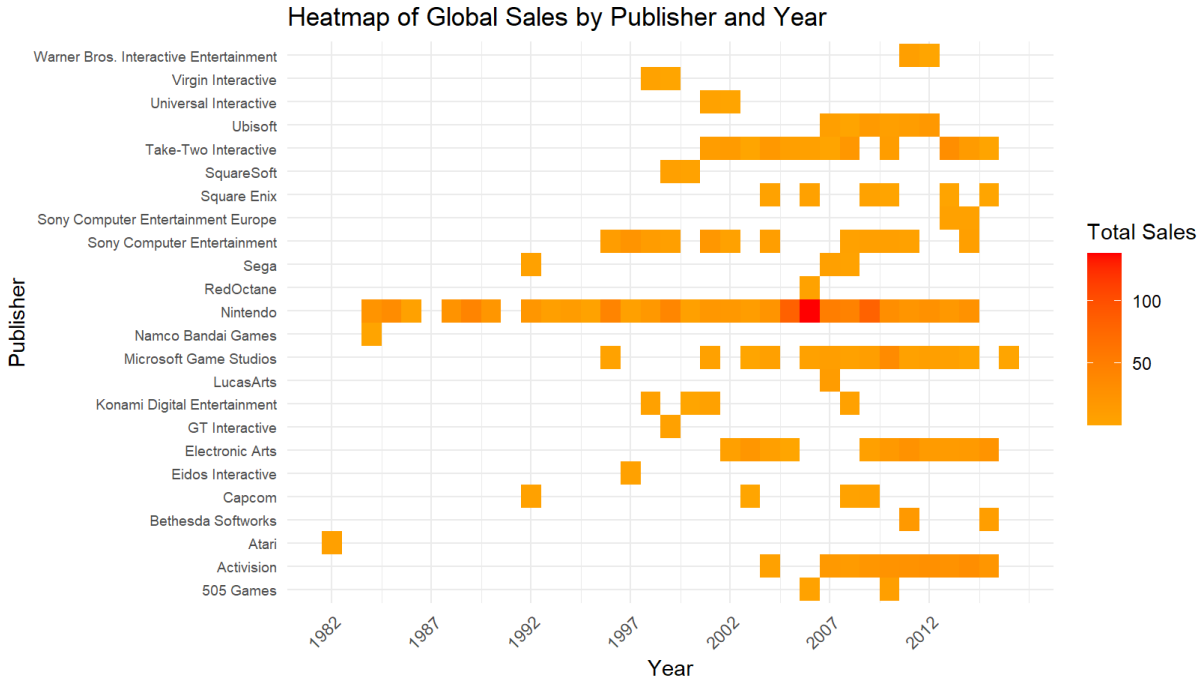


- Nintendo is the most popular publisher globally.

Heatmaps



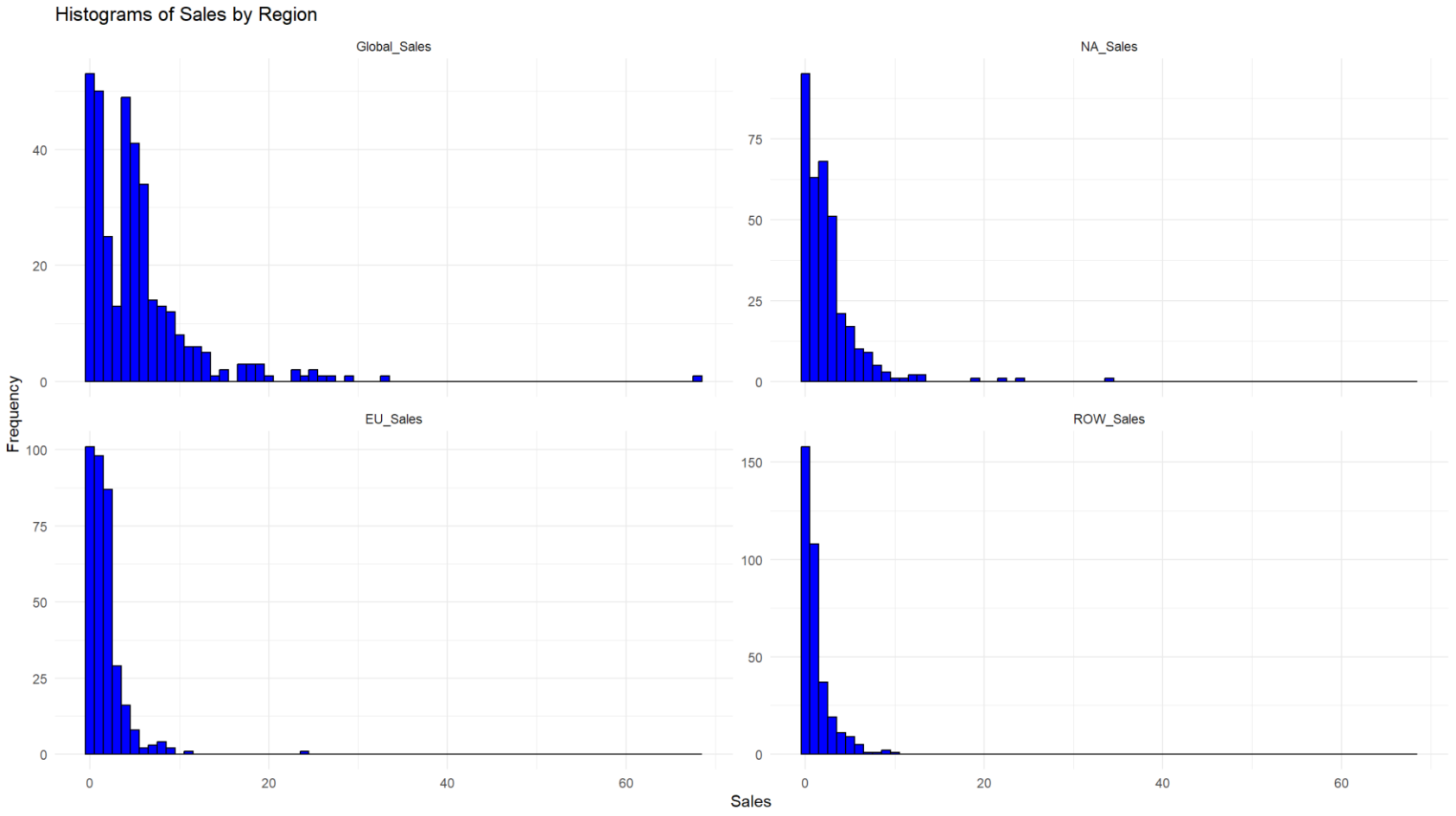
- The most consistent platforms in terms of global sales are X Box 360, Wii, PC and DS. It is evident that Platforms are introduced, perform well and then are superseded by new platforms.



- Nintendo is clearly the most consistently successful publisher over time.

Distributions

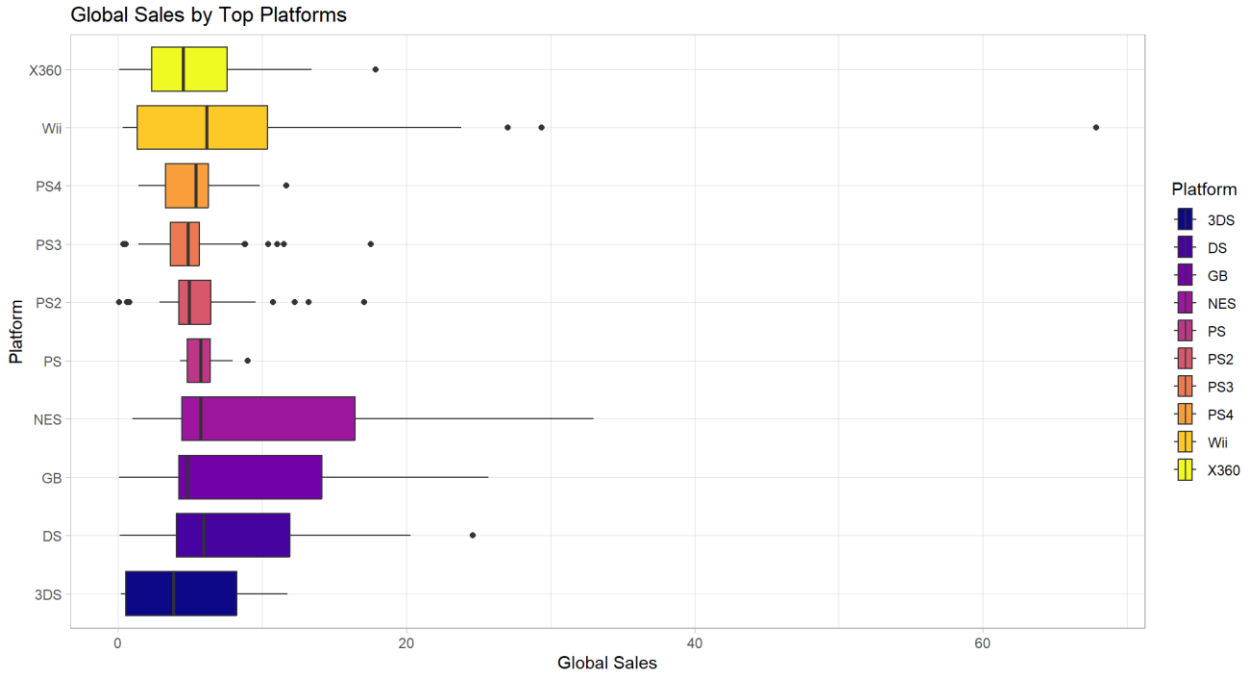
The data was pivoted, and a faceted histogram created.



