**Project Written Report: Amazon Consumer Behavior**

Group 8

**Executive Summary**

In the course of its operations Amazon has collected a vast amount of data about its customers and their preferences. The purpose of this study is to investigate and analyze the vast amount of consumer data that Amazon has collected, and from the analysis, the hope is that Amazon may be able to tailor its marketing and promotional strategy to better their customer satisfaction and increase their sales. Our aim is to derive data driven insights about the customers of Amazon and from the dataset that we have acquired. This dataset is located on Kaggle and is free to download by anyone. The dataset is quite large with over 60 different variables that can be analyzed.

For the purpose of this study, we have conducted a KNN analysis as well as regression tree, but the most important analysis was done using a Multiple linear regression analysis. In this context, we have designated the Purchasing frequency as the variable that we are trying to analyze. Using a linear regression model, we have been able to derive insights into which variables have had a positive and negative impact on the purchasing frequency. Based on the regression model, Cart browsing, cart completion and recommendation provided by amazon make up the most significant factors that had a positive impact on the purchasing frequency. Gender also played a positive role, with women being the deciding factor for that variable.

In conclusion, Amazon has to tailor its approach to increase their customer base as well as their bottom line. By formulating a strategy that takes into account the positive impacts on purchasing frequency while reducing the impact of negative variables, Amazon can improve customer satisfaction and improve the shopping experience on their website.

**Introduction**

Our topic was analyzing trends in Amazon consumer behavior. Amazon, originally founded in 1994, is an e-commerce giant and one of the largest technology companies in the world. Some of Amazon’s core business functions include cloud computing, artificial intelligence, and streaming. However, it is best known for its digital marketplace which is considered a go-to shop for millions of consumers because of its fast and reliable shipping and convenience. Its marketplace sells a wide range of products from books, apparel, electronics, and household items.

With such a vast selection of products there are many different types of Amazon consumers with different behaviors. Understanding consumer behavior is critical in developing marketing strategies and enhancing the customer experience, so we analyzed an Amazon consumer behavior dataset to uncover trends and identify which variables affected a customer’s purchase frequency. Using this dataset, we could explore how several different variables such as age, gender, the customer’s search method, shopping satisfaction, whether the customer left a review or not, and more, affected how frequently a customer made a purchase on Amazon. After analyzing this dataset, our goal was to derive data-driven insights and actionable recommendations to help Amazon optimize their marketing strategies and enhance the customer experience based on our understanding of consumer behavior.

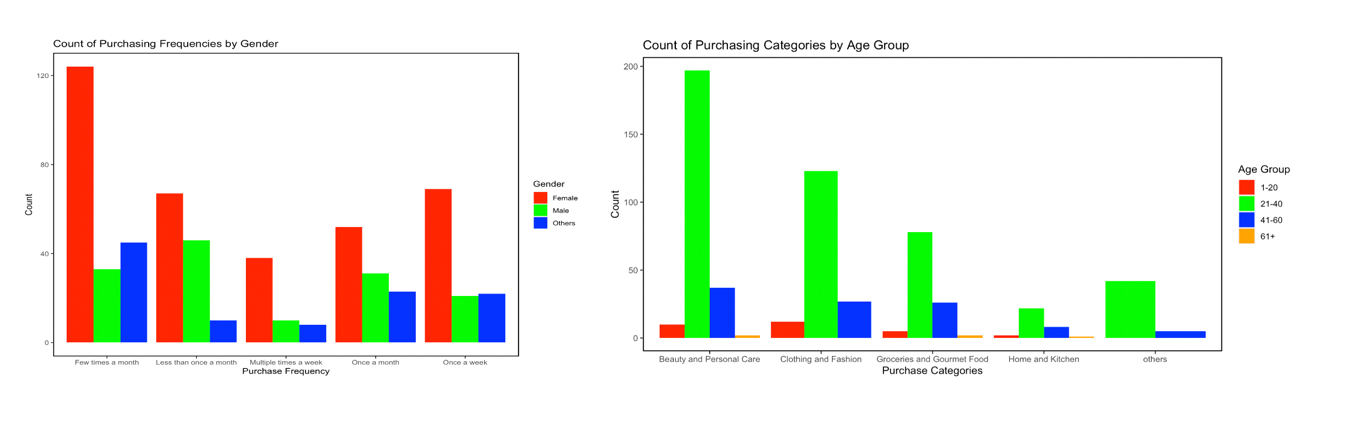
**The Data**

The dataset used for this project was retrieved from Kaggle.com and contains insights on the behaviors of Amazon’s consumers including information on customer interactions, browsing patterns, demographics, interactions, preferences, shopping habits, and decision-making processes. According to the author of the dataset, the data was collected through a Google Forms survey that was distributed and shared through social media platforms. The survey ended up receiving 602 individual responses where each single respondent represents a row in the dataset along with their answers where each question represents a variable. In total, the dataset, prior to performing any cleaning and transformations, contained 602 rows of individual responses and 23 columns of the following variables:

* Timestamp: Time and date at which each response was recorded.
* Age: The age of each respondent.
* Gender: The gender of each respondent.
* Purchase\_Frequency: How frequently each respondent made purchases on Amazon.
* Purchase\_Categories: What product category each respondent typically purchased from on Amazon.
* Personalized\_Recommendation\_Frequency: Whether each respondent has made a purchase based on personalized product recommendations from Amazon or not.
* Browsing\_Frequency: How often each respondent browses Amazon’s website or app.
* Product\_Search\_Method: How each respondent searches for products on Amazon.
* Search\_Result\_Exploration: Whether each respondent tends to explore multiple pages or just the first page of search results.
* Customer\_Reviews\_Importance: Each respondents’ rating on the importance of customer reviews in their decision-making process.
* Add\_to\_Cart\_Browsing: Whether each respondent adds products to their cart while browsing Amazon.
* Cart\_Completion\_Frequency: How often each respondent completes their purchase after adding products to their cart.
* Cart\_Abandonment\_Factors: The factors that influence each respondents’ decision to abandon a purchase in their cart.
* Saveforlater\_Frequency: How often each respondent uses Amazon’s “Save for Later” feature.
* Review\_Left: Whether each respondent has left a product review on Amazon or not.
* Review\_Reliability: How much each respondent relies on product reviews when making a purchase on Amazon.
* Review\_Helpfulness: Whether each respondent has been able to find helpful information from other customers’ reviews or not.
* Personalized\_Recommendation\_Frequency.1: How often each respondent receives personalized product recommendations from Amazon.
* Recommendation\_Helpfulness: Whether each respondent finds the recommendations helpful or not.
* Rating\_Accuracy: How each respondent rates the relevance and accuracy of the recommendations they receive.
* Shopping\_Satisfaction: How each respondent is satisfied with their overall shopping experience on Amazon.
* Service\_Appreciation: The aspects of Amazon’s services that each respondent appreciates the most.
* Improvement\_Areas: The areas that each respondent thinks Amazon can improve in.

An exploratory data analysis was also conducted to get a brief understanding of the data to help identify and understand any patterns within the dataset. As seen in Figure 1, the most popular demographic of Amazon shoppers who responded to the survey are Females who make a few purchases a month, which is also the most popular purchasing frequency in the dataset. Furthermore, most Amazon customers within the dataset are between the ages of 21-40, which also makes up the majority of the dataset’s most popular purchasing category of Beauty and Personal Care products.

**Figure 1**: *Bar graphs of purchasing frequency per gender and purchasing categories per age group.*

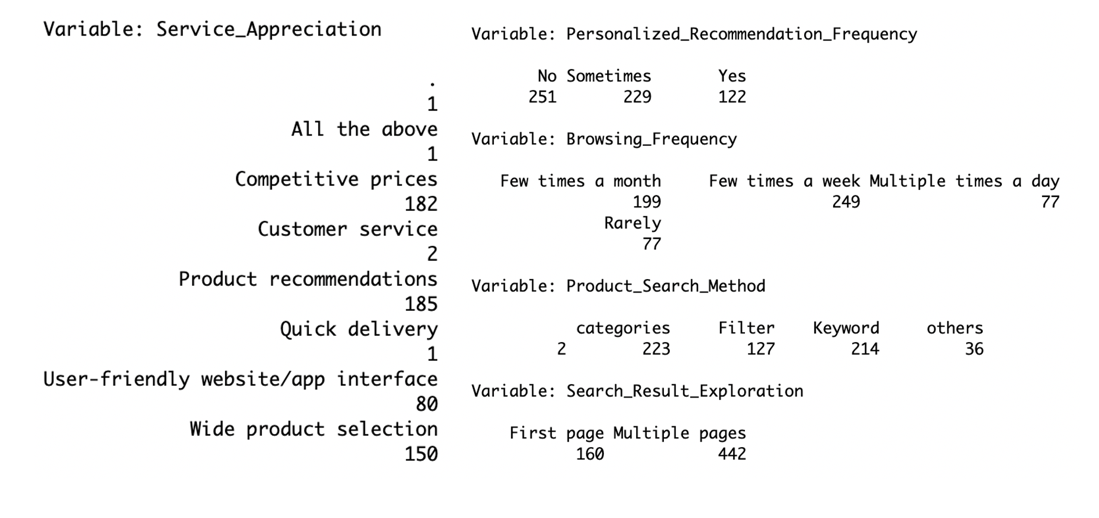


*Note*. The Age Group variable was created by from the Age variable solely for the exploratory data analysis and was subsequently removed prior to performing analyses.

