Data Science Capstone project on Retail Store Revenue Analysis & Forecasting

This interim report outlines our Data Science capstone project focusing on Multi Mart, a retail store. It includes analysis and predictions in two key areas:

Analysis/visualization

1. Evaluation of store sales performance and revenue generation.
2. Assessment of opportunities for expanding the Loyalty Card program to new regions.

Prediction/Forecasting

1. Anticipating the total number of purchases to forecast store revenue.

The Approach:

Every analytical Machine learning project is structured around a sequence of steps aimed to achieving its goal. These steps can be outlined as follows:

1. Identification of the business problem
2. Data Acquisition
3. Data Pre-processing and Cleaning
4. Exploratory Data Analysis
5. Analysis/Visualization to help business understand the problem
6. Identification of features and target variables for machine learning
7. Model development
8. Prediction

All the above-mentioned steps are explained thoroughly.

Identification of the Business Problem:

For the Multi Mart retail store, the challenge was to grasp sales trends, revenue patterns, and customer behavior for making better informed decisions. Additionally, the store initiated a Loyalty card program in one region and now seeks to extend it to new regions.

The primary objective of the visualization project for Multi Mart Retail store is to provide a comprehensive and visually intuitive view of the sales that happened in the store and revenue generated from sales from the year 2019 to 2023. This comprehensive analysis aims to shed light on the store's sales performance and revenue generation. With a focus on customer behavior and preferences, our goal is to guide the store in making informed decisions about the potential expansion of its Loyalty Card program to new regions.

Data Acquisition:

The next step after Identifying the business problem is obtaining relevant data. There are many data sources available on the internet that are easily accessible, for example Kaggle. It is relatively easy to search for data on Kaggle as well as look at examples of how other data scientists have utilized the same dataset. This is particularly helpful for comparative purposes i.e. if you want to compare your analysis of a dataset with how someone else used it.

The dataset used for this project comprises various customer-centric variables including purchase history, revenue, churn indicators, product categories, campaign responses and loyalty card participation. The data set was sourced from British Library.

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.

And below is the sample dataset. The granularity of the data is at customer level, each row refers to an individual customer, and our target variable in this case is 'totalpurchases' which indicates the sum of purchases made by a customer.

customerid	firstpurchasedate	lastpurchasedate	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 Pre-processing and Cleaning:

Once we have the data, the next step is to clean the same. Being from a Data Engineering background, our experience is that the analysis and visualization will showcase good results only if the data quality is good. And cleaning the data involves tasks such as checking for null values, imputing missing values, checking for outliers, or making sure columns are named correctly. Our data set was clean in terms of Nulls, missing values, or duplicates. We derived two additional columns for the simplifying the analysis. Below are the cleaning steps performed.

- The dataset comprises of 2 Date-Time, 6 Categorical, and 11 Numerical columns.
- None of the columns in our dataset have any missing values. Also, there were no duplicate records.
- We added below two additional columns to the data frame:
- Rounded Feedback Score: This column contains the feedback score rounded to the nearest whole number.
- Products: This column contains products by grouping them according to the initial letter of their names listed under the product categories column.
- Finally, we exported the data to an excel file and then imported to Tableau for data visualization.

Exploratory Data Analysis:

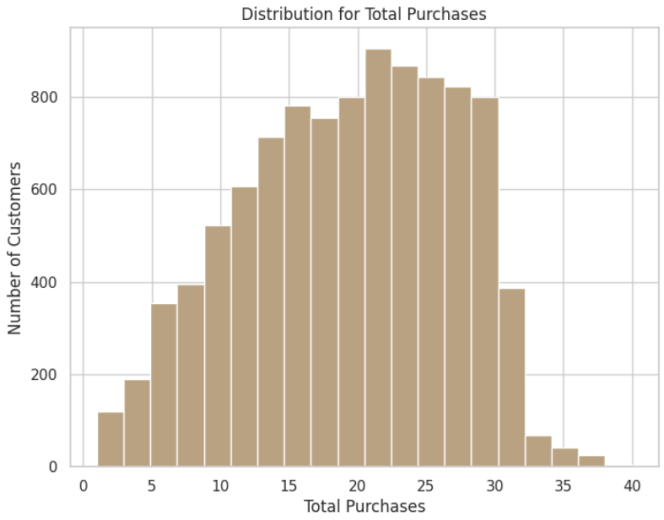
Next, we explore the data to gain insights into our dataset and what it contains. This includes, but is not limited to, checking for outliers or unusual data by looking at data distributions; and examining the relationship between the response and predictor variables using a correlation matrix.

