1. Introduction

Through an EDA process, the associations of given variables with the hourly coffee sale ($) were addressed. The findings and conclusions for the EDA process will be demonstrated in detail with following sections. On the basis of the EDA findings, a linear regression model, in order to predict the hourly coffee sales, was built, evaluated and explained in a statistical analysis. In the last part of the report, a critical thinking reflection for the EDA and modelling works are summarized.

2. EDA

2.1. Pre-processing dataset

By joining all data from 3 tables in the original database, selecting the pertinent variables, and calculating total charges per order by multiplying 'unit_price' and 'quantity' variables, a data frame with 27,171 rows and 10 columns was eventually obtained, named by 'cafe'.

	id	days_after_open	day_of_week	hours_after_open	drink_id	raining	name	dist_to_cafe	name	oder_price
0	0	0	Mon	0	5	No	Abercrombie (H70)	700	Flat White (L)	4.9
1	1	0	Mon	0	5	No	Law Library (F10)	84	Flat White (L)	4.9
2	2	0	Mon	0	2	No	Carslaw (F07)	150	Macchiato	3.8
3	3	0	Mon	0	2	No	Peter Nicol Russell (PNR)	950	Macchiato	3.8
4	4	0	Mon	0	5	No	Carslaw (F07)	150	Flat White (L)	4.9
...	...	...	...	...	...	...	...	...	...	...
27166	27709	153	Sun	5	14	No	Abercrombie (H70)	700	Mocha (S)	9.0
27167	27710	153	Sun	5	6	No	Abercrombie (H70)	700	Flat White (S)	4.0
27168	27711	153	Sun	5	12	No	Fisher Library (F03)	70	Chai Latte (S)	3.6
27169	27712	153	Sun	5	9	No	Brennan MacCallum (A18)	350	Cappuccino (L)	4.9
27170	27713	153	Sun	5	9	No	Abercrombie (H70)	700	Cappuccino (L)	4.9

27171 rows × 10 columns

To pertinently analyze questions, branches of data frames will be grouped from the major data frame, based on variables needed.

2.2. Data

A list of explanations for variables is give as follows:

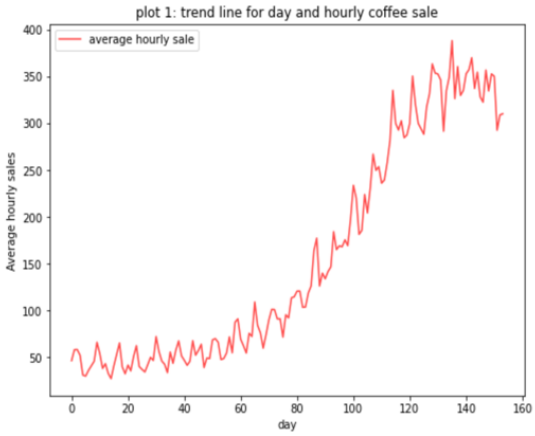
- **id**: unique transaction id (int)
- **date**: transaction date (string)
- **days_after_open**: number of days since opening the pop-up store on 2019-07-22 (int)
- **day_of_week**: day of the week (string, 'Mon' - 'Sun')
- **hours_after_open**: number of hours since opening at 7pm (int, 0 - 5)
- **drink_id**: id of the drink being purchased (int, 0 - 16)
- **raining**: whether it is raining at the time of purchase (string, 'Yes', 'No', missing indicated by 'NA')
- **name**: building name (string)
- **dist_to_cafe**: distance to cafe insomnia (in meters) (int)
- **name**: drink name (string)
- **order_price**: unit_price * quantity (float)

