TDS3301 DATA MINING

Trimester 1, 2022/2023

Group Project (30%)

Lecturer: Prof. Ts. Dr. Ting Choo Yee

Student ID	Student Name	Phone Number
1181103362	Chang See Jie	014-349 0382
1181103230	Loo Chen Zhi	011-1678 9079
1181103501	Lim Wei Jie	012-568 1547

Table of Contents

1 Exploratory Data Analysis

1.1 External Dataset

Extra data was found and incorporated into our original dataset, which is the weather data in Cyberjaya. Data Source: https://www.visualcrossing.com/weather/weather-data-services/cyberjaya/metric/2015-10-01/2016-03-31

1.2 Data Transformation

A new column is created using the *buyDrinks* variable. It is named *Drinks* and it indicates whether or not the customer purchased a drink or not. It only has two values with 1 meaning the customer bought a drink and 0 meaning the customer did not buy any drinks. Basically, any value of *buyDrinks* variable which is higher than 0 will return 1, whereas the rest will return 0 for the *Drinks* column.

In addition, binning will be done on the *Age_Range* variable to gain more insight. A new column named *Age_Bin* is created with the grouping of *Age_Range*.

1.3 Data Imbalance

A bar chart is plotted for the *Drinks* column mentioned earlier to check for imbalances. Surely enough, the bar chart shown in Figure 1.3.1 suggests that the data is imbalanced. Oversampling will be done on them in the tasks later on.

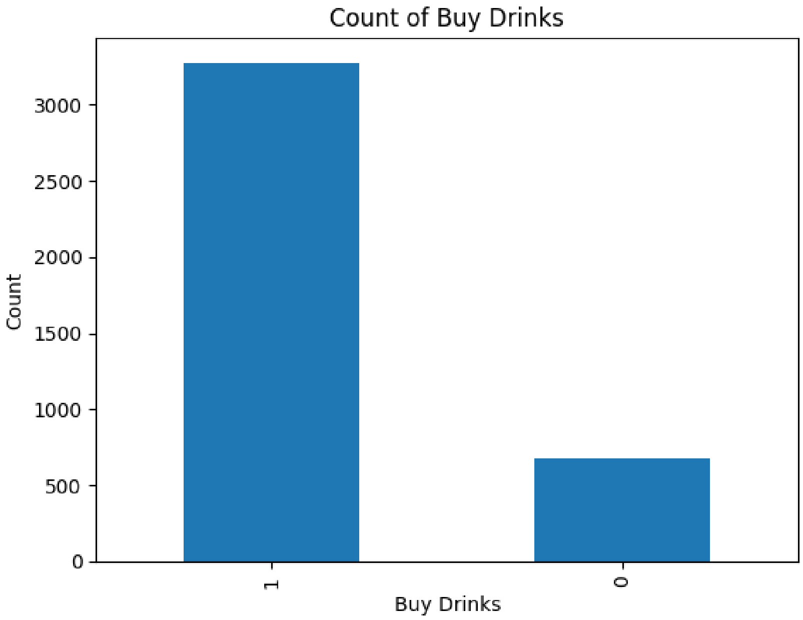


Figure 1.3.1: Count of Buy Drinks

1.4 Dealing with Missing Values

There are quite a lot of missing values in the dataset. The missing values in different columns are handled in different ways that suit the dataset as best as possible. For columns with the object data type, the missing values are imputed with ‘unknown’. For *Age_Range* and *TimeSpent_minutes*, they are replaced with the median value. The mode value will be used to replace missing values in *Num_of_Baskets* columns. Missing values are replaced with 0 for the *buyDrinks* column. Rows with missing values in *TotalSpent_RM* are dropped.

1.5 Data Visualization

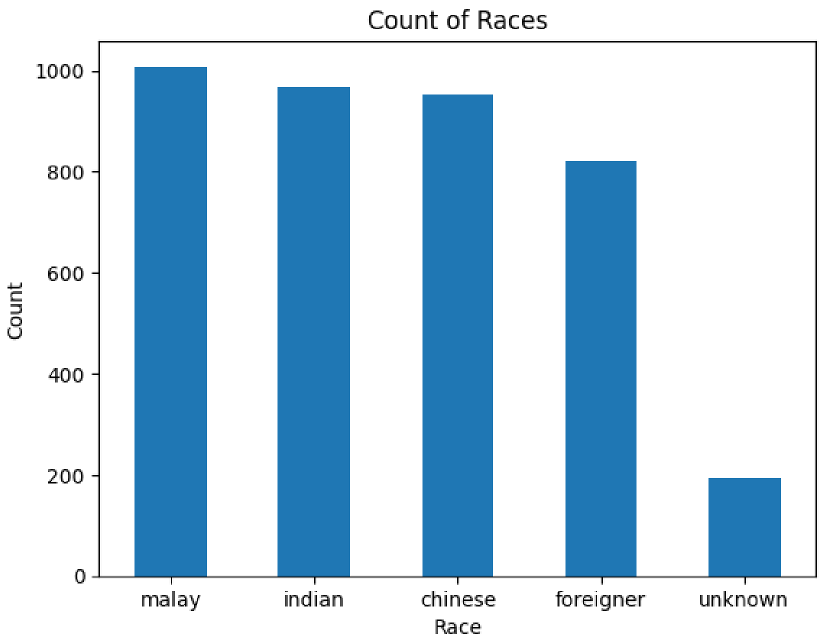


Figure 1.5.1: Count of Races

A bar chart is plotted for the *Race* variable. From it, we know that the majority of the customers are Malay, with a significant amount of Indian and Chinese customers too not falling behind.

