Project 1

Youth Asthma Clinic Data Set

Group 2

TABLE OF CONTENTS

01 **Intro**
About our data

02 **Our Questions**
Questions we want to answer

03 **Cleaning**
Steps to clean the dataset

04 **Analysis**
The work done to find our answers

About Our Data Set

Our data is from a Chicago-based youth asthma clinic that partners with schools to provide care to patients across the Chicagoland area.

About Our Data Set

The clinic has CSV files tracking patient appointments over the course of a fiscal year– We are looking at FY23, and FY24. On the day of an appointment each patient fills out a survey, which is used to calculate an “Asthma Control Test” (or ACT) score. Scores range from 0 - 26. Higher scores indicate better management of asthma symptoms.

patient id	check in date	appt date	patient age	patient zip	school code	appt type	act score	school days missed	er visits	hospitalizations
11115	1/31/23	1/31/23	17	60639	331	FOLLOW UP 30	24	0	0	0
11115	1/31/23	1/31/23	17	60639	331	FOLLOW UP 30	24	0	0	0
11115	1/31/23	1/31/23	17	60639	331	FOLLOW UP 30				
11115	1/31/23	1/31/23	17	60639	331	FOLLOW UP 30				
11115	5/24/23	5/24/23	17	60639	331	SKIN TEST 45	24	0	0	0
11115	5/24/23	5/24/23	17	60639	331	SKIN TEST 45	24	0	0	0
11115	5/24/23	5/24/23	17	60639	331	SKIN TEST 45	24	0	0	0
11115	5/24/23	5/24/23	17	60639	331	SKIN TEST 45	24	0	0	0
11115	5/24/23	5/24/23	17	60639	331	SKIN TEST 45				
11115	5/24/23	5/24/23	17	60639	331	SKIN TEST 45				

Questions

01

Demographics

What can we learn about patient demographics, from the given data?

02

Correlations

Is there a correlation between ACT score and visits, age, school days missed, and medical visits?

03

Seasons

Does the time of the year have a significant impact of ACT score?

04

API - Application Programming Interface

How could we utilize API to analyze environmental factors and how it relates to the patient data?

Data Cleaning

We started by concatenating our CSV files into a single data frame, and removed rows with duplicate values. We then converted appointment dates into a consistent format, in order to sort our rows chronologically by date.

When making our calculations we made sure to ignore invalid and non-numerical values, where necessary.

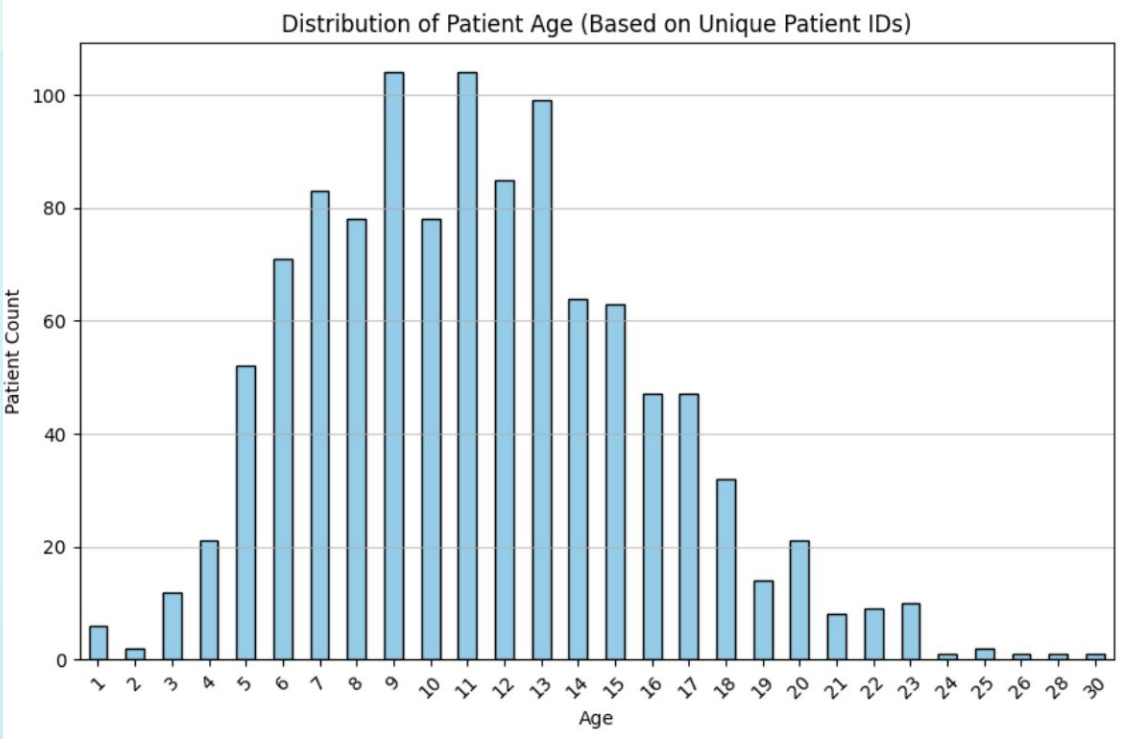
For example: ignoring ACT scores over 26, which is the max. We also combined “ER Visits” and “Hospitalizations” into a new column when analyzing correlations with ACT scores.

```
FY23 = "Resources/FY23.csv"
FY24 = "Resources/FY24.csv"

# Read the CSV files into DataFrames
FY23_df = pd.read_csv(FY23)
FY24_df = pd.read_csv(FY24)

# Concatenate the DataFrames
combined_df = pd.concat([FY23_df, FY24_df])

# Remove rows with exactly the same values
combined_df = combined_df.drop_duplicates()

# Convert 'appt date' column to datetime format
combined_df['appt date'] = pd.to_datetime(combined_df['appt date'], format='%m/%d/%y')

# Sort the combined DataFrame by the 'appt date' column and reset index
sorted_df = combined_df.sort_values(by='appt date').reset_index(drop=True)

sorted_df
```

