****

**CST3440: Business Intelligence**

**Coursework 2 Report**

**Submission: Friday 19th April 2024**

**Student ID:M00810926**

**Name: Abdullahi Mohamed**

**CAMPUS: HENDON**

**Contents**

**Introduction…………………………………………………………………………………………………………………Page 3-7**

**Data Analysis & Visualizations ………………………………………………………………………………………Page 8-17**

**Selection of Data Mining Algorithm and Data Pre-Processing ………………………………………..Page 18**

**Data Mining ………………………………………………………………………………………………………………….Page 19-23**

**Data Ethics …………………………………………………………………………………………………………………….Page 24**

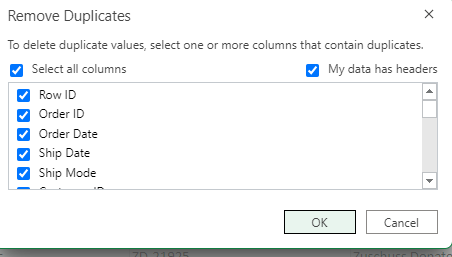
**Conclusions …………………………………………………………………………………………………………………...Page 24**

**Introduction**

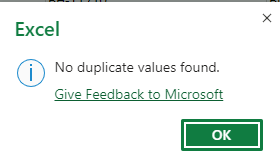
The selected dataset contains sales information from a superstore, a sizable retail store that provides a range of products and services. It includes important factors, including segmentation, quantity amount, profit, order date, and ship date. This dataset was downloaded from Kaggle, a website that provides datasets for experts and the public to use and share. The dataset includes data on product sales, client segmentation, order and shipment dates, and related earnings, and other quantitative and categorical characteristics. It offers an overview of sales activity within the superstore with around 10,000 rows of data. Standard data cleaning techniques were probably carried out before analysis to guarantee data quality. These processes include fixing any inconsistencies in the dataset, handling outliners, addressing missing numbers, and standardizing formats.

Data Cleaning

**Step 1: Remove Duplicates**

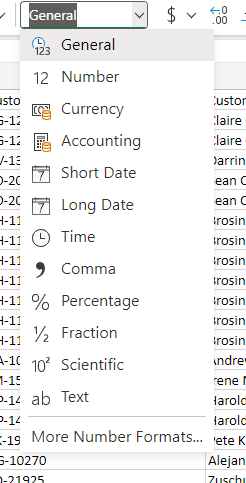
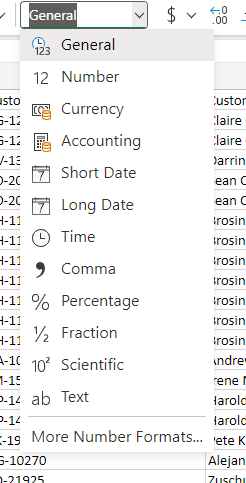


Checking all rows and columns and seeing if there is any duplication.



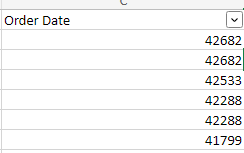
In the screenshot above there is no duplication for any rows meaning that every row is unique, but it is a key step in data cleaning.

**Step 2 Format date**



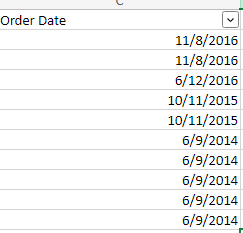
The function short date will be used to ensure the date is correctly formatted.

Before date



This is the date in general format to which is going to be changed into the correct format.

Correct Date Format



A few rows to show the change in date with the correct format.

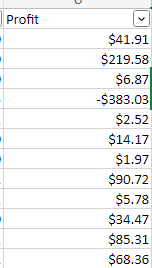
**Step 3 Change decimal point to 2 and currency to dollar sign**



Currently the profit row has a got many decimal points to which will be shortened to 2 and the dollar sign will be used as the dataset is based in the United States location therefore it must match the correct currency.

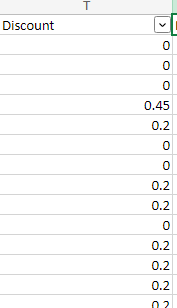


Below is the dollar sign and the symbol needed to convert to just to 2 decimal points.

