Final Project Proposal

Project Title: **PriceOP101**

Team Name: **FindMyPrice (212)**

Track: **F2 Price Optimization**

Defining the Problem

The research question for our project is “Which factors are the most influential in determining which pricing group the users get put in?” We are interested in users—those who hear about the product from various sources, who use mobile devices or web to make purchases, those who make purchases on different operative systems, and those who live in different locations in the US—and their different attributes in relation to the pricing model.

Data Collection

There is a dataset that has been provided which will be used for this project. The method for collecting the dataset was done by company ABC where they conducted a study, grouping the users into two groups. To understand how the experiment was conducted, one can observe the results.csv file, which has a column named test. The value for each row in this column can only be 0 or 1. These values help indicate which group the user was in by seeing the value in that column for the specific user row. There are only two groups, one group was offered a cheaper price of 39.0 while another group was offered a higher price of 59.0. The dataset also includes specific time, date, source used such as direct traffic or Facebook advertisements, operative system such as mobile or web, price offered, which group they were placed in, and whether they converted meaning if the user performed a desired action such as purchasing or signing up. Some limitations and biases include the population size or sample size being limited to specifically users in the United States. The sample that was taken also might have bias since we are not sure if the sample conducted for the experiment was randomized. This suggests potential bias in user selection such as specific geographic areas in the United States or specific source used such as Facebook ads, since we cannot say for certain that each user was randomly selected for this experiment. However, we are certain that 66% of the users were offered the cheaper price while 33% of the users were offered a higher price.

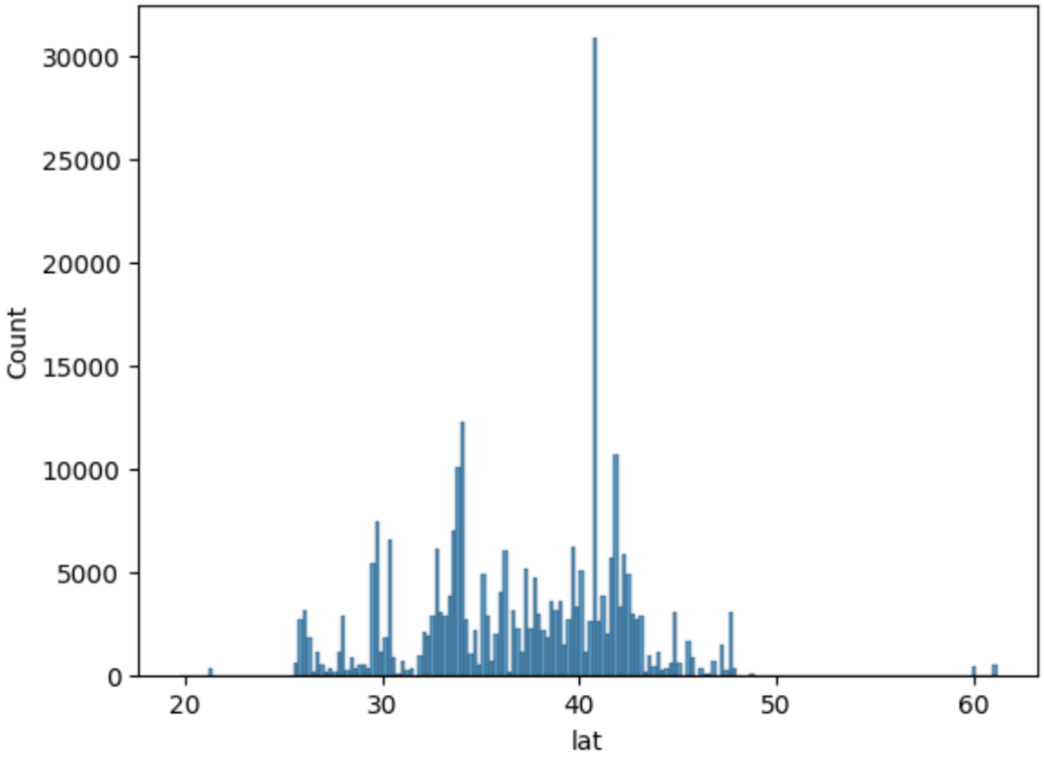
Data Preparation

To gain insight with the geographic location of our users, we must prepare the dataset by merging the user_table.csv with the test_results.csv table. This will provide an additional feature to our table by displaying the geographic location of each user. This is our main goal during data preparation is to essentially create a unified dataset. Since our dataset size is 316,800, we do not need to collect anymore data as we deem the collected data to be satisfactory.

Data Exploration

Based on the data provided in the track, a brief analysis of the dataset could yield some interesting findings for our team to design our study. By using the describe() function and the groupby() function, we were able to quickly identify the median and mean of the data, and helped us to first gain some knowledge about the differences between the two attributes that we planned to investigate on: the devices that the users were using and the geological location of the users. The result shows a 0.5% difference, however without further cleaning we still can’t make any firm conclusion about their relationship with the conversion rate.

In addition, we were able to use a simple histogram to find some interesting aspects of the population of the dataset.



We can observe from the data that a majority of the users who contributed to this dataset live at a latitude around 40. This may be a finding that worth further investigation in our study on the geological pattern of the amazon users.

Model Building

To test the influences of device and location on price sensitivity, we will conduct an A/B test by grouping users according to their device and location and comparing the conversion rate at $39 and $59 price levels. This model should reveal how price sensitivity varies across two groups.

Furthermore, we will build a decision tree model to identify the most significant factors on conversion rate. Using a decision tree combining results from A/B test, we can reveal whether device or location are the most critical factors for price sensitivity, and then giving out a best price optimization solution to the company.

Risks and Timeline

There are some potential risks such as uneven distributions of data, one group may have much more samples than another. Besides, there might be other unrecorded confounding variables that also influence conversion rate, which may bias our price optimization strategy.

Our group will spend the first two weeks on data exploration and cleaning, then writing code to conduct all necessary statistical tests, and to build primary models. Then, we will refine our codes and decision trees, and we will finish the paper and poster along the way.