Assignment #1: Charts and plots
COSC 6344 Visualization
Fall 2021
University of Houston

Raunak Sarbajna
1956665

September 6, 2021

Exercise 1

Information Data: Click-through rate (CTR) on browser-based advertisements. An example of such data can be found at this kaggle link.

CTR is quite obviously informational data as it is both discrete and abstract data that represents *behaviour*, rather than a physical phenomenon.

Scientific Data: Multi-spectral satellite imagery (MSI). A freely available example of multi-spectral data can be found in Sentinel imagery, courtesy ESA.

MSI is high-dimensional (4 visible bands, 6 Near-Infrared bands, and 3 Short-Wave Infrared bands) and represents ground-level reflectance of the Earth's surface. Each point is a vector of information representing an actual georeferenced point on the ground. All of this makes MSI a good example of Scientific data.

Exercise 2

Chosen datasets:

1. Titanic Data
2. Houston Weather Pollution Data

Platform used: Python - matplotlib, seaborn, Basemap.

2.1 Titanic

First, we try to plot the passenger data from the Titanic. By a casual exploration, we can see that the data contains several **NaN** values, which we have to ignore. The embarkation point is also labeled with a single letter, which is difficult to understand. We replaces the letters with the name of the actual embarkation port. No other data pre-processing is needed.

Next, we want to see if there is any obvious disproportionate discrepancies in who survived the sinking. We consider that the significant features would be *Sex, Age, Passenger Class*. We do a scatter plot to identify any plots. This can be seen in Figure 1.

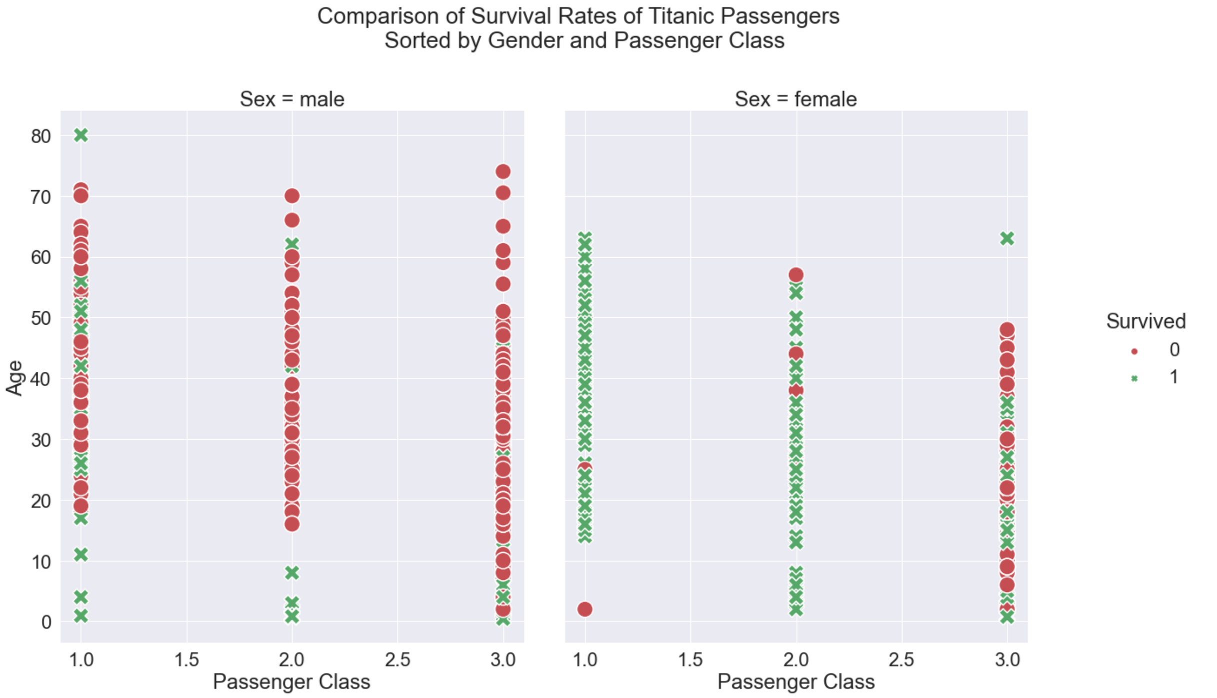

Figure 1: **Plot showing the distribution of passengers on board the Titanic as she sunk. The passengers are marked red if they did not survive and green otherwise.**

