Final Project

November 22, 2021

[19]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from matplotlib.pyplot import figure
from scipy.stats import pearsonr
```

1 Checkpoint 1: EDA

Group members: Alisha Mirapuri, Coco Sun, Nameera Faisal Akhtar, Riya Berry
Question 1: (Causal Inference) Do higher air pollution levels cause an increase in asthma mortality rates?

Question 2: (Comparing GLMs and nonparametric methods): How well does state location predict risk for asthma mortality?

[20]:
```
asthma = pd.read_csv("U.S._Chronic_Disease_Indicators__Asthma.csv")
pollution = pd.read_csv("Daily_Census_Tract-Level_PM2.
 ↪5_Concentrations__2011-2014.csv")
```

1.1 Asthma mortality rates by state

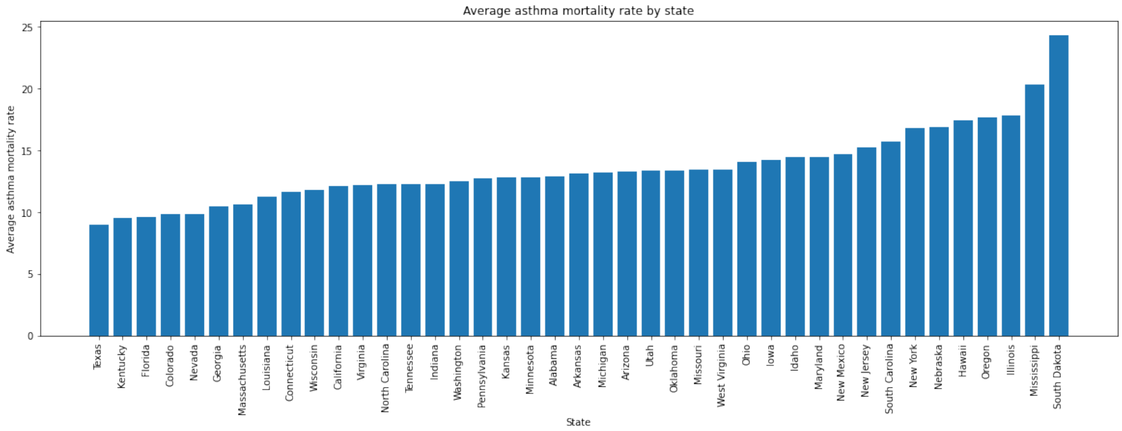
[21]:
```
new_asthma = asthma[~asthma['DataValue'].isna()]
new_asthma = new_asthma[new_asthma['Question'] == "Asthma mortality rate"]
new_asthma = new_asthma[new_asthma['DataValueType'] == "Crude Rate"]
new_asthma.columns
```

[21]:
```
Index(['YearStart', 'YearEnd', 'LocationAbbr', 'LocationDesc', 'DataSource',
       'Topic', 'Question', 'Response', 'DataValueUnit', 'DataValueType',
       'DataValue', 'DataValueAlt', 'DataValueFootnoteSymbol',
       'DatavalueFootnote', 'LowConfidenceLimit', 'HighConfidenceLimit',
       'StratificationCategory1', 'Stratification1', 'StratificationCategory2',
       'Stratification2', 'StratificationCategory3', 'Stratification3',
       'ResponseID', 'LocationID', 'TopicID', 'QuestionID', 'DataValueTypeID',
       'StratificationCategoryID1', 'StratificationID1',
       'StratificationCategoryID2', 'StratificationID2',
       'StratificationCategoryID3', 'StratificationID3'],
      dtype='object')
```

```
[22]: new_asthma = new_asthma[new_asthma['LocationDesc'] != "United States"]
      location = new_asthma[['LocationDesc', 'DataValue']]
      location = location.groupby('LocationDesc').mean()
      location = location.reset_index()
      location = location.sort_values(by='DataValue')
```

```
[23]: plt.figure(figsize=(20, 6))
      plt.bar(location['LocationDesc'], location['DataValue'])
      plt.xticks(location['LocationDesc'], rotation='vertical');
      plt.xlabel("State")
      plt.ylabel("Average asthma mortality rate")
      plt.title("Average asthma mortality rate by state")
```

