

**TDS 3301**

**DATA MINING**

**PROJECT: QUESTION 1**

**Profiling Customers in a Self-Service Coin Laundry Shop**

Prepared by

Nicholas Chee Jian Shen, 1171103441, 010-8289371

Chan Jun Ting, 1171103572, 010-9200138

Tee Wai Bing, 1171103537, 019-6289891

# 

**Table of Contents**

[Chapter 1: Exploratory Data Analysis and Data Pre-Processing](#_xjnu9tbzuxda) 2

[Chapter 2: Feature Selection](#_1jaqpw87bm58) 6

[Chapter 3: Appropriate Machine Learning Techniques Used](#_mwwxfqstf79m) 7

[3.1 Classification Models](#_2qca9bi84m5q) 7

[3.2 Regression Models](#_w3uy3lc8jeb8) 9

[3.3 Association rule](#_op8smaikq4gs) 10

[3.4 Clustering](#_fpu3dlm6wrxh) 11

[Chapter 4: Deployment](#_ids567ml7hu3) 12

[Summary (Questions Answered)](#_uuybmokc3y9f) 13

**Contributions**

| **Group Member** | **Work Done** |
| --- | --- |
| Everyone | * EDA & Data Pre-Processing * Report |
| Nicholas Chee Jian Shen  1171103441 | * Feature Selection * Streamlit Deployment |
| Chan Jun Ting  1171103572 | * Association Rule * Clustering |
| Tee Wai Bing  1171103537 | * Classification Model * Regression Model |

# 

# Chapter 1: Exploratory Data Analysis and Data Pre-Processing

For our EDA, we have performed the necessary steps:

1. Imported the required libraries.
2. Read the required data.

|  | From the head and tail of the dataframe, we can tell that there are null values and that the data mainly consist of categorical variables with age range being the only numerical value. |
| --- | --- |

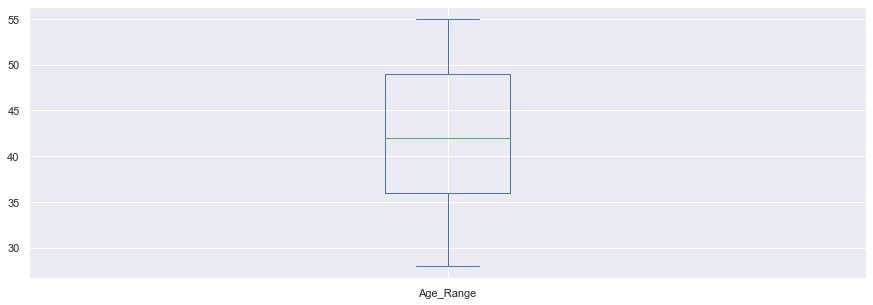
1. Look through the descriptive statistics.

|  |  |
| --- | --- |

1. Check for missing values.

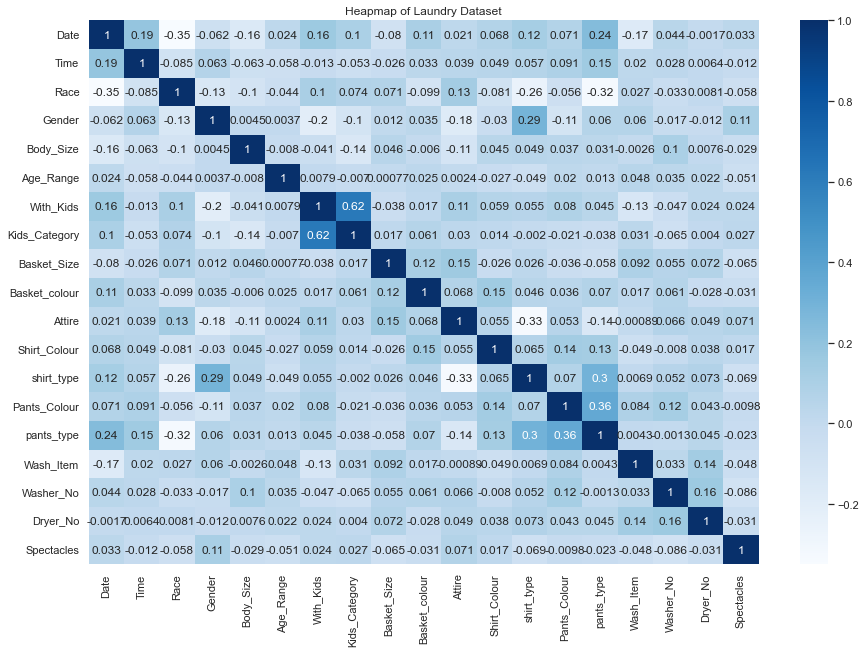
|  | The attribute count from df.info() and the isnull().any() function shows that some of the attributes are missing some data.  As the number of null values are relatively small, we filled in the missing values with the most frequently occuring value (mode) of each column. |
| --- | --- |

