TDS3301 Data Mining
Project

Lecturer:
Dr. Ting Choo Yee

QUESTION 1: Profiling Customers in a Self-Service Coin Laundry Shop

PREPARED BY :

Tan Kuang Chen
Chan Jun Yang
Sashrreik Menon

1.0 Question

1. Which is the top washer_no and dryer_no used by the customer?
2. When is the most popular period of a day that the customer visited the most?
3. Which day in a week has the highest sales based on the number of customers?
4. What are the peak hours that customers visit the self-service laundry shop?
5. What is the distribution of gender based on customer race?
6. What is the distribution of the basket size based on basket colour?
7. Which washer and dryer were often used together?
8. Did daily rainfall information and day impact the sales of the laundry shop?
9. What kind of relationships are there between all the attributes?
10. What are the Top-10 and Bottom-10 features based on 'parts_of_day' attributes?
11. What are the Top-10 and Bottom-10 features based on 'Basket_Size' attributes?
12. How well performed are the classification models in predicting attributes called 'parts_of_day' and "Basket_Size?
13. How well performed are the regression models in predicting attributes called 'Number_Customers'?
14. How many clusters are there in the dataset?

2.0 Data Acquisition

With the aid of the lecturer, we are provided with a dataset named LaundryData.csv that contains 19 attributes showing customers appearance and behaviour in a self-service coin laundry shop. Luckily, by referring to data.gov.my. (2019) we have also found a supplementary dataset that contains some information about the daily rainfall amount for each state from 2014 until 2020. Below are both dataset information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 807 entries, 0 to 806
Data columns (total 20 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   No             807 non-null    int64
 1   Date           807 non-null    object
 2   Time           807 non-null    object
 3   Race           797 non-null    object
 4   Gender         793 non-null    object
 5   Body_Size      790 non-null    object
 6   Age_Range      799 non-null    float64
 7   With_Kids      794 non-null    object
 8   Kids_Category  777 non-null    object
 9   Basket_Size    801 non-null    object
 10  Basket_colour  798 non-null    object
 11  Attire         776 non-null    object
 12  Shirt_Colour   798 non-null    object
 13  shirt_type     770 non-null    object
 14  Pants_Colour   802 non-null    object
 15  pants_type     798 non-null    object
 16  Wash_Item      784 non-null    object
 17  Washer_No      807 non-null    int64
 18  Dryer_No       807 non-null    int64
 19  Spectacles     807 non-null    object
dtypes: float64(1), int64(3), object(16)
memory usage: 126.2+ KB
```

```
1  df_rain.info() #rainfall dataset
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30684 entries, 0 to 30683
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   State          30684 non-null  object
 1   Year           30684 non-null  int64
 2   Month          30684 non-null  int64
 3   Day            30684 non-null  int64
 4   Rainfall (mm)  30684 non-null  float64
dtypes: float64(1), int64(3), object(1)
memory usage: 1.2+ MB
```

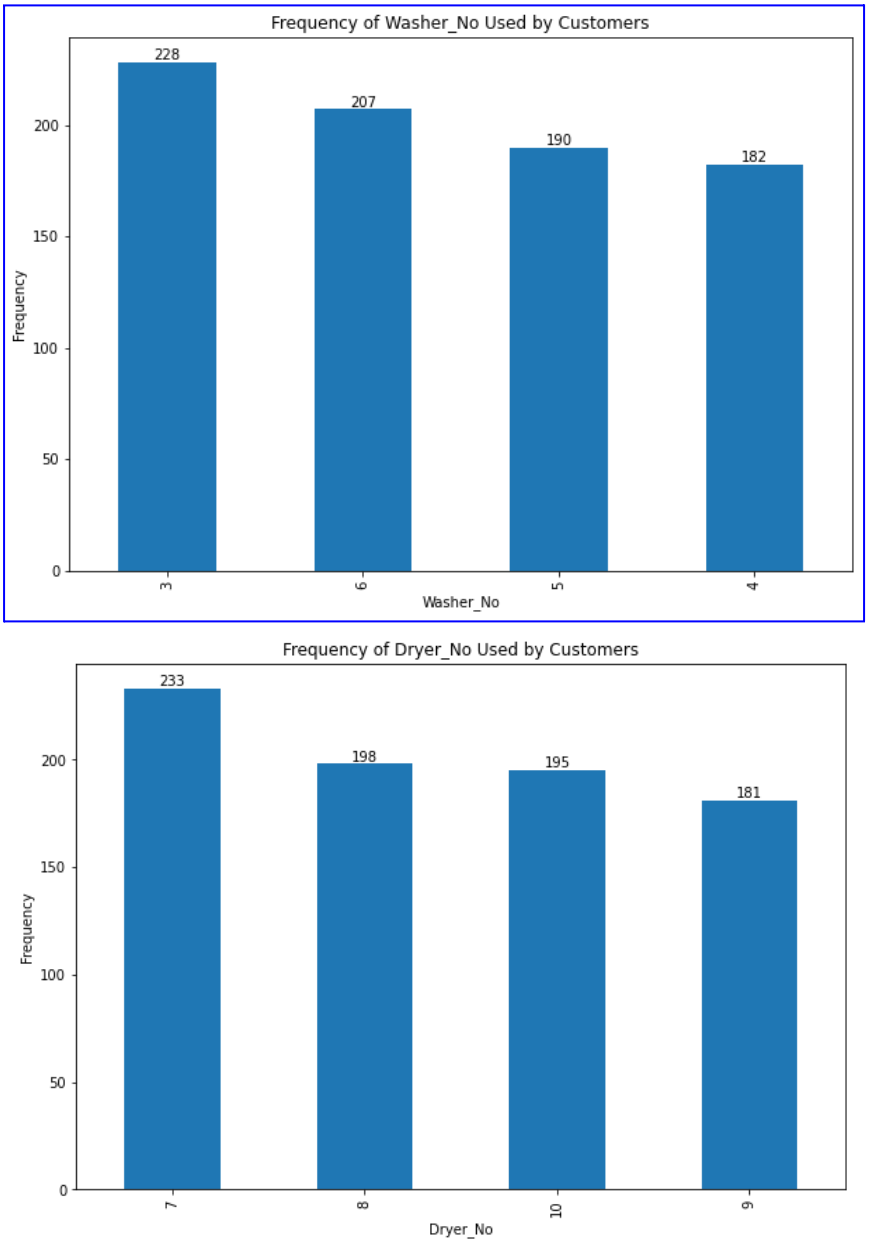
2.1 Data Preprocessing

For the data processing procedure, we discovered over 200 rows with missing values. To avoid reducing data, we performed a data cleaning procedure that replaced all missing data with mode values for all object type attributes. While the mean value is used to replace any missing values for integer and float types attributes. The reason for replacing mode values for all object type attributes with missing data is

because we are unable to get the mean value. While for the integer and float types attributes replaced with mean values is to avoid an unbalanced range of data.

