Air Quality and Health Outcomes in Kentucky in 2019

Jianing, Madhvi, Maggie, Sanha

PPOL564: Final Class Status Update. 11.30.2022

Outline

- Motivation
- Research Questions
- Data
- Methods
- Results thus far
- Limitations/Next Steps

Motivation

- Why air quality and sulfur?
  - PM 2.5 is the standard measure for air pollution
    - fine particulate matter of 2.5 microns or less in diameter (PM2.5)
  - Sulfur dioxide is a byproduct of burning fossil fuels (like coal and oil) used in domestic heating, power generation and motor vehicles.
- Why Kentucky?
  - One of largest coal-producing states
  - Ranked in bottom 5 states for health outcomes according to a report by NiceRx
- Why 2019?
  - Latest data available for both air quality and health outcomes
- Why certain health categories?
  - Initially focused on lung-related health categories (asthma, lung disease, cancer)
  - Research suggested cardiovascular and mental health outcomes may be connected to air quality too

Research Questions

- Is there a connection between air quality and certain health outcomes?
  - Ex: cancer, lung disease, and asthma?
  - Ex: mental health?
- Is there a connection between sulfur and certain health outcomes?
  - Ex: cancer, lung disease, and asthma?
  - Ex: mental health outcomes?

Data Sources 1/3

- Air Quality Data
  - Source: U.S. Environmental Protection Agency

```
Int64Index: 8336 entries, 102990 to 111325
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   State Name                8336 non-null   object
 1   county Name               8336 non-null   object
 2   State Code                8336 non-null   int64
 3   County Code               8336 non-null   int64
 4   Date                      8336 non-null   object
 5   AQI                       8336 non-null   int64
 6   Category                  8336 non-null   object
 7   Defining Parameter        8336 non-null   object
 8   Defining Site             8336 non-null   object
 9   Number of Sites Reporting 8336 non-null   int64
dtypes: int64(4), object(6)
memory usage: 716.4+ KB
```

Data Sources 2/3

- Sulfur Data
  - Source: U.S. Environmental Protection Agency

```
Int64Index: 7776 entries, 162862 to 170637
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   State Code           7776 non-null   int64
 1   County Code          7776 non-null   int64
 2   Site Num             7776 non-null   int64
 3   Parameter Code       7776 non-null   int64
 4   POC                  7776 non-null   int64
 5   Latitude             7776 non-null   float64
 6   Longitude            7776 non-null   float64
 7   Datum                7776 non-null   object
 8   Parameter Name       7776 non-null   object
 9   Sample Duration      7776 non-null   object
 10  Pollutant Standard   7776 non-null   object
 11  Date Local           7776 non-null   object
 12  Units of Measure     7776 non-null   object
 13  Event Type           7776 non-null   object
 14  Observation Count    7776 non-null   int64
 15  Observation Percent  7776 non-null   float64
 16  Arithmetic Mean      7776 non-null   float64
 17  1st Max Value        7776 non-null   float64
 18  1st Max Hour         7776 non-null   int64
 19  AQI                  7776 non-null   int64
 20  Method Code          0 non-null      float64
 21  Method Name          7776 non-null   object
 22  Local Site Name      7776 non-null   object
 23  Address              7776 non-null   object
 24  State Name           7776 non-null   object
 25  County Name          7776 non-null   object
 26  City Name            7776 non-null   object
 27  CBSA Name            5719 non-null   object
 28  Date of Last Change  7776 non-null   object
dtypes: float64(6), int64(8), object(15)
memory usage: 1.8+ MB
```

Data Sources 3/3

- Health Data
  - Source: U.S. Centers for Disease Control and Prevention

```
Int64Index: 33170 entries, 3596 to 857829
Data columns (total 23 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Year                         33170 non-null  int64
 1   StateAbbr                    33170 non-null  object
 2   StateDesc                    33170 non-null  object
 3   CountyName                   33170 non-null  object
 4   CountyFIPS                   33170 non-null  int64
 5   LocationName                 33170 non-null  int64
 6   DataSource                   33170 non-null  object
 7   Category                     33170 non-null  object
 8   Measure                      33170 non-null  object
 9   Data_Value_Unit              33170 non-null  object
 10  Data_Value_Type              33170 non-null  object
 11  Data_Value                   33170 non-null  float64
 12  Data_Value_Footnote_Symbol   0 non-null      float64
 13  Data_Value_Footnote          0 non-null      float64
 14  Low_Confidence_Limit         33170 non-null  float64
 15  High_Confidence_Limit        33170 non-null  float64
 16  TotalPopulation              33170 non-null  int64
 17  Geolocation                  33170 non-null  object
 18  LocationID                   33170 non-null  int64
 19  CategoryID                   33170 non-null  object
 20  MeasureId                    33170 non-null  object
 21  DataValueTypeID              33170 non-null  object
 22  Short_Question_Text          33170 non-null  object
dtypes: float64(5), int64(5), object(13)
memory usage: 6.1+ MB
```