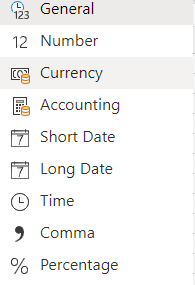
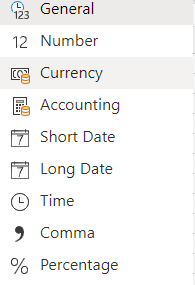


The profit row has now been changed to 2 decimal points each and the dollar sign before the first number.

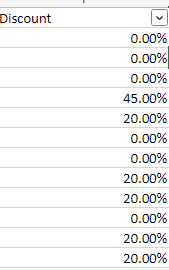
**Step 4 Change discount from general to percentage**



The discount row currently is in decimal points however will be changed for better understanding purposes.

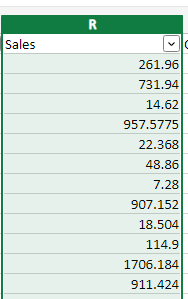


The row is currently in the general format but has been changed to the percentage symbol.

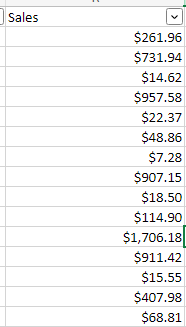


Now it is easier to understand the discounts as they are in the percentage format.

**Step 5 – Change Sales rows.**



This is before changing the rows and the same steps from step 2 will be applied.



Now the superstore sales have the correct currency and a 2-point decimal places.

**Data Analysis and Visualizations**

**Sales Trend Over Time**

A screenshot of a graph

Description automatically generated

The line graph, which allowed for an examination of annual performance, showed sales trends over four years. There were fluctuations, with decreased sales in 2015 despite a bigger number in 2014. On the other hand, 2017 had the largest sales, which would point to a rise in new clientele. Patterns were affected by filters for shipping mode, year, and region, which highlighted the importance of certain choices and regional preferences. This information offered insights into the dynamics of sales, highlighting patterns and the influence of factors on results. The graph's in-depth analysis clarified the impact of location and shipping mode on sales patterns.

**Profits by sub-category**

A graph of different colored bars

Description automatically generated with medium confidence

This Tableau visualization provides an analytical analysis of the superstore's inventory profitability across multiple subcategories. Initially, a visually appealing bar chart is used to present a thorough picture of profitability, successfully highlighting both high-performing and underperforming regions. By data sorting, it allows for more targeted research of subcategories, emphasizing notable performances like "Copiers." This strategic analysis is critical in directing decisions targeted at increasing income and reducing losses in the superstore's operations. Notably, the data shows that tables are the least profitable product by a large percentage, recommending caution against refilling this item.

**Top 10 regional sales cities**

A screen shot of a graph

Description automatically generated

This Tableau visualization presents the top ten regional sales cities from the east, south, and west regions. Initially, the bar chart provides a comprehensive picture of sales data, with the west leading and the south behind all regions. Further data analysis narrows the focus to the top ten cities in each region, demonstrating differences in sales results. Filtering by city reveals differences within areas, insights on local sales trends and city-specific impacts. Notably, the west has the best-performing sales city, with continuous annual increase, whereas the east and south have different tendencies, reflecting dynamic market dynamics within each region.