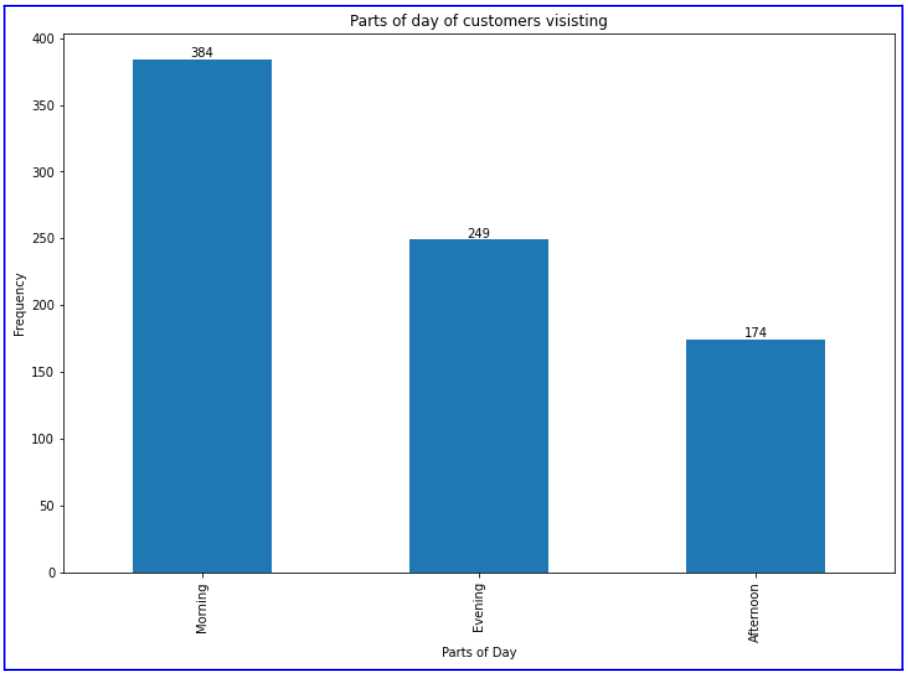
2.2 Exploratory Data Analysis

Question 1: Which is the top washer_no and dryer_no used by the customer?



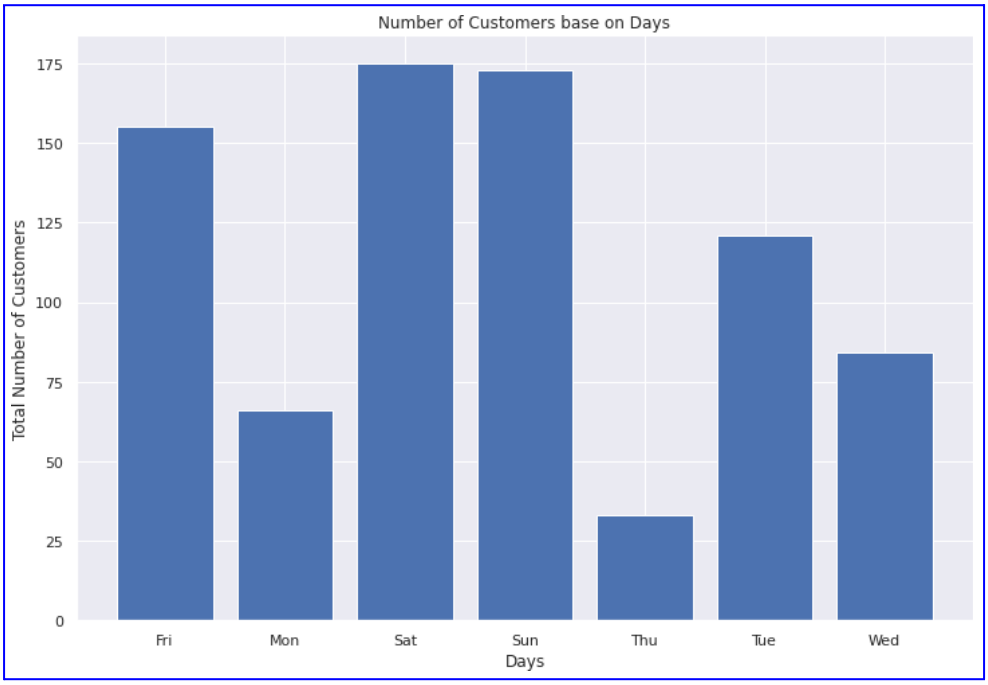
Figures above show the frequency of Washer_No and Dryer_No used by customers. Results show the Washer_No 3 appeared as the most popular washer used by the customers with 228 users records. While Dryer_No 7 is the most commonly used dryer among all the dryers with 233 users records.

Question 2: When is the most popular period of a day that the customer visited the most?



