Submitted by : Sarthak Sunil Dhanke Date : 10/16/2024

McDonald's Case Exploratory Data Analysis

1) What is the business problem as you understand it?

Ans : To ameliorate the falling profits and the customer McDonalds introduced 'All Day Breakfast' hoping that it would bring those customers back who prefer to have the breakfast items even after the official breakfast hours. Accordingly it was implemented in McDonald's restaurants in Michigan as well. However this s**trategy was met with mixed responses.** In some cases the sales indeed increased and customers came back, yet in other cases there was no marked improvement. Now McDonald's wants **to quantify the impact of introducing all day breakfast** using the data which they have collected over time because **they want to decide whether to continue promoting all day breakfast or not since the strategy is not without its drawbacks.** For one, there are technical difficulties in delivering breakfast items since it requires a different process than the regular items and in some cases this clutters the kitchen. And secondly the profit margin on the breakfast menu is low. Therefore even if the sales increase, there may not be a concomitant rise in revenue. Hence **the business problem is to figure out whether the monetary impact of the policy is worth the additional efforts or not.**

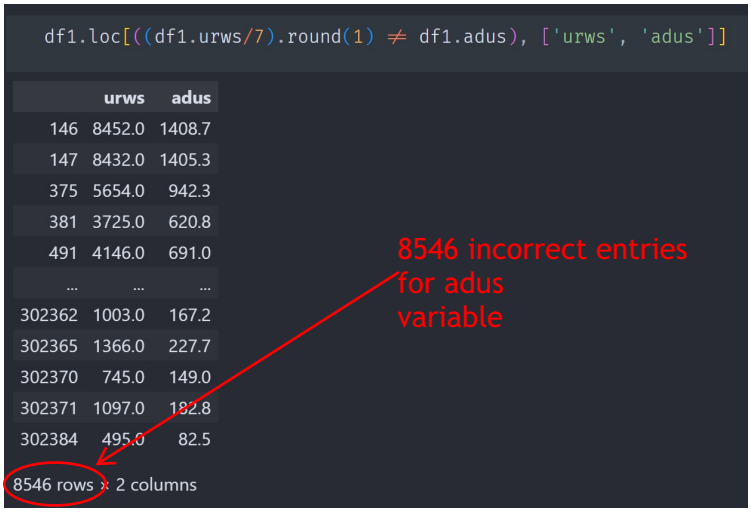
2) What variables in the data relate most directly to the business problem?

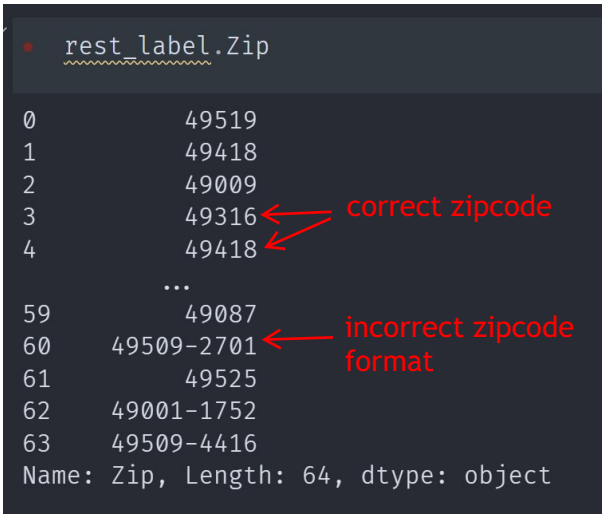
Ans : Since we want to measure whether the introduction of 'All Day Breakfast' increases the sales as well as the customer count over the time when the policy was being implemented, the most important variables are
itemdesc : the name of the item being sold to identity breakfast and non-breakfast items
wk_ending : to perform analysis in the time period we are interested in
urws: Total number of unit of an item sold over a week
wavg_price: Weighted average price of the item over the week
agc: The total count of customers during a week
Apart from these we can use all the variables in the rest_facts.csv file to identify non-performing restaurants and why the introduction of all day breakfast was not successful in those restaurants.

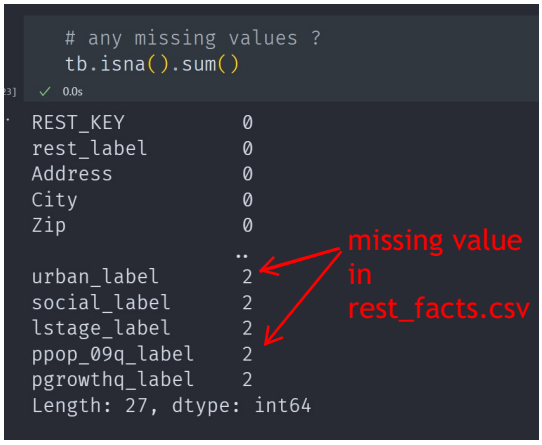
3) Are there any data quality concerns with these variables?

