Air Quality Index Analysis

GROUP 3

Soumya Patil
Bhavna Patadia
Kate Engard
Quecis Joshua
Michael Nosa

“You can have data without information, but you cannot have information without data.”

- *DANIEL KEYS MORAN*

Problem Statement

The study is about the Air Quality Index in the United States. The data that was used is dated 2019 as it has the most recent and complete monitoring of Air Indexes. This study is to show the quality of air that we breath and the chemicals that go with it.

People need to breathe, and so other living creatures in this world. The flow of life wherein a process called respiration. Respiration is when a living creature breathes in oxygen and breathes out carbon dioxide for living things that need it.

It also aims to answer if too much carbon dioxide plus other chemicals may be harmful as it reaches the atmosphere. Will this change the distribution of the heat into Earth and how does it affect global climate change.

Problem Statement

- Data is not collected for all 365 days by each city so an accurate picture cannot be attained with certainty. This has to be taken into consideration when looking at results.
- Looked at Cities with best data coverage for evaluation (atleast 70% or higher data collected)
- Looked at the Cities that has the best good days in respect to AQI
- The original imported dataset for AQI does not have Latitude and Longitude for each city. In order to map it, Latitude & Longitude is pulled from google.
- The dataset used includes information about the data captured for monitoring ozone and PM2.5 parameter on daily basis across counties by EPA's Air Quality System for the year 2019. The data in these datasets is delimited with a comma ','

Problem Statement

- What is the air quality index in the USA?
- Which states has the best air quality index?
- Which city in Arizona has the most hazardous days?
- Which city in Arizona has the highest ground-level ozone?
- What percentage of harmful chemicals in the air is considered harmful for sensitive groups?
- Which state has the most air pollution?
- Is there a correlation between population and air quality index?
- Is there a significant correlation between Latitude and Ozone or PM2.5?

Milestones

Team 3 finalized the subject during the 2nd meeting and started looking for resources for the data.

During the 2nd meeting of the team, the data was addressed.

The team collaborated and assigned each and everyone to manipulate the data.

The team assigned a representative to create a repository and instructed everyone to push their work in it.

The team collaborated on finalizing the codes and addressing the hypothesis.

The team assigned Joshua as the representative for team 3 to present the finished product.

Kate, Josh and Bhavna collaborated on the codes that we used in most parts of the project.

Michael volunteered to put the PowerPoint presentation together.

Bhavna proactively assisted and collaborated to finalize the PowerPoint presentation.

Everyone in the team came up with a hypothesis based on their focuses.

Quecis Joshua combined all the codes from the team's notebook into one notebook.