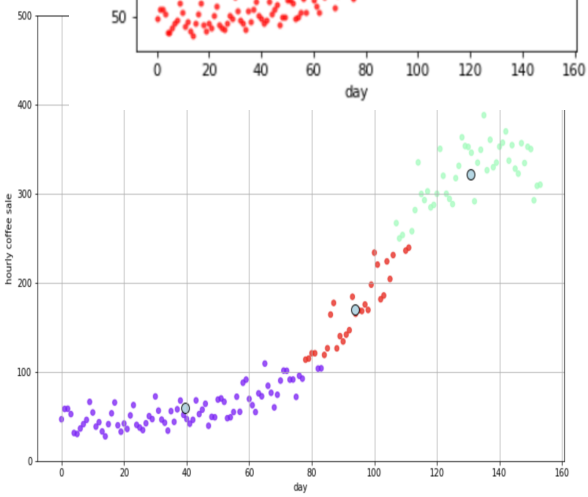
2.3. Number of Day opened with hourly coffee sale ($)

The pivotal variables regard to the day opened and hourly coffee sale are 'day_after_open' and 'order_price' and 'hours_after_open'. 'cafe' is reorganized by grouping the number of day opened, and calculating each hour's sale in each day, the following data frame is the end product.

	day	0	1	2	3	4	5	avg_hsale
0	0.0	49.0	83.6	37.3	50.6	16.1	43.5	46.7
1	1.0	65.8	41.4	63.3	95.3	46.8	37.3	58.3
2	2.0	92.2	67.1	81.7	33.4	34.1	42.4	58.5
3	3.0	39.1	73.2	67.3	67.0	56.9	11.6	52.5
4	4.0	26.2	46.4	48.8	20.6	41.2	4.4	31.3

The data frame is named by 'new_df', with 154 rows and 8 columns. The first and last columns are the number of days opened and average hourly sale daily separately. The line plot visualization of these 2 variables is demonstrated as left. Following with the scatter plot visualization. From the 2 plots, an upward trend between hourly coffee sales and number of days open can be observed clearly. Also, for the early stage of the Pop-up store opening, the growth of coffee sales ($) is relatively slow and stable. Between day around 65 and day around 130, there was an incredible growth of coffee sales ($). Since then, the trend reached a peak and might have come into a recession stage. The k-means clustering algorithm, with k = 3 generates clusters demonstrated on the left below (Plot 4). Each cluster represents a stage of the operational period - beginning stage, boom stage and peak stage. Even though the sale in $ grown fast in the past days, and sharply increased in the boom stage (Cluster in red color points). The growth rate, however, was calculated as a diminishing series of numbers, which are 4.2%, 3.42% and 0.99% separately in each stage. In accordance with the diminishing growth, the hourly coffee in the day after 154 might merely increase, even though with a high base of hourly sale, thus, there might be a non-linear relationship between hourly coffee sale and day opened in the long term. But in the short term, only linear relationships will be discussed. The statistical description for the linear association between hourly coffee sale and day open explains that one more day open is associated





OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.868
Model:	OLS	**Adj. R-squared:**	0.867
Method:	Least Squares	**F-statistic:**	997.4
Date:	Sat, 05 Jun 2021	**Prob (F-statistic):**	1.16e-68
Time:	19:59:20	**Log-Likelihood:**	-796.44
No. Observations:	154	**AIC:**	1597.
Df Residuals:	152	**BIC:**	1603.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-31.3079	6.883	-4.548	0.000	-44.907	-17.709
x1	2.4569	0.078	31.582	0.000	2.303	2.611

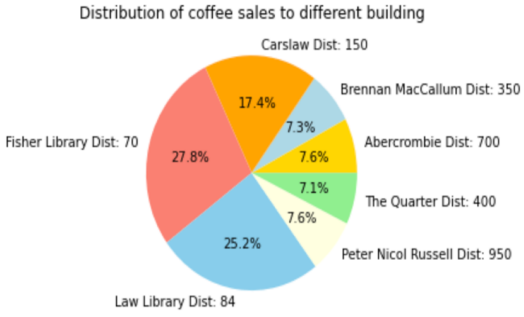
Omnibus:	24.578	**Durbin-Watson:**	0.225
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	7.031
Skew:	0.165	**Prob(JB):**	0.0297
Kurtosis:	2.006	**Cond. No.**	176.