	patient id	check in date	appt date	patient age	patient zip	school code	appt type	act score	school days missed	er visits	hospitalizations
0	13403	6/29/23	2022-07-13	9	60629	NaN	NEW PATIENT 45	15.0	0.0	0.0	0.0
1	17896	3/22/23	2022-07-13	7	60629	334	NEW PATIENT 45	24.0	0.0	2.0	8.0
2	20749	3/15/23	2022-07-13	5	60629	334	NEW PATIENT 45	99.0	0.0	0.0	0.0
3	20788	3/16/23	2022-07-13	9	60629	334	SKIN TEST 45	17.0	10.0	0.0	0.0
4	11825	6/29/23	2022-07-13	11	60629	334	NEW PATIENT 45	24.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...
5482	21723	3/28/24	2024-03-28	13	60632	310.0	Telehealth	NaN	NaN	NaN	NaN
5483	21723	3/28/24	2024-03-28	13	60632	310.0	Telehealth	25.0	0.0	0.0	0.0
5484	21443	3/28/24	2024-03-28	16	60632	310.0	Telehealth	20.0	0.0	0.0	0.0
5485	19705	3/28/24	2024-03-28	14	60632	310.0	Telehealth	24.0	0.0	0.0	0.0
5486	21443	3/28/24	2024-03-28	16	60632	310.0	Telehealth	NaN	NaN	NaN	NaN

Patient Demographics

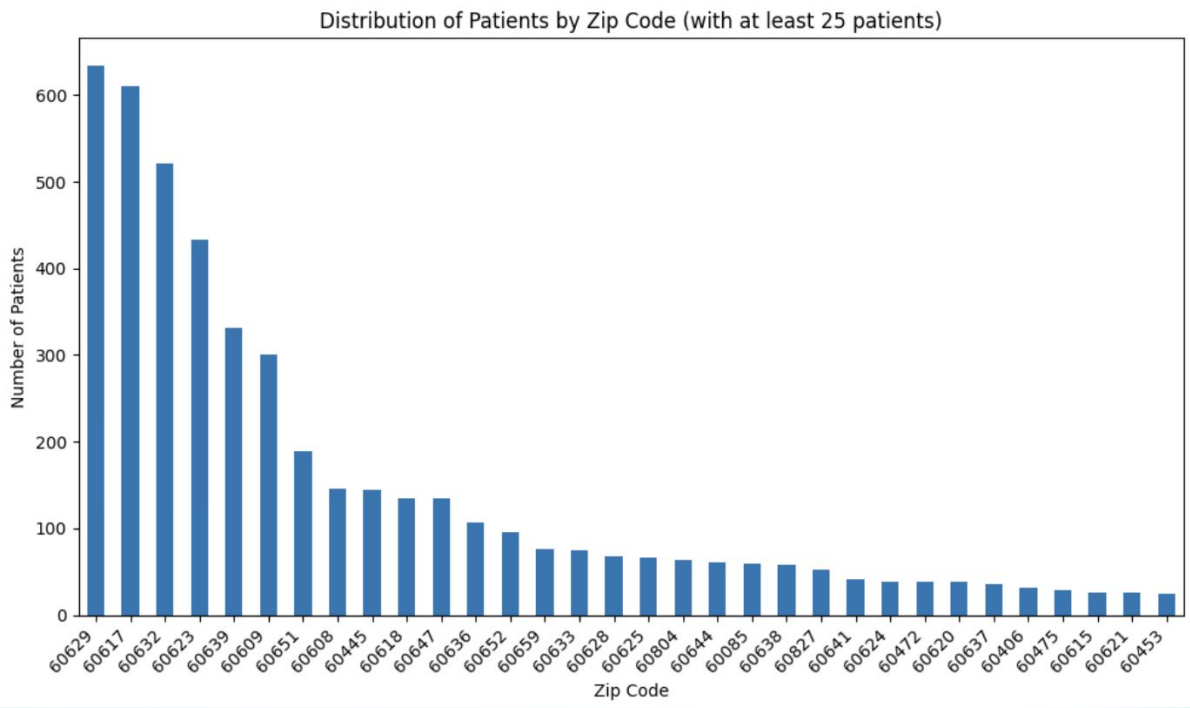
How old are the patients?