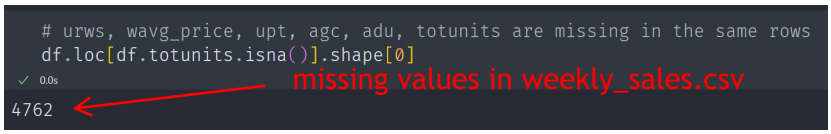
Ans : There are two main problems to be identified.

i) The first one is the **missing data** in the M395_weekly_sales.csv files; 6 variables have 4762 values missing in total.
ii) The other one is **data entry error.** The adus variable was supposed to be urws / 7, however there are 8546 rows where this rule is not followed.
iii) Likewise M395_rest_facts.csv has 2 missing values.
iv) There are data entry errors in the Zip code field. Consistent format has not been followed.









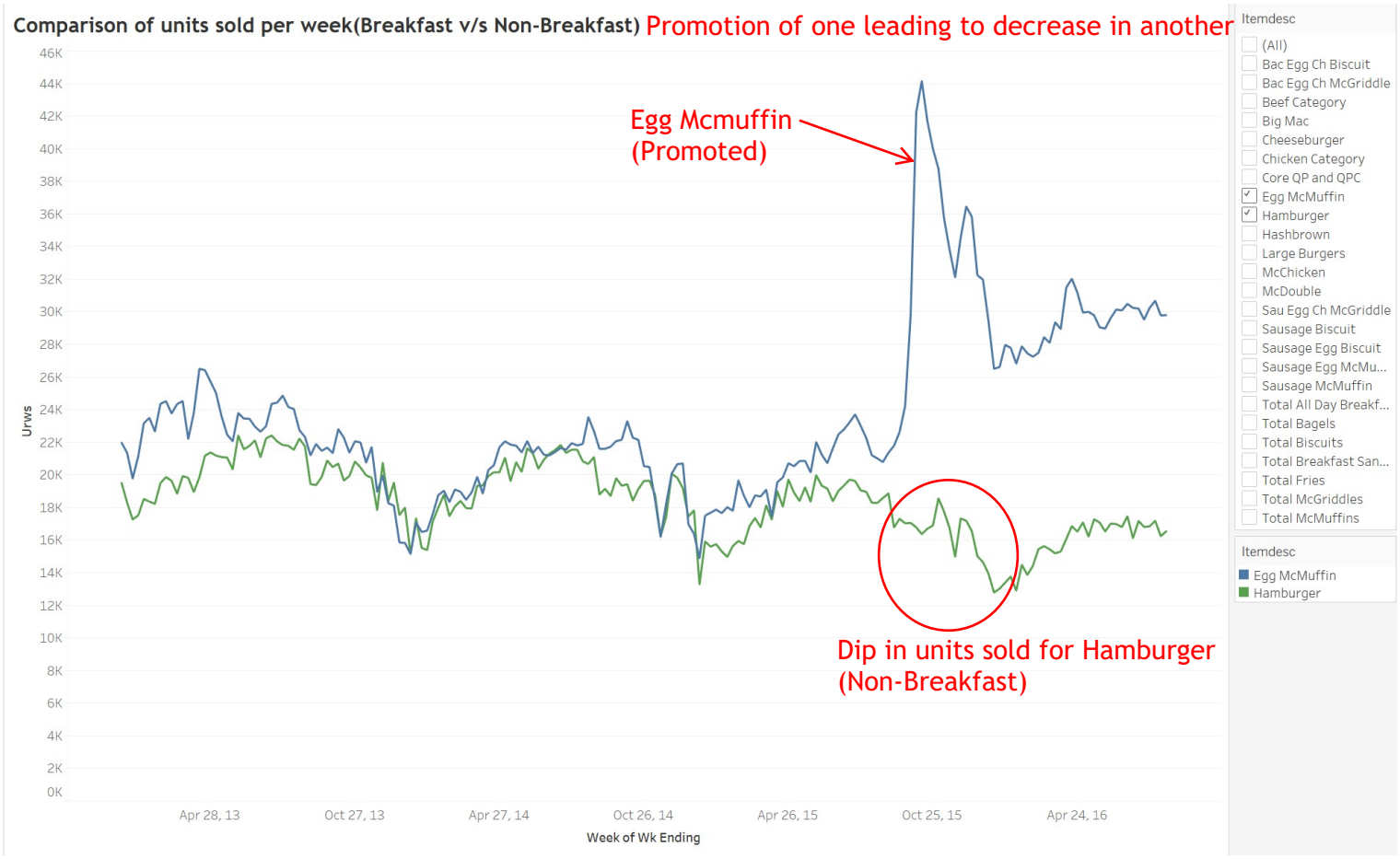
4) Do you have any hypotheses about how this specific data could answer the business problem?
Ans : Some of the hypotheses the can help use solve the business problems are as follows

i) Introduction of all day breakfast increases the units sold per week of all non breakfast items on average across all restaurants during the promotion period.
ii) Introduction of all day breakfast increases the units sold per week of breakfast items during the promotion period.
iii) Introduction of all day breakfast items increased the customer count on average across all restaurants during the promotion period.