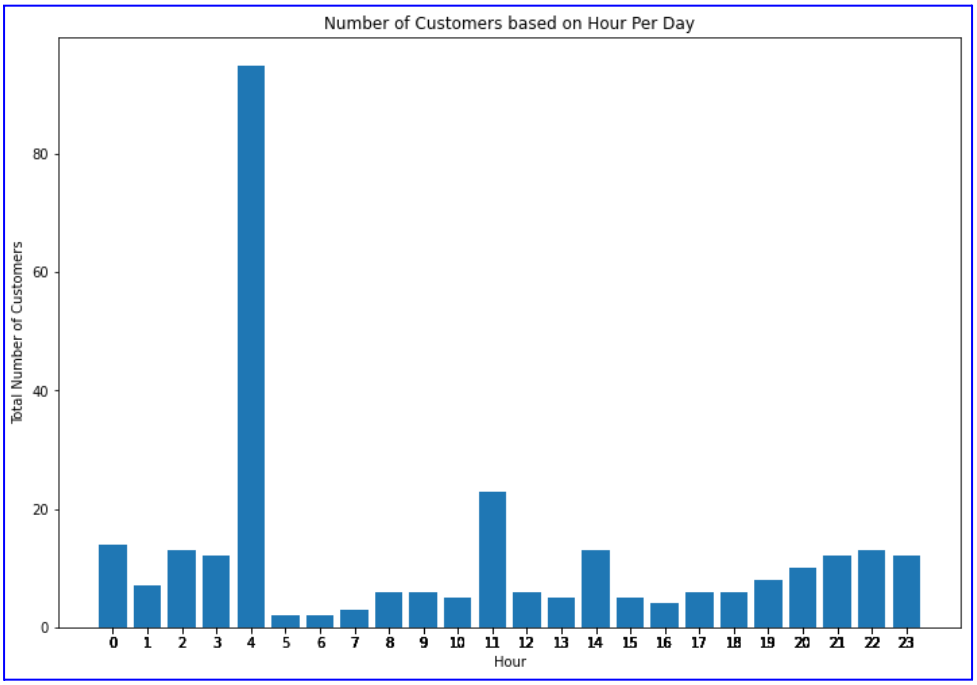
We have added a new column called 'parts_of_day' where we have converted the time of customer arrival from the dataset into different parts of the day as the figure above which are morning (00:00 - 11:59), afternoon (12:00 - 17:59) and evening (18:00 - 23.59). As we can see from the figure above, most of the customers visited the laundry shop during the morning, which is from midnight to the morning.

Question 3: Which day in a week has the highest sales based on the number of customers?



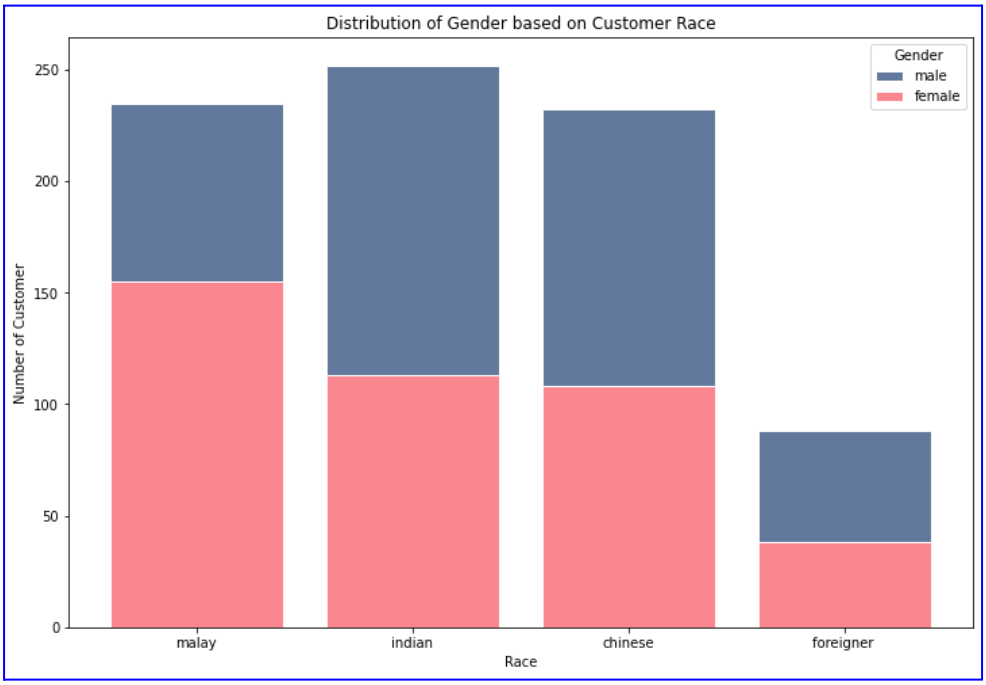
Due to our lack of skills in programming, we could not convert the date into the day of the week where we manually inserted the day by using python code. The days of the week are separated as Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday. As we can see from the bar chart above, Saturday and Sunday are having the most customers in a week. We believe that this is because most of the people are not working during the weekend which is why they could use their free time to visit the laundry shop and do their laundry.

Question 4: What are the peak hours that customers visit the self-service laundry shop?



