Café Insomnia:

Task 1:

1. Seasonality:

Starting our analysis, we will first look at seasonality and how it can impact the hourly revenue generated on different days of the week and at different hours of the day. (Note: Revenue generated was calculated by multiplying unit cost times quantity). Seasonality was analysed by the consultant; however, the consultant's analysis lacked some very important features as shown in figure 1.1. and figure 1.2, without those features, the figures are unreadable and unhelpful. The figures require a title, axes labels, bar labels and generally better formatting. After improving the two figures into figure 1.3 and figure 1.4, we notice from figure 1.3 that the hourly revenue is changing on different days of the week. By looking further, we can clearly note that on Wednesday the business is expected to have the highest hourly revenue of $177.29 while on Monday and Saturday the business is expected to have the lowest hourly revenue which is approximately $146 per hour. Moving to the line chart in figure 1.4, it can be noted that coffee sales increase to a high peak of 175$ after 1 hour of opening and then rapidly decrease toward $135 after 5 hours of opening.

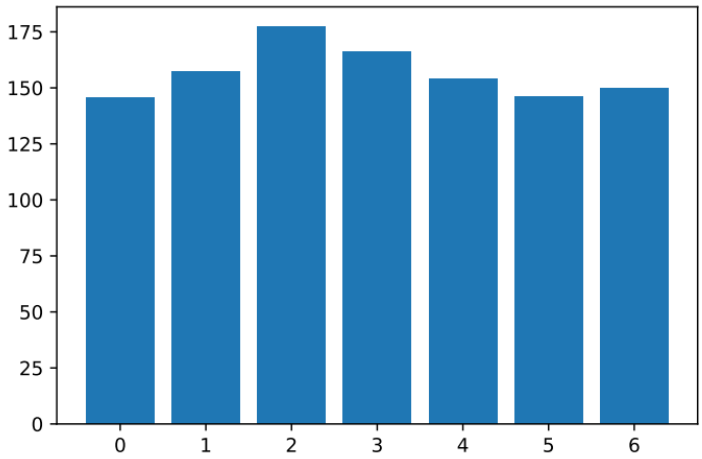

Figure 1.1. Previous Consultant's Boxplot.

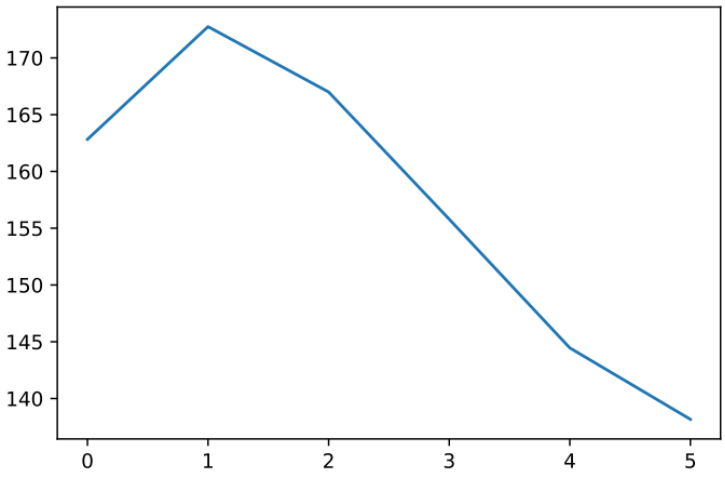

Figure 1.2. Previous Consultant's line chart.