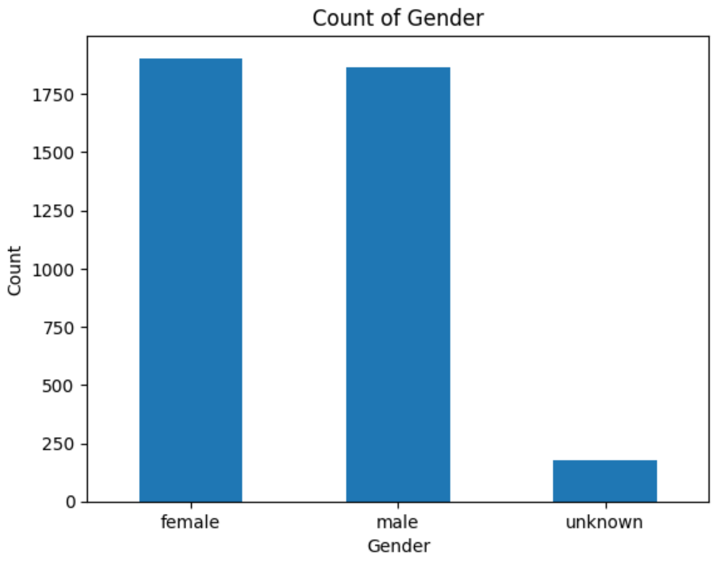


Figure 1.5.2: Count of Gender

Moreover, a bar chart for *Gender* is plotted as shown in Figure 1.5.2. The number of female customers is just slightly higher than the number of male customers.

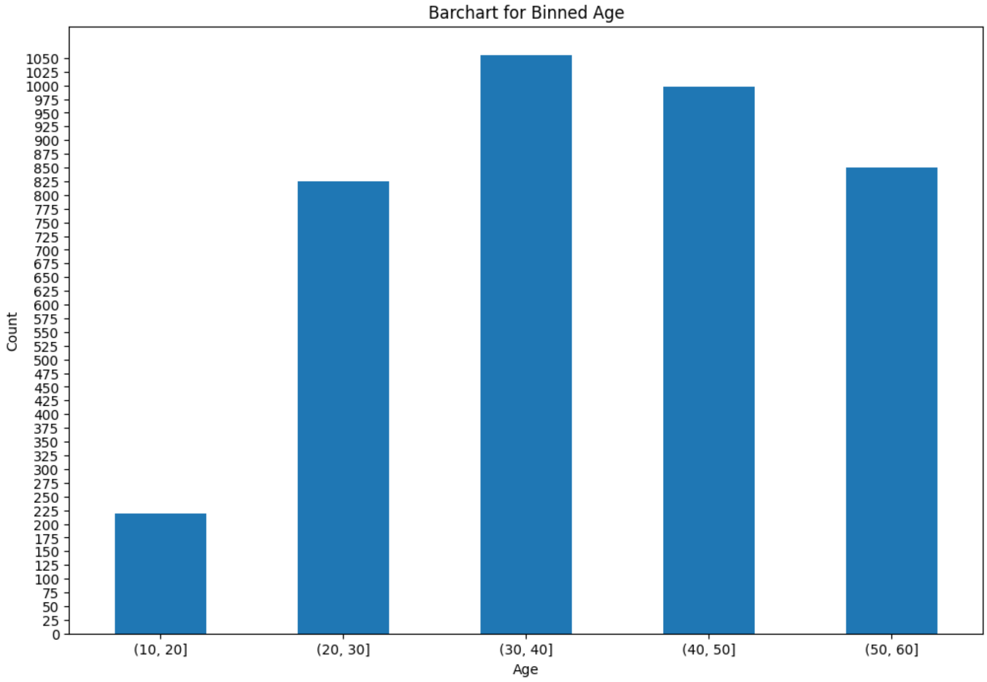


Figure 1.5.3: Barchart for Binned Age

The *Age_Rang*e variable is binned and used to plot a bar chart to gain insight. From the figure, we can determine that most of the customers are between 30 to 40 years old.