1. Represent our data using graphs.

**

The only numerical attribute in our dataset is the age range:

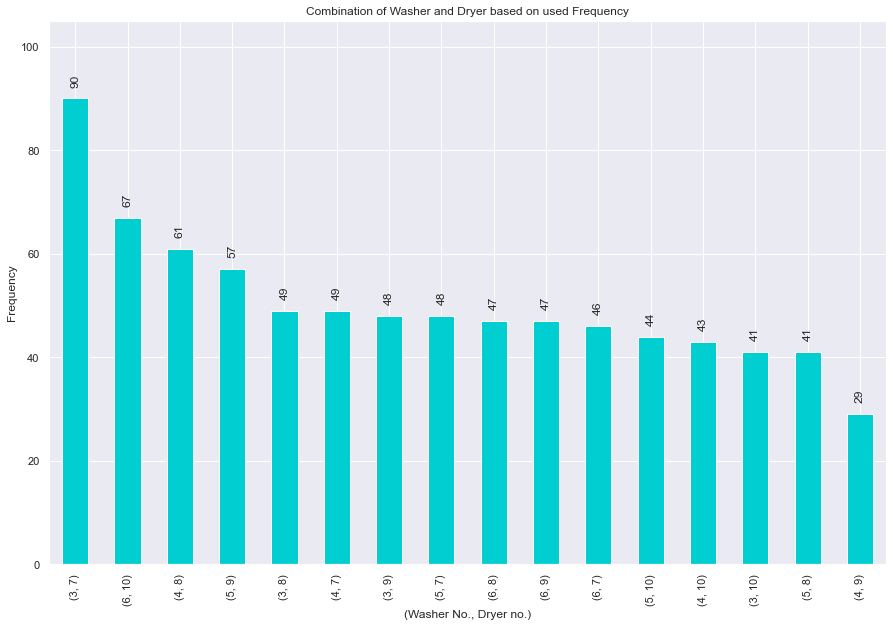
* The data that we used has no outliers and is approximately symmetrically distributed.

**

The heatmap shows the correlation between the different attributes in our dataset.

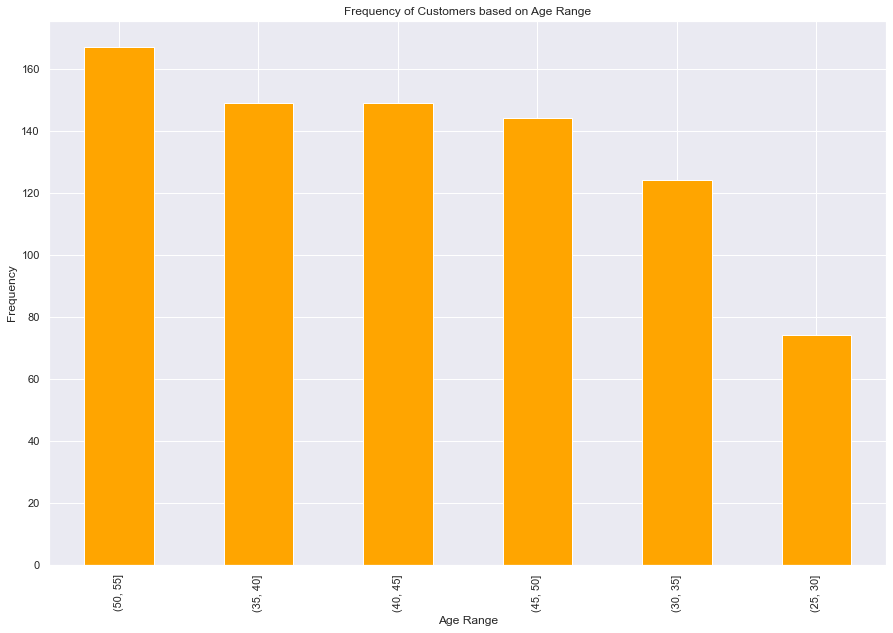
* No significant strong correlations are found.
* The highest correlation of 0.62 is between Kids\_Category and With\_Kids which does not obtain any new information.
* Insignificant weak correlation of 0.36 between Pants\_Colour and pants\_type.
* Insignificant weak correlation of 0.3 between shirt\_type and pants\_type.