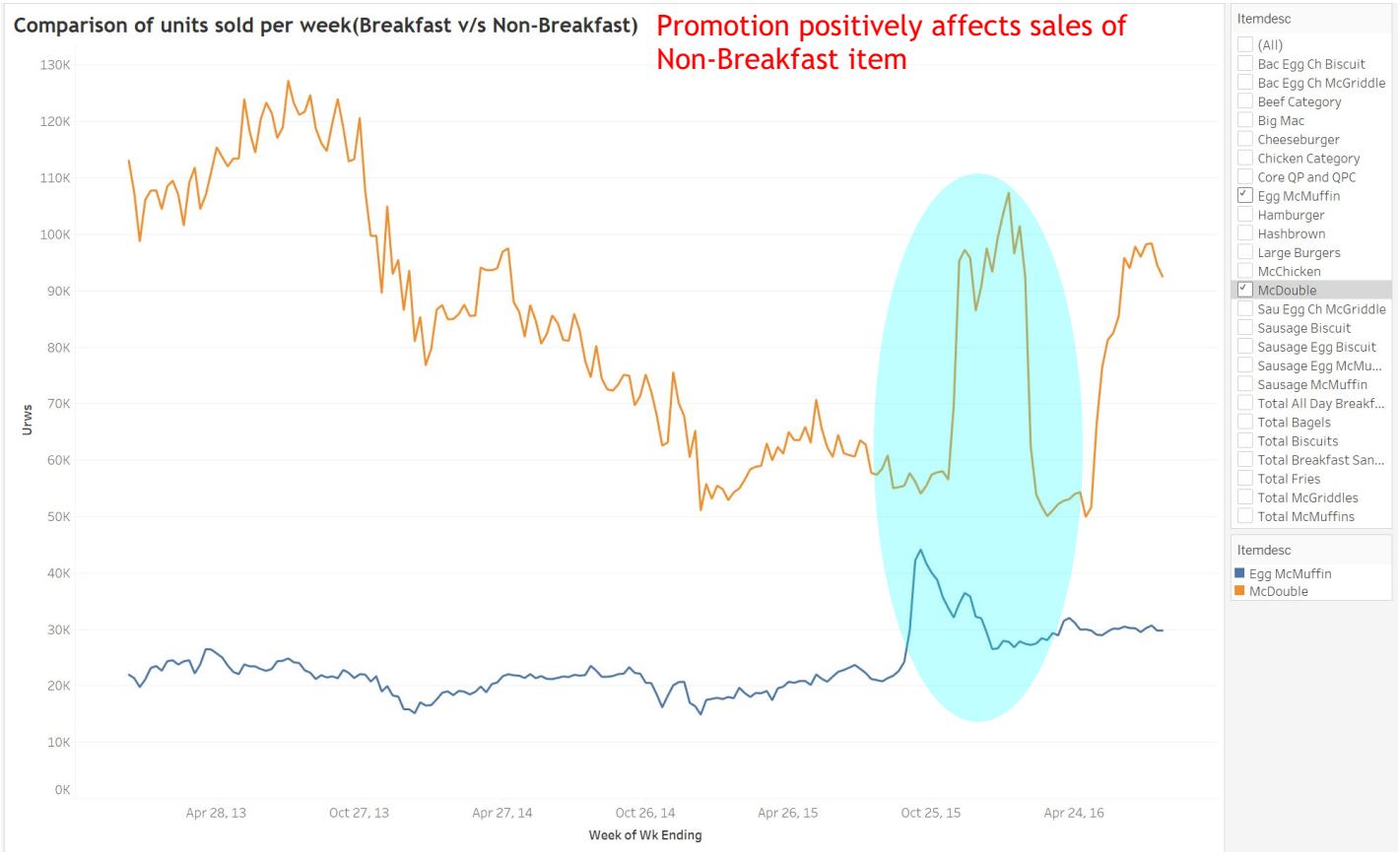
If we can test these hypothesis using the data we have we will come to know whether
i) The introduction of all day breakfast increases sales of breakfast or not as claimed by McDonalds
ii) Did the increased sales translated into equivalent dollars. This is important because even if sales for breakfast items increased, it does not mean that it will be profitable since the margin on breakfast items is low. Hence to break even we have to consider the actual profit, and the sales of non-breakfast items.
non-breakfast items

5) Is the data likely sufficient to confirm or reject your hypothesis? Consider data completeness, data quality, and methodological assumptions in your answer. If not, what additional data or corrective measures will be needed to answer the business problem?