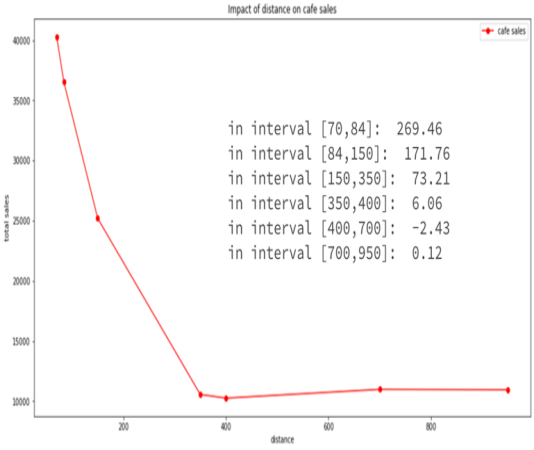
with $2.46 more hourly sale for coffee. The p-value of coefficient associated with day open is equal to 0, thus, the effect of day opens in hourly coffee sales is statistically significant and is not due to the random variation of the sample data. The R^square of the model take value of 0.868, which indicates that approx. 87% of error term and explainable, in additional, the value of adjusted R^square is almost equal to the value of R^square, thus the model is valid and free from the impact of error term on accuracy of prediction. Comprehensively consider all attributes, summarized by statmodel, above, the result is precise and the model is valid, thus the positive association between day open and hourly sale can be proved.

2.4. Where on campus their customers are studying and whether it is raining

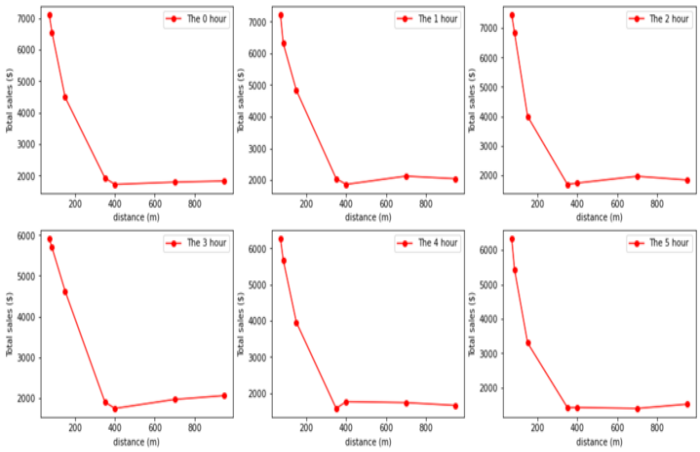
2.4.1. Distance

After extracting information from data frame ‘cafe’ and checking the total sales based on distance between buildings and cafe store, through the visualization by a pie chart, it can be concluded there is a clear relationship. Location does matter, showing as the closer the building to the cafe, the higher sales the shop can earn. A more suitable location has better response. However, the influence of sales for farer locations gradually decreases (from 7.6% to 7.1%). It seems the farer the store away with building, the smaller impact of distance on cafe sales. Hence, a further justification is presented by a line graph, complying with numerical findings. Before 350 meters, especially in intervals [70,84] and [84,150], the impact of distance is tremendously huge. On the grounds of those, there may be a strongly non-linear relationship between distance to cafe and coffee sales in interval [0,350] and might be slightly correlated





or not correlated after 350 meters. Further justification is required for whether the hourly sales would convey the same information.



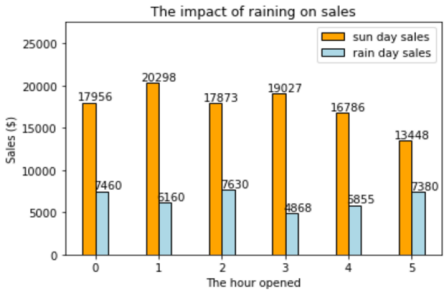
. Here the total sale is segmented by opening hour into 6 parts, but the trend lines show no significant differences.

