

DAB422 CAPSTONE PROJECT REPORT

Multi Mart: Loyalty Card & Total Purchases Prediction

**Prepared By,
Group 7**

Ikram Patel	0822315
Sujata Biswas	0832706
Alisha James	0811919
Gayathri Manju Jayasena Kurup	0836679
Srikanth Ayyalasomayajula	0808545

**Under the Guidance of,
Prof. Sodiq Shofoluwe**

ACKNOWLEDGMENTS

We would like to express our heartfelt gratitude to Prof. Sodiq Shofoluwe for his guidance and mentorship throughout this project. His insightful feedback and unwavering support played a pivotal role in successful completion of this project.

We would also like to thank St. Clair College for providing us with an environment to utilise and apply our knowledge and skills we gained through this course.

TABLE OF CONTENTS

1.	Abstract
2.	Introduction
3.	Data Acquisition
4.	Data Preprocessing and Cleaning
5.	Exploratory Data Analysis
6.	Data Visualisation
7.	Preprocessing For ML Model
8.	Prediction Using Machine Learning Model
9.	Final Product: Web Interface
10.	Conclusion
11.	References

ABSTRACT

This report presents a comprehensive analysis of Multi Mart Retail store's sales and revenue data from 2019 to 2023, focusing on two main objectives: revenue generation and the expansion of the Loyalty Card program. Utilizing data visualization and machine learning techniques, the study provides insights into the store's sales performance, revenue trends, and customer behavior.

By forecasting future sales trends and customer demands, the report empowers Multi Mart Retail to make data-driven decisions for optimizing revenue and guiding the expansion of its Loyalty Card program into new regions. The outcomes provide a clear roadmap for enhancing customer satisfaction and achieving long-term business success in the retail industry.

INTRODUCTION

The retail industry is a dynamic landscape that requires strategic insights to maintain a competitive edge. As customer expectations evolve and market trends shift, businesses must adapt quickly to remain relevant and profitable. This project is centered around providing a comprehensive and visually intuitive view of the sales and revenue generated at Multi Mart Retail store from 2019 to 2023. Through meticulous data analysis, this report sheds light on the store's sales performance and revenue trends over the years, offering valuable perspectives on the store's trajectory and its areas of strength and potential improvement.

By examining customer behavior and preferences, this project aims to guide the store in making informed decisions about potential expansion opportunities, particularly regarding the Loyalty Card program. Understanding customer purchasing patterns and preferences enables the store to tailor its offerings and promotions, thus enhancing customer satisfaction and loyalty. The project leverages advanced data visualization techniques to present complex data in a clear and accessible manner, making it easier for stakeholders to interpret trends and patterns. Furthermore, the use of machine learning predictions allows the project to forecast future sales and customer trends.

The outcomes of this project not only present a clear picture of the store's past and present performance but also pave the way for future growth and development in alignment with customer needs and business objectives. By capitalizing on data-driven insights, Multi Mart Retail can enhance its operational efficiency, optimize its product offerings, and strengthen its market position. Ultimately, this project serves as a roadmap for the store's journey towards sustained success and continued innovation in the retail industry.

DATA ACQUISITION

For this project, the data was sourced from the British Library, providing a dataset rich in customer-centric variables. The dataset includes key metrics such as purchase history, revenue, churn indicators, product categories, campaign responses, and loyalty card participation. These data points offer a detailed understanding of customer behavior and preferences, allowing for a comprehensive analysis of Multi Mart Retail store's sales and revenue performance.

Below is the data set description:

Variable	Description
CustomerID	Unique customer ID
first purchase date	The "First Purchase Date" refers to the date when a customer or user made their initial purchase or transaction with the organization.
last purchase date	The "Last Purchase Date" refers to the date when a customer or user made their most recent purchase or transaction with the organization.
total purchases	"Total Purchases" is the count or sum of all purchases made by a customer or user with the organization. It represents the aggregate number of transactions.
total revenue	"Total Revenue" is the sum of all revenue generated from customer or user transactions with the organization. It represents the aggregate monetary value of all transactions.
referral source	"Referral Source" refers to the origin or channel through which a customer or user was referred to your organization or website. It provides information about how individuals found out about your products or services.
churn indicator	The "Churn Indicator" is a binary flag that indicates whether a customer or user has churned (i.e., stopped using your products or services) or is still an active customer. Typically, a value of 1 or "Yes" is used to indicate churn, while 0 or "No" is used to indicate an active customer.
discount used	"Discount Used" indicates whether a customer or user has applied a discount or promotional code during a transaction. It provides information about whether a discount was utilized for a specific purchase or order.
product category	"Product Category" classifies products into specific categories or groups based on their characteristics or purpose. It helps organize and categorize products for various purposes, such as reporting and analysis.
responsetolastcampaign	"Response to Last Campaign" indicates whether a customer or user responded to the most recent marketing campaign. It provides information about whether the individual engaged with the campaign in some way.

feedbackscore	"Feedback Score" represents a numeric score or rating provided by customers or users as feedback for a product, service, or experience. It is often used to gauge satisfaction or quality.
preferredpaymentmethod	"Preferred Payment Method" indicates the payment method that a customer or user prefers to use for transactions. It provides information about the customer's preferred way to make payments.
supportticketsraised	"Support Tickets Raised" represents the number of customer or user support tickets that have been opened or raised by individuals seeking assistance, reporting issues, or making inquiries.
hasloyaltycard	"Has Loyalty Card" is a binary indicator that shows whether a customer or user possesses a loyalty card or membership with your organization. It helps identify individuals who are part of a loyalty program.
frequency	"Frequency of Customers" represents how often a customer or user interacts with your organization, such as making purchases, engaging with your services, or participating in activities. It is a measure of how frequently individuals interact with your business. The frequency column is based on the first purchase date and the last purchase date period. It shows how frequently the customer has purchased during this period.