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.

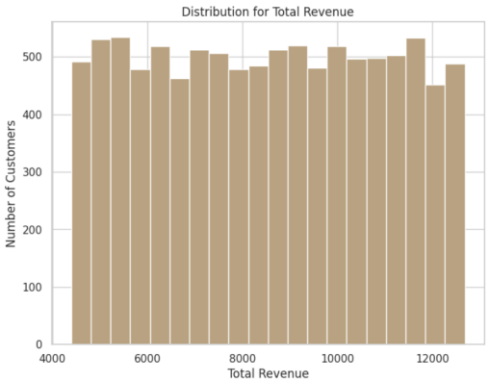
- To start with, we first looked at the basic descriptive statistics of all the numerical columns of our dataset to get a better idea about our dataset for identifying its competencies.

	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

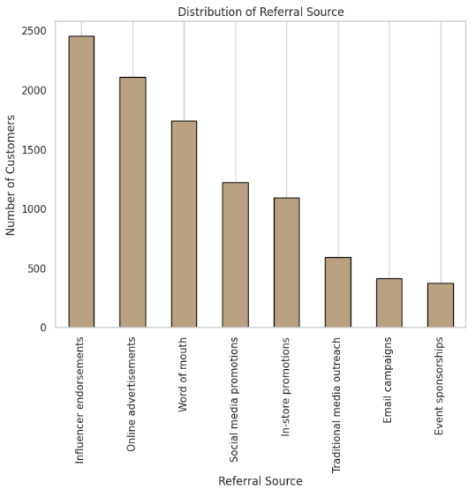
- Based on the provided statistics, it's apparent that the majority of columns exhibit no outliers, with only two exceptions.
- Among these exceptions, the columns identified with outliers are 'feedbackscore' and 'frequency'.
- In the case of 'feedbackscore', the lower bound of the Interquartile Range (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.



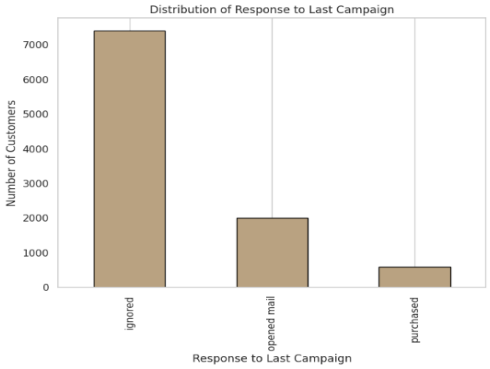
- Distribution of 'Total Purchase' is symmetric and unimodal.
- This implies a balanced and centered pattern in customer buying behavior.
- Most customers tend to make a consistent number of purchases, forming a bell-shaped curve around an average.



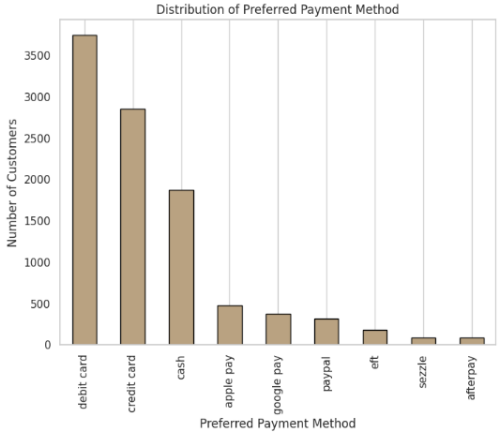
- 'Total Revenue' is uniformly distributed.
- This indicates an even and consistent spread of revenue across different customer segments or transactions.