Thus, above conclusion can be asserted: there may have strongly non-linear relationship between distance to cafe and coffee sales in interval [0,350] and might be slightly correlated or not correlated after 350 meters far away to cafe.

2.4.2. Raining

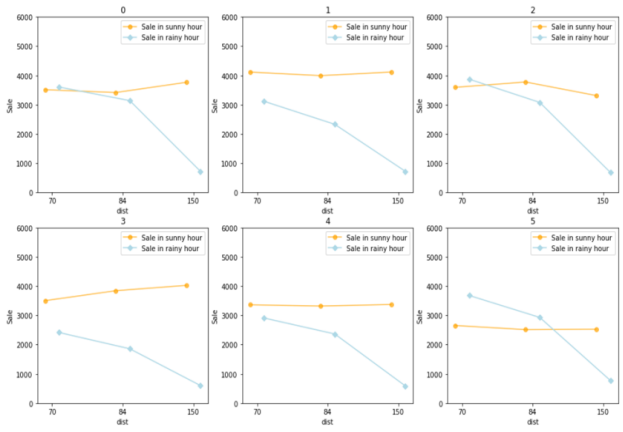
Again, after extracting useful information from data frame 'cafe', numerical calculations can be made and hence see the impacts of whether raining or not does affect coffee sales by using a bar chart for visualization. Here the gap between sunny days and rainy days is very huge, no matter how long the cafe opened. Although the difference varies to some extent, the weather condition does have a significant impact on coffee sales, with an average reduction of approximately 62% in hourly sales due to rain. For further exploration, whether the location (if students are studying in libraries near Pop-up stores) would affect sales on raining days, and whether the influences are still significant are taken into consideration. Through separation of related information, it shows that students only buy coffee when they are studying at these 3 libraries which are the closest to the Pop-up store on raining days. Otherwise, they will not buy coffee. Hence, only data for these 3 libraries will be used for the following discovery.

```
The average reduction in hourly sale due to rain is 61.68%
```



hours_after_open	raining	dist_to_cafe	order_price
0	No	70	3509.5
0	No	84	3413.1
0	No	150	3762.0

hours_after_open	raining	dist_to_cafe	order_price
0	Yes	70	3599.8
0	Yes	84	3137.0
0	Yes	150	723.0

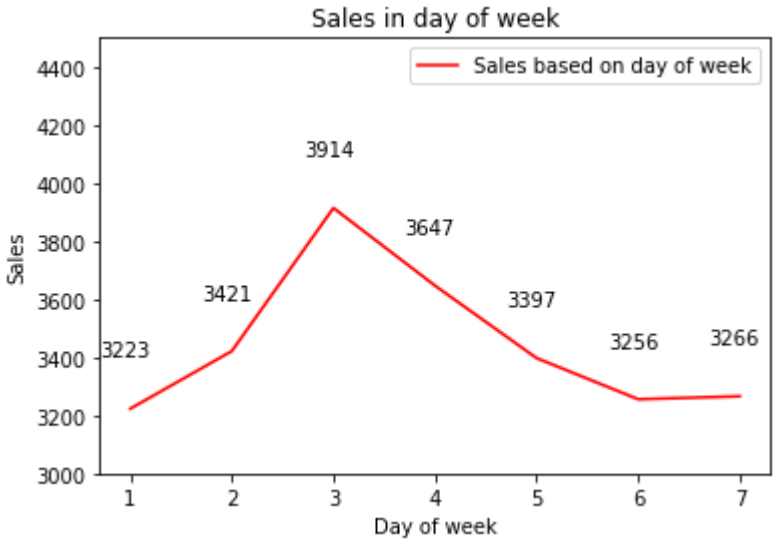


Through calculation, the average reduction in sales due to rain in the first 2 closest libraries (< 90 meters) is 2.76%, in the first 3 closest libraries (< 151 meters) is 28.77%, and the reduction in the third library is 80.78%. As shown in the line graph, if the Pop-store is close enough to the building, the raining weather has a slight impact on coffee sales, and also has a strong impact if the distance is too far. In conclusion, according to these graphs, the coffee sales in sunny hours are smooth to students in the 3 observed libraries. In the rainy hour, the farther from the library to the coffee shop, the fewer students buy coffee in the Pop-up store. When the distance exceeds 150 meters, or in other words, except for the Law library, Fisher library and Claw library, students will not buy coffee in the Pop-up store on a rainy day. Even on a sunny day, there are also few students who will walk a long way to buy coffee in the Pop-up store.