Below is the sample dataset. The granularity of the data is at customer level, each row refers to an individual customer.

customerid	firstpurchase date	lastpurchase date	totalpurchases	totalrevenue	referralsource	churnindicator	discountsused	productcategory	responsetolastcampaign	feedbackscore	preferredpaymentmethod	supportticketsraised	hasloyaltycard
8519	2021-12-31	2022-03-06	7	11670	Online advertisements	0	2	Q02	ignored	4.729998311	debit card	0	no
38152	2019-09-27	2023-02-02	20	5260	Traditional media outreach	1	6	F76	purchased	4.184511996	cash	0	no
19680	2021-06-13	2022-02-04	29	9790	Influencer endorsements	0	2	X04	opened mail	4.346639694	google pay	0	no
35744	2021-07-28	2022-08-21	15	9591	Influencer endorsements	0	5	A25	ignored	5	debit card	0	no
11663	2021-01-19	2022-03-10	13	10134	Word of mouth	0	3	A16	ignored	4.482089345	credit card	0	no
23498	2021-05-31	2022-02-03	8	10665	Word of mouth	1	2	O25	ignored	5	credit card	0	no
22735	2021-09-21	2021-12-14	29	4866	Word of mouth	0	4	V08	ignored	1.961207901	credit card	0	yes
41296	2019-06-28	2023-07-31	25	8826	In-store promotions	0	2	S31	opened mail	5	cash	0	no
14351	2019-06-05	2019-06-13	7	9978	Influencer endorsements	0	4	Y30	ignored	5	debit card	0	no
22571	2020-05-15	2023-06-30	26	6085	Influencer endorsements	1	6	C86	ignored	5	apple pay	0	no

DATA PREPROCESSING AND CLEANING

Once the data was acquired, the next step was to clean and prepare it for analysis. Ensuring high data quality is essential for producing reliable analysis and visualizations. Data cleaning involves several tasks, including checking for null values, imputing missing values, checking for outliers, and ensuring columns are named correctly.

Our dataset was in good condition, with no missing values, nulls, or duplicate records. This clean state of the data streamlined our analysis and allowed us to focus on deriving actionable insights. In addition to verifying data quality, we also derived four additional columns to simplify the analysis and enhance our understanding of the data:

- **Rounded Feedback Score:** This column contains the feedback score rounded to the nearest whole number, providing a clearer view of customer satisfaction.
- **Products:** This column groups products based on the initial letter of their names as listed under the product categories column, facilitating a more straightforward analysis of different product types.
- **Recency:** This field measures the time since the customer's last purchase, helping us assess the level of customer engagement and predict future purchasing behavior.
- **Tenure:** This field calculates the duration of the customer's relationship with the store, offering insights into customer loyalty and the potential for long-term retention.

The dataset comprises 2 Date-Time, 6 Categorical, and 11 Numerical columns, providing a comprehensive overview of customer interactions and behaviors. Once the data cleaning and preparation were complete, the data was exported to an Excel file and then imported into Tableau for data visualization. This step allowed for clear, interactive visual representations of the data, further aiding in the analysis and decision-making process.

EXPLORATORY DATA ANALYSIS

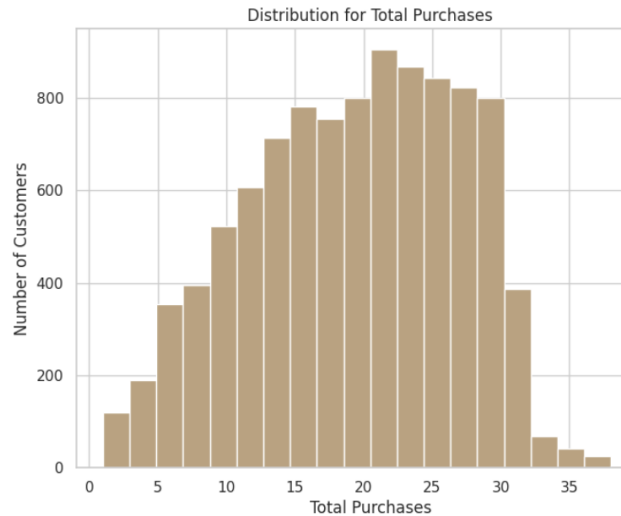
Next, we explored the data to gain insights into our dataset and what it contains. This includes, but is not limited to, checking for outliers or unusual data by looking at data distributions; and examining the relationship between the response and predictor variables using a correlation matrix.

To start with, we first looked at the basic descriptive statistics of all the numerical columns of our dataset to get a better idea about our dataset for identifying its competencies.

