# Final Project Tech Pro Academy | Data Science Stream

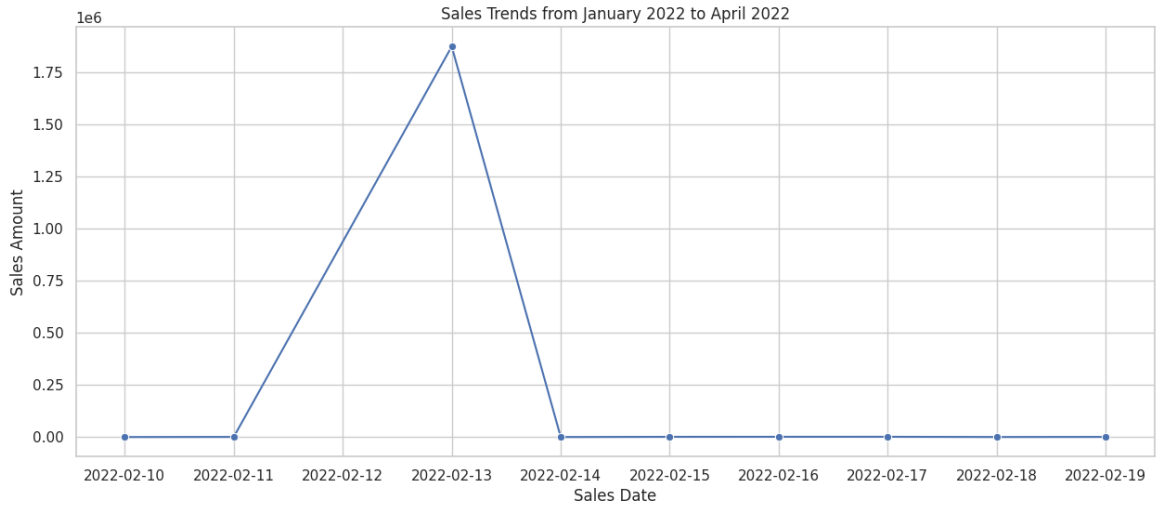
## DATA CLEANING & DATA PREPROCESSING

Upon receiving a dataset concerning a hypothetical company scenario, the primary objective was to analyze factors influencing the target variable, Sales Amount, with the aim of maximizing profits. The initial phase of the analysis involved data cleaning and preprocessing to ensure data integrity and prepare it for further analysis.

1. **SalesDate**: This column denotes the date of each product purchase. The data was transformed into the appropriate datetime format and subsequently segmented into yearly and monthly subsets for analytical purposes.

2. **ProductCategory**: Categorizing products into three distinct categories—Clothing, Electronics, and Home Appliances—was observed in this column. To convert this categorical variable into a numerical format, methods such as one-hot encoding and get dummies were applied. One-hot encoding facilitated the conversion of categories into numerical values, while get dummies created separate columns for each category, representing their presence(1) or absence(0) in the dataset.

3. **SalesAmount**: Representing the monetary value of each purchase, this column underwent preprocessing to retain only numeric characters. Additionally, an outlier observed on February 13, 2023, was removed to maintain data consistency and accuracy.




4. **CustomerAge**: This column records the age of each customer. To ensure statistical robustness, outliers were identified and removed, considering a plausible age range from 18 to 90 years. Furthermore, age groups were delineated to differentiate between customers aged 20- 22 and those aged 35, facilitating more targeted analysis.

5. **CustomerGender**: Reflecting the gender identity of customers, this column included three options: Male, Female, and Non-binary, alongside responses denoted as Unknown or Not Answered, combined. Additionally, erroneous entries such as "Clothing" were corrected based on contextual cues as there was “Female” in product category column, which means that there was a mistake. Utilizing one-hot encoding, categorical data was transformed into numerical format, ensuring compatibility with analytical techniques.

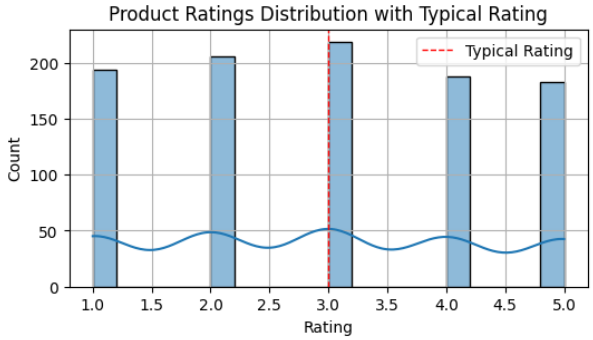
6. **CustomerLocation**: Describing the geographical location of customers, this column contained six distinct options: Japan, Australia, India, USA, UK, and Canada. To address missing values, imputation was performed using the most frequent response. Similar to gender, one-hot encoding was employed to convert categorical data into a numerical representation, enabling quantitative analysis.

7. **ProductRatings**: Reflecting the ratings assigned to each product, ranging from 1 (low) to 5 (high), this column provides insights into product quality and customer satisfaction.