**Cleaning and Formatting**

After downloading the dataset from Kaggle and importing the dataset into R Studio, cleaning and formatting procedures needed to be performed to prepare the dataset for analysis. The first step taken was to remove any variables deemed to be irrelevant which in this case, was only the Timestamp variable. After removing the Timestamp variable with R coding, a for-loop was used to produce a table that contained each variable along with the count of each of their respective values which can briefly be seen in Figure 2.

**Figure 2**: *Table output containing the count of each value per variable.*



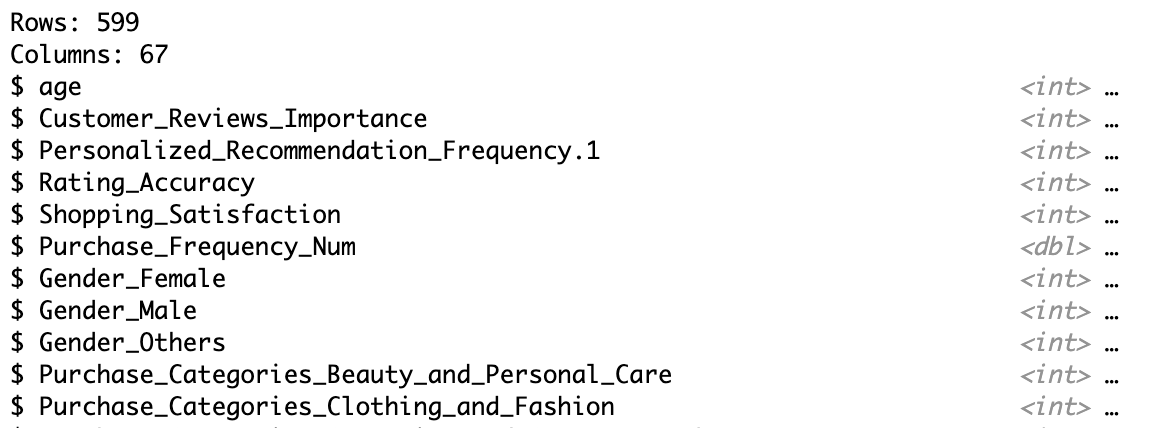
*Note*. This table was used to identify unusual values that needed to be removed, such as blanks, N/A’s, or periods, and to easily see the similar or infrequently occurring values for each categorical variable.

After using R coding to remove the rows that contained the unusual values, the mutate() function was used to group similar values and convert infrequently occuring values to an “others” category for the variables Gender, Purchase\_Categories, Service\_Appreciation, and Improvement\_Areas in order to reduce the total number of dummy variables that would be created. This process, however, experienced several complications as some values, such as “User-friendly website/app interface,” contained spaces or special characters that would produce errors in R. While the mutate() function was used again along with the gsub() function to rename most of the conflicting values, some values produced further errors and therefore had to be manually fixed within the dataset using Microsoft Excel. Once all the values were properly formatted, a new numerical variable called Purchase\_Frequency\_Num was created within R as we believed that there was valuable numerical information behind the values of the Purchase\_Frequency variable which can be seen below:

* Less than once a month = 0.5 purchases a month
* Once a month = 1 purchase a month
* Few times a month = 3 purchases a month
* Once a week = 5 purchases a month
* Multiple times a week = 7 purchases a month

Following the addition of the Purchase\_Frequency\_Num variable, the next step taken was the creation of dummy variables. This was done using the dummy\_cols() function and was applied to the categorical variables of Gender, Purchase\_Categories, Personalized\_Recommendation\_Frequency, Browsing\_Frequency, Product\_Search\_Method, Search\_Result\_Exploration, Add\_to\_Cart\_Browsing, Cart\_Completion\_Frequency, Cart\_Abandonment\_Factors, Saveforlater\_Frequency, Review\_Left, Review\_Reliability, Review\_Helpfulness, Recommendation\_Helpfulness, Service\_Appreciation, and Improvement\_Areas. Once all the dummy variables were created, all the original categorical variables that were used to create them were removed from the dataset along with the old Purchase\_Frequency variable. At this point, the dataset has been fully cleaned and transformed for analysis and a brief glimpse of it can be seen in Figure 3.