- '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 '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.

- 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.

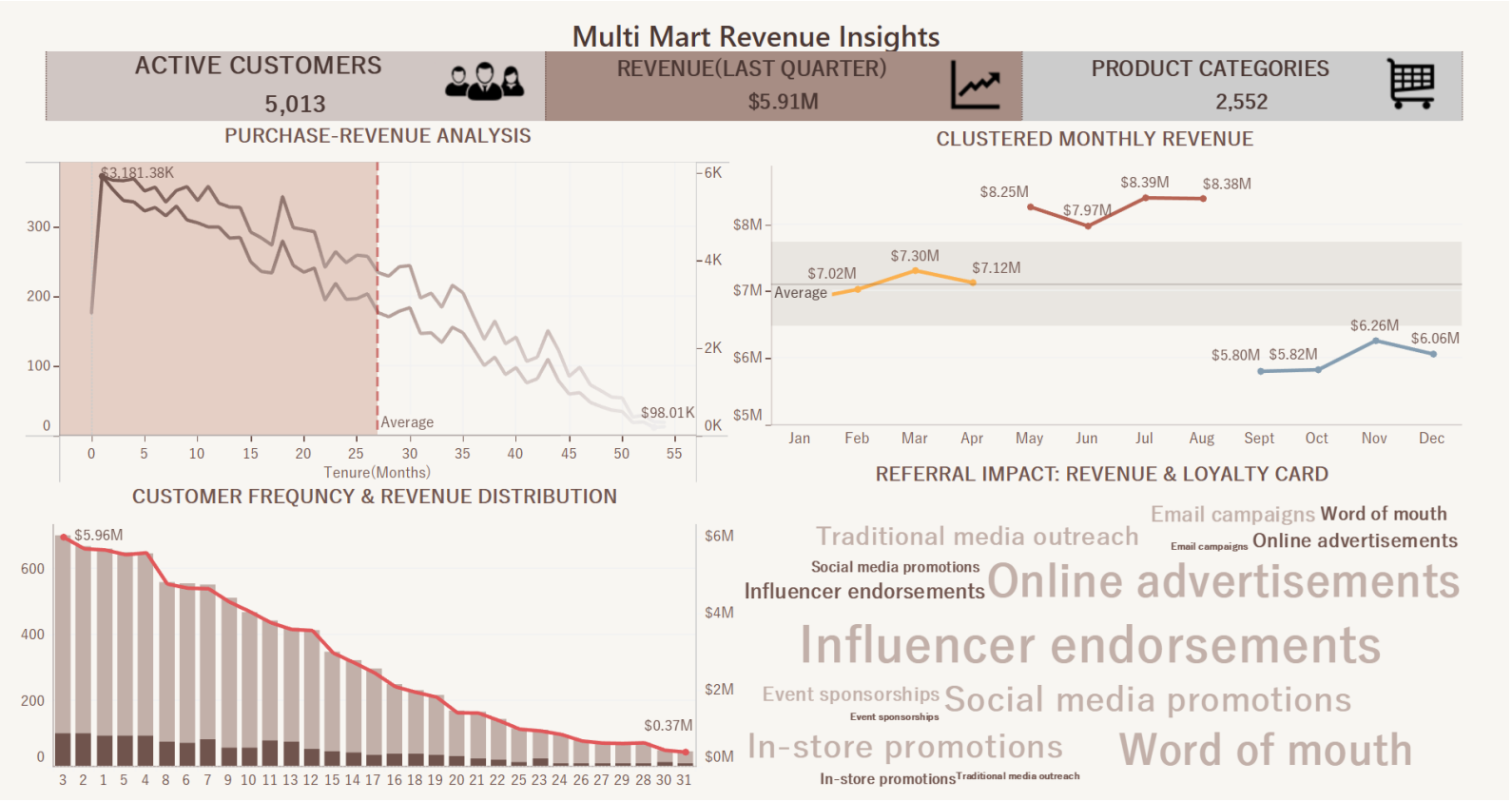
We use a heat-map to explore the variables in the dataset and check if there is a meaningful correlation between them. A correlation value close to -1 indicates strong negative correlation, values close to 0 indicate no/very weak correlation, and values close to 1 indicate a strong positive correlation between variables. The 'Exited' and 'Age' variables had the highest correlation value of 0.29, which does not suggest strong correlation between the pairs as the value is close to 0. All the other variables had very weak correlation values as shown in

- Customers making more purchases tend to engage with higher frequency.
- Other correlations do not exhibit very significant strengths 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.

