Café Insomnia Project Report

Author : Voon Ken Ren

Date : 6th June 2021
Date Updated: 3rd November 2023

Originally created as part of a Postgraduate Business Analytics University Project, it has been subsequently reviewed and enhanced through a self-revision initiative undertaken by the author.

Table of Contents

Introduction

Coffee is widely favoured by university students. This report conducts an in-depth analysis of factors impacting the hourly sales at Café Insomnia, situated within the University of Sydney. Employing the CRISP-DM process model, this study focuses on the precise determinants of hourly sales. Leveraging business analysis and statistical insights, the report employs exploratory data analysis (EDA) to identify influential factors. Additionally, it documents the methodology for selecting the optimal linear regression model, with an emphasis on minimising regression mean squared error (MSE) for accurate hourly sales predictions at Café Insomnia

Business understanding

Before starting our analysis, we first need to understand the business goals. Based on the project description given, it appears that the café owners seek to determine whether a café that only opens at night is a profitable business. Data collected from the pop-up store will be provided to us by the café owners to use. The owners have given us two tasks to accomplish. The first is to identify any key factors that may affect the hourly sales of the café, both positively and negatively. The second is to determine if it is possible to build a model that predicts hourly sales.

With this, we have determined that the business objective is to "***help the business identify factors that could maximise profitability".*** We will be performing our analysis under the assumptions that all customers were students studying at the university (not necessarily USYD students) and that all transactions present in the dataset were successful.

Before embarking on this analysis, it is essential to establish an understanding of the project's business goals. As outlined in the project brief, the objective is to evaluate the feasibility of operating a night-only café at the University of Sydney Camperdown campus. The café owner has generously provided access to data collected from the pop-up store, and two specific tasks have been assigned:

1. identify key factors influencing the café's hourly sales, both positively and negatively;
2. develop a model for predicting hourly sales.

With these objectives in mind, the primary business goal is to **support the establishment in optimising its profitability**. For the purpose of this analysis, it is assumed that all patrons are university students without specific university affiliations. This assumption is based on the close proximity of the Camperdown campus to several other educational institutions, and the higher likelihood of students being after-hours patrons compared to campus staff. Additionally, it is assumed that all transactions within the dataset indicate successful transactions.

We have established a set of success criteria based on the two key tasks:

1. **For Task 1:** To compile a list of variables influencing Café Insomnia's hourly sales through exploratory data analysis (EDA) and to provide corresponding analysis on how these variables impact hourly sales.
2. **For Task 2:** To develop a predictive model for hourly sales or provide an explanation if such a prediction is not feasible.

Constraints: For the EDA component, the analysis is limited to the use of Python packages covered in the relevant tutorials (BUSS6002).

Resource Overview and Ethical Considerations

Resources

- Café Insomnia dataset (cafe_insomnia.sqlite)
- Corresponding Data Dictionary (data_dictionary.txt)
- Python, sqlite3, NumPy, Matplotlib, Statsmodels, Scikit-learn
- Jupyter/Google Colab Notebooks
- BUSS6002 Teaching Team & Cohort (supervision/technical support, peer support)

Ethical Implications

In the context of this study, it is important to note that the dataset used does not contain any personally identifiable customer information, thereby mitigating potential ethical concerns. Moreover, customers willingly provided specific details, including their place of study and weather conditions, via an iPad at the point of sale. It is reasonable to assume that customers provided this information with informed consent.

Given the intended audience for this report, we include a technical glossary of terms frequently used throughout the document for clarity:

Technical Glossary (Definitions from Oxford Languages):

- **Data set:** A collection of interrelated pieces of information that, while consisting of separate elements, can be manipulated as a unified entity by a computer.
- **Attributes/Variables**: Data points that define the characteristics of a field or tag in a database or a string of characters in a display.
- **Data dictionary:** A comprehensive description of a database's content, format, structure, and the relationships among its elements, serving as a reference for database access and manipulation.
- **Records/Rows:** Groups of related information items treated as a single unit.
- **Data Cleaning:** The process of identifying and rectifying inaccurate/invalid, incomplete, duplicated, poorly formatted, irrelevant, or unnecessary data or customer records within a database (Experian, 2021).
- **Dummy Variable:** A numeric variable representing categorical data, such as gender, race, or political affiliation (Stat Trek, n.d.).
- **Multicollinearity:** The presence of significant intercorrelations among two or more independent variables in a multiple regression model, which can lead to skewed or misleading results (Hayes, 2021).

Subsequently, this report will proceed to delineate the step-by-step methodology employed in our analysis and the successful execution of our project plan.

EDA (Q1) : What affects Café Insomnia's hourly sales (in AUD)?

Data understanding

Prior to commencing our analysis, a preliminary meta-analysis was conducted on the supplied dataset. It was revealed that the dataset encompasses **27,714** records spanning from **July 22, 2019, to December 22, 2019 (5 months)**. In response to the primary inquiry, an exploration of the dataset's attributes is imperative. This section will expound upon the methodology employed to identify a subset of variables of significance and delineate the criteria for determining their suitability for further investigation.

Unit Price

Represented as *'unit_price'* in the dataset, and defined as 'unit price of the drink in AUD (float)' in the provided data dictionary, the 'price' variable assumes significance. Often, lower-priced or more cost-effective products tend to attract university students, potentially influencing Café Insomnia's sales performance. This prompted an initial assumption that the 'price' variable merited investigation.

However, it's worth noting that the initial assumption was based on the premise that competitive pricing might substantially impact sales, especially by drawing students as the primary customer base. Yet, it is essential to acknowledge that the dataset lacks specific details for distinguishing between student and non-student purchases. Consequently, further exploration of this variable has been deemed unfeasible.

In light of this limitation, the analysis will shift focus towards other variables that permit more reliable and comprehensive assessment, aimed at uncovering insights into Café Insomnia's sales performance.

Days after open

Represented as *'days_after_open'* in the dataset, and defined as the 'number of days since opening the pop-up store on 2019-07-22 (int)' in the provided data dictionary, this variable carries a noteworthy significance.

Café Insomnia, being a pop-up establishment, aligns with a prevailing trend among young consumers, notably students, who are increasingly drawn to pop-up shops, often driven by the fear of missing out (Baras, 2021). This trend exhibits a positive correlation between pop-up stores and their brand image, resulting in a sales growth pattern.

Research findings underscore the fundamental factors that drive the success of pop-up shops, even if these factors may seem self-evident (Gonzalez, n.d., as cited in retailTouchPoints):

- **Well-established brand images:** Pop-up shops with strong brand recognition are more likely to succeed. This principle forms the foundation of the analysis.
- **Sales growth:** Similarly, the importance of consistent sales growth for the success of pop-up shops may seem like stating the obvious. However, this principle guides the basis for understanding the contributing factors to success in this context.

This positive correlation between these factors and success is particularly noticeable during a pop-up shop's initial phases. In essence, when a pop-up shop is newly opened, it can greatly benefit from having a well-established brand image and achieving consistent sales growth.

Now, applying these insights to Café Insomnia, a newly established coffee shop without a well-established brand image, we can infer that it too may experience a positive correlation between its sales and the duration of its operation. While Café Insomnia lacks an established brand image, the general trend observed in the research suggests that the longer Café Insomnia remains open, the greater the potential for its sales to increase. This inference is based on the broader patterns seen in the research regarding pop-up shop success

With these insights in mind, the 'days_after_open' variable has been included in the analysis to investigate its potential impact on the sales performance of Café Insomnia.

Day of week

Represented as ‘day_of_week’ in the dataset, and defined as “day of the week (string, 'Mon' - 'Sun')” in the provided data dictionary.

According to industry analysis, hourly sales of cafés may vary depending on different days of the week. Cafés tend to have their lowest sales on Tuesdays and their highest on the weekends. This difference can be as high as 20% (Sobelman, 2021). Therefore, we have decided to include 'day_of_week' as part of our investigation.

Hours after open

Represented as 'hours_after_open' within the dataset, this variable is defined as the 'number of hours since opening at 7pm (int, 0 - 5)' in the provided data dictionary.

The operational hours of a café constitute a determinant capable of influencing its sales performance. As observed by Nong (2020), the timing of the establishment's opening plays a critical role in shaping customer perceptions. In the case of Café Insomnia, operational hours span from 7 PM to 1 AM, thoughtfully designed to cater to the convenience of students engaged in nocturnal studies within the campus library.

Situated on a university campus, the café's operational hours are strategically aligned with the presence of students engaged in evening academic pursuits. This alignment underscores the potential impact of the academic community on Café Insomnia's sales. Consequently, 'hours_after_open' emerges as a salient factor to be thoughtfully incorporated into the predictive framework for coffee shop sales analysis.

Weather and Distance

In the dataset, the variables 'raining' and 'dist_to_cafe' are represented, defined respectively as “whether it is raining at the time of purchase (string, 'Yes', 'No', missing indicated by 'NA')” the presence of rain at the time of purchase ('Yes', 'No', or 'NA' for missing values) and the distance to Café Insomnia in metres (int) according to the provided data dictionary.

Given the University of Sydney's campus layout, there may be limited shelter between Café Insomnia and the various campus libraries. Consequently, it is reasonable to assume that weather conditions could influence café sales due to their potential impact on consumers' mood and behaviour (Bujistic et al., 2019). Notably, research indicates that convenience is a significant factor affecting consumers' coffee purchase decisions (Burge, 2013). Thus, the presence of rain is expected to decrease café sales, which will equate with an increased 'distance' from a convenience perspective, as greater distance implies reduced convenience for consumers.

Drink Type

Notably, statistics indicate that Latte, Flat White, and Cappuccino are the most popular coffee choices among Australians, collectively constituting nearly 80% of preferences (Kilroy, n.d.). Consequently, it is reasonable to assume that the hourly sales at a café may be influenced by the different types of coffee offered. The variable selection process, informed by initial business understanding and metadata analysis, includes the following factors (variables):

- days_after_open
- day_of_week
- hours_after_open
- weather (raining)
- dist_to_cafe
- drink type (name)

Data preparation

The dataset, an SQLite file, consists of three tables: **'ci_transaction**,' **'study_area**,' and **'drink**.' The provided data dictionary indicates missing values in the **'raining'** and **'study_area_id'** variables.

The disjointed structure of the dataset poses challenges for meaningful data analysis and complicates the examination of relationships between the tables. The presence of missing values introduces data uncertainty.