|  |  |
| --- | --- |

**

As we can tell from the bar charts above:

* The most frequently used washer is washer no.3 with a count of 228, the least used washer is washer no.4 at 182 times.
* The most frequently used dryer is dryer no. 7 with a count of 233, the least used dryer is dryer no.9 at 181 times.
* The combination of the washer no.3 and dryer no.7 is most frequently used by the customers, with a total of 90 times. This is interesting as the most popular washer and dryer are also frequently used in unison.

**

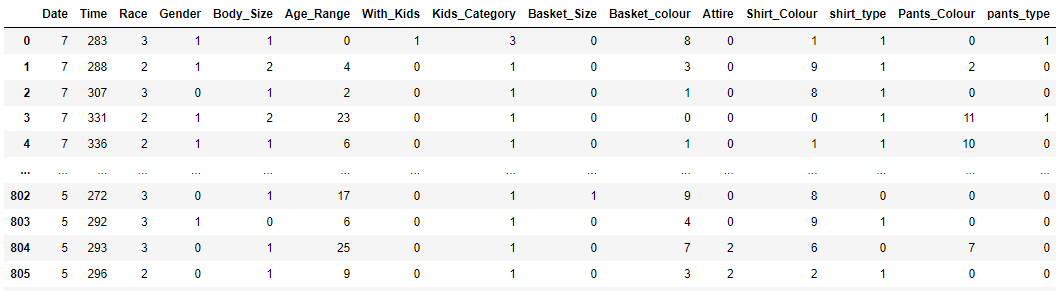
We have separated the age range into bins of 5, excluding values above 55 and below 25 as no records outside this range were recorded. The bar chart above shows that:

* A linear decrease in the number of customers as their age decreases.
* This shows that customers starting from the 35 to 40 years old range will have a significant increase in tendency to use the laundry shop.

# 

# Chapter 2: Feature Selection

Before the data can be used for any models and feature selection, label encoding is done to turn the categorical variables into numeric values so that potential number-related connections between the encoded data can be made.



We have chosen ‘Gender’ as the target variable as the attributes, like the customers’ attire, are closely related to the customer’s gender. To find the best set of features, we have done two feature selection techniques: Boruta and Recursive Feature Elimination (abbreviated as RFE).

| **Boruta** | **RFE** |
| --- | --- |

Based on both feature selections from Boruta and RFE:

* The top 5 features obtained are the same, with them being Washer\_No, Time, Pants\_Colour, Age\_Range and Shirt\_Colour.
* Washer\_No, Time and Pants\_Colour are the features we will be using in the upcoming models.

# Chapter 3: Appropriate Machine Learning Techniques Used

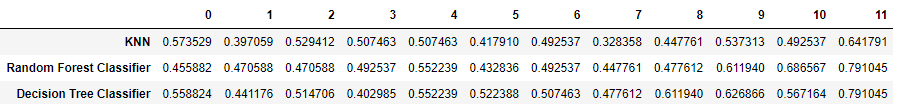
## 3.1 Classification Models

Based on the feature selection done, the washer’s number, the customer’s pants colour and the time of the customer using the shop’s service are the features that perform a high score to the customer’s gender. Thus, we set these features as X to classify the gender of the customer.