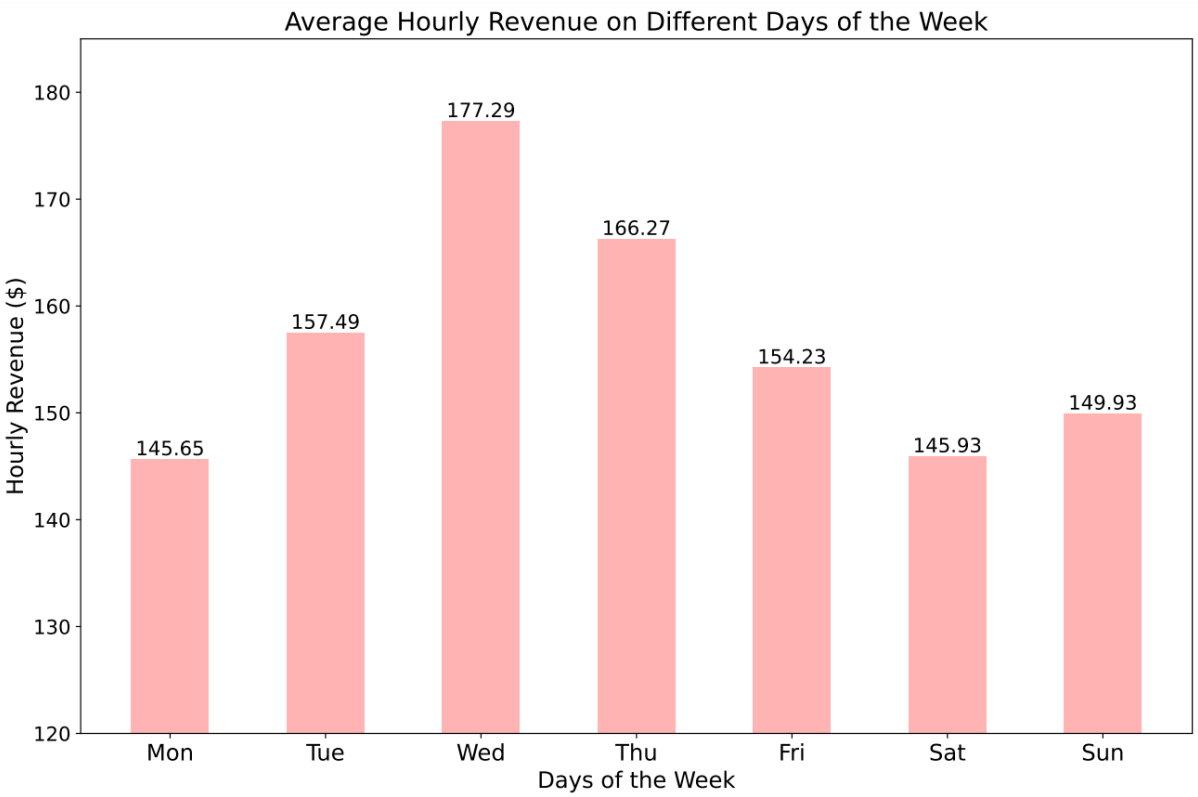

Figure 1.3. My Bar chart illustrates the changing hourly revenue on different days of the week.