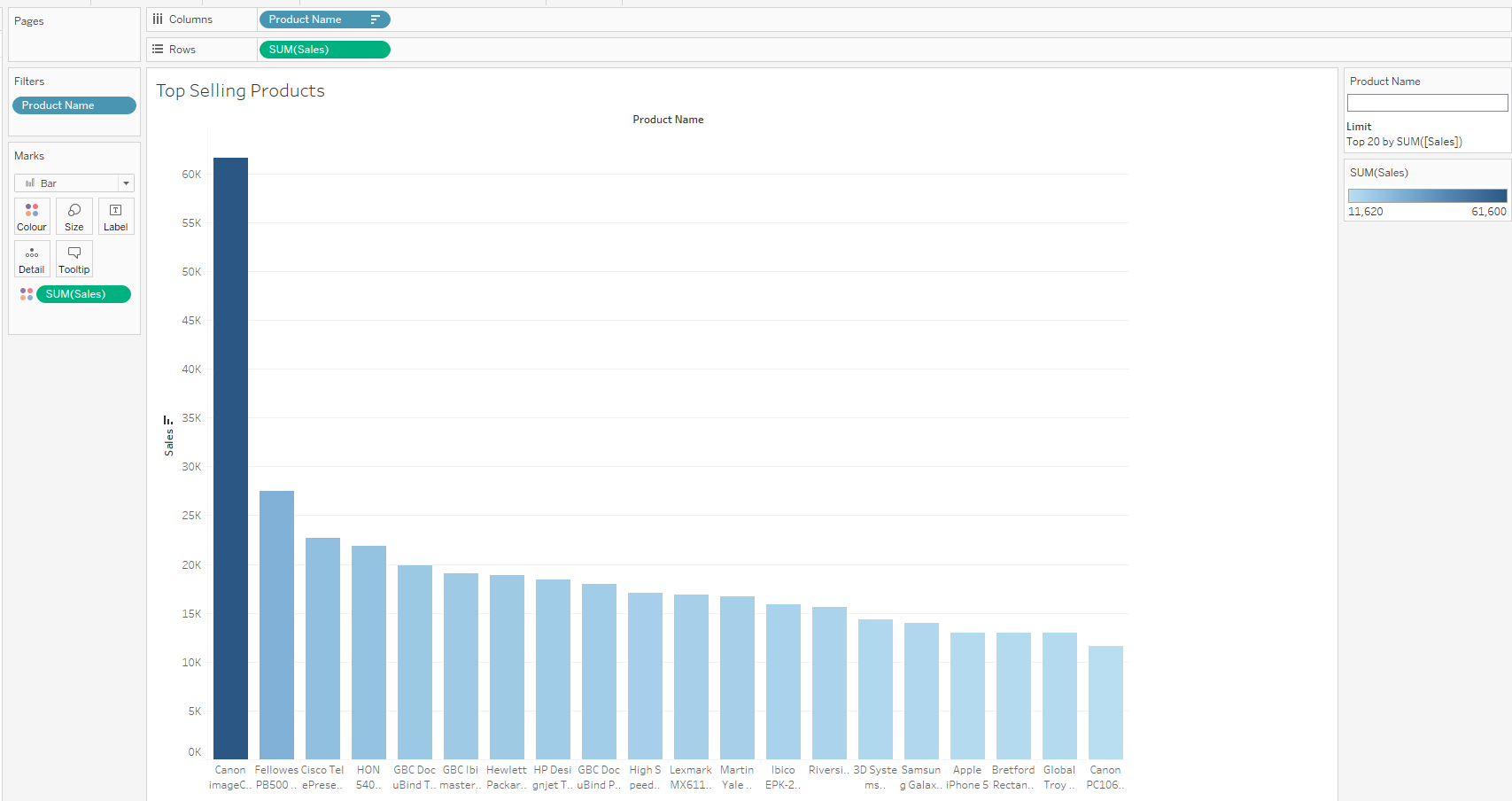
**Customer names by segment and sales**

A screenshot of a graph

Description automatically generated

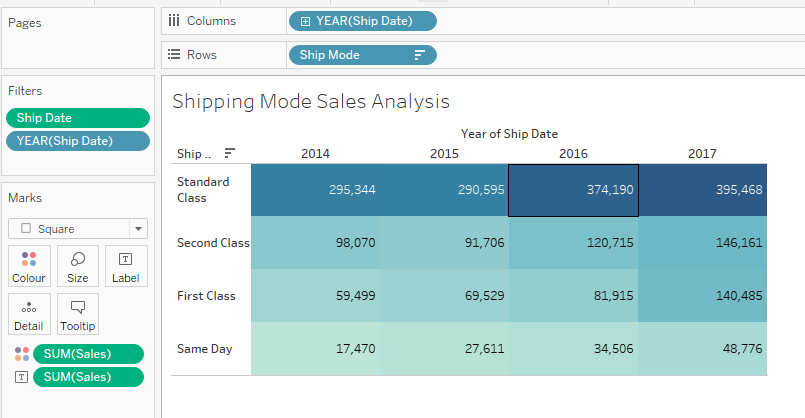
This Tableau visualization presents a study of the consumer, business, and home office client sectors, as well as their sales contributions. Initially, a statistical graph reveals the top ten clients, six of them are from the consumer sector, with Seth emerging as the key spender. This observation points to the potential effectiveness of using targeted incentives to retain high-spending consumers like Seth. Furthermore, the presence of sales, including two corporate and two home office identities, demonstrates the Superstore's diversified strategy. Overall, these visualizations provide useful insights into various consumer categories, driving the development of tailored strategies for specific audiences and optimizing sales results.