Data Analysis

Glossary

Annual Air Quality Index

Columns/Indicators	Description
State Code	The FIPS code of the state in which the monitor resides.
County Code	The FIPS code of the county in which the monitor resides.
Site Num	A unique number within the county identifying the site.
Parameter Code	The AQS code corresponding to the parameter measured by the monitor.
POC	This is the “Parameter Occurrence Code” used to distinguish different instruments that measure the same parameter at the same site.
Latitude	The monitoring site’s angular distance north of the equator measured in decimal degrees.
Longitude	The monitoring site’s angular distance east of the prime meridian measured in decimal degrees.
Location	Latitude and Longitude of the monitoring site
Datum	The Datum associated with the Latitude and Longitude measures.
Parameter Name	The name or description assigned in AQS to the parameter measured by the monitor. Parameters may be pollutants or non-pollutants.
Sample Duration	The length of time that air passes through the monitoring device before it is analyzed (measured). So, it represents an averaging period in the atmosphere (for example, a 24-hour sample duration draws ambient air over a collection filter for 24 straight hours). For continuous monitors, it can represent an averaging time of many samples (for example, a 1-hour value may be the average of four one-minute samples collected during each quarter of the hour).
Pollutant Standard	A description of the ambient air quality standard rules used to aggregate statistics.
Date Local	The calendar date for the summary. All daily summaries are for the local standard day (midnight to midnight) at the monitor.
Units of Measure	The unit of measure for the parameter. QAD always returns data in the standard units for the parameter. Submitters are allowed to report data in any unit and EPA converts to a standard unit so that we may use the data in calculations.
Event Type	Indicates whether data measured during exceptional events are included in the summary. A wildfire is an example of an exceptional event; it is something that affects air quality, but the local agency has no control over. No Events means no events occurred. Events Included means events occurred and the data from them is included in the summary. Events Excluded means that events occurred but data form them is excluded from the summary. Concurred Events Excluded means that events occurred but only EPA concurred exclusions are removed from the summary. If an event occurred for the parameter in question, the data will have multiple records for each monitor.
Observation Count	The number of observations (samples) taken during the day.
Observation Percent	The percent representing the number of observations taken with respect to the number scheduled to be taken during the day. This is only calculated for monitors where measurements are required (e.g., only certain parameters).
Arithmetic Mean	The average (arithmetic mean) value for the day.
1st Max Value	The highest value for the day.
1st Max Hour	The hour (on a 24-hour clock) when the highest value for the day (the previous field) was taken.
AQI	The Air Quality Index for the day for the pollutant, if applicable.
Method Code	An internal system code indicating the method (processes, equipment, and protocols) used in gathering and measuring the sample. The method name is in the next column.
Method Name	A short description of the processes, equipment, and protocols used in gathering and measuring the sample.
Local Site Name	The name of the site (if any) given by the State, local, or tribal air pollution control agency that operates it.
Address	The approximate street address of the monitoring site.
State Name	The name of the state where the monitoring site is located.
County Name	The name of the county where the monitoring site is located.
City Name	The name of the city where the monitoring site is located. This represents the legal incorporated boundaries of cities and not urban areas.
CBSA Name	The name of the core bases statistical area (metropolitan area) where the monitoring site is located.
Date of Last Change	The date the last time any numeric values in this record were updated in the AQS data system.

Data Analysis

CREATING DATA BY STATE

- Get the Average data by State
- Focused on Top 5 States with high pollutants

CREATING DATA BY CITY

- Collected data percentage
- Cities with best data coverage for evaluation (at least 70% or higher data collected in 365 days)
- Top 5 cities with best air quality

Pulling Latitudes and Longitudes

- Loop through all the cities from the imported dataset
- Handle spaces in city names
- Gathered city data
- Acquired state from city name
- Acquired lat & Lng from json response
- Collected medianAQI for found cities
- Collected good days for found cities
- Collected hazardous days for found cities
- Collected percentile AQI
- Printed city data as its aquired

Data Analysis

CREATING DATA BY CITY

aqi_city["Year Coverage"] = aqi_city["Days with AQI"] / 365

aqi_city["Good Day Percent"] = aqi_city["Good Days"]/aqi_city["Days with AQI"]

	CBSA	CBSA Code	Year	Days with AQI	Good Days	Moderate Days	Unhealthy for Sensitive Groups Days	Unhealthy Days	Very Unhealthy Days	Hazardous Days	...	90th Percentile AQI	Median AQI	D
282	Malone, NY	31660	2019	272	272	0	0	0	0	0	...	0	0	
27	Augusta-Waterville, ME	12300	2019	270	266	4	0	0	0	0	...	45	33	
254	Lake Havasu City-Kingman, AZ	29420	2019	273	268	5	0	0	0	0	...	29	15	
477	Utica-Rome, NY	46540	2019	267	260	7	0	0	0	0	...	40	27	
31	Bangor, ME	12620	2019	304	294	10	0	0	0	0	...	45	33	

```
# Top 5 States with Good Days
aqi_state_summary.sort_values(by='Good Days', ascending=False).head()
```

State	Good Days	Moderate Days	Unhealthy Days	Very Unhealthy Days	Hazardous Days	Days CO	Days NO2	Days Ozone	Days SO2
Maine	257.600000	13.300000	0.1	0.0	0.0	0.300000	5.000000	209.400000	0.000000
North Dakota	250.000000	22.500000	0.1	0.0	0.0	0.100000	2.000000	221.000000	5.700000
New York	237.000000	22.903226	0.0	0.0	0.0	0.258065	3.709677	179.419355	27.258065
Massachusetts	215.538462	25.153846	0.0	0.0	0.0	0.153846	4.923077	192.923077	0.000000
Mississippi	211.300000	42.300000	0.0	0.0	0.0	0.100000	0.100000	144.300000	0.000000