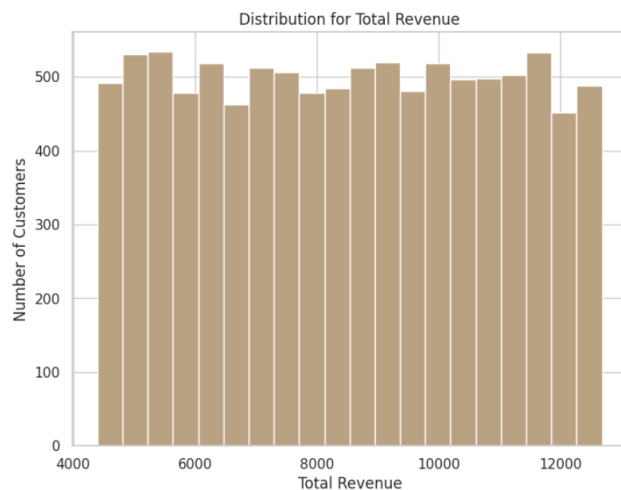
	customerid	totalpurchases	totalrevenue	churnindicator	discountsused	feedbackscore	frequency	rounded_feedbackscore
count	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	27519.237400	19.28050	8521.876100	0.498700	2.993600	4.433905	10.053700	4.438000
std	13118.347463	7.82962	2388.452322	0.500023	2.005283	0.920760	7.098652	0.963353
min	5000.000000	1.00000	4401.000000	0.000000	0.000000	1.006071	1.000000	1.000000
25%	16144.750000	13.00000	6427.750000	0.000000	1.000000	4.251318	4.000000	4.000000
50%	27617.000000	20.00000	8543.000000	0.000000	3.000000	4.985672	9.000000	5.000000
75%	38967.250000	26.00000	10589.250000	1.000000	5.000000	5.000000	14.000000	5.000000
max	49994.000000	40.00000	12678.000000	1.000000	6.000000	5.000000	31.000000	5.000000

- Based on the provided statistics, it's apparent that most columns exhibit no outliers, with only two exceptions.
- Among these exceptions, the columns identified with outliers are 'feedbackscore' and 'frequency'.
- In the case of 'feedbackscore', the lower bound of the Interquartile Range (IQR) was calculated as 2.5, while the minimum value in this column was found to be 1, indicating an outlier below the lower bound.
- Conversely, for the 'frequency' column, the upper bound of the IQR was determined to be 29, yet the maximum observed value was 31, signifying an outlier above the upper bound.
- However, it's noteworthy to acknowledge that while these observations may qualify as outliers statistically, they may not necessarily be considered outliers within the context of the dataset.

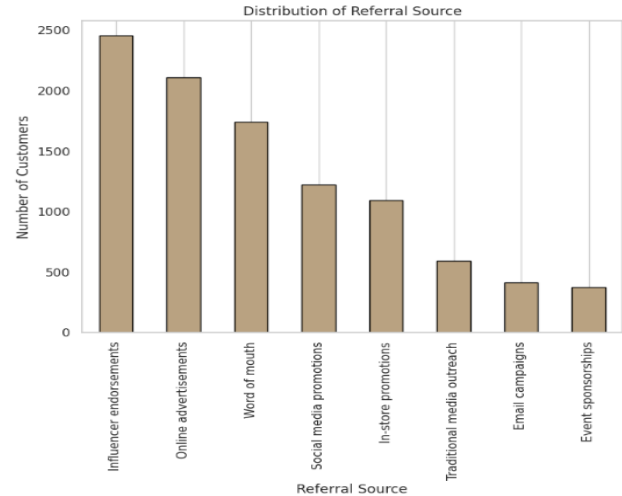
The plots generated provide a lot of information about the dataset. As can be seen below, plots provide effective visual summaries of data. This is especially useful for large datasets.



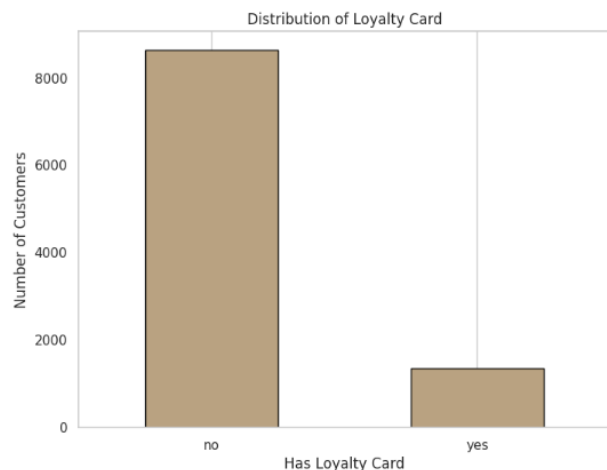
- Distribution of 'Total Purchase' is symmetric and unimodal.
- This implies a balanced and centered pattern in customer buying behavior.
- Most customers tend to make a consistent number of purchases, forming a bell-shaped curve around an average.



- 'Total Revenue' is uniformly distributed.
- This indicates an even and consistent spread of revenue across different customer segments or transactions.