**Top Selling Products**



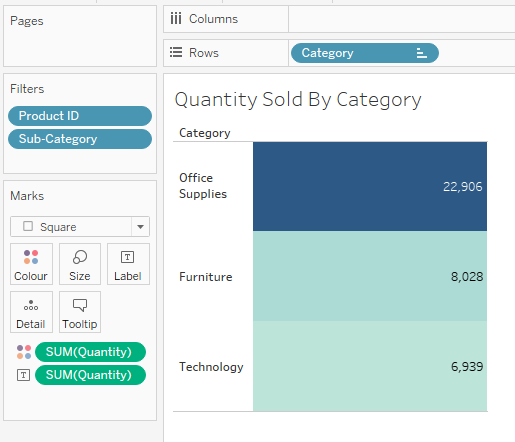
The graph highlights the top 20 selling products, with each product's name, sorted order, and associated sales revenue. Notably, the Canon image Class 2200 Advanced Copier is the best-selling product, with sales more than double those of the second-placed product. This significant margin emphasizes the need to prioritize this high-performing commodity, which constantly earns a sizable amount of the superstore's revenue.

**Shipping mode yearly sales**



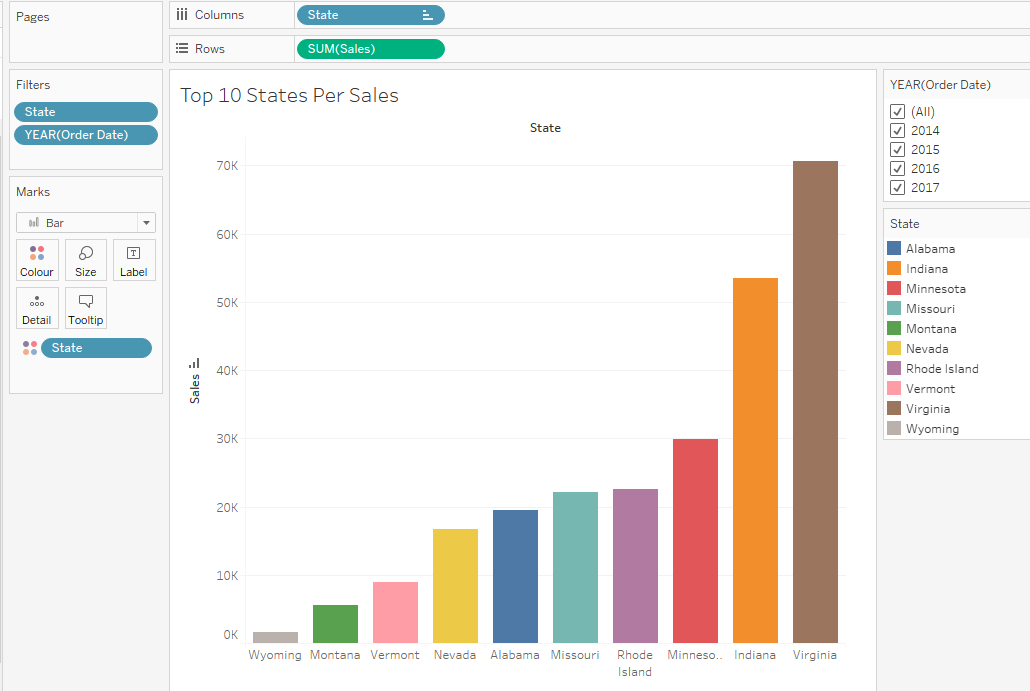
The information graphics present a complete analysis of shipping mode sales on an annual basis, outlining consumer preferences for various delivery methods. Across all years, standard class shipping has consistently been the greatest revenue generator, most likely due to its free or low-cost nature and slightly longer delivery schedule. In contrast, same-day delivery has the lowest annual sales, probably due to logistical challenges or greater expenses. The second-class option generates large sales by striking a balance between cost-effectiveness and prompt delivery, which aligns well with consumer expectations. Notably, first-class shipping experienced significant growth, more than doubling its sales between 2014 and 2017, indicating increased consumer usage. This growing trend implies great potential for first-class shipping as a favored mode in the future, as highlighted by its revenue growth over the years.