Codes by Bhavna Patadia and Quecis Joshua

Data Analysis

Pulling Latitudes and Longitudes

```
#loop through all the cities from the imported dataset
for i in range (len(aqi_city['CBSA'])):
    try:
        #handle spaces in city names
        response = requests.get(map_query_url + aqi_city['CBSA'][i].replace(" ","+")).json()

        #gather city data
        available_cities.append(aqi_city['CBSA'][i])

        #get state from city name
        state.append(aqi_city['CBSA'][i].rsplit(", ")[1])

        #get lat & Lng from json response
        lat = round(response['results'][0]['geometry']['location']['lat'], 2)
        lng = round(response['results'][0]['geometry']['location']['lng'], 2)
        lats.append(lat)
        lngs.append(lng)

        #collect medianAQI for found cities
        medianAQI.append(aqi_city['Median AQI'][i])

        #collect good days for found cities
        good_days.append(aqi_city['Good Days'][i])

        #collect hazardous days for found cities
        hazardous_days.append(aqi_city['Hazardous Days'][i])

        #collect percentile AQI
        percentile.append(aqi_city['90th Percentile AQI'][i])

        #print city data as its aquired
        print(aqi_city['CBSA'][i] + ", Lat:" + str(lat) + ", Lng:" + str(lng))

    except Exception:
        #print city name that was not found
        print(aqi_city['CBSA'][i] + " not found!")
```

```
----------Pulling City Lat and Lng data------------
Aberdeen, SD, Median AQI = 24 Lat:45.46, Lng:-98.49
Aberdeen, WA, Median AQI = 18 Lat:46.98, Lng:-123.82
Adjuntas, PR, Median AQI = 18 Lat:18.16, Lng:-66.72
Adrian, MI, Median AQI = 40 Lat:41.9, Lng:-84.04
Akron, OH, Median AQI = 44 Lat:41.08, Lng:-81.52
Albany, GA, Median AQI = 44 Lat:31.58, Lng:-84.16
Albany, OR, Median AQI = 26 Lat:44.64, Lng:-123.11
Albany-Schenectady-Troy, NY, Median AQI = 39 Lat:42.76, Lng:-73.65
Albuquerque, NM, Median AQI = 54 Lat:35.08, Lng:-106.65
Alexandria, LA, Median AQI = 29 Lat:31.31, Lng:-92.45
Allentown-Bethlehem-Easton, PA-NJ, Median AQI = 42 Lat:40.58, Lng:-75.5
Altoona, PA, Median AQI = 42 Lat:40.52, Lng:-78.39
Amarillo, TX, Median AQI = 45 Lat:35.22, Lng:-101.83
Americus, GA, Median AQI = 40 Lat:32.07, Lng:-84.23
Anchorage, AK, Median AQI = 31 Lat:61.22, Lng:-149.9
Ann Arbor, MI, Median AQI = 39 Lat:42.28, Lng:-83.74
Appleton, WI, Median AQI = 38 Lat:44.26, Lng:-88.42
Ardmore, OK, Median AQI = 44 Lat:34.17, Lng:-97.14
Arkadelphia, AR, Median AQI = 33 Lat:34.12, Lng:-93.05
Asheville, NC, Median AQI = 46 Lat:35.6, Lng:-82.55
Ashtabula, OH, Median AQI = 38 Lat:41.87, Lng:-80.79
Athens, OH, Median AQI = 27 Lat:39.33, Lng:-82.1
Athens, TN, Median AQI = 29 Lat:35.44, Lng:-84.59
Athens-Clarke County, GA, Median AQI = 44 Lat:33.95, Lng:-83.36
Atlanta-Sandy Springs-Roswell, GA, Median AQI = 55 Lat:33.85, Lng:-84.44
Atlantic City-Hammonton, NJ, Median AQI = 40 Lat:39.64, Lng:-74.8
Augusta-Richmond County, GA-SC, Median AQI = 48 Lat:33.47, Lng:-82.01
Augusta-Waterville, ME, Median AQI = 33 Lat:44.45, Lng:-69.7
Austin-Round Rock, TX, Median AQI = 45 Lat:30.12, Lng:-97.61
Bakersfield, CA, Median AQI = 44 Lat:35.37, Lng:-119.02
Baltimore-Columbia-Towson, MD, Median AQI = 47 Lat:39.37, Lng:-76.64
Bangor, ME, Median AQI = 33 Lat:44.8, Lng:-68.77
Baraboo, WI, Median AQI = 37 Lat:43.47, Lng:-89.74
Barnstable Town, MA, Median AQI = 39 Lat:41.7, Lng:-70.3
Baton Rouge, LA, Median AQI = 45 Lat:30.45, Lng:-91.19
```