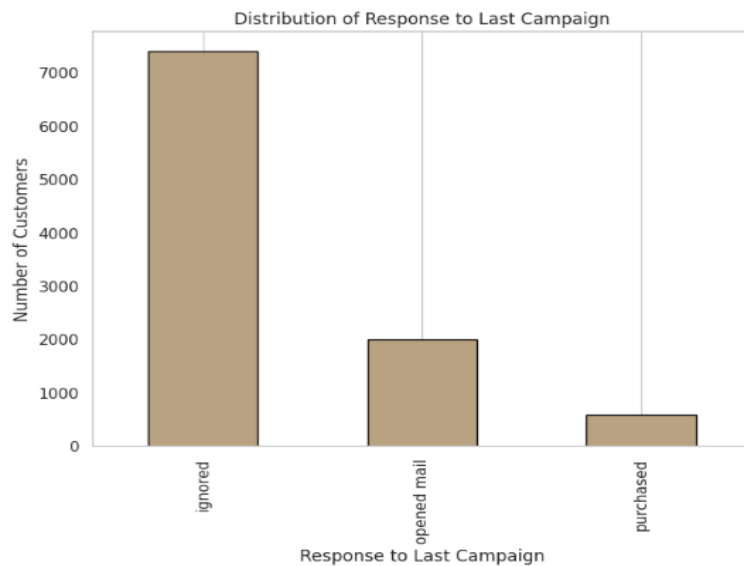


- 'Influencer endorsements' significantly outweigh other referral campaigns followed by online advertisements.
- This suggests a strong impact of influencer endorsements in driving customer interest or sales.
- Allocating resources or strategies to leverage this influential referral source might capitalize on its effectiveness in attracting customers.

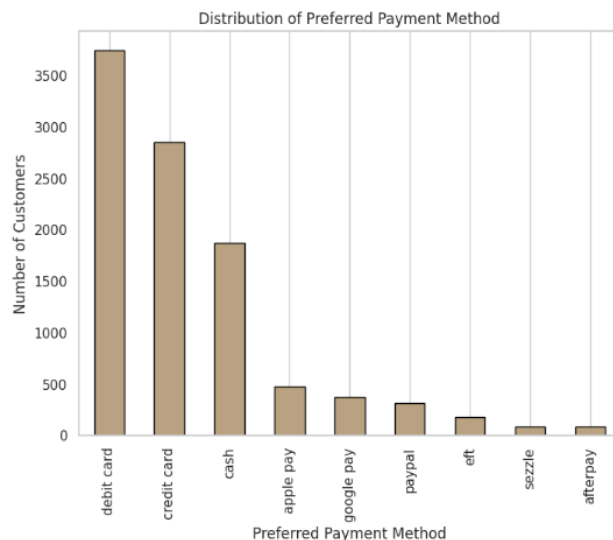


- In column 'has loyalty card', the dominating category is 'no' with 8632 occurrences.
- Above graph indicates that a significant majority of customers do not possess a loyalty card.

- Implementing strategies to encourage enrollment in loyalty programs could foster stronger customer retention and engagement.



- In column 'Response to Last Campaign', the dominating category is 'ignored' with 7401 occurrences.
- Understanding this dominant response aids in refining future campaign strategies to enhance engagement or address factors leading to disinterest.



- In column 'Preferred Payment Method', the dominating category is 'debit card' with 3746 occurrences.
- Above bar chart signifies a strong preference among customers for using debit cards and credit over other payment methods.

- This can be potentially due to its convenience, widespread acceptance, or associated perks like cashback or security features.



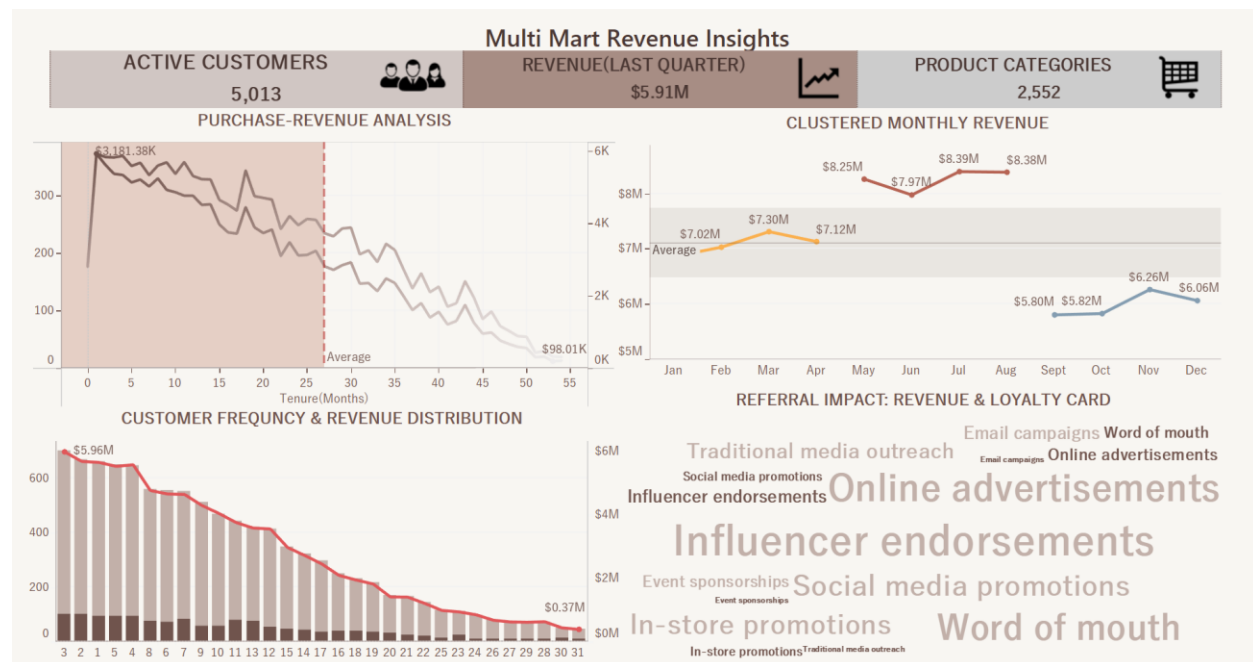
We used heat-map to explore the variables in the dataset and check if there is a meaningful correlation between them. A correlation value close to -1 indicates strong negative correlation, values close to 0 indicate no/very weak correlation, and values close to 1 indicate a strong positive correlation between variables. The variables total purchase and frequency has highest frequency of 0.49 which suggests that customers making more purchases tend to engage with higher frequency. All other variables do not exhibit significant strength to be highlighted. Further, in Machine learning modelling we will be converting few categorical columns to numerical columns so that those columns can also be used in the correlation matrix.