**Quantity Sold by Category**



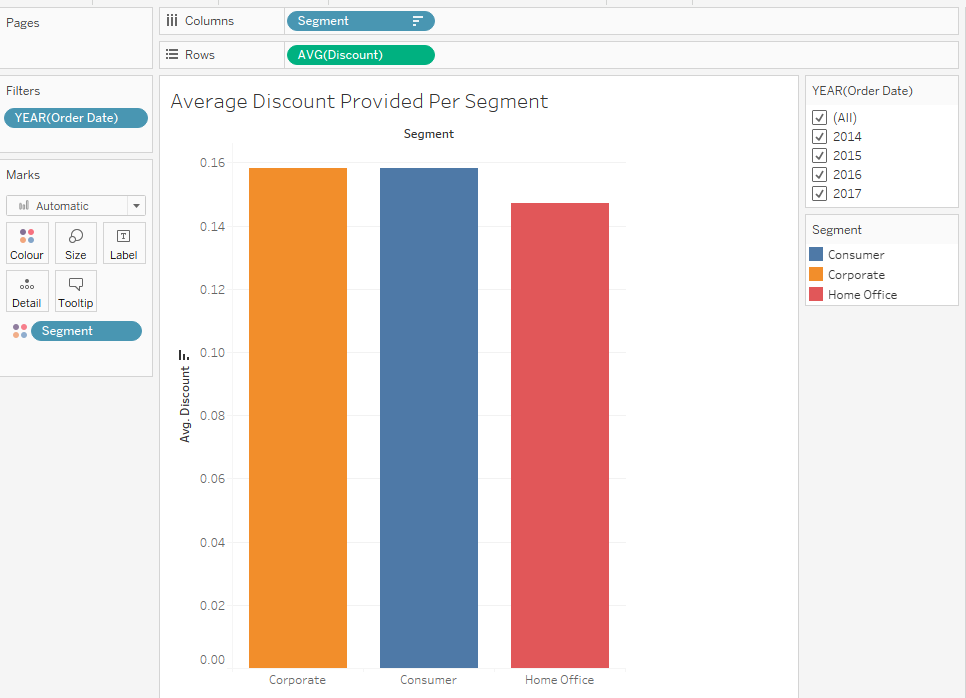
The office supplies category outsells the technology sector by a wide margin, indicating strong demand in this market segment. This insight gives light on the widespread demand for office supplies among numerous consumer categories, including consumer, corporate, and home office. Unlike technology, which may not appeal to all consumer segments, office supplies are necessary and frequently used in a variety of settings. The data highlights the office supplies category's dominance, which is underpinned by its global relevance and consistent demand among consumers and enterprises.