All sales data exhibit strong right skews with a notable outlier in Global, NA and EU Sales.

Top 10 platforms

The data was grouped by platform and filtered for the top 10 by global sales:



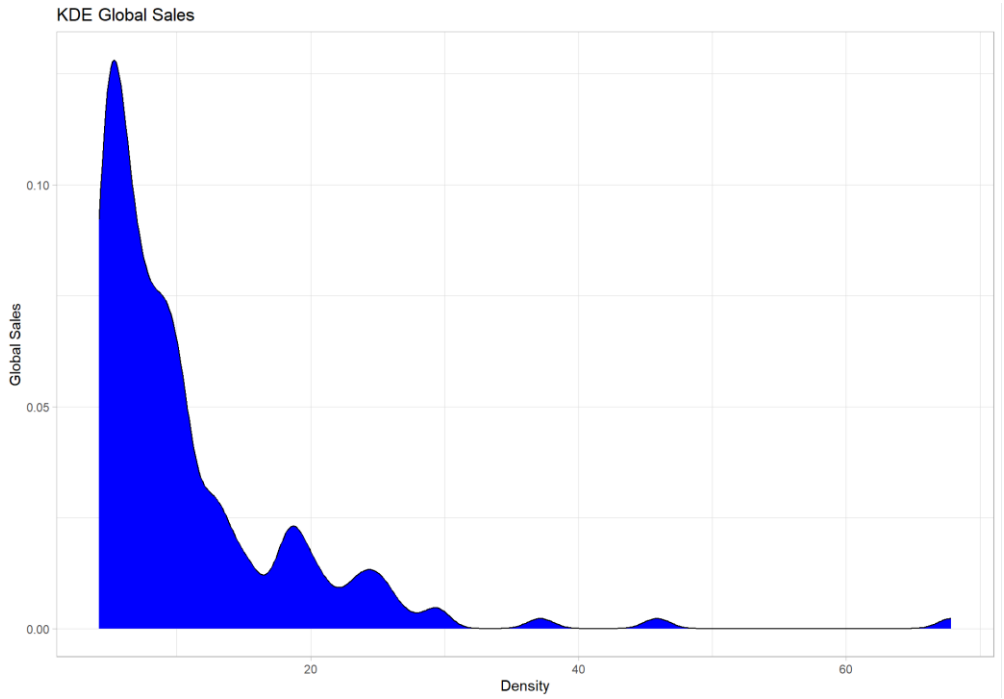
The notable outlier is in the Wii platform. The highest median value is within the DS platform category and greatest spread in terms of global sales is the NES.

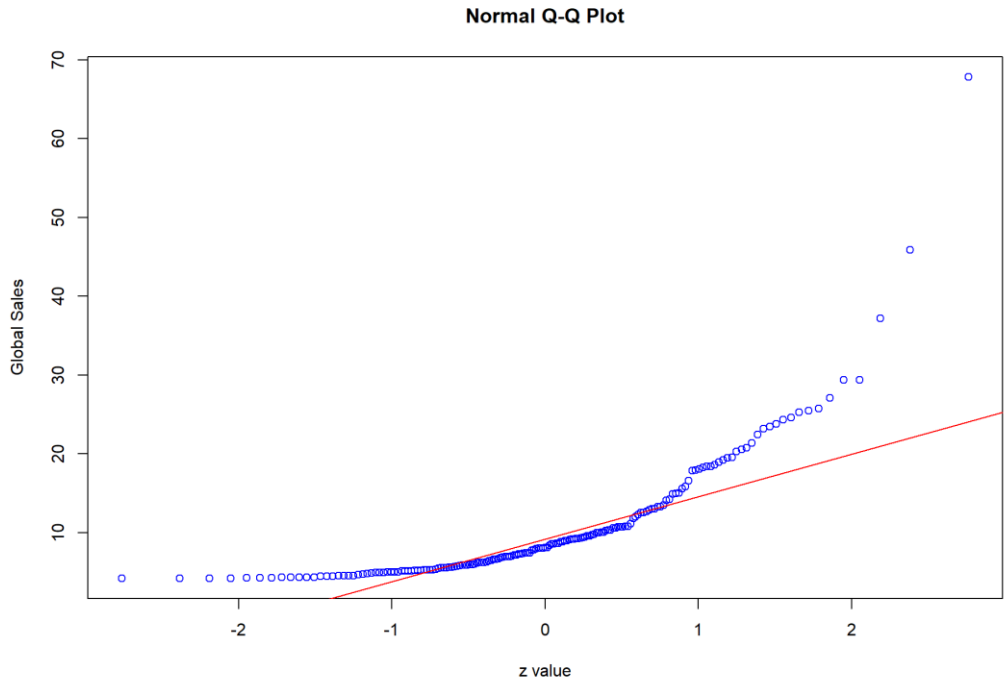
Correlations in Regional Sales Data – see section 4.2

4.2 Data Reliability and Suitability for Regression Modelling

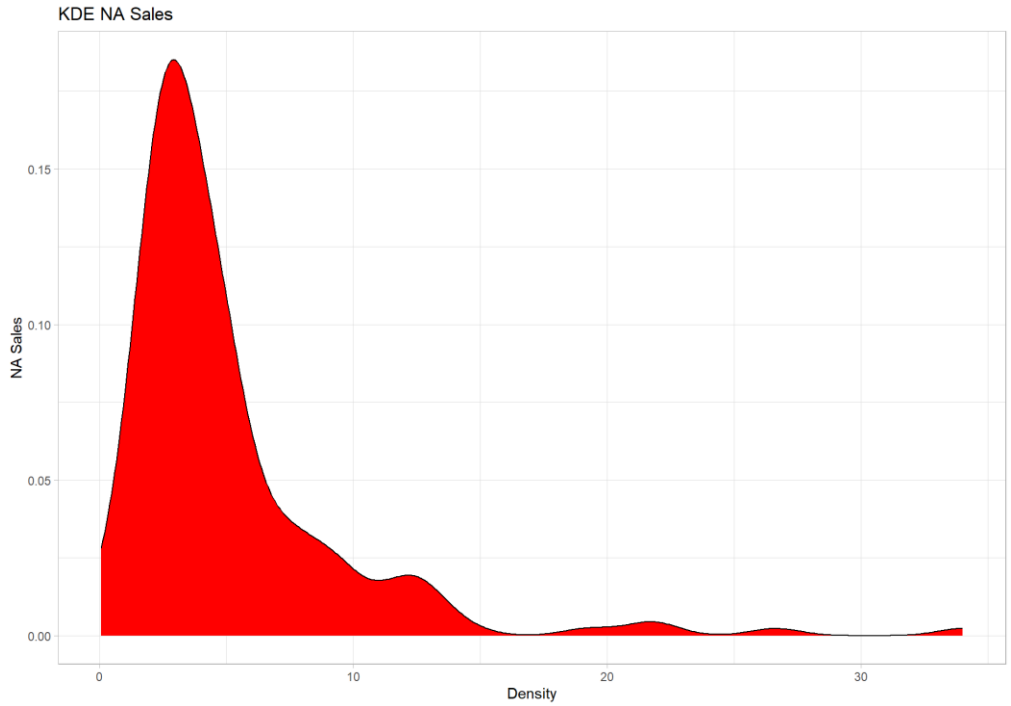
Determine the Normality of the Sales Data

