Project Outline

By: Group 8 (Loc Le, Briana Patel, Mohammed Seifeldin, Quynh Vo)

The project aims to look at Amazon consumer behavior by investigating trends that are prevalent in the data, how they correlate with each other and how they fit within a cluster as well as in a regression model. All data visualization will be conducted using the R program. The dataset was acquired using [Kaggle](https://www.kaggle.com/datasets/swathiunnikrishnan/amazon-consumer-behaviour-dataset) and it includes data on:

* Age
* Gender
* **Purchase\_Frequency (Dependent Variable)**
* Purchase\_Categories
* Personalized\_Recommendation\_Frequency
* Browsing\_Frequency
* Product\_Search\_Method
* Search\_Result\_Exploration
* Customer\_Reviews\_Importance
* Add\_to\_Cart\_Browsing = Do you add products to your cart while browsing on Amazon?
* Cart\_Completion\_Frequency = How often do you complete the purchase after adding products to your cart?
* Cart\_Abandonment\_Factors = What factors influence your decision to abandon a purchase in your cart?
* Saveforlater\_Frequency = Do you use Amazon's "Save for Later" feature, and if so, how often?
* Review\_Left = Have you ever left a product review on Amazon?
* Review\_Reliability = How much do you rely on product reviews when making a purchase?
* Review\_Helpfulness = Do you find helpful information from other customers' reviews?
* Personalized\_Recommendation\_Frequency = How often do you receive personalized product recommendations from Amazon?
* Recommendation\_Helpfulness = Do you find the recommendations helpful?
* Rating\_Accuracy = How would you rate the relevance and accuracy of the recommendations you receive
* Shopping\_Satisfaction = How satisfied are you with your overall shopping experience on Amazon?
* Service\_Appreciation = What aspects of Amazon's services do you appreciate the most?
* Improvement\_Areas = Are there any areas where you think Amazon can improve?

Using this dataset, we hope to uncover the dependence of the purchase frequency on the other variables using regression analysis. We have come up with some predictions:

* H0(Null Hypothesis) “The independent variables do not have a corresponding impact on the dependent variable “
* H1 (Alternative Hypothesis) “The purchasing frequency is dependent on the independent variables”

Using a regression analysis, we can also look at which of the independent variables had a statistical impact on the purchasing frequency if the null hypothesis is rejected. We can also use the forecasting aspect of the regression model to map out the future purchasing frequency. Because the purchasing frequency variable is a factor, we will likely use logistic regression and may create classification trees to illustrate our results. In addition to possible splitting the dataset into training and testing data, we may also test several different models that each have different variables to see which model is the best at making predictions based off classification rates and/or RMSE/MSE.

In our opinion, we believe that the null hypothesis will be rejected while the alternative hypothesis is accepted. This would allow us to understand the mechanism by which amazon is able to incorporate different aspects and variables to increase the purchasing frequency and thus overall sales and revenue.

We also intend to use k-means clustering to determine the main clusters within the dataset as that would help us better understand the underlying aspects of the dataset. Clustering will be conducted using R with the determinant k being adjusted as we continue with the dataset. We plan to use k-means clustering to group the datapoints to better understand how each of the variables correspond to each other. This will be further explained using non-numerical and qualitative methods to explain the underlying connections.