This already gives us an indication that class and gender are good predictors of survival. We have not considered *Embarkation Port, Number of Siblings/Spouse, Number of Parents/Children, Fare*. We can assume that the Ticket Number or Cabin Name would not produce any interesting visualization. We split up the visualization of Figure 1 by Embarkation Port. This new visualization can be seen in Figure 2. This also gives us further insight into who died. We finish the plots here, as adding in family members and fare values were starting to make the plot too crowded.

2.2 Pollution

The Houston Weather Station data has 3 files, each track one aspect of pollution: PM 2.5 particles, Ozone, and Carbon Monoxide. The data includes daily Air Quality Index information as well. Most importantly, the dataset contains *geospatial time-series data*, which means dates, latitude and longitude are essential markers. Each file has a limited number of stations collecting data, over the period of January 1, 2018 to March 31, 2018.

There are two approaches to take here:

- We could focus on individual stations and follow how their pollution level changes over the period of 3 months.
- We could look at the geographic distribution of the stations and see which areas are more polluted.

For a complete assessment, we would take both approaches. For the purposes of this assignment I've only taken the latter approach. I have plotted each of the areas along with the values of the corresponding pollutant. The colouration of the points shows the passage of time with the interior being from the oldest date and the outer ring being the newest. The 3 maps can be seen in Figure 3, Figure 4 and Figure 5

Exercise 3

3.1 Titanic

3.1.1 Patterns in the dataset

From the Titanic data visualization, the most obvious point that stands out is the gender imbalance. Obviously, due to the prevailing attitudes of the time, more women and children were allowed to get into the lifeboats, and this is reflected in the data. Also, the plots clearly show that children were more likely to survive, no matter their gender or class.

After breaking out the data along the lines of class and embarkation port, some more interesting patterns emerge. Upper and middle class women, were far more likely to survive compared to lower class women. This same pattern emerges for me, but not quite as stark. After looking at the boarding data, we can see that women boarding from Cherbourg had the highest chance of surviving, while men boarding from Queenstown the lowest. However, almost 70% of all passengers boarded from Southampton, so that skews the plot somewhat.

If we had charted out the family member feature as well, we could have compared if being the member of a family affected the chance of survival, but that was not attempted for this assignment.

3.1.2 Most effective plot

Scatter plots are the most effect dataviz tool in this case, considering the number of discrete features. They make the patterns stand out quickly.

3.1.3 Hidden Information

It is difficult to combine 6 different essential features within a single scatter plot, the cognitive load becomes too much. Even 4 features, as has been done here, makes for difficult reading.

3.2 Pollution

3.2.1 Patterns in the dataset

The Houston Pollution Weather station data has a *lot* of extraneous information. We focused on 3 features, and plotted them using the provide latitude/longitude. The clearest pattern to emerge from this is that the air is better the closer the station is to the Gulf. Conversely, the closer to the chemical factories near La Porte/Deer Park the station, the worse the air quality is. However, Carbon Monoxide and Ozone concentrations are high downtown as well, meaning they are byproducts of vehicular emissions.

3.2.2 Most effective plot

Scatter plots were not as effective as I had hoped. A line chart would perhaps have been a better way of showing the change in pollution over time.

3.2.3 Hidden Information

This dataset is ideal for an animated visualization. Keeping just the linechart for the time-series data meant we would lose the geospatial information and vice-versa. Combining both line plots and a map would not make a good visualization. Animated points, where the colour of the points change as the intensity of the pollution varies over time would be the best method to convey the information in this dataset.