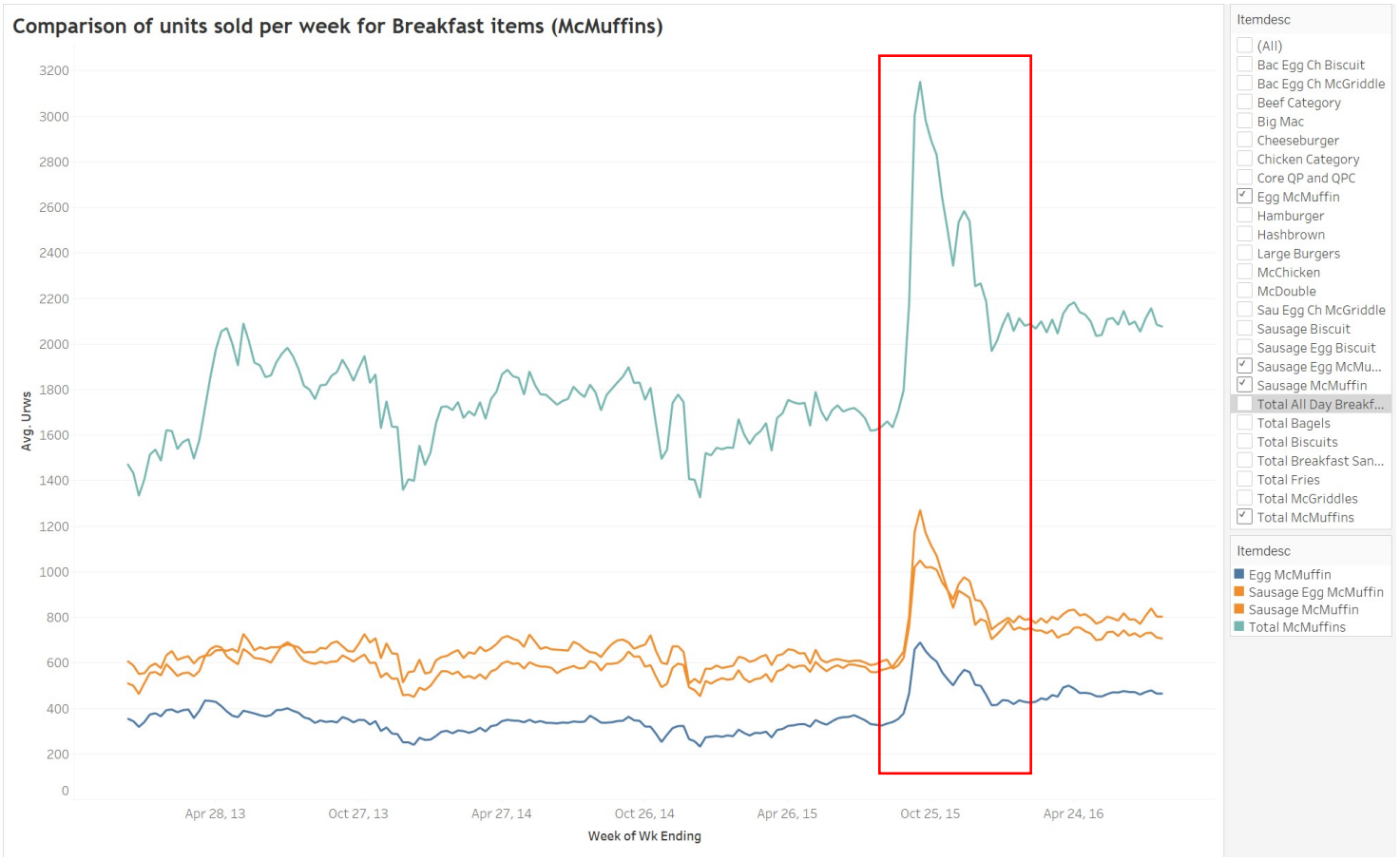
Ans: From a completeness perspective, the missing values are not from the period promotion was done, therefore it should not pose a problem to the analysis. We do have columns which have data entry errors, but they are redundant for example urws and adus. Once urws is known, we can calculate adus = urws / 7. One assumption is that **we only need to consider the period when the promotion was applied**, and another period does not help in answering the question. If **the series are autocorrelated then this may be false.** And there is a good chance that it will be. For example, a particular restaurant might be doing badly due to other reasons, and this will affect their performance even when the promotion is applied.
Apart from that **we need the exact date when the promotion was started** so that we can consider only that part. This is an additional information we need. In absence of this, we can simple use the spike in sale of breakfast items across all the restaurant as a crude indicator.
Furthermore we are looking at the average measure is sensitive to outlier so outstanding performance in one restaurant might affect the overall statistics. It is better to look at each and every individual restaurant. What might also help is having data on how difficult it was for each restaurant to implement this program. Besides, **there is the analytical difficulty of analyzing data at the restaurant level to understand why a particular population responded better to the promotion and others did not.**



