

**Data Analytics Career Accelerator**

**Course 3 - Assignment**

**Advanced Analytics for Organisational Impact**

(Using Python and R)

**Technical Report**

Somayeh Davodi Manesh

October 2023 - Cohort 2

London School of Economy and Politics

Contents

[1 Introduction 3](#_Toc147683973)

[About the project: 3](#_Toc147683974)

[Business questions: 3](#_Toc147683975)

[2 Data analyst approach 3](#_Toc147683976)

[**2.1. First phase: Analysis in Python** 3](#_Toc147683977)

[**2.2. Second phase: Analysis in R** 5](#_Toc147683978)

# Introduction

## About the project:

Turtle Games, a gaming company, both creates and sells their own products and sources and sells items made by other firms. They offer a wide range of products, including books, board games, video games, and toys, to a global customer base. Their primary goal is to enhance sales by leveraging customer trends.

## Business questions:

Marketing department:

1. How customers accumulate loyalty points.
2. How useful are remuneration and spending scores data.
3. Can social data (e.g. customer reviews) be used in marketing campaigns?

Sales department:

1. What is the impact on sales per product?
2. The reliability of the data (e.g. normal distribution, Skewness, Kurtosis)
3. If there is any possible relationship(s) in sales between North America (NA), Europe (EU), and global sales.

# Data analyst approach

In this project, we utilised both Python and R for data analysis. Python was applied to address the first three business questions of marketing team, while R was employed to handle the remaining inquiries for sales team.

## **2.1. First phase: Marketing department (Analysis in Python)**

### **2.1.1 Exploratory data Analysis:**

**Part 1: Loyalty point influencers**

The marketing department wants to better understand how users accumulate loyalty points. Therefore, the possible relationships between the loyalty points, age, remuneration, and spending scores were investigated.

Data exploration and data cleansing:

Necessary libraries imported and review dataset was loaded. Data set contain 2000 rows and 11 columns. Neither missing data nor duplication detected. Redundant columns removed and remaining column names changed to informative names.

Descriptive analysis:

The existence of relationship between loyalty point and each of spending score, remuneration and age features was analysed using OLS (ordinary least square) linear regression model and scatter plots.

A moderate positive correlation was detected between loyalty-Spending score (R-sq = 0.45) and loyalty-Renumeration (0.38) but no correlation between loyalty point and age (R-sq = 0.002).

Recommendations:

* Reward Program Focus: Concentrate loyalty rewards on customers with higher spending scores and remunerations due to their stronger loyalty correlation.
* Targeted Marketing: Design personalised marketing for these high-value customer segments to drive loyalty and repeat business.
* Age Consideration: While age shows a weak correlation with loyalty, explore generational preferences and adapt strategies accordingly.
* Continuous Adaptation: Monitor loyalty program impact and customer behaviours, adjusting strategies as needed.

**Part 2: Usefulness of renumeration and spending scores for customer segmentation**

In order to assist marketing department to better understand the usefulness of renumeration and spending scores to target specific market segments, k-means clustering was employed to group customers based on income and spending score characteristics.

Analysis approach:

To achieve this objective, we began by cleaning the data and removing unnecessary columns. Next, the relationship between renumeration and spending score was examined using pair-plots, allowing us to observe how data points were distributed relative to each other.

Subsequently, the Elbow and Silhouette methods were employed to identify the optimal number of clusters. Based on the results, we determined that using five clusters was the best approach for the final prediction model.

Recommendations:

* Customise marketing campaigns for each cluster to align with their preferences and behaviours.
* Suggest products that resonate with each cluster's interests and needs.
* Adjust prices based on the price sensitivity of each cluster.
* Use preferred communication channels for each cluster to maximize engagement.

**Part 3: Polarity, and sentiment analysis (NLP)**

The marketing department aims to discover the 15 most frequently used words in online reviews. Additionally, they seek to compile lists of the top 20 positive and negative reviews received from the website.

Analysis approach:

After data cleansing, including removing duplicates and redundant columns, we focused on the review and summary sections for NLP analysis. Both columns, "reviews" and "summary," underwent sentiment analysis steps like punctuation removal and tokenization. Word clouds were generated to identify frequent words, and results favoured the removal of English stop words.

