STAT-515

FINAL PROJECT REPORT

GROUP 14
Akhil Arekatika
Nithish Bilasunur Manjunatha Reddy
Pritham Mahajan

Dataset Selection:

Customer satisfaction is a key factor that influences the success of any business. It refers to the degree to which customers are satisfied with the quality, service, and value of a product or service. Customer satisfaction can be measured by various methods, such as surveys, feedback, ratings, reviews, and loyalty programs. Starbucks is one of the leading coffee companies in the world, with over 32,000 stores in 83 countries. Starbucks aims to provide a high level of customer satisfaction by offering a variety of products, such as coffee, tea, pastries, sandwiches, and merchandise. Starbucks also strives to create a welcoming and comfortable environment for its customers, by providing free Wi-Fi, music, and seating areas. Starbucks has a loyal customer base, with over 19 million active members in its rewards program. According to a recent survey, Starbucks ranked first among coffee chains in customer satisfaction, with a score of 82 out of 100. Starbucks' customer satisfaction strategy has helped the company to increase its sales, reputation, and market share. We have used Starbucks Customers Survey dataset from github.com which is composed of 20 survey questions of 122 respondents at Starbucks in Malaysia collecting from October 1st to October 5th, 2019. This dataset compiled from a survey asking over 100 respondents about their Starbucks purchasing habits.

The dataset from the Starbucks satisfaction survey contains various aspects of customer experience and demographics. A snapshot how columns in the dataset looks like attached screenshot below.

```
> colnames(starbucks_data)
 [1] "Timestamp"
 [2] "Your.Gender"
 [3] "Your.Age"
 [4] "Are.you.currently....."
 [5] "What.is.your.annual.income."
 [6] "How.often.do.you.visit.Starbucks."
 [7] "How.do.you.usually.enjoy.Starbucks."
 [8] "How.much.time.do.you.normally..spend.during.your.visit."
 [9] "The.nearest.Starbucks.s.outlet.to.you.is...."
[10] "Do.you.have.Starbucks.membership.card."
[11] "What.do.you.most.frequently.purchase.at.Starbucks."
[12] "On.average..how.much.would.you.spend.at.Starbucks.per.visit."
[13] "How.would.you.rate.the.quality.of.Starbucks.compared.to.other.brands..Coffee.Bean..Old.Town.White.Coffee....to.be."
[14] "How.would.you.rate.the.price.range.at.Starbucks."
[15] "How.important.are.sales.and.promotions.in.your.purchase.decision."
[16] "How.would.you.rate.the.ambiance.at.Starbucks...lighting..music..etc...."
[17] "You.rate.the.WiFi.quality.at.Starbucks.as.."
[18] "How.would.you.rate.the.service.at.Starbucks...Promptness..friendliness..etc..."
[19] "How.likely.you.will.choose.Starbucks.for.doing.business.meetings.or.hangout.with.friends."
[20] "How.do.you.come.to.hear.of.promotions.at.Starbucks..Check.all.that.apply."
[21] "Will.you.continue.buying.at.Starbucks."
>
```

We performed extensive data cleaning on the dataset using R and Excel, which included renaming columns to make them more understandable as the original column names were too long to insert in the code and removing duplicates.

```
> colnames(starbucks_data)
 [1] "Gender"            "Age"               "Status"
 [4] "Income"            "Frequency"         "Method"
 [7] "timepervisit"      "nearest"           "membership"
[10] "fequencyofpurchase" "spending"          "comparerate"
[13] "pricerate"         "promotion"         "rateambiance"
[16] "Wifi"              "rateservice"       "situational"
[19] "source"            "loyalty"           NA
>
```

Research Questions:

1. What are the preferences of students compared to employed individuals on choosing Starbucks?

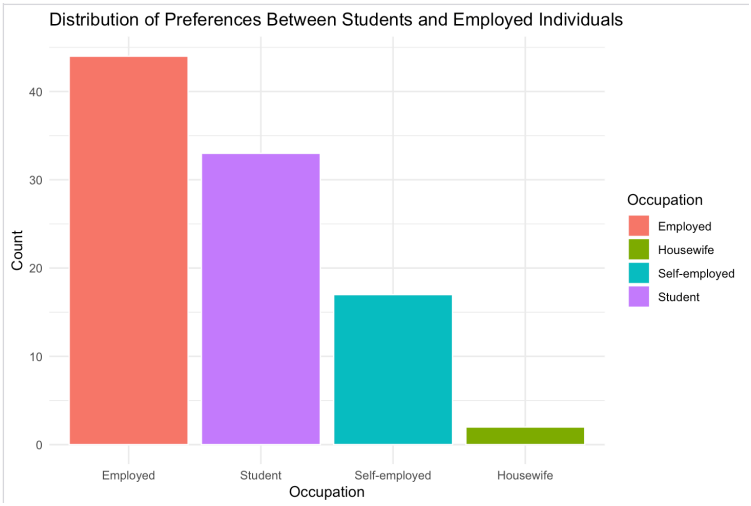
2. What impact does the proximity to the closest Starbucks location have on dining preferences?