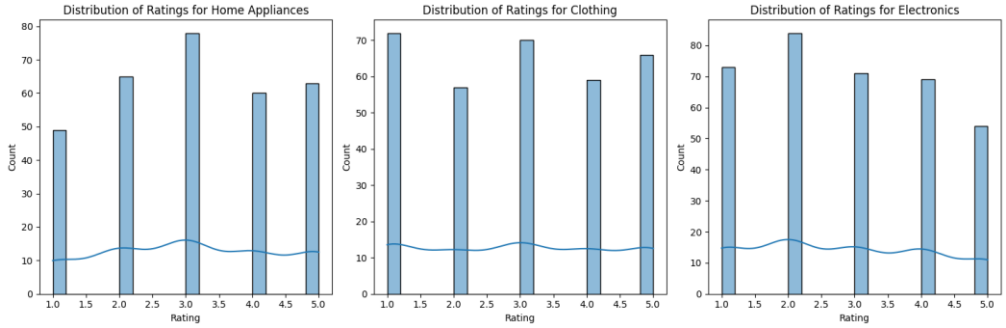
By systematically cleaning and preprocessing the dataset, we have ensured data quality and prepared it for subsequent analysis aimed at maximizing profits and identifying key factors influencing sales performance.

## EXPLORATORY DATA ANALYSIS

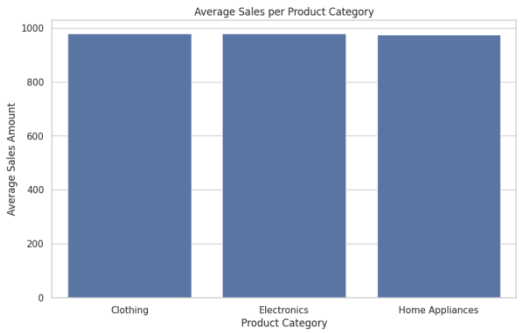
During the exploratory data analysis (EDA) phase, descriptive statistics were computed to understand the distribution of the sales amount variable. It was observed that the typical range of sales amount is approximately 3 units, and the distribution deviates from normality due to the discrete nature of the variable.



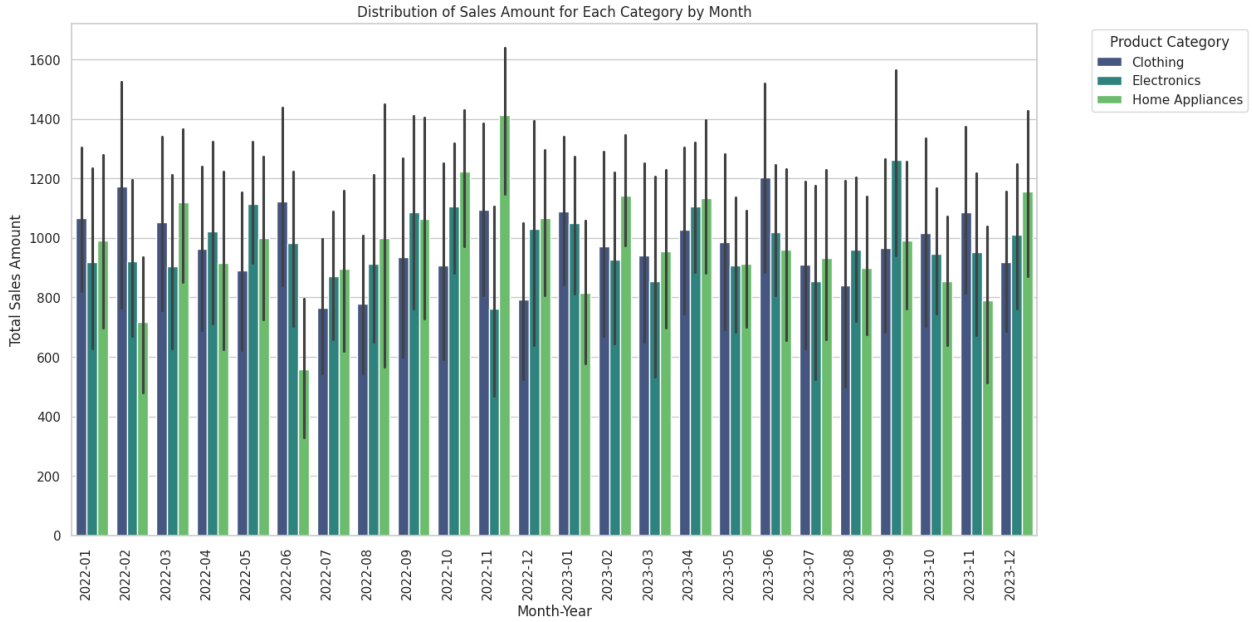

Furthermore, histograms were constructed to visualize the distribution of sales amount across different product categories. These histograms revealed notable differences in distribution patterns among the product categories.




Subsequently, a new histogram depicting the mean sales amount for each product category was generated. The analysis indicated that the mean sales amounts across the product categories were relatively similar and did not exhibit significant disparities.



Examining the distribution of sales by month for each product category revealed distinct purchasing trends. Specifically, home appliances were predominantly purchased in November 2022, clothing in June 2023, and electronics in September 2023. Additionally, comparative analyses were conducted to identify the most commonly purchased product category for each month.




Further investigation was conducted to ascertain the relationship between the quantity of products purchased and the resulting profits. Contrary to expectations, instances were noted where exceeding the mean product count did not correspond to proportionally higher sales amounts. This observation suggests that the mere increase in the number of products purchased does not necessarily translate to increased profitability.

