**Customer’s Choice of Lyft Ride Products Analysis**

Team 6: Wanyue Xiao, Yingxue Gao, Yiyuan Cheng

**Research Question**

When people are using Lyft, there may be some patterns relating to their choice of whether taking an economy or luxury ride. Possible features that might influence their choices can be weather, time, price or others. In order to have a better understanding on **1) which factors have effects on the customer’s choice in Lyft ride product** and **2) how these factors affect the customer’s choice**, the group will use part of the [Uber and Lyft Dataset Boston, MA](https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma) dataset as training dataset to create different models and then use these models to **3) make prediction on the customer’s choice of the type of ride** for the testing dataset (also generated from the original dataset).

**Dataset Description**

The original dataset consists of 57 different types of variables and approximately 0.7 million records regarding the Uber and Lyft rides in Boston, MA. Among these records, 0.3 million are from Lyft while others are from Uber. As the entire dataset is too big and distracting for the proposed analysis, the group decides to extract 20,000 observations about merely Lyft ridership while keeping all initial variables. Those variables mainly describe the information of each ride such as date, cab type, price, distance, and destination, as well as the concurrent weather such as precipitation, wind and temperature. Variable types vary from integer and numeric to categorical and text.

The target variable is *product\_id*, which defines different types of rides a customer can choose. It can be divided to two classes: Economy (Shared Saver, Shared, Lyft, Lyft XL) and Luxury (Lux, Lux Black, and Lux Black XL).

**Experiment Design**

The group will first extract a smaller dataset from the original dataset that required for this research. Random selection will be used to avoid introducing biases. After gaining the dataset, the group will remove the irrelevant and duplicate attributes such as ID and excessive temperature description. Outliers and missing values will be discarded or replaced with meaningful and unbiased values. Effective data visualization will be applied to help understanding the target dataset better.

**Association rule mining:** The group will first conduct supervised association rule mining to find out what kinds of LHS will result in a choice of economy ride and luxury ride, respectively. Since Association Rule can efficiently identify those conditions that are most likely to co-occur, it is possible that two or more factors (such as weather and distance) often show up together to affect the customer’s choice. Thus, these useful patterns can be identified using the association rule mining.

**Classification:** The group will use two different classification methods to analyze the target dataset.

1. Decision tree: This model will develop a structure emulating human’s decision making flow with each node showing the most significant variables and their contribution to the final outcomes. It provides an intuitive way revealing how people come to different choices and can be utilized to predict a possible decision given the pattern.
2. Support vector machines (SVM): The SVM is a supervised machine learning model that uses classification algorithms for two-group classification problems. The group will use SVM to build a model trying to separate economy ride and luxury ride, and then use this model to make predictions on the testing dataset.

**Logistic regression:** Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. The group will create a logistic regression model to separate the data of economy ride and luxury ride as two possible classes. The model will also be used to make predictions on the testing dataset.

The group will conduct model performance evaluation for each model generated above.