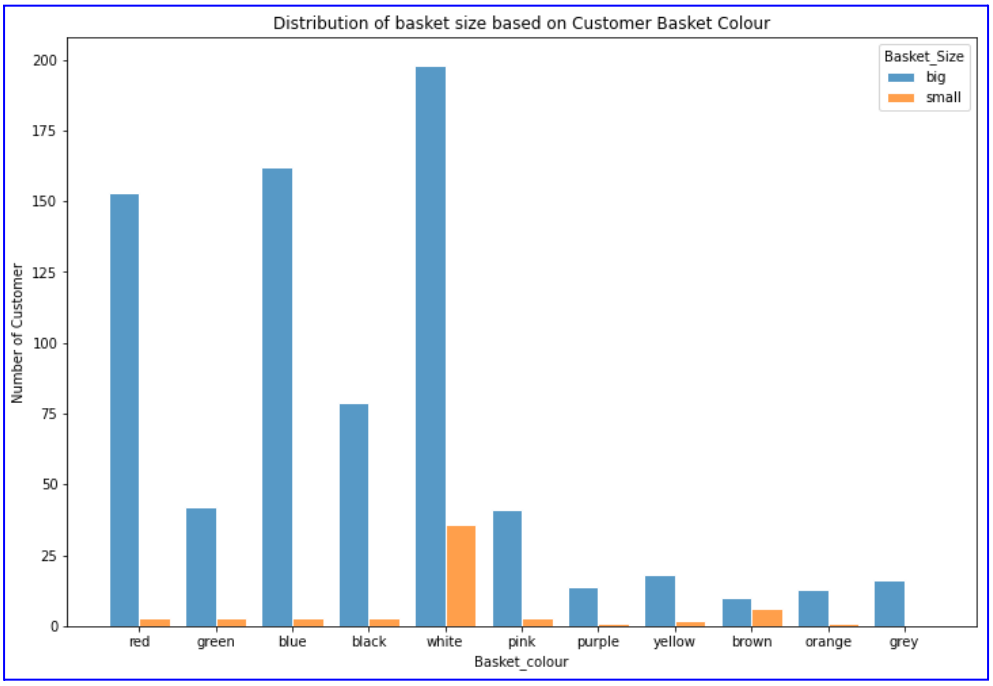
Based on the figure, shows the total number of customers visiting the self-service laundry shop every hour from October 2015 to December 2015. Surprisingly, most of the customers visited the laundry shop at 04:00 in the early morning with a record of over 80 customers while 05:00 in the early morning had the fewest customers visiting the laundry shop. There are more customers visiting the laundry shop during the nighttime compared to the late morning as we think that most of the customers work during the daytime.

Question 5: What is the distribution of gender based on customer race?



The figure above shows the distribution of gender based on customer race. Indians visited the most frequently compared to other races which recorded about 250 customers within the two months. Chinese and Malay have an approximately equal number of customers visiting the laundry shop with about 230 customers. However, there were less than 100 foreign customers who visited the laundry shop. Overall, the number of male and female customers received a fair distribution across all races but the Malay female customer recorded more visitors than the Malay male customer.

Question 6: What is the distribution of the basket size based on basket colour?



The figure above shows the distribution of basket size based on basket colour. In general, a high amount of customers use big baskets and white colour baskets are widely used by the customers. However, the least amount of customers are using the brown big basket and grey small basket which record the lowest number among all kinds of basket. Hence, we can see that it is heavily unbalanced in this case where there is a large amount of customers using big size baskets.

2.3 Association Rule Mining Algorithm

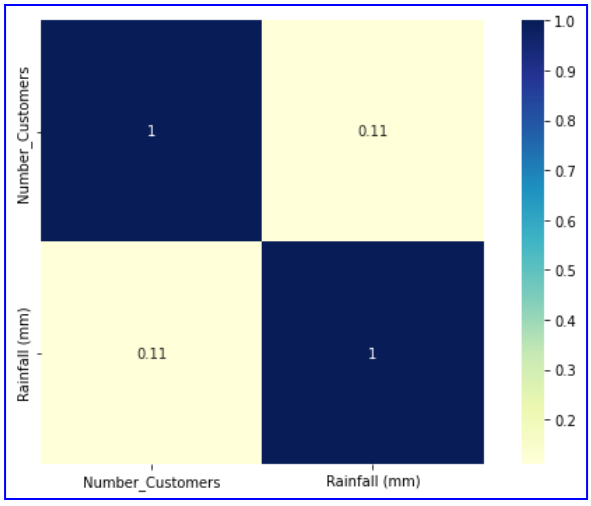
Question 7: Which washer and dryer were often used together?

```
(Rule 1) 7 -> 3
Support: 0.112
Confidence: 0.3947
Lift: 1.3672
====================================
```

We used the apriori algorithm to determine which washer and dryer are most often used together by customers. We have set the threshold of minimum support and minimum confidence to 0.1 and 0.3 to determine the association between different washers. The results show that washer number 3 and dryer number 7 received much attention from the customers with a 0.112 support ratio, 0.3947 confidence ratio and 1.3672 lift score. Therefore this result shows a total sense whereby referring to question 1, the result analysed showed the most used washer and dryer by the customer were washer 3 and dryer 7.