In summary, the location / distance that students are studying to the café influences the coffee sales, and weather conditions (rainy days) also have a tremendous impact on coffee sales (about 62% reduction in sales compared to sunny days). Also, hourly coffee sales ($) has a nonlinear relationship with weather conditions and distance. In this consideration, the statistical model will not be discussed in this part. When considering the distance, if it's < 100 meters between building: weather condition has slight impact on coffee sale (2.76% reduction of sale), students are more likely to make consumption as usual; if it's 100 meters < distance < 150 meters: weather condition has huge impact on coffee sale (about 81% reduction), students are more likely to refuse to go the café and make further consumption; if distance > 150 meters: students do not buy coffee at Pop-up store (100% reduction). In this case, in order to increase the coffee sale and attract more potential customers, the café should pay more attention to students positioning at farer locations and take the weather condition into consideration. Further activities such as promotion may be required.

2.5 The impact of seasonal factor on hourly sales - day of week

In this graph, it's easy to observe that daily sales gradually increased from Monday to Wednesday, peaking on Wednesday and then gradually decreasing till Sunday. Monday and the weekends have lower sales in a week. The correlation coefficient between daily sales and day of the week is 0.3514 which is close to zero, suggesting that there is only a weak linear relationship. The p-



value is 0.022 with a two-tail test, it's smaller than 0.05, therefore the hypothesis of there is no linear relationship can be rejected. In terms of hourly sales, it's the same story as daily sales. We can clearly observe from the line plot and summary that the linear association between hourly sales and day of the week is not strong. Thus, this variable (day of the week) is not going to be the first consideration when modeling a linear regression.

3. Modelling

3.1. Pre-work

The initial data frame applied to the model to predict hourly coffee sale ($) is 'new_df'. Different with 'new_df' previously referenced in section 2.3, the columns containing coffee sales in each hour are deleted and only 2 variables are remaining which are 'avg_hsale' and 'day'. These variables are implemented to build a simple linear regression model by introducing OLS methodology. To improve the performance of the model by lowering the value of mean squared error and avoiding multicollinearity, more variables, but in a limited amount will be taken into consideration, which will be introduced in corresponding sections.

	day	avg_hsale
0	0.0	46.7
1	1.0	58.3
2	2.0	58.5
3	3.0	52.5
4	4.0	31.3
...	...	...
149	149.0	352.5
150	150.0	350.0
151	151.0	292.5
152	152.0	308.7
153	153.0	310.0

154 rows × 2 columns

3.2. Linear Regression

3.2.1. Ordinary Least Squares (OLS) regression

OLS is a common way in machine learning algorithms to directly calculate the exact value of coefficient and intercept for selected
variables. It outputs the same result as the gradient descent algorithm does. The reasons why OLS is eventually applied rather than gradient descent is as follow:

1. OLS requires higher space complexity and time complexity when dealing with large scales of data. Technically speaking, the gradient descent algorithm takes advantage in dealing data with scales over 10,000. But the row dimension in this case is 154.