1.6 Relationship between Variables

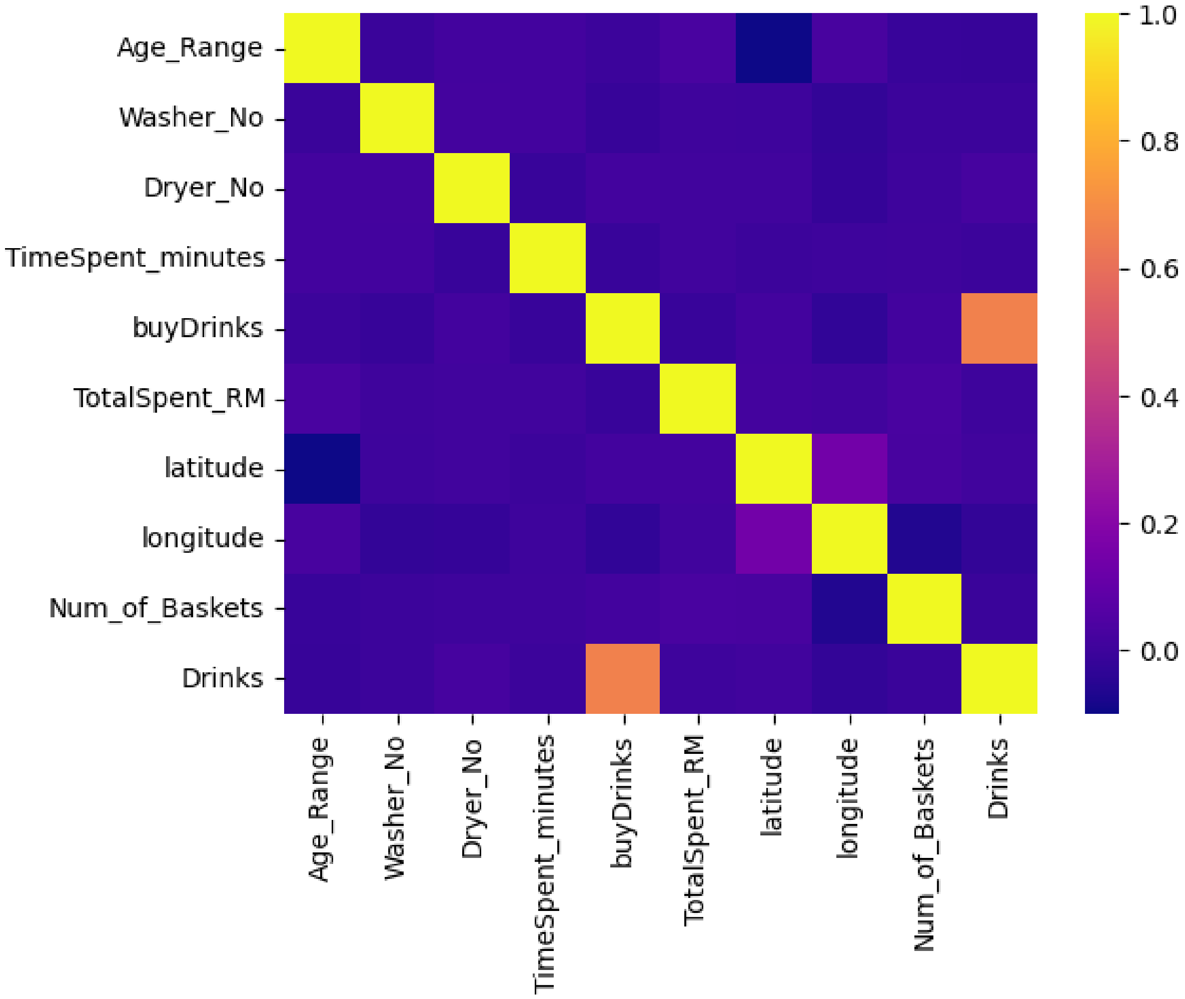
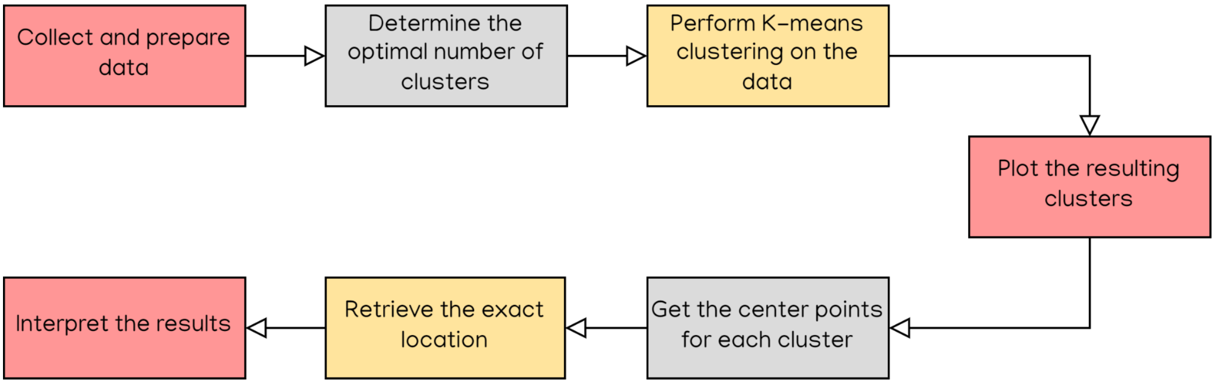


Figure 1.6.1: Correlation heatmap between the dataset variables

From the correlation heatmap above, it can be seen that there are not many relationships between the variables. Since the *Drinks* column is created using *buyDrinks*, it may seem to have a correlation between them.

2 Question 1: Where are the customers located?

2.1 Flowchart



Collect and prepare data
Determine the optimal number of clusters
Perform K-means clustering on the data
Plot the resulting clusters
Get the center points for each cluster
Retrieve the exact location
Interpret the results

Figure 2.1.1: Steps for answering Question 1

To answer this question, **K-means clustering** is performed to group the customers based on their geographic location, which is the latitude and longitude columns. The Elbow Method is used to identify the optimal k value, shown in Figure 2.1.2.