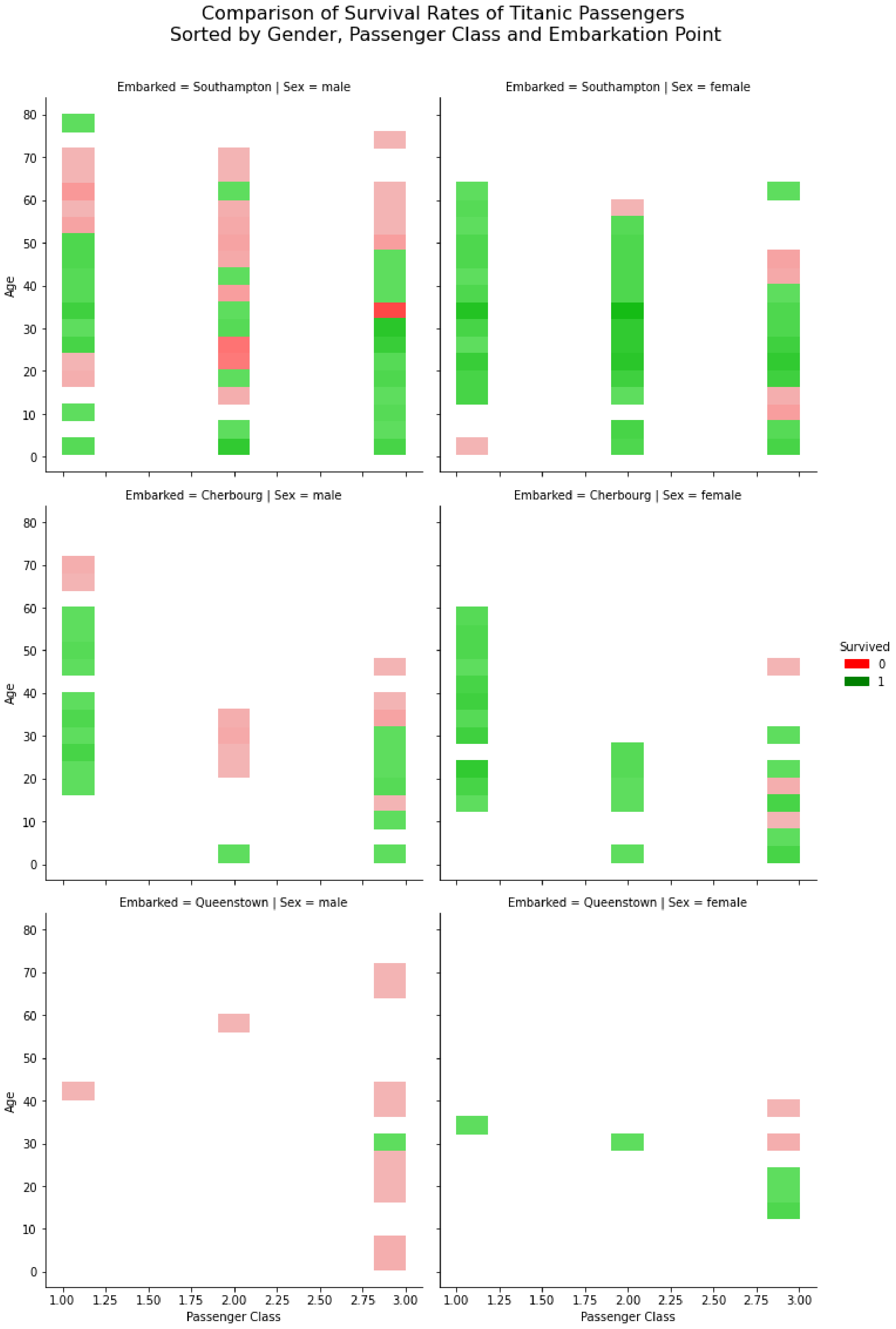

Figure 2: **Plot showing the distribution of passengers on board the Titanic as she sunk, split up by Gender, Passenger Class and Embarkation port.**
The red shows passengers who did not survive.