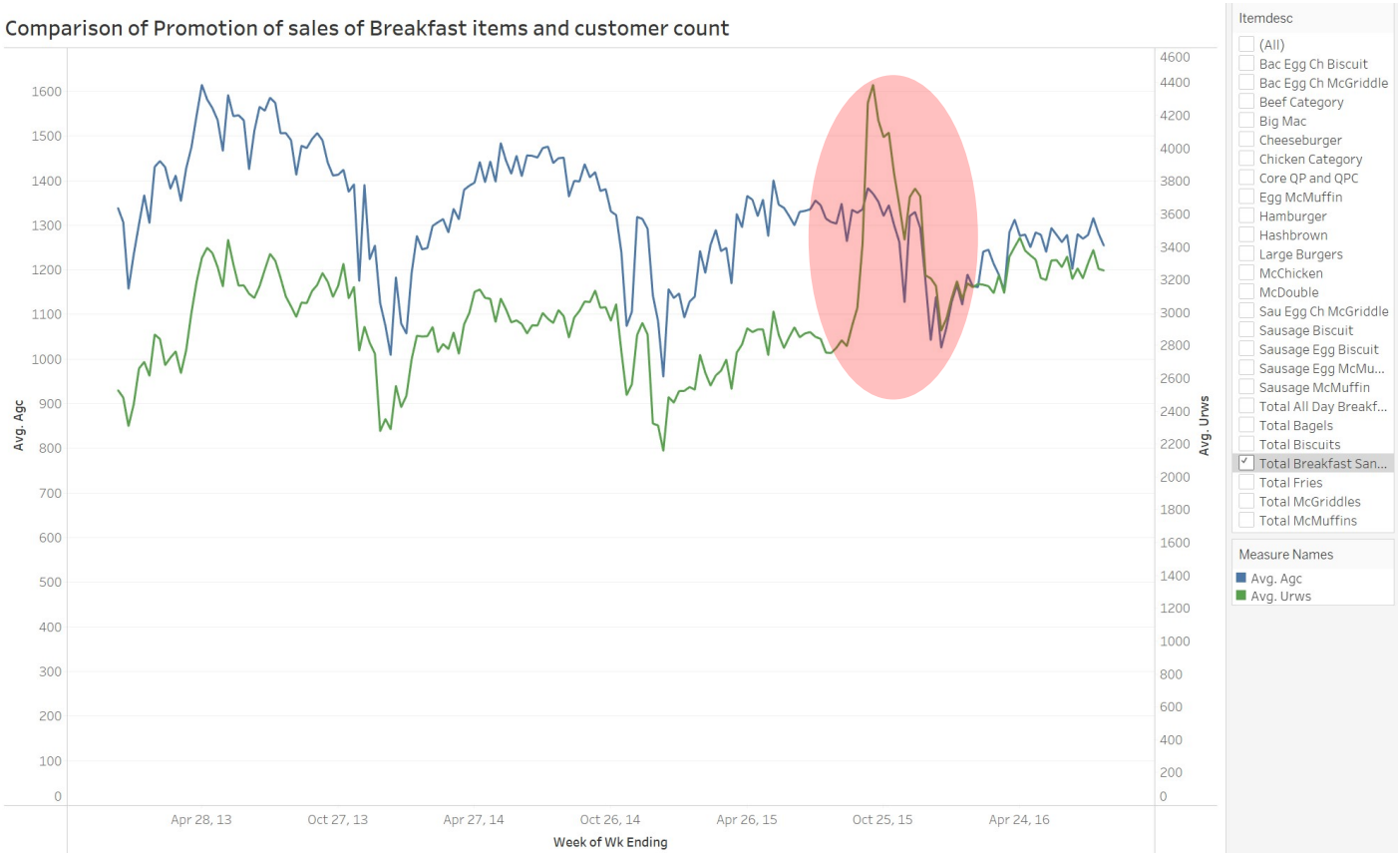
If we look at the simple case of Egg Mcmuffin which was promoted during all day breakfast we can immediately observe that its sales rocketed, however Hamburger sales dipped. Again note that the figures for these items were comparable before promotion, therefore this presents an example where our hypothesis 1 is not true.



In other case (Egg McMuffin vs McDouble) we can see the sales of McDouble went up sharply during the promotion period, which supports our hypothesis. However we should all consider the regional factors. This could be due to increase in McDouble sales in few restaurants, which affects the entire group



To test the second hypothesis we can check the individual sales of promoted items vs their total sales. This shows that all the promoted itesm received a boost, which confirms the hypothesis. However, this increase was short lived, perhaps because promotion was discontinued.