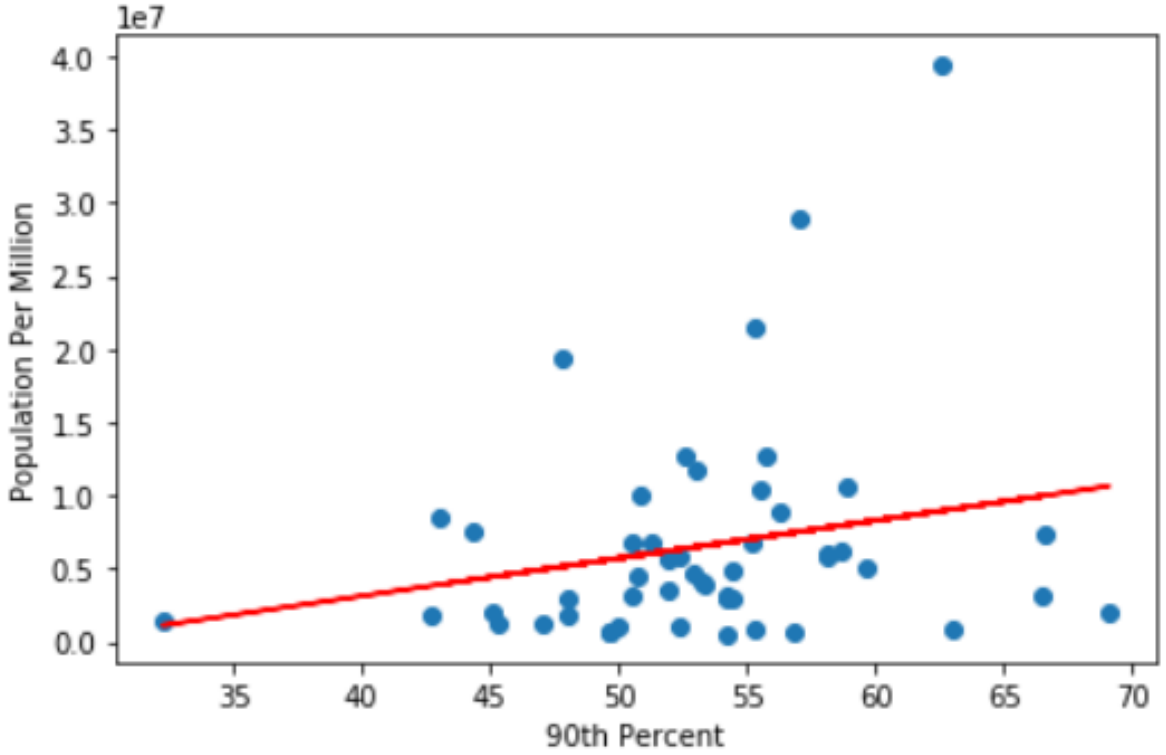
Codes by Bhavna Patadia and Kate Engard

Data Analysis

Scatter Plot & Linear Regression

```
# Create a Scatter Plot for temperature vs latitude
newest_df = pd.merge(state_df, states_updated, on="State")
x_values = newest_df["90th Percentile AQI"]
y_values = newest_df['Population']
(slope, intercept, rvalue, pvalue, stderr) = linregress(x_values, y_values)
regress_values = x_values * slope + intercept
line_eq = "y = " + str(round(slope,2)) + "x + " + str(round(intercept,2))
plt.scatter(x_values,y_values)
plt.xlabel('90th Percent')
plt.ylabel('Population Per Million')
plt.plot(x_values,regress_values,"r-")
plt.tight_layout()
plt.annotate(line_eq, (2,5), fontsize=15,color="red")
print(line_eq);
```

```
y = 258122.01x + -7181262.44
```

