Walking in Clean Los Angeles Air

Data Scientists:

Airton Kamdem, Morgan Hofmann, Sarah Sturgeon & Steven Litvack-Winkler

Agenda

- Problem & context
- Data collection & exploration
- Modeling
- Recommendations
- Questions

Problem statement

Can we determine if areas of high walkability in Los Angeles are areas of good air quality?

Walkability

Air Quality Measures

Five major pollutants:

1. Ground-level ozone
2. Particle pollution
3. Carbon monoxide
4. Sulfur dioxide
5. Nitrogen dioxide

Air quality index (AQI) basics

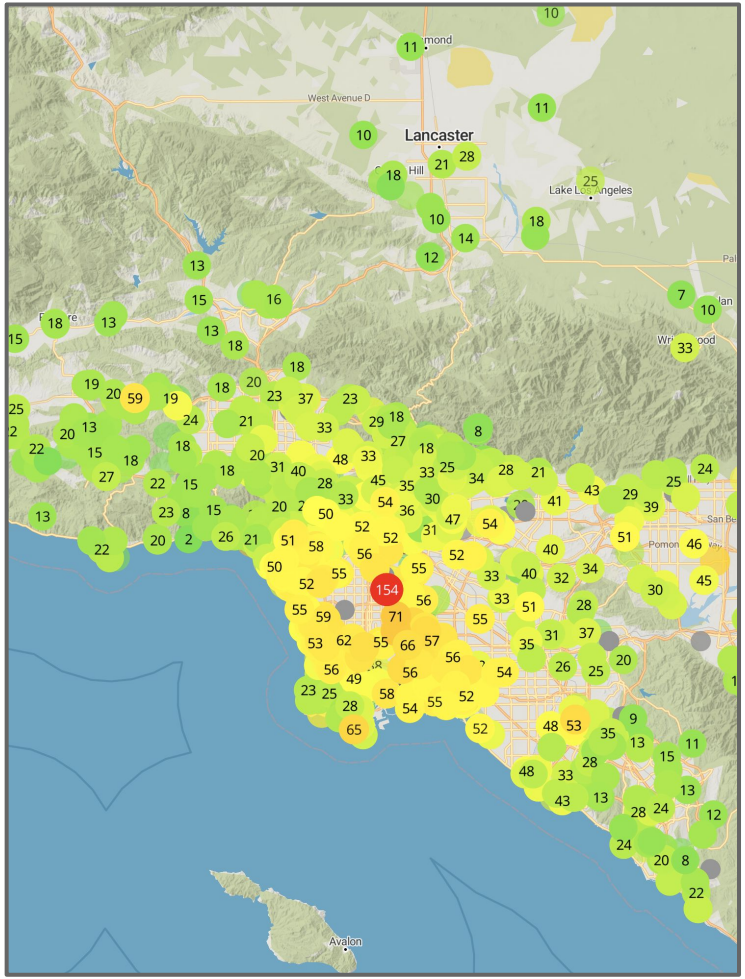
Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