Analysis/Visualization to help business understand the problem.

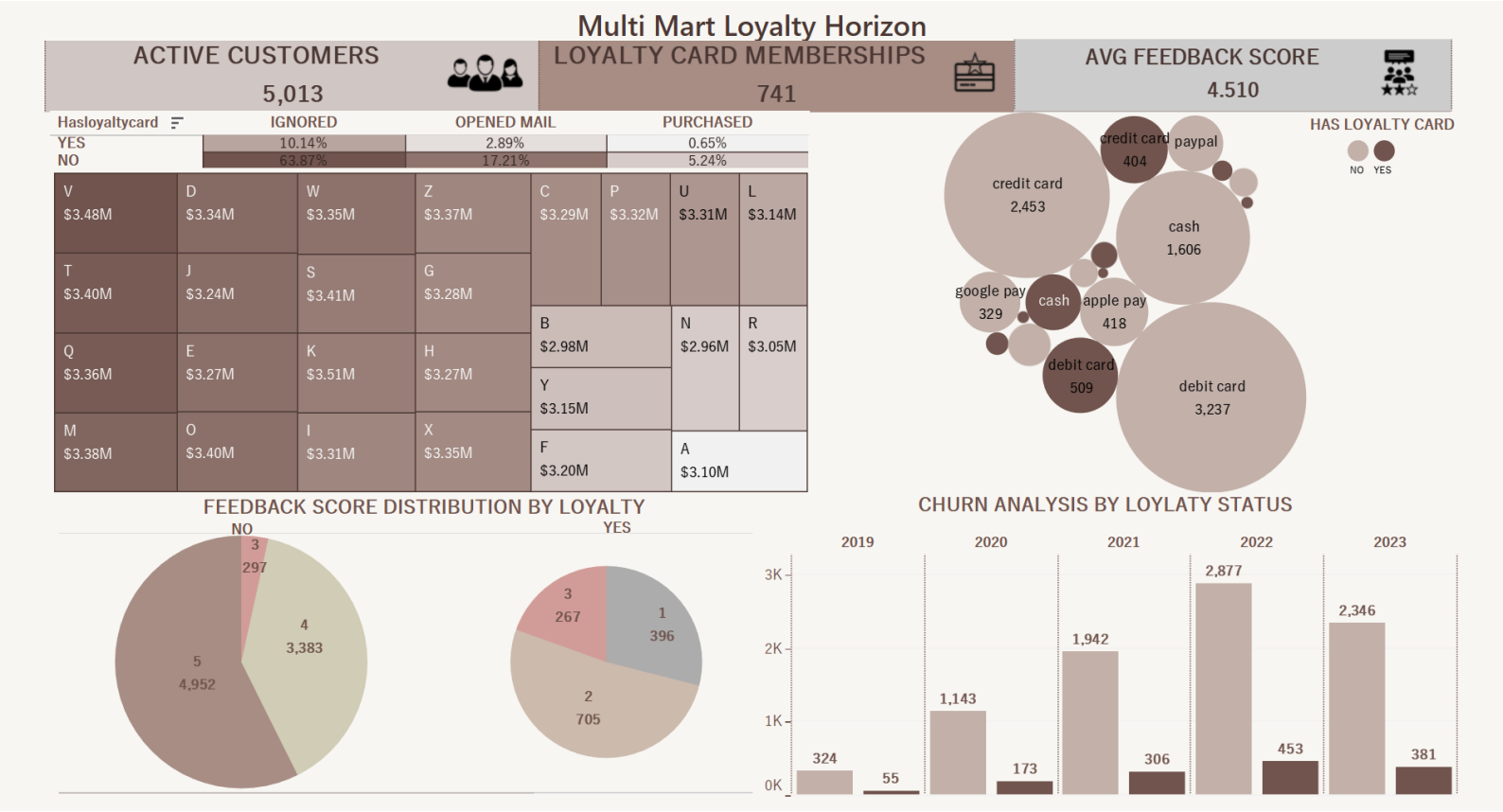
In this step we have used tableau to display trends and analysis in our data sets through visualization. There are two dashboards we have designed to explain below two problems /objectives to the business.

