Report

Business Need and Importance

[US retailers confirmed another 169 closures last week - bringing the total so far this year up to almost 2,600](https://www.dailymail.co.uk/yourmoney/consumer/article-13367703/store-closures-walmart-foxtrot-urban-outfitters.html). In today's data-driven business landscape, understanding customer behaviour and market trends is paramount for strategic decision-making. Extracting insights from vast datasets is central to this endeavour. The provided code snippets showcase a systematic approach to analysing sales data, beginning with data loading, cleaning, and summarization. By examining key metrics such as sales revenue over time and clustering geographical sales patterns, businesses can discern actionable intelligence. Moreover, employing predictive modelling techniques like time series forecasting enables organizations to anticipate future sales trends, empowering them to allocate resources efficiently and optimize inventory management. This comprehensive data analysis pipeline not only enhances operational efficiency but also facilitates informed business strategies, ultimately fostering competitiveness and sustained growth in a dynamic marketplace. With the ability to adapt quickly to changing market conditions, businesses can stay ahead of the curve and capitalize on emerging opportunities, ensuring long-term success and resilience.

Statistical Methodology

The dataset used for analysis comprises several variables critical to understanding sales transactions. These include Order.Date and Ship.Date, which are essential for tracking the timing of orders and shipments, respectively. The Sales variable quantifies the monetary value of each sale, serving as a key metric for analysis. Geographic information is represented by City and State, providing insights into regional sales patterns. Additionally, the Postal.Code variable, utilized in the analysis, is included to facilitate geographical clustering. Other variables such as Segment, categorizing customers into different segments, and Product.Category and Product.Sub.Category, detailing product classifications, are vital for segmenting and analysing sales data effectively. These variables collectively offer a comprehensive view of sales activities, facilitating in-depth exploratory analysis and forecasting.

Preprocessing steps are crucial to ensure the quality and integrity of the data before analysis. In this dataset, several preprocessing steps were undertaken. Firstly, missing values were identified and addressed. For instance, missing postal codes were imputed to maintain the completeness of the dataset. Secondly, date variables (Order.Date and Ship.Date) were converted from character to date format using the as.Date function, enabling temporal analysis. Moreover, numeric variables were scaled to ensure uniformity in their ranges, a crucial step for certain clustering algorithms. Additionally, irrelevant variables or those not pertinent to the analysis were removed to streamline the dataset. These preprocessing steps are fundamental for preparing the data for subsequent analysis, ensuring accurate and reliable results.

For time series analysis, I aggregated the data into quarterly revenue values using the lubridate and dplyr packages. This resulted in a time series object covering the years 2015 to 2018. I then applied time series cross-validation by partitioning the data into training and validation sets. Next, I fitted linear and quadratic models to the training data using the forecast package's tslm function, capturing trends and seasonal patterns. Subsequently, I generated forecasts for the validation set and evaluated model accuracy using performance measures.

On the other hand, I conducted hierarchical clustering to identify geographical sales patterns based on postal codes. The analysis involved grouping sales data by postal code and calculating total sales for each postal code using the dplyr package. I then applied hierarchical clustering using the agnes function from the cluster package, employing the Ward method for merging clusters. The resulting dendrogram allowed me to cut the tree into five clusters, representing distinct geographical regions with similar sales patterns. I computed summary statistics for each cluster to analyse the distribution of observations across the clusters. Overall, these methods provided valuable insights into sales trends and geographical patterns, enabling informed decision-making in business strategy and resource allocation.

Results and Interpretation

Hierarchial Clustering Analysis

In the hierarchical clustering analysis, I first grouped the dataset by postal codes and calculated the total sales for each postal code. This allowed me to aggregate sales data at a regional level. Next, I scaled the numerical columns and computed the Euclidean distance to measure similarity between postal codes based on their total sales.

Using the agnes function with the Ward method, I performed agglomerative clustering to identify clusters of postal codes with similar sales patterns. The resulting dendrogram visually depicted the hierarchical relationships between the postal codes, revealing distinct clusters.

A barcode on a white background

Description automatically generated

The agglomerative coefficient, a measure of clustering strength, was found to be 0.9979445, indicating robust clustering tendencies within the dataset.