**Sales per Top 10 States**



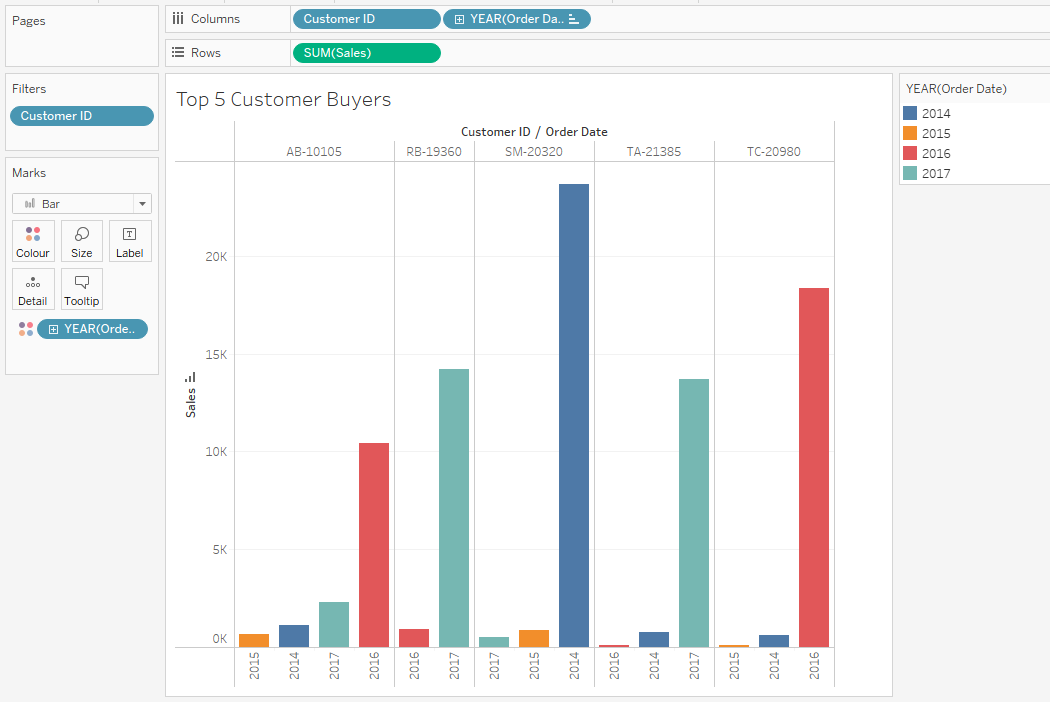
The visuals highlight the top ten states by sales in the United States, revealing a distinct pattern of progressively increasing sales until Minnesota. Notably, the increase in sales from Minnesota to Indiana has nearly doubled, highlighting the largest client base. Virginia emerges as the state with the most sales, demonstrating significant economic activity, while Wyoming stands out as the least in the top ten, with sales in the low thousands. This visual representation sheds light on regional sales trends, emphasizing the differences in market significance among states.

**Average Discount Per Segment**



The data shows the average discount per segment offered to customers when they make purchases from the superstore. Corporate enterprises and consumers both receive about 16% of the total discount, demonstrating a balanced allocation between both categories. In contrast, the home office group obtains the lowest average discount, around 14%, over all four years of available data. This difference highlights the discount rates across segments, with home Office clients receiving less favorable discounts than their counterparts in the Corporate and Consumer segments possibly to them being greater spenders and have been rewarded.

**Top 5 Customer Buyers**



The data exposes the top five superstore customers, with individuals whose customer ID begins with "SM" leading in overall sales, notably in 2014, with a staggering $25,000. However, consumer performance in 2015 was muted, with "SM" remaining at the top of the list but with just over a thousand purchases. The next year saw a huge increase, with "TC" spending $18,000 and "AB" spending more than $10,000. In the final year of the dataset, there was a tight margin, with user "RB" outspending user "TA" by only a thousand dollars, indicating the superstore's growth trajectory, in its most successful sales year when all aggregated results were considered.

**Selection of Data Mining Algorithm and Data Pre-Processing**

The chosen algorithm is K-means that unsupervised clustering method that partitions data into k clusters based on similarity.  It starts by randomly selecting k centroids, which represent the initial cluster centers. After the assignment step, the centroids are recalculated as the meaning of all data points assigned to each cluster.  This process shows until the centroids no longer change significantly, or a specified number of iterations is reached.

Why k-means is suitable for this dataset:

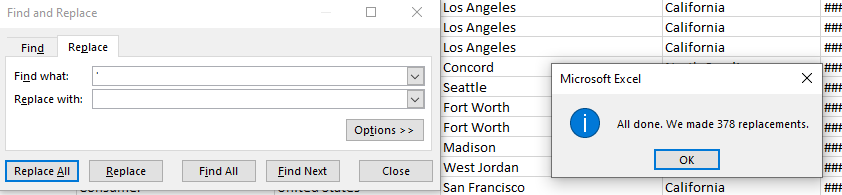
* K-means is suitable for this dataset as it aims to identify natural groupings or clusters within the data based on similarity.
* Since the dataset contains information about sales, profits, and various attributes of products and customers, k-means can help identify clusters of products, customers, or regions with similar characteristics or purchasing behaviours.