3. What factors including price, Wi-Fi quality, ambiance rating, and service rating, help predict whether customers will return to Starbucks?"

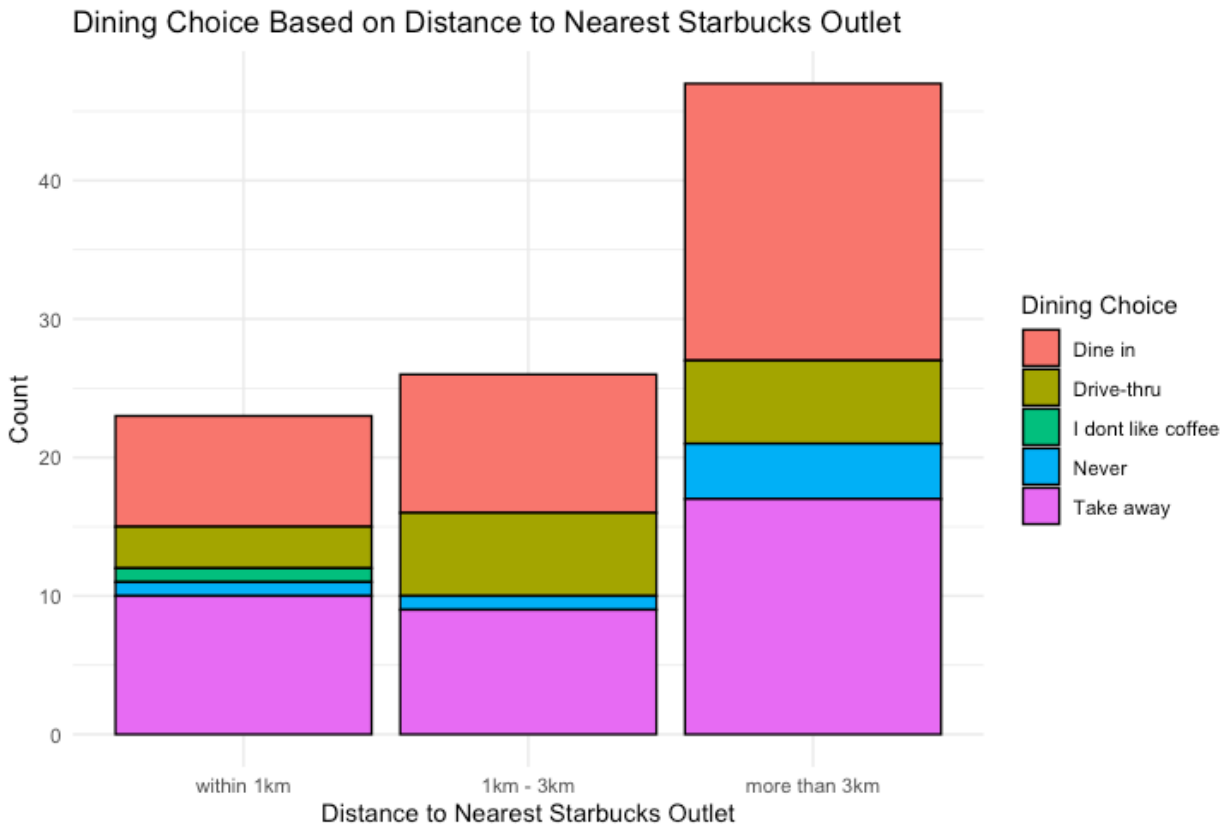
4. How does the quality of service at Starbucks impact the probability of customers choosing it for meetings or hangouts?

Data Exploration:

The dataset contains information about Starbucks customers, including their gender, age, employment status, income, frequency of visits, method of purchase, time spent per visit, proximity to the nearest Starbucks, membership status, frequency of purchase, spending, comparison rate, price rate, promotion rate, ambiance rating, Wi-Fi rating, service rating, situational factors, source of information, and loyalty. The dataset includes 9 rows of data, each representing a different customer. Most customers are female students with an income of less than RM25,000 who visit Starbucks rarely and spend less than RM20 per visit. Most customers dine in or take away, with a few customers using the drive-thru option. Most customers rate the ambiance and service as good, and they use social media and Starbucks' website/apps as their primary sources of information. The dataset provides valuable insights into the behavior and preferences of Starbucks customers, which can be used to improve the customer experience and drive business growth.