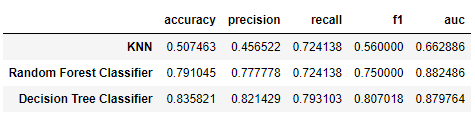
As for the classification models, we will be using three different classifiers:

1. K-Neighbors Classifier
2. Random Forest Classifier
3. Decision Tree Classifier

We start by obtaining the cross validation score of the three classification models. We chose to do 12 times of cross validation.



We proceed to use the K-fold cross-validation with 12 splits to divide our dataset and fit them with the three classification models to obtain the performance evaluation results. The larger the value, the more accurate the classification result.



* *Accuracy: A = (TP + TN)/(TP + FN + FP + TN), the value range is [0,1].*
* *Precision: P = TP/(TP+FP), the value range is [0,1].*
* *Recall: R = TP/(TP+FN), the value range is [0,1].*
* *F-score(f1) : F1 = 2 ＊ (P ＊R) / (P + R), the value range is [0,1].*
* *Area Under Curve(auc): The area under the ROC curve, the ROC curve is generally above y=x, so the value range of AUC is generally between 0.5 and 1.*

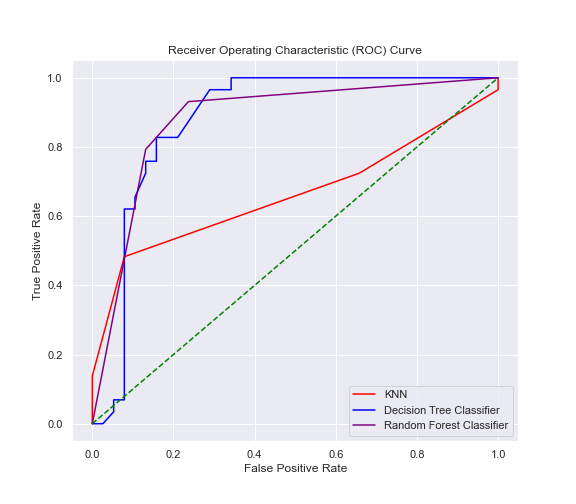
*# TP(True Positive) = original positive, classified positive*

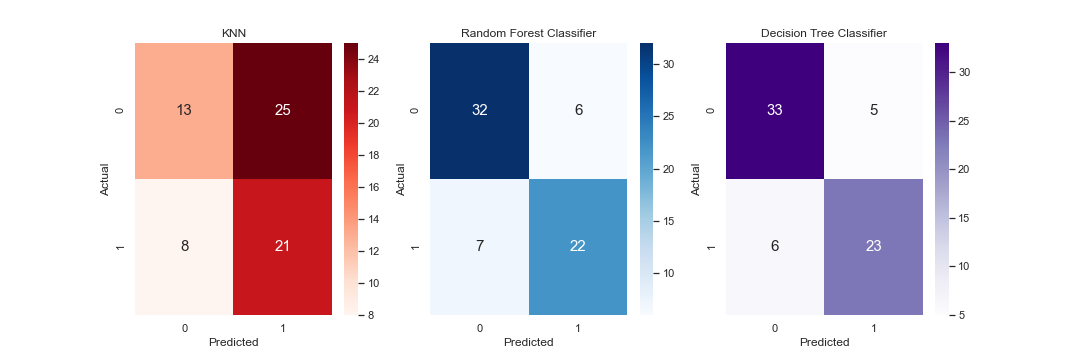
*# TN(Ture Negative) = original negative, classified negative*

*# FP(False Positive) = original negative, classified positive*

*# FN(False Negative) = original positive, classified negative*

We then obtain the Receiver Operating Characteristic (ROC) curve and the confusion matrix for these classifiers.





*Confusion matrix: TN FP*

*FN TP*

According to both the performance evaluation results and confusion matrix of the classifier:

* The Decision Tree Classifier is good in classifying the customer’s gender based on the washer’s number, customer’s pants colour and time of service.

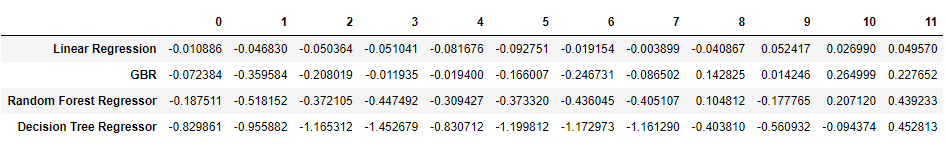
## 3.2 Regression Models

Same with the classification models, X will be the three features obtained from feature selection to classify the customer’s gender.