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.

Identification of features and target variables for machine learning

As per the initial steps in machine learning, it is necessary to consider as many features as possible, and since this is a regression problem, we have converted the non-numerical columns to numeric using one hot encoding. After encoding, features and target variables selection was done for machine learning model creation, for that we first analyzed the correlation between various variables in dataset. Our goal was to identify the most significant features and determine an appropriate target variable to drive the predictive model which ultimately benefit the store for their operations and decision-making processes.

While analysing the correlation, we observed that only two variables, total purchase and frequency of engagement demonstrated a significant correlation of 0.49 suggesting that customers who engage with higher frequency tend to make more purchases. This finding underscores the interconnectedness between customer engagement and purchasing behavior, shedding light on a pivotal aspect of our analysis. Other variables lacked in significant correlation strength, so we decided to go select total purchase as our target variable and understand the factors driving total purchase of store for further analysis.

Moving forward, our focus will be on delving deeper into the features that impact total store purchases. Leveraging techniques such as feature importance and Principal Component Analysis (PCA), we aim to discern the key drivers behind purchasing behavior within the store environment. By meticulously unraveling these underlying factors, we endeavor to construct a robust machine learning model capable of accurately predicting total store purchases.

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]
```

Model Development and Prediction

Any machine learning model requires the dataset to be divided into training as well as testing data so as to train the model using one and then applying the model on the new test dataset to understand how well the model is able to 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.

Because the problem at hand is a regression problem, we would have to go ahead with supervised machine learning models that are compatible with regression problems. The list of possible models for the type of problem we are working on are as follows:

1. Linear regression
2. Decision tree
3. Random forest
4. Ridge regression
5. Lasso
6. Support Vector regression
7. KNN Regressor
8. Logistic Regression
9. Bayesian linear regression
10. Gradient boosting algorithms
11. Partial least squares regression
12. ElasticNet regression
13. Gaussian Regression
14. Polynomial Regression

Moving forward we started researching about the models and if and how they can be useful for our dataset.

We experimented with supervised regression models, like Random Forest Regressor and Linear Regressor, aiming to capture the underlying patterns in the data. The models were trained using the training dataset, allowing them to learn the relationships between features and target variable (total purchases). To assess the performance of our models, we employed various evaluation metrics including R2 score, Mean Absolute Error (MAE), and Mean Squared Error (MSE). These metrics were computed using the test set to gauge the models' ability to generalize to unseen data.

The metrics output for both the models are as follows:

```
R-squared value using Linear Regression model: 0.5215026410302918
MAE: 4.4233458686559395
MSE: 29.37090370287445
RMSE: 5.419492937800957


R-squared value using Random Forest model: 0.3136134162751568
MAE: 5.05467
MSE: 42.131464
RMSE: 6.490875441725869
```

Considering the low performance of the models, we intend to explore alternative supervised regression models that may better suit the characteristics of our dataset. To achieve the goal of gaining optimal performance of a machine learning model for our dataset, we further wish to keep researching while also applying different models on our data. Once the best model is chosen, we will further make it better by tuning the hyperparameters so as to get the optimal performance from this model that will successfully be able to identify and capture most of the variations and patterns in our dataset so as to perform better in the new unseen data thus giving the best insights to the business.

Conclusion and Further Work

From the analysis/visualization, we were able to highlight problems in the existing business operations in terms of Revenue and Loyalty card expansion.

Further using Machine learning and deep learning if compatible with our data set, we will be helping the business to predict the total purchases so that business can plan a head satisfying their customers needs and wants.