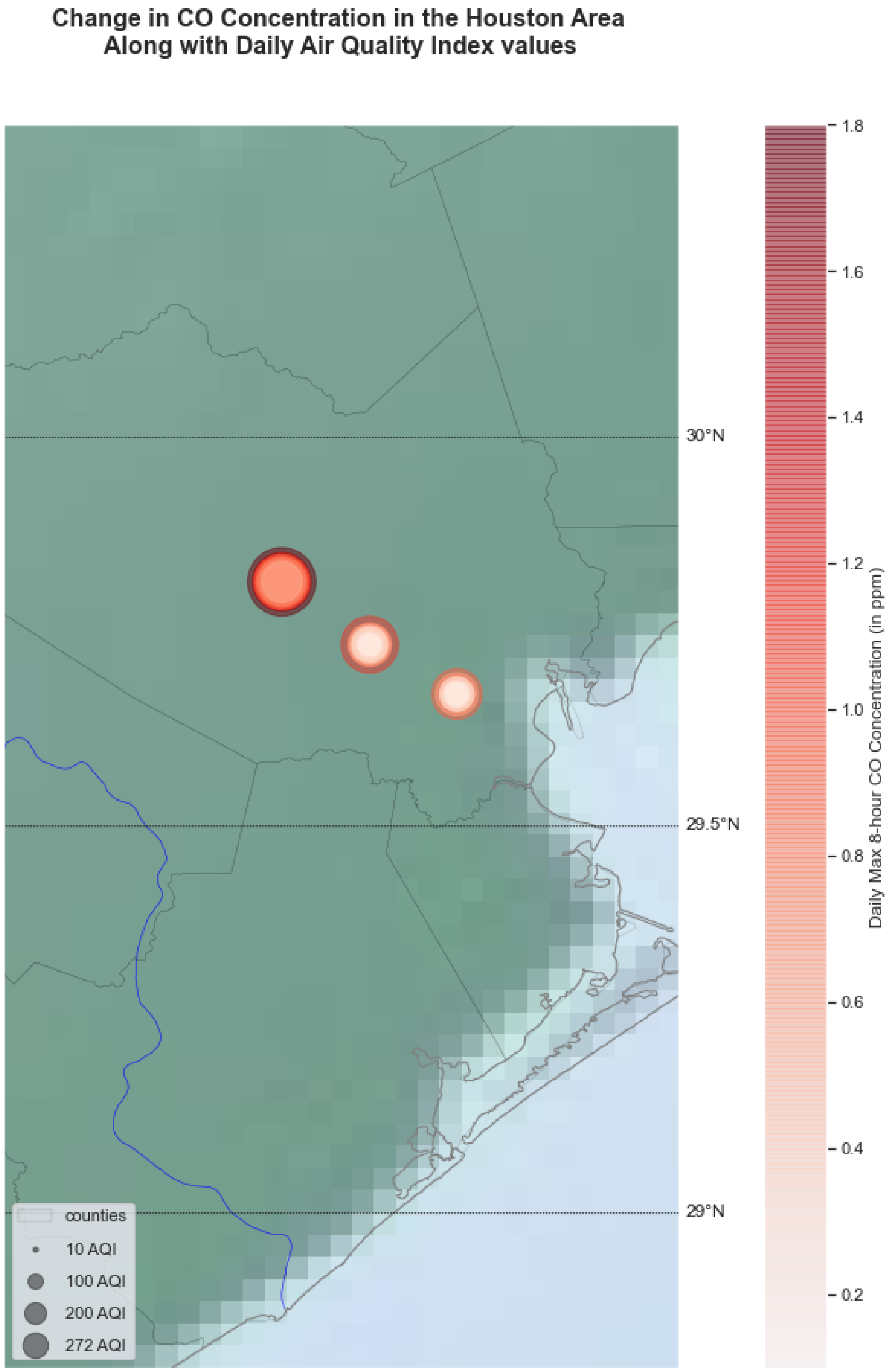
Change in CO Concentration in the Houston Area
Along with Daily Air Quality Index values
30°N
29.5°N
29°N
counties
10 AQI
100 AQI
200 AQI
272 AQI
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Daily Max 8-hour CO Concentration (in ppm)

Figure 3: **Map of the Greater Houston area, showing the concentration of Carbon Monoxide**
Each point shows the change in Daily CO level, with the oldest being in the interior and newer on the exterior.