### Count of Products Bought Each Month:

| | Month Year | Clothing | Electronics | Home Appliances |
|---|---|---|---|---|
| 0 | 2022-01 | 16 | 13 | 16 |
| 1 | 2022-02 | 11 | 11 | 11 |
| 2 | 2022-03 | 13 | 18 | 13 |
| 3 | 2022-04 | 13 | 12 | 16 |
| 4 | 2022-05 | 20 | 09 | 12 |
| 5 | 2022-06 | 11 | 14 | 13 |
| 6 | 2022-07 | 16 | 19 | 10 |
| 7 | 2022-08 | 07 | 11 | 09 |
| 8 | 2022-09 | 10 | 08 | 14 |
| 9 | 2022-10 | 10 | 20 | 11 |
| 10 | 2022-11 | 14 | 09 | 12 |
| 11 | 2022-12 | 14 | 10 | 12 |
| 12 | 2023-01 | 14 | 19 | 11 |
| 13 | 2023-02 | 14 | 12 | 08 |
| 14 | 2023-03 | 14 | 10 | 17 |
| 15 | 2023-04 | 14 | 20 | 14 |
| 16 | 2023-05 | 14 | 20 | 16 |
| 17 | 2023-06 | 12 | 20 | 13 |
| 18 | 2023-07 | 13 | 15 | 13 |
| 19 | 2023-08 | 11 | 18 | 15 |
| 20 | 2023-09 | 16 | 10 | 13 |
| 21 | 2023-10 | 12 | 16 | 18 |
| 22 | 2023-11 | 16 | 16 | 13 |

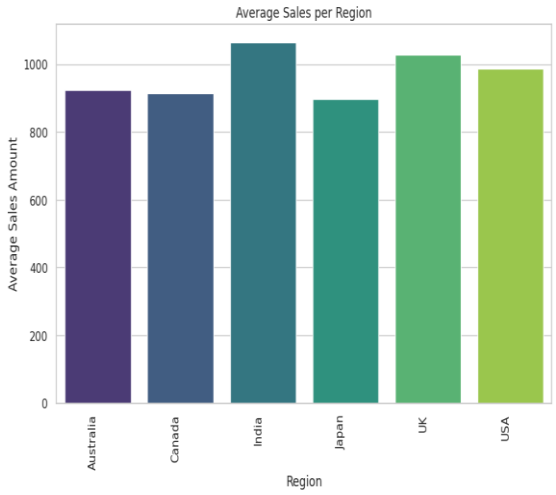
| | Month Year | Clothing | Electronics | Home Appliances |
|---|---|---|---|---|
| 0 | 2022-01 | 17040 | 11942 | 15848 |
| 1 | 2022-02 | 12909 | 10139 | 07880 |
| 2 | 2022-03 | 13680 | 16272 | 14538 |
| 3 | 2022-04 | 12520 | 12244 | 14657 |
| 4 | 2022-05 | 17796 | 10035 | 11988 |
| 5 | 2022-06 | 12349 | 13739 | 07239 |
| 6 | 2022-07 | 12248 | 16557 | 08969 |
| 7 | 2022-08 | 05451 | 10038 | 08987 |
| 8 | 2022-09 | 9358 | 08689 | 14899 |
| 9 | 2022-10 | 09059 | 22130 | 13464 |
| 10 | 2022-11 | 15317 | 06851 | 16963 |
| 11 | 2022-12 | 11083 | 10301 | 12811 |
| 12 | 2023-01 | 15235 | 19944 | 8958 |
| 13 | 2023-02 | 13614 | 11116 | 9134 |
| 14 | 2023-03 | 13155 | 08553 | 16250 |
| 15 | 2023-04 | 14400 | 22110 | 15871 |
| 16 | 2023-05 | 13786 | 18137 | 14582 |
| 17 | 2023-06 | 14452 | 20383 | 12473 |
| 18 | 2023-07 | 11811 | 12823 | 12136 |
| 19 | 2023-08 | 09251 | 17303 | 13460 |
| 20 | 2023-09 | 15451 | 12610 | 12898 |
| 21 | 2023-10 | 12189 | 15128 | 15389 |
| 22 | 2023-11 | 17390 | 15224 | 10257 |

Keeping only above average:

| | Month Year | Product Count |
|---|---|---|
| 0 | 2022-01 | 45 |
| 2 | 2022-03 | 44 |
| 6 | 2022-07 | 45 |
| 12 | 2023-01 | 44 |
| 15 | 2023-04 | 48 |
| 16 | 2023-05 | 50 |
| 17 | 2023-06 | 45 |
| 19 | 2023-08 | 44 |
| 21 | 2023-10 | 46 |
| 22 | 2023-11 | 45 |
| 23 | 2023-12 | 54 |

| | Month Year | Sales Amount |
|---|---|---|
| 0 | 2022-01 | 44830 |
| 2 | 2022-03 | 44490 |
| 9 | 2022-10 | 44653 |
| 12 | 2023-01 | 44137 |
| 15 | 2023-04 | 52381 |
| 16 | 2023-05 | 46505 |
| 17 | 2023-06 | 47308 |
| 20 | 2023-09 | 40959 |
| 21 | 2023-10 | 42706 |
| 22 | 2023-11 | 42871 |
| 23 | 2023-12 | 54984 |

Additionally, statistical tests were performed to examine the distribution of sales across different customer locations. The analysis revealed that India, UK, and USA exhibited higher sales amounts compared to other locations.