```
[23]: Text(0.5, 1.0, 'Average asthma mortality rate by state')
```



Written analysis

Describe any trends you observe, and any relationships you may want to follow up on. We observe that states in the north (South Dakota, Illinois, New York) tend to have higher asthma mortality rates than states in the south (Texas, Kentucky, Florida). Although this trend is not perfect and there are outliers (e.g. Mississippi in the South having a high mortality rate), this is a general trend that can be observed for the majority of the states. After this, we would want to follow up on the relationship between asthma morality rates in rural/urban/suburban areas. I would also want to explore why the north has higher asthma mortality rates than the south by looking at other demographic and social factors.

Describe any data cleaning steps you took. How will these decisions impact your model and inferences? We filtered out rows that had a null value for the Asthma Mortality prevalence column. We also removed rows that had a location description of "United States", rather than belonging to a specific state. By removing values with a "United States" location description, we ensure that location is considered on a state-wide basis and granularity is not lost in our model.

Explain how your visualizations are relevant to your research questions: do they motivate the question you're asking? Do they suggest a potential answer? They motivate our question about how state location predicts risk of asthma mortality. E.g., this visualization helps us analyze trends in asthma mortality by geographic location and region in the United States. They movitate larger questions about why certain regions of the US have higher asthma rates; is it due to those areas having less environmental pollutants and being less industrialized?

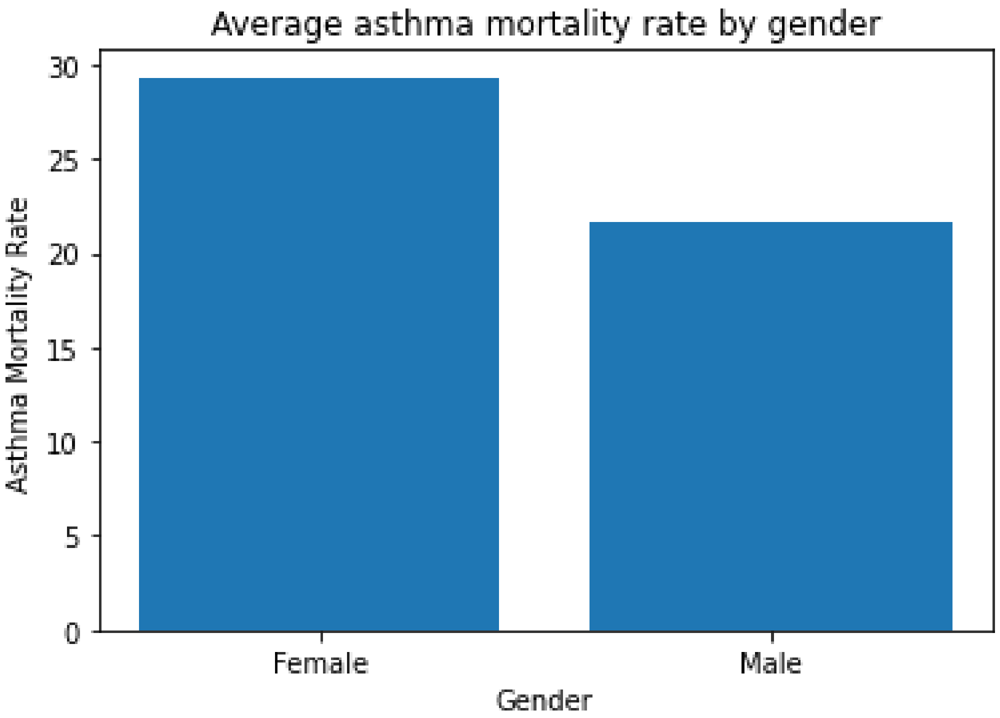
1.2 Asthma mortality rates by gender

[24]:

```
gender = asthma[asthma['StratificationCategoryID1'] == 'GENDER']
gender = gender[gender['Question'] == "Asthma mortality rate"]
gender = gender[['StratificationID1', 'DataValue']]
gender = gender[~gender['DataValue'].isna()]
gender = gender.groupby('StratificationID1').mean()
gender = gender.reset_index()
```

[25]:

```
plt.bar(gender['StratificationID1'], gender['DataValue'])
labels = ["Female", "Male"]
plt.xticks(gender['StratificationID1'], labels, rotation='horizontal')
plt.ylabel("Asthma Mortality Rate")
plt.xlabel("Gender")
plt.title("Average asthma mortality rate by gender")
plt.show()
```



Written Analysis

Describe any trends you observe, and any relationships you may want to follow up on. One trend we observed was that females have a higher average mortality rate from asthma than males. We may want to follow up on the stratificiations of asthma morality rates by race and other demographic factors WITHIN gender; i.e. are women of a certain race more likely to die from asthma than women of other races?

Describe any data cleaning steps you took. How will these decisions impact your model and inferences? We filtered out values that are null for the gender column, and only looked at values relevant to asthma mortality rates. We also changed the x-axis labels to be "Female" and "Male" in order to make the model more interpretable. Because we only focused on the question of asthma mortality, we were able to gather information to build our model solely on mortality rates rather than prevalence and hospitalization rates.

Explain how your visualizations are relevant to your research questions: do they motivate the question you're asking? Do they suggest a potential answer? While we were exploring demographic factors such as state location, we realized that it would be helpful to consider other factors such as gender. This will allow us to analyze how well gender predicts risk for asthma mortality, maybe even in conjuction with state. Our data suggests that women tend to have a higher risk for asthma mortality than men.

1.3 Asthma mortality rates versus pollution levels

[26]:

```
pollution = pollution[['statefips', 'ds_pm_pred']]
```

[27]:

```
pollution = pollution.groupby('statefips').mean()
pollution = pollution.reset_index()
pollution['statefips'] = pollution['statefips'].astype(int)
pollution = pollution[~pollution['statefips'].isna()]
```

[28]:

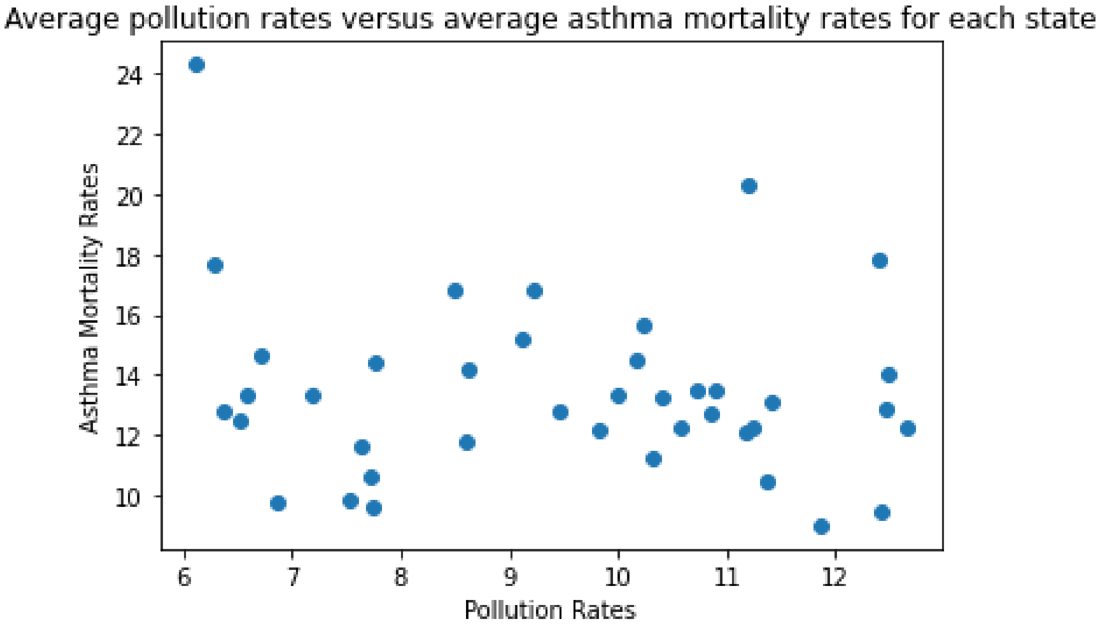
```
mortality_rates = new_asthma[['LocationID', 'DataValue']]
mortality_rates = mortality_rates.groupby('LocationID').mean()
```

[29]:

```
merged = pd.merge(mortality_rates, pollution, how = 'inner', left_on =␣
 ↪'LocationID', right_on = 'statefips')
```

[30]:

```
plt.scatter(merged['ds_pm_pred'], merged['DataValue'])
plt.xlabel("Pollution Rates")
ax = plt.ylabel("Asthma Mortality Rates")
plt.title("Average pollution rates versus average asthma mortality rates for␣
 ↪each state");
```



```
[31]: np.corrcoef(merged['ds_pm_pred'], merged['DataValue'])
```

```
[31]: array([[ 1.        , -0.15138083],
             [-0.15138083,  1.        ]])
```

Written analysis

Describe any trends you observe, and any relationships you may want to follow up on. We notice that there is a weak negative correlation between pollution rates and asthma mortality rates. Though this result is surprising, it motivates us to follow up on whether confounding variables exist that may influence both pollution and asthma mortality rates.

Describe any data cleaning steps you took. How will these decisions impact your model and inferences? We looked at row that had a non-null state ID. Including these in our model would introduce outliers and cause us to make incorrect inferences.

Explain how your visualizations are relevant to your research questions: do they motivate the question you're asking? Do they suggest a potential answer? This visualization motivates the question we are asking because it introduces the relationship between pollution rates and mortality rates. However, it does not suggest a potential answer because the result is firstly, not what we expected. Secondly, it is not a strong correlation, so it will not allow us to draw reasonable conclusions. However, this prompts us to think about confounding variables such as access to healthcare and median income of region.

```
[ ]:
```