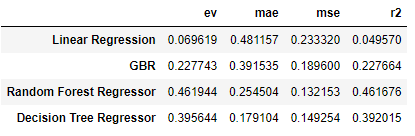
For the regression models, we will be using four different regressors:

1. Linear Regression
2. Gradient Boosting Regressor
3. Random Forest Regressor
4. Decision Tree Regressor

We start the same by obtaining the cross validation score of the three regression models. We chose to do 12 times of cross validation.

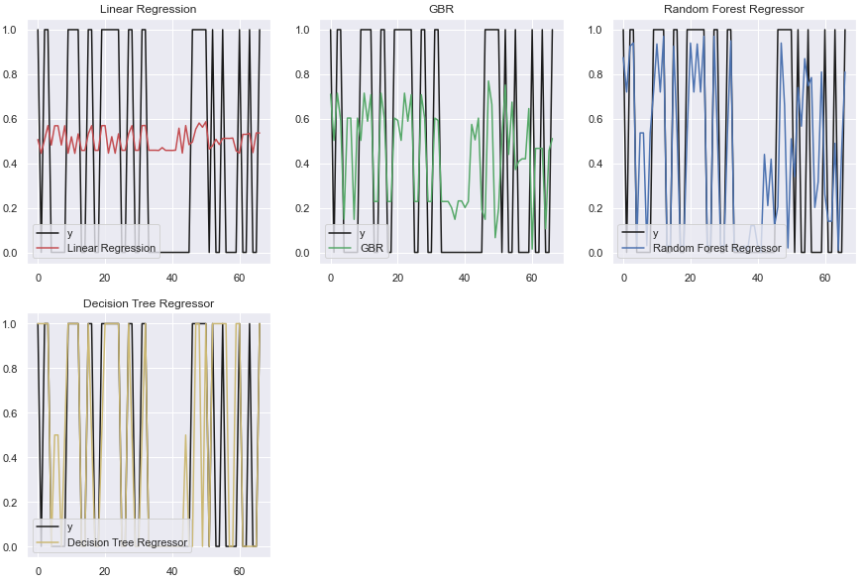


We then use the K-fold cross-validation again with 12 splits to divide our dataset and fit them with the three classification models to obtain the performance evaluation results.



* *Explained Variance Score(ev): Explain the variance score of the regression model, the value range is [0,1]. The smaller the value, the worse the effect.*
* *Mean Absolute Error(mae): Used to evaluate the degree of closeness between the predicted result and the real data set. The smaller the value, the better the fitting effect.*
* *Mean Squared Error(mse): This indicator calculates the mean value of the sum of squares of the errors between the fitted data and the corresponding sample points of the original data. The smaller the value, the better the fitting effect.*
* *R2: Explain the variance score of the regression model. Its value range is [0,1], The smaller the value, the worse the effect.*

The comparison of the predicted data and test set for these four models is shown below.

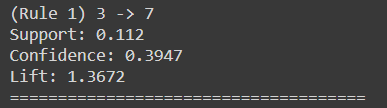


According to the performance evaluation results of the regressors:

* The Random Forest Regressor is good in predicting the customer gender based on the washer’s number, customer’s pants colour and time of service.
* A thing to note is even though the variance score and the R2 score of the Random Forest Regression is the highest among the models, it is still low comparatively to real world values.

## 3.3 Association rule

For the association rule, we will only be using the apriori algorithm.