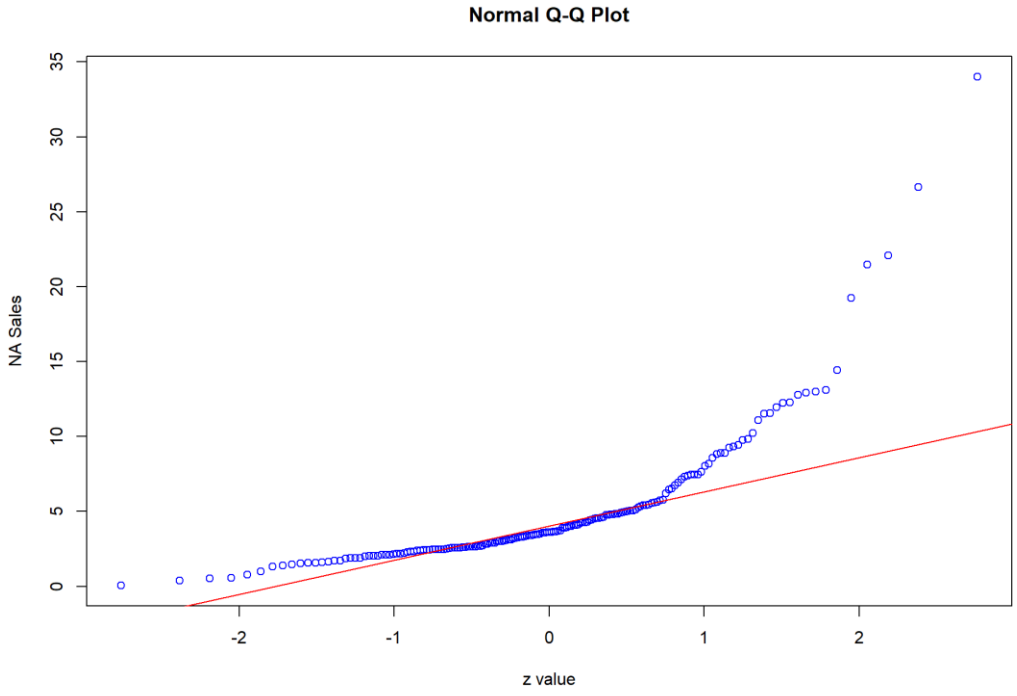
A. Global Sales



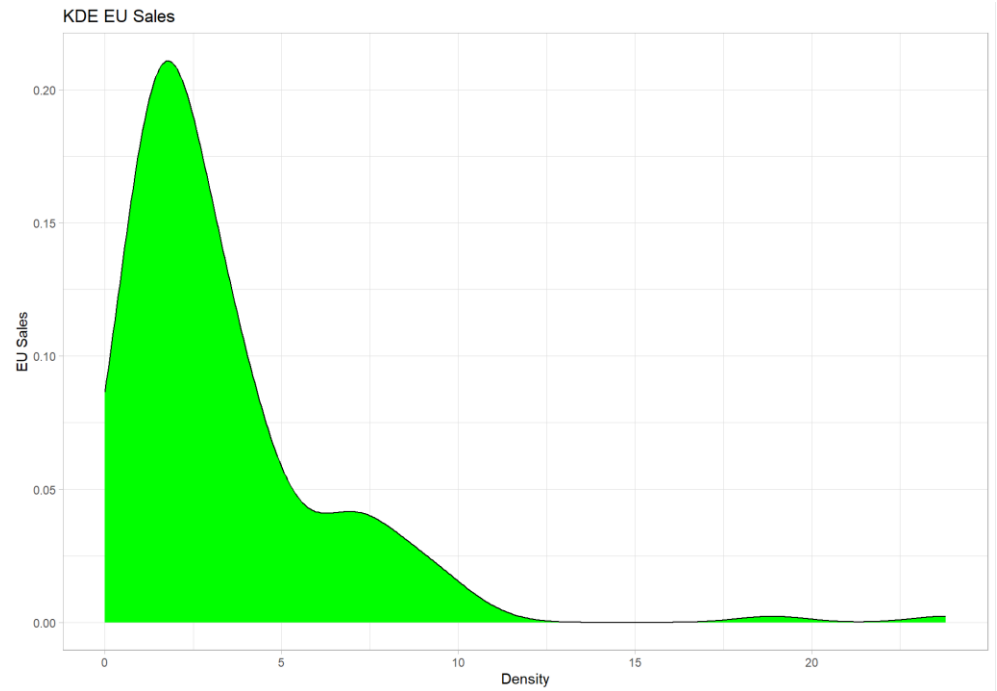


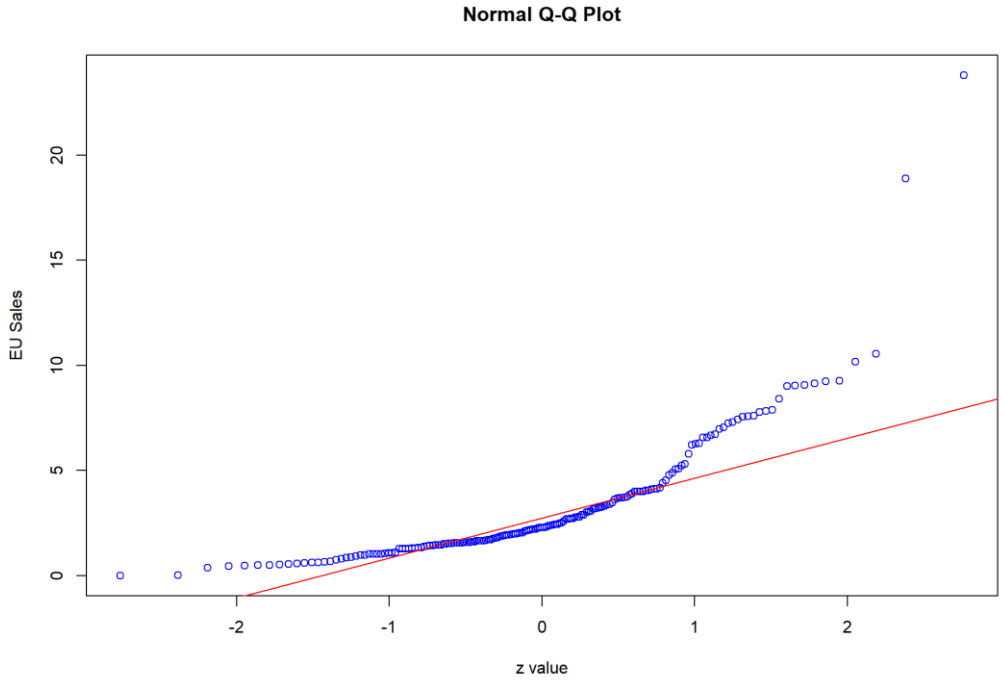
B. North American Sales



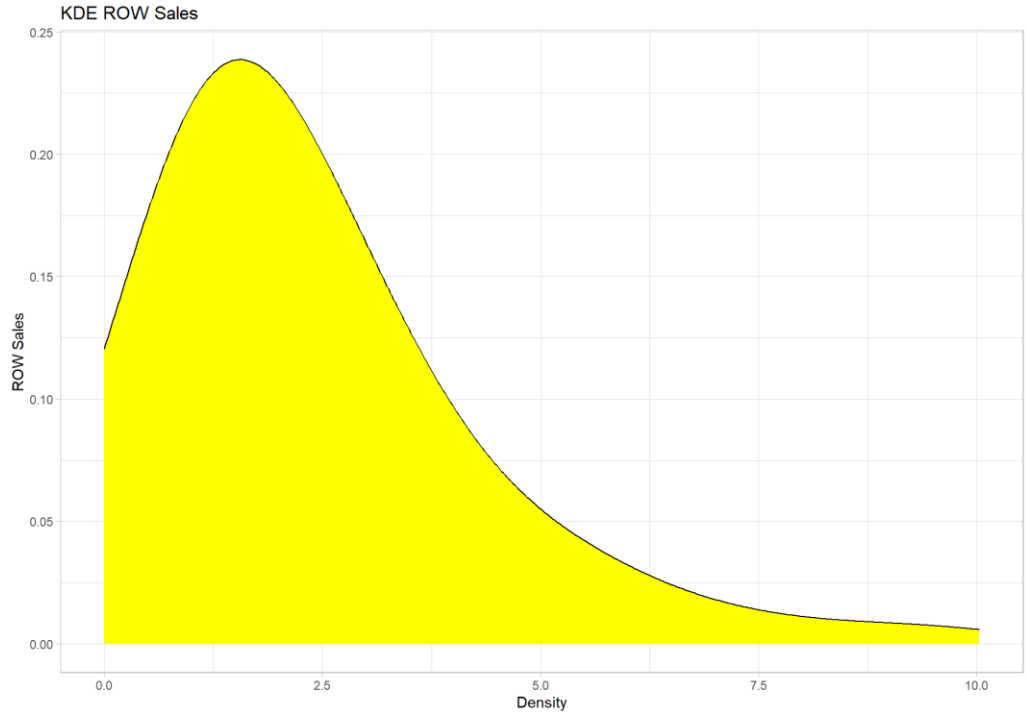


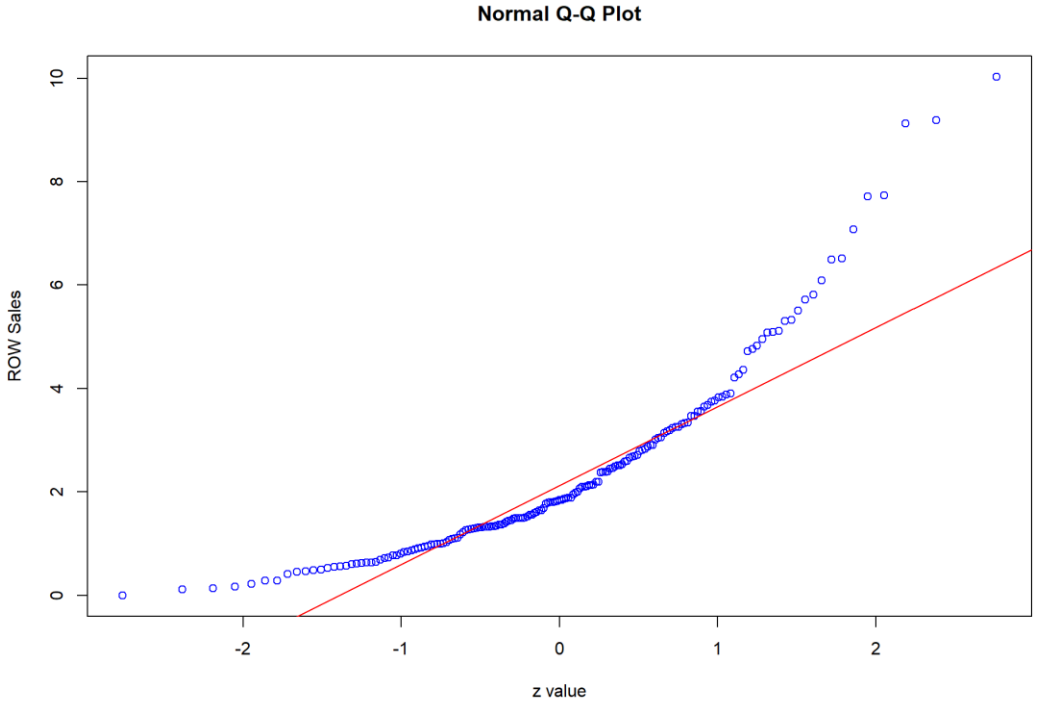
C. European Sales





D. Rest of the World Sales





Metric	Global Sales	NA Sales	EU Sales	ROW Sales
Shapiro-Wilk				
Statistic	0.71	0.698	0.741	0.858
P-value	3.47E-17	1.64E-17	2.99E-16	9.64E-12
Skewness	3.07	3.05	2.89	1.63