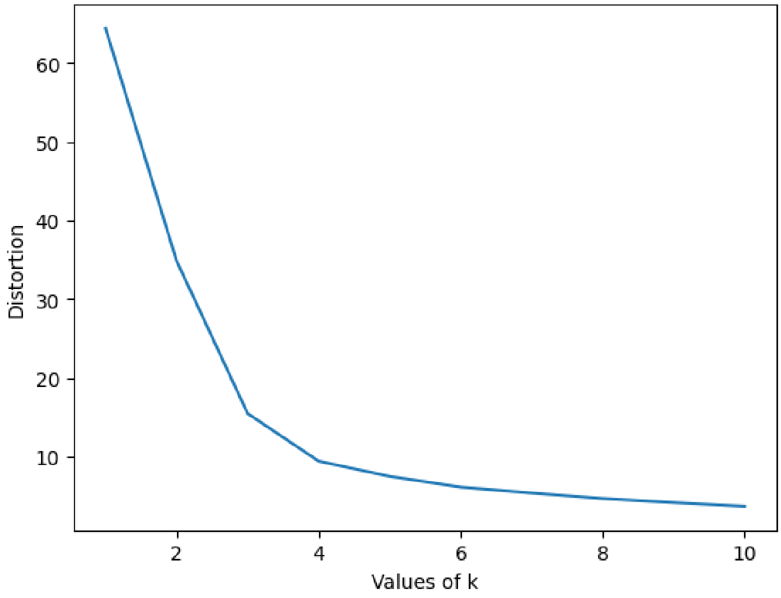


Figure 2.1.2: Determine optimal k-value by using Elbow Method

Observing the plot above, the optimal k-value is identified as 4 from the elbow point, where distortion starts to decrease at a slower rate. Therefore, K-means clustering is performed with a k-value of 4 in Figure.

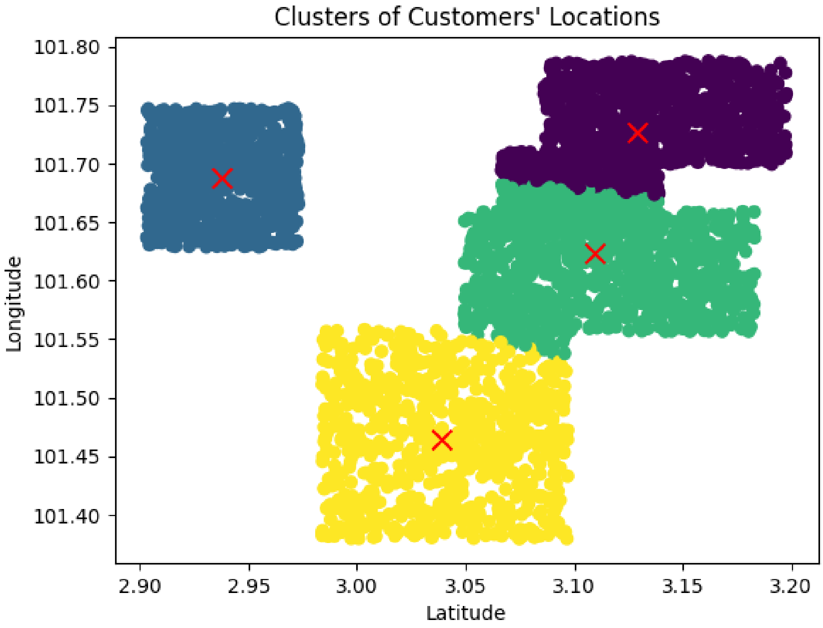


Figure 2.1.3: Clusters of Customers' Locations

Figure 2.1.3 shows the clustering of the customers' locations. The locations have been grouped into 4 clusters, with an 'X' symbol at the centre of each cluster to represent the centre points. After that, the exact locations are retrieved from the centre point:

- Petaling Jaya: 3.1097539467184703, 101.6235128716843
- Putrajaya: 2.9378713243802816, 101.68811912202305
- Kuala Lumpur: 3.129212255976744, 101.72651149009475
- Majlis Perbandaran Klang: 3.03908829348, 101.46461657832258

2.2 Insight

To use this to answer the question, it can be said that the customers are majorly located in Kuala Lumpur, Putrajaya, Petaling Jaya and Majlis Perbandaran Klang.

3 Question 2: Are there any frequent patterns that occur between the various attributes of the customers and the laundry machine usage?

3.1 Flowchart

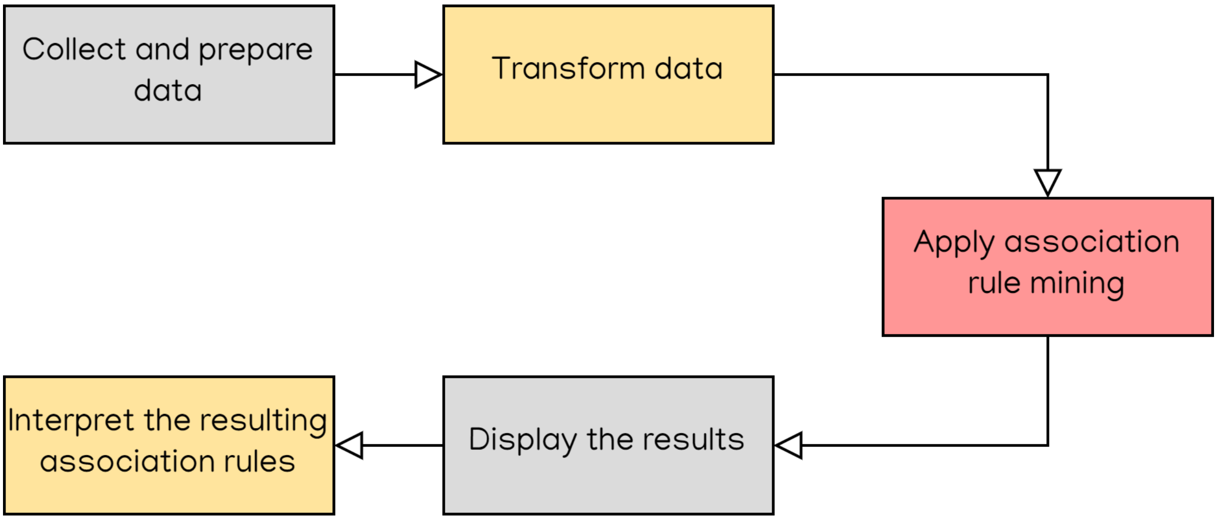

Collect and prepare data
Transform data
Apply association rule mining
Display the results
Interpret the resulting association rules