In conclusion, the exploratory data analysis provided valuable insights into the distribution of sales amounts across various dimensions, including product categories, months, and customer locations. These findings contribute to a deeper understanding of purchasing trends and profitability factors, which can inform strategic decision-making processes within the organization.

## HYPOTHESIS TESTINGS:

In addition to the aforementioned analyses, further investigation was conducted utilizing ANOVA and Kruskal-Wallis tests to examine the potential impact of other variables on sales amount. However, the results of these tests did not yield statistically significant findings, leading to the retention of the null hypothesis.

These outcomes indicate that there is insufficient evidence to reject the null hypothesis, suggesting that the variables under consideration do not exert a significant influence on sales amount. Such findings underscore the importance of rigorous statistical testing and interpretation in discerning meaningful relationships within the dataset.

Despite the lack of statistical significance, these analyses contribute valuable insights by elucidating the absence of notable associations between the variables examined and sales amount. This underscores the need for comprehensive exploration and consideration of various factors when assessing factors affecting profitability within the organization.

However, it was observed that only the factor of Customer Location exhibited a statistically significant influence on sales amount. The results of these tests are presented below:

1. One-way ANOVA:
   - F-statistic: 2.5075884796973975
   - p-value: 0.028797924547020327
   - Interpretation: The one-way ANOVA test yielded a statistically significant result ($p < 0.05$), indicating a notable difference in average sales among different Customer Location groups.

Advantages of ANOVA:
- ANOVA is suitable for comparing means across multiple groups simultaneously.
- It assumes normality and homogeneity of variance, which can provide more precise results when met.

Disadvantages of ANOVA:
- ANOVA is sensitive to outliers and assumes the data follows a normal distribution.
- It may produce inaccurate results if the assumption of homogeneity of variances is violated.

2. Kruskal-Wallis test:
   - H-statistic: 12.678848796465873
   - p-value: 0.026581555427246828
   - Interpretation: The Kruskal-Wallis test also yielded a statistically significant result ($p < 0.05$), indicating a significant difference in average sales among Customer Location groups.

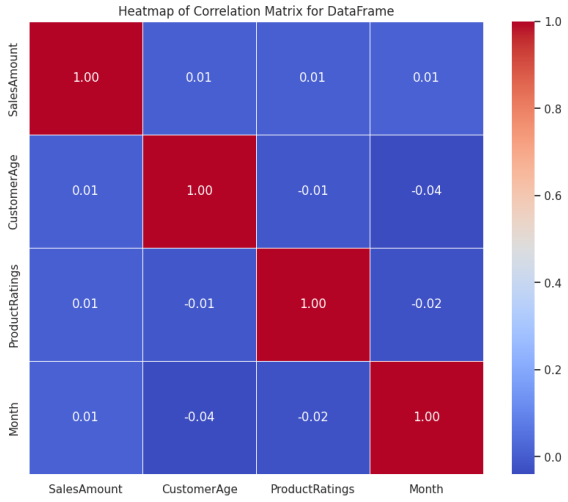
Advantages of Kruskal-Wallis test:
- Kruskal-Wallis is a non-parametric alternative to ANOVA, making it robust against violations of assumptions such as normality and homogeneity of variances.
- It can be applied to ordinal or non-normally distributed data.

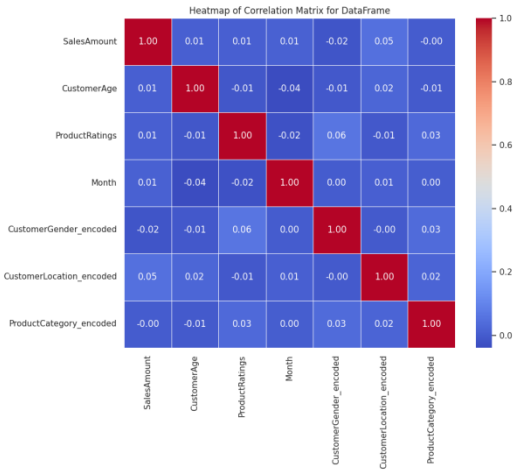
Disadvantages of Kruskal-Wallis test:
- It may have less power compared to ANOVA when assumptions are met.
- Interpretation can be more complex compared to ANOVA when dealing with multiple groups.

