Data Collection Report

Team Members: Olivia Walker, Trek Rhodehouse, Milena Belianovich.
Emails: olivia.walker@utah.edu, trek.rhodehouse@gmail.com, milushka.trend@gmail.com.

Based on our data collection process and comments from the proposal, we have decided to change the direction of our project due to limited resources. The new direction we are taking will be reported in more detail in a newly formed proposal, located at the end of our data collection report.

The meteorological data was obtained from kaggle.com. The direct link is the following: https://www.kaggle.com/datasets/muthuj7/weather-dataset . This dataset was chosen due to its size and complexity, which tends to work well with the methods preferred for further analysis.

The picked dataset consists of 12 attributes and 96,454 entries, which makes it a broad spectrum of possibilities for data analysis. The large amount of entries proves dependable for proper data mining and further experiments with the data, which influenced the group's decision on choosing this exact dataset. The further explanation of attributes and entries can be seen in the following section.

The data is currently stored in a pandas dataframe. As previously mentioned above, the data set contains 96,454 rows and 12 columns. Each row in the data frame corresponds to the date and time the weather metrics were taken. Specifically, data was recorded each hour from April 1st 2006 to September 9th 2016. Each column in the data frame is organized by a key word or phrase, which indicates the data being recorded. The columns in the data frame are named as follows: *formatted date (date, time)*, *summary (partly/mostly cloudy)*, *precip type (precipitation type)*, *temperature (C)*, *apparent temperature (C)*, *humidity (0-1), wind speed (km/h)*, *wind bearing (degrees)*, *visibility (km)*, *loud cover, pressure (millibars)* and *daily summary (Mostly cloudy throughout the day/Partly cloudy throughout the day/Partly cloudy until night/Partly cloudy starting in the morning/Foggy in the morning)*. Most of the columns in the data frame contain numeric values (float64) however columns like *summary*, and *precip type* contain categorical data.

We obtained the data by downloading a csv file from kaggle.com. As a result the data was very clean and did not require much pre-processing. After viewing the contents of the data on the kaggle website, we downloaded the csv and converted the contents into a pandas data frame (using python). After further review of the file contents we decided to remove the column labeled *loud cover* as it contained only zeros. As of now, these are the only preprocessing steps that have been performed.

Revised Project Proposal

Group members: Olivia Walker, Trek Rhodehouse, Milena Belianovich.
Emails: olivia.walker@utah.edu, trek.rhodehouse@gmail.com, milushka.trend@gmail.com.

We have decided to change the topic of our project based on some of the feedback we received from the teaching staff. The teaching staff warned us that collecting data may be a struggle as they weren't sure how public the data we needed to collect would be. As we tried to gather enough data for our project, we found this to be true, so we have created a new proposal to go along with our data collection information report.

What data do you plan to use and where do you plan to get it from?
https://www.kaggle.com/datasets/muthuj7/weather-dataset.
The website above contains over 10 years of weather information from the same location. It contains over 90,000 data points and various weather information including cloud cover, precipitation type, humidity, temperature, wind speed and more.

What structure do you want to mine from the data?
We plan to isolate information from the data like temperature, humidity, windspeet, etc. and use this information to make a model that predicts whether or not precipitation will occur. To do this we plan on performing a multivariable linear regression. This process will likely involve splitting the data into training, testing and validation sets. By performing a multivariable linear regression we hope to be able to draw conclusions about what factors correlate with precipitation. Despite our decided method, we also considered comparing the results to other potential data mining techniques, exp. clustering.

Why is this problem interesting?
This problem is interesting because weather prediction is important for everyday living, which includes travel plans, event scheduling, commuting, etc.

What is new, or what I (the instructor) will learn?
At the end of the project, we hope to be able to use our model to predict whether or not precipitation will occur under a set of conditions. Using this model we will be able to showcase weather trends, and improve our understanding of multivariable linear regression along the way.