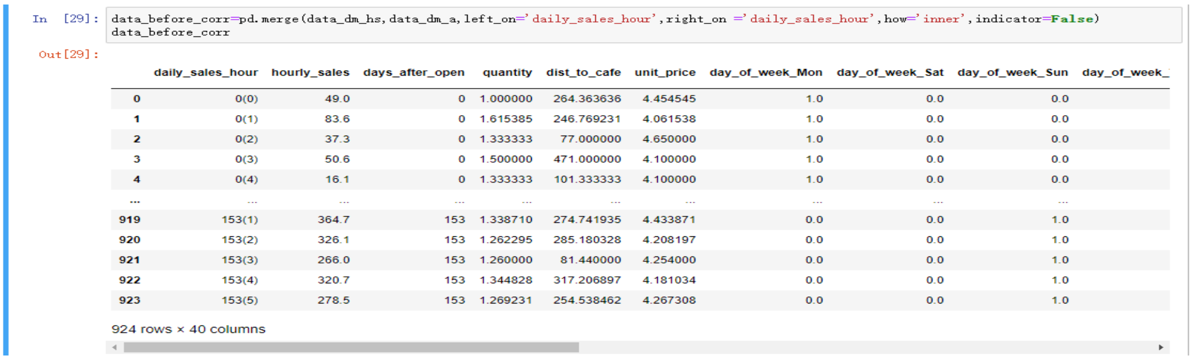
Before initiating the analysis, it is essential to perform data cleaning and transformation. The SQLite file format allows this to be accomplished in a single step. The merging of the three tables was executed using an **SQL INNER JOIN** query, which facilitated the selection of desired variables and their consolidation into a **single pandas dataframe**. This approach also effectively removed missing values by selecting only rows that match between the tables (W3Schools, n.d.).

Upon examination, **543 missing values** were identified, which, upon removal, resulted in a remaining sample size of **27,171 transaction records.**

The analysis is aimed at understanding the factors influencing the hourly sales at Café Insomnia. This involves combining '**days after open**' and **'hours after open**' into a new variable called '**daily sales hour**.' Hourly sales were calculated by multiplying the '**unit price**' by the **quantity** for **each transaction**. The variable of interest was then associated with 'daily sales hour' to compute total hourly sales and average hourly sales.

To mitigate bias during the regression process, certain categorical variables, including **'day_of_week**,' **'raining**,' and **'drink_name**,' were transformed into dummy variables. Notably, 'hours_after_open' was retained as a continuous independent variable and not converted into dummy variables (LAERD Statistics, 2021).

The next step involves comparing the correlations between hourly sales and other factors listed in the table. To conduct these comparisons, certain variables need to be transformed into an hourly basis, similar to the procedure applied to 'hourly sales.' An example of the typical process for this task is displayed in the screenshot below.



```
In  [29]: data_before_corr=pd.merge(data_dm_hs,data_dm_a,left_on='daily_sales_hour',right_on ='daily_sales_hour',how='inner',indicator=False)
          data_before_corr
```

Out[29]:

	daily_sales_hour	hourly_sales	days_after_open	quantity	dist_to_cafe	unit_price	day_of_week_Mon	day_of_week_Sat	day_of_week_Sun	day_of_week_
0	0(0)	49.0	0	1.000000	264.363636	4.454545	1.0	0.0	0.0	
1	0(1)	83.6	0	1.615385	246.769231	4.061538	1.0	0.0	0.0	
2	0(2)	37.3	0	1.333333	77.000000	4.650000	1.0	0.0	0.0	
3	0(3)	50.6	0	1.500000	471.000000	4.100000	1.0	0.0	0.0	
4	0(4)	16.1	0	1.333333	101.333333	4.100000	1.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	
919	153(1)	364.7	153	1.338710	274.741935	4.433871	0.0	0.0	1.0	
920	153(2)	326.1	153	1.262295	285.180328	4.208197	0.0	0.0	1.0	
921	153(3)	266.0	153	1.260000	81.440000	4.254000	0.0	0.0	1.0	
922	153(4)	320.7	153	1.344828	317.206897	4.181034	0.0	0.0	1.0	
923	153(5)	278.5	153	1.269231	254.538462	4.267308	0.0	0.0	1.0	

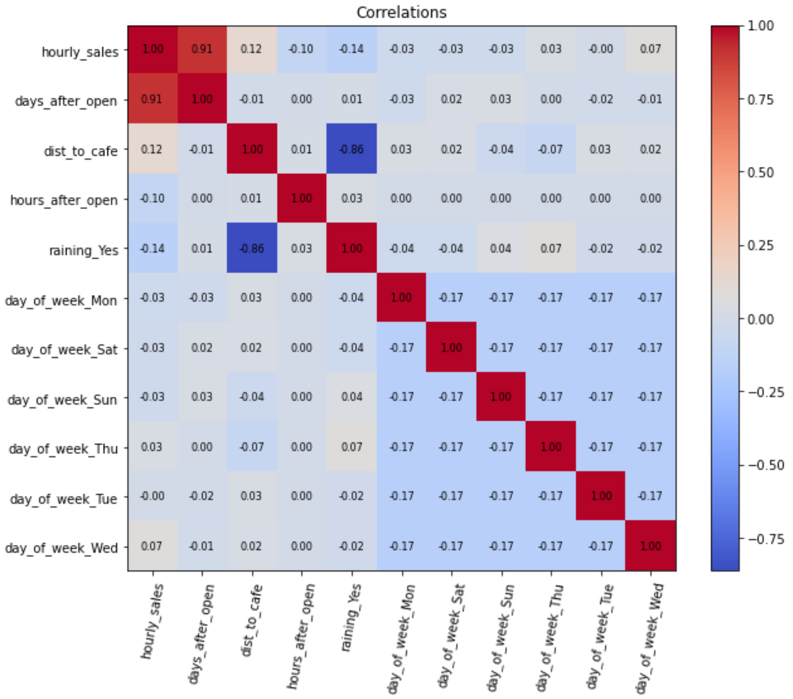
924 rows × 40 columns

Figure 1. Screenshot Snippet of Jupyter Notebook Code

The subsequent section outlines the methodology employed to identify variables that exert a significant influence on hourly sales. Furthermore, this process serves as a means of validating the assumptions posited in the Business Understanding section. The analysis assesses the impact of several variables on Café Insomnia's hourly sales. These variables encompass the number of days after the establishment's opening, seasonal elements (such as day of the week and hour of the day), drink name, the customer's study location, and the presence of rain.

Correlation analysis

Hourly sales result from the multiplication of quantity and unit price. Consequently, both these variables are not included as independent factors in the model used to predict Café Insomnia's hourly sales. Correlations were then computed, and a heatmap was generated to illustrate these findings, as shown below.



	hourly_sales	days_after_open	dist_to_cafe	hours_after_open	raining_Yes	day_of_week_Mon	day_of_week_Sat	day_of_week_Sun	day_of_week_Thu	day_of_week_Tue	day_of_week_Wed
hourly_sales	1.00	0.91	0.12	-0.10	-0.14	-0.03	-0.03	-0.03	0.03	-0.00	0.07
days_after_open	0.91	1.00	-0.01	0.00	0.01	-0.03	0.02	0.03	0.00	-0.02	-0.01
dist_to_cafe	0.12	-0.01	1.00	0.01	-0.86	0.03	0.02	-0.04	-0.07	0.03	0.02
hours_after_open	-0.10	0.00	0.01	1.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
raining_Yes	-0.14	0.01	-0.86	0.03	1.00	-0.04	-0.04	0.04	0.07	-0.02	-0.02
day_of_week_Mon	-0.03	-0.03	0.03	0.00	-0.04	1.00	-0.17	-0.17	-0.17	-0.17	-0.17
day_of_week_Sat	-0.03	0.02	0.02	0.00	-0.04	-0.17	1.00	-0.17	-0.17	-0.17	-0.17
day_of_week_Sun	-0.03	0.03	-0.04	0.00	0.04	-0.17	-0.17	1.00	-0.17	-0.17	-0.17
day_of_week_Thu	0.03	0.00	-0.07	0.00	0.07	-0.17	-0.17	-0.17	1.00	-0.17	-0.17
day_of_week_Tue	-0.00	-0.02	0.03	0.00	-0.02	-0.17	-0.17	-0.17	-0.17	1.00	-0.17
day_of_week_Wed	0.07	-0.01	0.02	0.00	-0.02	-0.17	-0.17	-0.17	-0.17	-0.17	1.00

Fig 2. Correlation Heatmap of selected variables.

Note that the variables for name (drinks) were not included due to limited spacing that would result in hard to read labels. The following analysis will be conducted with the variables selected from the heatmap (including drink name). These variables include ***days_after_open, day_of_week, raining, dist_to_cafe, hours_after_open, drink_name*.**

days_after_open

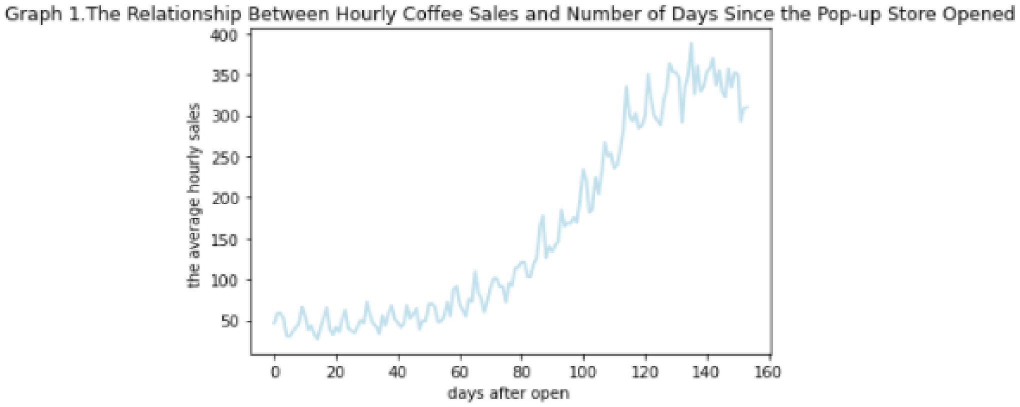


Fig 3. Screenshot snippet of a line graph of days after open vs average hourly sales

Graph 1 demonstrates that within the first 140 days after Café Insomnia's opening, there is a noticeable increase in average hourly sales as the number of days after opening accumulates. Beyond the 140-day mark, the average hourly sales seem to stabilise. This observed trend aligns with the strong positive correlation evident between **'days_after_open'** and '**hourly sales**' in the graph. Consequently, it is advisable to incorporate 'days_after_open' as a relevant factor in the model for predicting Café Insomnia's hourly sales.

day_of_week

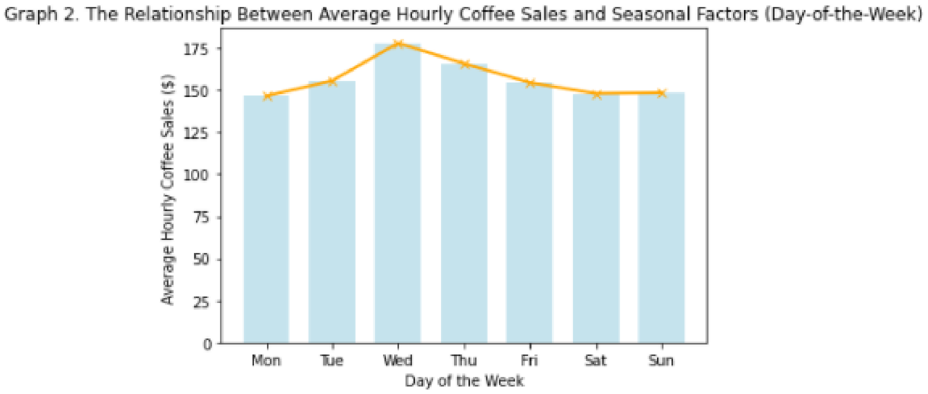


Fig 4. Bar plot with accompanying Line plot for Days of the Week against Average Hourly Sales in ($ AUD).

