**Project Milestone 1**

**Identify Datasets**

Nathanael I. Ochoa

Bellevue University

DSC 540: Data Preparation

Professor C. Williams

April 7, 2024

**Project Subject Area**

I live in Southern California and to my surprise we have actually *had* rain these past couple of years during the rainy season, *a lot* of rain actually. I have always enjoyed the rain so I thought it would be fun to pick a city/area and use previous rain data and use that to determine the days it will likely rain this year. For example using past data to predict if it will rain at all tomorrow, in said area of course.

**Data Sources**

* Flat file:
* I came across a dataset on [kaggle.com](http://kaggle.com) that contains records of daily rainfall from January 1, 1948 to December 12, 2017. The dataset contains 25,551 records and 5 columns. The dataset’s website page states that the data was collected at the Seattle-Tacoma International Airport.
* The file I will be downloading and using can be found [here](https://www.kaggle.com/datasets/rtatman/did-it-rain-in-seattle-19482017).
* API:
* [OpenWeatherMap.org](http://openweathermap.org) has a free Current Weather Data API that returns the current weather of the latitude and longitude coordinates entered into an API call. There are a total of 35 possible data variables that are returned with each API call.
* The docs for the Current Weather Data API can be found [here](https://openweathermap.org/current).
* Website:
* The remaining source is the website, [Weather Underground](https://www.wunderground.com/), which is an online weather service. I liked this source because it is possible to select a city and view the hourly breakdown of the weather. There are some visualizations available, as well as a smaller statistical summary. The Daily Observations table provided contains 24 rows (1 for each hour in a day) and 10 columns. It is also possible to view data from previous days.
* The exact Seattle page I will be using from the website can be found [here](https://www.wunderground.com/history/daily/us/wa/seatac/KSEA/date/2024-4-7).

**Relationships**

I can use the file data to view all of the days that it rained in Seattle. I can use the website to view the hourly breakdown on specific days that I select. The API will be used to either prove or disprove any predictions made.

**Plan/Possible Challenges/Ethical Concerns**

I will first have to double check the .csv file and check that there are no errors in the data. I actually noticed a few when I was skimming the file so I will have to go back and correct them. I am currently unsure if I will need all of the records dating back to 1948. I may drop the records at a certain point just to slim the file down a bit. The website data does not seem to go back that far either so that may be how I decide how far back I will go. I chose a random date from the 90’s and the hourly data only had 22 records instead of the usual 24. To no surprise it seems like the farther back I go the more errors/missing values I may find.

I would like to train some of the data and use that to predict future rainy days and as the days pass compare the current results, using the API, to the predicted results. I am still unsure about how to go about doing this but it is my current plan to make my predictions.

The API service also has a free API that returns the coordinates of an entered city name. This means that I can either use this API to return the coordinates for Seattle, WA or find the coordinates of the Seattle-Tacoma International Airport online and use them instead. I plan on comparing the weather difference between the coordinates. If they are similar enough I will use the city coordinates to simplify things.

I am currently unable to think of any ethical implications in my project. I am using previous data to predict the weather and cross checking with the weather API. If any implications come to mind they will be added/addressed in the future.