Polarity and sentiment scores were computed for both columns and visualized using bar plots and histograms. To validate the analysis, 20 negative and 20 positive reviews/summaries were cross-checked, revealing slight discrepancies in scores. Interestingly, the "summary" column exhibited higher accuracy compared to the "review" column.

Technical recommendations:

* Consider filtering observations where the polarity of reviews and summaries conflict, allowing for a more in-depth analysis.
* Exploring the potential benefits of calculating the mean or weighted average of sentiment scores from both columns to enhance accuracy is advisable.
* Furthermore, conducting granular sentiment analysis based on specific game types is valuable due to their distinct characteristics and applications.

Recommendations to marketing team:

Based on the results showing predominantly positive sentiments with noticeable neutral ones and smaller number of negatives:

* Leverage positive sentiments to reinforce the positive aspects in marketing campaigns.
* Engage with customers who express neutral and negative sentiments to gather feedback and better understand their needs and expectations.
* Compare sentiments with competitors for advantages and areas to differentiate in your products/services.
* Establish a system for continuous sentiment monitoring which can help in adapting marketing strategies as sentiments evolve.

## **2.2. Second phase: Sales departments (Analysis in R)**

### **2.2.1 Exploratory data analysis**

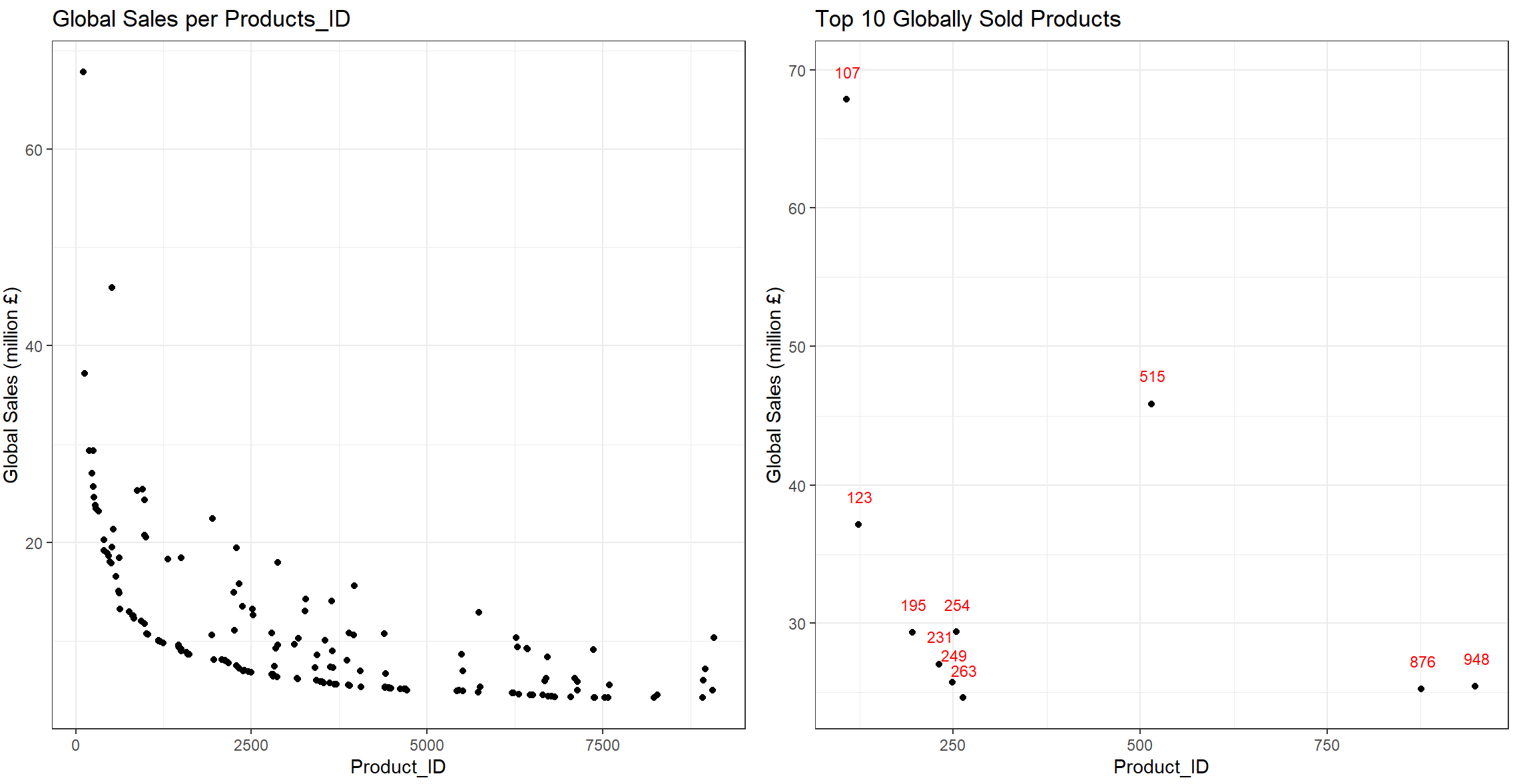
Descriptive analysis:

Initially, we imported essential R packages like 'tidyverse' and 'ggplot2', loaded the "turtul\_sales" dataset, and explored the data. The dataset initially had dimensions of 352 x 9, but we simplified the analysis by removing redundant columns and creating a subset data frame. There were no missing or duplicate values. For descriptive statistics, we calculated the minimum, maximum, and mean values for each sales region and conducted a preliminary analysis using the summary function.

**Part 1: Impact of products on sale**

Analysing global sales by product\_ID revealed a declining trend, which is intriguing because product\_ID is categorical. We can discuss with the client if there's a convention defining product\_ID.

To identify the best-selling products, we displayed the top 10 products. Further analysis can be done by filtering the dataset based on other characteristics like publisher, platform, or genre to pinpoint the top-selling products in each category. This is valuable for stakeholders, especially in marketing, planning product-specific advertisements.



To enhance the analysis of product impact on sales, we can encode product\_ID, create dummy variables, and apply machine learning models like multiple linear regression to explore potential relationships with specific products.

**Part2: The reliability of the data (e.g. normal distribution, Skewness, Kurtosis):**

Data visualisation, sales distribution and their possible relationship with each other:

Various plot types were utilised, including scatter plots, histograms, and boxplots, using qplot to gain a broad understanding of the data and select the most informative visuals for business purposes.

* Scatter plots revealed a strong positive correlation among EU, NA, and global sales.