Graph 2 reveals that the highest hourly sales occur on Wednesdays, whereas the average hourly sales for the other days of the week remains relatively stable around $150. This finding diverges from the initial assumptions outlined in the business understanding section. However, it aligns more closely with the observed correlation pattern. Since the fluctuations in hourly sales are not substantial enough to decisively conclude whether this variable should be included in the model based solely on visualisation, a hypothesis test is needed to determine the significance of this variable.

raining and distance to cafe

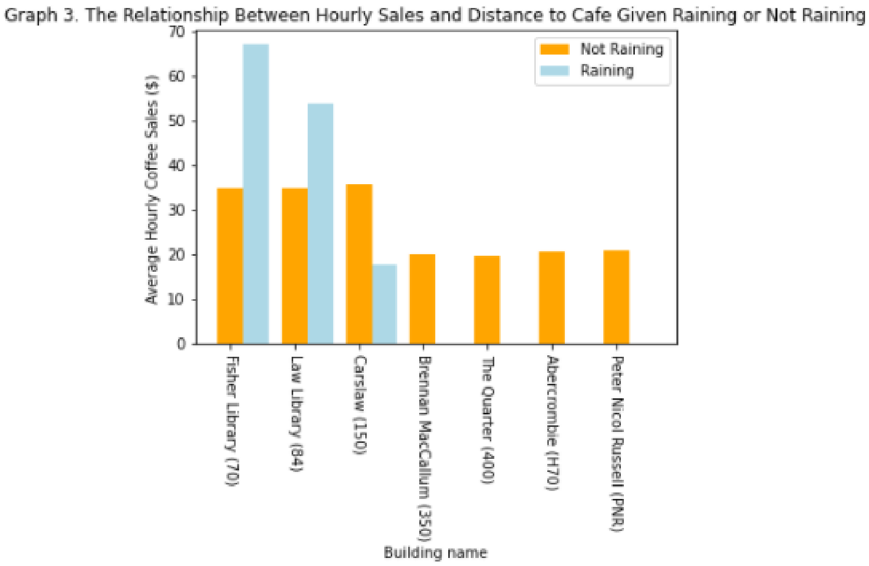


Fig 5. multi-bar plot illustrating the relationship between average hourly sales and distance to the cafe, categorised by rainy and non-rainy conditions.

Graph 3 illustrates the relationship between rain and the distance from the study area to Café Insomnia in relation to average hourly sales. Surprisingly, rain is linked to a surge in Café Insomnia's hourly sales, challenging the initial assumptions detailed in the Business Understanding section. It was previously assumed that a shorter distance to the cafe, which implies less walking, would lead to increased convenience and subsequently higher hourly sales.

Furthermore, students residing within 150 metres from the café significantly contribute to its hourly sales compared to those at greater distances. Conversely, students at specific locations (Brennan MacCallum, The Quarter, Abercrombie, and Peter Nicol Russell) do not purchase coffee from Café Insomnia during rainy conditions, showing a strong correlation (0.81). Hence, one of these factors will be excluded from the model for predicting the café's hourly sales.

hours_after_open

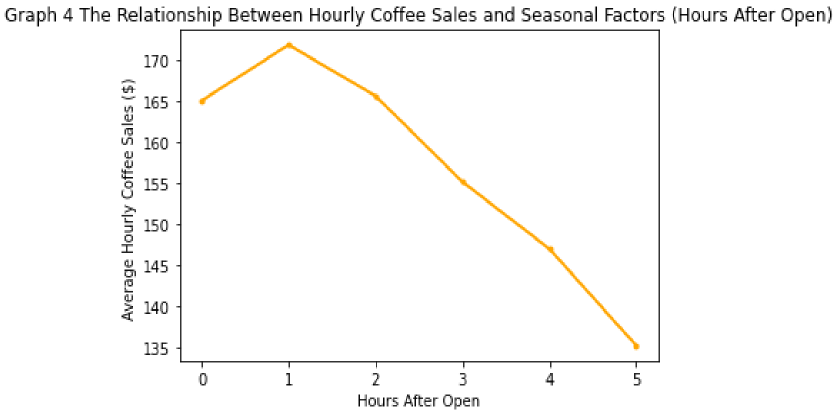


Fig 6. Line plot depicting the relationship between Average Hourly Sales and the Hours After the Cafe Opens.

Graph 4 exhibits a notable trend, illustrating an initial surge in hourly sales during the first hour of opening, followed by a sharp decline as time progresses. The pronounced fluctuations in this pattern align with the correlation pattern observed, affirming that "hours_after_open" significantly influences hourly sales.

drink name

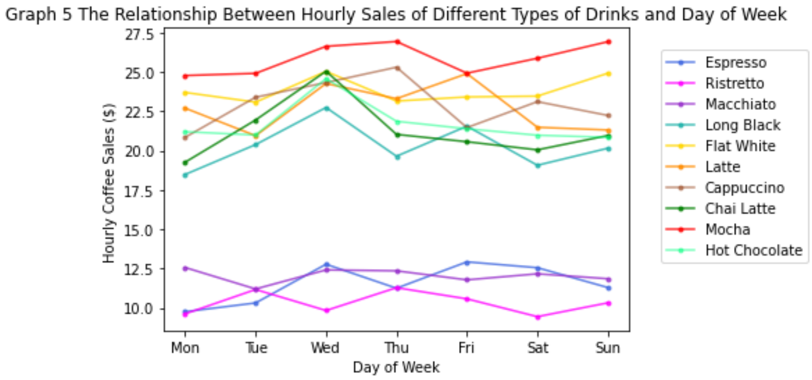
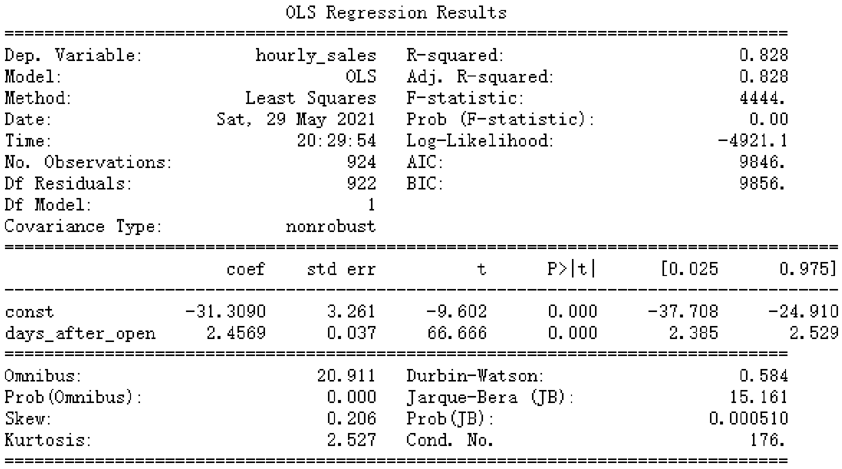


Fig 7. Multiple Line Graph depicting the relationship between Hourly coffee Sales of Different types of Drinks sold at Cafe Insomnia and the Day of the Week.

Graph 5 depicts the weekly sales performance of different coffee types. Macchiato and Espresso have the lowest sales figures, suggesting a limited impact on overall sales. However, Espresso exhibits a significant positive correlation, while Macchiato, with similar sales, shows a negative correlation. This inconsistency justifies excluding these variables from the regression model.