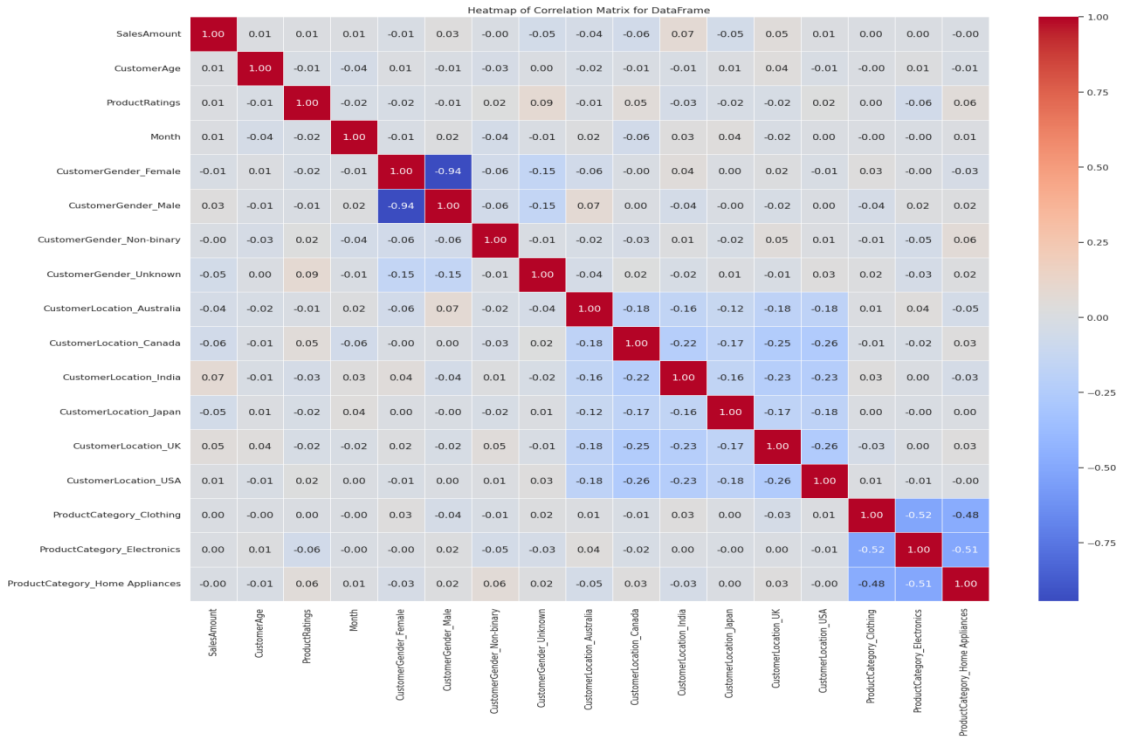
In conclusion, both ANOVA and Kruskal-Wallis tests provided evidence of a significant relationship between Customer Location and sales amount. While ANOVA assumes normality and homogeneity of variances, Kruskal-Wallis offers a non-parametric alternative that is robust to these assumptions, making it suitable for datasets that do not meet parametric assumptions. Therefore, the choice between ANOVA and Kruskal-Wallis depends on the nature of the data and the fulfillment of parametric assumptions. So as the provided dataset is about 1000 values, so this is not too large, and most variables are not normally distributed, Kruskal-Wallis is a better choice but as it has less power than ANOVA, I ran both tests to ensure that there was a coefficiency for the datasets.

## PREDICTIVE MODELING:

Initially, a heatmap was constructed to assess correlations among variables. It was observed that several variables contained missing values, prompting the utilization of one-hot encoding and get dummies techniques to handle categorical data. Subsequent correlation analysis revealed that correlations were generally below 0.1, except for correlations inherent within variables that originated from the same column prior to the application of get dummies.




Following this initial exploration, linear regression models were fitted to further investigate potential relationships between the variables and the target variable, Sales Amount. However, the obtained results indicated negligible correlations, as evidenced by R-squared values near 0 and large mean squared errors. These results are presented below:

## LINEAR & MULTILINEAR REGRESSION:

Linear Regression Results:

Feature: ProductCategory_encoded
Mean Squared Error: 292048.7779756151
R-squared: -0.003919459381162227

Feature: CustomerAge
Mean Squared Error: 291834.4058554063
R-squared: -0.003182553907637553

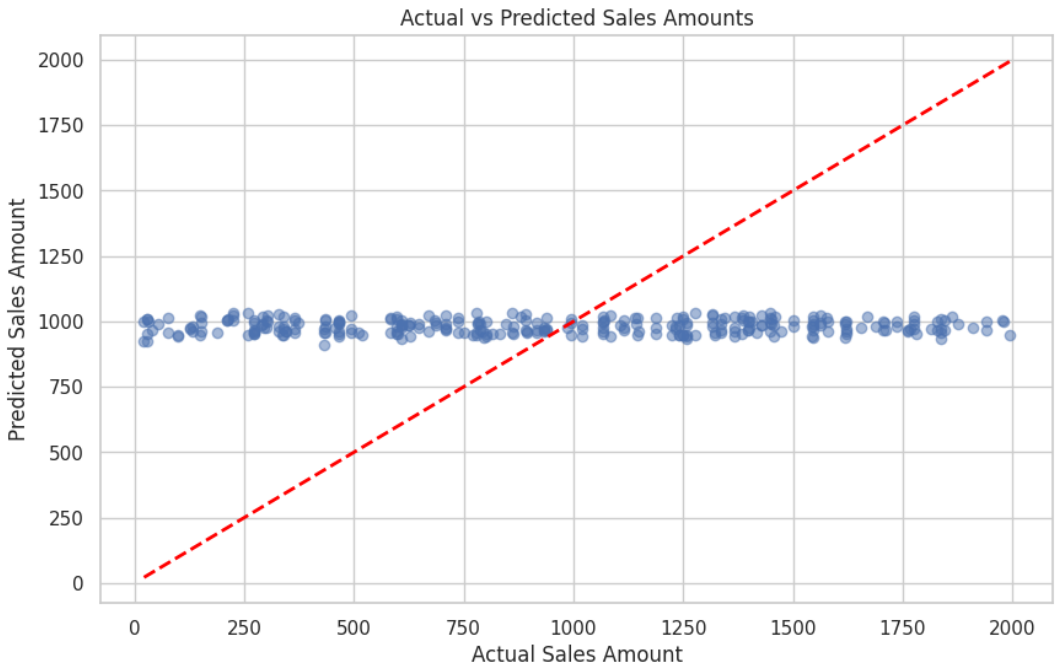
Feature: CustomerGender_encoded
Mean Squared Error: 291671.4819556473
R-squared: -0.0026225020063759263