Kurtosis	17.79	15.60	16.23	6.08

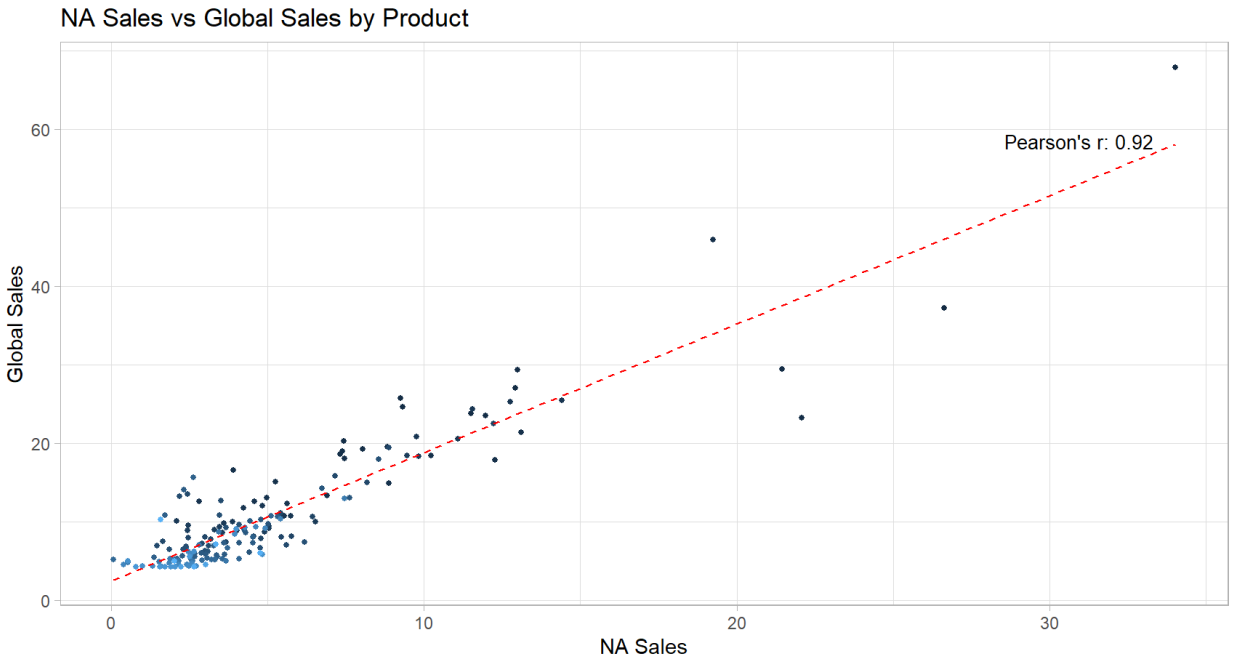
Correlation and Linearity

A correlation matrix was produced:

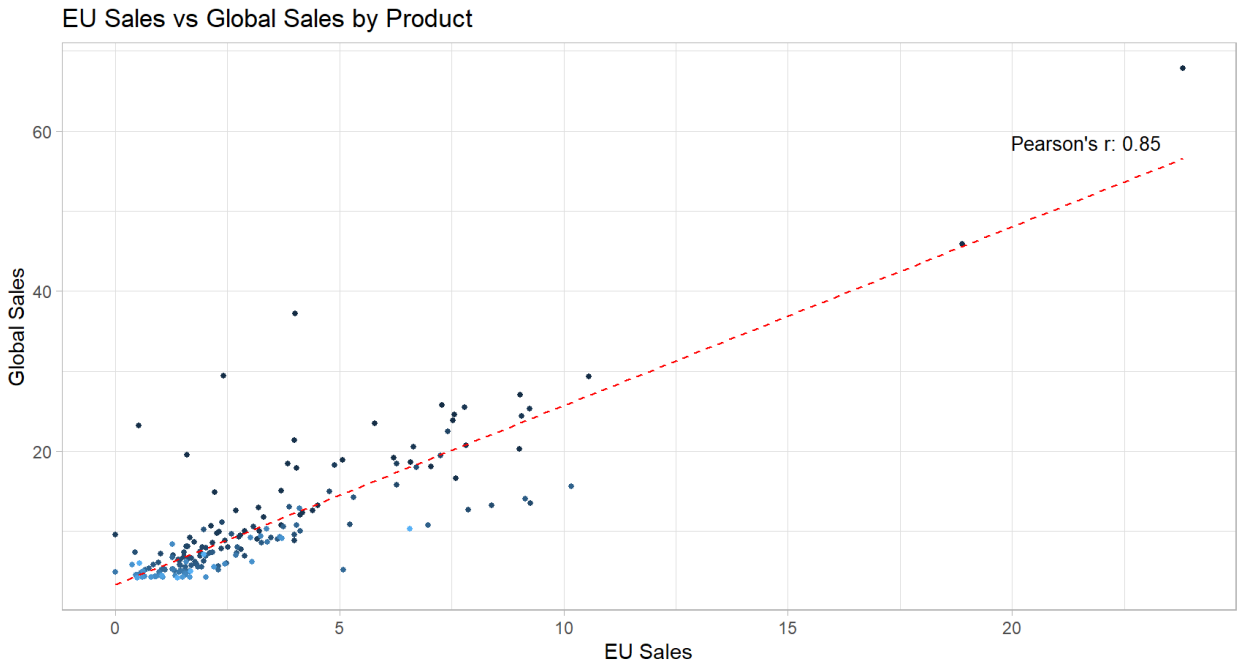
	NA_Sales_Sum	EU_Sales_Sum	ROW_Sales_Sum	Global_Sales_Sum
NA_Sales_Sum	1.00	0.62	0.53	0.92
EU_Sales_Sum	0.62	1.00	0.54	0.85
ROW_Sales_Sum	0.53	0.54	1.00	0.73
Global_Sales_Sum	0.92	0.85	0.73	1.00

Scatterplots

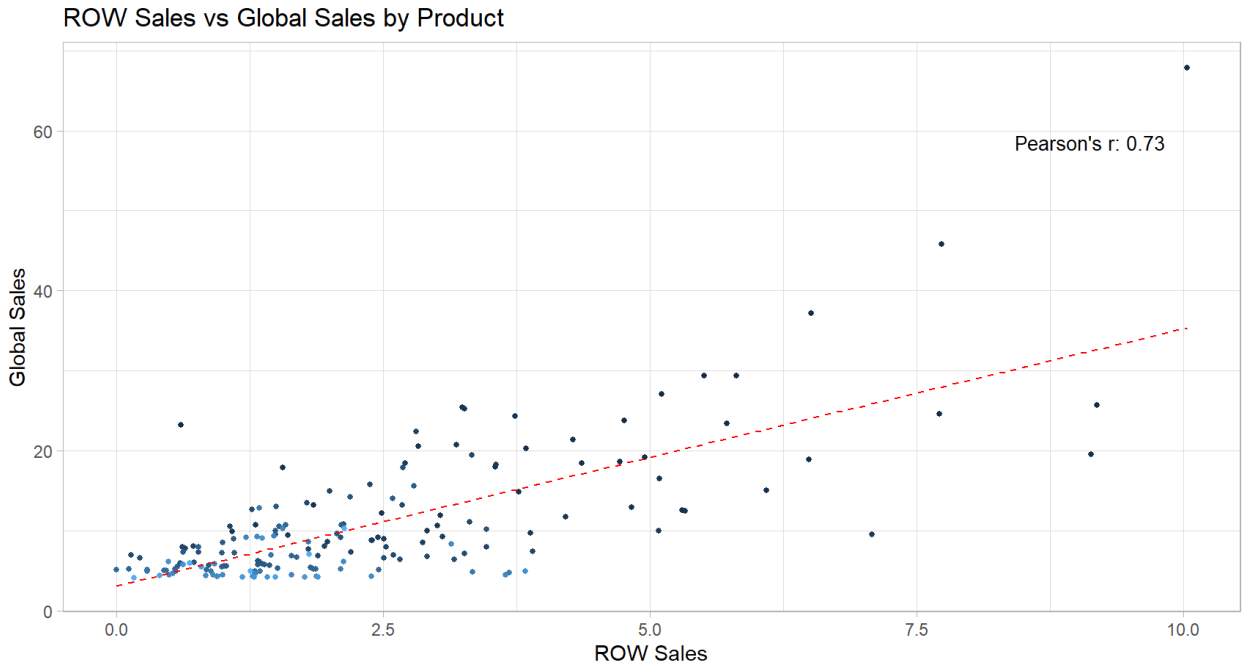
North American and Global sales are positively and strongly correlated.



