MULTIMEDIA UNIVERSITY

FACULTY OF COMPUTING AND INFORMATICS

TDS 2201 DATA MINING

TRIMESTER 2210

Project

Data Mining in Self-Service Laundry Shop

Prepared by:

Ng Kong Jun, 1181202889, 0165016893

Lai Yong Siang, 1211300514, 0124275212

Chin Wei Song, 1191100961, 0127300644

Yoong Yu Hong, 1181203116, 0143378434

Introduction

This project is to study the exploratory data analysis of a self-service laundry shop on the customer visiting and activities records in the shop. We are expected to make insights based on the data collected, to provide data that could help to improve the business and customer experiences. In this study, we use external data which is a weather data report of the according dates to support and form several hypotheses that might be related to the business. We apply a few different methods to prove and compare the accuracy of the model. In this research , we brought up a few hypotheses such as which generation of customer spends the most time in the shop, does the weather have an impact on the sales and what type of customer will bring kids along to find valuable insights that will be beneficial to the self-service laundry shop.

Data Preparation

Data Collection

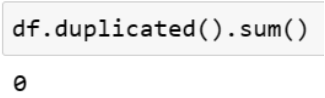
A total of two datasets were collected to complete this project. The first dataset used is the dataset with information about a self-service laundry shop. While the second dataset used, which is the external dataset, is a dataset about the weather. Both datasets are in csv format.

The weather dataset is collected from a website called Visual Crossing. The dataset consists of 5483 rows and 33 columns of data. There is a lot of useful information to be found in the dataset that will be a great help in completing the project and finding out more useful information and valuable insights.

Data Pre-Processing

Self-Service Laundry Shop Dataset

First, we checked if there is any duplicated data in the dataset.

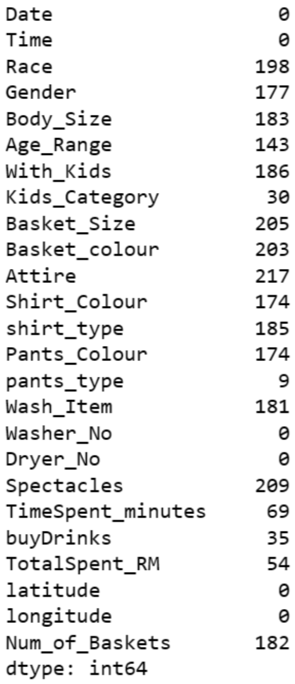


```
df.duplicated().sum()

0
```

From the figure above we can see that there are no duplicate values in the dataset.

We moved on to check the missing values in each of the columns.

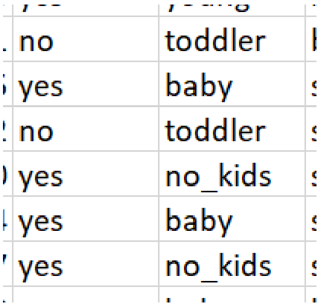


```
Date                   0
Time                   0
Race                 198
Gender               177
Body_Size            183
Age_Range            143
With_Kids            186
Kids_Category         30
Basket_Size          205
Basket_colour        203
Attire               217
Shirt_Colour         174
shirt_type           185
Pants_Colour         174
pants_type             9
Wash_Item            181
Washer_No              0
Dryer_No               0
Spectacles           209
TimeSpent_minutes     69
buyDrinks             35
TotalSpent_RM         54
latitude               0
longitude              0
Num_of_Baskets       182
dtype: int64
```

From the figure above, we can see that there are a lot of missing values in the dataset. To solve this problem, we decided to fill in the missing values instead of removing all the rows with missing values.

For the missing values of the numerical data, we used the mean of the data to fill in the missing values and rounded up the mean to get an integer value. This is to prevent any unwanted errors when we proceed on solving the questions. The reason we decided to use mean for our dataset is because mean value is more suitable to fill in the missing values when the data is symmetric.

While for the missing values of categorical data, we filled in all the missing values with the mode of the data, except for the “With_Kids” and “Kids_Category” column. For these two columns, we found out a few issues in the data. For example, some of the rows show the customer comes with their kids but the kids category listed no kids; or the data shows that the customer does not come with their kids but the kids category listed out the type of the kid. Figure below is a proof of noisy data in the dataset.



no	toddler
yes	baby
no	toddler
yes	no_kids
yes	baby
yes	no_kids

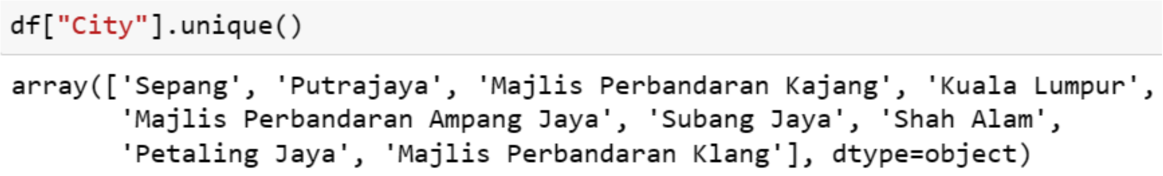
We solved this problem by using several if statement:

1) If the “With_Kids” and “Kids_Category” both do not have any value then we fill in with "no" and "no_kids".
2) If the “With_Kids” shows "yes" but there is no value in the "Kids_Category" then we fill in the value by using the mode of "Kids_Category" after filtering out the "no_kids" value in the column which is "young". The reason we filter out the "no_kids" value in "Kids_Category" is because the mode of "Kids_Category" is "no_kids", so in this case if we want to use mode value, we need to filter the value out otherwise it will also cause problems.
3) If the “With_Kids” shows "no" and no value in the "Kids_Category", we fill in the value with "no_kids".
4) If there is no value in “With_Kids” only and the "Kids_Category" shows any value other than "no_kids" then we will fill in “With_Kids” with "yes".
5) If there is no value in “With_Kids” only and the "Kids_Category" shows "no_kids" then we will fill in “With_Kids” with "no".

We decided to use mode to fill in the missing values of the categorical data because mode is the most common value of the data. It is easy and fast and it changes the statistical nature of the data.