A black text with black text

Description automatically generated with medium confidenceThe statistics provided under "Height (summary)" offer valuable insights into the hierarchical clustering process applied to the data. The minimum height of 0.000699 indicates that some clusters are extremely close together, suggesting the presence of tightly grouped data points. Conversely, the maximum height of 28.753787 suggests that there are significant distances between certain clusters, possibly indicating outliers or distinct groupings. The median height (0.055861) and mean height (0.401908) provide central tendency measures, offering an understanding of the typical distances between clusters. Additionally, the first quartile (0.021867) and third quartile (0.165141) show the spread of distances, highlighting how clustered or dispersed the data points are within the hierarchical structure. These statistics collectively aid in interpreting the clustering results and identifying meaningful patterns within the data.

Cluster 1 has a mean sales value of $2227.76, with 94 observations. Cluster 2 shows a higher mean sales value of $46626, but with only 13 observations. Cluster 3 follows with a mean sales value of $17499 and 39 observations. In contrast, Cluster 4 has a lower mean sales value of $1681.88 but a larger number of observations at 242. Finally, Cluster 5 has a mean sales value of $1491.72 with 239 observations.

Forecasting using cross validation and logistic regression

In this analysis, I started by loading the dataset and summarizing its structure. Then, I extracted the year and quarter from the Order Date to create a time series dataset. Next, I split the data into training and validation sets, and I estimated two different linear models using the training data. After making forecasts for the validation set using both models, I evaluated their performance using accuracy measures such as Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE). Finally, I estimated a final model using the entire dataset and generated forecasts for the next four quarters.

A close up of numbers

Description automatically generated

When comparing the performance of two forecasting models, fReg1 and fReg2, interesting insights emerge. Initially, fReg1 appears promising, showcasing minimal error on the training set with a MAPE of 3.6% and an RMSE of 3863.065. However, its performance drastically deteriorates when tested on unseen data, displaying a much higher MAPE of 35.5% and an RMSE of 64374.779. Conversely, fReg2 exhibits similar training set performance to fReg1 but outperforms it on the test set, boasting a MAPE of 22.4% and an RMSE of 40078.519. This discrepancy suggests that while both models excel during training, fReg2's more complex structure allows it to generalize better to new data. It implies that while fReg1 might overfit the training data, fReg2's sophistication facilitates improved forecasting accuracy on unseen data, underscoring the significance of model robustness in time series forecasting.

A screenshot of a computer

Description automatically generated  
This forecast, generated by the model RegFin, offers predictions for each quarter of the year 2019. In each row, you'll find the anticipated value for that quarter, labelled as the "Point Forecast." Additionally, the "Lo 80" and "Hi 80" columns outline the lower and upper limits of the 80% prediction interval, indicating where the actual value might lie with 80% certainty. Similarly, the "Lo 95" and "Hi 95" columns present the boundaries of the wider 95% prediction interval, offering a broader range within which the actual values are likely to fall with 95% confidence. These intervals provide insights into the potential variability and uncertainty surrounding the forecasted values, aiding in decision-making and risk assessment.

Alternative Approaches

The methods I implemented, including time series cross-validation with linear regression and hierarchical clustering, were chosen for their ability to provide robust and interpretable insights into sales data. Time series cross-validation allowed for accurate forecasting by capturing temporal patterns, while linear regression models provided a clear understanding of the relationship between variables. Hierarchical clustering facilitated the identification of geographic sales patterns, enabling regional segmentation for targeted marketing strategies. These methods offer a comprehensive approach to analysing sales data, ensuring that insights gained are both reliable and actionable.

Conclusion

In today's dynamic retail landscape, where closures continue to rise, leveraging data analytics is imperative for success. Our comprehensive analysis offers valuable insights into customer behaviour, market trends, and regional sales patterns, empowering businesses to make informed decisions. By understanding past trends and anticipating future developments, organizations can optimize their strategies, allocate resources effectively, and drive sustainable growth. The methodologies employed, including time series forecasting and hierarchical clustering, provide reliable and actionable insights that enable businesses to stay ahead of the curve and thrive in an ever-evolving market environment. With the ability to adapt quickly and capitalize on emerging opportunities, businesses can navigate challenges with confidence and build a resilient foundation for long-term success. This strategic approach not only enhances operational efficiency but also fosters competitiveness, ensuring businesses remain agile and resilient in the face of uncertainty. By harnessing the power of data analytics, businesses can unlock new opportunities, mitigate risks, and position themselves for sustained growth and prosperity.