Nathan Alvarez Olson
Rohan Chaudhry
Aparna Kakarlapudi
Daniel Diaz

Group 25: Dataset Selection for Darwin project

1. **Describe the dataset. (ex: Information about Traffic violations in Montgomery County, MD; or The number of microaneurysms found in a patient's eye and whether or not they have diabetic retinopathy.)**

   We will be using two datasets. The first dataset, which is the one we will be most concerned with and will serve as the basis of our data analysis, details B-Cycle usage in Austin from the end of 2013 to the end of 2018. The dataset can be found here: https://data.austintexas.gov/Transportation-and-Mobility/Austin-B-Cycle-Trips/tyfh-5r8s/data

   The second, supplementary dataset details weather in Austin from end of 2013 to the middle of 2017. It can be found here: https://www.kaggle.com/grubenm/austin-weather. If possible, we would like to find a dataset detailing the weather in Austin by the hour, rather than by the day.

2. **How many records does the dataset have?**

   The B-Cycle dataset has 1.08 million rows. The weather dataset has 1319 rows.

3. **How many features does the dataset have? List or describe a few of them.**

   The B-Cycle dataset has 12 features. Some of these include start and stopping location, trip duration, and check-out time.

   The weather dataset has 21 features. Some of them include: date (YYYY-MM-DD), high, low, and average temperature, humidity, visibility, wind, and precipitation (in).

4. **What can you try to predict in this dataset? (ex: We can try using the features, including age, race, gender, car make and model, etc, to predict the type of traffic violation; or We can use the number of microaneurysms measured in the patient's eye to predict whether or not they have diabetic retinopathy.)**

   Given a specific set of initial conditions, such as check-out station, check-out time, date / season, or weather (from the other dataset), can we predict the trip duration? Can we predict with some accuracy a potential destination?

   Ultimately, we are most concerned with getting a sense of the "types" of trips people take: i.e. how do the check-in location and the weather determine (or not determine) where the person is going, and how long they'll take?

5. **Is this a labeled dataset, appropriate for a supervised learning classification problem? (In other words, if you are trying to predict whether or not someone has a disease, does your dataset contain whether or not each record has the disease?)**

   With some feature engineering, we can combine both datasets so that the B-Cycle dataset includes weather as a feature. Then, we can try to predict the trip duration and the trip destination. With partitioning for either of these features (such as dividing the possible trip durations into five minutes chunks, or dividing the destinations into different regions), this becomes a *multi-class* problem. Even without binning, the different destinations serve as multiple classes (though there are many of them).

   Even more simply, we can try to predict the number of rides on a day. This would require creating a new table with the weather and date data, and a feature that counts up the total rides on that day.

6. **Provide a link to the dataset, if there is one. If you are getting your data from somewhere other than a link, where are you getting it from?**

   The B-Cycle dataset can be found here: https://data.austintexas.gov/Transportation-and-Mobility/Austin-B-Cycle-Trips/tyfh-5r8s/data

   The weather dataset can be found here: https://www.kaggle.com/grubenm/austin-weather.