Results of Hypothesis Test

To ascertain the correlation between the selected variables and Café Insomnia's hourly sales, we will perform t-tests on their correlation coefficients. The null hypothesis assumes beta equals zero, while the alternative hypothesis states that beta does not equal zero.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           hourly_sales   R-squared:                       0.828
Model:                            OLS   Adj. R-squared:                  0.828
Method:                 Least Squares   F-statistic:                     4444.
Date:                Sat, 29 May 2021   Prob (F-statistic):               0.00
Time:                        20:29:54   Log-Likelihood:                -4921.1
No. Observations:                 924   AIC:                             9846.
Df Residuals:                     922   BIC:                             9856.
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const            -31.3090      3.261     -9.602      0.000     -37.708     -24.910
days_after_open    2.4569      0.037     66.666      0.000       2.385       2.529
==============================================================================
Omnibus:                       20.911   Durbin-Watson:                   0.584
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               15.161
Skew:                           0.206   Prob(JB):                     0.000510
Kurtosis:                       2.527   Cond. No.                         176.
==============================================================================
```

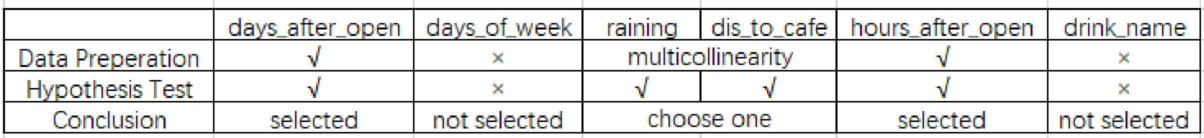
Fig 8. Screenshot snippet of OLS Regression Results of hourly_sales against days_after_open.

Hypothesis testing for "**days_after_open**" establishes its significant relationship with Café Insomnia's hourly sales, consistent with insights from the Business and Data Understanding sections and the correlation analysis.

The results of hypothesis tests on ***raining_Yes***, ***dist_to cafe***, ***day_of_week, hours_after_open*** and ***drink_name*** are shown in Appendix Table A.

From the hypothesis tests, we observed significant correlations between "**raining_Yes**," "**dist_to_cafe**," and "**hours_after_open**" with Café Insomnia's hourly sales, warranting their inclusion in the predictive model.

Conversely, the hypothesis tests indicated that "**day_of_week**" and "**drink_name**" exhibit insignificant relationships with Café Insomnia's hourly sales, leading to their exclusion from the model.



	days_after_open	days_of_week	raining	dis_to_cafe	hours_after_open	drink_name
Data Preperation	√	×	multicollinearity		√	×
Hypothesis Test	√	×	√	√	√	×
Conclusion	selected	not selected	choose one		selected	not selected

Fig 9. Matrix depicting variables to be selected for the regression model.

These tests have led to the decision to include "**days_after_open**" and "**hours_after_open**" in the predictive model. However, due to multicollinearity between "**raining**" and "**dist_to_cafe**," only one of these variables will be included.

Q2: Can a model be built to predict hourly sales (in AUD)?

For the purpose of building a predictive model for hourly sales, a **multiple linear regression** model has been selected. This supervised machine learning technique facilitates the description of the relationship between hourly sales and key independent variables, including "days_after_open," "raining," "dist_to_cafe," and "hours_after_open" (Kurama, 2019). The model allows estimation of the influence of these independent variables on the dependent variable. The choice of a linear regression model is driven by its interpretability and suitability for non-technical audiences.

Multicollinearity between "raining" and "dist_to_cafe" initiated the model selection process. Two distinct models were developed, each incorporating one of these variables and including the significant variables identified through hypothesis testing.

In the context of building a regression model, evaluation was conducted using the mean squared error (MSE). The ultimate choice was determined by selecting the model with the lowest MSE, signifying the most optimal fit.

The model prototype can be defined as:

$$\widehat{Hourly\ Sales}_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i$$

- x_1= days_after_open
- x_2= raining_Yes or dist_to_cafe
- x_3= hours_after_open

The presented model indicates that a change in β1, for instance, leads to a corresponding change in hourly sales while keeping all other independent variables constant.

The model with x_2= raining_Yes was selected as the superior choice due to its lower MSE of **2030** compared to **2105.90**. Therefore, our final model can be defined as follows:

$$\widehat{Hourly\ Sales}_i = -3.14 + 2.46x_1 - 39.18x_2 - 6.35x_3 + \varepsilon_i$$

Where x_1= **days_after_open**, x_2= **raining_Yes**, x_3= **hours_after_open.**

In the above model, Ŷ is the expected value of 'Houry sales'. The regression coefficients associated with **days_after_open**, **raining_Yes** and **hours_after_open** is **2.46, -39.18, -6.35** respectively. For example, this means that for each unit increase in *days_after_open* is associated with a *2.46* increase in *hourly sales* when all other independent variables are held constant.

Assumption Checking

For our tests to be valid, four principle assumptions must hold.

- **Linearity:** Upon examining the residual plot, it is challenging to discern any signs of linearity. A distinct pattern, resembling a parabolic shape, is evident in the graph. Consequently, the linearity assumption is violated, raising the possibility of model misspecification.

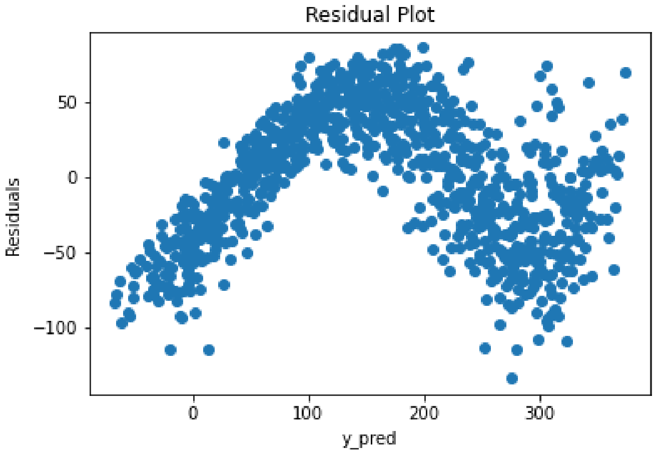


Fig 10. Residual Scatter Plot used to verify for Homoscedasticity and the Linearity Assumption.

- **Independence:** Given that the dataset primarily comprises customer transaction data, an initial assumption was made that each transaction is independent of others, and customer visits occur randomly. While this suggests independence among errors, the presence of time-series data in the dataset elevates the risk of violating the independence assumption. Nonetheless, it can be reasonably assumed that the independence assumption is approximately met.
- **Homoscedasticity:** Based on the scatter plot, the residuals for each fitted value appears to be spread unevenly. Therefore, the constant error variance assumption is violated.
- **Normality:** Upon examining the QQ-plot, it's evident that the points deviate from the diagonal line, with only a small portion aligning with it. This indicates a violation of the normality assumption. However, considering the dataset's size (n = 27171), the Central Limit Theorem applies. Consequently, for the purposes of this analysis, we assume that the normality assumption is not compromised.

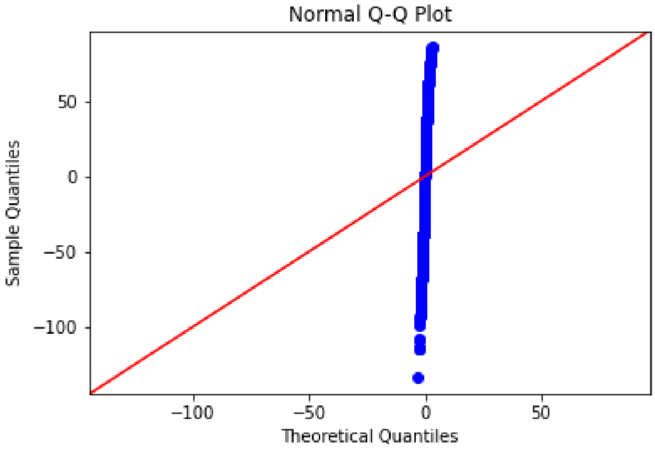


Fig 11. QQ plot used to verify the Normality Assumption.

Regrettably, our model has contravened half of the assumptions. While initially normality assumption was justified through the Central Limit Theorem, the QQ-plot contradicts this assertion, suggesting it may be inaccurate. Transformation attempts to address these violations but, despite these efforts, the assumptions persistently appear to be infringed. It is conceivable that building a **linear regression** model to predict hourly sales with this dataset may not be feasible, a topic further explored in the reflection section.

Reflection