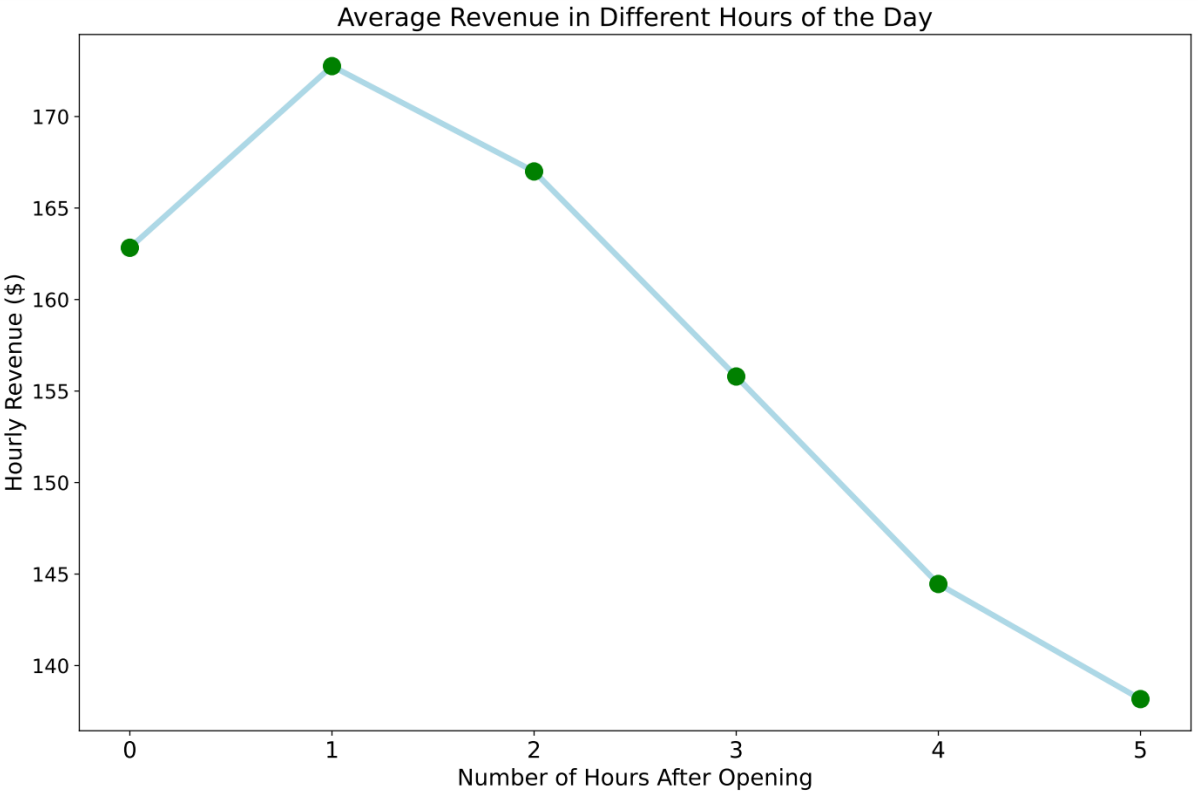

Figure 1.4. My line chart demonstrates the change in hourly revenue at different hours of opening.

2. Days Since Opened:

When plotting the revenue generated against days after opening, we can notice that the data takes on a convex shape at the beginning of the curve and a concave shape at the end of it. This data set is hence not very linear (figure 1.5). When we want to form a regression model, we will assume that the

dependent variable is approximately linearly related to the independent variable. Therefore, for this assumption to be maintained, we will need to transform our values of revenue by using the logarithmic function. The Logarithmic function is concave so it will help us properly transform our data. Figure 1.6 represent the newly transformed response or dependent variable. By looking at the scatter plot, we can observe that the data is more spread out for values of x < 40 and so the curve became less convex toward the beginning. This indicates that our transformation was successful in filtering out the nonlinearity of the data.

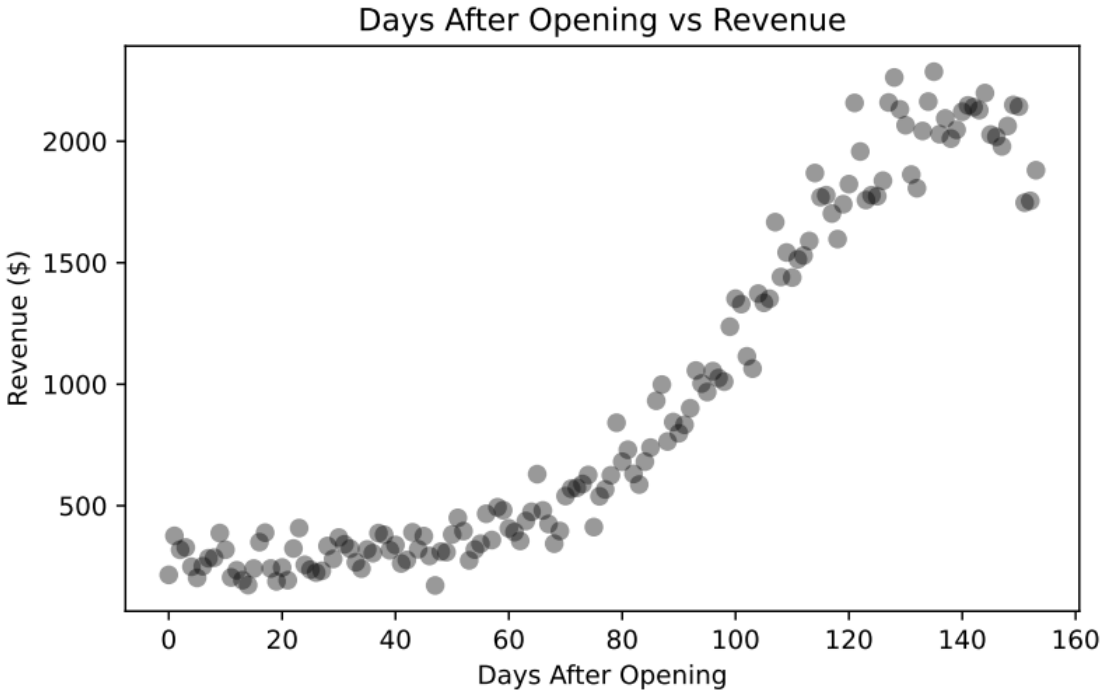

Figure 1.5. The pre-transformed plot of data.

