**Final Report Group Project**

Elias El Metennani

Trongha Phan

**Extract:** We decided to examine how the water and air quality would be effected on a certain date of the year where there were Ecoli outbreaks on beaches in Chicago. We collected .csv files for water and air quality from data world website. We also challenged ourselves by using a different format, .json file for Ecoli measurements throughout a day from a government website.

**Transform:** we cleaned up data by dropping certain columns that we did not need such as measurement ID, which was the timestamp and the transducer that did the measurement. We renamed a certain column such as “Measurement Timestamp label” to be “Measurement Timestamp”, and transform the date and time column to use date.

Next we used groupby function to groupby two columns, “beach name” and “timestamp” then we averaged the “turbidity” and “water temperature” because there were different measurements in a date and thus if we were to merge on timestamp, which not work since the timestamp for each measurement would not match.

Because the file for Ecoli measurement were in .json format, but the way it was formatted, it was “dict within a dict” and thus, using the commands that we learned in the class couldn’t open it. When we were able to open it, the format came out very strange and thus through many different trials we were able to read it correctly. Finally, we transformed .json file to the format that we wanted which were three columns: “Beach Name”, “Timestamp”, and “Ecoli ppm”.

\*The json file was the unexpectedly difficult, and was the roadblock and it took us a while to get through.

Another (smaller)challenge was that the weather station wasn’t at the beach. In another words, the names didn’t mach. Thus, we decided to name the weather station that is closet to a beach name, for example, we called

“oak street station” as “Ohio Street Beach” because “oak street station” is close to “Ohio Street beach”

**Load:** After many transformation steps, we had three beaches that had all the data for water, weather, and Ecoli. Also we dropped any row that did have NA in any columns.