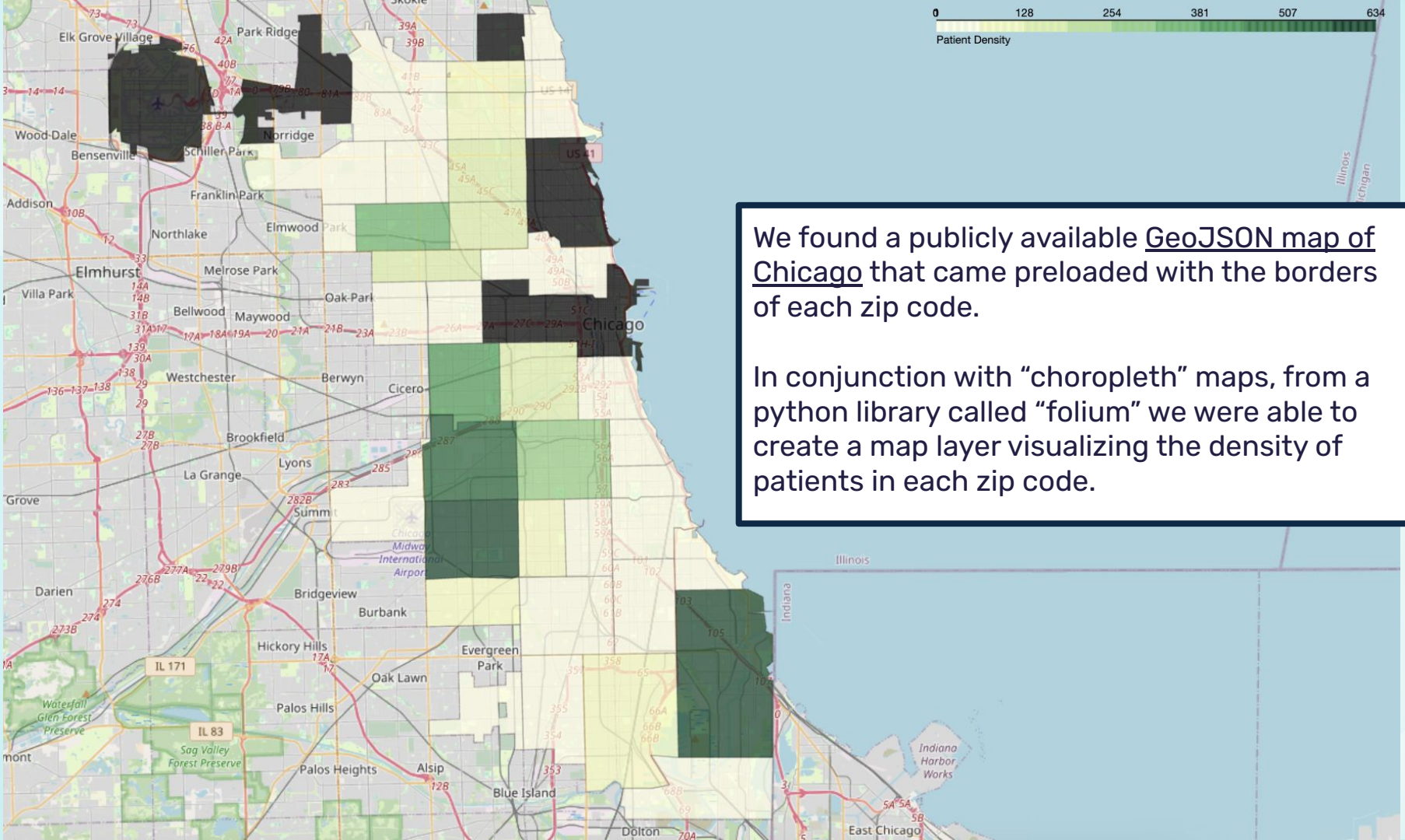


The clinic treated patients ages 1 - 30 years old, with the bulk of patients seen ranging from grade school to high school age.

In what areas do the patients live?

The clinic saw patients across 94 different zip codes in the Chicagoland area. The majority of patients seen were concentrated within the top 10 zip codes, all located on the South and West sides.





We found a publicly available GeoJSON map of Chicago that came preloaded with the borders of each zip code.

In conjunction with “choropleth” maps, from a python library called “folium” we were able to create a map layer visualizing the density of patients in each zip code.