External Dataset (Weather Dataset)

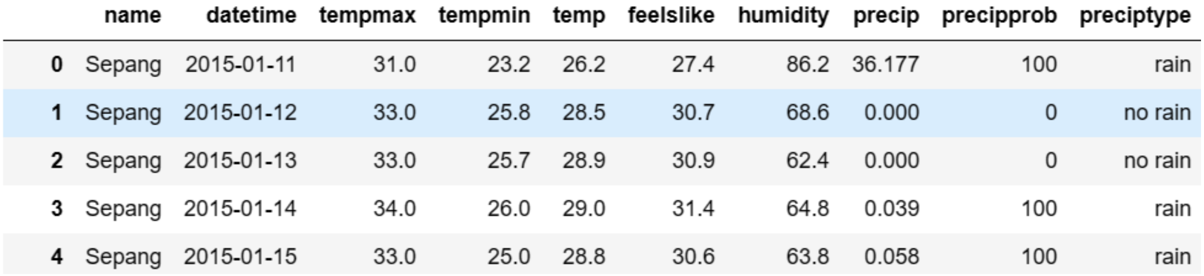
We started to find the location of each customer base on the latitude and longitude data in the dataset to find the weather of the particular location. The cities shown in the figure below are some of the cities that are located based on the latitude and longitude.



```
df["City"].unique()
```

```
array(['Sepang', 'Putrajaya', 'Majlis Perbandaran Kajang', 'Kuala Lumpur',
       'Majlis Perbandaran Ampang Jaya', 'Subang Jaya', 'Shah Alam',
       'Petaling Jaya', 'Majlis Perbandaran Klang'], dtype=object)
```

Then we get the weather data of each of the locations from the previous analysis. The unwanted columns are dropped and only useful weather information is taken to merge with the first dataset.



	name	datetime	tempmax	tempmin	temp	feelslike	humidity	precip	precipprob	preciptype
0	Sepang	2015-01-11	31.0	23.2	26.2	27.4	86.2	36.177	100	rain
1	Sepang	2015-01-12	33.0	25.8	28.5	30.7	68.6	0.000	0	no rain
2	Sepang	2015-01-13	33.0	25.7	28.9	30.9	62.4	0.000	0	no rain
3	Sepang	2015-01-14	34.0	26.0	29.0	31.4	64.8	0.039	100	rain
4	Sepang	2015-01-15	33.0	25.0	28.8	30.6	63.8	0.058	100	rain

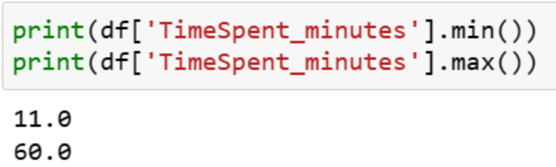
The above dataset is then merged with the cleaned first dataset to form the final dataset that will be useful when answering the questions.

Question 1

<u>Which generation of customer spend the most time in the shop?</u>

Exploratory Data Analysis

First and foremost, the lowest and highest time spent in the self-service laundry shop of the customers are analysed.



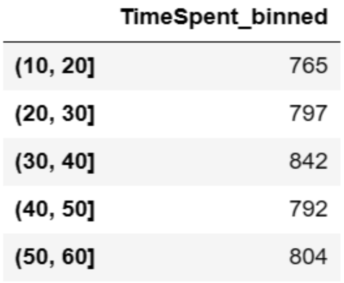
```
print(df['TimeSpent_minutes'].min())
print(df['TimeSpent_minutes'].max())
```

```
11.0
60.0
```

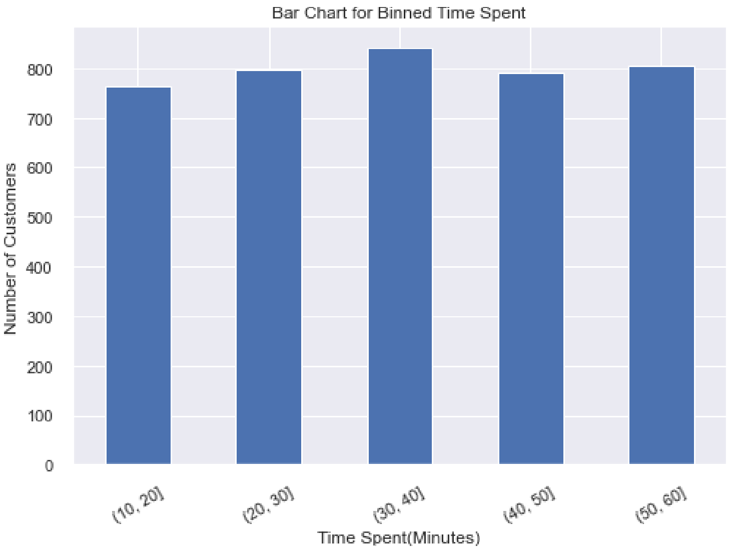
Based on the figure beside we can know that the least time spent by a customer is 11 minutes and the longest time spent by a customer is 60 minutes.

Moving on, the time spent by the customers are cut into bins of 10,20,30,40,50 and 60.

From the table, it has shown that the total number of customers that spend around 30 to 40 minutes in the shop is 842.



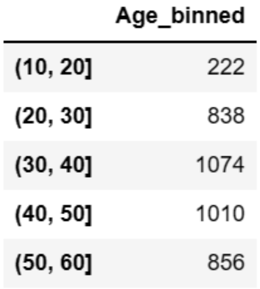
	TimeSpent_binned
(10, 20]	765
(20, 30]	797
(30, 40]	842
(40, 50]	792
(50, 60]	804



The binned data is then visualised in a bar chart with the number of customers.

Based on the bar chart we can see that most of the customers spend around 30 to 40 minutes in the shop.

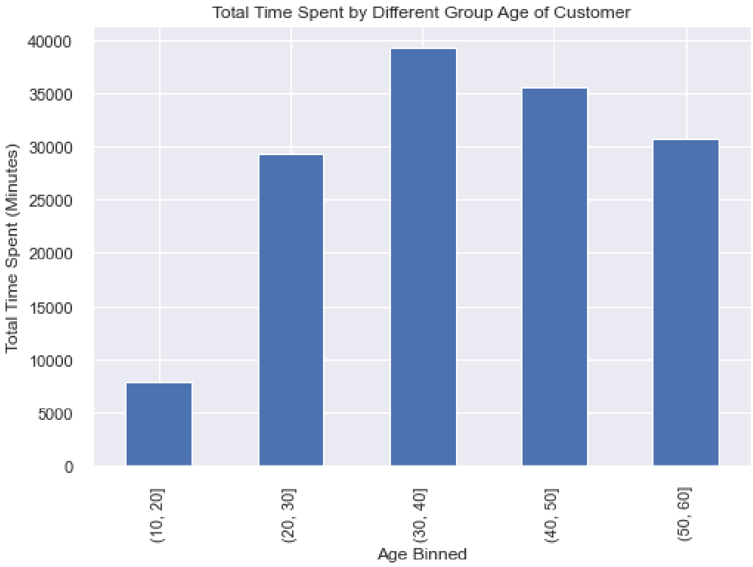
The age attribute is also cut into bins of 10,20,30,40,50 and 60.



	Age_binned
(10, 20]	222
(20, 30]	838
(30, 40]	1074
(40, 50]	1010
(50, 60]	856

From the table shown, it has shown that most of the customers that visited the shop were aged between 30 to 40. While the customers between age 10 to 20 are the least likely to visit the shop.