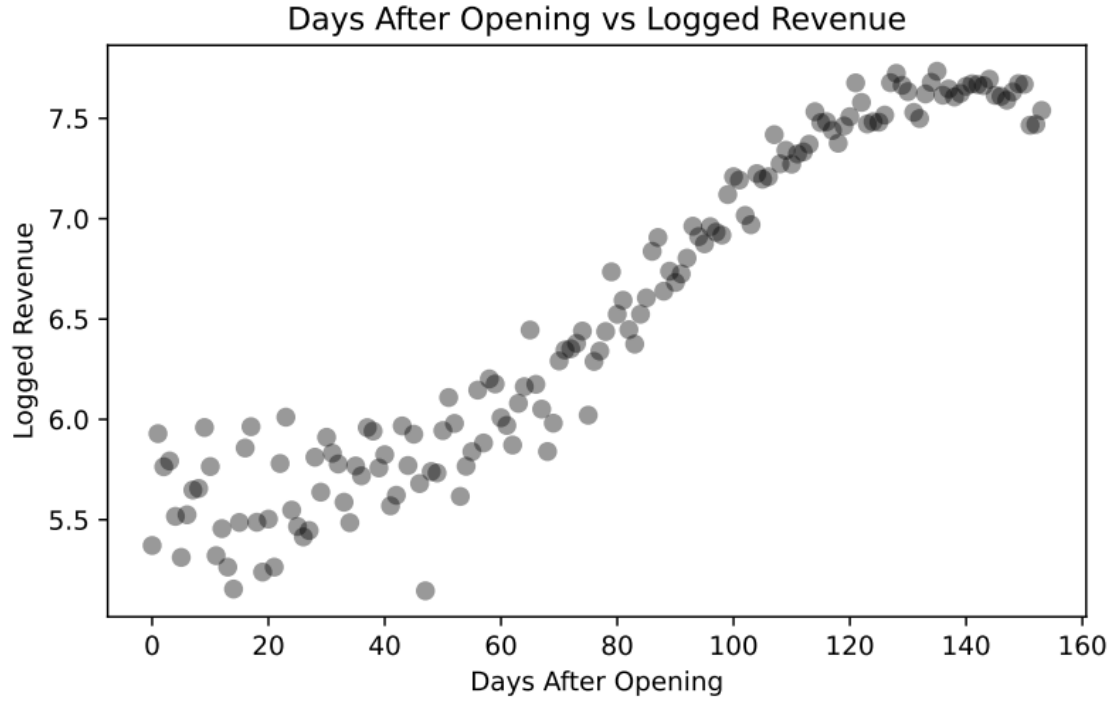

Figure 1.6. The scatter plot of data after transformation with logarithmic transformation.

3. Customer Location:

There are numerous variables in the datasets that might have a potential interaction together that produces an impact on the revenue generated by the business. Two important variables that have a high chance of displaying such interaction are the study area from which the students are coming and the rain patterns. There is a very logical reason why an interaction should exist between those two variables and that reason, is because students from study areas that are far away from the cafe can be expected to show up less when it's raining. The previous consultant has analysed the relationship between the study area and the generated revenue as shown in figure 1.7. Nonetheless, the bar chart is lacking a title, labels, and proper formatting. Figure 1.8 makes sufficient adjustments and the relationship between the variables becomes clearer. From Figure 1.8 we can clearly notice that Law Library, Fisher Library and Carslaw building are the main study areas where customers come from. Moreover, the interaction between study locations and rain patterns was studied using a clustered bar chart as shown in figure 1.9. The clustered bar chart presents some very interesting observations, for example, when it's raining, all the revenue is generated from Carslaw, Fisher Library and the Law Library, while all other study areas have 0 generations of revenue for the cafe. An interesting fact that might explain this observation is that all those study locations are linked with an underground tunnel.