Correlations

Correlations - ACT Score vs. Number of Visits

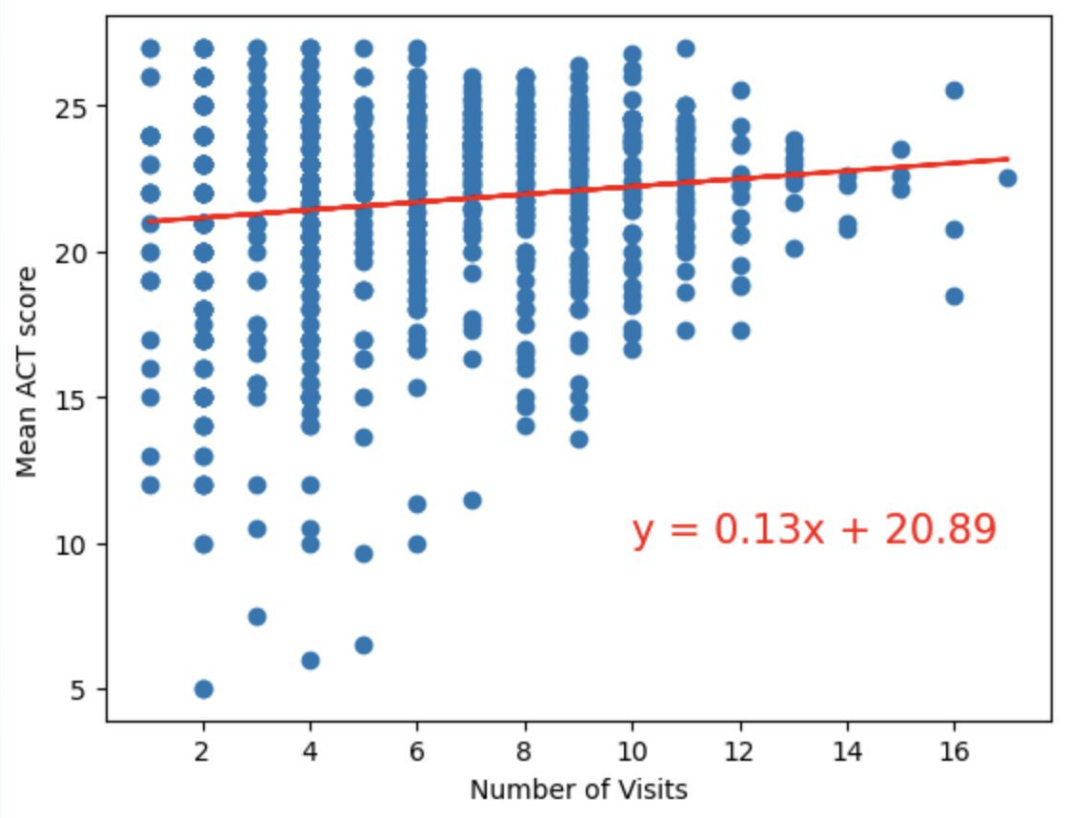
Slope: 0.13

R-value: 0.12

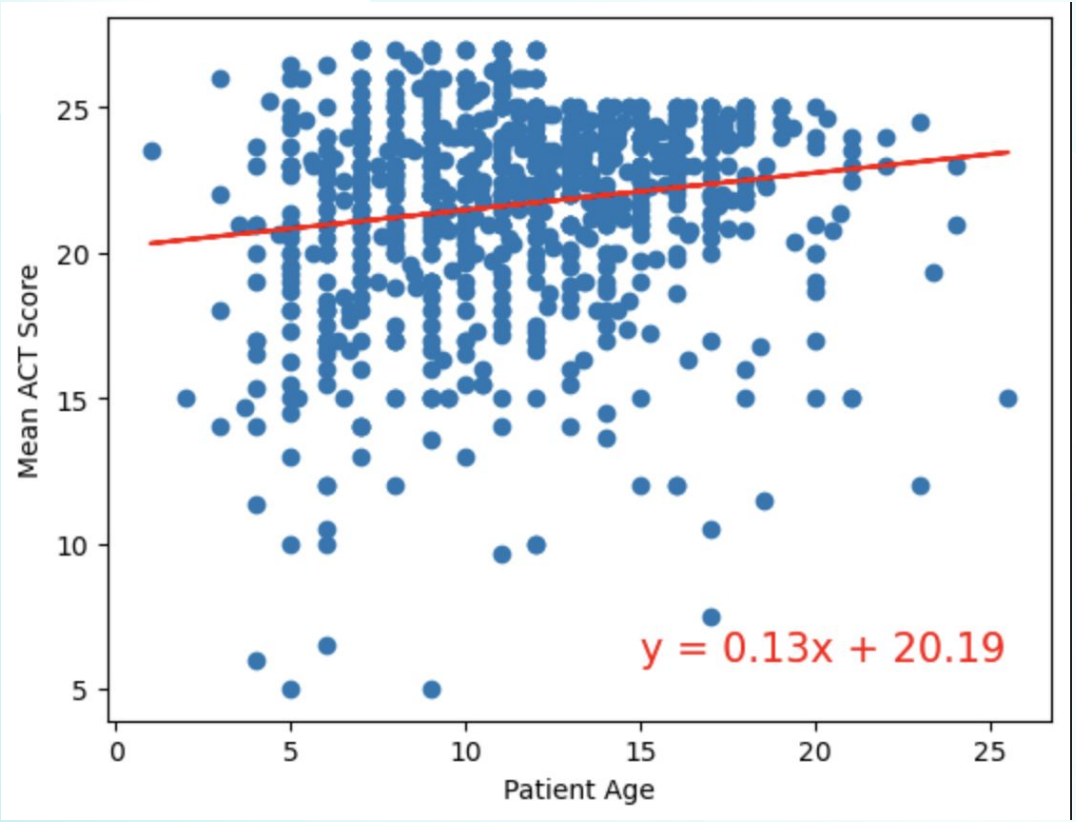
P-value: 0.000328

ACT Score and Number of Visits have a significant correlation

The more a patient visits, the better their ACT score.