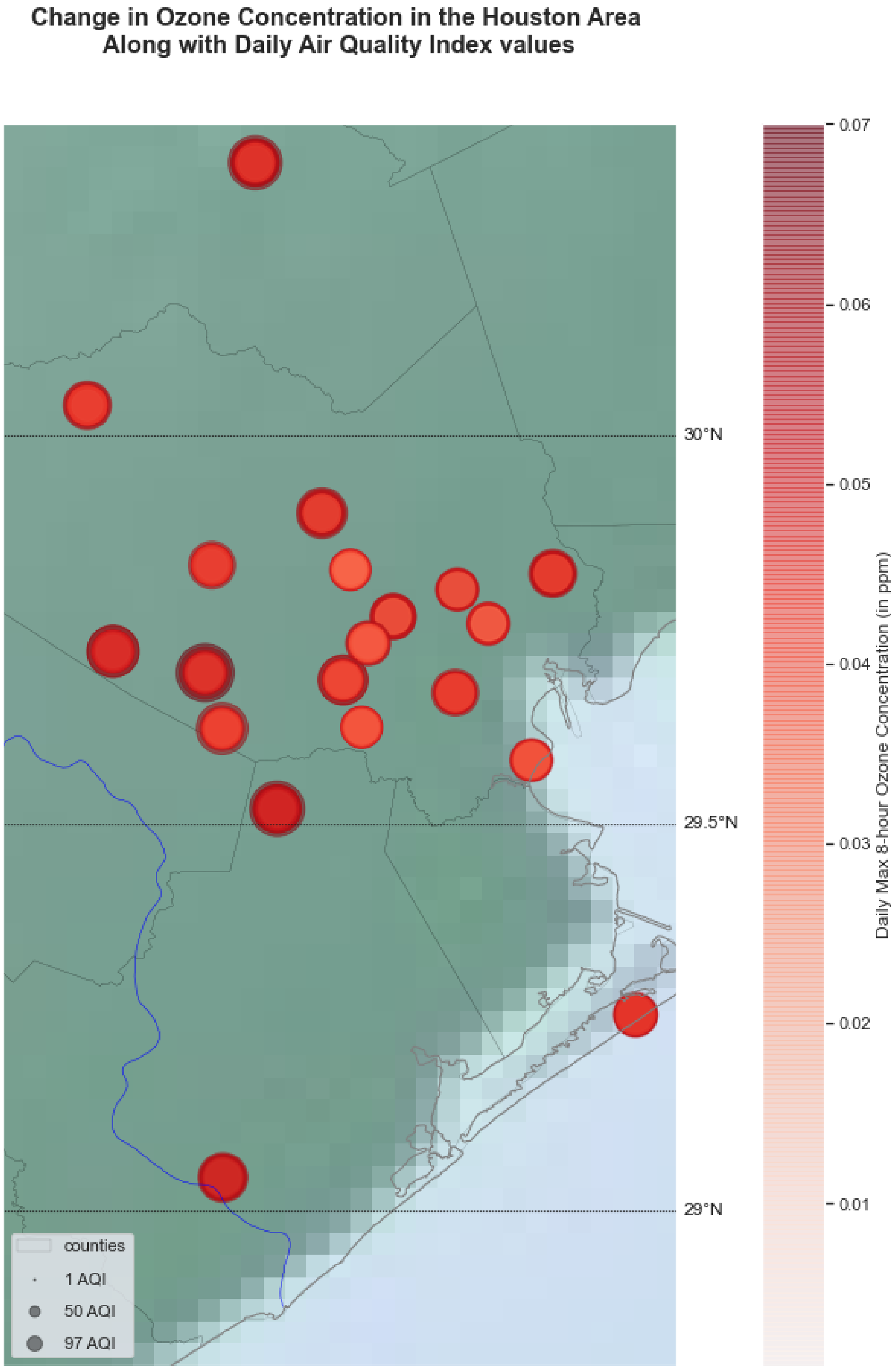
Change in Ozone Concentration in the Houston Area
Along with Daily Air Quality Index values
30°N
29.5°N
29°N
counties
1 AQI
50 AQI
97 AQI
0.07
0.06
0.05
0.04
0.03
0.02
0.01
Daily Max 8-hour Ozone Concentration (in ppm)

Figure 4: **Map of the Greater Houston area, showing the concentration of Ozone Each point shows the change in Daily O_3 level, with the oldest being in the interior and newer on the exterior.**