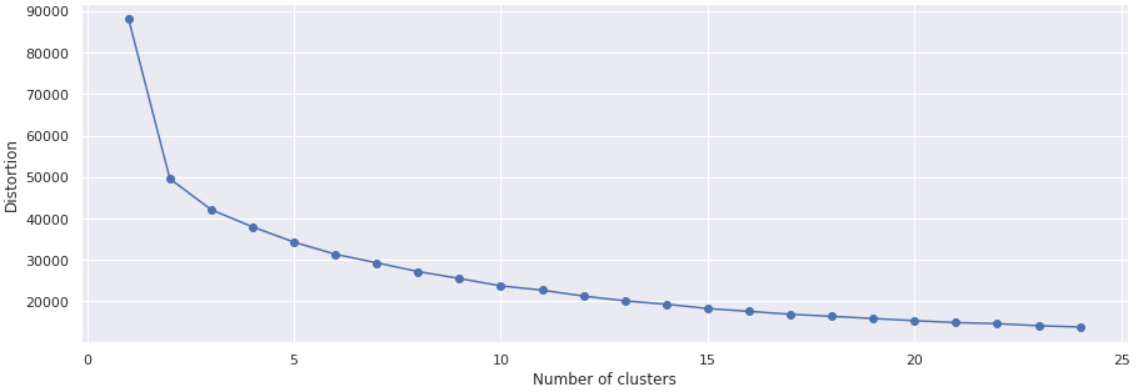
Based on the results which show the association between washer number and dryer number:

* Only one rule is shown, which is between washer no.3 and dryer no.7.
* The support value of 0.112, the confidence value of 0.3947 and the lift of 1.3672 represents a strong association between washer no.3 and dryer no.7.
* This goes along with the high frequency of their combined use which means that customers tend to use these two machines together in particular.

## 3.4 Clustering

For clustering, we will be using k-means to find out the characteristics of customers that will visit the laundry store based on the time. The features obtained from the feature selection (Pants\_Colour, Age\_Range, Shirt\_Colour) are used.

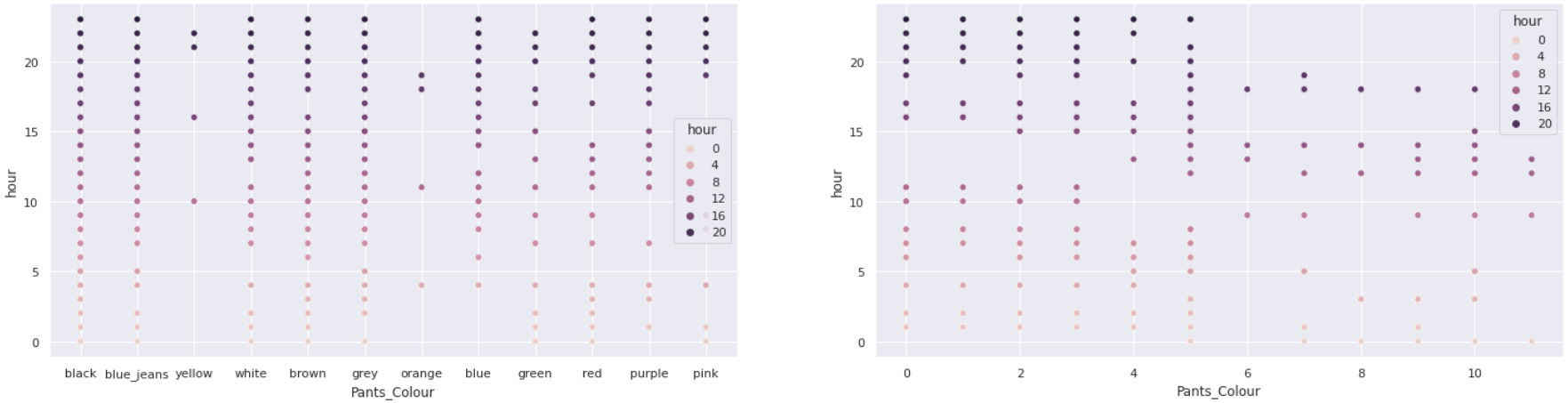
**Elbow method**



We have set 24 as the number of clusters to represent the hours in a day. This will help determine the optimal number of clusters into which the data may be clustered.

* The clustering number is 5 based on the graph above.