A bar chart is then plotted to show which age group has the highest total time spent in the laundry shop.

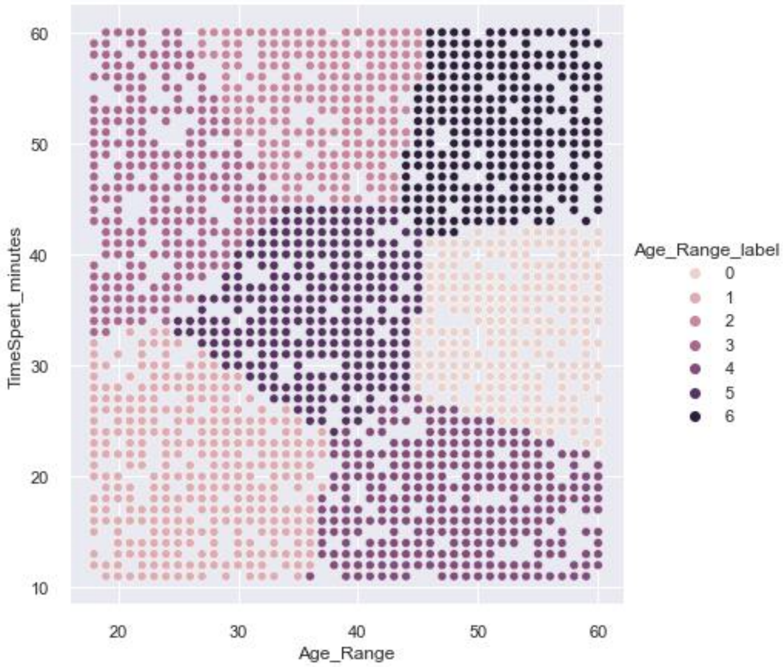


Based on the bar chart, we can see that the customers who are in the age range from 30 to 40 have the highest time spent in the laundry shop.

Cluster Analysis

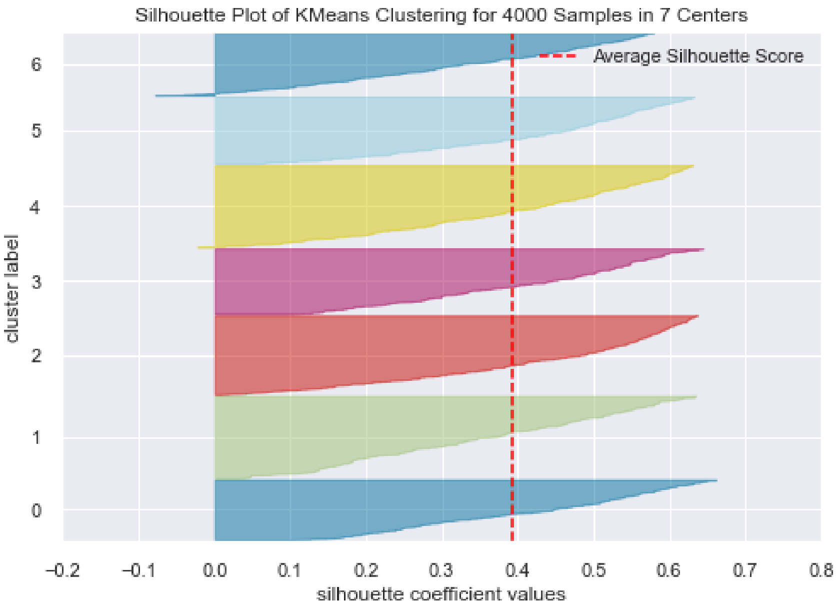
The cluster analysis technique that is used in this analysis is KMeans. The reason for choosing KMeans is because it is easy to implement and it scales to large data sets. The age range of the customers and the time spent of the customers are the two variables that are used to perform the cluster analysis. A relplot is plotted to show findings of the cluster analysis.

From the plot shows, we can see that the age range of the customers are separated into 7 different clusters. The distribution is rather spread around the whole graph.



The silhouette score of the data when n=7 is 0.3933169987159571. Since the score is 0.3933169987159571 which is below 1 and above 0, therefore we can say that the result is considered below average and are able to do better.

The silhouette coefficient is visualised with a silhouette visualizer.

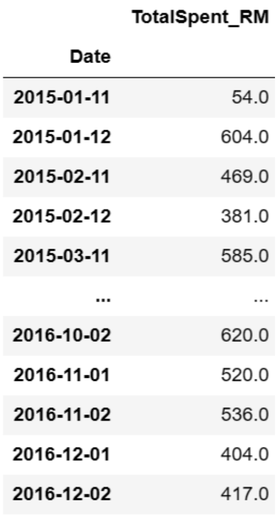


Question 2

<u>Does the weather has an impact on the sales?</u>

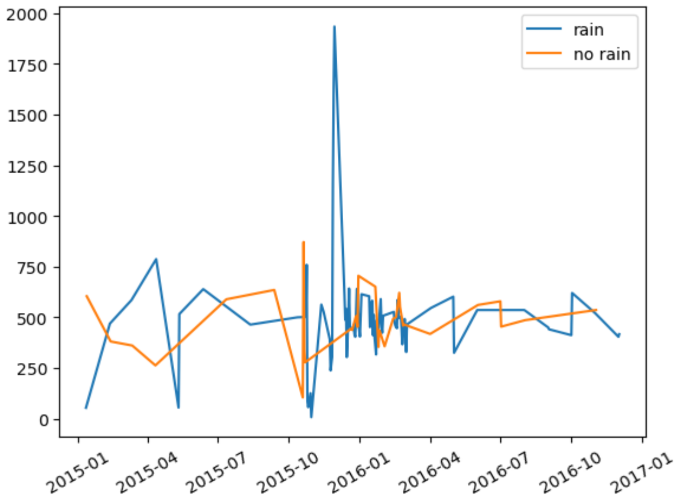
Exploratory Data Analysis

The relationship of the weather and the sales is visualised in a line graph to see the connection between the two variables. The sales will be the total sum up of the attribute, “TotalSpent_RM” based on the “Date”. Figure below shows the table of the sales of each day after summing them up together.



	TotalSpent_RM
Date	
2015-01-11	54.0
2015-01-12	604.0
2015-02-11	469.0
2015-02-12	381.0
2015-03-11	585.0
...	...
2016-10-02	620.0
2016-11-01	520.0
2016-11-02	536.0
2016-12-01	404.0
2016-12-02	417.0

The weather attribute that is taken to answer this question is the “Is_rain” attribute. Two line plots are plotted to show the sales when it is raining and when it does not rain. The two line graphs are then joined together to have a clear visual on the impact of the weather towards the sales.