Correlations - ACT Score vs. Patient Age



Slope: 0.13

R-value: 0.15

P-value: 1.27e-5

ACT Score and Patient Age have a significant correlation

The older a patient is, the better their ACT score.

Correlations - ACT Score vs. School Days Missed

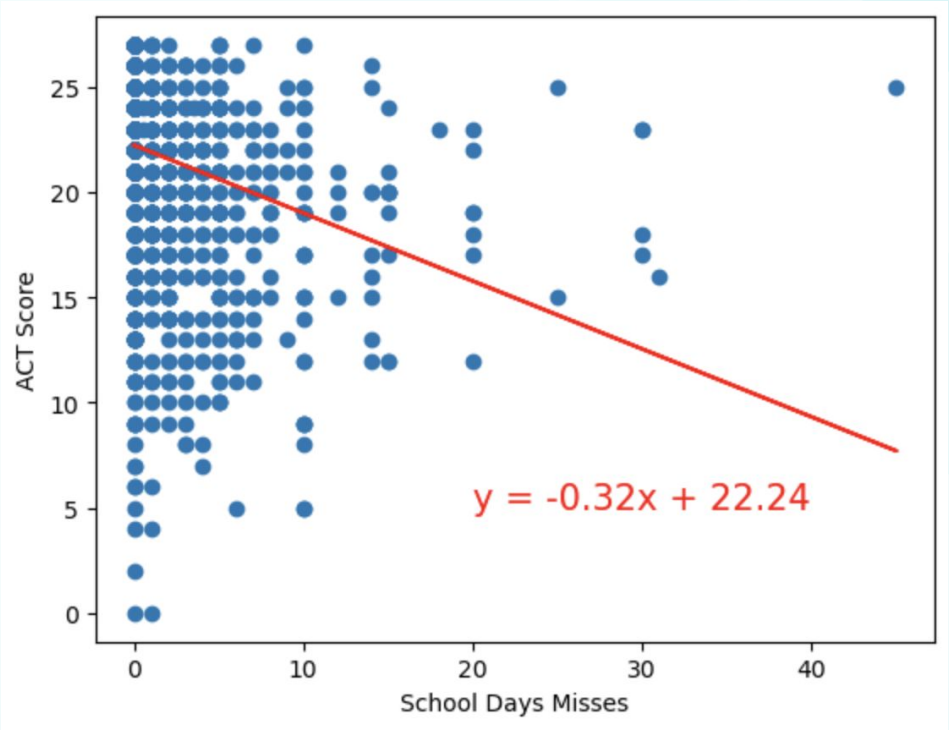
Slope: -0.32

R-value: -0.25

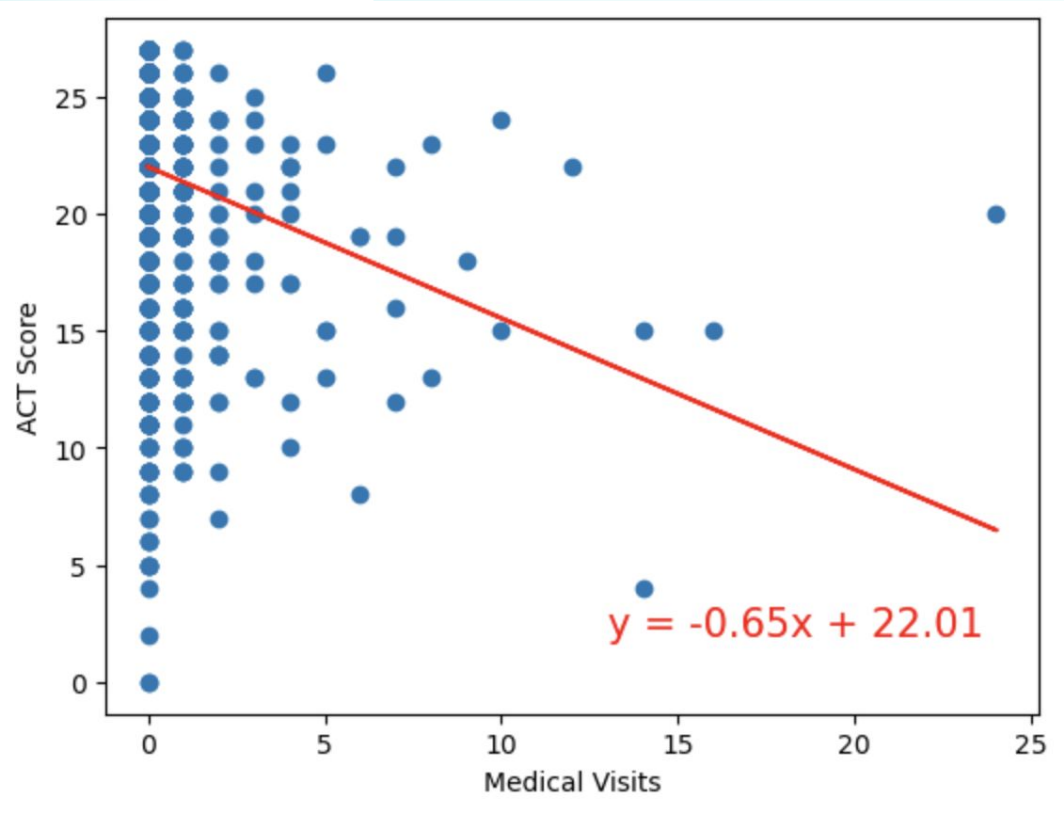
P-value: 2.16e-35