We built a bar plot to illustrate the distribution of preferences among students, employed individuals, self-employed, and housewives. The y-axis represents the count, while the x-axis denotes different occupations. Employed individuals show the highest preference, followed by students, self-employed, and housewives in decreasing order.



Based on the distance to the outlet, we represented customers' favored dining locations in a stacked bar plot across various radius categories. Takeaway is slightly more popular than dine-in within a kilometer's radius, which indicates that customers nearby prefer convenience. Dining becomes the most favored option in the 1-3 km radius category, showing that customers are willing to drive a couple of kilometers to enjoy a dine-in meal. Dine-in and takeout preferences appear almost evenly distributed over a 3-kilometer radius, showing that customers living farther away value both options. It's interesting to note that the proportion of drive-throughs stays the same for all distance categories, indicating that people frequently prefer this practical choice regardless of the distance of the outlet.

In conclusion, the stacked bar plot shows a shift in preferences, with people near the outlet choosing takeout, dine-in becoming more common at 1-3 km, and people choosing between dine-in and takeout in a proportionate way as distance increases. Drive-through is still an adequate choice, emphasizing the need for a quick and readily available meal.

Model Analysis:

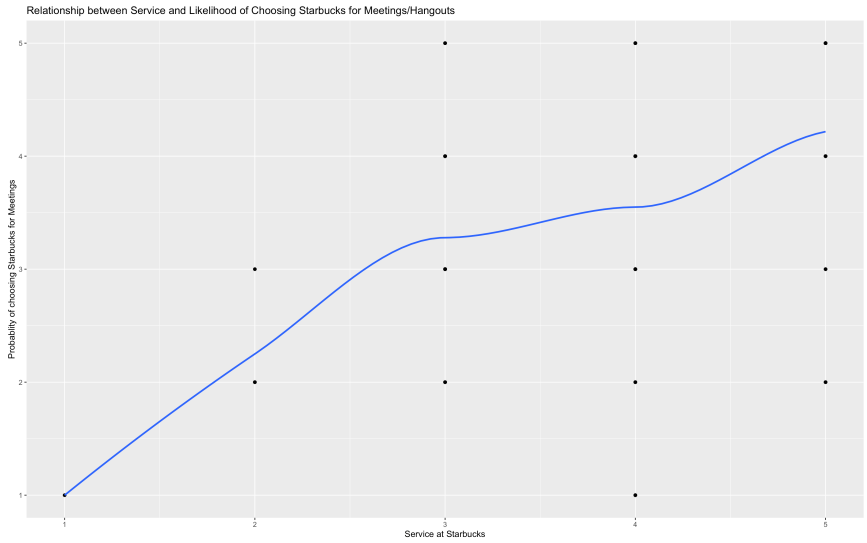
We fitted ordinal logistic regression model to understand the relationship between service ratings at Starbucks (including promptness and friendliness) and the likelihood of choosing Starbucks for business meetings or hanging out with friends. Using logistic regression model Customer's loyalty towards coming to Starbucks based on few predictor variables.

Ordinal Logistic Regression:

We decided to go with ordinal logistic regression model because the dependent variable, "How likely you will choose Starbucks for doing business meetings or hangout with friends," is ordinal, featuring categories with a meaningful order.

```
> summary(ordinal_model)

Re-fitting to get Hessian

Call:
polr(formula = as.ordered(rateservice) ~ situational, data = starbucks_data)

Coefficients:
            Value Std. Error t value
situational 0.8948     0.2023   4.424

Intercepts:
    Value   Std. Error t value
1|2 -1.9749  1.1713    -1.6860
2|3 -0.2515  0.7802    -0.3224
3|4  2.6770  0.7312     3.6612
4|5  4.8441  0.8308     5.8309

Residual Deviance: 269.8349
AIC: 279.8349
```

This model aims to predict the rating of the service provided by Starbucks based on the situational factors. The output shows the coefficients of the model, which indicate the relationship between the dependent variable (rating of service) and the independent variable (situational factors). The coefficient for situational factors is 0.8948, which means that as the situational factors increase, the rating of service also increases. The intercepts represent the threshold values for each rating category. For example, the threshold value for rating categories 1 and 2 is -1.9749, which means that if the predicted value is less than -1.9749, the rating will be 1 or 2. The residual deviance is 269.8349, which measures the difference between the observed and predicted values. The AIC (Akaike Information Criterion) is 279.8349, which is a measure of the quality of the model. A lower AIC value indicates a better model fit. Overall, the ordinal regression model provides valuable insights into the factors that influence the rating of service provided by Starbucks.