Methods: Data Acquisition

- How did we acquire the data?
  - Relatively simple: downloaded publicly-available data
  - Official data: csv data compiled by the federal government
- How did the federal government acquire the data?
  - Health Data (CDC Places)
    - CDC's Behavioral Risk Factor Surveillance System (BRFSS), Census 2010 population counts, annual Census county population estimates, and the Census American Community Survey (ACS) estimates
  - Air Quality Data (EPA AQI) and Sulfur (EPA SO2)
    - "The Clean Air Act requires that state, local, and tribal air pollution control agencies monitor the air for ambient levels of certain pollutants. . . In addition to the required monitoring, many agencies perform additional and/or voluntary monitoring of substances and meteorological parameters." - EPA.gov
    - Daily and yearly measures

Methods: Data Cleaning

- Subset all three datasets to Kentucky (and 2019 data for health)
- For health data, create a function to subset to measures we'd like to focus on
  - Health Outcomes:
    - current asthma (CASTHMA)
    - chronic obstructive pulmonary disease (COPD)
    - cancer (excluding skin cancer) (CANCER)
    - depression (DEPRESSION)
  - Prevention:
    - routine checkup within the past year (CHECKUP)
    - current lack of health insurance (ACCESS2)
    - taking medicine for high blood pressure control (BPMED)
  - Health risk behavior:
    - current smoking (CSMOKING)
    - no leisure-time physical activity (LPA)
  - Health status:
    - physical health not good (PHLTH)
    - mental health (MHLTH)
    - fair or poor self-rated health status (GHLTH)

Methods: Data Cleaning Code Highlight (Health)

```
# create a function to subset to focus measures for each category
def focus_measures(df, category, measures):
  focus_category = df[df['Category'].isin(category)]
  focus_measures = measures
  focus_category_measures = focus_category[focus_category["MeasureId"].isin(focus_measures)]
  return focus_category_measures

# create df subset to our focus measures
all_categories = ["Health Outcomes","Prevention","Health Risk Behaviors","Health Status"]
all_focus_measures = ["CANCER",'COPD','CASTHMA',"DEPRESSION", "CHECKUP",'ACCESS2','BPMED',"CSMOKING",'LPA',"PHLTH",'MHLTH','GHLTH']
health_ken_2019_fm = focus_measures(health_ken_2019_clean, all_categories, all_focus_measures)
```

Methods: Data Cleaning Code Highlight (all)

```
# Print number of Kentucky counties in each dataset
print("Number of unique county in air data: " + str(
    air_ken_2019['County'].nunique()))

# Number of unique counties in air data: 27

print("Number of unique county in sulfur data: " + str(
    so2_ken["County Name"].nunique()))

# Number of unique county in sulfur data: 10

print("Number of unique counties in health data: " + str(
    health_ken_2019['CountyName'].nunique()))

# Number of unique county in health data: 120
```

Methods: Data Cleaning Code Highlight (SO2)

```
## Create a function to categorize AQI to categorical values
filter_method = lambda x: 'Good' if x < 50 else 'Moderate' if (x > 50 and x <= 100) else 'Unhealthy for Sensitive Groups' if (x > 100 and x <= 150) else "NA" if (x == "NaN") else "Unhealthy"

so2_ken_f["AQI_category"] = so2_ken_f["AQI"].apply(filter_method)
```

Methods: Data Cleaning Code Highlight

```
## creating new columns on these categories
so2_ken_ff["good_aqi"] = (so2_ken_ff["AQI_category"] == "Good")

so2_ken_ff["moderate_aqi"] = (so2_ken_ff["AQI_category"] == "Moderate")

so2_ken_ff["unhealthy_aqi"] = (so2_ken_ff["AQI_category"] == "Unhealthy")

so2_ken_ff["unhealthy_sens_aqi"] = (so2_ken_ff["AQI_category"] == "Unhealthy for Sensitive Groups")
```

Methods: Data Cleaning

- Before merging, we reshaped the sulfur data to create a wide dataframe with new columns showing key statistical info (max, min, etc) for the daily counts

```
so2_reshape = so2_ken_ff.groupby(["County Name", "Site Num"]).agg(so2_n_dates = ("Date Local", "count"),
    so2_good_days = ("good_aqi", "sum"),
    so2_moderate_days = ("moderate_aqi", "sum"),
    so2_unhealthy_days = ("unhealthy_aqi", "sum"),
    so2_unhealthy_sens_days = ("unhealthy_sens_aqi", "sum"),
    so2_avg_aqi = ("AQI", "mean"),
    so2_max_aqi = ("AQI", "max"),
    so2_min_AQI = ("AQI", "min"))
```