2.4 Correlation Matrix

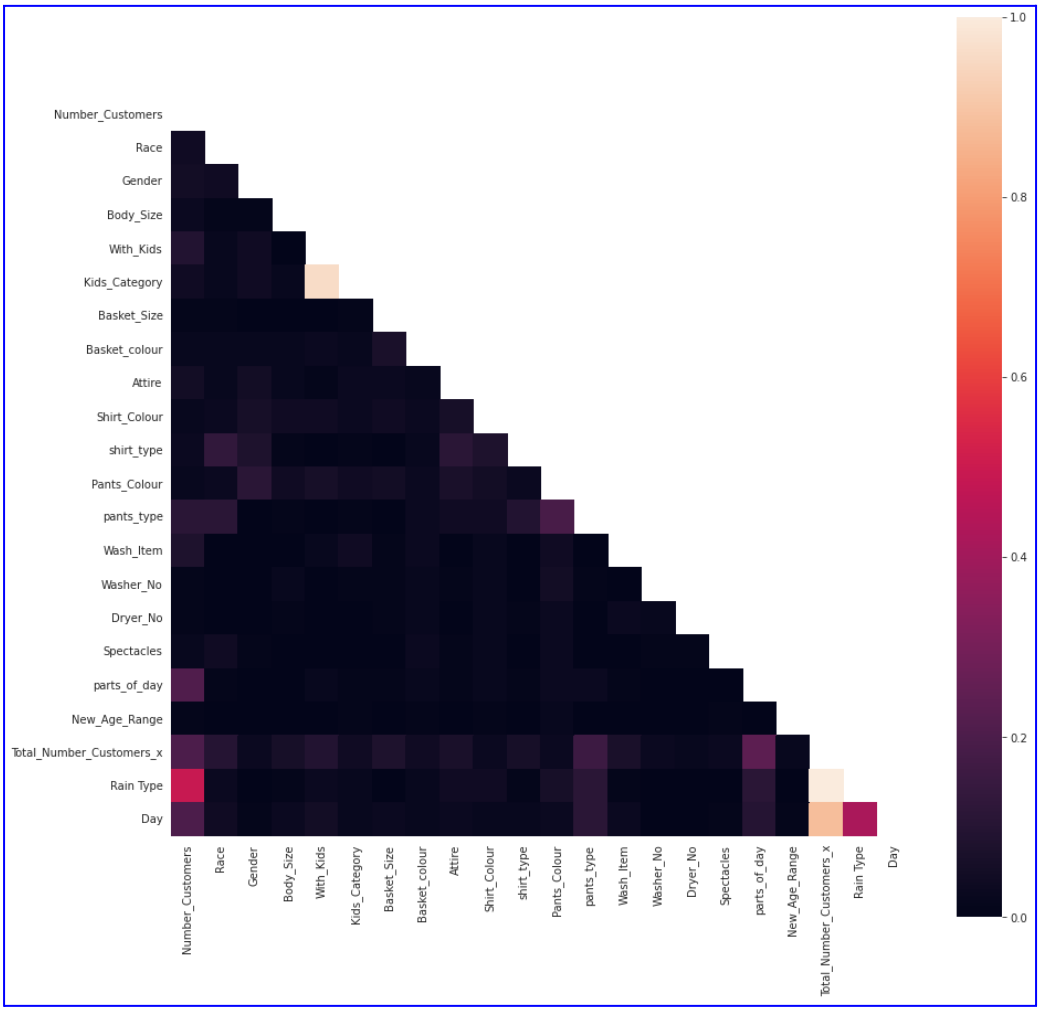
Question 8: Did daily rainfall information impact the sales of the laundry shop?



We have collected the rainfall dataset of 2015 from data.gov.my. (2019) that provided the rainfall information of all states in Malaysia from 2014 to 2020. Hence, we are able to find the relationship between the sales of the laundry shop and the daily rainfall of Selangor. Firstly, we determine the sales of laundry based on the number of customers of the laundry shop according to the dates provided. By merging both the dataset and the dataframe in question 4 which is total customer based on day, we used a correlation matrix to find the relationship between the attributes found. First, we use the correlation

matrix to find the relationship between rainfall per day(mm) and the number of customers per day. Based on the figure above, we can conclude that there is a weak positive relationship between both variables where the rainfall is only having a minor impact on the sales of the laundry shop.

Question 9: What kind of relationships are there between the attributes?



After merging the dataset available, we have used Cramer's V correlation matrix to find the relationship between all the attributes in the dataset used. According to Baidu (2008), we have split the 'rainfall(mm)' attribute into different rain types such as light rain, moderate rain and heavy rain. The code of correlation matrix was referred from chrisbss1. (2019, September 25) because it is suitable for finding the correlation between the categorical attributes. Based on the figure above, most of the attributes show a very low ratio value of correlation. This concludes that most of the attributes are independent variables which mean there is no relationship between the attributes. However, Kid_Category and With_Kids received the highest correlation score among all the correlations with a 1.0 ratio because Kid_Category can only be determined by With_Kids in this scenario. We can also see that the Rain type and Day are having a correlation score higher than 0.8 with the Number of customers where we believe that both the attributes will affect the number of customers in a day.

2.5.1 Feature Selection (parts_of_day)

Question 10: What are the Top-10 and Bottom-10 features based on 'parts_of_day' attributes?

---------Top 10----------

	Features	Score
0	Number_Customers	1.00
1	Race	1.00
19	Rain Type	1.00
18	Total_Number_Customers_x	1.00
12	pants_type	1.00
11	Pants_Colour	1.00
20	Day	1.00
5	Kids_Category	1.00
9	Shirt_Colour	0.92
4	With_Kids	0.92

---------Bottom 10---------

	Features	Score
8	Attire	0.75
10	shirt_type	0.67
13	Wash_Item	0.58
14	Washer_No	0.50
6	Basket_Size	0.42
3	Body_Size	0.33
17	New_Age_Range	0.25
15	Dryer_No	0.17
2	Gender	0.08
16	Spectacles	0.00

For the feature selection, we used the package called Boruta to predict a class to maximize performance by removing irrelevant features. We used all the attributes that score above 0.5 in the Boruta feature selection model to proceed with our classification model. Based on the figures Basket_Size, Body_Size, New_Age_Range, Dryer_No, Gender and Spectacles attributes will not be used in the prediction model.

2.5.2 Feature Selection (Basket_Size)

Question 11: What are the Top-10 and Bottom-10 features based on 'Basket_Size' attributes?