PurpleAir outdoor monitors

PurpleAir PA-II
$249

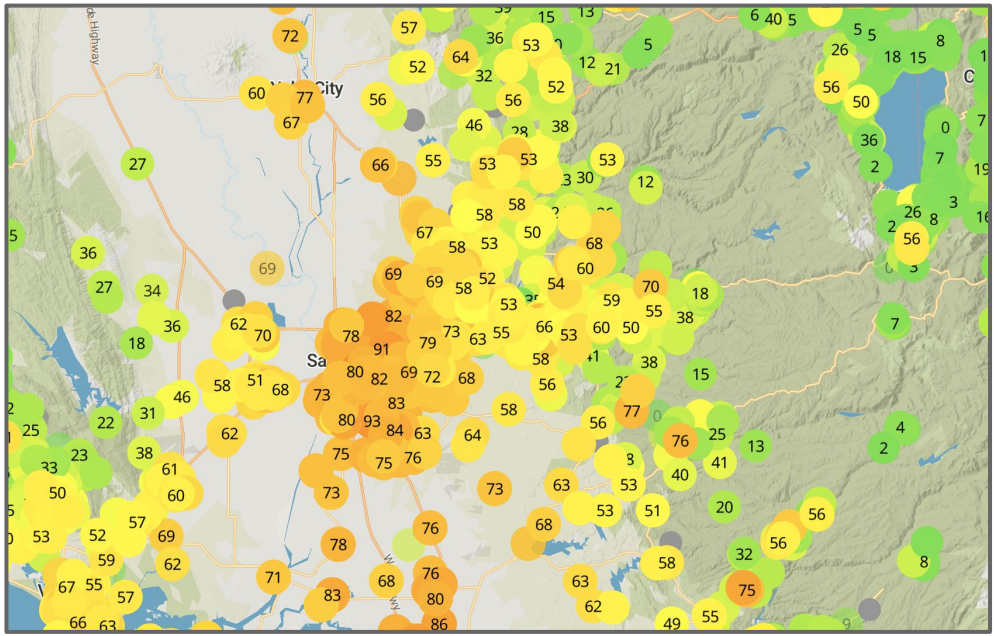
PurpleAir PA-II-SD
$279

AQ REGIONS OF INTEREST

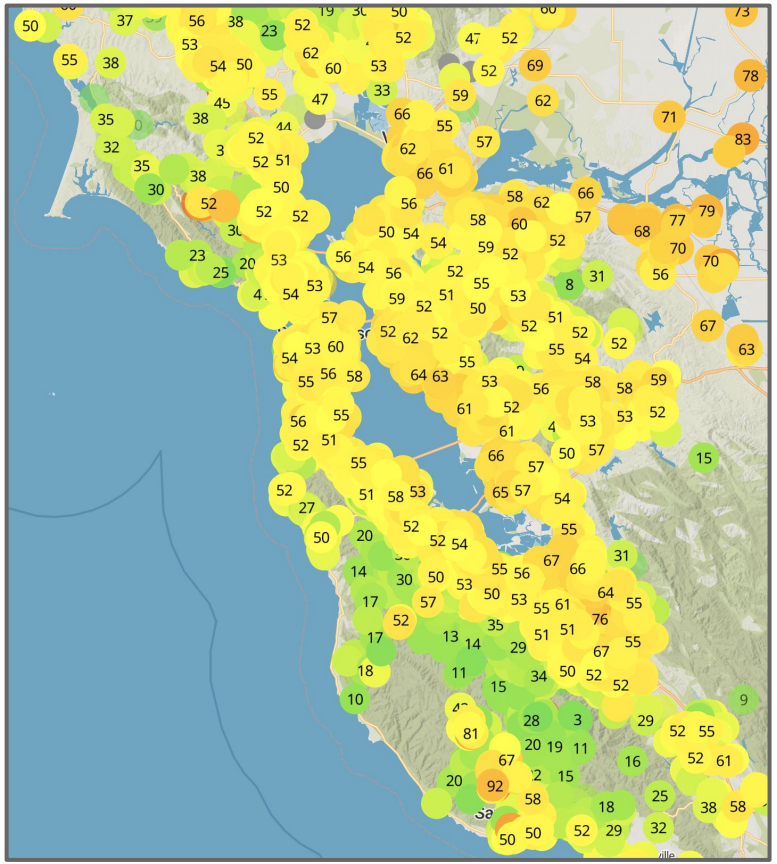


Los-Angeles-Long Beach

AQ REGIONS OF INTEREST



Sacramento-Roseville



San Jose-San Francisco-Oakland

Data Collection

- Walkability data
  - U.S. EPA Walkability Index CSV
- Air Quality (AQ) data
  - PurpleAir API

Merging Data

- FCC Block API
  - Walkability: FIPS code
  - AQ: latitude/longitude
- Merged dataset:
  - 140 features
  - 9,538 observations across USA