ACT Score and School Days Missed have a significant correlation

The more school days a patient misses, the worse their ACT score.



Correlations - ACT Score vs. Medical Visits



Slope: -0.65

R-value: -0.17

P-value: 4.99e-17

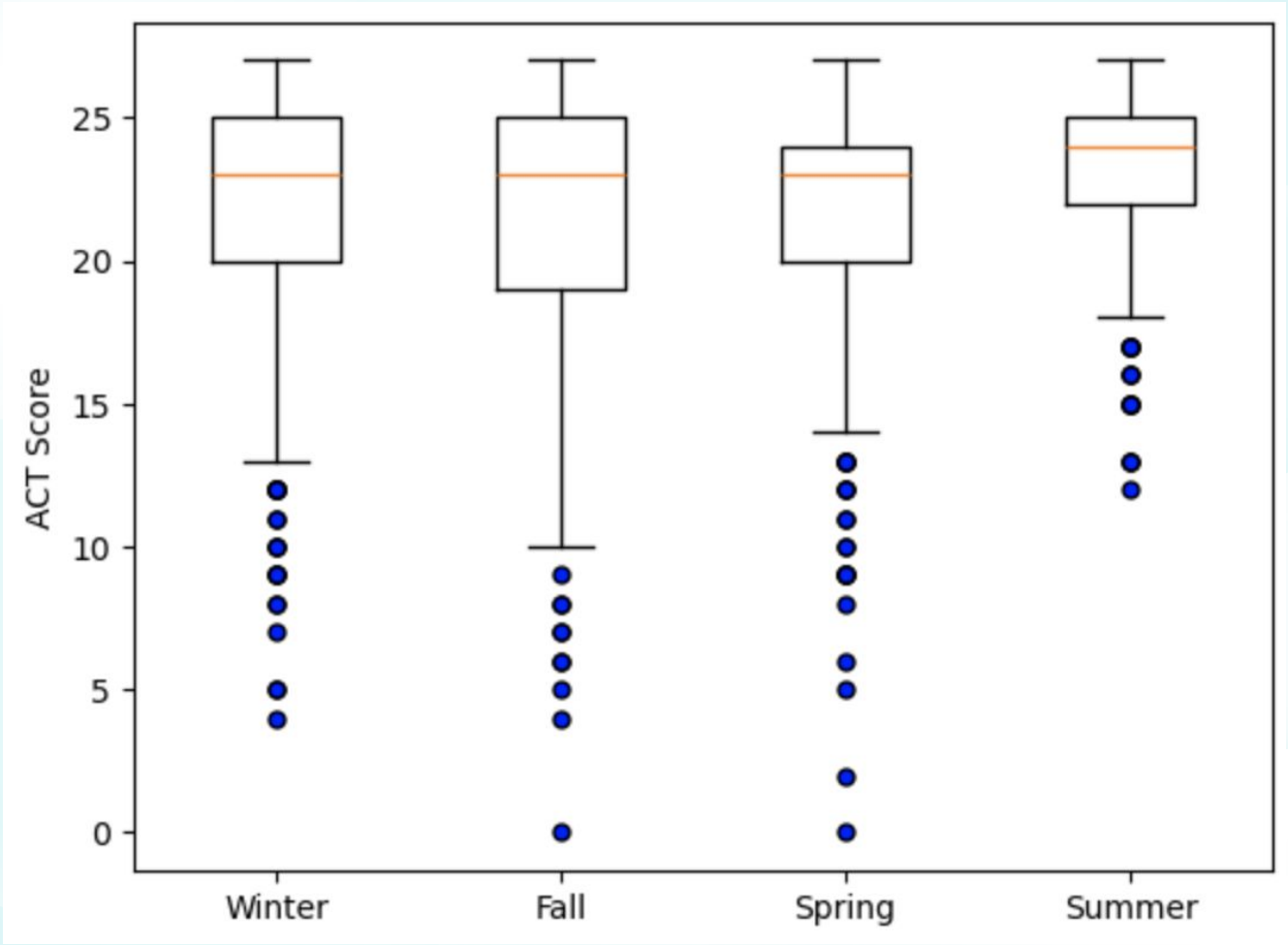
ACT Score and Medical Visits have a significant correlation

The more medical visits, the worse their ACT score.