From the graph, we can see that there is a slight difference in the sales based on the weather. There is a slightly higher sales revenue when it is raining compared to when it is not. This shows that when the weather is raining, it can help the shop to generate more sales.

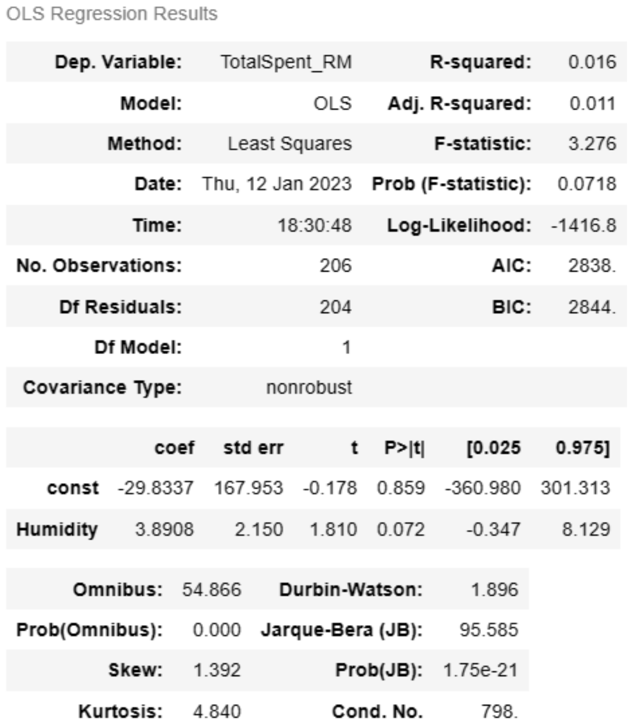
Regression Modelling

For the regression modelling of this question, we decided to use the humidity attribute to perform the modelling instead of the temperature and precipitation. This is because the result using the two other variables is misleading and is not suitable to answer the question. We are interested in finding the relationship between humidity and the sales of the shop.

Linear Regression

We started off with the first regression model which is linear regression. We used two different libraries to construct the regression model, which are statsmodels and sklearn.

Statsmodels



OLS Regression Results

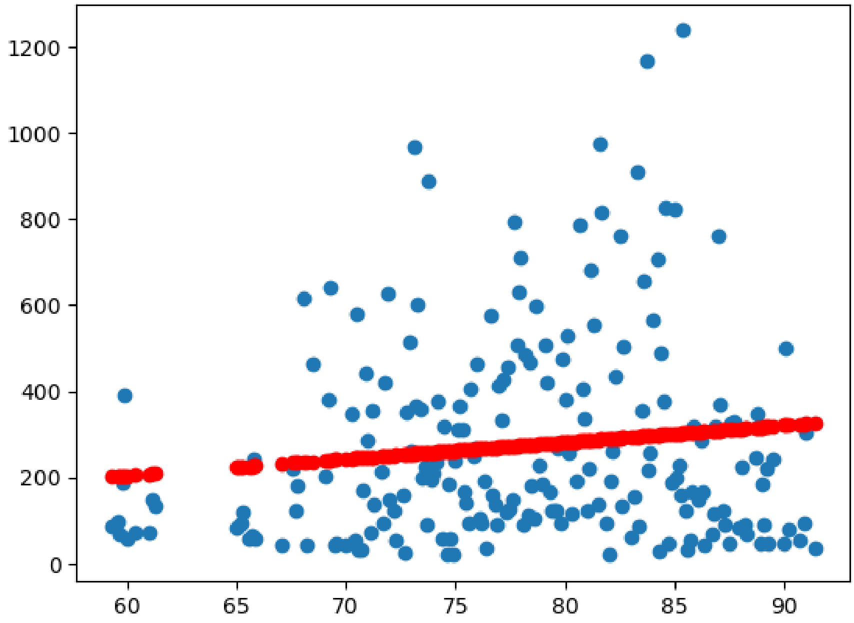
Dep. Variable:	TotalSpent_RM	**R-squared:**	0.016
Model:	OLS	**Adj. R-squared:**	0.011
Method:	Least Squares	**F-statistic:**	3.276
Date:	Thu, 12 Jan 2023	**Prob (F-statistic):**	0.0718
Time:	18:30:48	**Log-Likelihood:**	-1416.8
No. Observations:	206	**AIC:**	2838.
Df Residuals:	204	**BIC:**	2844.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-29.8337	167.953	-0.178	0.859	-360.980	301.313
Humidity	3.8908	2.150	1.810	0.072	-0.347	8.129

Omnibus:	54.866	**Durbin-Watson:**	1.896
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	95.585
Skew:	1.392	**Prob(JB):**	1.75e-21
Kurtosis:	4.840	**Cond. No.**	798.

The figure shows the result of the summary of OLS regression results. We acquired:

- R-squared: 0.016
- Coefficient: 3.8908



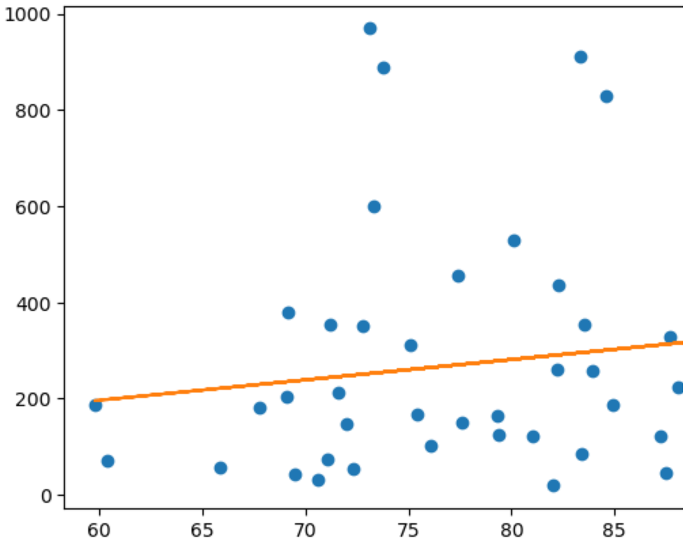
The scatter plot above shows the result more clearly. From the scatter plot above, we can conclude that there is just a slight impact of humidity on the sales of the shop. As the increase in the x-axis has only very little increase in the y-axis. This means that there is not much influence between X and Y.