Variables being used in the algorithm:

The main variables used in the k-means algorithm for this dataset include:

* Sales
* Profit
* Quantity
* Discount
* Product Category
* Customer Segment
* Region

Dealing with anomalies:

* Commas from names being removed so that the dataset can be.
* Anomalies could be dealt with by either removing them if they are data errors or by treating them separately if they represent valid but rare cases.

 Other changes to the dataset:

One transformation that could be carried out on the dataset is feature scaling, especially for variables with different scales, such as sales and quantity. Standardizing or normalizing these variables can ensure that they have a comparable influence on the clustering process.

**Data Mining**

1st iteration

1. Data Input and Configuration:

* The SimpleKMeans algorithm was applied with specific configurations.
* The dataset used was from the Superstore, with 9994 instances and 11 attributes.

2. Clustering Model (Full Training Set):

* The algorithm was trained on the full dataset.
* Two initial starting points (random) were chosen for the clusters.
* The within-cluster sum of squared errors was calculated, resulting in a value of approximately 59555.42.

A close-up of a text

Description automatically generated

3. Model and Evaluation on Test Split:

* + The trained model was evaluated on a test split of the dataset (66% training, 34% test).
  + Two initial starting points (random) were chosen for the clusters in the test split.
  + The within-cluster sum of squared errors for the test split was calculated, resulting in a value of approximately 39365.89.

A close-up of a computer screen

Description automatically generated

4. Clustered Instances:

* The instances in the test split were clustered into two groups, with Cluster 0 containing 48% of the instances and Cluster 1 containing 52%.

A white background with black text

Description automatically generated

2nd Iteration

1. Removal of Attributes:

* + Three attributes, namely ROW ID, ORDER DATE, and CUSTOMER NAME, were removed from the dataset. These attributes were considered irrelevant for clustering purposes and were removed to simplify the dataset and potentially improve clustering performance.

2. Reconfiguration of Algorithm:

* + The SimpleKMeans algorithm was applied again with the modified dataset, now containing 8 attributes instead of the original 11.

3. Clustering Model (Full Training Set):

* + The algorithm was trained on the full modified dataset.
  + The number of iterations was reduced to 3 compared to the first iteration.
  + The within-cluster sum of squared errors was calculated, resulting in a value of approximately 38915.46.

A close-up of a computer screen

Description automatically generated

4. Model and Evaluation on Test Split:

* + The trained model was evaluated on a test split of the modified dataset.
  + Two initial starting points (random) were chosen for the clusters in the test split.
  + Missing values were globally replaced with mean/mode.
  + The final cluster centroids for the test split were computed.
  + The within-cluster sum of squared errors for the test split was calculated, resulting in a value of approximately 25744.49.

A screenshot of a computer

Description automatically generated

5. Clustered Instance

* The instances in the test split were clustered into two groups, with Cluster 0 containing 48% of the instances and Cluster 1 containing 52%.

A white background with black text

Description automatically generated

6. Interpretation:

* + The removal of irrelevant attributes led to a reduction in the within-cluster sum of squared errors, indicating an improved clustering performance.
  + The clusters were evaluated for their cohesion and interpretability, considering the attributes Ship Mode, Country, State, Region, Category, Product Name, Quantity, and Profit.

3rd Iteration

1. Data Input and Configuration:

* + The SimpleKMeans algorithm was applied with the same configurations as the previous iterations.
  + The dataset used was from the Superstore, with 9994 instances and 6 attributes after removing "Product Name" and "Quantity".

2. Clustering Model (Full Training Set):

* + The algorithm was trained on the fully modified dataset.
  + Two initial starting points (random) were chosen for the clusters.
  + The within-cluster sum of squared errors was calculated, resulting in a value of approximately 29,092.0.

A computer error message

Description automatically generated

3. Model and Evaluation on Test Split:

* + The trained model was evaluated on a test split of the modified dataset.
  + Two initial starting points (random) were chosen for the clusters in the test split.
  + The within-cluster sum of squared errors for the test split was calculated, resulting in a value of approximately 19,022.0.