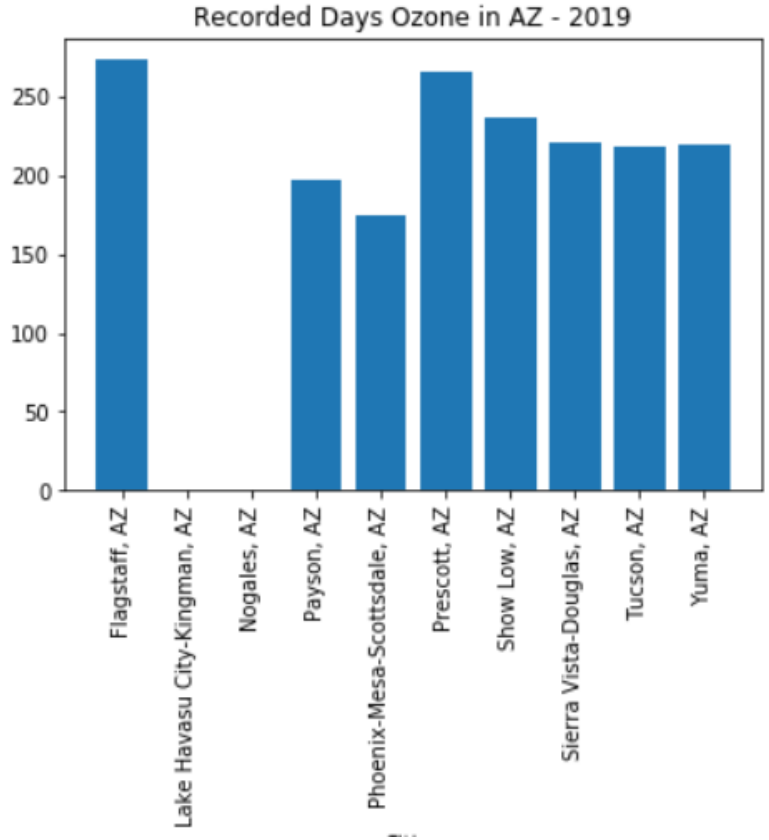


Codes by Quecis Joshua and Kate Engard

Data Analysis

Ozone Barchart for Arizona

```
#create bar chart
x_axis = np.arange(len(AZ_data))
plt.bar(x_axis, AZ_data["Days Ozone"], align="center")
tick_locations = [value for value in x_axis]
plt.xticks(tick_locations, AZ_data["City"], rotation="vertical")
plt.title('Recorded Days Ozone in AZ - 2019')
plt.xlabel('Cities')
plt.show()
AZ_data
```