**Figure 3**: *Brief overview of the fully cleaned dataset using the glimpse() function.*

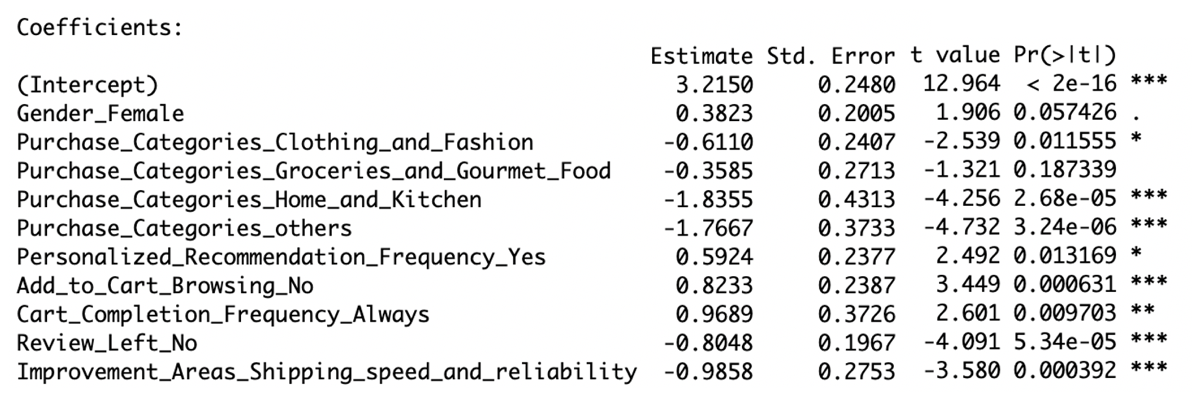


*Note*. The dataset has 599 observations and 67 variables after being fully cleaned and formatted.

**Analysis**

The three analyzation methods we planned to use were linear regression, k-nearest neighbors, and regression trees for our dependent variable of Purchase\_Frequency\_Num. While all three methods were used to make predictions, linear regression was also used to find the most impactful variables on our dependent variable and was the first analyzation method we conducted after performing a 60/40 split for the training and testing data. The first linear regression model constructed included all variables in the dataset minus a baseline for each dummy variable, resulting in a total of 50 predictor variables which yielded an RMSE of 1.99. However, the model needed to be optimized as 40 of the predictor variables had p-values that were above 0.05, indicating that they were not statistically significant and therefore needed to be removed. After removing those variables for our second model, we had an improved RMSE of 1.97 which was better than our first model but there were still two more predictor variables that had p-values above 0.05 and needed to be removed for another model. Despite our third model having no predictor variables that had p-values above 0.05, it yielded an RMSE of 1.98 which was worse than our previous model. Therefore, this meant that our second model was the best overall linear regression model as it was the most accurate for making predictions with the lowest RMSE of 1.97 and can be seen in Figure 4 which also contains the variables that are the most impactful on the dependent variable.

**Figure 4**: *Results of our second linear regression model using the summary() function*



*Note*. Reduced linear regression model with 10 predictor variables that yielded an RMSE of 1.97

The next analyzation method we conducted was k-nearest neighbors to see if it would produce a better RMSE than our best linear regression model. Rather than testing multiple individual models, a for-loop was used to find an optimal value of k that would produce the lowest RMSE which in this case, was 1.95 with a k value of 25. Although the best k-nearest neighbors model already outperformed the best linear regression model in terms of RMSE, we decided to also create a regression tree to ensure that our findings could not get any better. However, after using similar optimization methodology as k-nearest neighbors with a for-loop for mindev instead of k, the best regression tree model had an RMSE of 2.04 with a mindev value of 0.045 which was significantly worse than all the previous models tested previously. Therefore, this meant that the k-nearest neighbors model was the best overall model for making accurate predictions.

**Discussion and Conclusion**

Amazon has a breadth of information and analysis that they can utilize and leverage to increase their purchases. Based on our analysis, Amazon must reorient its marketing and promotional strategy to further the purchasing frequency. The results of our linear regression analysis led to insights on what variables have had an impact on purchases, namely 10 independent variables that had a positive as well as a negative impact on purchases.

Gender is an important variable as it positively impacts the purchasing frequency, Women had a moderate positive impact on purchases. That makes more sense when coupled with some of the segments that have strong showing in purchases namely the beauty and personal segment that we uncovered using exploratory data analysis. However, some variables like Fashion, Groceries, and home appliances had a negative impact on purchases. This specific analysis is important for amazon because it informs them of the categories that are harming their bottom line. By reducing these categories, Amazon can further improve user satisfaction in their products and services.

In conclusion, Amazon can use data driven approaches to increase purchases and sales. KNN has shown to have the most accurate predictable model for amazon, but due to the nature of the study Multiple linear regression is much more important. The MLR model produced positive and negative impacts on purchases and by focusing on the positive impacts, Amazon can positively increase their revenues. Lastly, in the course of the study one thing became paramount, by improving the user experience on their website amazon is able to increase purchasing frequency and thus sales for their company.