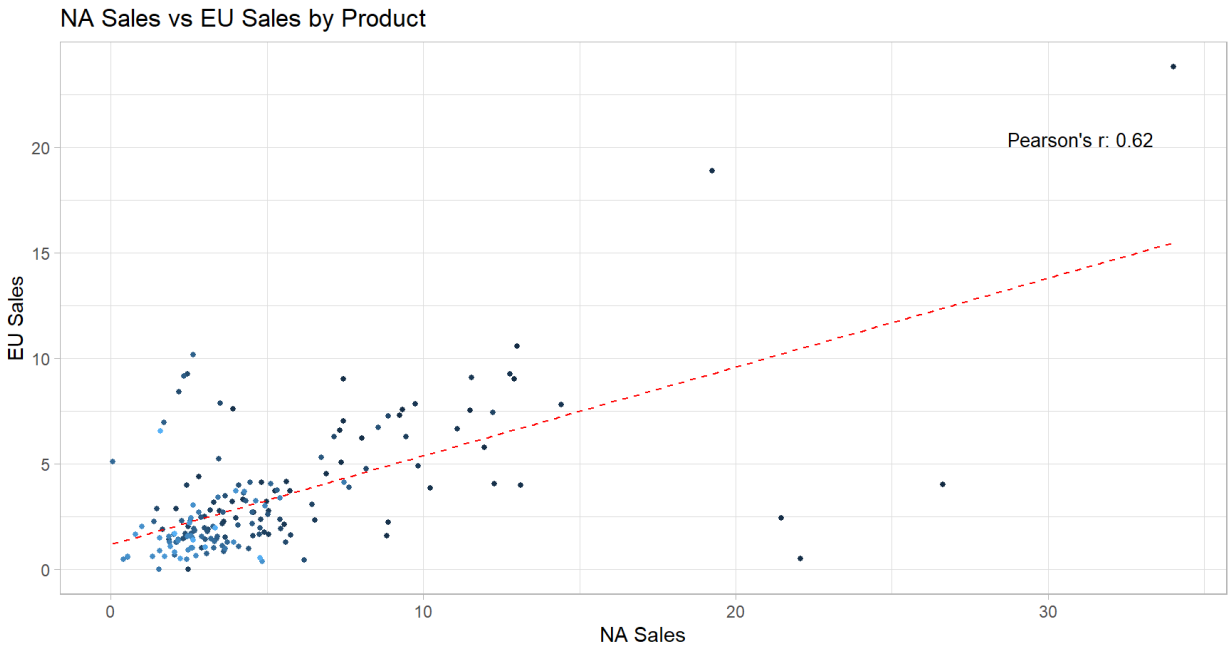
European and Global sales are positively and strongly correlated.



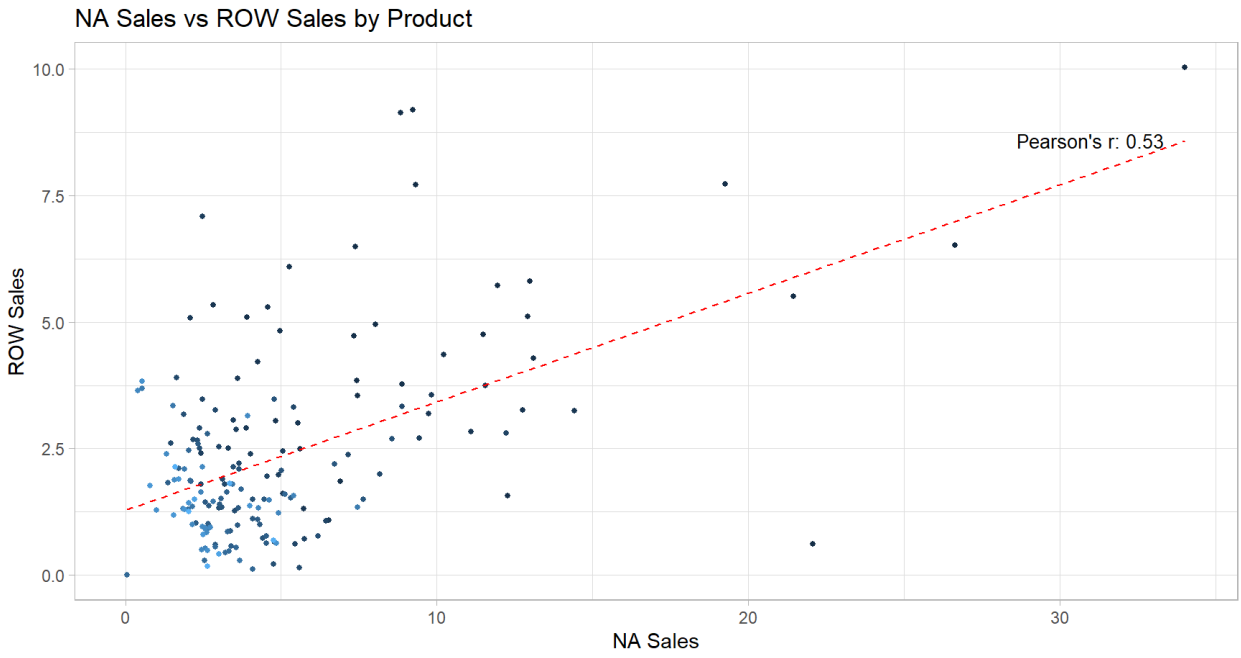
Sales outside Europe and American are also positively and strongly correlated although to a lesser extent.



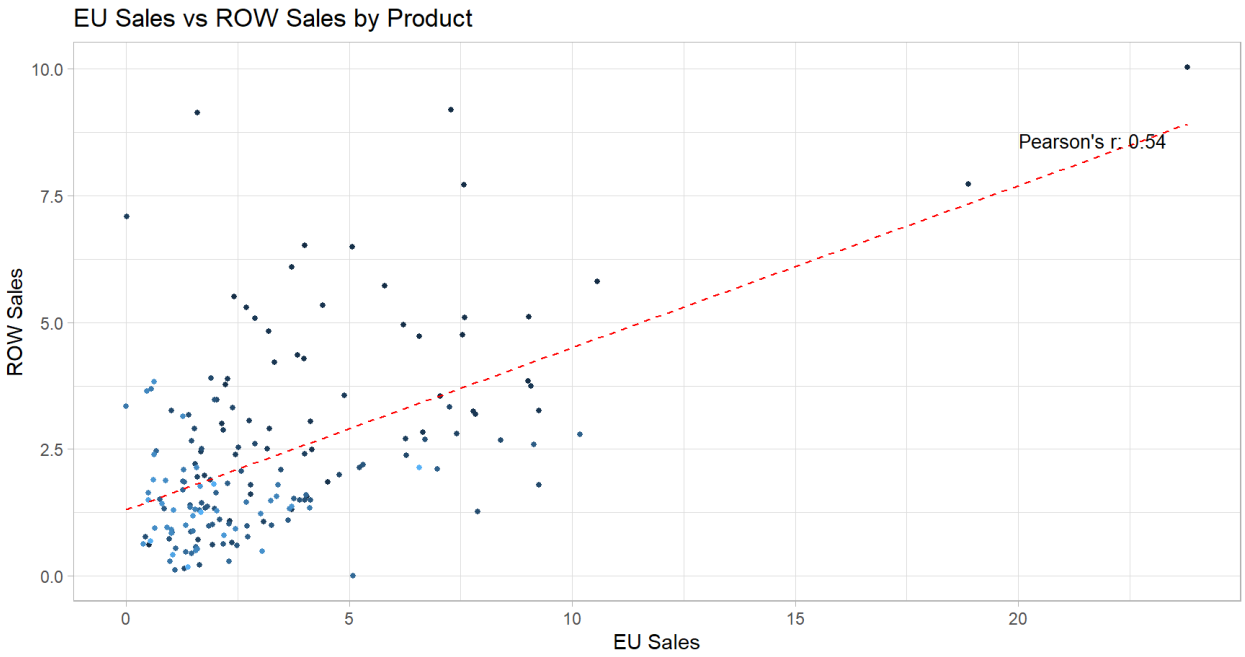
North American and European sales are positively and moderately correlated.