A screenshot of a computer program

Description automatically generated

4. Clustered Instances:

* + The instances in the test split were clustered into two groups, with Cluster 0 containing 68% of the instances and Cluster 1 containing 32%.

A screenshot of a computer program

Description automatically generated

5. Interpretation:

* + The removal of additional attributes ("Product Name" and "Quantity") further simplified the dataset and potentially improved clustering performance.
  + The clusters were evaluated for their cohesion and interpretability, considering the attributes Ship Mode, Country, State, Region, and Category, along with Profit.

Further discussion on best Iteration

Based on the trend of decreasing WCSS values and the improved interpretability of clusters, the third iteration appears to be the best one. It achieved the lowest WCSS values on both the full training set and the test split, indicating more compact and well-separated clusters. Additionally, removing more irrelevant variables led to clearer distinctions between clusters, enhancing the overall quality of the clustering results.

* + The WCSS measures the compactness of clusters, with smaller values indicating tighter clusters and better separation between clusters.
  + In the best iteration (third iteration), the WCSS for the full training set is 29,092.0, and for the test split, it is 19,022.0.
  + These WCSS values are lower than those of the previous iterations, indicating improved clustering performance and better fit of the model to the data.

2. Comparison of Percentage in Each Cluster:

* + This comparison assesses the consistency of cluster assignments between the training set and the test set.
  + In the best iteration, Cluster 0 comprises 68% of the instances in the test split, while Cluster 1 comprises 32%.
  + This distribution is consistent with the percentages observed in the full training set (Cluster 0: 66%, Cluster 1: 34%), indicating stability in cluster assignments across different subsets of data.

Overall Fit of the Model:

* + The best iteration demonstrates improved clustering performance, as evidenced by lower WCSS values compared to previous iterations.
  + These findings indicate a good fit of the K-means clustering model to the data, with clear and well-separated clusters that are consistent across different subsets of the dataset.

**Data Ethics**

Ethical considerations for superstore datasets are critical to ensuring that the data is handled and analyzed responsibly. To begin, when analyzing sales data from the superstore, sensitive client data must be anonymized and protected to ensure their privacy rights are upheld. Transparency regarding how customer data is used and ensuring that it is only used for an agreed-upon objectives and is critical to preserving consumer trust. From a legal standpoint, the superstore must follow data protection regulations such as GDPR, particularly when dealing with customer information. Compliance with these standards is required to avoid legal ramifications and defend the rights of the people whose data is being analyzed. Professionally, analysts using the superstore dataset must ensure that their procedures are transparent and unbiased. Any conclusions drawn from the data should be reliable and free of bias. Furthermore, the data should be handled ethically, benefiting both the company and its customers while avoiding harm or unfair treatment. Ethical data analysis supports justice, openness, and accountability, which helps to create a favorable business environment and fosters trust among stakeholders. Overall, analyzing the superstore dataset with a significant emphasis on ethics, legality, and professionalism is critical for preserving trust, compliance, and integrity in data.

**Conclusion**

Finally, the study of the superstore dataset provided useful insights into sales patterns, consumer behaviour, and operational performance. Data visualisation and mining approaches have shown some meaningful results and patterns.

The visualisations gave a full overview of long-term sales patterns, profitability by product sub-category, regional sales performance, and client segmentation. Notable findings include changing sales trends over time, the dominance of specific product categories in driving profits, and variations in sales performance and consumer demographics.

The K-means clustering algorithm showed a unique cluster in the dataset, allowing for the identification of meaningful patterns and groupings. The model suited the data well, with distinct clusters and consistent assignments across subgroups.

Managers and stakeholders can use the analytical findings to support strategic decision-making and drive corporate success. Understanding sales trends, for example, can aid in inventory management and resource allocation, whilst knowledge of customer segmentation can help drive focused marketing campaigns.

Overall, the combination of data visualization and mining techniques has provided actionable intelligence that can be used to optimize operations, improve customer engagement, and drive competitive advantage in the retail industry. Businesses can use data analytics to explore new opportunities and efficiently manage problems in an increasingly complicated environment.