A graph with a line and a dotted line

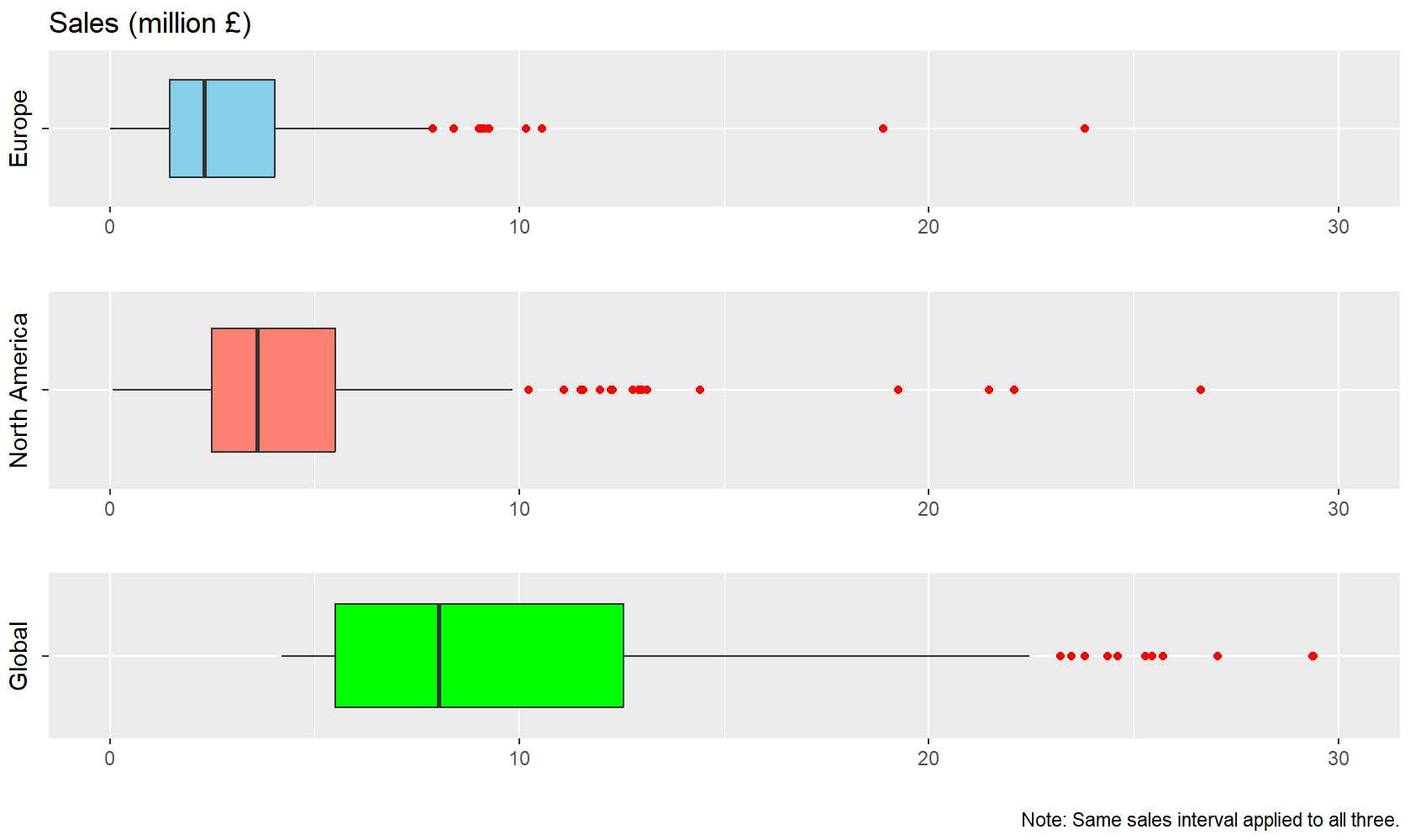
Description automatically generated

* Histograms displayed similar sale distributions for all three regions, with a right skewness and a noticeable left-side peak (asymmetric distribution).