The x-axis represents the service at Starbucks and the y-axis represents the predicted probability of choosing Starbucks for meetings/hangouts. The blue line represents the predicted probability of choosing Starbucks for meetings/hangouts based on the level of service at Starbucks. As the level of service at Starbucks increases, the predicted probability of choosing Starbucks for meetings/hangouts also increases.

At last, the ordinal regression model, implemented with the "polr" function, aimed to analyze the relationship between situational factors and the likelihood of choosing Starbucks for meetings or hangouts. The model output exhibited a significant positive effect of situational factors on the ordinal response variable, indicating that as these factors increase, the odds of selecting Starbucks for meetings or hangouts also increase (coefficient = 0.8948, t value = 4.424). The intercepts provided baseline odds for transitioning between adjacent response categories. The model fit, assessed through a Residual Deviance of 269.8349 and an AIC of 279.8349, indicated a reasonable fit to the data. The subsequent plot reinforced the model's findings, illustrating a positive relationship between the level of service at Starbucks and the predicted probability of choosing Starbucks for meetings or hangouts. The visual representation enhanced the interpretation, showing that an improvement in service quality corresponded to an increased likelihood of customer preference for Starbucks in these scenarios. In conclusion, the ordinal regression model and accompanying plot emphasized the significance of situational factors, particularly service quality, in influencing customer choices for Starbucks in the context of meetings or hangouts.

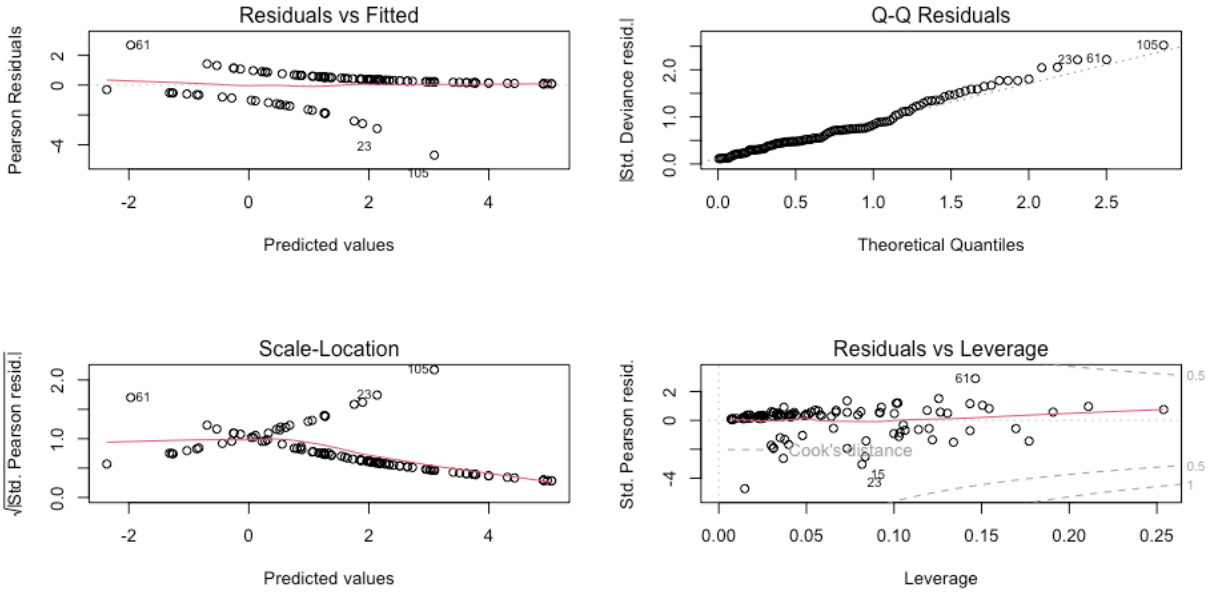
Logistic Regression:

Using data from Starbucks customer surveys, we ran an analysis using logistic regression which is used to analyze the relationship between customer loyalty and various factors that predicted customers experiences on the ambiance of the outlet and quality of the service provided.

```
> # Print coefficients
> print(coef(model))
 (Intercept)  comparerate    pricerate    promotion rateambiance         Wifi
  1.76169481   0.49042475   1.12775328   0.11276695   0.34651019  -0.06498359
 rateservice
 -0.14901572
```

When all predictors are at their normal levels, the intercept of the logistic regression model, that is, 1.76169481, signifies the log-odds of customer loyalty. This intercept gives an outline to analyze the influence of the predictor variables by acting as a starting point to calculate log-odds. The Positive coefficients like 0.3465109 for rateambiance, 0.11276695 for promotion , 1.2775328 for pricerate, and 0.49042475 for comparerate , show that an increase in these variables will result in an increase in the log-odds of the outcome, assuming the rest of the predictors remain fixed. It indicates that an increased likelihood of customer loyalty is correlated with higher perceived rates of comparison on pricerate, rateservice, promotion, and ambiance.

On the other hand, negative coefficients for rateservice (-0.14901572) and wi-fi (-0.06498359) show that the log-odds of the outcome decrease with each unit increase in these variables. Which means the clients who place higher importance on Wi-Fi and service are not as likely to be loyal. The analysis of the results of the logistic regression model is made simpler by these coefficients, that provide insightful data about the direction as well as strength of each predictor's impact on customer loyalty.



Significant variables related to the logistic regression model's fit, and reliability have been highlighted by the diagnostic plots. A substantial variance between predicted and observed values is shown in the Residuals vs. Fitted plot, which indicates potential flaws in enclosing the nuanced nature of the data. For better model performance, a more random scatter around a horizontal line at zero is ideal. The Q-Q plot shows further problems as it indicates the residuals diverge from

normal distribution, which goes against a key principle of logistic regression and casts doubt on the precision of the model's predictions. Moreover, the Residuals vs. Leverage plot shows significant variations for lesser leverage values, indicating that results for observations with a smaller impact may not be precisely predicted. High influence observations are highlighted by Cook's distance plot, which might suggest the presence of outliers or significant points of data that might negatively impact the model's efficacy. In summary, the diagnostic plots together indicate that, to enhance reliability and predictive accuracy, the logistic regression model needs to be improved and more accurately adapted to the data, despite the fact it provides insights into predictor-outcome ties.

Conclusion:

In conclusion our study explored customer preferences and decision-making dynamics at Starbucks, our initial analysis was visually represented by a simple bar plot which identified differences between different occupational groups. The employed individuals showed the highest preference. Students, self-employed, and housewives ranked lower. This study result shows the significant impact of occupation on customer preferences and provides insights for the development of customized customer experiences and specialized promotions. Customer choices was the focus of our further research, that used a stacked bar plot to show the preferences with respect to the locations of Starbucks locations. Customers who were nearby favored takeout, but those who were one to three kilometers away preferred dining in. An equivalent balance between dine-in and takeout options was observed as distances grew. Interestingly, drive-throughs remained popular over all distances, showing an ongoing desire for quick and conveniently available meals, despite the fact they are not in close range of the outlet.

on our further study, we looked at the complex relationship between specific variables and the likelihood that customers will choose Starbucks for hangouts or meetings utilizing an ordinal regression model. The results of the model showed a strong positive impact of specific circumstances on the ordinal response variable, indicating that higher situational factors contributed to higher likelihoods of selecting Starbucks. A well-fitted model supported these subtle insights, as revealed by an AIC of 279.8349 and an acceptable residual deviation of 269.8349. The model's results were further supported by a following visualization that showed a positive correlation between Starbucks' service quality and the likelihood that customers would prefer to meet or hang out. This highlights the important these factors are, especially rate service impacting customer decisions and setting a strong basis for smart business choices. But in the process of our analysis, the logistic regression model's diagnostic plots gave important issues about range and reliability. A significant variation in expected and observed values was shown in the Residuals vs. Fitted plot, showing feasible limits in completely capturing the complexity of the dataset. The Q-Q plot revealed a deviation from the residual distribution's assumed normality, raising doubts about the model's capacity to predict. Meanwhile, variations for lower leverage values were shown by the Residuals vs. Leverage plot, suggesting possible errors in predictions for less significant observations. The Cook's distance plot indicated high influence observations, which could be important information that might harm the model's effectiveness.

References:

Prasertcbs. (n.d.). *basic-dataset/Starbucks satisfactory survey.csv at master · prasertcbs/basic-dataset*. GitHub. https://github.com/prasertcbs/basic-dataset/blob/master/Starbucks%20satisfactory%20survey.csv