Feature: CustomerLocation_encoded
Mean Squared Error: 291014.55162712437
R-squared: -0.0003642999867226049

Feature: ProductRatings
Mean Squared Error: 291902.07620086893
R-squared: -0.003415171133786865

Additionally, a multilinear regression model was employed to explore potential interactions among the variables. However, the resultant model also demonstrated poor performance, with negligible R-squared values and large mean squared errors.

Multilinear Regression Results:




Mean Squared Error: 291141.22045834357
R-squared: -0.00079972486827673
Coefficients:
ProductCategory_encoded: -8.379964633090628

CustomerAge: -0.20283148319814792
CustomerGender_encoded: -10.55349594970362
CustomerLocation_encoded: 13.944788265612374
ProductRatings: -1.2208577302984207
Intercept: 970.7175458149605

In summary, despite the rigorous analysis conducted, no significant correlations were found between the examined variables and sales amount. These results underscore the complex nature of the relationships within the dataset and highlight the need for further exploration and refinement of analytical approaches to elucidate factors influencing sales performance more comprehensively.

## DECISION TREE & RANDOM FOREST

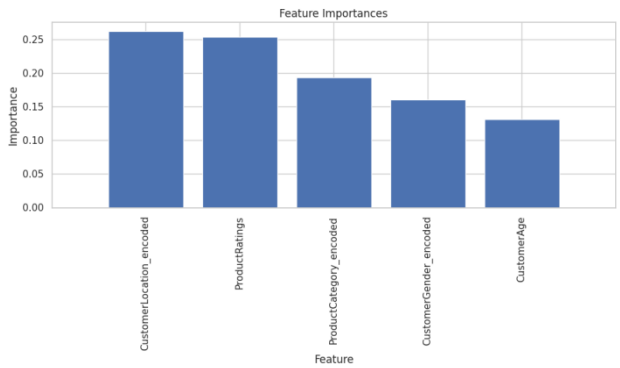
Subsequently, a decision tree analysis was conducted to explore potential nonlinear relationships between the variables and sales amount. However, the outcomes revealed unsatisfactory performance metrics, with a mean squared error of 296380.2492918259 and an R-squared value of -0.005816002254210861. These results indicate that the decision tree model did not effectively capture the variation in sales amount explained by the predictor variables.

Despite the utilization of a more flexible modeling approach with decision trees, the inability to attain a satisfactory level of predictive accuracy suggests the presence of inherent complexities or noise within the dataset that are not adequately captured by the selected features. This underscores the need for further refinement of the analytical methodology or consideration of alternative modeling techniques to better elucidate the factors influencing sales performance.

Following the exploration of various modeling techniques, a random forest algorithm was employed to discern significant predictors of sales amount. The analysis indicated that customer location and product ratings emerged as more influential factors compared to product category, customer age, and gender. However, despite these notable observations, the model's performance metrics were not optimal, with a mean squared error of 348149.8202164874 and an R-squared value of -0.18150470954923814.

While the random forest model exhibited promising insights regarding variable importance, the relatively high mean squared error and negative R-squared value suggest that the model's predictive capability remains limited. These results imply that the model may not adequately capture the underlying relationships within the data or may be influenced by factors not accounted for in the analysis.

Therefore, further refinement of the random forest model or exploration of alternative modeling approaches may be warranted to improve predictive accuracy and better understand the determinants of sales amount within the dataset. Additionally, careful consideration of feature selection, parameter tuning, and potential interactions among variables may enhance the model's performance and provide deeper insights into the factors driving sales performance.



In accordance with the sales distribution across various customer locations, the following recommendations can be made regarding advertising strategies for each product category:

Electronics:

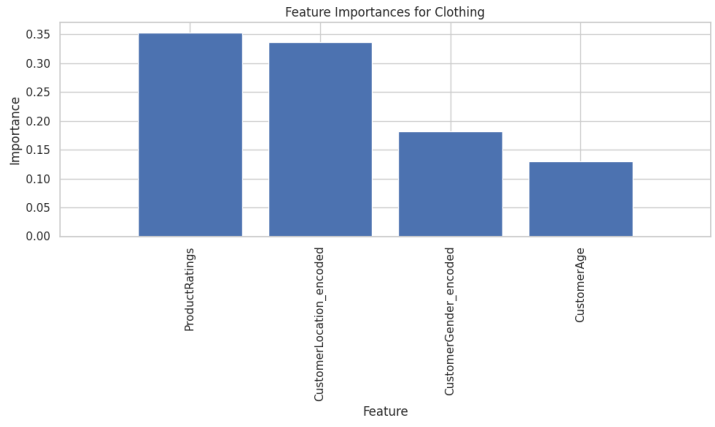
As Electronics demonstrate a consistent sales trend across multiple locations, advertising efforts should prioritize channels that reach a broad audience.
Emphasize the technological advancements, reliability, and performance of Electronics products to appeal to a wide range of consumers.
Utilize diverse marketing platforms and strategies to effectively showcase the value proposition of Electronics across different geographic regions.
Home Appliances:

Home Appliances exhibit varying degrees of popularity across different locations, suggesting a need for targeted advertising initiatives.
Focus advertising efforts on regions where Home Appliances have shown higher sales amounts, such as the USA and Canada.
Highlight the practicality, efficiency, and convenience of Home Appliances in marketing campaigns tailored to the preferences and needs of specific geographic markets.
Clothing:

While Clothing sales may not be as widespread as Electronics or Home Appliances, targeted advertising in select regions can still yield favorable results.
Tailor advertising messages and campaigns to resonate with the fashion preferences and lifestyle aspirations of consumers in key markets, such as the UK and India.
Leverage digital marketing channels and platforms that are popular among fashion-conscious demographics to increase brand visibility and drive Clothing sales.
By aligning advertising strategies with sales data and leveraging insights from geographic sales distribution, businesses can optimize their marketing efforts to effectively reach target audiences and capitalize on sales opportunities in different regions.

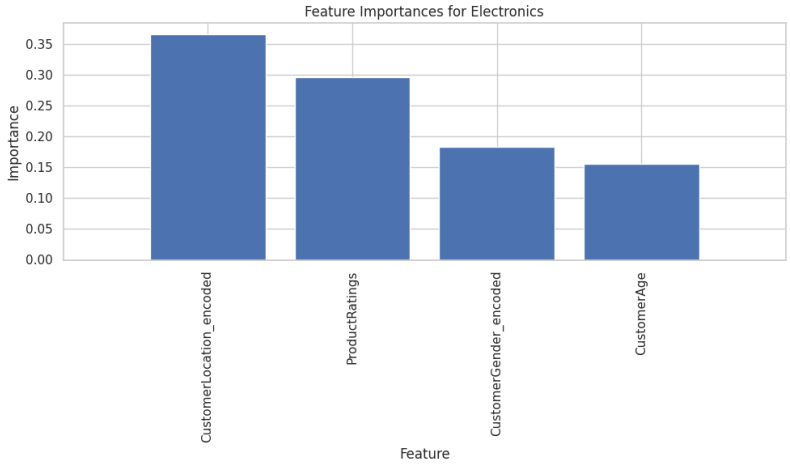
The random forest analysis specifically conducted for the Clothing product category yielded the following results:

Mean Squared Error: 360503.9100115954
R-squared: -0.2118558992642514



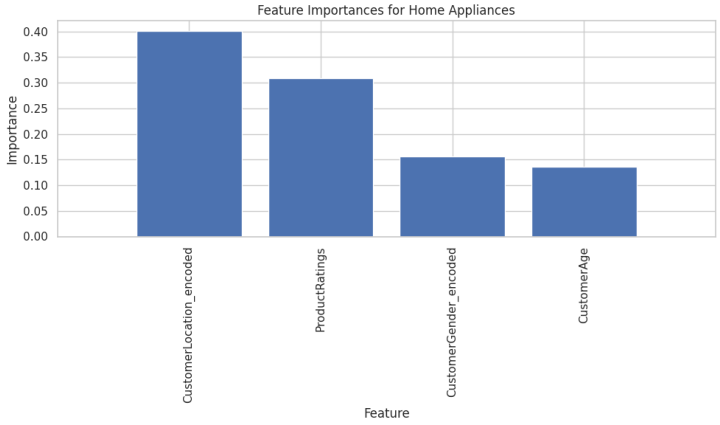

For the Electronics product category, the random forest analysis yielded the following results:

Mean Squared Error: 359172.600326731
squared: -0.4333286092034332




For the Home Appliances product category, the random forest analysis yielded the following results:

Mean Squared Error: 369755.66455983877
R-squared: -0.5343107565828833




Upon splitting the dataset into each product category and applying linear regression with Location and Ratings as predictor variables, the following results were obtained:

Clothing Data:

Coefficients: [7.18083813, 40.23426086]
Intercept: 869.0592573960871
Mean Squared Error: 296056.35403438134
R-squared: 0.004788494139520205
Interpretation: The positive coefficient for Ratings indicates that higher product ratings are associated with higher sales amounts for Clothing. The positive intercept suggests a base level of sales, independent of Location and Ratings. However, the low R-squared value indicates that the model explains only a small proportion of the variability in Clothing sales.

Electronics Data:

Coefficients: [-7.35073874, -9.97178076]
Intercept: 1010.7079758679865
Mean Squared Error: 256537.39793860834
R-squared: -0.023748447018293595
Interpretation: The negative coefficients for both Location and Ratings suggest that neither variable significantly contributes to Electronics sales. The positive intercept indicates a base level of sales,

irrespective of Location and Ratings. However, the negative R-squared value indicates that the model performs poorly in explaining the variability in Electronics sales.