Seasons

Seasons - Average ACT Scores

Fall	Winter
21. 539310	21.821566
Spring	**Summer**
21.681275	23.253378


ACT Score
25
20
15
10
5
0
Winter
Fall
Spring
Summer

ANOVA

Overall:	F-stat: 14.41	P-value: 2.65e-9

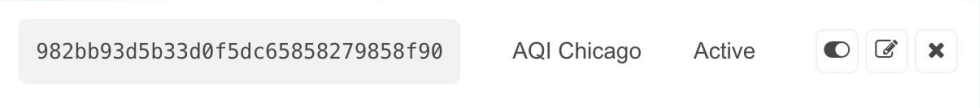
	p-value	Reject
Fall Spring	0.9246	False
Fall Summer	0.0	True
Fall Winter	0.5026	False
Spring Summer	0.0	True
Spring Winter	0.924	False
Summer Winter	0.0	True

Air Quality Data Collection

In our project, we implemented code to monitor Air Quality Index (**AQI**) as a significant factor influencing patient well-being, using the OpenWeatherMap API to help identify health risks for asthma patients. With real-time air pollution data for specific locations in Chicago, we monitored AQI levels by zip code. Regular monitoring allows healthcare providers to offer timely warnings, advising patients to avoid outdoor activities during periods of high pollution or to use prescribed medications.

OpenWeather Air Quality Index Levels Scale Description

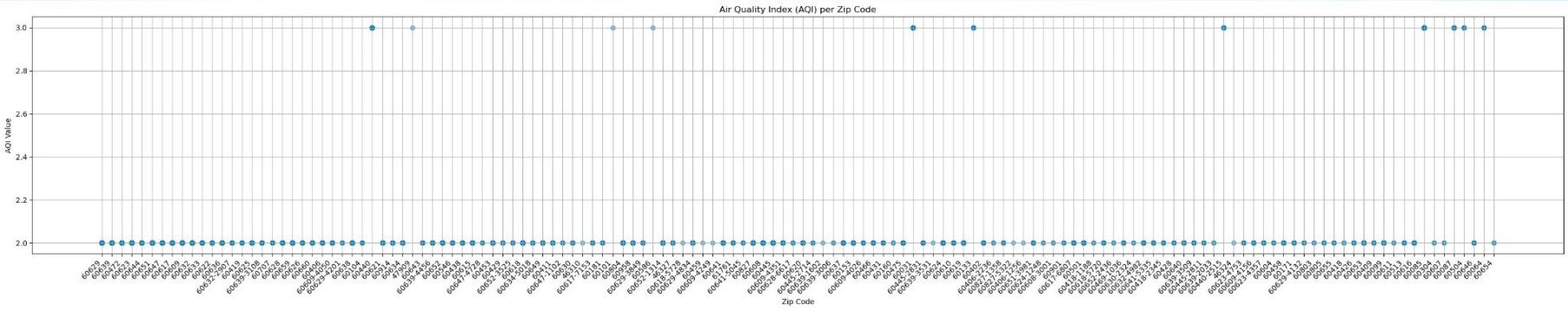
Qualitative name	Index	Pollutant concentration in μg/m^3					
		SO_2	NO_2	PM_{10}	$PM_{2.5}$	O_3	CO
Good	1	[0; 20)	[0; 40)	[0; 20)	[0; 10)	[0; 60)	[0; 4400)
Fair	2	[20; 80)	[40; 70)	[20; 50)	[10; 25)	[60; 100)	[4400; 9400)
Moderate	3	[80; 250)	[70; 150)	[50; 100)	[25; 50)	[100; 140)	[9400-12400)
Poor	4	[250; 350)	[150; 200)	[100; 200)	[50; 75)	[140; 180)	[12400; 15400)
Very Poor	5	⩾350	⩾200	⩾200	⩾75	⩾180	⩾15400



Air Quality Data Collection - Results

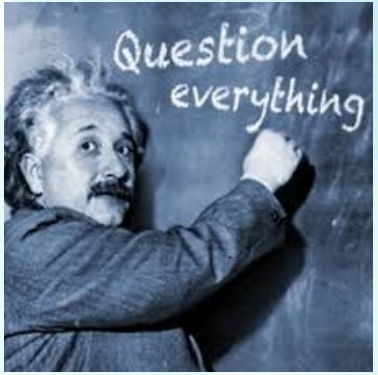
Analyzing the graph over time can reveal temporal trends in air quality within different zip codes. For example, certain areas may experience seasonal variations in AQI due to factors such as weather changes, seasonal emissions, or variations in human activities.

The most common AQI level observed in the city is 2, which is categorized as "Fair" according to the OpenWeather scale for Air Quality Index levels. This indicates moderate air quality. Based on this information, we can conclude that the overall air quality in the city is acceptable. However, it's important to continue tracking AQI levels to ensure that air quality stays within safe limits and to take necessary steps, especially for individuals with respiratory conditions like asthma.

