Ground-Level Ozone In U.S. Counties and Air Quality Health

By

Spencer Rubin

Springboard: Capstone2

Background - Problem:

- Ground-level Ozone (O3) - hazardous air-pollutant
- Result of anthropogenic and natural forces
- Monitored by the EPA
- O3 can result in respiratory disease and illness

- **How does O3 health concern differ between major counties in the U.S?**

Applying Data Science

- Classifying county-level O3 concentration:
  - Not of health concern (<0.06ppm)
  - Health concern (>0.06ppm)
- County population as a feature contributing to O3 health quality
- 10-most populous counties in U.S.

Acquiring Datasets

- EPA Daily Air Quality data for 2019 in 10-counties (.csv): https://www.epa.gov/outdoor-air-quality-data/download-daily-data

- Population estimation data for each county in 2019 (.csv): https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html

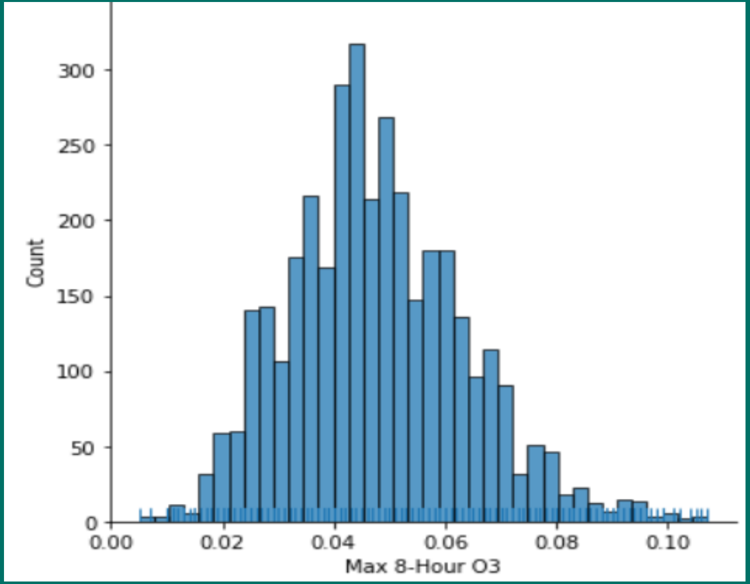
- Counties:
  - Los Angeles County, CA
  - Orange County, CA
  - San Diego County, CA
  - Riverside County, CA
  - Dallas County, TX
  - Harris County, TX
  - Miami-Dade County, FL
  - Cook County, IL
  - Queens County, NY
  - Maricopa County, AZ

Data Wrangling

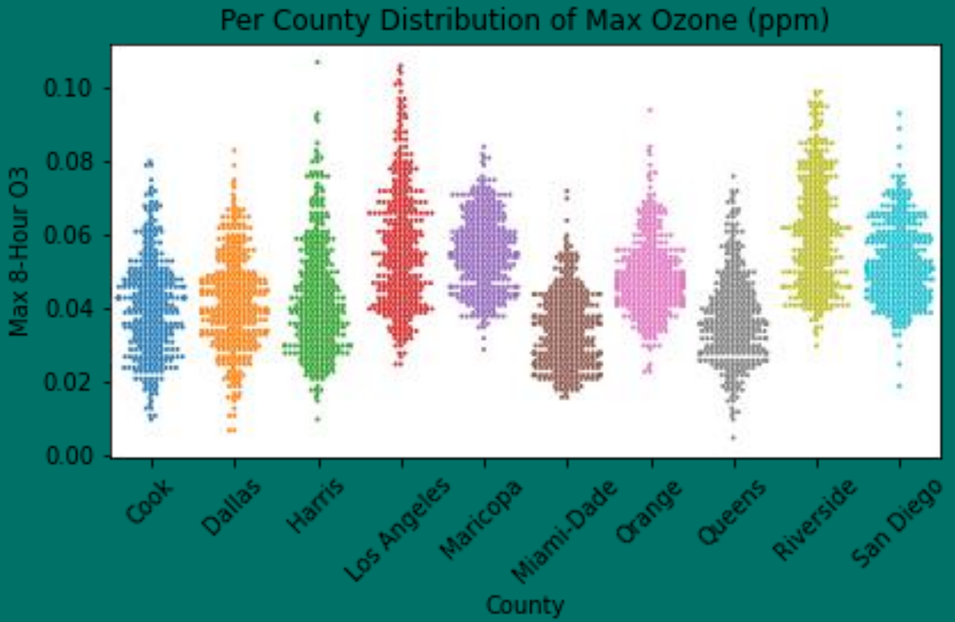
- Maximum Daily 8-hour Ozone averaged for county-wide EPA air measurement stations.
- One Ozone concentration value (ppm) per day (observation) per county
- Population (per million/people) add for each observation by county
- Time series (date) indexed.
- Dataframe:
  - 3606 observations
  - No missing values
  - Variance with number of averaged EPA air measurement stations per county