DATA VISUALISATION

We did analysis by performing data visualization using Tableau. This approach allowed us to create intuitive and insightful visual representations of the data, helping us to better understand the relationships and trends within the dataset.

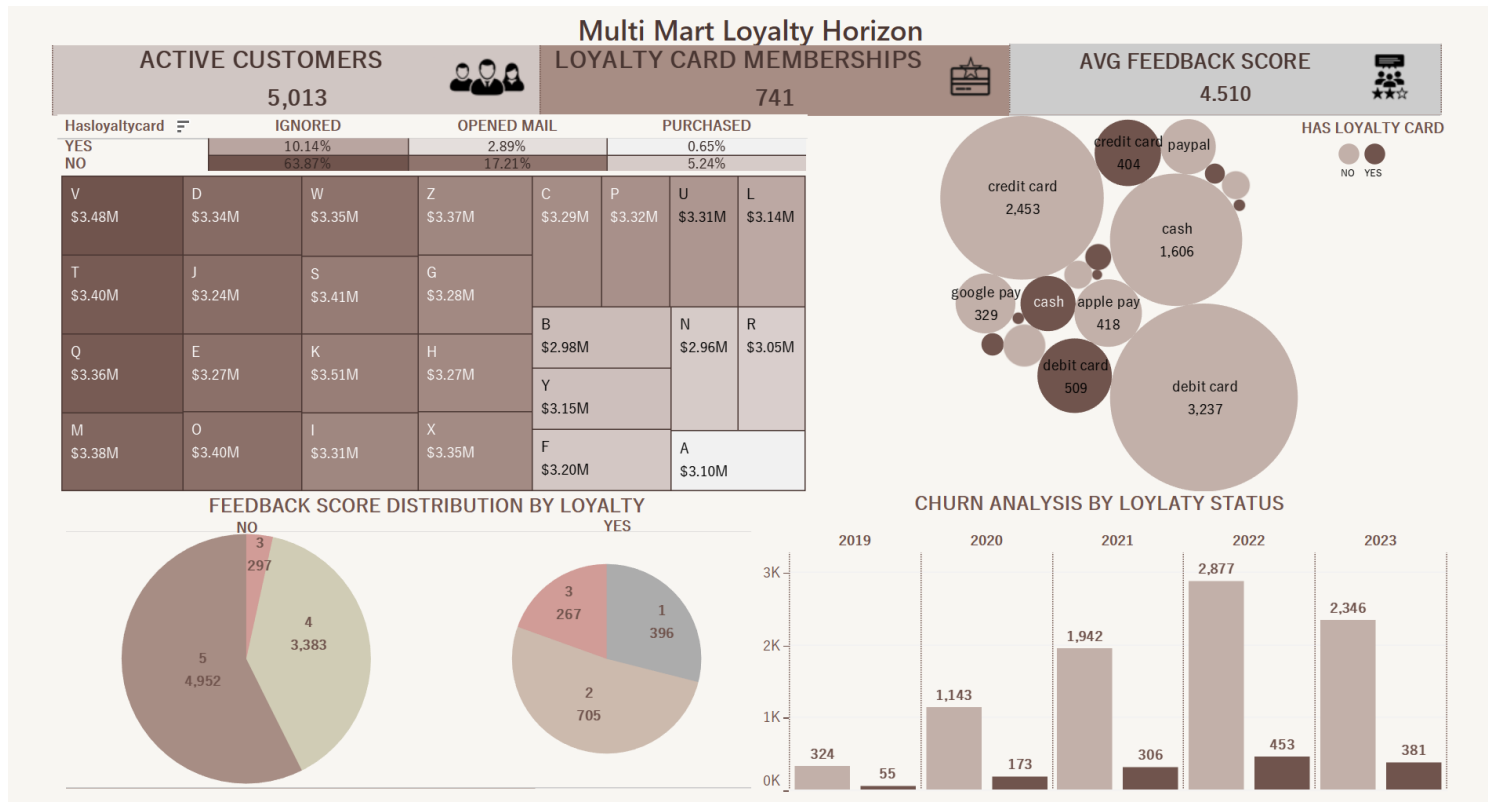
There are two dashboards we have designed to explain below objectives to the business.

1. Evaluation of store sales performance and revenue generation.



- The dashboard offers insights into the revenue trends of Multi Mart retail stores spanning four years from 2019 to 2023.
- Notably, revenue peaks during the months of May to September and dips towards the end of the year. This observation aids management in strategizing additional offers and promotions during the year-end period.
- Another noteworthy finding is that customers who visit the store less than 10 times contribute significantly to revenue generation. This suggests an opportunity for management to prioritize promoting loyalty cards to this segment of customers.
- Furthermore, the data indicates that influencer endorsements and online advertisements referrals are driving higher revenue compared to other channels. This insight underscores the potential effectiveness of these marketing strategies and warrants further investment.