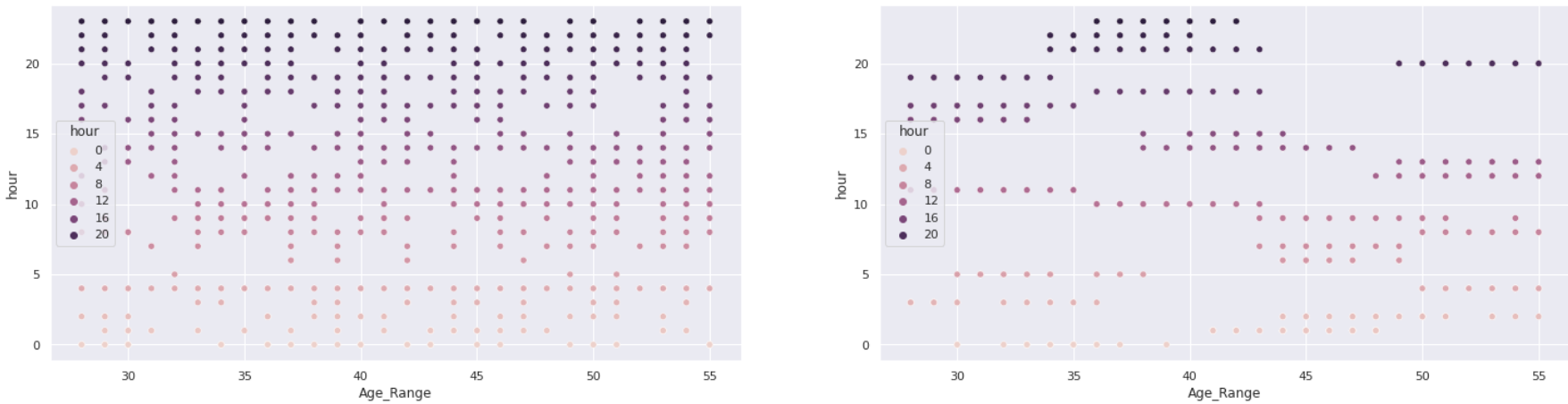
**Pants\_Colour**



A difference can be seen between the two scatter plots. The left plot uses the original data to do clustering, while the right plot is prior to doing the k-means clustering. Based on the scatter plot above:

* The customers who wear black, blue\_jeans, yellow, white, brown and grey pants mostly appear in the morning or during night hours.
* The customers who wear orange, blue, green, red, purple, pink pants mostly appear in the morning and in the afternoon.

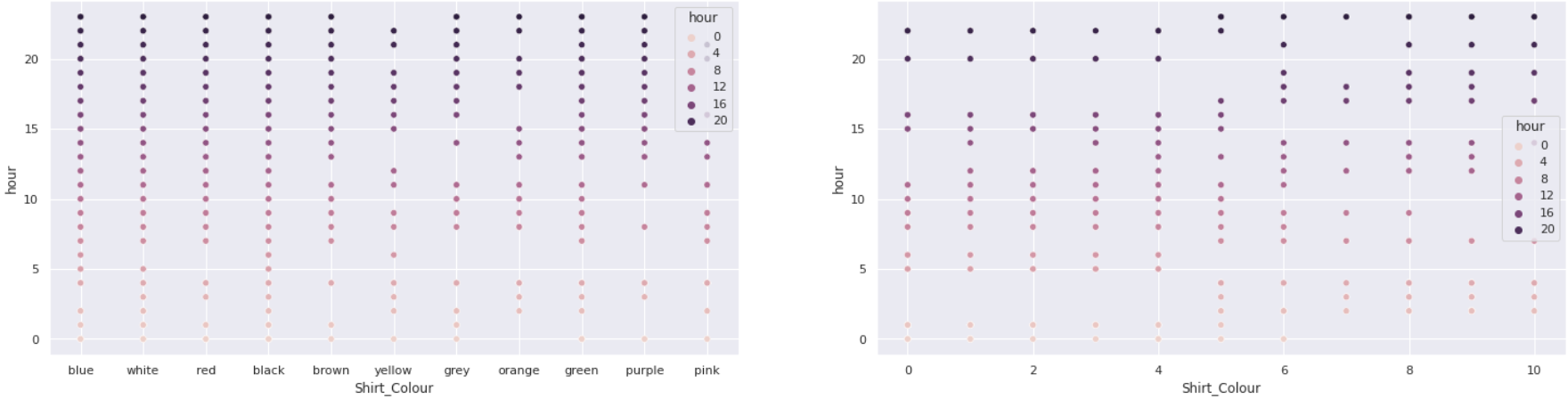
**Age\_Range**



Based on the scatter plot above:

* The customers with ages between 45 to 55 appear early in the morning.
* The customers with ages between 30 to 40 appear at night.