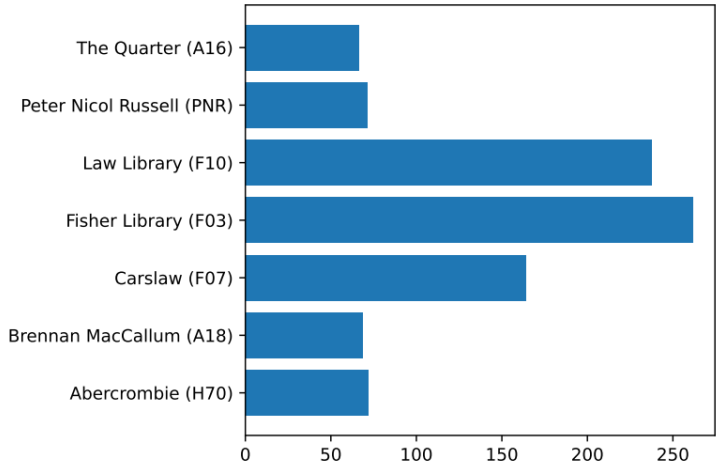

Figure 1.7. The previous consultant's horizontal bar chart.

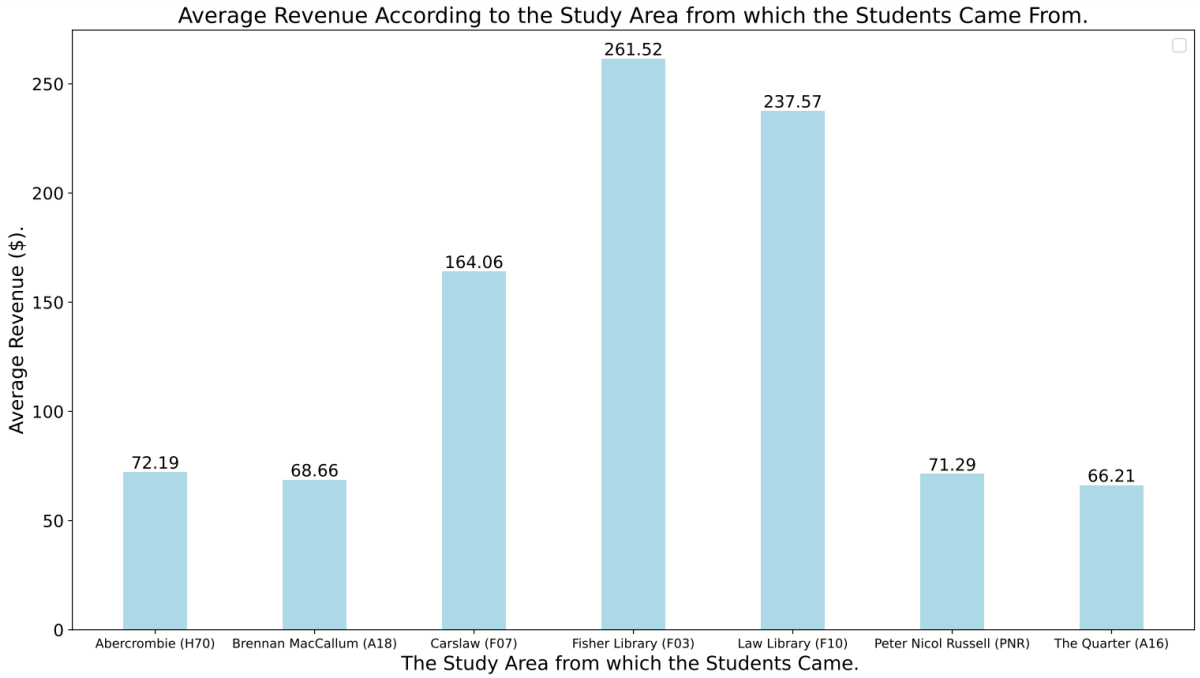

Figure 1.8. My bar chart of the average revenue vs the study area from which the customers are coming.