2. Assessment of opportunities for expanding the Loyalty Card Program to New Regions.



- The Loyalty Horizon Dashboard gives us an insight on the performance of the store in the aspect of the Loyalty Card Program the store currently runs. It will help the business look at how the customers' purchase behaviors, their preferred payment methods, the feedback score the customers give and if the customer is churning or not will impact their Loyalty Card Program expansion plan.
- The dashboard above displays the key information of the store at the top being the number of active customers, the number of customers holding a loyalty card and the average feedback score that they rate the store with.
- The most prioritized information about how the customers are reacting to the loyalty card campaign tells the management that most of their efforts are not being paid back for as the crowd chooses to ignore the campaign suggesting that the business should implement a different way of campaigning so as to increase customer engagement.
- A number of offers can be provided on loyalty card holders on the products that are being bought the most by the customers including strategies like product bundling and clustering so as to attract customers towards buying more.
- Most of the customers prefer to pay using a debit or a credit card.
- Further we can also see that the customers holding a loyalty card do not provide a very good feedback score to the store suggesting improvement in that field as well.

PREPROCESSING FOR ML MODEL

As per the initial steps in machine learning, it is necessary to consider as many features as possible, so we converted the non-numerical columns to numeric using one hot encoding. This transformation allows for the inclusion of a broader range of data attributes in the model, ensuring a comprehensive analysis. Following the encoding process, we carefully selected features and target variables to build our machine learning model. To achieve this, we analyzed the correlation between various variables within the dataset to identify the most impactful features and determine an appropriate target variable for the predictive model.

Our analysis revealed a notable correlation of 0.49 between total purchase and frequency of engagement, suggesting a strong link between customer engagement and purchasing behavior. This finding highlights the importance of customer interaction in driving sales and offers insights for the store's strategic planning. Furthermore, the variable "hasloyaltycard" exhibited a high negative correlation with feedback score (-0.873). Other variables lacked in significant correlation strength.

Given these findings, we selected total purchase and hasloyaltycard as our target variables for further analysis. Our focus is on understanding the key factors influencing total purchase and the expansion of the loyalty card program to guide the store's future operations and decision-making.

To further investigate the factors impacting total store purchases, we utilized techniques such as feature importance analysis and Principal Component Analysis (PCA). These approaches enable us to uncover the key drivers behind purchasing behavior within the store environment. By meticulously exploring these underlying factors, we strive to develop a robust machine learning model capable of accurately predicting total store purchases and providing valuable insights for strategic decision-making.

Below is the feature importance captured after performing one hot encoding.

Feature Importance:		
	Feature	Importance
6	frequency	0.279176
8	tenure	0.247433
0	totalrevenue	0.122990
7	recency	0.121734
3	feedbackscore	0.064687
2	discountsused	0.047605
1	churnindicator	0.014716
13	referralsource_Online advertisements	0.009851
24	preferredpaymentmethod_debit card	0.009540
12	referralsource_Influencer endorsements	0.009152
14	referralsource_Social media promotions	0.007844
23	preferredpaymentmethod_credit card	0.007790
16	referralsource_Word of mouth	0.007579
9	referralsource_Email campaigns	0.005133
11	referralsource_In-store promotions	0.005074
18	responsetolastcampaign_opened mail	0.004770
21	preferredpaymentmethod_apple pay	0.004653
22	preferredpaymentmethod_cash	0.004437
17	responsetolastcampaign_ignored	0.004123
27	preferredpaymentmethod_paypal	0.003897
10	referralsource_Event sponsorships	0.003357
26	preferredpaymentmethod_google pay	0.003253
15	referralsource_Traditional media outreach	0.002332
4	supportticketsraised	0.002272
19	responsetolastcampaign_purchased	0.001949
20	preferredpaymentmethod_afterpay	0.001911
28	preferredpaymentmethod_sezzle	0.001377
25	preferredpaymentmethod_eft	0.001363
5	hasloyaltycard	0.000000

And below is the variance ratio for PCA, the explained variance ratio quantifies the contribution of each principal component to the overall variance in the dataset. For example, the first principal component (PC1) captures approximately 15.17% of the total variance, indicating that it explains a significant portion of the variability in the data.

By examining the explained variance ratio, we can identify the most influential principal components in terms of capturing variance. In this case, the first few principal components with higher ratios, such as PC1, PC2, and PC3, contribute more significantly to the total variance and are therefore more important for summarizing the dataset's variability. And the below 20 PC's constitutes to 100% which denotes that the model can be built on these 20 features.

```
Explained variance ratio: [0.15174215 0.10592869 0.08651564 0.06847016 0.05423612 0.04400922
0.04288627 0.03812719 0.0378965 0.03756142 0.03701504 0.03686911
0.03645765 0.03579143 0.03505019 0.03455506 0.03383383 0.0250289
0.01908267 0.01781942]
```

PREDICTION USING MACHINE LEARNING MODEL

Any machine learning model requires the dataset to be divided into training as well as testing data to train the model using one and then applying the model on the new test dataset to understand how well the model can capture variations in the data and based on the same how well can it predict the target using this new test dataset. For this reason, we divided the dataset into training and testing sets, with 80% of the data allocated for training the models and 20% reserved for evaluating model performance.