North American and rest of the world sales are positively and moderately correlated.



European and the rest of the world sales are positively and moderately correlated.



Is the sales data suitable for modelling purposes?

(Note: Several of the assumptions were tested after modelling.)

Observations

Shapiro-Wilk

For all the sales data (NA, EU, ROW, and Global Sales), the test statistics are significantly less than 1, and the p-values are extremely small (far below the typical alpha level of 0.05). This strongly suggests that the sales data for each region do not follow a normal distribution.

Skewness

For all variables, the skewness is greater than 0, indicating that all distributions are right-skewed. In practical terms, this suggests that a small number of products have very high sales compared to the majority.

Kurtosis

The kurtosis for all variables is significantly greater than 3, indicating that all distributions are leptokurtic and far from normally distributed. This suggests a higher likelihood of outliers in the sales data, which can impact statistical analyses and model performance.

Q-Q Plots

All Q-Q plots display deviations and therefore point towards non-normality of the sales data.

However, non-normality of the sales data does not invalidate linear regression analysis. Normality of residuals is a fundamental assumption of regression analysis, which will be tested in due course.

Normality of Residuals

See section 4.3.

Correlations

The correlations indicate that sales in NA, EU, and ROW are all positively associated with each other and have a strong positive influence on Global Sales. The strongest correlations are between regional sales and Global Sales, particularly for NA and EU, highlighting their significant contributions to Global Sales. The moderate correlations between NA, EU, and ROW Sales suggest interdependence between these markets, albeit to a lesser degree than their individual contributions to Global Sales.

Therefore, the sales data fulfils the linear regression requirement for linearity.

Autocorrelation

The Durbin Watson is less than 2, which suggests the presence of positive autocorrelation in the residuals. The p-value is less than 0.05, therefore the null hypothesis - that there is no autocorrelation, is rejected.

This is an indication that the residuals are not independent.

Test for Independence – VIF for Multi-collinearity

See section 4.3.

Homoscedasticity

See section 4.3.

Suitability for modelling requires that the following prerequisites are fulfilled.

Assumption	Explanation	Result
Linear Model	This assumption posits a linear relationship between independent and dependent variables, as non-linear relationships in real-world data would lead to inaccurate predictions with significant variance from actual observations in a linear regression model.	✓
No multicollinearity	Multicollinearity occurs when predictor variables are correlated, leading to redundancy in the dataset as these variables contain similar information. This redundancy increases model complexity without contributing new information or patterns, making it advisable to avoid highly correlated features even in complex models.	✓
Homoscedasticity of Residuals	Homoscedasticity refers to the condition where residuals from a linear regression model are uniformly spread, indicating a satisfactory model.	✓
No Autocorrelation in residuals	Absence of autocorrelation, where residuals are independent of each other.	✗
Number of observations Greater than the number of predictors	To enhance model performance, the quantity of training data should exceed that of test data. Specifically, in linear regression, the number of observations must be greater than the number of independent variables.	✓
Unique Observations	Each observation in the dataset should be independent, implying that it is uniquely measured for each occurrence of the event causing the observation.	✓
Normal distribution of Residuals	This assumption states that the residuals (errors) from the regression model should be normally distributed. This is particualrly important if the data is to be used in hypothesis testing.	✗

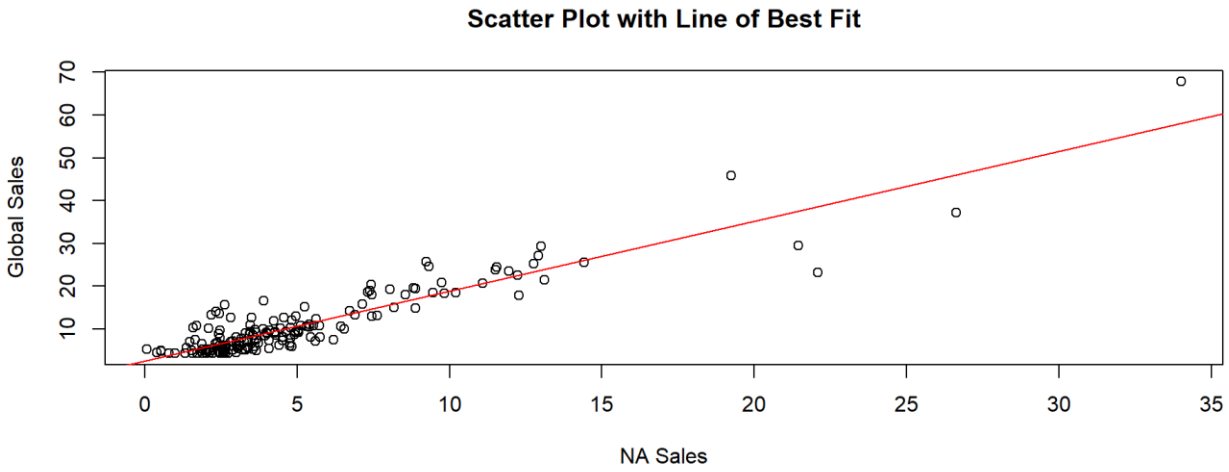
4.3 Relationships between North American, European and Global Sales

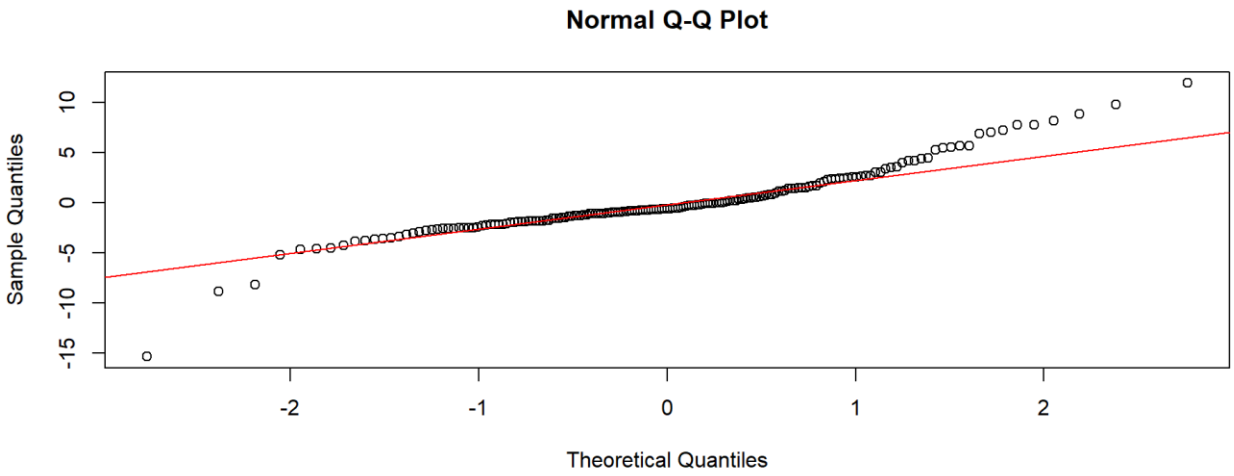
Simple Linear Regression

NA Sales vs Global Sales

Summary

```
Call:
lm(formula = sales_product$Global_Sales_Sum ~ sales_product$NA_Sales_Sum)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3417  -1.8198  -0.5933   1.4322  11.9345

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.45768    0.36961   6.649 3.71e-10 ***
sales_product$NA_Sales_Sum  1.63469    0.05435  30.079  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.266 on 173 degrees of freedom
Multiple R-squared:  0.8395,    Adjusted R-squared:  0.8385
F-statistic: 904.7 on 1 and 173 DF,  p-value: < 2.2e-16
```