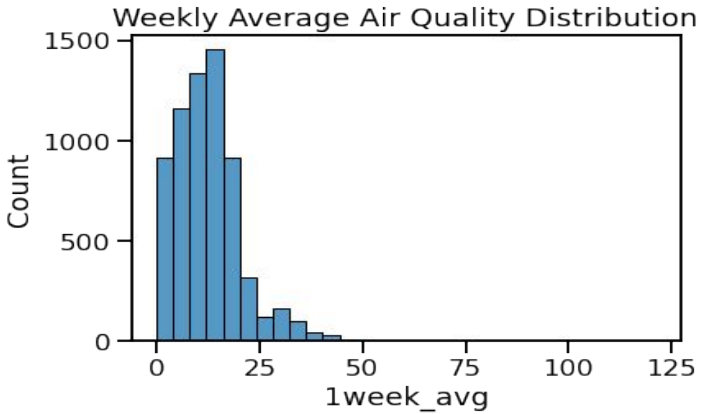
Cleaning Data

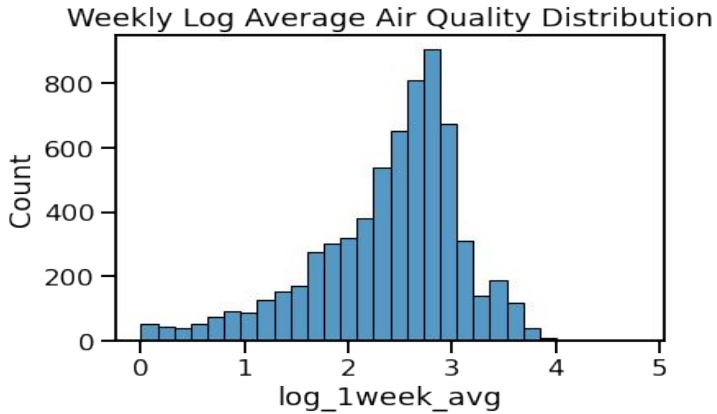
- Filtering for observations that are:
  - In Los Angeles, San Francisco, & Sacramento
  - Outside AQ sensors
- Drop redundant & stagnant features
- Impute nulls & irregular values
- Consideration of outliers

Feature Creation

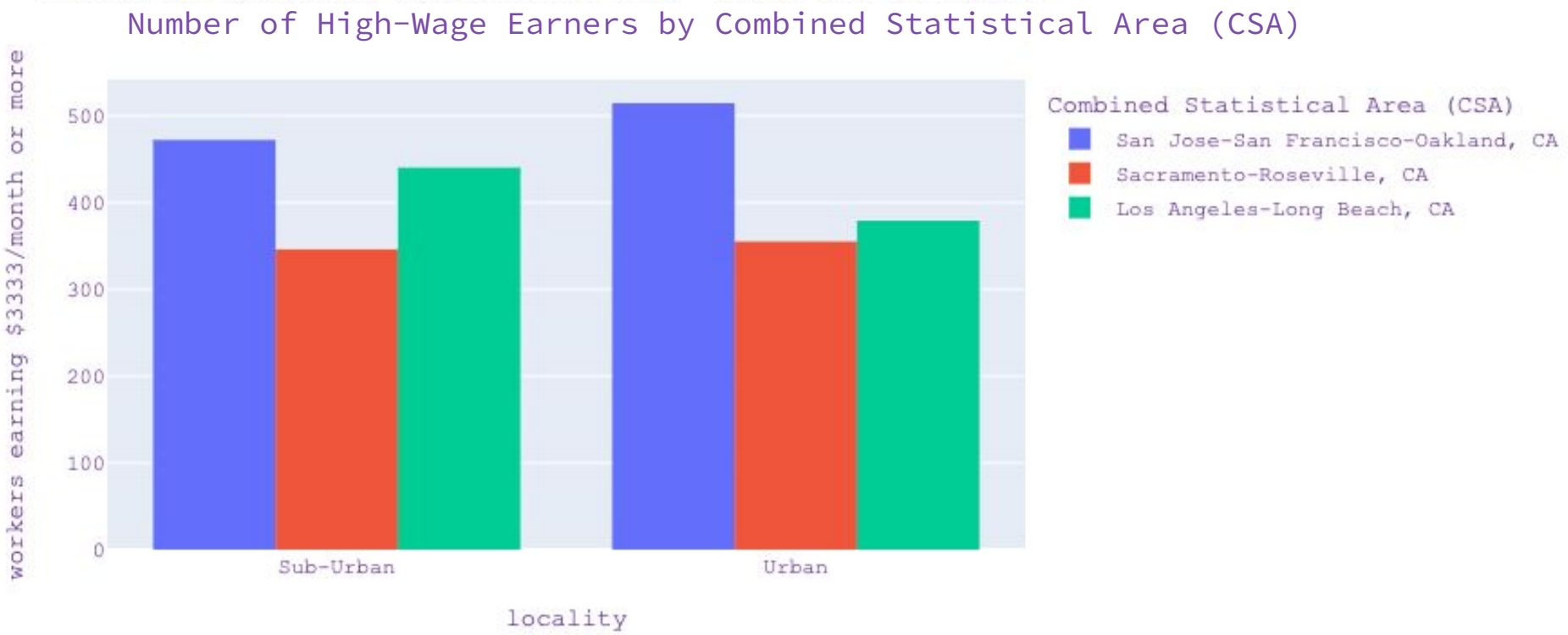
New features:

- Percentage variables
- Logged variables



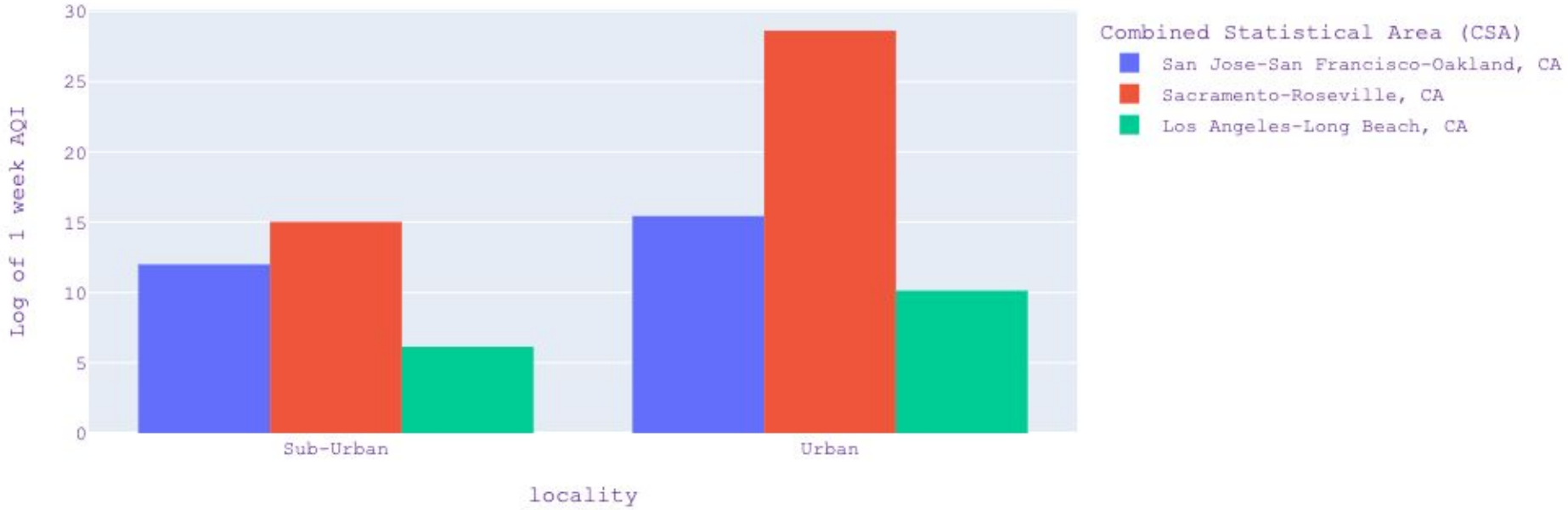


EDA

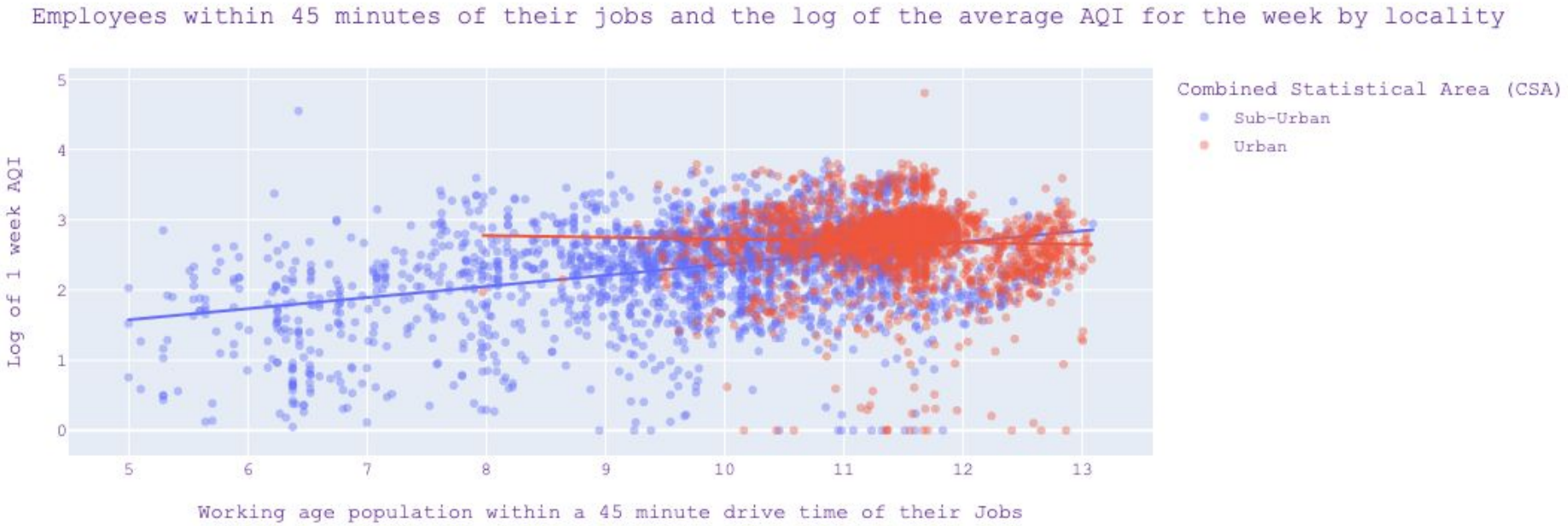

Number of High-Wage Earners by Combined Statistical Area (CSA)
workers earning $3333/month or more
0
100
200
300
400
500
Sub-Urban
Urban
locality
Combined Statistical Area (CSA)
San Jose-San Francisco-Oakland, CA
Sacramento-Roseville, CA
Los Angeles-Long Beach, CA

EDA

Log of 1-Week Air Quality Average by Locality & Combined Statistical Area (CSA)



EDA


Employees within 45 minutes of their jobs and the log of the average AQI for the week by locality
Log of 1 week AQI
0
1
2
3
4
5
5
6
7
8
9
10
11
12
13
Working age population within a 45 minute drive time of their Jobs
Combined Statistical Area (CSA)
Sub-Urban
Urban

EDA


Working age population within 45 minutes of their jobs and the log of the average AQI for the week
Log of 1 week AQI
0
1
2
3
4
5
5
6
7
8
9
10
11
12
13
Working age population within a 45 minute drive time of their Jobs
Combined Statistical Area (CSA)
San Jose-San Francisco-Oakland, CA
Sacramento-Roseville, CA
Los Angeles-Long Beach, CA

EDA


Walkability and the log of AQI
Log of Weekly AQI
0
1
2
3
4
5
5
10
15
20
Walkability Index Rating (Higher numbers are more walkable)
Combined Statistical Area (CSA)
San Jose-San Francisco-Oakland, CA
Sacramento-Roseville, CA
Los Angeles-Long Beach, CA

Model Comparison

- Linear regression models
- Focus on building sparse models
- LA region

Model	% Explained Variance