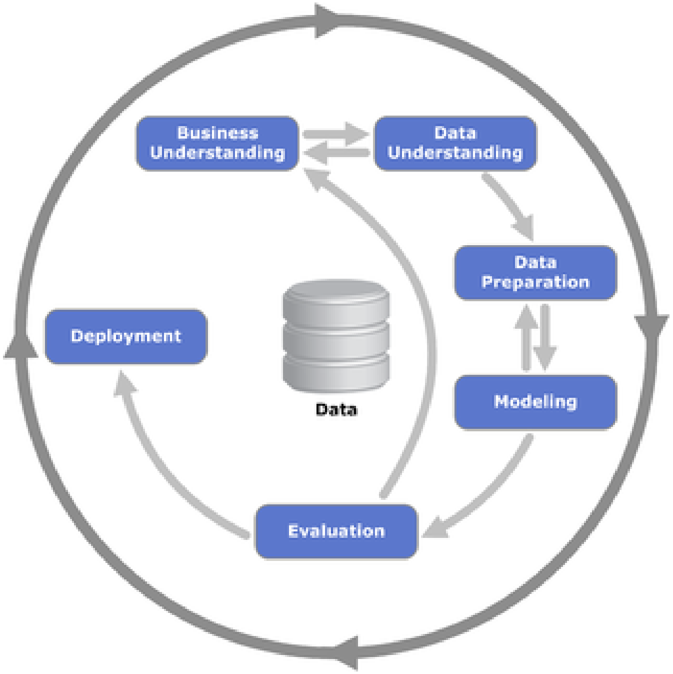


Fig 12. Diagram depicting the CRISP-DM model project life cycle.

Evaluation

In this report, the project plan aligns with the structure of the CRISP-DM model, a well-established data mining process model comprising six stages, as depicted in the diagram above. Throughout this project, the first five stages have been meticulously pursued, with this section representing the fifth stage, Evaluation. Given that this report serves as a project deliverable, it omits formal documentation of the deployment stage, as the report's content effectively embodies the deployment stage.

Through the initial EDA, various factors with potential influence on Cafe Insomnia's hourly sales were identified. These variables were employed in constructing the predictive model, although the model's feasibility is addressed in the Modelling section. The insights gained from the EDA can inform the business decision-making process. For instance, the statistical significance of "hours_after_open" could guide decisions on optimal operating hours for maximising profits. Similarly, the statistical significance of "raining_Yes" might influence decisions on opening during rainy conditions.

Finally, given the violation of crucial assumption principles in our final model, its suitability for business applications is compromised. The predictions may lack reliability and could be substantially biassed or inaccurate.

Business understanding :

In this section, the business objectives were identified, aligning with the project's primary goals as outlined in the assignment specifications. Constraints were stated, and a set of success criteria was defined. Additionally, a basic situation assessment within the project's scope was presented, encompassing a resource list, assumptions (including disclosure of ethical implications), and a glossary of data-related terms, recognising that non-technical individuals may engage with this report.

Data understanding :

Traditionally, this section comprises four key tasks: gathering, describing, exploring, and verifying data quality (Brown, n.d.1a). Since the dataset was provided, there was no need for data gathering. Each variable of interest was described, and their relevance to the objectives was verified. In this case, the aim was to identify variables suitable for constructing a predictive model for hourly sales.

Limitations: Research objectives impose constraints. Use of the provided dataset is mandatory, and additional sources cannot be accessed. This limitation results from both project requirements and the scarcity of data specifically related to "cafes operating on Australian university campuses." Expanding data sources beyond these parameters may introduce subjective bias and significant data distortion.

Data preparation :

As per Brown (n.d.1b), this stage encompasses five key tasks: selecting, cleaning, constructing, integrating, and formatting data.

In this report section, data selection and cleaning were executed through an **SQL INNER JOIN** query. Hourly sales and the 'daily sales hour' variable, pertaining to the construction of data, were described. Data integration involved merging components from three raw SQL dataset tables into a single Pandas dataframe, followed by integrating derived attributes into the same dataframe. Data formatting was crucial, particularly in transforming data frames into an "hourly basis" for the analysis of independent variables' effects on hourly sales. Categorical variables were addressed by converting them into dummy variables, aligning with the analysis objectives of model building.

Limitations & Next Steps:

1. **Handling of dummy variables**

Two issues arose in dealing with dummy variables. Initially, "**hours_after_open**" was treated as an ordinal categorical variable with **six classes**, representing a range from 0 to 5 hours. However, when converted to dummy variables, only **five classes** were generated (hours_after_open_1 to hours_after_open_5), potentially because the class for "hours_after_open = 0" was omitted.

Initially, the assumption was that the choice of the benchmark variable among the five dummy variables would not affect the analysis. However, hypothesis tests revealed the inability to establish the significance of the base variable concerning hourly sales. Typically, the variable with the highest frequency in all categories is used as the benchmark (Upadhyay, 2021). Nevertheless, all classes in "**hours_after_open**" have the same frequency. Additionally, only some dummy variables were statistically significant at the **5% significance level** (see Appendix Table A). Notably, when "**hours_after_open**" was not treated as a dummy variable, it showed significance. This prompts consideration of which approach would have been more suitable. Future studies could conduct further hypothesis tests to determine the preferred option.

2. **raining_Yes and dist_to_cafe**

The correlation analysis revealed a significant association among specific independent variables. This issue of multicollinearity was addressed by selecting one variable while excluding the other from the model.

Contemplation was given to the creation of four new variables by multiplying dummy variables. For example, the first two dummy variables pertained to proximity to the three nearest buildings, and the remaining two related to rainfall conditions. However, a potential concern arises regarding the emergence of new correlations between these freshly created variables and existing ones, which remains unresolved. In a prospective study, an examination of the impact of the aforementioned strategy on model performance is advisable.

3. **Unbalanced samples for raining data**

In the data cleaning phase, a significant portion of records pertaining to 'raining' lacked information. Consequently, there is an apparent imbalance between "Not Raining" and "Raining (Yes)" instances, potentially introducing bias into our findings. A potential remedy involves acquiring or sourcing historical weather data to populate these missing values. This approach is feasible due to our knowledge of Cafe Insomnia's location and the presence of time-series data (date) within the dataset.

Modelling:

The initial phase involved conducting a correlation analysis and subsequent hypothesis testing to select variables for model construction. A linear regression model was chosen as the most suitable for our objectives and target audience, with a focus on verifying the four principal assumptions.

In the main report, it was acknowledged that despite efforts to rectify assumption violations, a model compliant with all assumptions could not be established.

One transformation experiment involved converting certain independent variables into squared forms, such as days_after_open into days_after_open$^{2.}$ However, the residual plot continued to exhibit non-linearity with a discernible pattern. Additional attempts to address heteroscedasticity through cube transformations and log transformations of the dependent variable yielded unsatisfactory results.

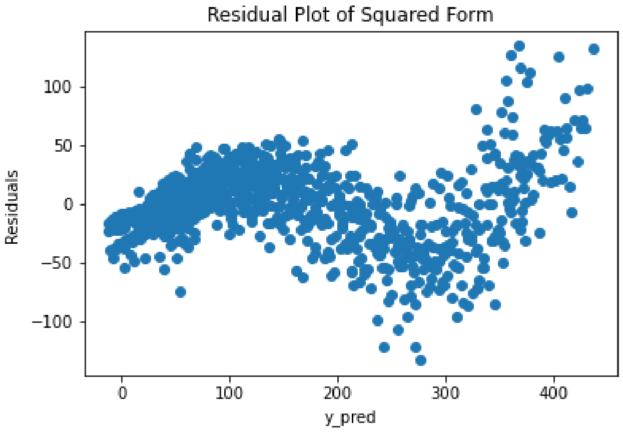


Fig 13. Residual Scatter Plot (Squared) used to verify for Homoscedasticity and the Linearity Assumption