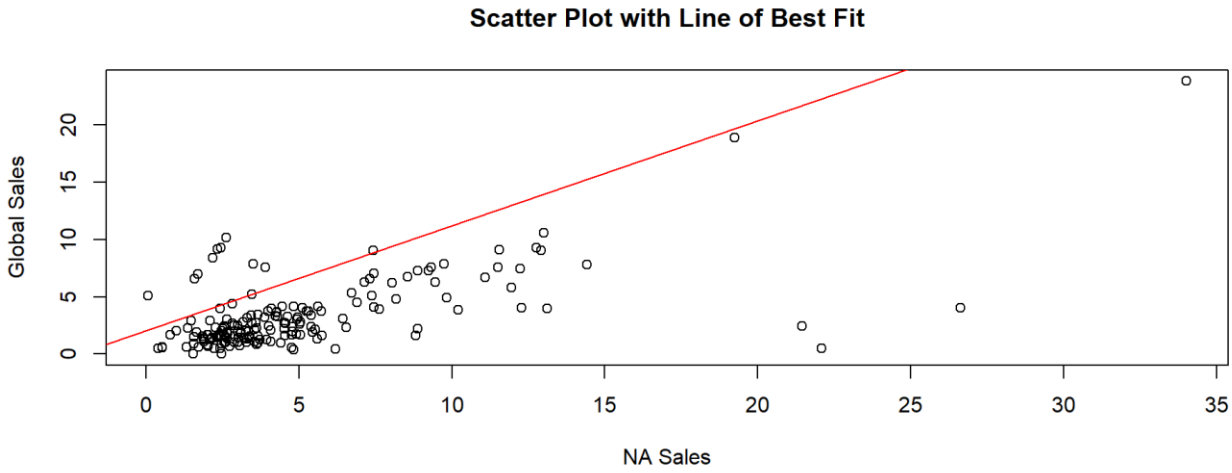


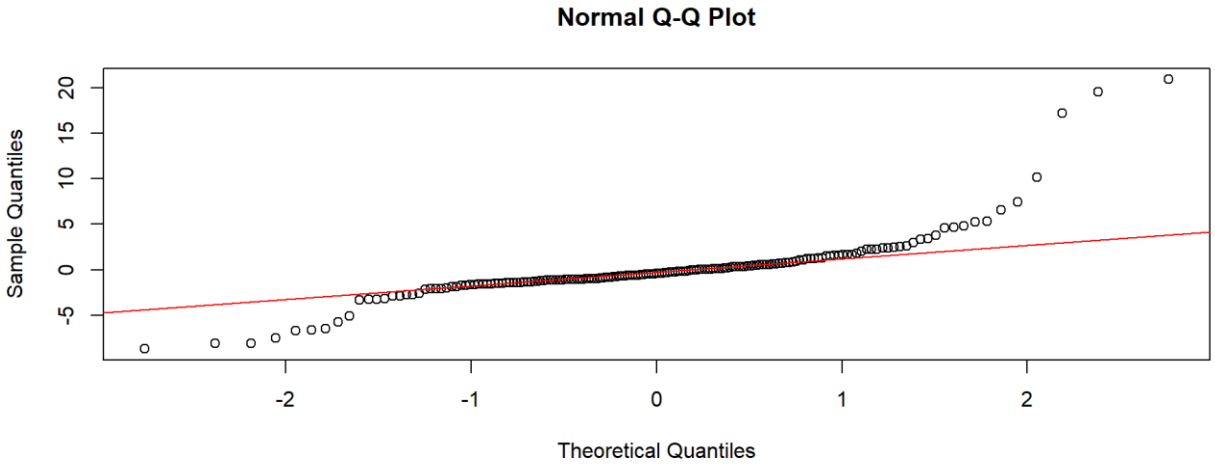


NA Sales vs EU Sales

Summary

```
Call:
lm(formula = sales_product$NA_Sales_Sum ~ sales_product$EU_Sales_Sum)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7273 -1.2982 -0.3932  0.7136 20.9338

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.02748    0.39757    5.10 8.87e-07 ***
sales_product$EU_Sales_Sum  0.91739    0.08805   10.42  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.582 on 173 degrees of freedom
Multiple R-squared:  0.3856,    Adjusted R-squared:  0.382
F-statistic: 108.6 on 1 and 173 DF,  p-value: < 2.2e-16
```


Scatter Plot with Line of Best Fit
Global Sales
NA Sales
0
5
10
15
20
25
30
35


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-5
0
5
10
15
20
-2
-1
0
1
2

Multiple Linear Regression Analysis

Dependent Variable – Global Sales

Independent Variables – NA Sales and EU Sales

(Note: ROW Sales cannot be used alongside NA Sales and Eu Sales as it would result in a perfect fit as combined, they are equivalent to Global Sales.)

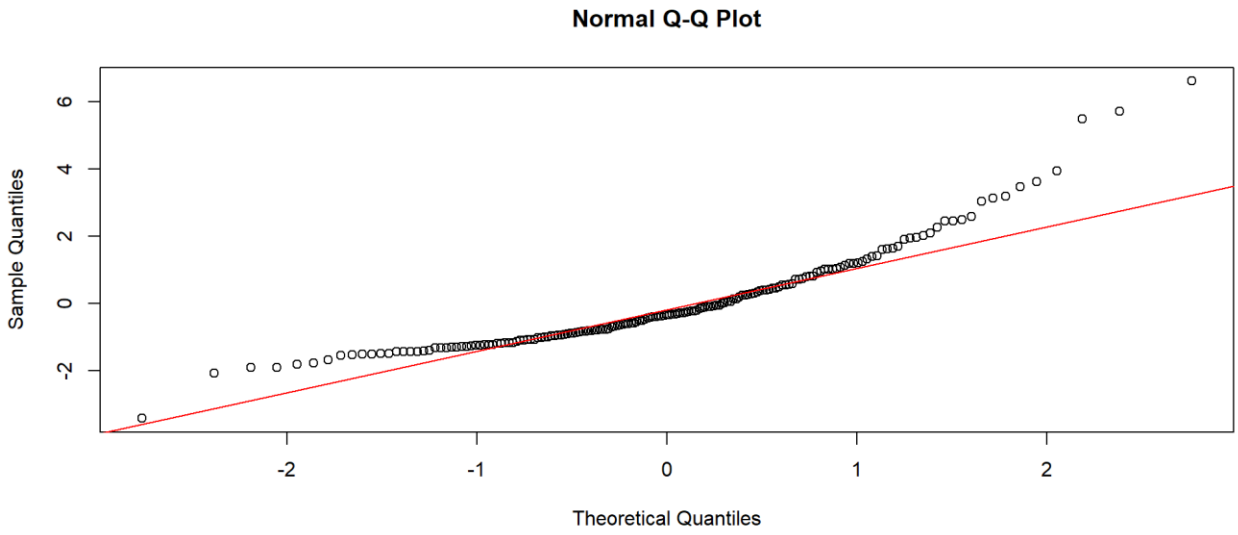
Summary

```
lm(formula = Global_Sales_Sum ~ NA_Sales_Sum + EU_Sales_Sum,
    data = sales_cols2)

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 ***
NA_Sales_Sum  1.13040    0.03162  35.745  < 2e-16 ***
EU_Sales_Sum  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
```

Model Evaluation

Independent Variables	R-squared	F-stat and Prob (F-statistic)	Coefficients	Residual Standard Error	Summary
NA Sales and EU Sales	Approximately 96.68% of the variability in Global_Sales_Sum can be explained by the linear relationship with NA Sales and EU Sales. This is a very high R-squared value, indicating a strong model fit.	The F-test is highly significant (p-value < 2.2e-16), indicating that the model is statistically significant, and there is a linear relationship between the response and the predictors.	Both predictors are statistically significant (p-values < 2e-16), indicating a strong linear relationship between each predictor and the response variable	1.49 on 172 degrees of freedom, indicating a good fit.	Significant and positive relationship between NA Sales + Eu Sales and Global Sales.

Test for Normal Distribution of Residuals

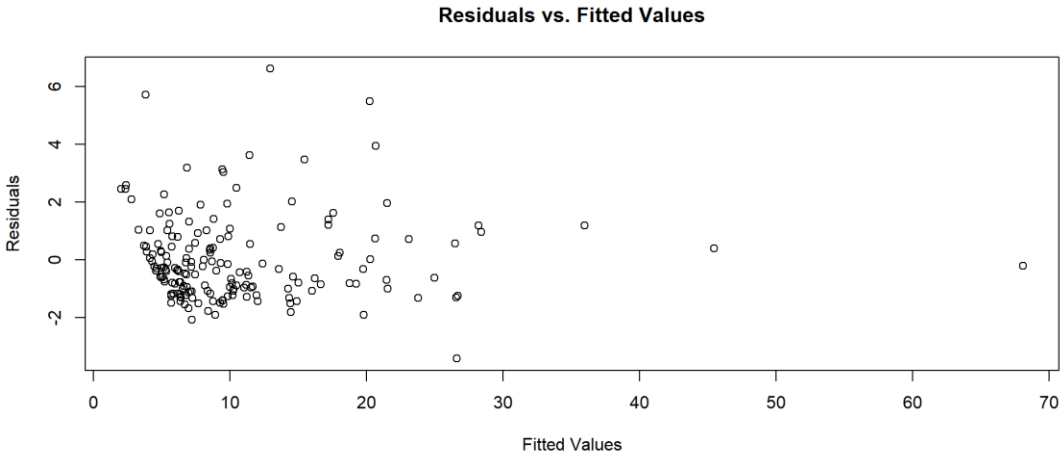


There is some deviation of the residuals from the line of best fit indicating that the residuals are not normally distributed.

```
        Shapiro-Wilk normality test

data:  residuals(MLR1)
W = 0.88431, p-value = 2.125e-10
```

The Shapiro Wilk normality test illustrates that the residuals are not normally distributed. The p value is excessively small indicating a solid rejection of the null hypothesis that the residuals are normally distributed.

Test for Heteroscedasticity - Residual Plot



The residual plots show a degree of randomness and no clear pattern, indicating homoscedasticity of residuals.

Test for Independence – VIF for Multi-collinearity.

```
NA_Sales_Sum EU_Sales_Sum
    1.627488     1.627488
```

Both independent variables show a low to moderate but acceptable level of Multi-collinearity, well below the threshold of 5.