Codes by MN

Data Analysis

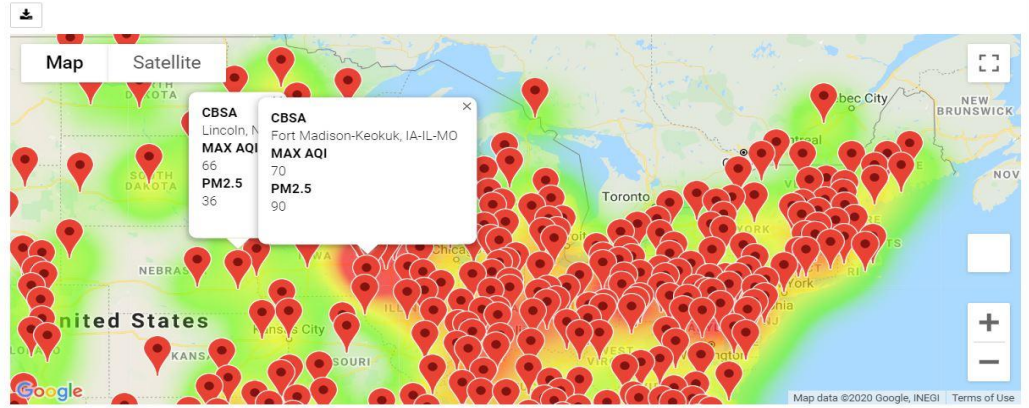
Ozone Barchart for Arizona

Cities

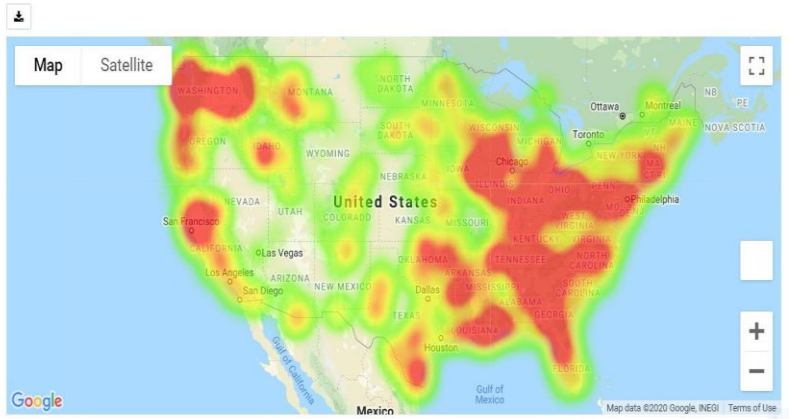
	City	State	Lat	Lng	AQI	Days Ozone	Hazardous Days	90th Percentile AQI
162	Flagstaff, AZ	AZ	35.20	-111.65	46	273	0	61
254	Lake Havasu City-Kingman, AZ	AZ	34.48	-114.32	15	0	0	29
332	Nogales, AZ	AZ	31.34	-110.93	29	0	0	56
354	Payson, AZ	AZ	34.23	-111.33	67	197	0	112
358	Phoenix-Mesa-Scottsdale, AZ	AZ	33.42	-111.83	67	175	7	104
372	Prescott, AZ	AZ	34.54	-112.47	44	266	0	58
433	Show Low, AZ	AZ	34.25	-110.03	45	237	0	61
435	Sierra Vista-Douglas, AZ	AZ	31.83	-109.95	47	221	0	71
469	Tucson, AZ	AZ	32.22	-110.97	49	218	0	74
519	Yuma, AZ	AZ	32.69	-114.63	46	219	0	71

Codes by MN

Data Analysis




```
# Using the template add PM2.5 ,MAX AQI and CBSA name to the heatmap
info_box_template = """
<dl>
<dt>CBSA</dt><dd>{CBSA}</dd>
<dt>MAX AQI</dt><dd>{Max AQI}</dd>
<dt>PM2.5</dt><dd>{PM2_5}</dd>
</dl>
"""

# Store the DataFrame Row

info = [info_box_template.format(**row) for index, row in dataMap_df.iterrows()]

locations = dataMap_df[["Latitude", "Longitude"]]
```