Nonetheless, the model was tested by splitting the data into training and test sets. Initial steps involved parameter setting, model description, and the final model construction. Model evaluation was conducted using Mean Squared Error (MSE) as the metric. The latter task involved extensive model fine-tuning to minimise MSE. However, given the model's violation of most assumptions, efforts were redirected toward addressing these violations. Verification was carried out to determine if it was indeed impossible to create a linear regression model for hourly sales prediction using this dataset.

Despite exhaustive efforts, a model that satisfies all four principal assumptions could not be established. Consequently, based on the analysis, it can be concluded that constructing a **linear regression model** for **hourly sales prediction (in AUD)** using **this dataset is not feasible.**

Limitations & Next Steps:

1. **Procedure used to deal with model assumption violations**

Addressing model assumption violations involved a trial-and-error approach, which, admittedly, was not exhaustive. In future studies, adopting a more systematic procedure to handle these violations is recommended. One possible approach is to adhere to MacEwan University's (n.d.) recommended transformations for addressing model assumption violations (see Appendix for details).

2. **Only made use of linear regression models**

Time and technical constraints constrained us to the development of a single model type. In future studies, an extensive exploration and testing of various regression and advanced models, including Polynomial, SVM, and Random Forest, could be pursued.

Evaluation and deployment :

The final two stages of the CRISP-DM model encompass evaluation and deployment. As mentioned earlier, the deployment stage is where findings are concluded and delivered, aligning with the report's objective. The focus of the evaluation stages is on the model's adaptation to the business and its processes, which is the subject of this section.

The evaluation phase consists of three tasks: evaluating results, reviewing the process, and determining the next steps (Brown, n.d.1d). Initial findings were briefly assessed at the beginning of this section, culminating in the determination that the model is unsuitable for business use. Each project phase was then dissected, elucidating the process and associated limitations. These breakdowns encompass descriptions of potential next steps or solutions applicable to future analyses or studies.

References

Baras. (2021). Pop-Up Stores Become More Than Just A Trend - Retail TouchPoints. Retrieved from https://retailtouchpoints.com/resources/pop-up-stores-become-more-than-just-a-trend

Brown, M.S. (n.d.1a). Phase 2 of the CRISP-DM Process Model: Data Understanding. Retrieved from https://www.dummies.com/programming/big-data/phase-2-of-the-crisp-dm-process-model-data-understanding/

Brown, M.S. (n.d.1b). Phase 3 of the CRISP-DM Process Model: Data Preparation. Retrieved from https://www.dummies.com/programming/big-data/phase-3-of-the-crisp-dm-process-model-data-preparation/

Brown, M.S. (n.d.1c). Phase 4 of the CRISP-DM Process Model: Modelling. Retrieved from https://www.dummies.com/programming/big-data/phase-4-of-the-crisp-dm-process-model-modeling/

Brown, M.S. (n.d.1d). Phase 5 of the CRISP-DM Process Model: Evaluation. Retrieved from https://www.dummies.com/programming/big-data/phase-5-of-the-crisp-dm-process-model-evaluation/

Bujisic, et al. (2019). It's Raining Complaints! How Weather Factors Drive Consumer Comments and Word-of-Mouth. *Journal Of Hospitality & Tourism Research*, *43*(5), 656-681. doi: 10.1177/1096348019835600

Burge, S (2013) . The Motivational Reasons behind Consumer Choice in Branded Coffee Shops. Retrieved from https://warwick.ac.uk/fac/cross_fac/iatl/reinvention/archive/bcur2013specialissue/burge/

Experian. (2021). Data Cleansing - 3 Easy Steps to Maximise Your ROI. Retrieved from https://www.experian.co.uk/blogs/latest-thinking/data-quality/data-cleansing-3-easy-steps-to-maximise-your-roi/

Hayes, A. (2021). Multicollinearity. Retrieved from https://www.investopedia.com/terms/m/multicollinearity.asp

Kotler, P., & Keller, K. L. (2009). *Marketing management*. Upper Saddle River, N.J: Pearson Prentice

Kilroy, K. (n.d.). THE 5 MOST POPULAR COFFEE ORDERS IN AUSTRALIA. Retrieved from https://www.chefworks.com.au/blog/most-popular-coffee-orders

Kurama, V. (2019). REGRESSION IN MACHINE LEARNING: WHAT IT IS AND EXAMPLES OF DIFFERENT MODELS. Retrieved from https://builtin.com/data-science/regression-machine-learning

LAERD Statistics. (2021). Creating dummy variables in SPSS Statistics | Laerd Statistics. Retrieved from https://statistics.laerd.com/spss-tutorials/creating-dummy-variables-in-spss-statistics.php

Upadhyay. (2021). Dummy Variable Trap In Regression Models: Everything in 5 Simple Points. Retrieved from https://www.jigsawacademy.com/blogs/data-science/dummy-variable-trap/#Dummy-Variables-and-the-Dummy-Variable-Trap

MacEwan University (n.d.). *STAT378 Dealing with Model Assumption Violations Lecture/Tutorial handout.* Retrieved from https://academic.macewan.ca/burok/Stat378/notes/remedies.pdf

MARS Foodservices. (n.d.). Catering to Consumer Behaviour Throughout the Week. Retrieved from https://www.marsfoodservices.com/trends/catering-to-consumer-behavior-throughout-the-week

Nong, N.P.N. (2020). A Market Analysis and Marketing Plan on Green Café Business: A Case of Gwangju City, South Korea (Bachelor’s thesis). Retrieved from https://www.theseus.fi/bitstream/handle/10024/333903/Nong%2C%20Thi%20Nghi%20Phuong.pdf?sequence=3&isAllowed=y

retailTouchPoints. (n.d.). Pop-Up Stores Become More Than Just A Trend. Retrieved from https://retailtouchpoints.com/resources/pop-up-stores-become-more-than-just-a-trend

Sobelman, D. (2021). Catering to Consumer Behavior Throughout the Week. Retrieved from https://www.marsfoodservices.com/trends/catering-to-consumer-behavior-throughout-the-week

Stat Trek. (n.d.). Statistics Dictionary: Dummy Variable. Retrieved from https://stattrek.com/statistics/dictionary.aspx?definition=dummy-variable

W3Schools. (n.d.) Retrieved from https://www.w3schools.com/sql/sql_join_inner.asp

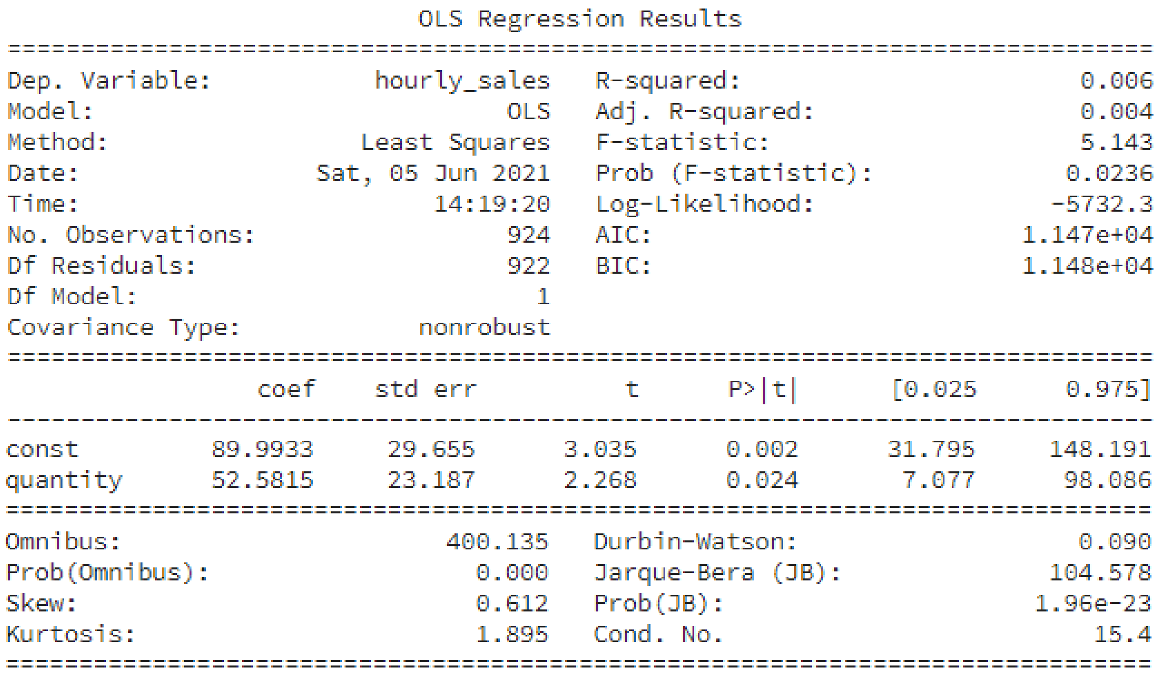
Appendix

Appendix A.

Table A.1. Hypothesis Test Results

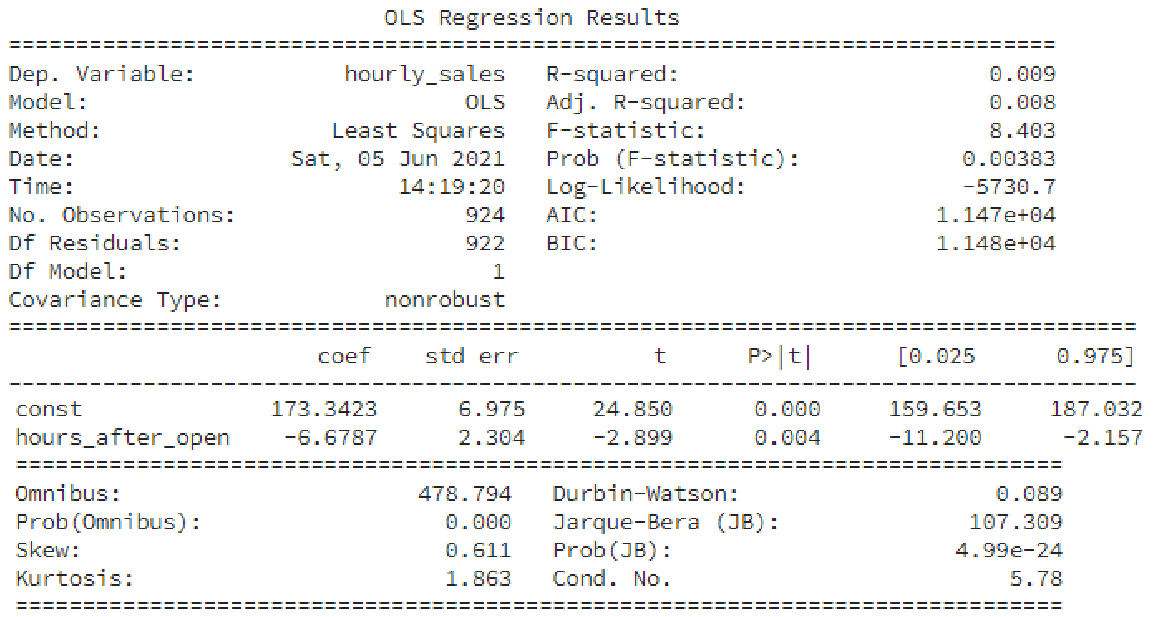
This section contains all the hypothesis test results from our code, excluding **days_after_open** as it already is present in the main body of the report**.**

quantity



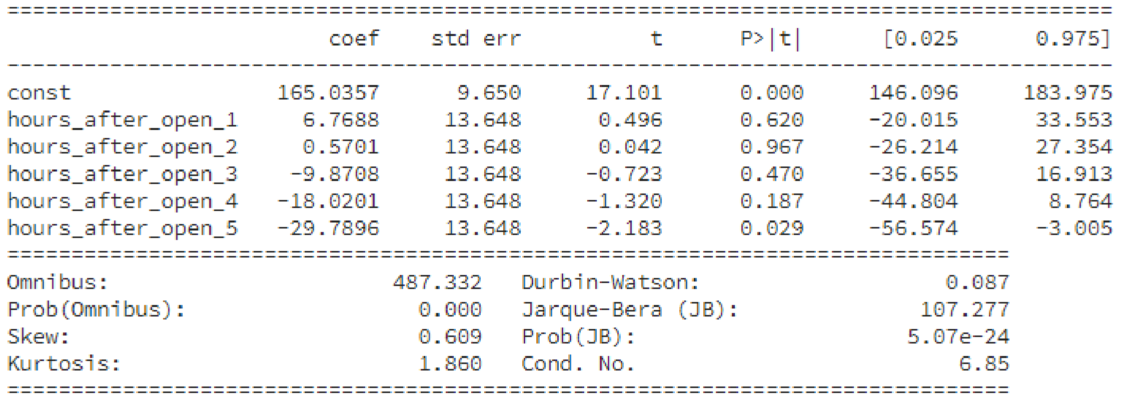
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           hourly_sales   R-squared:                       0.006
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     5.143
Date:                Sat, 05 Jun 2021   Prob (F-statistic):             0.0236
Time:                        14:19:20   Log-Likelihood:                -5732.3
No. Observations:                 924   AIC:                         1.147e+04
Df Residuals:                     922   BIC:                         1.148e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         89.9933     29.655      3.035      0.002      31.795     148.191
quantity      52.5815     23.187      2.268      0.024       7.077      98.086
==============================================================================
Omnibus:                      400.135   Durbin-Watson:                   0.090
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              104.578
Skew:                           0.612   Prob(JB):                     1.96e-23
Kurtosis:                       1.895   Cond. No.                         15.4
==============================================================================
```