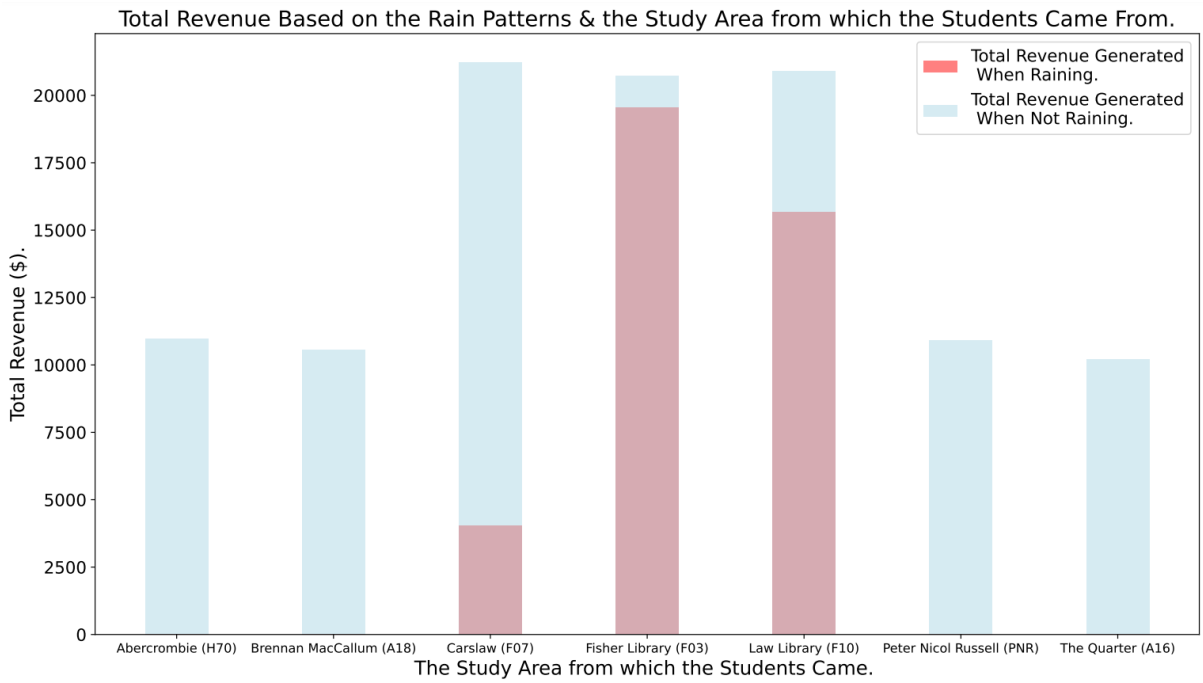

Figure 1.9. This is a clustered bar chart illustrating the total revenue based on different study areas and the impact of rain interactions on the total revenue.

Task 2:

A. Assumptions and pre-processing of variables:

Since the data set contains many variables, we have made a series of assumptions to assist us in narrowing our selection of the best variables for our regression model.
The Assumptions made are as follows:

1. We assume that the time taken to make any drink is constant regardless of the type of the drink.
2. We assume that the location can't be changed, that is, the café is fixed in place.
3. We assume that daily demand is a more important factor than hourly demand.

There are 13 variables in our dataset, 2 variables are the distinguishing or key variables, they are used to group or query other variables. Three variables are concerned with time. Four variables are concerned with the product. One variable (raining) is concerned with the weather. And finally, three variables are concerned with location.

Based on our previous assumptions, the location is no longer important in making the predictions because whatever the results are, the location is to stay fixed, and no action would be taken. Additionally, we assumed that daily demand is more important than hourly demand and so (hours_after_open) will not be of any interest to us. We also assumed that making any drink will take the same time and since revenue is not the focus of this regression model then we can remove all three variables (drink_id, name, unit_price) from our selection.

This will leave us with 4 variables that will be the focus of our regression model, those variables are (days_after_open, day_of_week, quantity, raining). Quantity is the dependent variable the other 3 variables will be the independent variables. The variables were grouped by days_after_open and day_of_week. Raining was converted to 1s for "Yes" and 0s for "No" to allow for grouping. Additionally, since many people would respond differently to rain questions on the same day, converting this to a numerical value would provide more insightful information about the true effect

of rain on demand. day_of_week variable was converted to a dummy variable with the Friday coefficient being removed to prevent collinearity.