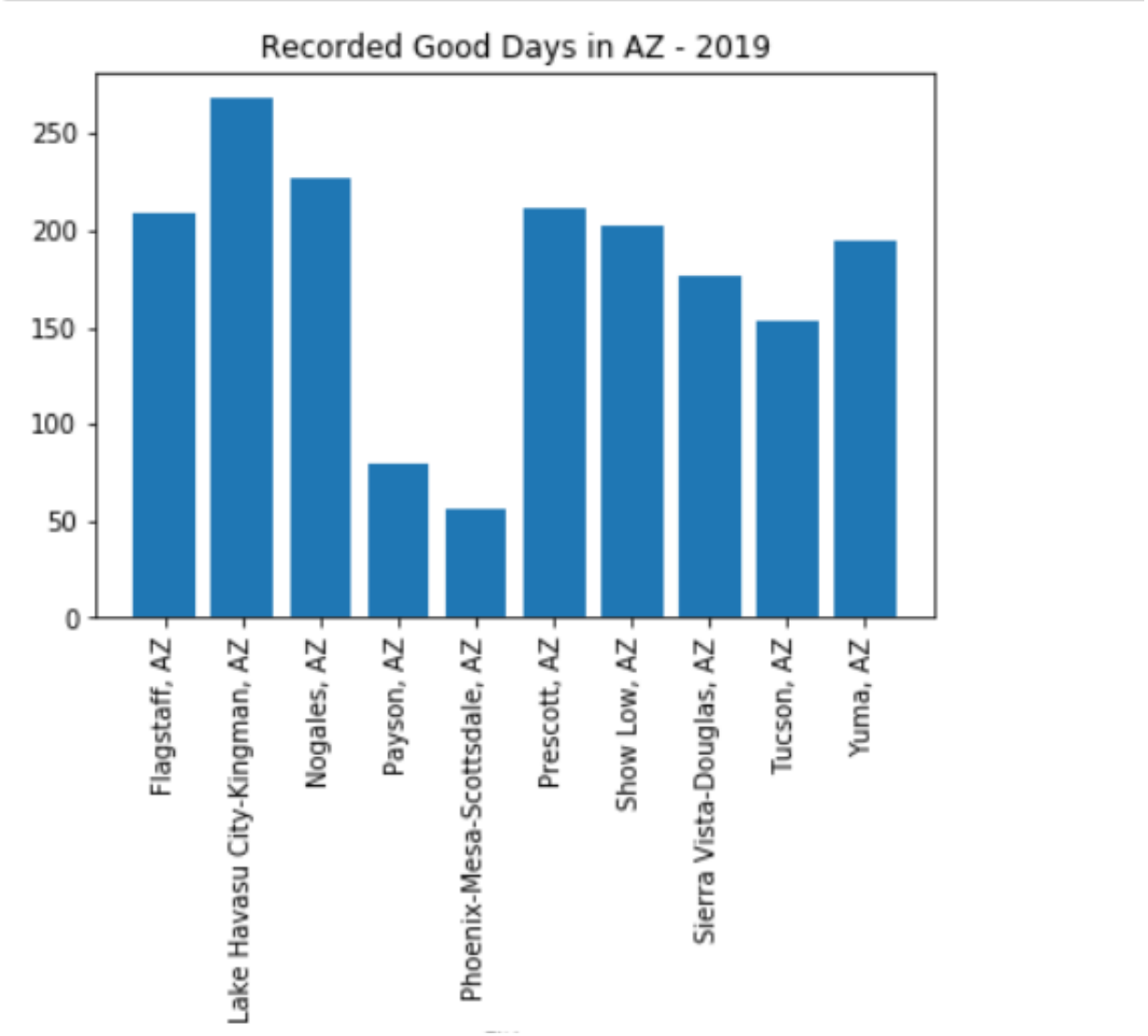

```
# Add marker layer ontop of heat map
markers = gmaps.marker_layer(locations, info_box_content= info)
fig.add_layer(markers)

# Display Map
fig
```

Heat maps

Codes by Soumya Patil and Kate Engard

Annual Air Quality Index

Good days in found cities



Cities

	City	State	Lat	Lng	AQI	Good Days	Hazardous Days	90th Percentile AQI
162	Flagstaff, AZ	AZ	35.20	-111.65	46	209	0	61
254	Lake Havasu City-Kingman, AZ	AZ	34.48	-114.32	15	268	0	29
332	Nogales, AZ	AZ	31.34	-110.93	29	227	0	56
354	Payson, AZ	AZ	34.23	-111.33	67	79	0	112
358	Phoenix-Mesa-Scottsdale, AZ	AZ	33.42	-111.83	67	56	7	104
372	Prescott, AZ	AZ	34.54	-112.47	44	211	0	58
433	Show Low, AZ	AZ	34.25	-110.03	45	203	0	61
435	Sierra Vista-Douglas, AZ	AZ	31.83	-109.95	47	176	0	71
469	Tucson, AZ	AZ	32.22	-110.97	49	153	0	74
519	Yuma, AZ	AZ	32.69	-114.63	46	195	0	71