Exploratory Data Analysis

- Overall distribution of Maximum 8-hour O3 ppm



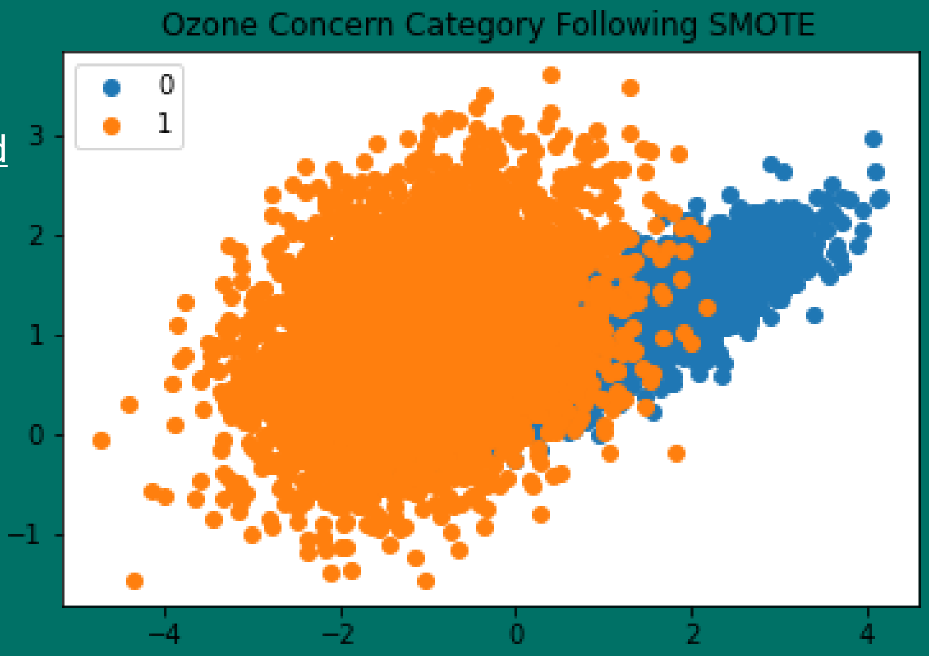
- County-level distribution of Maximum 8-hour O3 ppm



Dataset Balancing

- Classifying days per county O3 >0.06ppm:
- Most days are not above this threshold

- imblearn - SMOTE for resampling an unbalanced dataset
- Synthetic resampling of unbalanced dataset to reduce bias



Dataset Preprocessing

- Categorical variables (pandas get_dummies()):
    - County
    - Population (above or below the median of 10-counties
    - “ozone_concern”: above or below 0.06ppm
    - “ozone_threshold”: above or below 0.007ppm

- Scale “Max 8-hour Ozone (ppm)”:
    - sklearn’s StandardScaler()

Model Selection

- Supervised learning models for classification:
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier

- Metric for selection: “Accuracy”

  - *Which model can classify best if a county has O3 related air quality health concern?*

Model Selection

- Logistic Regression: ROC_AUC

- *Model accuracy:*
  - *classifying "ozone_concern" by county relative to population*

classifier model type	**model metric**
	Accuracy Score
Logistic Regression	0.8876
Decision Tree	0.8856
Random Forest	0.899

- Random Forest: *90% Accuracy*

Conclusions

- Supervised learning classification models can have high accuracy for determining O3 health concern
    - Random Forest Classifier performed best
- Balancing the dataset is critical to developing classification model
    - Most days of the year are not O3 related air quality issues
- County population is a strong indicator of high concentrations of O3
    - Larger population ~ more anthropogenic forcing of O3 production

Future Considerations

- Feature selection considerations:
    - Population density
    - Seasonality
    - Meteorological components: sunshine, temperature, wind
    - County-size
    - Geography of county; EPA measurement stations

- O3 < 0.06ppm of no significant health concern
- O3 0.06 – 0.07ppm: health concern!
- O3 >0.07ppm: EPA non-attainment zones; serious respiratory health concerns
    - Los Angeles and Riverside counties have notable days with O3 above 0.07ppm!

Questions?