Figure 3.1.1: Steps for answering Question 2

For the question, the variables *Race, With_Kids, Kids_Category, Basket_Size, Washer_No, Dryer_No, TimeSpent_minutes, buyDrinks, TotalSpent_RM and Num_of_Baskets* will be used to perform **Association Rule Mining** using the Apriori algorithm. In the Apriori function, the 'min_support' parameter is set to 0.009, which means that the itemset must appear at least 0.9% of the dataset to be considered frequent. The 'min_confidence' is set to 0.2, which means that a rule must have the confidence of at least 20% to be considered interesting. The 'min_lift' is set to 2. The output with top 3 lift is as follows:

- (Rule 7) Basket_Size:big -> Num_of_Baskets:1.0, Support: 1.4%, Confidence: 24.89%, Lift: 2.2268
- (Rule 8) Basket_Size:big -> buyDrinks:0.0, Support: 1%, Confidence: 65%, Lift: 2.1608
- (Rule 6) Basket_Size:big -> Washer_No:3, Support: 1.6%, Confidence: 28.57%, Lift: 2.1153

3.2 Insight

From the output, it can be concluded that customers who use big baskets tend to only use 1 basket, are not likely to purchase drinks and usually use the washer no.3.

4 Question 3: Will a customer purchase drinks in the laundry shop?

4.1 Flowchart

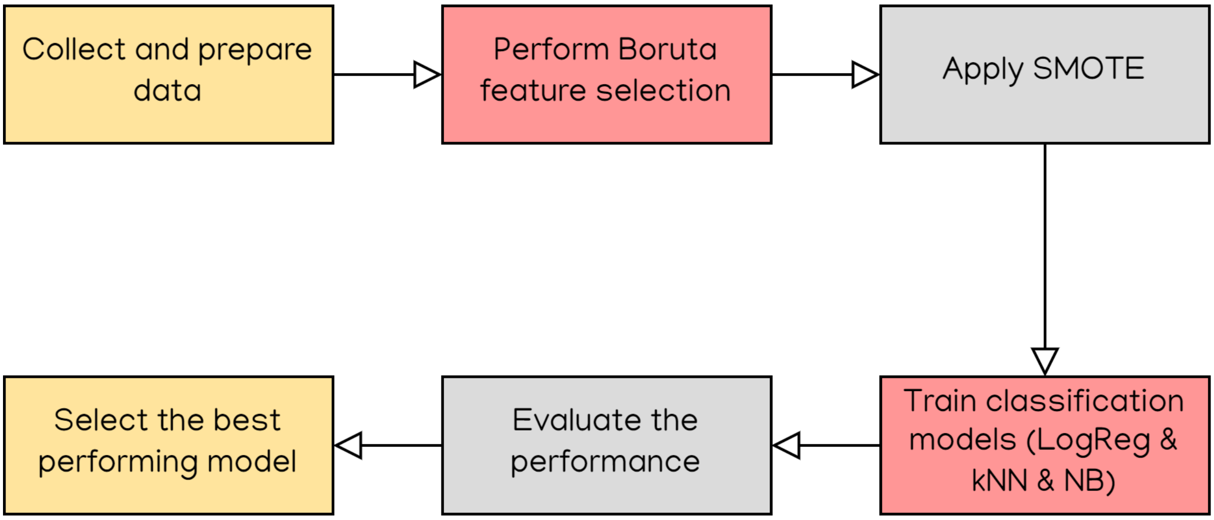

Collect and prepare data
Perform Boruta feature selection
Apply SMOTE
Train classification models (LogReg & kNN & NB)
Evaluate the performance
Select the best performing model

Figure 4.1.1: Steps for answering Question 3

In this question, the *Drinks* variable will be predicted to know whether or not the customers will buy drinks, where yes=1 and no=0. Different classification models, which are **Logistic Regression**, **k-Nearest Neighbors** and **Naive Bayes**, are used to predict the target variable, *Drinks*. Before starting, unnecessary columns such as *Date, Time, latitude* and *longitude* are dropped since they don't contribute to the prediction.

Then, **Boruta** is used to performing feature selection to identify the most important features from the dataset.

As mentioned earlier, the *Drinks* variable is imbalanced, so oversampling has to be applied by using **SMOTE**. After the top 10 features are selected, SMOTE is applied. Then, the dataset is trained using the three classification models and a comparison is done. The presence of SMOTE will also be compared to see if SMOTE made a significant difference.

For each model, **Accuracy**, **Confusion Matrix**, **Area Under the Curve (AUC)** and **Precision-Recall** will be used to evaluate the performance.

4.2 Without SMOTE

First, we will do the model evaluation for the dataset without performing SMOTE.