Baseline	0.0
6 Features (National Model)	26.8
5 Interpretable Features (LA)	34.7
8 Features (LA)	36.8
57 Features (LA)	44.3

Our model

- Top 5 features selected to optimize for score and interpretability
- Los Angeles-Long Beach region only

Feature	Expected Change in log AQI per unit
Log of Working Age Population Within 45 Minutes Car Commute	0.1843
Proportion of Jobs in Office Sector	0.0036
Proportion of Jobs in Industrial Sector	0.0028
2+car Households	-0.0046
Log of Census Block Group Acreage	-0.0799

LIMITATIONS TO OUR PROCESS

- PurpleAir sensors are less likely in low income areas
- EPA walkability dataset features were engineered from multiple datasets
  - This data is not updated regularly
- Air quality varies considerably by time, weather, season, and local wildfire status

NEXT STEPS

- Deeper analysis of factors correlated with air quality
- Analysis of tradeoffs between air pollution and other concerns such as cost of living

Takeaways & Recommendations

- Air quality varies considerably across California
  - LA > San Francisco > Sacramento

- Suburban regions have better air quality than urban ones
- Despite car pollution, more vehicles are associated with better air quality

Sources + More Info

- EPA Walkability Index dataset
- PurpleAir API
- PurpleAir website
- FCC Block API
- AirNow website
- Proximity One FIPS info
- Air Pollution from Vehicles in California: People of Color Bear the Biggest Burden (article by David Reichmuth)

Thank you!

Questions?