Home Appliances Data:

Coefficients: [20.44298859, 8.50456411]
Intercept: 899.5055035856494
Mean Squared Error: 245946.77095235296
R-squared: -0.02056252922651991
Interpretation: The positive coefficients for both Location and Ratings suggest that both variables positively influence Home Appliances sales. The intercept indicates a base level of sales, regardless of Location and Ratings. However, the negative R-squared value indicates that the model performs poorly in explaining the variability in Home Appliances sales.

In summary, while there are some insights gained from the coefficients, the R-squared values indicate that the linear regression models have limited explanatory power for all product categories. Further exploration or refinement of the models may be necessary to improve their predictive accuracy and better understand the factors driving sales for each product category.

Total Purchases per Product Category for each Location:

Clothing Location Summary:

| CustomerLocation_encoded | Total Purchases | Percentage of Total Purchases |
|---|---|---|
| Australia | 29858 | 9.42 |
| Canada | 54121 | 17.07 |
| India | 67354 | 21.25 |
| Japan | 26493 | 8.36 |
| UK | 66203 | 20.88 |
| USA | 72971 | 23.02 |

Electronics Location Summary:

| CustomerLocation_encoded | Total Purchases | Percentage of Total Purchases |
|---|---|---|
| Australia | 44481 | 12.99 |
| Canada | 67101 | 19.59 |
| India | 59513 | 17.38 |
| Japan | 34185 | 9.98 |
| UK | 66981 | 19.56 |
| USA | 70209 | 20.50 |

Home Appliances Location Summary:

| CustomerLocation_encoded | Total Purchases | Percentage of Total Purchases |
|---|---|---|
| Australia | 29153 | 9.50 |
| Canada | 58720 | 19.13 |
| India | 50003 | 16.29 |
| Japan | 34479 | 11.23 |
| UK | 72688 | 23.68 |
| USA | 61934 | 20.18 |

Based on the summarized data of total purchases and their respective percentages across different locations for each product category, several observations can be made:

Clothing Location Summary:

The UK and USA exhibit the highest percentages of total purchases for Clothing, followed by Canada and India.
Australia and Japan have relatively lower percentages of total purchases for Clothing.

Electronics Location Summary:

Similar to Clothing, the UK and USA demonstrate the highest percentages of total purchases for Electronics, followed by Canada and India.
Australia and Japan have comparatively lower percentages of total purchases for Electronics.
Home Appliances Location Summary:

The UK and USA also lead in percentages of total purchases for Home Appliances, followed by Canada and India.
Australia and Japan show lower percentages of total purchases for Home Appliances.
Based on these observations, it may seem beneficial to allocate advertising and marketing efforts towards locations with lower percentages of total purchases, such as Australia and Japan, for all product categories. However, it's essential to consider additional factors before making conclusive decisions:

Market Potential: Evaluate the market potential and consumer behavior in Australia and Japan to determine if there are untapped opportunities for increased sales.

Competitive Landscape: Assess the competitive landscape in these regions and consider the presence of competitors and their marketing strategies.

Cultural Factors: Take into account cultural preferences, purchasing habits, and socio-economic factors that may influence consumer behavior in Australia and Japan.

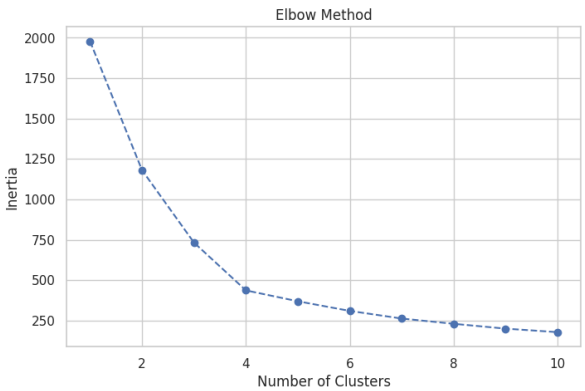
Target Audience: Identify the target audience and tailor advertising and marketing campaigns to resonate with the preferences and needs of consumers in these regions.

In conclusion, while it may seem advantageous to focus advertising efforts in Australia and Japan based on the provided data, a comprehensive analysis considering market potential, competition, cultural factors, and target audience preferences is necessary to formulate effective marketing strategies and maximize sales opportunities in these regions.

## Clustering, K-Means

Following the linear regression analysis, the dataset underwent clustering using the K-means algorithm. The optimal number of clusters was determined using the elbow method, which identifies the point at which the rate of decrease in within-cluster sum of squares (WCSS) slows down significantly, indicating the optimal number of clusters.

The elbow method indicated that the optimal number of clusters is 4, as this point on the plot demonstrated the greatest change in WCSS. This suggests that partitioning the data into 4 clusters captures a substantial amount of variation in the dataset while avoiding overfitting.



Utilizing K-means clustering with 4 clusters facilitates the segmentation of the dataset into distinct groups based on similarities in the features. This approach enables the identification of patterns or subgroups within the data, which may provide valuable insights for decision-making processes such as targeted marketing strategies or customer segmentation.

Overall, employing the elbow method to determine the optimal number of clusters ensures a balanced trade-off between capturing meaningful patterns in the data and avoiding excessive complexity in the model.

Thank you for your time & your attention,
Lilian Vitsa