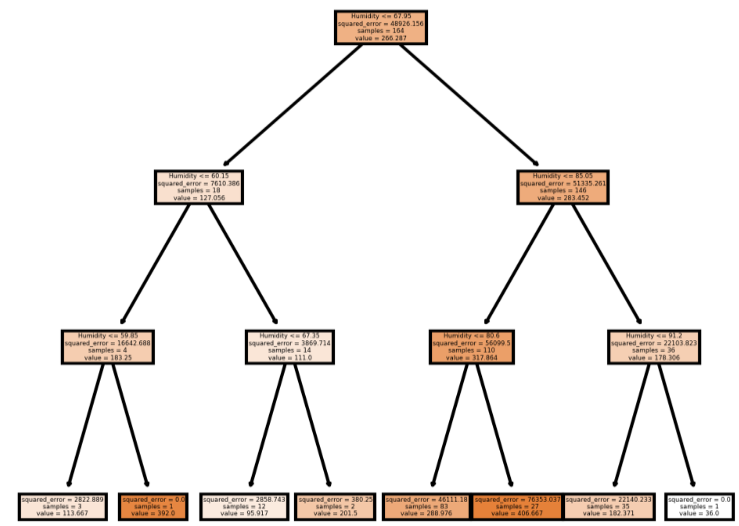
SKLearn

Train test split is performed before constructing the SKLearn regression model. The result of this model are as follows:

- Coefficient: 4.23936722
- Intercept: -57.78884424191318

We also predicted the sales based on the humidity data. The scatter plot above shows the result of the prediction. The predicted sales are in blue dots while the actual sales are in yellow crosses. From the scatter plot above we can see that the model did not perform well as the predicted sales are mostly incorrect.





Decision Tree Regressor

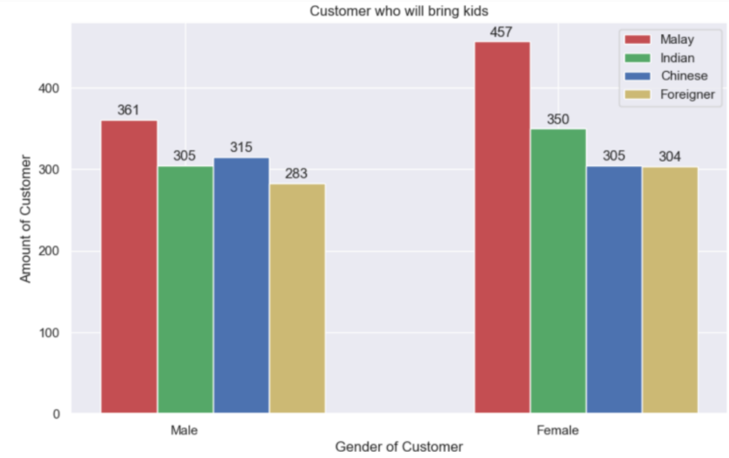
The result of the decision tree regressor is not ideal as the value of the mean absolute error is 213.58, which is way higher than the mean value of the sales. The results of the decision tree are as shown in the figure below.

Question 3

<u>What type of customer will bring kids along?</u>

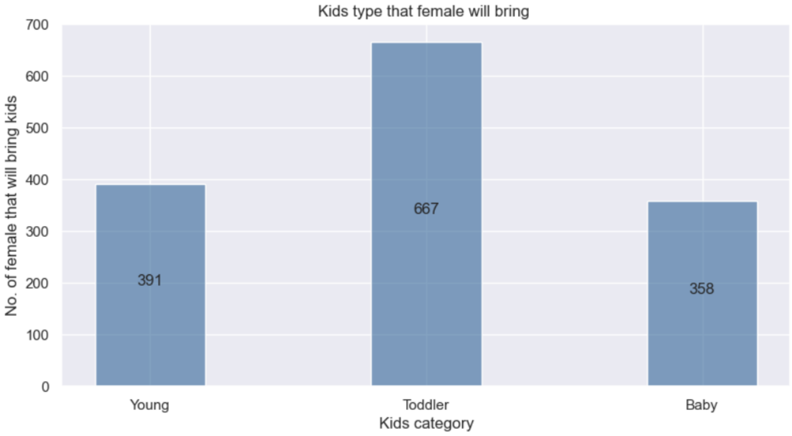
Exploratory Data Analysis

In this hypothesis, the race and gender of the customers are analysed in terms of customers who will bring kids.

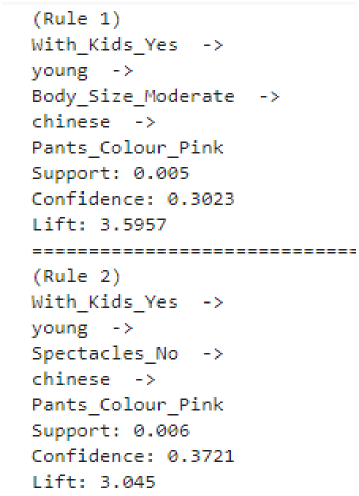


Based on the Group Bar plot, we can find out that the majority of the females, especially Malay females, like to bring kids to the laundry. The amount of female malay is 457, higher than male malay. The Indian female amount is 350, compared with Indian male is higher. However, when comparing between Chinese male and females, Chinese male amount slightly higher than Chinese females. At last, Foreigner female amount is higher than foreigner male. The total number of females that will bring kids is higher than males.

Next, let's look at the data of females who bring kids. Based on the bar chart, we found out that the majority of the females whose kid's is toddler have higher opportunity to visit laundry. The total amount of customers who will bring a toddler are 667. Compared with the young and baby category, the female whose kid is toddler will come to the laundry.



Association Rule Mining



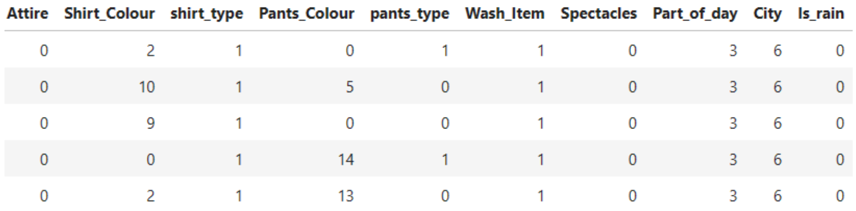
```
(Rule 1)
With_Kids_Yes  ->
young  ->
Body_Size_Moderate  ->
chinese  ->
Pants_Colour_Pink
Support: 0.005
Confidence: 0.3023
Lift: 3.5957
==============================
(Rule 2)
With_Kids_Yes  ->
young  ->
Spectacles_No  ->
chinese  ->
Pants_Colour_Pink
Support: 0.006
Confidence: 0.3721
Lift: 3.045
```

To make a clean and focused environment, we make a list of combining each of the association values together by using the following columns. Which are Race of customer, Gender of customer, Body size of customer, With kids status, their kids category, Customer attire, Shirt Color, Shirt Type, Pants Color, Pants types, and are the customer wearing spectacles.

After a series of model training with Association Rule Mining. We gained the list of the confidence of the data and model. The model stores the related values into each list as an association rule by using apriori model. We set the min_support as 0.0045, min_confidence as 0.2, min_lift as 3, and min_length as 2. The following parts are the sample output of Association Rule Mining lists.