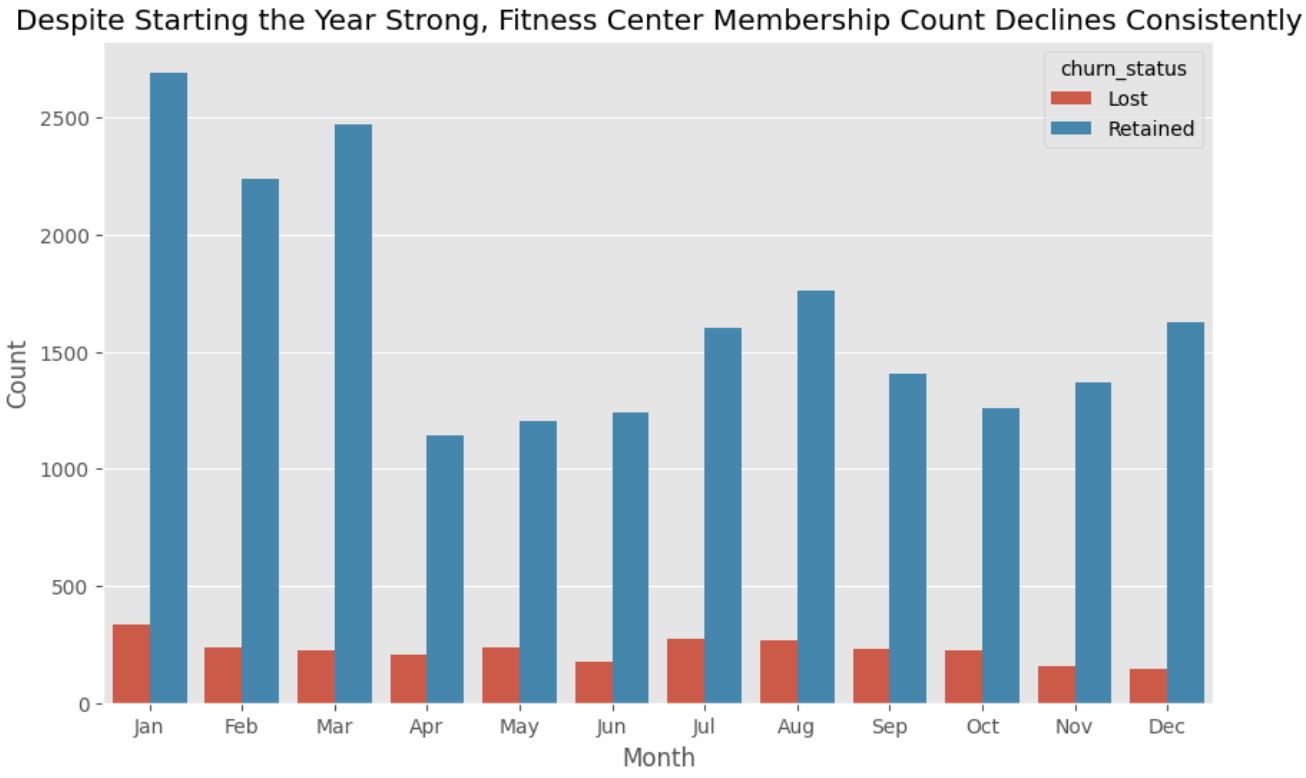


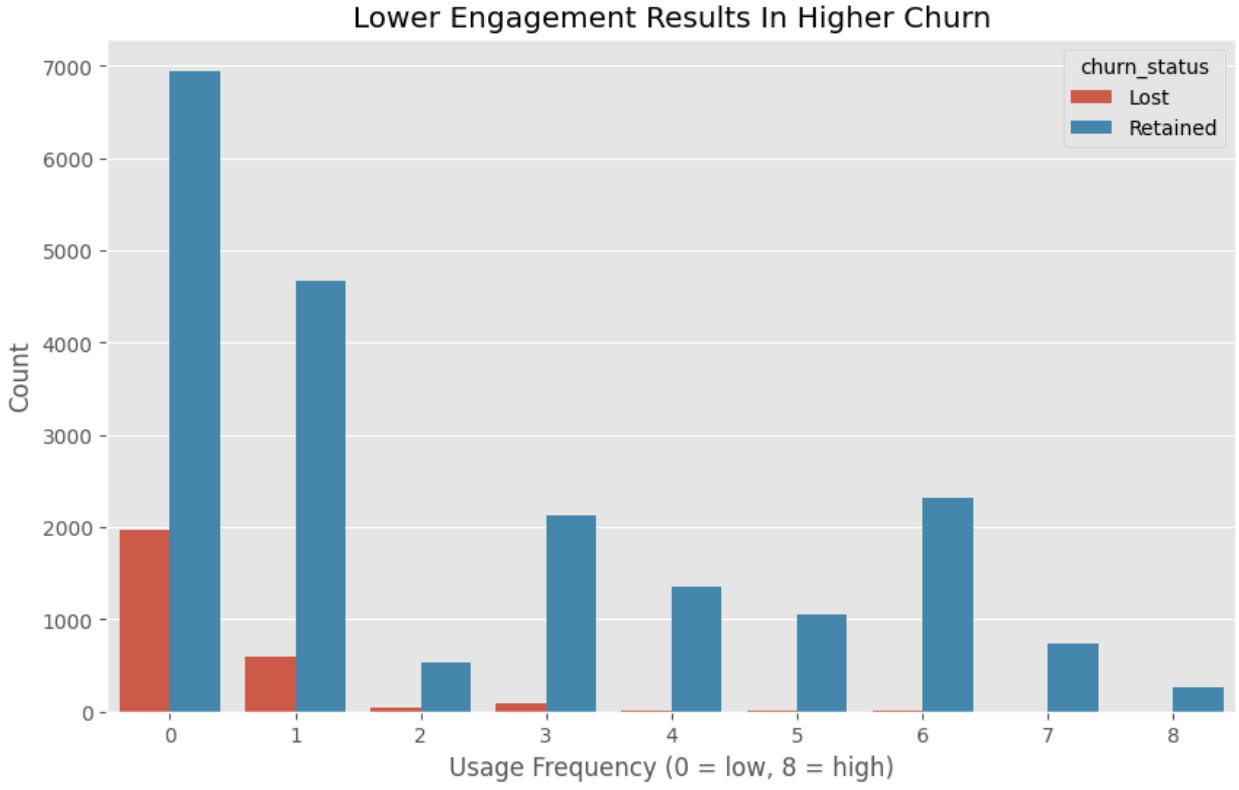
If we compare the average customer count during the promotion period, it is apparent that the it did not increase proportionally to the sales figures. Furthermore there is a strong seasonal component to the customer count data. Perhaps if all day breakfast was introduced during peak of customer traffic then perhaps we could have seen better results. From the graph we can deduce that there is no strong evidence that customer count increase as suggested by the third hypothesis.