To predict total purchases, we leveraged supervised regressor models to harness the relationships between features and the target variable. We utilized the Lazy Predict library to efficiently identify the most suitable models for our dataset. This approach helped streamline the model selection process and provided insights into the performance of various algorithms. Based on the Lazy Predict results, we selected two models: Gradient Boosting and Random Forest Regressor. These models were trained on the training dataset, enabling them to learn complex patterns and relationships between the features and the target variable (total purchases). Their ensemble methods offer robust performance, leveraging the power of multiple decision trees for more accurate predictions.

To evaluate the performance of our models, we employed a range of evaluation metrics, including R2 score, Mean Absolute Error (MAE), and Mean Squared Error (MSE). These metrics were calculated using the test set to measure the models' ability to generalize to unseen data. The metrics for both the models are as follows:

```
OOB using Random Forest model: 0.3120450374474153
R-squared: 0.3089101013928247
Mean Squared Error (MSE): 42.42016070000001
Root Mean Squared Error (RMSE): 6.513076131905723
Mean Absolute Error (MAE): 5.04474
```

```
R-squared value using Gradient Boosting Regressor model: 0.34894021660993824
Mean Squared Error (MSE): 39.9630506716639
Root Mean Squared Error (RMSE): 6.321633544556652
Mean Absolute Error (MAE): 5.004418481531846
```

Among the two selected models, Gradient Boosting demonstrated superior performance based on evaluation metrics such as R2 score, MSE, and MAE. However, the model's initial results were not optimal. Therefore, we proceeded with hyperparameter tuning using Grid Search to refine the model and enhance its predictive capabilities. The Grid Search approach systematically explores a range of parameter combinations for the Gradient Boosting model, aiming to identify the optimal set of parameters that maximizes performance and minimizes error. This process helps fine-tune the model, allowing it to better capture the complex relationships within the data and improve its overall accuracy and reliability.

```
Best parameters found: {'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 100}
R-squared value using the best estimator from GridSearchCV: 0.3539284297382498
```


Even after conducting hyperparameter tuning for the Gradient Boosting model, the improvement in the R2 score was minimal. The score only increased marginally from 0.34 to 0.35, indicating that the model's ability to explain the variance in the target variable did not significantly improve.

Next, we developed a model to predict the likelihood of a customer using a loyalty card. Considering the nature of the problem at hand, we opted for a supervised classification model. To facilitate model selection, we utilized the Lazy Predict library, which quickly provided a range of potential models for our dataset. Based on the Lazy Predict results, we selected the Random Forest Classifier model for its superior performance and robustness in classification tasks. Once the model was chosen, we trained it on the training set and evaluated its performance on the test set. The model's efficacy was assessed using a variety of metrics, including accuracy score, precision, recall, and F1 score. The output of the model is as follows:

```
Accuracy using RandomForest Classifier: 0.9765
Precision: 0.8960573476702509
Recall: 0.9328358208955224
F1-score: 0.9140767824497258
```

These results illustrate the model's strong capability in classifying customer loyalty card usage, providing a reliable tool for strategic planning and decision-making.

The proportion of people with a loyalty card was significantly lower compared to those without one, indicating a data imbalance in the dataset. This imbalance can lead to biased model predictions that favor the majority class over the minority class. To address this issue and improve the model's performance, we employed under-sampling techniques to balance the dataset. Under-sampling involves reducing the number of samples from the majority class to match the size of the minority class. This technique can help create a more balanced dataset, allowing the model to learn more effectively from both classes.

Additionally, we applied k-fold cross-validation to ensure robust evaluation of the model's performance. K-fold cross-validation involves dividing the dataset into k equally sized folds, using each fold as a test set once while the remaining folds are used for training. This method provides a more comprehensive assessment of how well the model generalizes across different data subsets, helping mitigate the risk of overfitting to any segment.

```
Cross-Validation Scores:
Accuracy:
  Mean: 0.9715
Precision:
  Mean: 0.9004132357415939
Recall:
  Mean: 0.8903344830352129
F1_score:
  Mean: 0.8952209869777444
```

Following the successful validation, the two models—one to predict total purchases and the other to estimate the likelihood of a customer signing up for a loyalty card—were deployed for making predictions on new, unseen data. These models offer the potential to transform raw data into actionable insights in real-time, assisting the store in decision-making processes.

To facilitate user-friendly interaction with the models, a user interface (UI) was created for seamless predictions. The UI was integrated with the models using Python Flask, enabling intuitive access and efficient execution of predictions. This integration allows store personnel and stakeholders to leverage the predictive capabilities directly from a web interface, enhancing the usability of the models.

The real-time deployment of these models provides the store with a powerful tool for immediate decision-making, allowing them to optimize operations and tailor customer engagement strategies on the go. By leveraging this technology, the store can respond proactively to changing customer preferences and market conditions, positioning itself for sustained growth and success.