---------Top 10----------

	Features	Score
20	Day	1.00
6	Basket_colour	1.00
18	Total_Number_Customers_x	1.00
7	Attire	1.00
0	Number_Customers	0.94
10	Pants_Colour	0.94
13	Washer_No	0.88
16	parts_of_day	0.81
1	Race	0.75
12	Wash_Item	0.69

---------Bottom 10---------

	Features	Score
8	Shirt_Colour	0.56
3	Body_Size	0.50
5	Kids_Category	0.44
9	shirt_type	0.38
17	New_Age_Range	0.38
15	Spectacles	0.25
11	pants_type	0.19
4	With_Kids	0.12
2	Gender	0.06
19	Rain Type	0.00

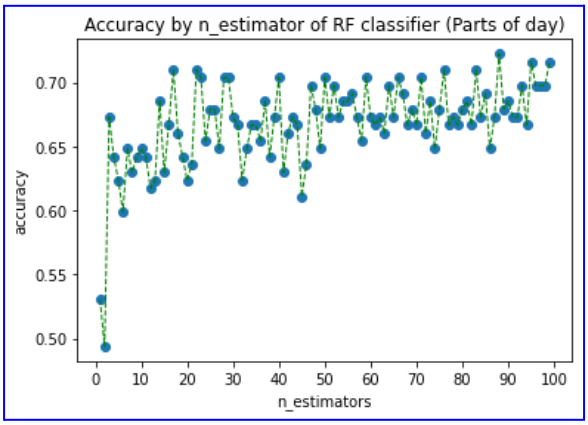
Again Boruta was used to predict the 'Basket_Size' class and all the attributes that score above 0.5 in the model will proceed with our classification model. Based on the figures, Kids_Category, shirt_type, New_Age_Range, Spectacles, pants_type, With_Kids, Gender and Rain Type attributes will not be used in the prediction model.

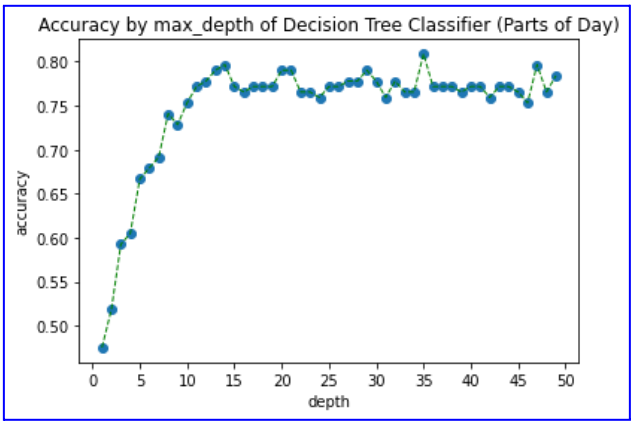
3.1 Classification

Question 12: How well performed are the classification models in predicting attributes called ‘parts_of_day’ and ‘Basket_Size’?

For the prediction section, we are going to predict two types of classes which are ‘parts_of_day’ and ‘Basket_Size’. First of all, the reason we predict ‘parts_of_day’ attributes is to define what type of customers would like to visit the self-service laundry shop in which part of the day. Next, ‘Basket_Size’ attributes are also predicted in order to find out what size of the basket and what type of customer is likely to bring the big size of the basket to the laundry shop.

3.1.1 Random Forest and Decision Tree Classifier (parts_of_day)





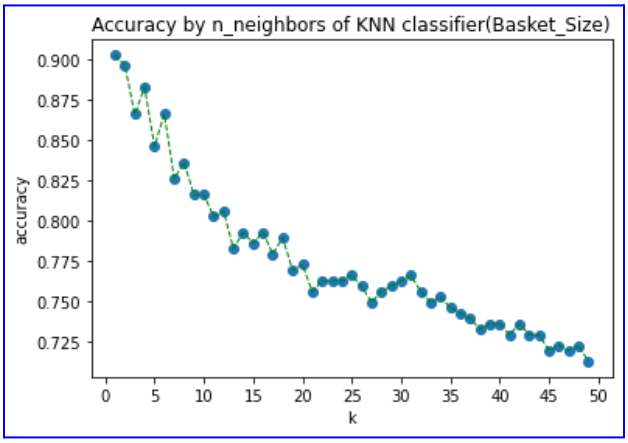
We employed two distinct classifiers for the 'parts of day' attribute prediction which are random forest (RF) and decision tree (DT) classifiers. As we can see, the DT classifier outperformed the RF classifier in terms of accuracy and consistency at different depths. The DT classifier received a peak accuracy of 0.8 and a constant accuracy of about 0.75-0.8, whereas the RF classifier had a peak accuracy of 0.72 and a steady accuracy of around 0.6-0.72. Hence, DT is better at predicting the class.

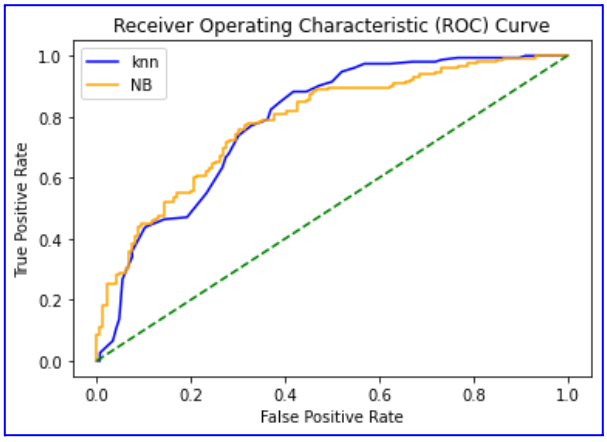
3.1.2 Naive Bayes and K Nearest Neighbour Classifier (Basket_Size)

Naive Bayes (NB) Precision Score:

score: 0.7290969899665551

	precision	recall	f1-score	support
0	0.74	0.68	0.71	146
1	0.72	0.77	0.74	153
accuracy			0.73	299
macro avg	0.73	0.73	0.73	299
weighted avg	0.73	0.73	0.73	299





The findings in question 6 shows the basket size is heavily unbalanced where there is a large amount of big size basket used by the customer. Hence, an oversampling method has been used to balance out the distribution of big and small baskets. Naive Bayes (NB) and K Nearest Neighbour (KNN) classifiers were used to predict the 'Basket Size' attribute. The KNN classifier outperformed the NB classifier in terms of accuracy, with an accuracy ratio of 0.85 for the KNN classifier and 0.73 for the NB classifier. We are unable to compare the model evaluation section since the NB classifier is unable to do tuning. However, we did calculate the AUC and plotted the ROC Curve for both classifiers to compare the classifiers. To determine whether classifier performance is superior, we must first determine which classifier curves are higher or closer to 1. As we can see KNN classifier ROC curve and AUC score is somewhat higher than the NB classifier, where the KNN classifier AUC score is 0.79 and the NB classifier is 0.78. Thus, KNN is better at predicting the class.