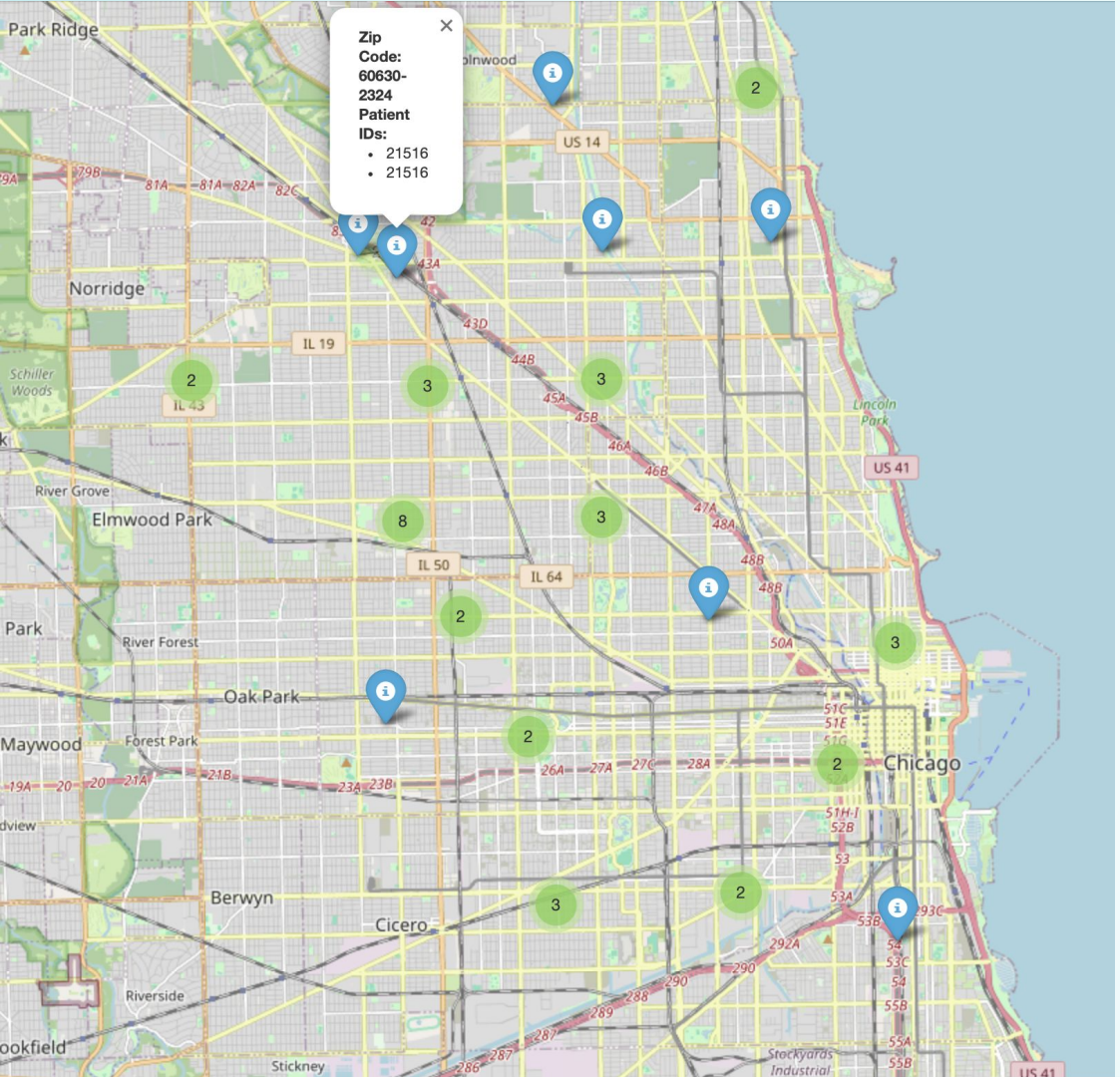


The API implementation was driven by our desire to answer the question: "In the case of a poor or very poor (AQI), can we identify exposed patients and alert them to the situation?" This would enable them to take preventive steps to reduce the impact on individuals' health.


```
Zip Code: 60472
--------------------------------
AQI Level: Fair
Patient IDs:
19498
19503
20232
19147
19148
20270
12752
22348
22398


Zip Code: 60623
--------------------------------
AQI Level: Fair
Patient IDs:
15755
11184
17102
17885
15421
20423
13204
```




```
Zip Code: 60633
--------------------------------
AQI Level: Fair
Patient IDs:
21089
19295
15317
14539
14536
11332
22394
22449
22446
22447
22445
23026


Zip Code: 60622
--------------------------------
AQI Level: Fair
Patient IDs:
16712
17062
22851
22942
```



We also implemented code to track the Air Quality Index (AQI) on the map of Chicago, locating patients with their respective IDs based on their zip code. We defined the following colors according to the AQI:

- Green - Good
- Yellow - Fair
- Orange - Moderate
- Red - Poor
- Dark Red - Very poor

THANK YOU!