hours_after_open



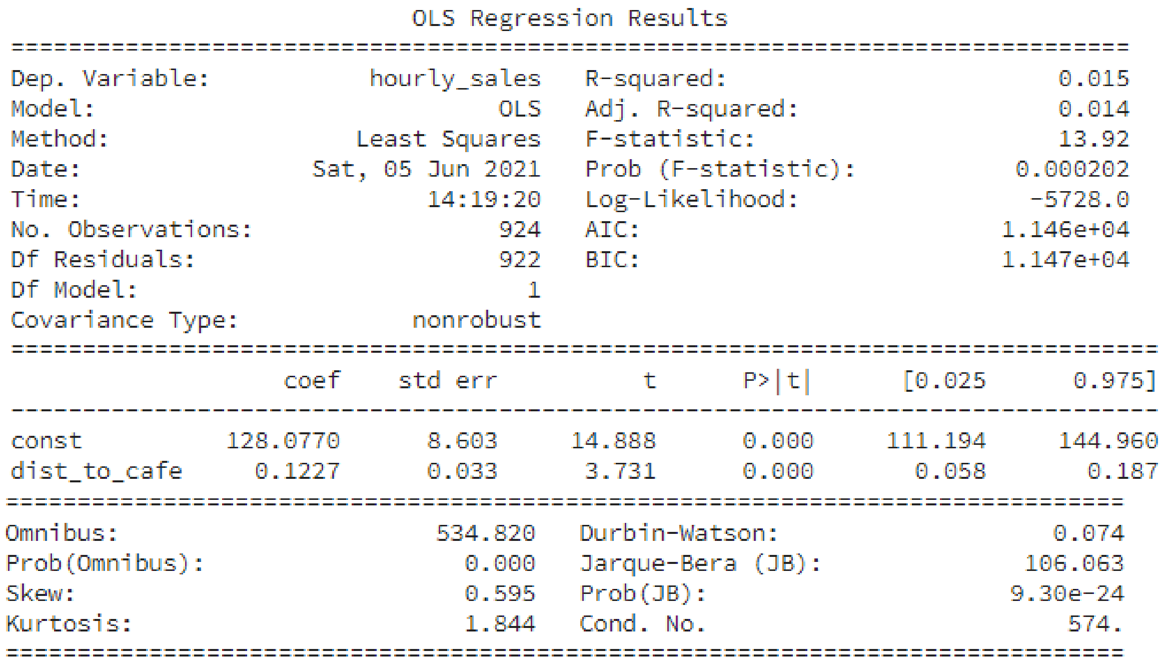
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           hourly_sales   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     8.403
Date:                Sat, 05 Jun 2021   Prob (F-statistic):            0.00383
Time:                        14:19:20   Log-Likelihood:                -5730.7
No. Observations:                 924   AIC:                         1.147e+04
Df Residuals:                     922   BIC:                         1.148e+04
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
 const            173.3423      6.975     24.850      0.000     159.653     187.032
 hours_after_open  -6.6787      2.304     -2.899      0.004     -11.200      -2.157
 ==============================================================================
 Omnibus:                      478.794   Durbin-Watson:                   0.089
 Prob(Omnibus):                  0.000   Jarque-Bera (JB):              107.309
 Skew:                           0.611   Prob(JB):                     4.99e-24
 Kurtosis:                       1.863   Cond. No.                         5.78
 ==============================================================================
```

hours_after_open_1 to _5



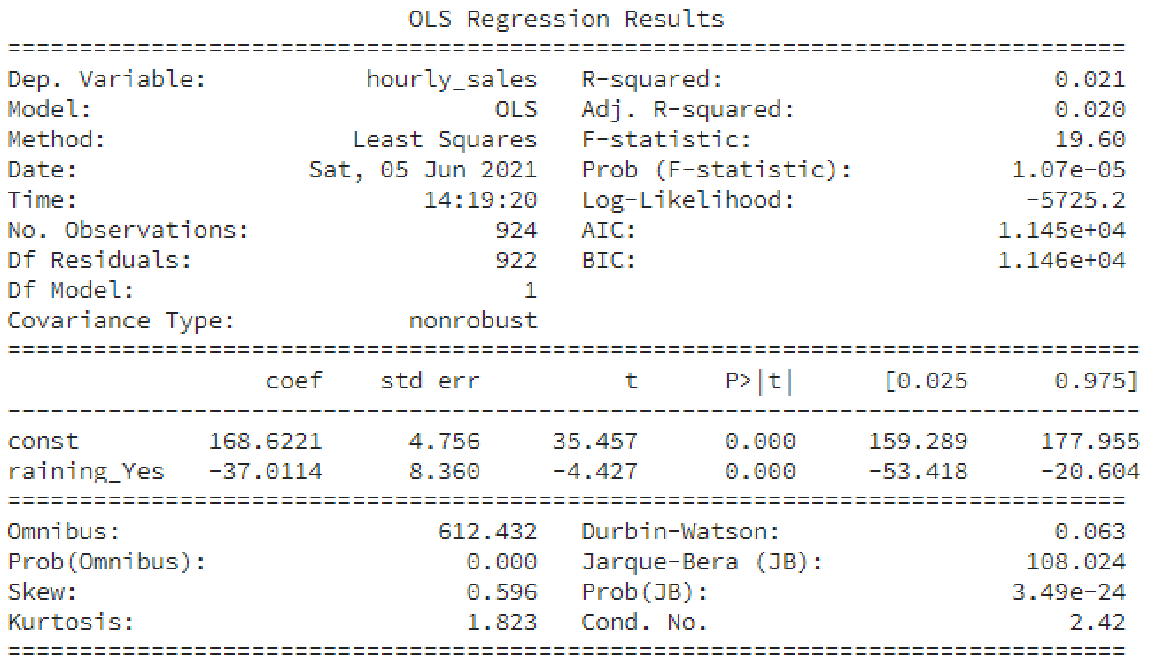
```
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const               165.0357      9.650     17.101      0.000     146.096     183.975
hours_after_open_1    6.7688     13.648      0.496      0.620     -20.015      33.553
hours_after_open_2    0.5701     13.648      0.042      0.967     -26.214      27.354
hours_after_open_3   -9.8708     13.648     -0.723      0.470     -36.655      16.913
hours_after_open_4  -18.0201     13.648     -1.320      0.187     -44.804       8.764
hours_after_open_5  -29.7896     13.648     -2.183      0.029     -56.574      -3.005
==============================================================================
Omnibus:                      487.332   Durbin-Watson:                   0.087
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              107.277
Skew:                           0.609   Prob(JB):                     5.07e-24
Kurtosis:                       1.860   Cond. No.                         6.85
==============================================================================
```

dist_to_cafe



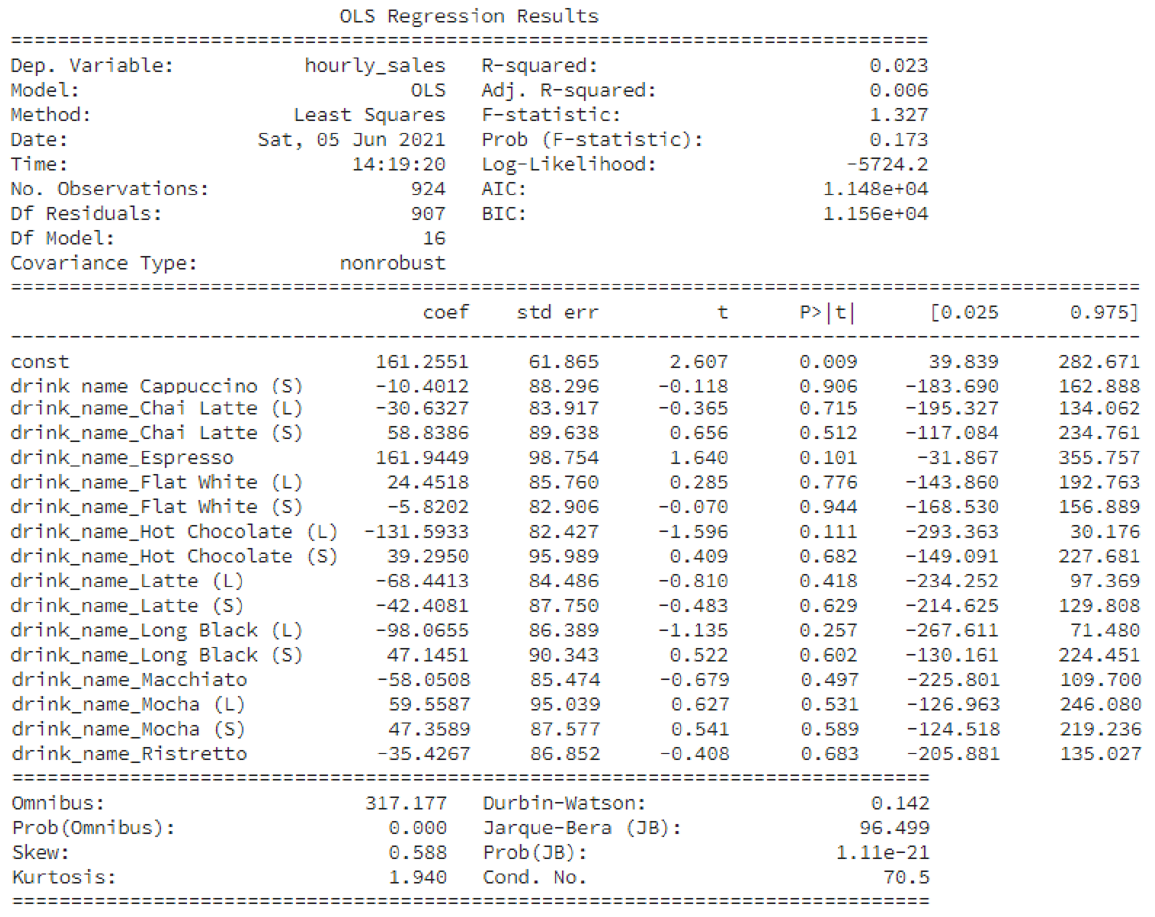
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           hourly_sales   R-squared:                       0.015
Model:                            OLS   Adj. R-squared:                  0.014
Method:                 Least Squares   F-statistic:                     13.92
Date:                Sat, 05 Jun 2021   Prob (F-statistic):           0.000202
Time:                        14:19:20   Log-Likelihood:                -5728.0
No. Observations:                 924   AIC:                         1.146e+04
Df Residuals:                     922   BIC:                         1.147e+04
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          128.0770      8.603     14.888      0.000     111.194     144.960
dist_to_cafe     0.1227      0.033      3.731      0.000       0.058       0.187
==============================================================================
Omnibus:                      534.820   Durbin-Watson:                   0.074
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              106.063
Skew:                           0.595   Prob(JB):                     9.30e-24
Kurtosis:                       1.844   Cond. No.                         574.
==============================================================================
```