3.2 Regression

Question 13: How well performed are the regression models in predicting attributes called ‘Number_Customers’?

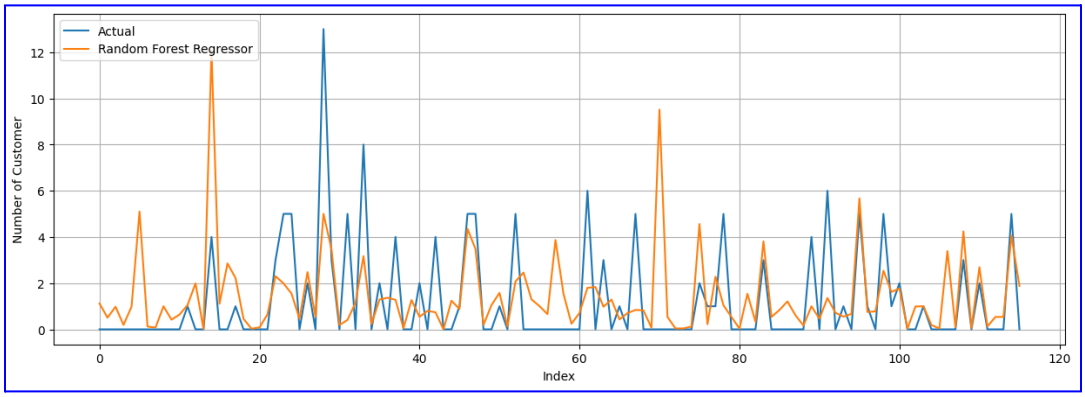
In this section, we are going to use the three different regression models which are RF, DT and Lasso Regression to do prediction on the attribute called ‘Number_Customers’. The reason for doing the regression process is to predict the number of customers by every hour of a day.

3.2.1 Random Forest, Decision Tree and Lasso Regression

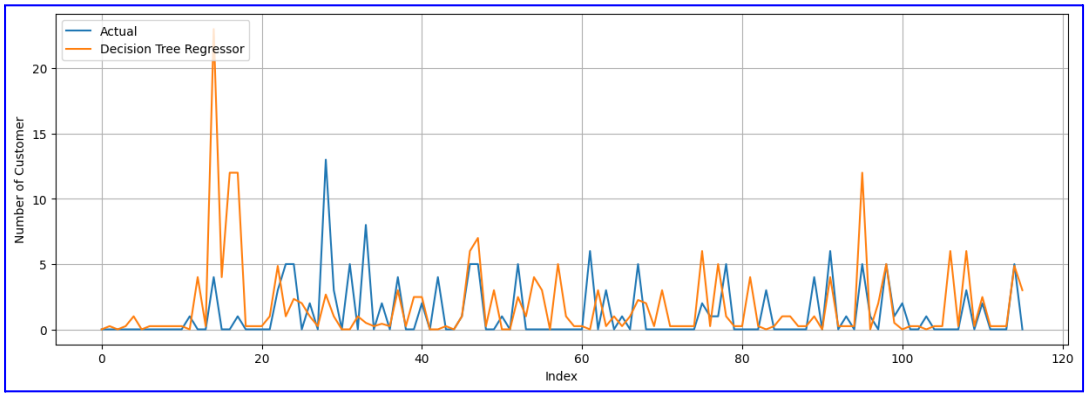
	Number_Customers	num_cus-1	num_cus-2	num_cus-3	num_cus-4	num_cus-5	num_cus-6	num_cus-7	num_cus-8	num_cus-9	num_cus-10	num_cus-11	num_cus-12	num_cus-13	num_cus-14
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
19	1.0	1.0	4.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
20	6.0	1.0	1.0	4.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
21	0.0	6.0	1.0	1.0	4.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
22	0.0	0.0	6.0	1.0	1.0	4.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
23	0.0	0.0	0.0	6.0	1.0	1.0	4.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0

576 rows × 25 columns

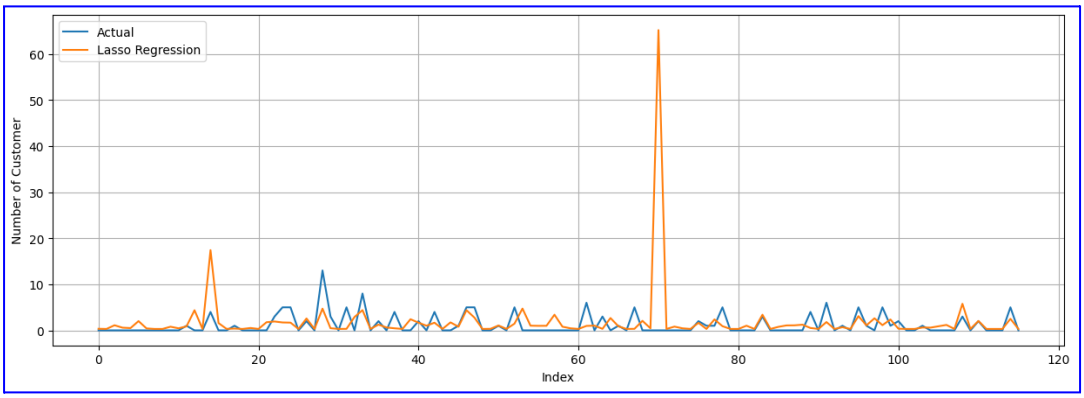
Random Forest Regression:



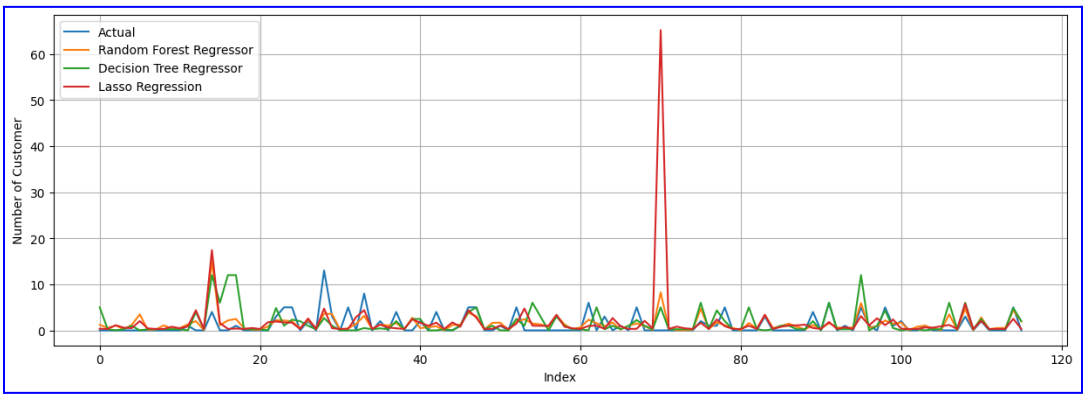
Decision Tree Regression:



Lasso Regression:



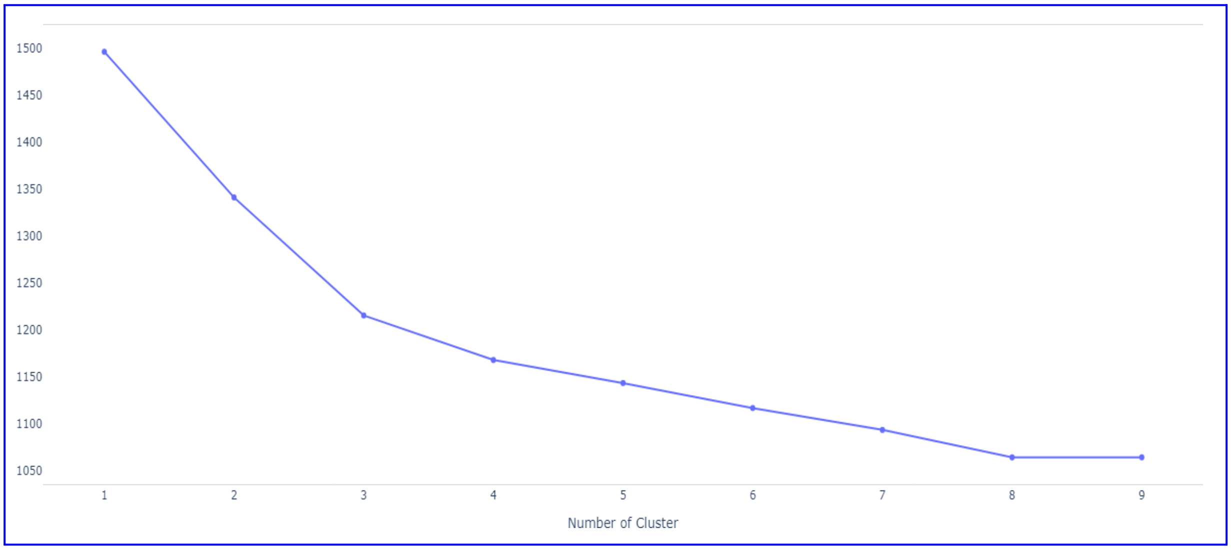
Regression Models Evaluation:



We have used the previous 24 hours of the number of customers as the features of our regression model to predict the upcoming number of customers visiting the laundry shop. The regressor used were RF, DT and Lasso Regression where each of them respectively scored a mean squared error with 2.22, 3.45, 6.47. As we can see from the figure above, RF and DT regression predicted values were roughly the same which is why the mean squared error