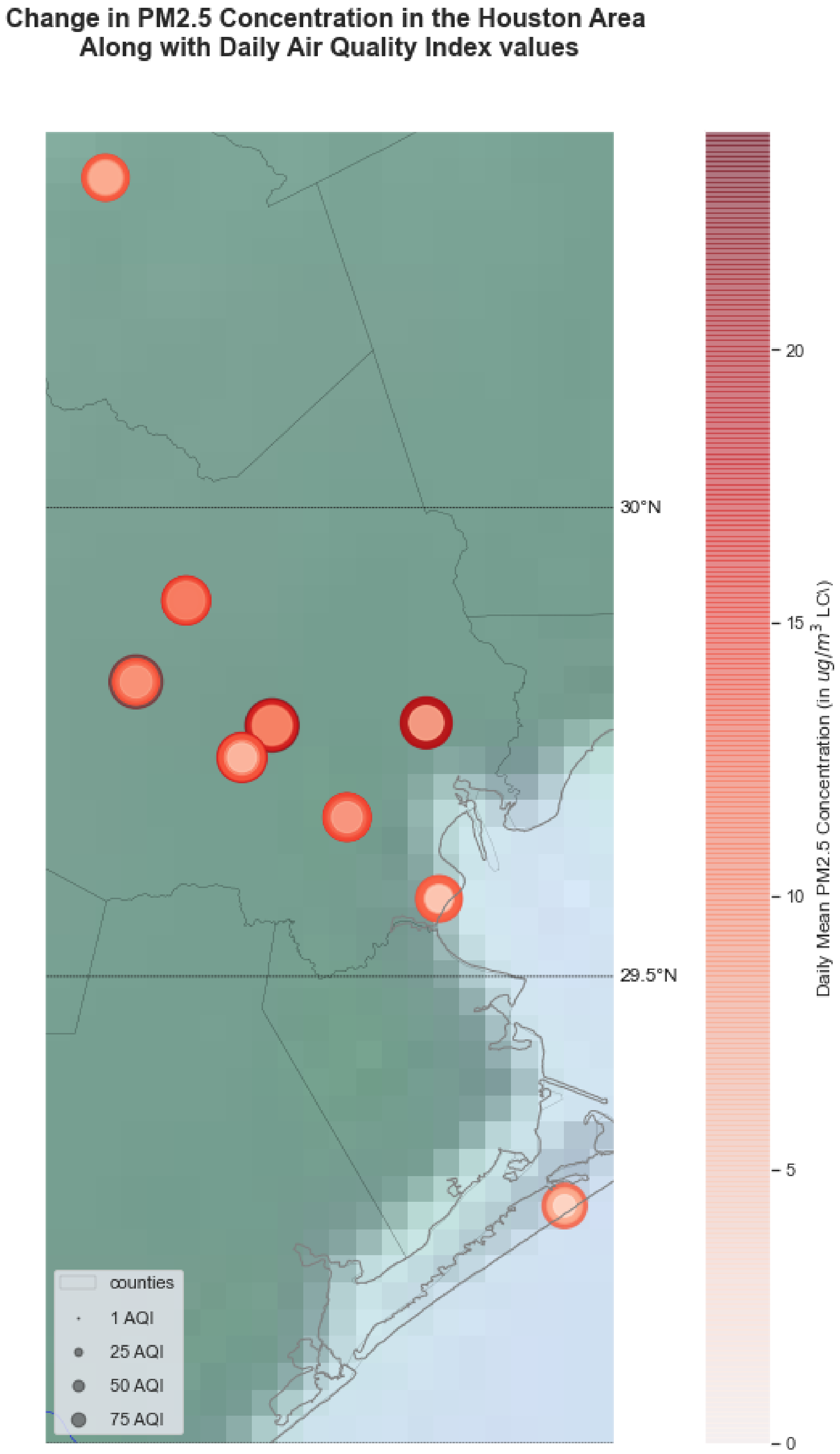
Change in PM2.5 Concentration in the Houston Area
Along with Daily Air Quality Index values
30°N
29.5°N
counties
1 AQI
25 AQI
50 AQI
75 AQI
20
15
10
5
0
Daily Mean PM2.5 Concentration (in ug/m³ LC)

Figure 5: **Map of the Greater Houston area, showing the concentration of *PM*2.5 Each point shows the change in Daily *PM*2.5 level, with the oldest being in the interior and newer on the exterior.**