2. Iteration and gradient are not required with OLS

The normal equation of OLS is:

$$(X.T@X)^{-1}@X.T@y$$

The result of simple variable linear regression is presented as follow:

$$\widehat{Hourly\ Sale} = -25.38 + 2.41 * DayOpen_i$$

The mean squared error (mse) of the model is calculated as 1552.45. The model is extremely simple and has no foresight. Assuming the cafe has opened for 1,000 days, the model then will predict the hourly coffee sale as $2,429. Under the premise of a limited number of students in a university, the predicted result is unrealistic. As section 2.3 previously mentioned, the growth rate of hourly coffee sales was diminishing, as more days opened. Thus, a polynomial term of day could be considered to depict the non-linear relationship. By introducing the quadratic square of day into 'new_df', and substituting into normal equation, the function can be written as:

$$\widehat{Hourly\ Sale} = 40.14 - 0.32 * DayOpen_i + 0.02 * DayOpen_i^2$$

The mse for the polynomial regression is calculated as 880.81. After confirming with no existence of multicollinearity, the performance of prediction for the model shows tremendous improvement than the simple variable linear regression.

In the previous EDA section, there were some other variables, such as raining, day of week and distance to the café, that have been proved that correlated with hourly coffee sales. In order to strengthen the accuracy of the model and reduce the mse further, variable 'raining' is taken into consideration. To fit 27,000 more rows of dummy variables 'Yes' and 'No' into 154 rows based on day, the raining hours on a day are summed up as discrete variables. For instance, there were 2 hours raining on the first day (0), thus, the 'rainingHours' column stores value 2, and presents 2 hours of rain on that day. Introducing normal equation, the regression can be written as:

day	avg_hsale	day^2	rainingHours
0.0	46.7	0.0	2.0
1.0	58.3	1.0	0.0
2.0	58.5	4.0	2.0
3.0	52.5	9.0	2.0
4.0	31.3	16.0	3.0
...	...	...	...
149.0	352.5	22201.0	3.0
150.0	350.0	22500.0	1.0
151.0	292.5	22801.0	5.0
152.0	308.7	23104.0	2.0
153.0	310.0	23409.0	1.0

$$\widehat{Hourly\ Sale} = 52.25 - 0.29 * DayOpen_i + 0.02 * DayOpen_i - 6.66 * rainingHours_i$$

The mse is 760.65 with the model. However, there is strong multicollinearity in the model, thus making the model sensible to little change of variables. Dropping variables such as day^2 incurs a huge cost with MSE of the model, which increase from around 760 to 1400, thus the model eventually will be not used.

In order to find a best solution of mean squared error, some advanced models are tried to be implemented. In consideration of the reason that lasso, ridge and random forest can intelligently balance the weigh average of each variables, in accordance with the associations to dependent variable, distance feature will then be add into the model.

day	avg_hsale	day^2	rainingHours	sum_dist
0.0	46.7	0.0	2.0	134.90
1.0	58.3	1.0	0.0	157.00
2.0	58.5	4.0	2.0	179.22
3.0	52.5	9.0	2.0	161.22
4.0	31.3	16.0	3.0	75.46
...	...	...	...	...
149.0	352.5	22201.0	3.0	793.96
150.0	350.0	22500.0	1.0	1151.88
151.0	292.5	22801.0	5.0	427.32
152.0	308.7	23104.0	2.0	912.04
153.0	310.0	23409.0	1.0	848.34

The data frame is presented as above. The ‘sum_dist’ variable is calculated by summing up the total distance that all orders sold on a day and being divided by 100 to normalize the value.

3.2.2 Lasso

Lasso regression is a type of linear regression that uses shrinkage. Unlike OLS where each parameter item has the same weight, the lasso model encourages a simple, sparse model. After performing the train test split and defining the x and y. Train set is fitted by lassoCV which can generate the sales according to the test set. After fitting the data to the model, coefficients relating to each variable are generated. Variable ‘day’ and ‘rainingHours’ have 0 coefficients which means that the model penalizes these two variables considering that they have no relationship to the sales. The lasso model has performed a variable selection by only retaining day square(day^2) sum of distance (sum_dist) and as the variable. The mean square error of the Lasso model is 433.31 which is smaller than OLS. However, just using one variable cannot predict the sales accurately.