B. Brief Interpretation of the Model.

We will be building both multiple Linear regression (MLR) and a polynomial regression model (PRM). The two models aim to predict the number of sales based on days after opening, rain patterns and day of the week, the equations are:

MLR:

$Y(\text{Quantity}) = -33.390901 + 2.465328X_1 + 1.301894X_2 + 2.374209X_3 - 10.162555X_4 - 2.425455X_5 + 12.395820X_6 + 10.066645X_7 + 28.489780X_8$

PRM:

$Y(\text{Quantity}) = 109.357449 - 5.299093X + 0.107955X^2 - 0.00036X^3$

The model equations can be interpreted by looking at the coefficients. Each coefficient could be treated like a weighing factor, the higher the coefficient is, the more important the X next to it is. Each X represents a different variable. For example, in the MLR, beta 8 is the dummy variable for day_of_week and specifically when it's a Wednesday and so we can notice that by being on a Wednesday and while fixing all other variables, it is expected that Y would increase by around 28.5.

Note: Additional EDA were conducted however, they didn't contribute to our choice of the regression model.

C. Model diagnostic.

For our model selection, we will be using MSE. This is the most common accuracy measure; we use it because it penalises large errors more heavily than small errors. Values of 0 are the best, however, for our case, we only need to select the model with an MSE closer to 0. We will also be using R^2 which is another famous measure of accuracy, and we use this measure to learn which model has captured more of the true data. Both accuracy measures indicate that PRM is better than MLR with an R^2 and Validation MSE of 0.9548 and 920.5098 respectively, compared to an R^2 of 0.9232 and an MSE of 2166.6204.

Note: For that, for PRM we used the training data to obtain R^2 while we used both training data and validation data to obtain MSE.

D. Further Interpretation of the Models:

As mentioned in section B, both MLR and PRM aim to predict sales or quantity based on other variables like the day of the week and rain patterns. When the coefficients are negative, the factor has a negative load on the demand, which means as that factor increases, the demand decreases. For example, when looking at beta 4 of MLR which corresponds to the dummy variable for days of the week and specifically Saturday, we can note that the demand decreases on this day and to be more exact, the demand is expected to decrease by 10.16 on Saturday. This observation agrees with our EDA and specifically Figure1.8 which shows that the Revenue (which is quantity times price where the price is fixed) decreases significantly on weekends.

Our MLR model shows that demand increases significantly on Wednesdays, Thursdays, and Tuesdays while it decreases significantly on the weekends. To prevent overstaffing or understaffing, it is recommended to move some employees' shifts from weekends to weekdays.
Our PRM model, on the other hand, predicts that the demand will slowdown in the future by -5.2 for every "Day After Open", therefore, it is recommended that new selling strategies are introduced and if possible, it is also recommended that the number of operating hours are increased.