	Logistic Regression	**k-NN**	**Naive Bayes**
Accuracy	0.804	0.804	0.787
Confusion Matrix	Majority TN = 0 Majority FP = 155 Majority FN = 0 Majority TP = 635	Majority TN = 0 Majority FP = 155 Majority FN = 0 Majority TP = 635	Majority TN = 13 Majority FP = 142 Majority FN = 26 Majority TP = 609
AUC	0.56	0.54	0.55
Precision-Recall	0.83	0.84	0.82

Table 4.2.1: Summary of Regression models scores without SMOTE

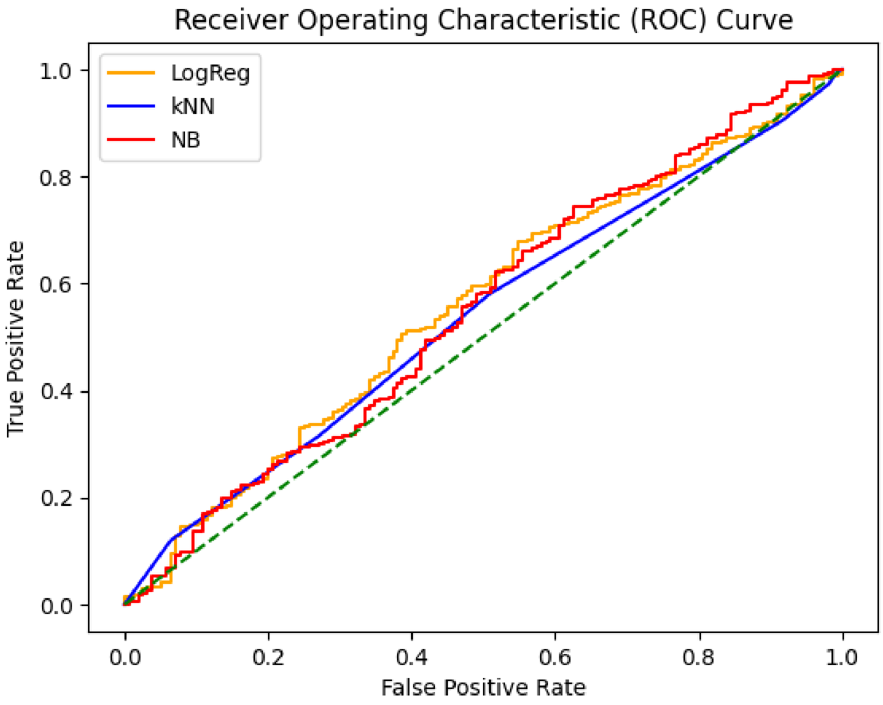


Figure 4.2.1: Receiver Operating Characteristics (ROC) Curve for three models without SMOTE

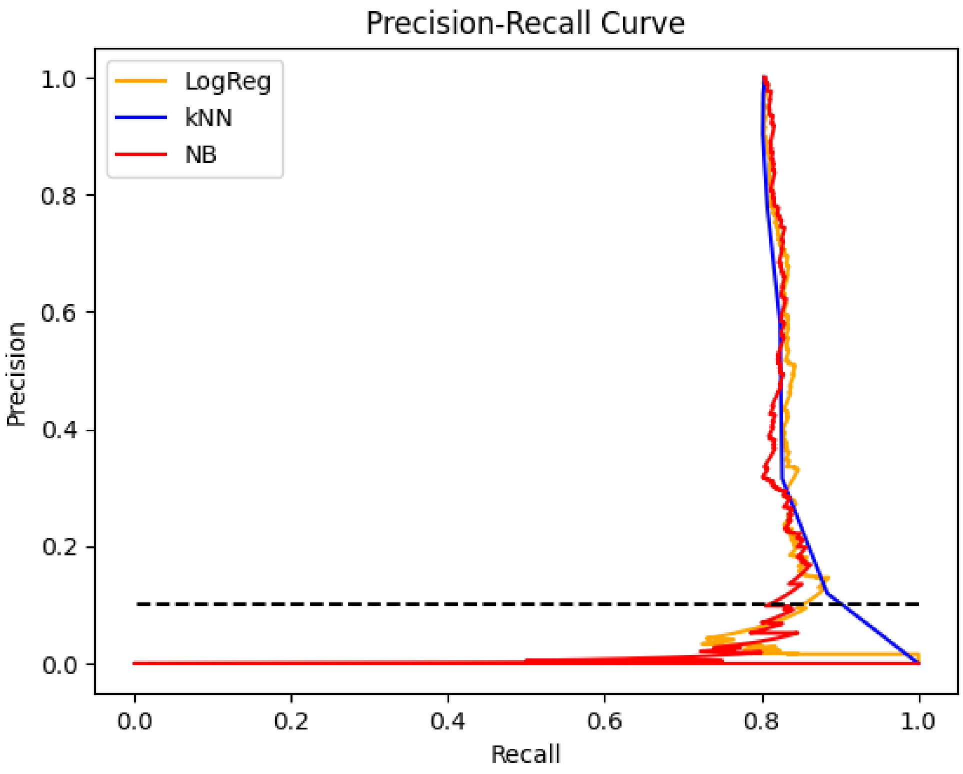


Figure 4.2.2: Precision-Recall Curve for three models without SMOTE

The above shows the Receiver Operating Characteristics (ROC) Curve and Precision-Recall Curve for models without SMOTE. From the results shown above, all three models are having similar accuracy scores and similar AUC scores. The accuracy of all three models is relatively similar, which is around 0.787 to 0.804, while the AUC scores are around 0.54 to 0.56, and precision_recall scores are around 0.02 to 0.84.