Methods: Merging

- Next, we merged the three datasets together on county names

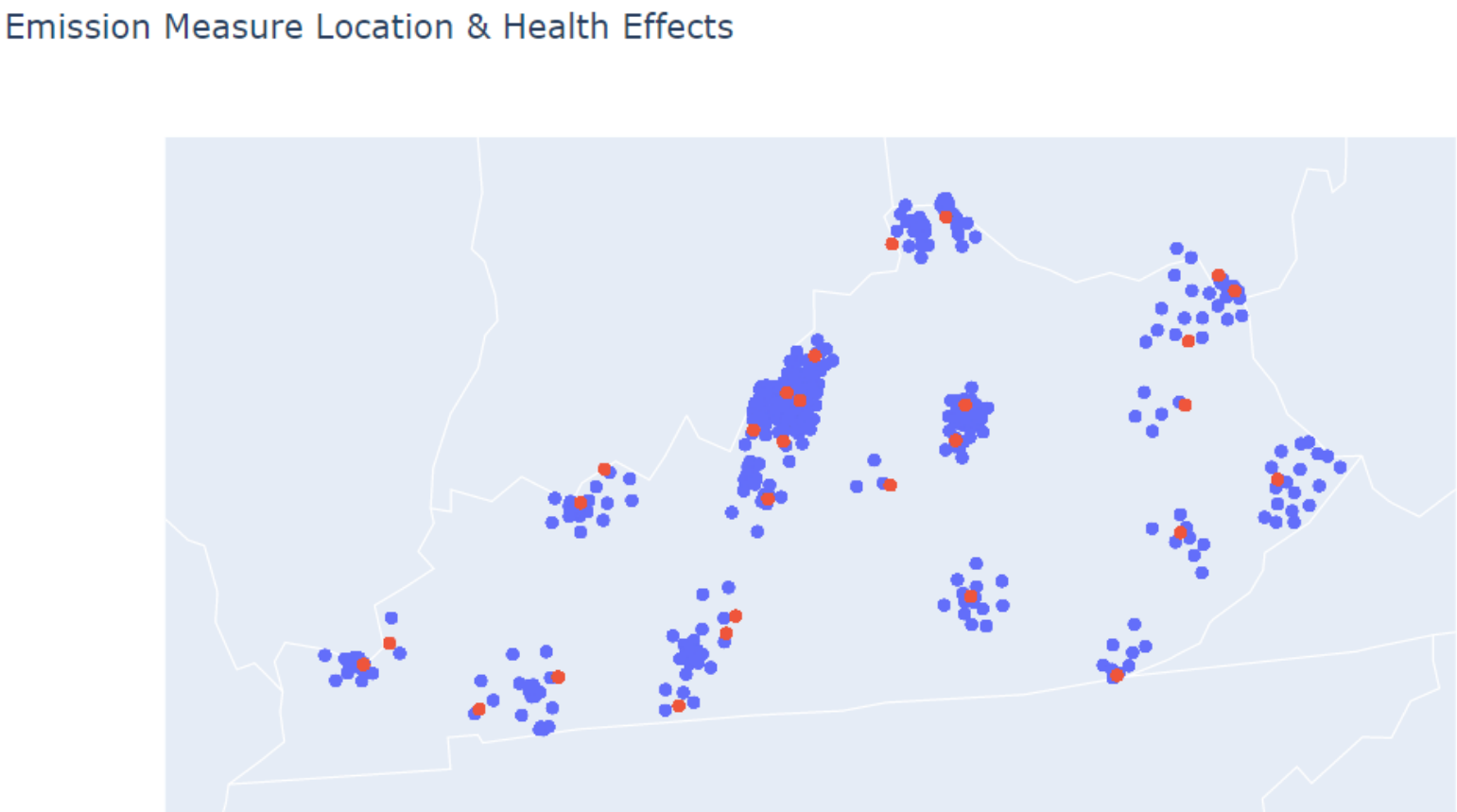
```
## merge the airqualtiy and health data
merged_ah = pd.merge(air_ken_2019, health_ken_2019_fm,
    how = "left", left_on = "County", right_on = "
    CountyName",suffixes=('_air', '_health'))

## merge this new df with the SO2 data
merged_ahs = pd.merge(merged_ah, so2_reshape2, how = "
    inner", left_on = "County", right_on = "County Name",
    suffixes=('_ah', '_Sulfur'))
```