FINAL PRODUCT: WEB INTERFACE

Loyalty Card Prediction :

Multi Mart Retail Store

Is the Customer Active?
☐ Yes ☐ No

What was the last feedback provided by customer ?
Choose One

What is the total revenue generated by the customer ?
Enter a Numeric Value

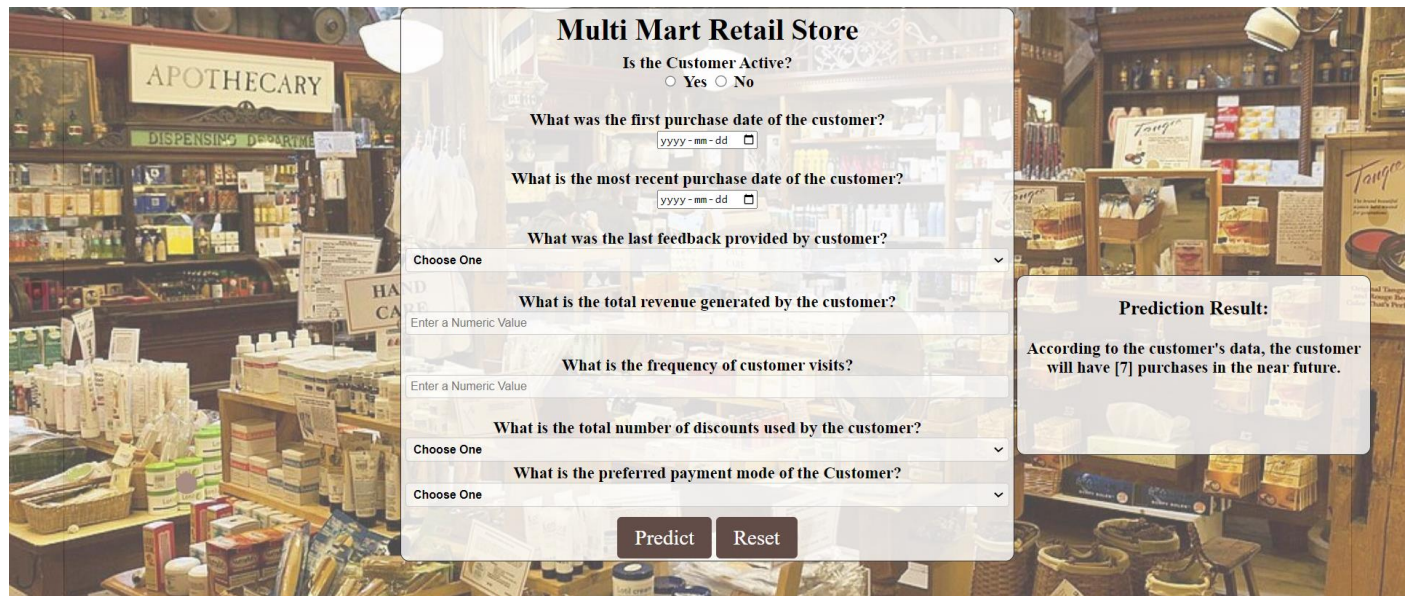
What is the total no of purchases made by the customer ?
Enter a Numeric Value

What is the frequency of customer visits ?
Enter a Numeric Value

Predict **Reset**

Prediction Result:
According to the customer's patterns: There is a **HIGH LIKELIHOOD** that the customer might opt for a loyalty card when promoted with one.

Total Purchases Prediction :



Multi Mart Retail Store

Is the Customer Active?
☐ Yes ☐ No

What was the first purchase date of the customer?
yyyy-mm-dd

What is the most recent purchase date of the customer?
yyyy-mm-dd

What was the last feedback provided by customer?
Choose One

What is the total revenue generated by the customer?
Enter a Numeric Value

What is the frequency of customer visits?
Enter a Numeric Value

What is the total number of discounts used by the customer?
Choose One

What is the preferred payment mode of the Customer?
Choose One

Prediction Result:
According to the customer's data, the customer will have [7] purchases in the near future.

Predict **Reset**

CONCLUSION

The predictive models developed for this project provide useful insights into customer behavior and purchasing trends, serving as important tools for the store's strategic decision-making. The model for total purchases achieved moderate accuracy, which suggests it captures some of the key factors influencing customer spending. However, the moderate performance indicates limitations in its ability to provide reliable predictions. The store should consider these limitations when using the model's output for decision-making and explore additional features or advanced techniques to improve its accuracy.

On the other hand, the model for predicting the likelihood of a customer signing up for a loyalty card demonstrated better results, with strong accuracy and balanced evaluation metrics. This model can be confidently utilized by the store to enhance customer engagement strategies and optimize loyalty programs. Its predictions offer actionable insights that can help the store tailor its marketing efforts and improve customer retention.

In conclusion, while the total purchase model shows potential, further refinement is needed to achieve more reliable results. In contrast, the loyalty card model presents robust outcomes that can be effectively leveraged by the store. Both models lay a strong foundation for the store's ongoing data-driven initiatives, which can lead to better-informed decisions and sustained growth.

REFERENCES

- *Undersampling*. CORP-MIDS1 (MDS). (2023, December 15). <https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/>
- Wisam, E. (2023, October 14). *Class imbalance: Exploring undersampling techniques*. Medium. <https://towardsdatascience.com/class-imbalance-exploring-undersampling-techniques-24009f55b255>
- *3.1. cross-validation: Evaluating estimator performance*. scikit. (n.d.). https://scikit-learn.org/stable/modules/cross_validation.html
- *A step-by-step explanation of principal component analysis (PCA)*. Built In. (n.d.). <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>
- *SKLEARN.DECOMPOSITION.PCA*. scikit. (n.d.-b). <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>