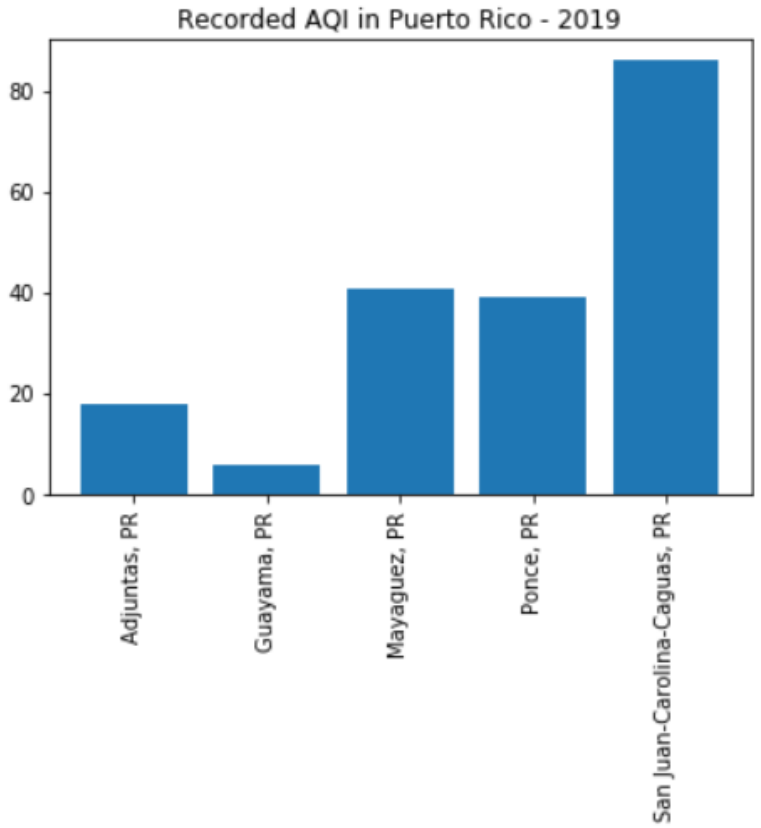
Annual Air Quality Index

Good days in found cities

Annual Air Quality Index

Cities

	City	State	Lat	Lng	AQI	Good Days	Hazardous Days	90th Percentile AQI
2	Adjuntas, PR	PR	18.16	-66.72	18	53	0	44
192	Guayama, PR	PR	17.98	-66.11	6	74	0	7
289	Mayaguez, PR	PR	18.20	-67.15	41	12	0	71
366	Ponce, PR	PR	18.01	-66.61	39	104	0	67
411	San Juan-Carolina-Caguas, PR	PR	18.24	-66.04	86	37	20	316



Good days in found cities


CONCLUSION

Annual Air Quality Index

Conclusions
(an excerpt)

An assumption can be made that region with higher concentration of population & industry can reduce air quality

Congested metropolitan city like Phoenix with high traffic volume and more industries would show lower number of Good Days

(AQI between 0-50) than cities that have lower population, low traffic and no pollution creating industry

Phoenix metropolitan area does experiences lower number of Good Days. Lake Havasu City-Kingman reported

Top 5 States with Hazardous Days resulted in Puerto Rico as a territory with the highest number of Hazardous Days.

Based on the graphs shown the city that had the most Hazardous days is in Puerto Rico.

Annual Air Quality Index

Hypothesis

(an excerpt)

“The Air Quality Index is affected by the amount of people in an area and the average temperature of the area.”

- Quecis Joshua

“Through systematically organizing the data and carefully analyzing it. It was determined that there was no significant correlation between Latitude and Ozone or PM2.5”

- Kate Engard

“The higher the pollutants and the cities with a long history of pollution in the air are most likely fatal for conditions or disease relating to pulmonary system.”.

- Soumya Patil

"In general, as the concentrations of ground-level ozone increases, the number of people with pre-existing conditions are affected. This causes more people with lung diseases to visit doctors or emergency rooms, some are more likely admitted to the hospitals.“

- Bhavna Patadia

“The higher the Ozone gets, the more it’s going to affect the temperature adding to the worsening effect to global climate change.”

- Michael Nosa

Annual Air Quality Index

Team 3

Soumya Patil

Bhavna Patadia

Kate Engard

Quecis Joshua

Michael Nosa


THANK
YOU