Label Encoding

Before feature selection, we use the Label Encoding model to change all the categorical values into integer numbers. The reason that we used label encoding is because to avoid overfitting issues that happen on the classification model. The following part are the sample output of after label encoding:



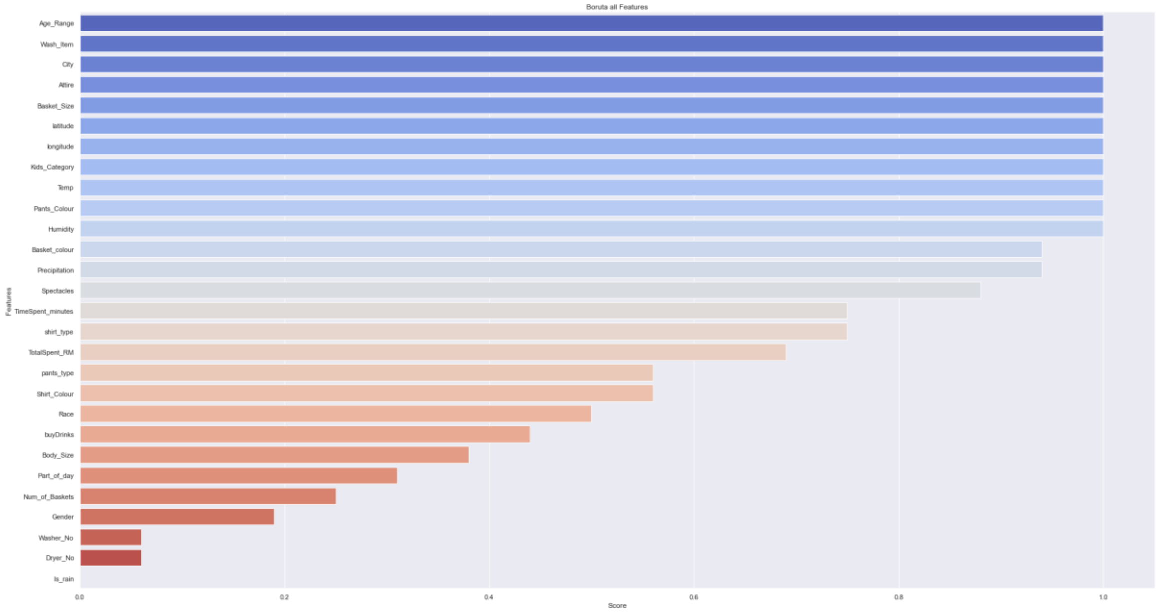
Attire	Shirt_Colour	shirt_type	Pants_Colour	pants_type	Wash_Item	Spectacles	Part_of_day	City	Is_rain
0	2	1	0	1	1	0	3	6	0
0	10	1	5	0	1	0	3	6	0
0	9	1	0	0	1	0	3	6	0
0	0	1	14	1	1	0	3	6	0
0	2	1	13	0	1	0	3	6	0

Feature Selection

Feature selection is one of the important parts in data mining. The purpose of feature selection is to select the related features from datasets. Because unrelated features will cause the result of the model to become worse. In this question, we used Boruta and Information Gain for feature selection.

Boruta

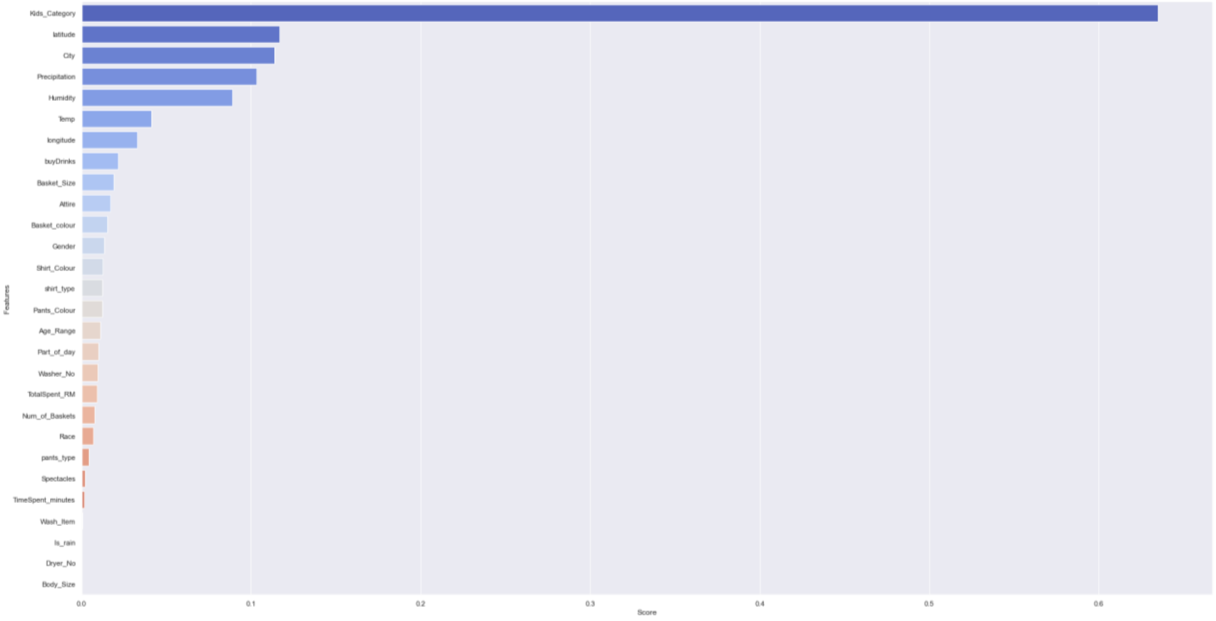
Here is the data visualisation graph of the Boruta score of all features



Based on the graph above, the top Boruta score features with score 1.0 are Age_range, Wash_item, City, Attire, Basket_Size, latitude, longitude, Kids_Category, Temp, Pants_Colour, and Humidity. Age_range gets the highest range between other features..

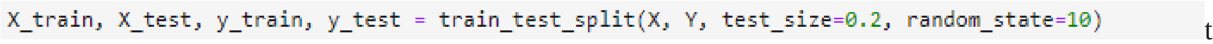
Information Gain

The following part is the data visualisation graph of the Information Gain of all features



Based on the Information Gain graph, Kids_Category has the highest score of 0.6 instead of other features with scores less than 0.2.

After feature selection with Boruta and Information Gain, we are going to use two classification models which are Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) to do prediction on this question. The reason that we choose this two models is because they will not have overfitting issue for this dataset. When we try other model, they had overfitting issue.