Submitted By: Sarthak Sunil Dhanke For Fitness Center Exploratory Data Analysis Exercise

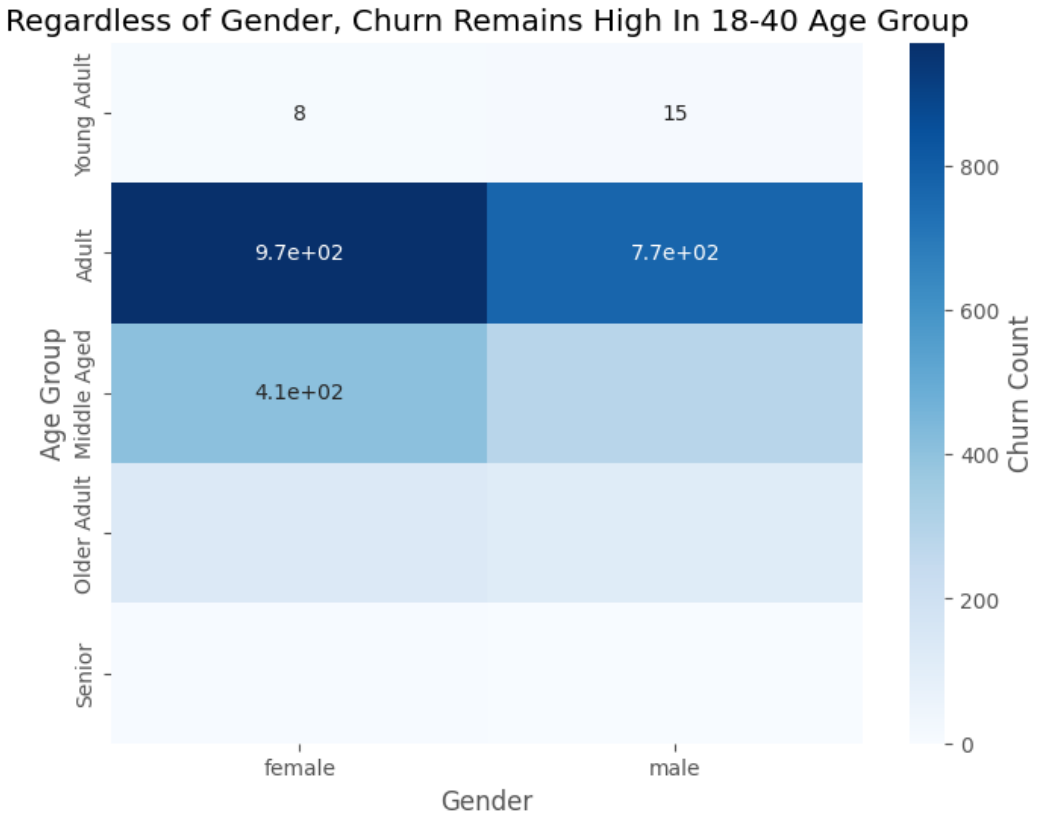
1) What do you consider the specific business problem and the matching goal of analysis?