from both the models are similar. Lasso regression performed the worst out of the 3 regression models here which had a significantly high mean squared error compared with another 2 models. It also has the worst result on a graph where the predicted values it gets is having a high similarity with the actual values.

3.3 Clustering Techniques

Question 14: How many clusters are there in the dataset?



As for the clustering technique, we have used K-Prototype to find the clusters available in our dataset. Due to the dataset available having different types of variables such as numerical and categorical, K-means and K-modes were not suitable for clustering the dataset. Hence, K-Prototype was used on clustering the variables because it is able to deal with categorical and numerical data at the same time. As we can see from the scatter plot above, the elbow criterion occurred at the number cluster of 3. Hence, we can conclude there are 3 different clusters in total for the attributes used.

Deployment

We have hosted the our visualization result in heroku website and the link is shown below:
https://tds3301-data-mining-project.herokuapp.com/

Reference

[1] data.gov.my. (2019, October 24). Daily Projected Rainfall for Scenario MRIB1 by State in Peninsular Malaysia - 2014–2020 : Daily Projected Rainfall (MRIB1) by State in Peninsular Malaysia - MAMPU. Retrieved from https://www.data.gov.my/data/ms_MY/dataset/daily-projected-rainfall-for-scenario-mrib1-by-state-in-peninsular-malaysia/resource/d9f7f4fd-6b22-4669-9402-60090e54242c

[2] Baidu. (2008, May 30). 涓嬮洦鐨勫欐姘撮噺澶氬皯澶氬皯mm鏄. Retrieved from https://zhidao.baidu.com/question/55313793.html

[3] chrisbss1. (2019, September 25). Cramer's V correlation matrix. Kaggle. Retrieved from October 25, 2021, from https://www.kaggle.com/chrisbss1/cramer-s-v-correlation-matrix.