A group of graphs showing different sizes of graphs

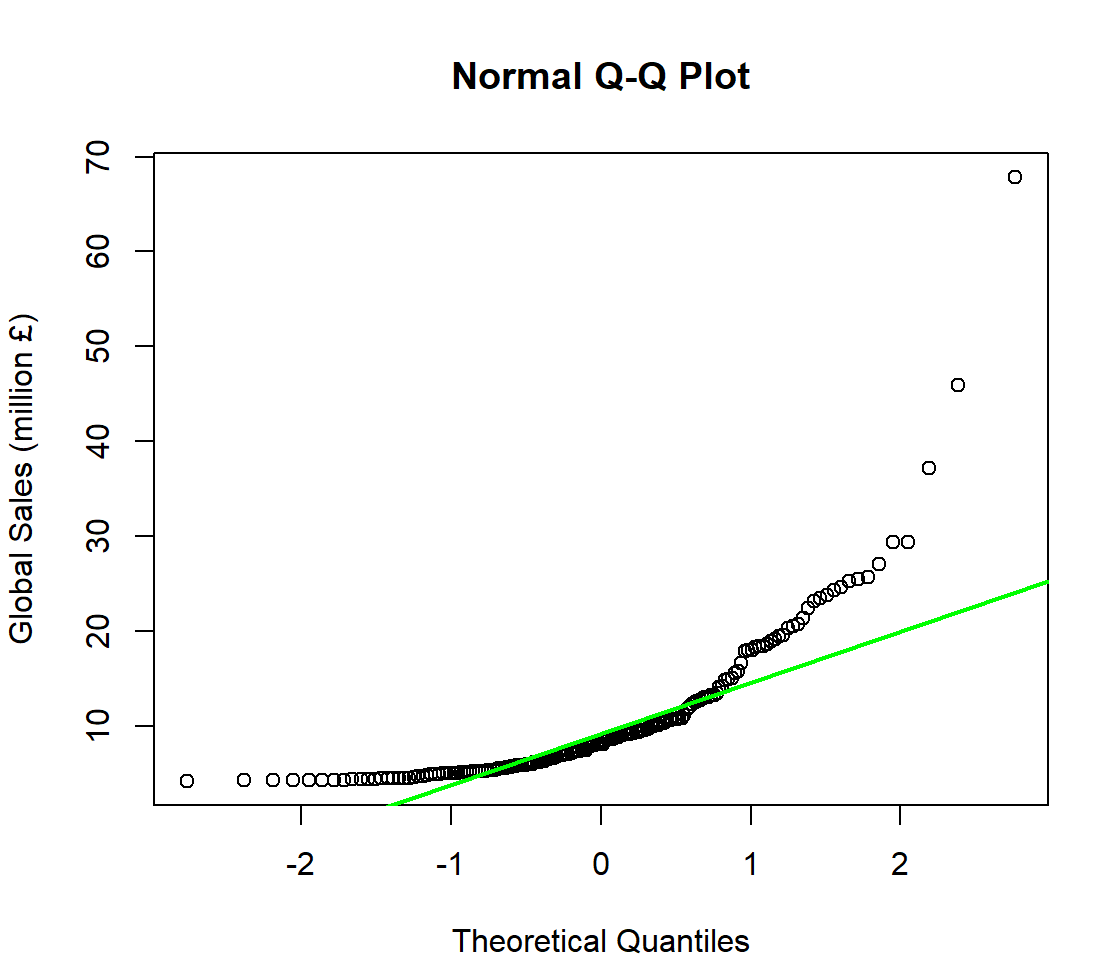
Description automatically generated with medium confidence

* Boxplots indicated that Europe's sales distribution had less variability (smaller IQR and range) compared to North America, with generally lower sales amounts and a smaller median. Global sales exhibited greater variability than regional sales.



To assess data normality, we examined Q-Q plots, conducted Shapiro tests, and calculated skewness, kurtosis, and correlation coefficients for each sales region.

Neither EU, NA, nor global sales exhibited normal distribution based on Q-Q plots and Shapiro tests. The QQ plots deviated from the normal distribution trend, and the Shapiro test p-values for all three regions were less than 0.05 (example Q-Q plot for global sales shown below).



All three regions displayed relatively high positive skewness (around 3), confirming the observed right skewness in histograms. Large kurtosis values (approximately 16) aligned with the prominent left-side peaks in the histograms.

A number of numbers on a white background

Description automatically generated

**Part 3: Modelling the global sales based on sales in other regions:**

Correlation coefficients and scatter plots indicated a positive linear relationship between each pair of regions. Notably, global sales exhibited a stronger correlation with either NA sales or European sales compared to the correlation between EU and NA sales themselves.

A graph of different colored lines

Description automatically generated with medium confidence

Conducted data visualization and employed two simple linear regression (SLR) models and one multiple linear regression (MLR) model to determine the best-fitting model.

The SLR between global sales and EU revealed that EU sales significantly predict global sales (p-value < 0.05). The R-squared value of 0.77 suggests that 77% of the data can be explained by the Global\_EU SLR model.

Similarly, the SLR between global sales and NA showed that NA sales significantly predict global sales (p-value < 0.05) with an R-squared value of 0.87.

The Global\_EU\_NA MLR model demonstrated that both EU sales and NA sales are significant predictors for global sales (both coefficients with p-values < 0.05). The high R-squared value of 0.97 indicates that 97% of the data fits well with the Global\_EU\_NA MLR model. Consequently, the multiple linear regression model was chosen as the best model.

Using five new data points with known EU and NA sale values model accuracy was measured by MAPE (mean absolute percentage error). A low value of 10.2%, obtained, indicating excellent model performance.