Test for Autocorrelation – Durbin Watson

```
        Durbin-Watson test

data:  MLR1
DW = 1.6605, p-value = 0.01067
alternative hypothesis: true autocorrelation is greater than 0
```

The DW is less than 2, which suggests the presence of positive autocorrelation in the residuals. The p-value is less than 0.05, therefore the null hypothesis - that there is no autocorrelation, is rejected.

This is an indication that the residuals are not independent.

Steps to improve the model - Log of Global Sales

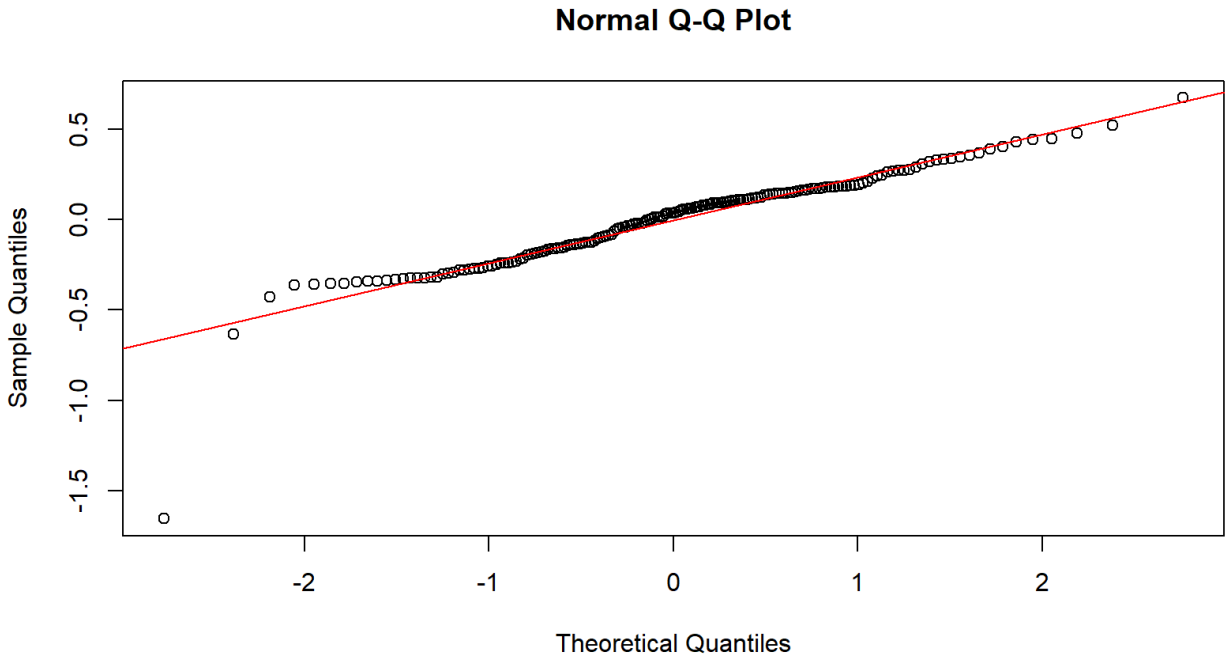
To reduce the deviation of residuals, global sales were transformed using a log function.

```
Call:
lm(formula = log_Global_Sales_Sum ~ NA_Sales_Sum + EU_Sales_Sum,
    data = sales_cols3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.65630 -0.16334  0.03887  0.15755  0.67551

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.565133   0.030710   50.96   <2e-16 ***
NA_Sales_Sum 0.067736   0.005476   12.37   <2e-16 ***
EU_Sales_Sum 0.084205   0.008090   10.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.258 on 172 degrees of freedom
Multiple R-squared:  0.7994,    Adjusted R-squared:  0.7971
F-statistic: 342.8 on 2 and 172 DF,  p-value: < 2.2e-16
```

Model Evaluation

Independent Variables	R-squared	F-stat and Prob (F-statistic)	Coefficients	Residual Standard Error	Summary
NA Sales and EU Sales	This suggests that approximately 79.94% of the variability in log-transformed global sales is explained by the model. It's a high value, indicating a good fit of the model to the data. **Adjusted R squared** - also indicates a good fit, confirming that the model explains a substantial portion of the variance in log-transformed global sales.	This tests the null hypothesis that all regression coefficients are equal to zero versus at least one of them is not. The very low p-value (< 2.2e-16) indicates we can reject the null hypothesis, meaning the model as a whole is statistically significant.	Both predictors are statistically significant (p-values < 2e-16), indicating a strong linear relationship between each predictor and the response variable	In this context, it indicates that the actual log-transformed global sales values typically deviate from the predicted values by about 0.258 on the log scale.	Significant and positive relationship between NA Sales + EU Sales and Log of Global Sales.

Test for Normal Distribution of Residuals




```
        Shapiro-Wilk normality test

data:  residuals(MLR2)
W = 0.90592, p-value = 3.916e-09
```

The transformed data improves the normality of the residuals. However, the r squared is significantly reduced. If Turtle Games intended to conduct hypothesis testing, it might consider using transformed

data. Our interest here is predicting future outcomes and therefore a higher r squared might be preferential.

Predicting Future Values Using MLR1

For given values of NA sales and EU Sales, the model provides the following predictions for Global Sales:

	NA_Sales_Sum	EU_Sales_Sum	Predicted_Global_Sales_Sum
1	34.02	23.80	68.06
2	3.93	1.56	7.36
3	2.73	0.65	4.91
4	2.26	0.97	4.76
5	22.08	0.52	26.63

Testing the Model

The data was split into training and test sets to evaluate the model's adaptability to unseen data.

The following results were obtained:

Root Mean Squared Error: 1.65557

R Squared: 0.9573971

This implies the model adapts well to unseen (test) data.

Are there any possible relationships between North American, European, and global sales?

NA Sales and Global Sales

The results suggest that North American sales are a strong predictor of global sales. The model shows a good fit and predictive power. However, the presence of significant residuals suggests that while the model is useful, it might not capture all aspects of the global sales dynamics.

- **R-squared**: The R-squared value is 0.8395, meaning approximately 83.95% of the variance in global sales can be explained by North American sales. This is a high R-squared value, suggesting a strong fit of the model to the data.
- **F-statistic and p-value**: The F-statistic is 904.7 with a p-value of less than 2.2e-16, suggesting the model is statistically significant. This means that the relationship observed is very unlikely to be due to chance.

- **Predictive Power**: The model has strong predictive power due to the high R-squared value, but care should be taken in predicting values far outside the range of the data used to fit the model, especially considering the potential outliers hinted at by the residuals.
- **Use in Decision Making**: The strong relationship between North American sales and global sales can inform strategic decisions, such as focusing marketing efforts or product availability in North America to boost global sales.
- **Limitations and Considerations**: Despite the strong relationship, it's important to consider other factors that might affect global sales, as this model only accounts for North American sales. External factors, market conditions, or other regional sales could also play significant roles and might need to be included in a more comprehensive model for more accurate predictions.

NA Sales and EU Sales

The model shows a statistically significant positive relationship between European sales and North American sales. However, the model explains less than half of the variability in North American sales, suggesting that other factors not included in the model may also play a significant role. The residuals indicate that while the model fits many observations well, there are outliers or extreme cases where the model's predictions are less accurate. This analysis provides valuable insights but also highlights the importance of considering additional variables or more complex models to improve understanding and prediction of sales patterns.

Residuals

- The residuals, or differences between the observed and predicted values, have a wide range. The minimum is -8.7273 and the maximum is 20.9338, which might indicate some outliers or extreme values in the data.
- The 1st quartile (Q1) and the 3rd quartile (Q3) suggest that 50% of the residuals fall between -1.2982 and 0.7136, indicating that the model's predictions are relatively close for many observations, but there may be some exceptions with large errors.

R-squared (0.3856): This value tells us that approximately 38.56% of the variability in North American sales can be explained by European sales. While this indicates a moderate level of explanatory power, more than 60% of the variability is still unexplained by the model.

F-statistic (108.6): This tests the null hypothesis that all regression coefficients are equal to zero (no linear relationship). The very small p-value (< 2.2e-16) associated with the F-statistic suggests that the model is statistically significant.

Multiple Linear Regression

Independent Variables	R-squared	F-stat and Prob (F-statistic)	Coefficients	Residual Standard Error	Summary
NA Sales and EU Sales	Approximately 96.68% of the variability in Global_Sales_Sum can be explained by the linear relationship with NA Sales and EU Sales. This is a very high R-squared value, indicating a strong model fit.	The F-test is highly significant (p-value < 2.2e-16), indicating that the model is statistically significant, and there is a linear relationship between the response and the predictors.	Both predictors are statistically significant (p-values < 2e-16), indicating a strong linear relationship between each predictor and the response variable	1.49 on 172 degrees of freedom, indicating a good fit.	Significant and positive relationship between NA Sales + Eu Sales and Global Sales.

The results of the multiple linear regression indicate a strong relationship between global video game sales and the sales within North America and Europe.

The model is highly significant with a strong fit, as indicated by the high R-squared values and the significance of the coefficients for sales in North America and Europe. It suggests that these regional sales are very good predictors of global sales. However, the presence of outliers and the size of residuals might suggest exploring more sophisticated models or adding other variables to improve predictions further.