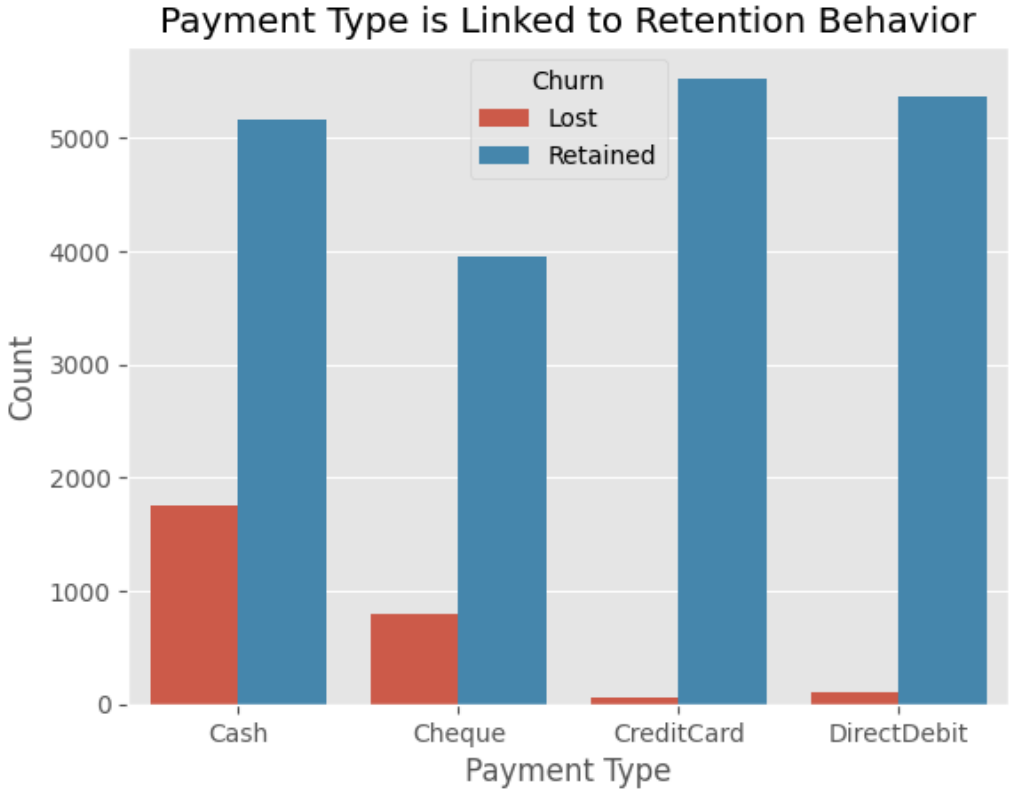
Business Problem: The fitness center is **losing customer throughout the year** despite many people signing up in beginning of the new year. Furthermore, the fitness center wants to understand **why customers are leaving**, and what can be done to reduce the churn(as can be seen in Fig 2, **lower engagement** is one key reason).





Goals of Analysis: To identify patterns and trends in the membership behavior, payment preferences, demographics, and usage frequency correlated with churn and suggest strategies to reverse the trend and reduce revenue losses. (Figure below shows trends between churn and payment type, gender, and age group)





2) What modeling approach do you suggest and what did you notice in the data that suggests it will be helpful here?

Given the categorical churn target (1 = Churn, 0 = No Churn), a **classification model** would be ideal. We can use **Logistic Regression** because it is easier to implement and interpret, Besides it makes is easier to handle the quantitative and qualitative predictors present in the data.

Notable observations that suggest this method are:

i) Imbalanced distribution in usage frequency: Low usage is linked to higher churn, making it a key feature.

ii) Categorical nature of payment types and demographic data: These can be handled easily by using assuming a linear relationship between the log odds of probability of churn and the predictors.

iii) Right-skewed distributions with outliers: With logistic regression we can handle transformed variables as these can be used as predictors in the linear relationship between the log odds of churn-probability and the predictors.

3) What questions, if any, do you have about the client organization, or its broader industry?

i) Does the fitness center has a retention program or member engagement program, and is it effective? This will help evaluate current retention program and why it is inadequate. Moreover, it might reveal more data which can adjunct the analysis.

ii) What services are preferred by the customers and why? This will help us understand the strengths and weaknesses of the fitness center and how to leverage them to design a better retention program.

iii) Are there any strategies implemented by the competition which help them retain customer better than us ? Understanding competitors could offer insights into gaps in the organization's own retention strategies.

iv) Does fitness center run promotion or offer incentives to customers who join in the beginning of the year ? If the fitness center does indeed offer incentives, it might be the reason for spike in new membership in the beginning of the year. It might that member only join for the sake of incentives, and never intended to stay for a long time.

4) What questions, if any, do you have about the data?

i) CMO mentioned that data collection process varies according to location, so one question to ask is if there are any large gaps in the data or if the data might be entered in different format. Such as different symbols or values for missing data.

ii) How is the use measured on the scale of 0-8 ? Is it tracked weekly or daily

iii) How is the churn measured ? Does churn = 1 mean that the customer has cancelled his membership or simply did not attend for a long time.

iv) Is there any feedback data available which can be used to learn why customer left ?

v) How accurate is the data ? Did the fitness center validated the data which they have ?