raining_Yes



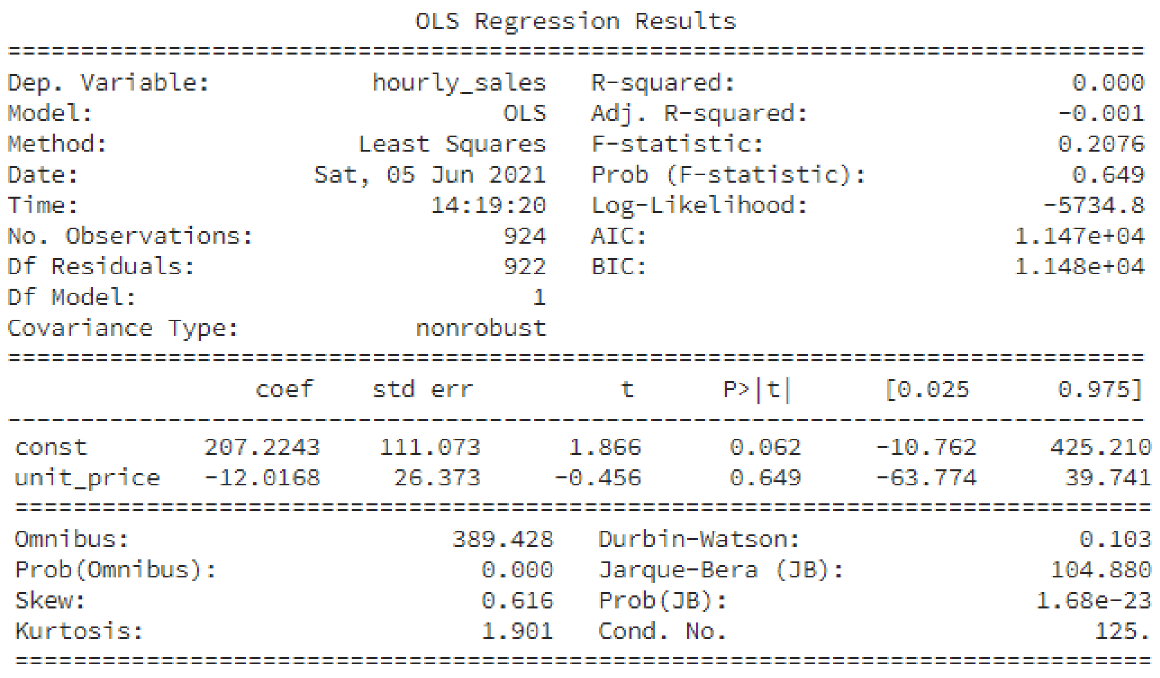
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           hourly_sales   R-squared:                       0.021
Model:                            OLS   Adj. R-squared:                  0.020
Method:                 Least Squares   F-statistic:                     19.60
Date:                Sat, 05 Jun 2021   Prob (F-statistic):           1.07e-05
Time:                        14:19:20   Log-Likelihood:                -5725.2
No. Observations:                 924   AIC:                         1.145e+04
Df Residuals:                     922   BIC:                         1.146e+04
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const         168.6221      4.756     35.457      0.000     159.289     177.955
raining_Yes   -37.0114      8.360     -4.427      0.000     -53.418     -20.604
==============================================================================
Omnibus:                      612.432   Durbin-Watson:                   0.063
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              108.024
Skew:                           0.596   Prob(JB):                     3.49e-24
Kurtosis:                       1.823   Cond. No.                         2.42
==============================================================================
```

name (drinks)



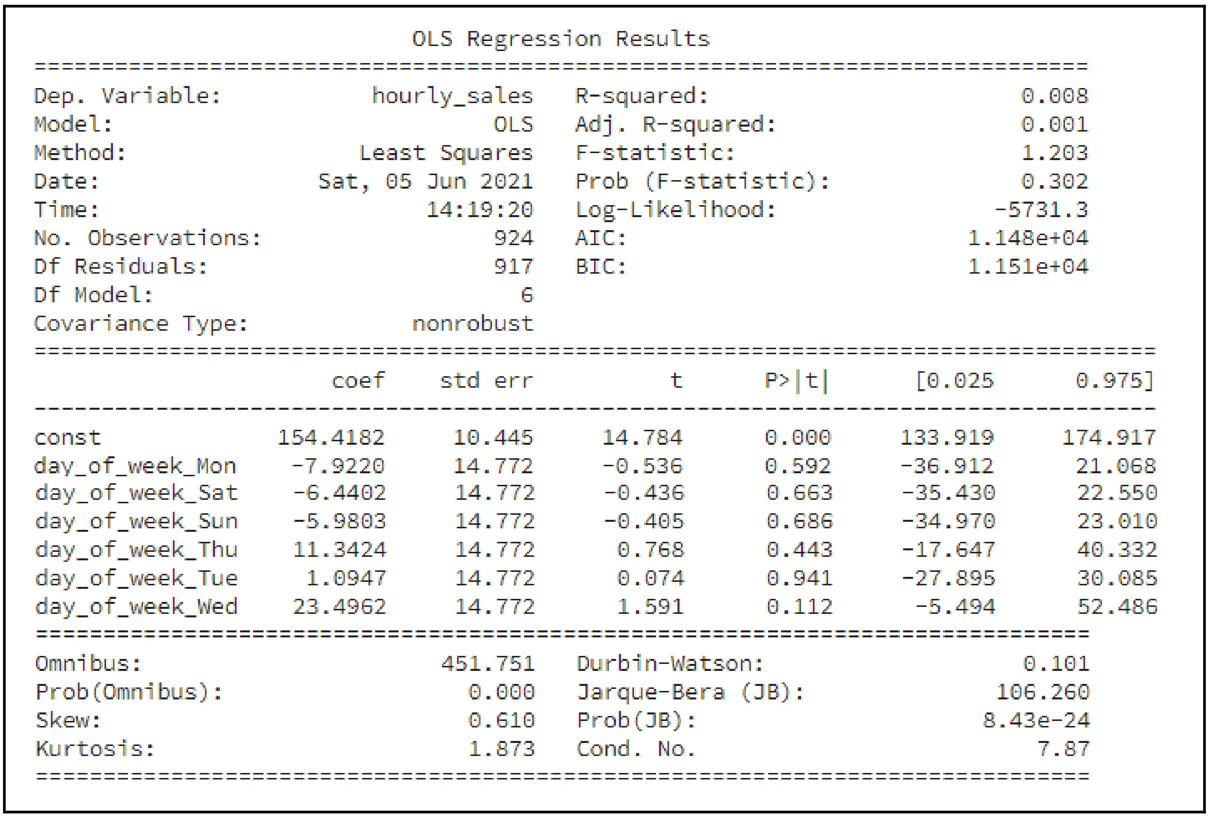
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           hourly_sales   R-squared:                       0.023
Model:                            OLS   Adj. R-squared:                  0.006
Method:                 Least Squares   F-statistic:                     1.327
Date:                Sat, 05 Jun 2021   Prob (F-statistic):              0.173
Time:                        14:19:20   Log-Likelihood:                -5724.2
No. Observations:                 924   AIC:                         1.148e+04
Df Residuals:                     907   BIC:                         1.156e+04
Df Model:                          16
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                            161.2551     61.865      2.607      0.009      39.839     282.671
drink name Cappuccino (S)        -10.4012     88.296     -0.118      0.906    -183.690     162.888
drink_name_Chai Latte (L)        -30.6327     83.917     -0.365      0.715    -195.327     134.062
drink_name_Chai Latte (S)         58.8386     89.638      0.656      0.512    -117.084     234.761
drink_name_Espresso              161.9449     98.754      1.640      0.101     -31.867     355.757
drink_name_Flat White (L)         24.4518     85.760      0.285      0.776    -143.860     192.763
drink_name_Flat White (S)         -5.8202     82.906     -0.070      0.944    -168.530     156.889
drink_name_Hot Chocolate (L)    -131.5933     82.427     -1.596      0.111    -293.363      30.176
drink_name_Hot Chocolate (S)      39.2950     95.989      0.409      0.682    -149.091     227.681
drink_name_Latte (L)             -68.4413     84.486     -0.810      0.418    -234.252      97.369
drink_name_Latte (S)             -42.4081     87.750     -0.483      0.629    -214.625     129.808
drink_name_Long Black (L)        -98.0655     86.389     -1.135      0.257    -267.611      71.480
drink_name_Long Black (S)         47.1451     90.343      0.522      0.602    -130.161     224.451
drink_name_Macchiato             -58.0508     85.474     -0.679      0.497    -225.801     109.700
drink_name_Mocha (L)              59.5587     95.039      0.627      0.531    -126.963     246.080
drink_name_Mocha (S)              47.3589     87.577      0.541      0.589    -124.518     219.236
drink_name_Ristretto             -35.4267     86.852     -0.408      0.683    -205.881     135.027
==============================================================================
Omnibus:                      317.177   Durbin-Watson:                   0.142
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               96.499
Skew:                           0.588   Prob(JB):                     1.11e-21
Kurtosis:                       1.940   Cond. No.                         70.5
==============================================================================
```

unit price



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           hourly_sales   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                    0.2076
Date:                Sat, 05 Jun 2021   Prob (F-statistic):              0.649
Time:                        14:19:20   Log-Likelihood:                -5734.8
No. Observations:                 924   AIC:                         1.147e+04
Df Residuals:                     922   BIC:                         1.148e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        207.2243    111.073      1.866      0.062     -10.762     425.210
unit_price   -12.0168     26.373     -0.456      0.649     -63.774      39.741
==============================================================================
Omnibus:                      389.428   Durbin-Watson:                   0.103
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              104.880
Skew:                           0.616   Prob(JB):                     1.68e-23
Kurtosis:                       1.901   Cond. No.                         125.
==============================================================================
```

day_of_week (Mon to Sun)

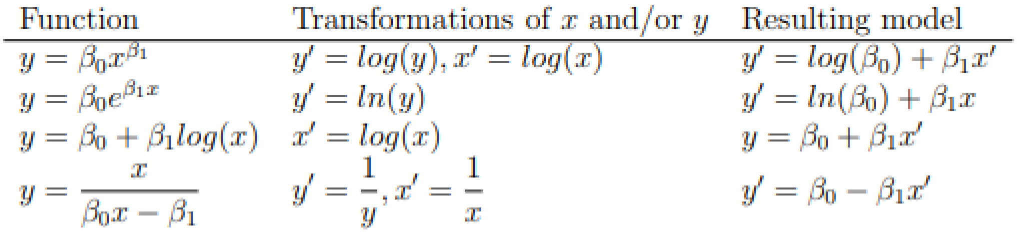


```
                            OLS Regression Results
==============================================================================
Dep. Variable:            hourly_sales   R-squared:                       0.008
Model:                             OLS   Adj. R-squared:                  0.001
Method:                  Least Squares   F-statistic:                     1.203
Date:                 Sat, 05 Jun 2021   Prob (F-statistic):              0.302
Time:                         14:19:20   Log-Likelihood:                -5731.3
No. Observations:                  924   AIC:                         1.148e+04
Df Residuals:                      917   BIC:                         1.151e+04
Df Model:                            6
Covariance Type:             nonrobust
=====================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const             154.4182     10.445     14.784      0.000     133.919     174.917
day_of_week_Mon    -7.9220     14.772     -0.536      0.592     -36.912      21.068
day_of_week_Sat    -6.4402     14.772     -0.436      0.663     -35.430      22.550
day_of_week_Sun    -5.9803     14.772     -0.405      0.686     -34.970      23.010
day_of_week_Thu    11.3424     14.772      0.768      0.443     -17.647      40.332
day_of_week_Tue     1.0947     14.772      0.074      0.941     -27.895      30.085
day_of_week_Wed    23.4962     14.772      1.591      0.112      -5.494      52.486
==============================================================================
Omnibus:                      451.751   Durbin-Watson:                   0.101
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              106.260
Skew:                           0.610   Prob(JB):                     8.43e-24
Kurtosis:                       1.873   Cond. No.                         7.87
==============================================================================
```

Table A.2. Methods to handle model assumption violations.

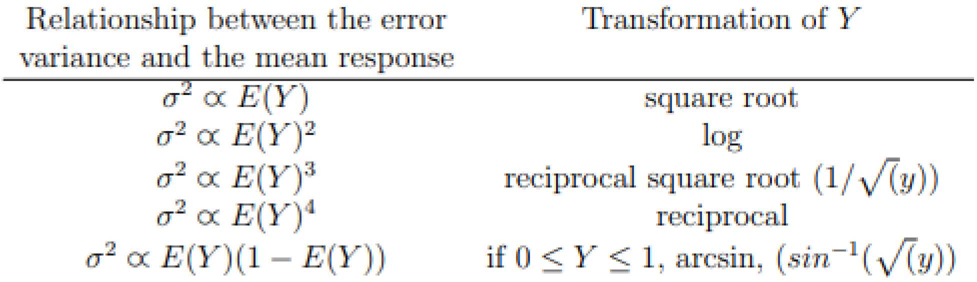
Note: All diagrams and information were obtained from a MacEwan University lecture/tutorial handout.

Dealing with non-Linearity



Function	Transformations of x and/or y	Resulting model
$y = \beta_0 x^{\beta_1}$	$y' = log(y), x' = log(x)$	$y' = log(\beta_0) + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = ln(y)$	$y' = ln(\beta_0) + \beta_1 x$
$y = \beta_0 + \beta_1 log(x)$	$x' = log(x)$	$y = \beta_0 + \beta_1 x'$
$y = \dfrac{x}{\beta_0 x - \beta_1}$	$y' = \dfrac{1}{y}, x' = \dfrac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

Dealing with Heteroscedastic



Relationship between the error variance and the mean response	Transformation of Y
$\sigma^2 \propto E(Y)$	square root
$\sigma^2 \propto E(Y)^2$	log
$\sigma^2 \propto E(Y)^3$	reciprocal square root $(1/\sqrt{(y)})$
$\sigma^2 \propto E(Y)^4$	reciprocal
$\sigma^2 \propto E(Y)(1 - E(Y))$	if $0 \leq Y \leq 1$, arcsin, $(sin^{-1}(\sqrt{(y)})$