**Shirt\_Colour**



Based on the scatter plot above:

* The customers who wear blue, white, red, black, brown and yellow shirts will mostly appear during the afternoon hour.
* The customers who wear grey, orange, green, purple and pink shirts will mostly appear in the morning and during night hours.

# Chapter 4: Deployment

The plots and a summarized version of the report are deployed and hosted on Heroku.com using the Python Streamlit package.

The link to the deployment: <https://dm-streamlit-project.herokuapp.com>

# Summary (Questions Answered)

*Chapter 1: Exploratory Data Analysis and Data Pre-Processing*

**Q1. How to visualize the content?**

We used the seaborn library and matplotlib library to create boxplots, heatmaps, bar charts, line charts and scatter plots.

**Q2. Do I need to preprocess the dataset?**

Yes, because the data set provided has missing values and is incomplete.

**Q3. How about outliers and missing values?**

The only numerical variable, age range, has no outliers. There are missing values but the number of null values are relatively small. We filled in the missing values with the most frequently occuring value (mode) of each column.

**Q4. Which washer and dryer is the most popular among the customers?**

Washer no.3 and dryer no.7 are the most popular respectively.

**Q5. What age group of customers tend to visit the laundry shop?**

Customers starting from the 35 to 40 years old range.

*Chapter 2: Feature Selection*

**Q6. Justify why that feature selection techniques were selected?**

Two different feature selection techniques were used to ensure that we can obtain the most reliable best set of features.

Boruta is a wrapper algorithm around Random Forest that focuses on the feature sets related to the target variable instead of trying to minimize the cost function of the model. It can help us understand the influencing factors of the dependent variable more comprehensively, so as to perform feature selection better and more efficiently.

RFE is also another wrapper algorithm that removes the weakest feature and repeats the process on the remaining features until every feature is eliminated. The descending order in which the features are eliminated becomes the ranking of the best features.

*Chapter 3: Appropriate Machine Learning Techniques Used*

**Q7. How do you obtain the optimal accuracy?**

K-Fold cross-validation is a Model Selection method. The step of K-fold cross-validation is to divide the initial dataset into K folds, one fold is used for performance evaluation, and the remaining K-1 folds are used as the training set.

The step will repeat K times and combine the results to obtain the final performance evaluation result. K-fold cross-validation can obtain the optimal hyperparameter value with satisfactory generalization performance, and has higher accuracy and robustness.

The reason for the good performance of K-fold cross-validation is that it has more training samples, and each training sample is verified exactly once, which can produce lower variance. K-Fold cross-validation will be used on training classification models and regression models.

**Q8. What classification model is suitable to classify the gender of the customer from the dataset?**

Decision Tree Classifier is suitable to classify the gender of the customer from the dataset. This is because it has obtained the highest scores in accuracy, precision, recall, f1 and auc among the models in this project.

**Q9. What regression model is suitable to predict the gender of the customer from the dataset?**

Random Forest Regressor is suitable to predict the gender of the customer from the dataset. This is because it has obtained the lowest scores in mae and mse, as well as obtained the highest scores in ev and r2 among the models in this project.

**Q10. Which dryer is often used after the customer uses the washer?**

Only one rule is obtained from apriori algorithm. Dryer no.7 is often used after washer no.3.

**Q11. What type of customers tend to use the laundry shop based on per hour?**

Morning:

* Wear black, blue\_jeans, yellow, white, brown and grey pants.
* Wear orange, blue, green, red, purple, pink pants.
* Wear grey, orange, green, purple and pink shirts.
* Age between 45 to 55.

Afternoon:

* Wear orange, blue, green, red, purple, pink pants.
* Wear blue, white, red, black, brown and yellow shirts.

Night:

* Wear black, blue\_jeans, yellow, white, brown and grey pants.
* Wear grey, orange, green, purple and pink shirts.
* Age between 30 to 40.

*Chapter 4: Deployment*

**Q12. What is used for deployment?**

It is deployed and hosted on Heroku.com using the Python Streamlit package.