3.2.3 Ridge

Ridge is a similar model to Lasso, the only difference is that Ridge won’t perform a variable selection by penalizing the coefficient of variable to 0. Ridge will retain all variables but just shrink those variables which show less relationship to the sales. Not surprisingly, the Ridge model returned three coefficients (-0.29, 0.02, -5.41). The prediction of Ridge has a slightly low mean square error (355.64) compared with Lasso

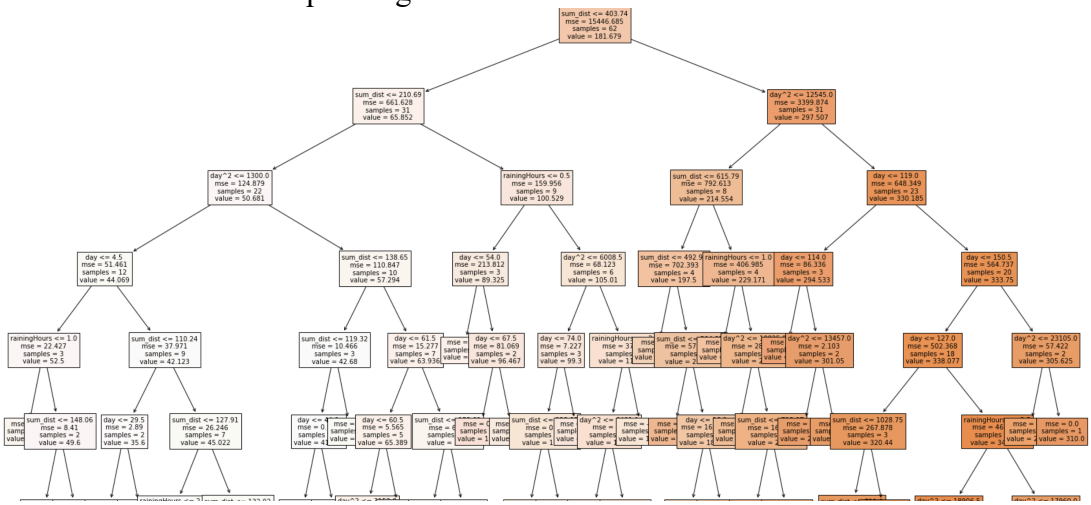
(433.31). The accuracy of model prediction outperforms Lasso and OLS, simply because Ridge captures more characteristic variables, while Lasso removes most of the potentially irrelevant variables through altering the coefficients, which enhances model robustness but may cause severe bias.

3.3 Advanced tree based model

3.3.1 Random Forest

Random Forest is a powerful algorithm which is capable of performing regression. It's an ensemble method in a tree based model which reduces the variance and improves performance. This is a nonlinear model, it's a combination rule of if, then and else. By fitting it with the train data set, the model returns the best estimator. After applying those estimators, the forest is constructed. The decision tree model can predict the daily sales. When evaluating the model, it has the lowest mean square error of 174.35, this is better than all the other models. For this reason, it's the final model that can be used to predict hourly sales. (Note: All mean square errors mentioned above are subject to fluctuations ±5% because of the volatility of the model.)

This graph is a single decision tree extract from the random forest. The model flow can be easily visualized and applied to the future decision of the cafe. Such as location selection and modified operating hours.



The more data and higher dimensionality in the dataset, the less likely that Random Forest will over fit the model. For our data set, only 62 observations available for model fitting, the size of observation and dimensionality are tended to be small. So, there is still a risk of overfitting the model.

4. Reflection

Overall, this report does follow the process of the CRISP-DM model to some extent. This model is found to be relatively effective, since it clearly shows different procedures for distinct working requirements with a convincing logic base.

A data mining project requires lots of different skills and knowledges to address strange problems. CRISP-DM (Cross Industry Standard Process for Data Mining) suggested a comprehensive methodology to process data mining projects (Wirth and Hipp, 2000). Wirth and Hipp in 2000, outlined 6 sections of CRISP-DM process, which are:

1. Business Understanding

The very beginning step of a data mining project. The objective of this step is to understand and address the requirements of the needed information from a business perspective. And then, transfer those needs into data mining problems.