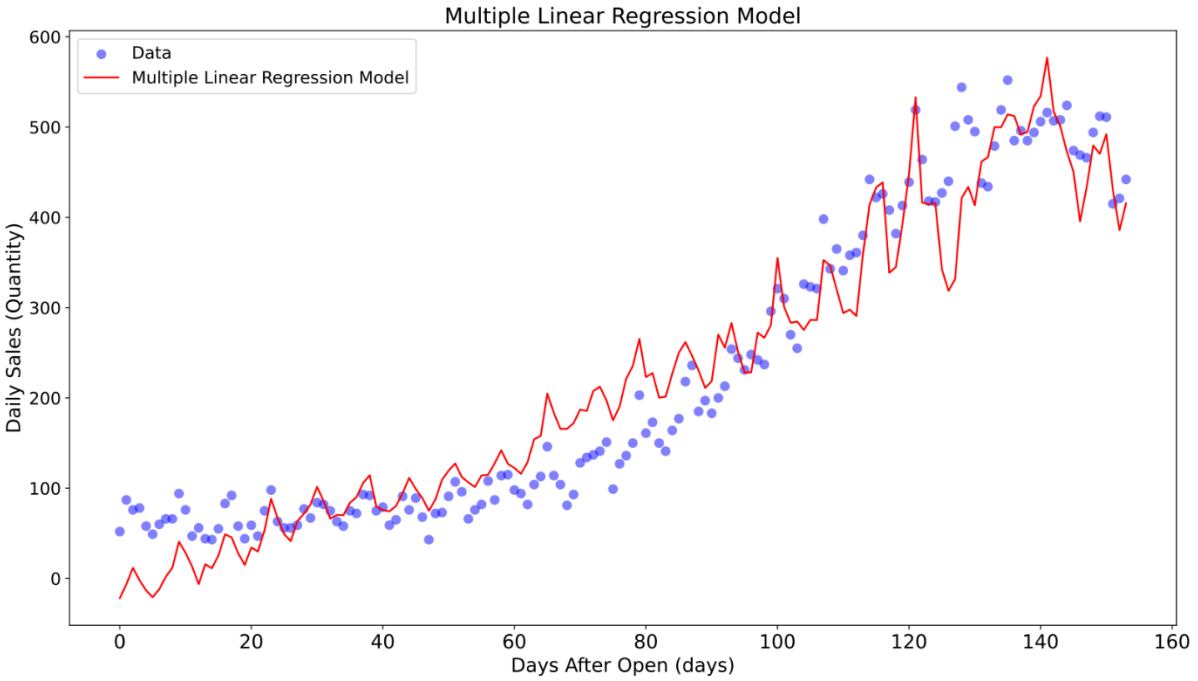

Figure 2.1. This is the Multiple Linear Regression Model (MLE).

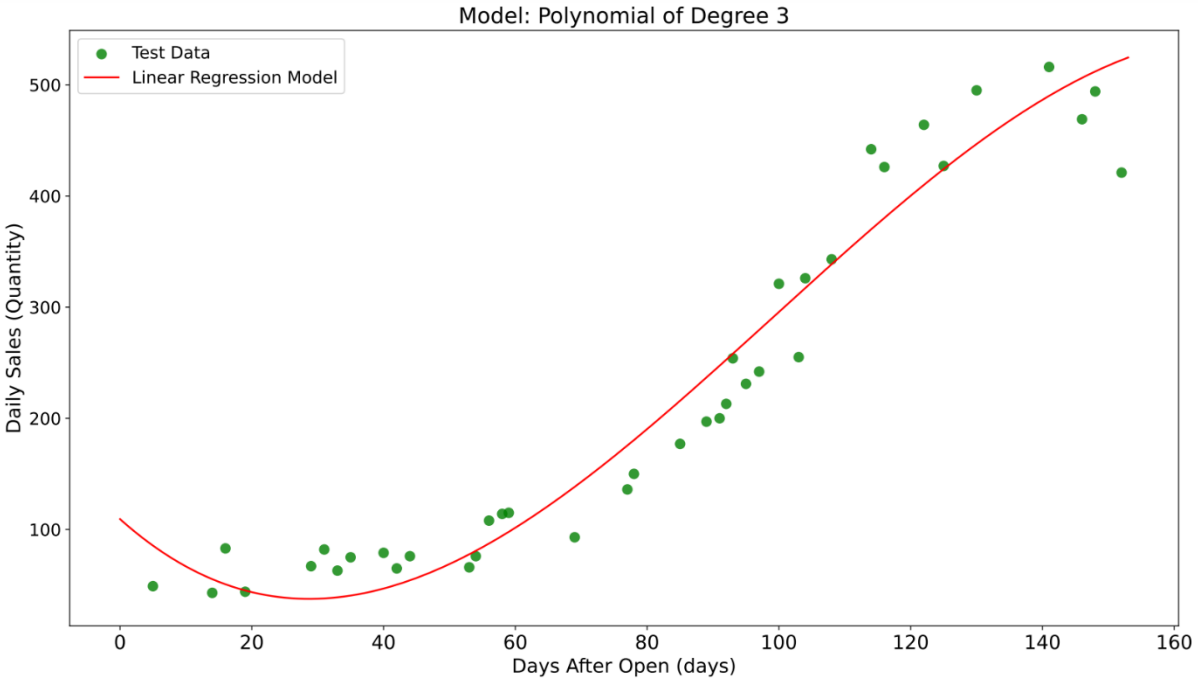

Figure 2.2. This is the Polynomial Regression Model (PRM).