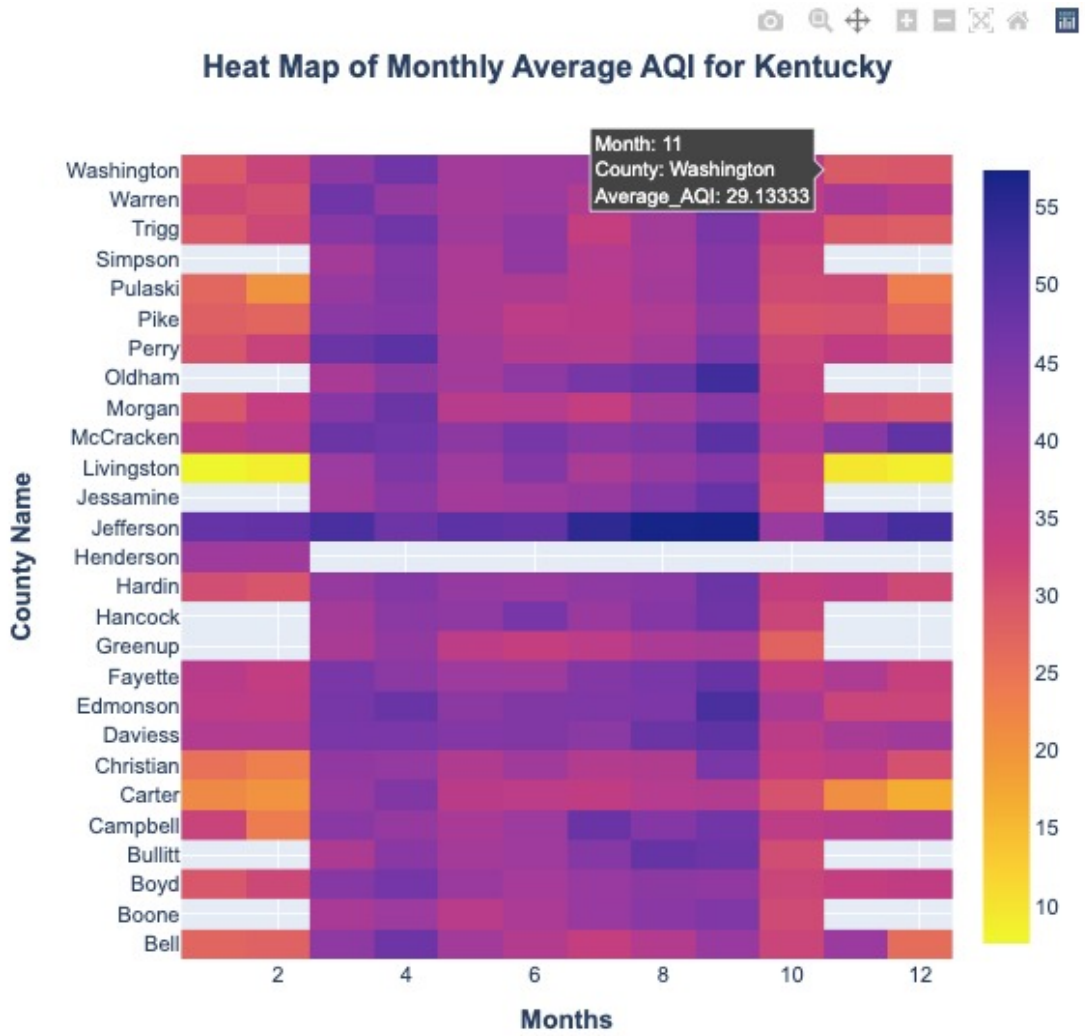
Describing the analytic sample

- We wanted to better understand the location of our health data versus our sulfur data
  - Step 1: create new columns with the lat and long data for the health data and the sulfur data
  - Step 2: Write a function to calculate the distance (mi) between the two points
  - Step 3: Use the function to create a new column

Results: visualization



Results: visualization



Results: visualization

Results: in words

- We are still exploring the data for the relationship between air quality and health but we can already note some conclusions on the air quality dataset
  - Average AQI is worse in Jefferson County (where the largest city, Louisville, is located)
  - Average AQI is better in colder months as compared to warmer months
- Challenges
  - Limited county data
  - Data from measurement site, not necessarily the source of the pollution

Next steps: Our Project

- Additional steps
  - Additional data visualizations to better understand the potential relationship between unhealthy air quality and health outcomes (including lung-related outcomes and mental health)
  - Potential ideas
    - Regression on health data and health outcomes, controlling for health prevention
    - Text scraping press release on air quality

Next steps: Recommendations for Future Work

- Recommendations for future analysis
  - Analyze data in additional states, particularly states with no or little coal production
    - Example, a national analysis with dummy variables for coal producing states and non coal producing states
  - Fill data gaps for Kentucky counties
  - Longitudinal research
- Potential Policy Applications
  - Expand air quality measuring across additional Kentucky counties
    - This may face resistance from coal companies and lobbies