2. Data Understanding

Collecting data, getting familiar with data, addressing data quality issues, and exploring the data to search the hidden information in the dataset. This step and data preparation step correspond to the EDA (Explanatory Data Analysis) process in this report, more detail will be talked further.

3. Data Preparation

The final target of data preparation is to construct the final dataset.

4. Modelling

Appling technical tools (regression, clustering, support vector machine, neural networks etc.) to train the dataset, which is constructed in previous stages.

5. Evaluation

Estimating the model through the accuracy of prediction and mean squared error between predicted values and actual values. Also considering the advantages and disadvantages of the model, and accessing the possibility of overfitting or under fitting. In order to reduce the risk of overfitting the model, there is a potential solution by stacking several models together and output the prediction based on the proportion of each model's MSE.

6. Deployment

Summarizing and presenting the findings from above stages in the form of a report, may give advice for future operation and improvement.

In this report's case, the ultimate goal for the client is to open a cafe that only operates at night. In order to predict and determine the potential outcome, the client opened a Pop-up store to see what actually affects the final (hourly) sales. By collecting various data given in the dataset and through the process of data mining, the data and outcomes from this report tries to see the relationship and extents between those factors and hence to improve the sales or give potential advice for doing that (if there is clearly an obvious problem which can be make up for).

Following the process of CRISP-DM, this report introduced relevant useful data (data understanding) in the beginning, and extracted useful information with quality for further investigation. After considering mainly 3 parts: number of days, weather condition & distance and seasonal factors, this report identified the preliminary correlation and deduced whether a corresponding model is suitable to make (data preparation). Through the model building (modelling) and detailed description about the models mentioned above (evaluation), this report did provide useful information required by the client. What needed to mention is this report proposed the visualization of model flow, the client can predict sales and maximize profits through the application of this figure.

For future operation and improvement (deployment), the client should consider the effects of the number of days opened, pay attention to the weather condition (deduce whether the opening hours should be reduced in this special circumstances), develop new activities or make promotions (price reduction) to attract specifically the students in farer locations

Overall, the logic and connection between each stage of CRISP-DM process are well considered in this report. But there is still space of improvement. For instance, because of the data frames that required in each analysis are different, and those are not fully prepared in a one segment. That is, when moving onto a new variable, a new data frame is needed to be grouped, rather than directly call a pre-prepared data frame. As Jaggia and Kelly in 2020 criticized that students generally fail to realize the interplays between stages between each stage (Jaggia and Kelly, 2020). The most serious bug in the report is the storytelling skill. As Jaggia and Kelly stated, data analysis as a technical field, students usually neglect the importance of storytelling, which can covert the process of data analytics process, such as numeric comparing, plotting, modelling, into interesting story. This is a way that is more likely to attract the attention of target audiences.

5. Reference

Wirth, R. & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, Manchester, United Kingdom, 29–39.

Jaggia, S., Kelly, A., Lertwachara, K., & Chen, L. (2020). Applying the CRISP-DM Framework for Teaching Business Analytics. *Decision Sciences Journal of Innovative Education, 18(4), 612–634*.

MINUTES TEMPLATE

Minutes of meeting for __3-5HOURS each time___

Date: __2021.5.14, 5.18,5.26,5.28,6.1,6.4,6.5_____ Time: _Afternoon~Night___

Location: ______zoom online____

Chairperson: YIYU DING, SHENHAO QIAN, HAORAN YAN

Minute-Taker: YIYU DING, SHENHAO QIAN, HAORAN YAN

Document tabled: python, word doc., google doc.

Present: YIYU DING, SHENHAO QIAN, HAORAN YAN

Apologies: No

Agenda Item	Key Points	Action	By Whom	When	Communication Strategy
3 parts in Q1 Modelling Report writing	*How to write *Which perspective *Which extent *Report quality	Discussion Code writing Report writing	ALL	Shown above	brainstorming

Source: TAFE Access Division “Communication for Business”, 2000