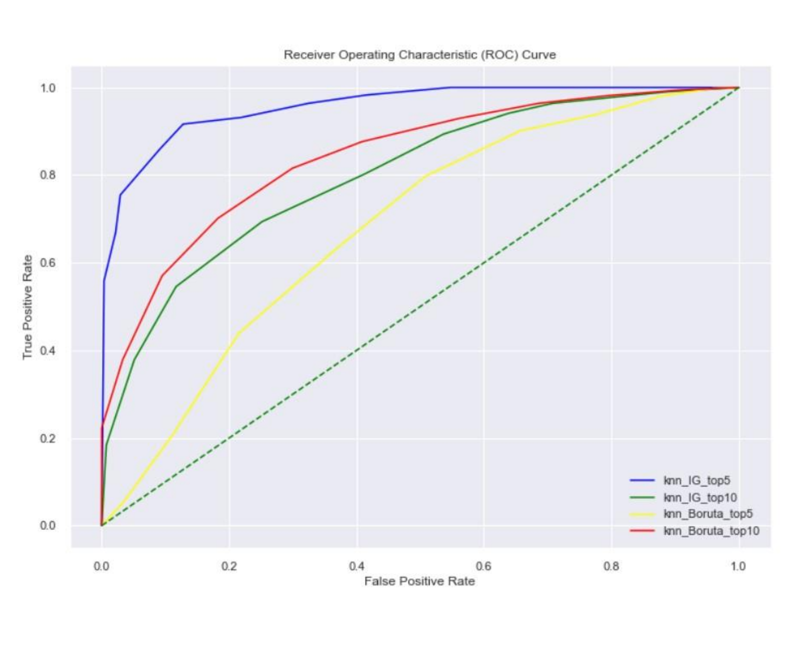
```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=10)
```

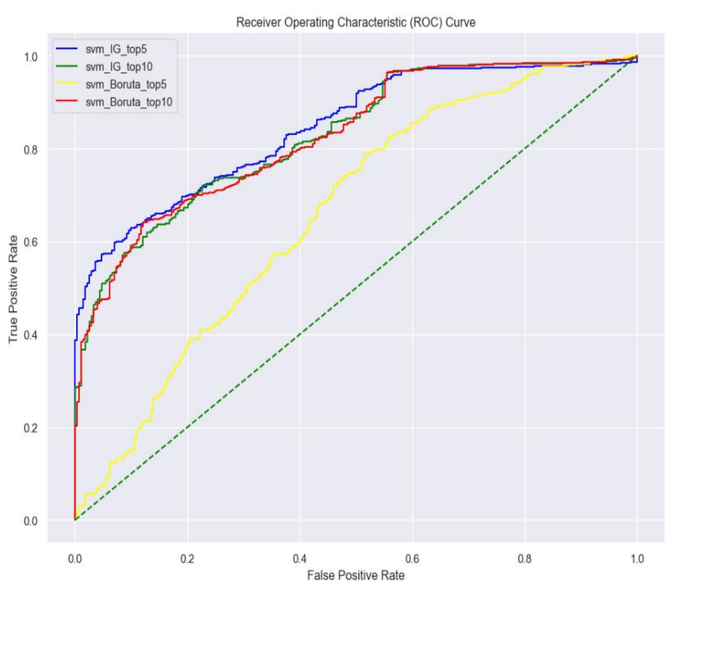
train-test-split with test size 20% and random state=10

With a series of calculations, we collect the AUC and ROC for KNN and SVM model and plot it to a graph.

KNN (AUC)	SVM (AUC)
KNN (Info Gain) Top 5 AUC: 0.96	SVM (Info Gain) Top 5 AUC: 0.85
KNN (Info Gain) Top 10 AUC: 0.85	SVM (Info Gain) Top 10 AUC: 0.83
KNN (Boruta) Top 5 AUC: 0.68	SVM (Boruta) Top 5 AUC: 0.66
KNN (Boruta) Top 10 AUC: 0.84	SVM (Boruta) Top 10 AUC: 0.83

KNN Model (ROC Curve)	SVM Model (ROC Curve)

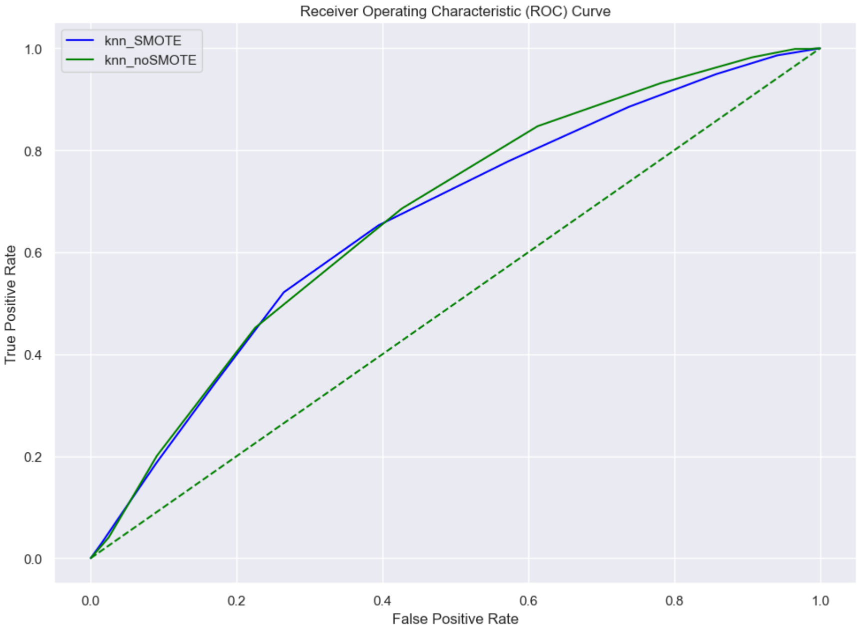




What is the impact of SMOTE and non-SMOTE datasets?

SMOTE- SMOTE (Synthetic Minority Over-sampling Technique) is a popular oversampling method used to balance the unbalanced dataset. Imbalanced datasets are those where the minority class has significantly less observations than the majority class. Below is the KNN classifier result comparison after with and without SMOTE dataset.