The logistic regression model has the highest accuracy at around 0.804, the highest AUC score at 0.56 and the highest precision recall at 0.83.

4.3 With SMOTE

	Logistic Regression	**k-NN**	**Naive Bayes**
Accuracy	0.570	0.552	0.504
Confusion Matrix	Majority TN = 78 Majority FP = 77 Majority FN = 263 Majority TP = 372	Majority TN = 64 Majority FP = 91 Majority FN = 263 Majority TP = 372	Majority TN = 87 Majority FP = 68 Majority FN = 324 Majority TP = 311
AUC	0.56	0.51	0.53
Precision-Recall	0.83	0.82	0.81

Table 4.3.1: Summary of Regression models scores with SMOTE

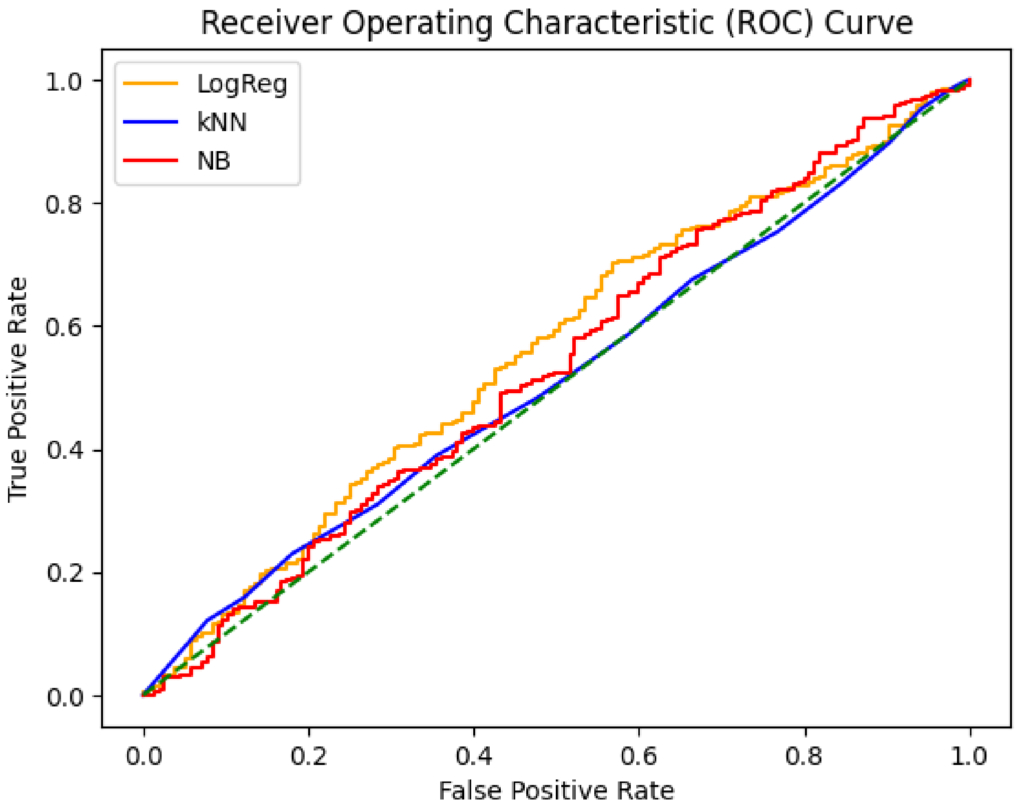


Figure 4.3.1: Receiver Operating Characteristics (ROC) Curve for three models with SMOTE

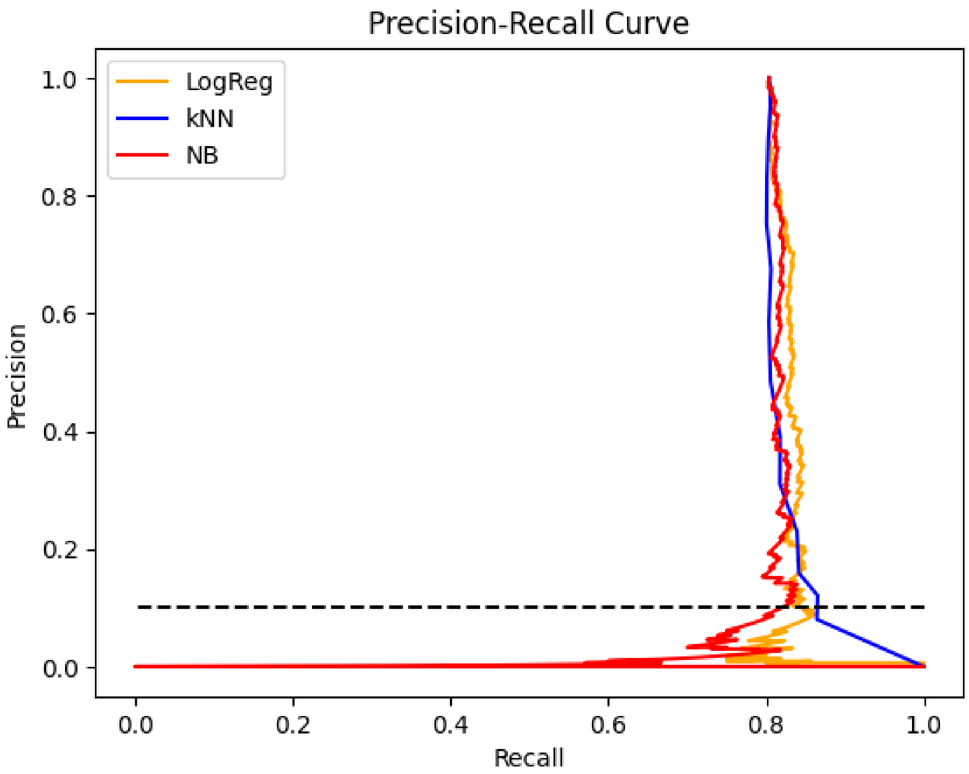


Figure 4.3.2: Precision-Recall Curve for three models with SMOTE

Above shows the Receiver Operating Characteristics (ROC) Curve and Precision-Recall Curve for models with SMOTE. Based on the results, it seems that the Logistic Regression performed the best with an AUC of 0.57 and a Precision-Recall of 0.83 when using SMOTE to balance the target variable.

4.4 Insight

To answer the question in this case, it can be said that Logistic Regression has a better performance than k-Nearest Neighbors and Naive Bayes. However, the conclusion can be made that all the models have relatively low accuracy and AUC scores, meaning that they may not be useful for this dataset. Thus, it is not easy to predict whether a customer will purchase drinks or not relying on these models.

5 Question 4: What is the relationship between the weather conditions and the number of customers at the laundry shop?

5.1 Flowchart

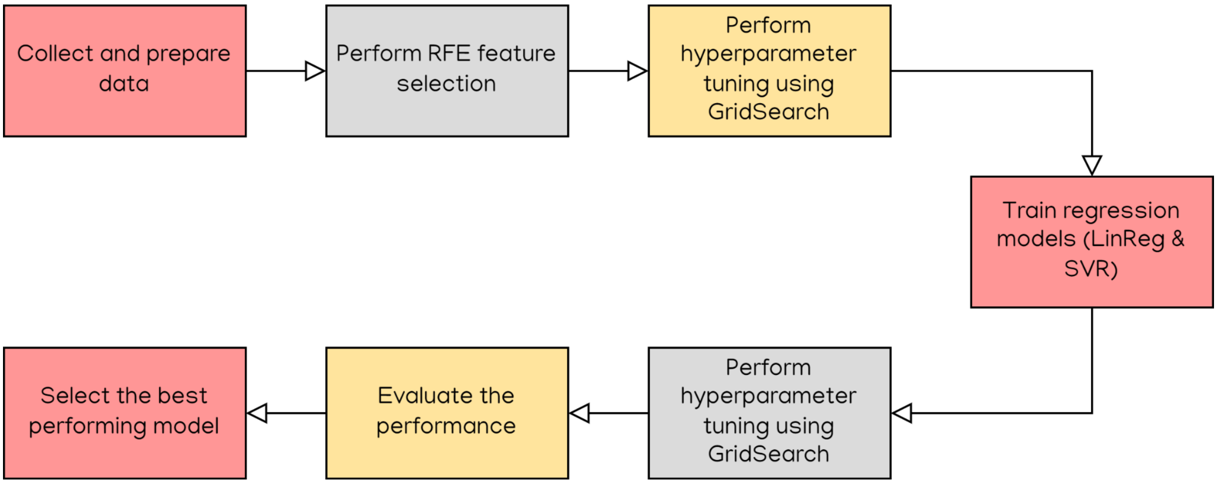

Collect and prepare data
Perform RFE feature selection
Perform hyperparameter tuning using GridSearch
Train regression models (LinReg & SVR)
Perform hyperparameter tuning using GridSearch
Evaluate the performance
Select the best performing model

Figure 5.1.1: Steps for answering Question 4

For this question, **Linear Regression** and **Support Vector Regression** are used to predict the number of customers based on the weather conditions using the external weather data. Firstly, only columns related to the weather are used and they are grouped by *date*. A column indicating the number of customers is also created. Next, one-hot encoding is done to convert categorical variables into numerical variables.

Recursive Feature Elimination with Cross-Validation (RFECV) is used as a feature selection technique to select the most relevant features for the predictive models. **Hyperparameter tuning is done using GridSearch** to get the optimal parameters for the RFECV. After that, the top and bottom 10 features are discovered. The top 5 features are used for the Linear Regression and Support Vector Regression to do prediction. Hyperparameter tuning using GridSearch is done on both models to be compared with models without tuning. The **Mean Absolute Error (MAE)** and **R-Squared (R2)** are used to evaluate the models.

Linear Regression:

	Before Tuning	**After Tuning**
MAE	9.149	9.348
R2	0.082	0.014

Table 5.1.1: Summary of Linear Regression scores before and after hyperparameter tuning

SVR:

	Before Tuning	**After Tuning**
MAE	9.384	9.385
R2	-0.024	-0.006

Table 5.1.2: Summary of SVR scores before and after hyperparameter tuning

5.2 Insight

To answer the question, both regression models are poorly performing. Both models are not suitable for predicting the number of customers based on the weather conditions. Therefore, we can conclude that the relationship between weather and the number of customers is weak.

6 Deployment

Streamlit Cloud: https://wjlim-14-dm-streamlit-dm-streamlit-iiizyk.streamlit.app/