SMOTE	**no SMOTE**
Precision= 0.79	Precision= 0.73
Recall= 0.52	Recall= 0.85
F1= 0.63	F1= 0.78
Accuracy= 0.59	Accuracy= 0.69
KNN - SMOTE AUC: 0.66	KNN - no SMOTE AUC: 0.67

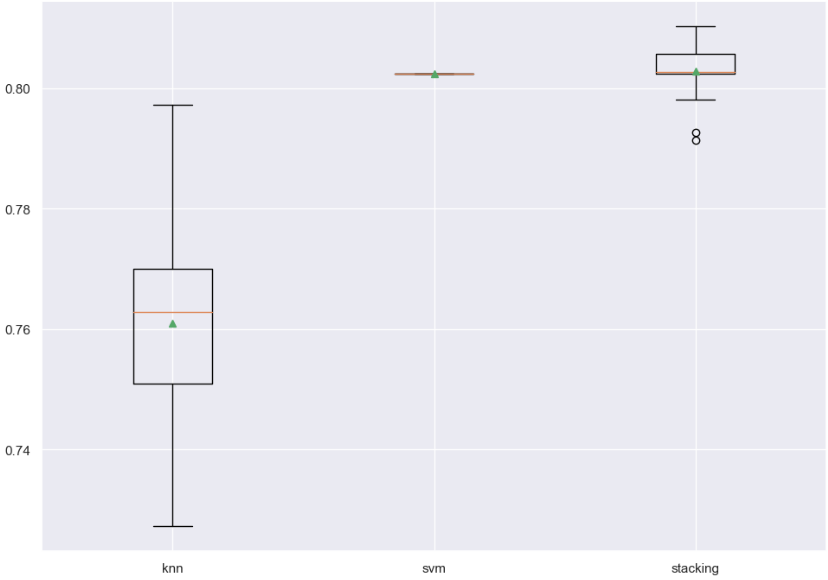


Plot above shows the comparison of the accuracy of the model with and without SMOTE technique.

From the result we can conclude that the SMOTE had oversampled the dataset that leads to decreased accuracy of the model.

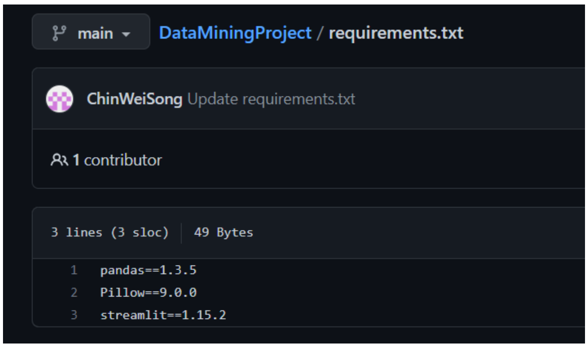
Does Stacked Ensemble Modeling work better?

The stacked Ensemble Modeling which is used here combines both KNN and SVM algorithms. The result for KNN model is 0.761, SVM model is 0.802. The Stacking Ensemble Modeling is 0.803. Therefore we can conclude that the Stacking Ensemble Modeling work better, because it’s score is higher than other two individual models.



Streamlit Deployment

The streamlit app is being hosted by using Streamlit cloud. Before we deploy the app to the cloud, we need to get the “requirements.txt” file. This is because most of the cloud platforms need to know what Python packages to install before they can start your app. If no requirements are being specified in a “requirements.txt” file it may cause errors when cloud platforms want to start your app. Therefore it is really important to do that.



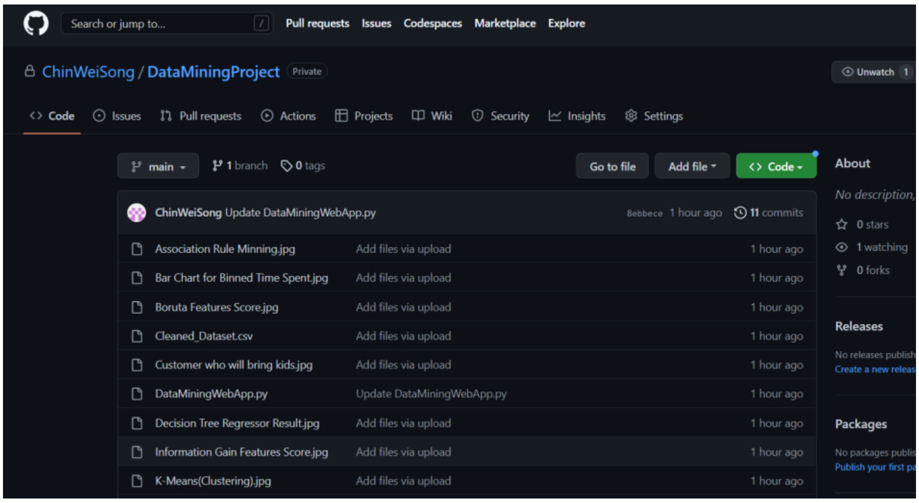
ChinWeiSong Update requirements.txt

1 contributor

3 lines (3 sloc) | 49 Bytes

```
pandas==1.3.5
Pillow==9.0.0
streamlit==1.15.2
```

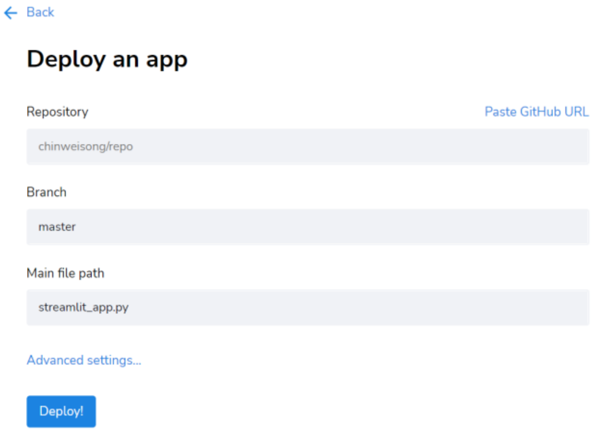
Additionally, a GitHub account is needed to upload your files by creating a new repository. Beside a GitHub account, a streamlit cloud account is also needed.



ChinWeiSong Update DataMiningWebApp.py 8ebbece 1 hour ago 11 commits

Association Rule Minning.jpg	Add files via upload	1 hour ago
Bar Chart for Binned Time Spent.jpg	Add files via upload	1 hour ago
Boruta Features Score.jpg	Add files via upload	1 hour ago
Cleaned_Dataset.csv	Add files via upload	1 hour ago
Customer who will bring kids.jpg	Add files via upload	1 hour ago
DataMiningWebApp.py	Update DataMiningWebApp.py	1 hour ago
Decision Tree Regressor Result.jpg	Add files via upload	1 hour ago
Information Gain Features Score.jpg	Add files via upload	1 hour ago
K-Means(Clustering).jpg	Add files via upload	1 hour ago

Once everything that is mentioned above is done, then we are able to create a new app at streamlit cloud and link it to the GitHub repository that has been created. After that, change the “Main file path” to your python file name and click “Deploy!”.



Deploy an app

Repository: chinweisong/repo

Branch: master

Main file path: streamlit_app.py

This is our group streamlit web app link: https://chinweisong-dataminingproject-dataminingwebapp-pl4vik.streamlit.app/

Reference

- https://www.visualcrossing.com/weather-history/Kuala%20lumpur