bayes

April 19, 2024

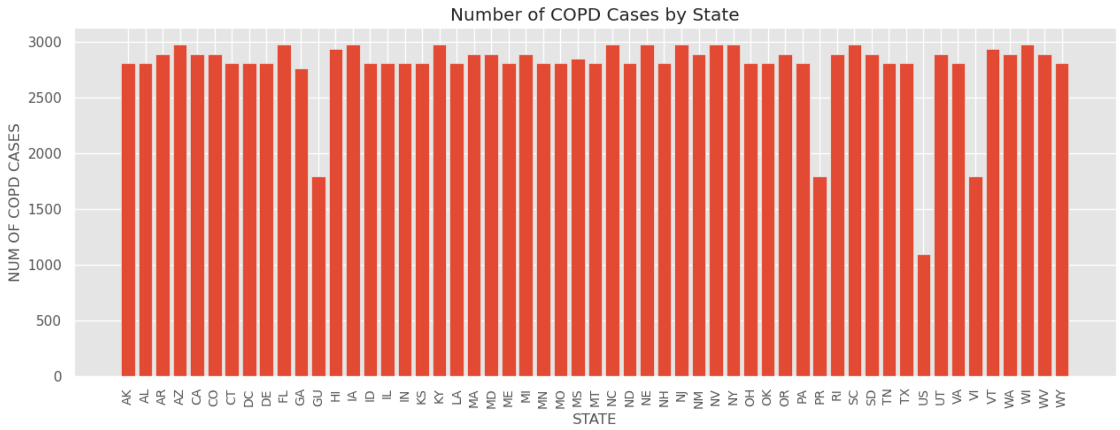
[1]:
```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import statsmodels.api as sm
import seaborn as sns
import itertools
from ipywidgets import interact, interactive
import re
import hashlib

sns.set(style="dark")
plt.style.use("ggplot")
%matplotlib inline
```

[38]:
```
copd = pd.read_csv("chronic_obstructive_pulmonary_disease.csv")
state_populations = pd.read_csv("pop.csv")
```

[39]:
```
copd_bystate = copd[['LocationAbbr','Topic']].groupby('LocationAbbr').count().
 ↪reset_index()
copd_bystate = copd_bystate.rename(columns={"LocationAbbr": "STATE", "Topic":
 ↪"NUM OF COPD CASES"})
```

[40]:
```
plt.bar(copd_bystate['STATE'], copd_bystate['NUM OF COPD CASES'])
locs, labels = plt.xticks()
plt.xticks(rotation=90, ha='center')
plt.title('Number of COPD Cases by State',y=1)
plt.xlabel('STATE')
plt.ylabel('NUM OF COPD CASES')
#plt.legend()
plt.tick_params(axis='x', which='major', labelsize=10)
plt.tight_layout(rect=[0, 0, 2, 1])
plt.subplots_adjust(bottom=0.1)
plt.show()
```



[41]:

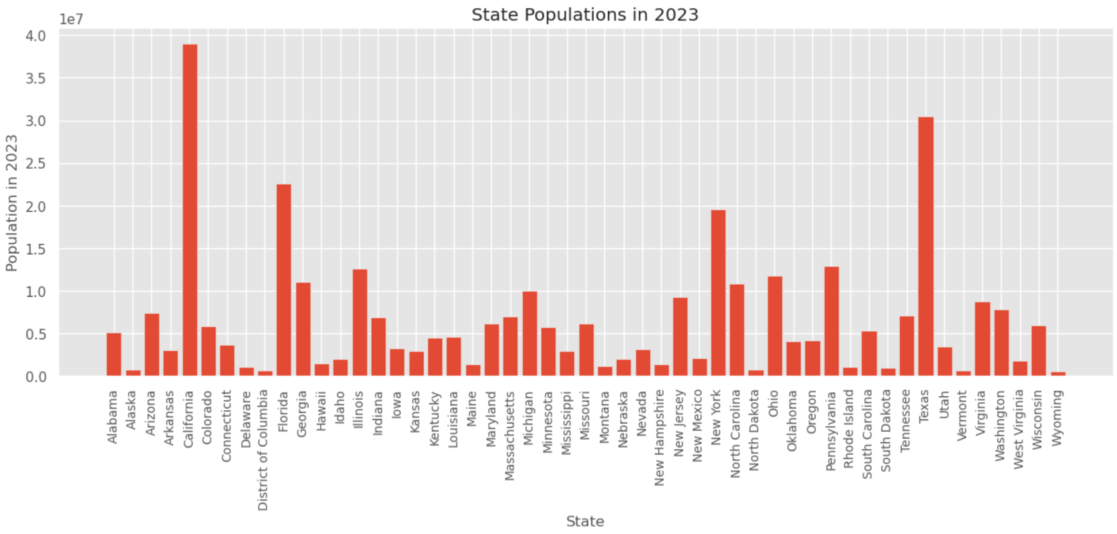
```
#rename state population columns and remove unnecessary rows
state_populations.columns = ["Geographic Area", "April 1 2020", "2020", "2021",␣
  ↪"2022", "2023"]
state_populations = state_populations.iloc[8:-8].reset_index()
state_populations = state_populations[["Geographic Area", "2023"]]
```

[42]:

```
#cleaned up the pop.csv dataset with regex to later merge by state name
state_populations['Geographic Area'] = state_populations['Geographic Area'].str.
  ↪extract(pat = '([\w ]+)')
state_populations['2023'] = state_populations['2023'].
  ↪replace(',','',regex=True).tolist()
state_populations['2023'] = pd.to_numeric(state_populations['2023'])
state_populations = state_populations.rename(columns={"Geographic Area":␣
  ↪"State"})
```

[44]:

```
plt.bar(state_populations['State'], state_populations['2023'])
locs, labels = plt.xticks()
plt.xticks(rotation=90, ha='center')
plt.title('State Populations in 2023',y=1)
plt.xlabel('State')
plt.ylabel('Population in 2023')
#plt.legend()
plt.tick_params(axis='x', which='major', labelsize=10)
plt.tight_layout(rect=[0, 0, 2, 1])
plt.subplots_adjust(bottom=0.1)
plt.show()
```



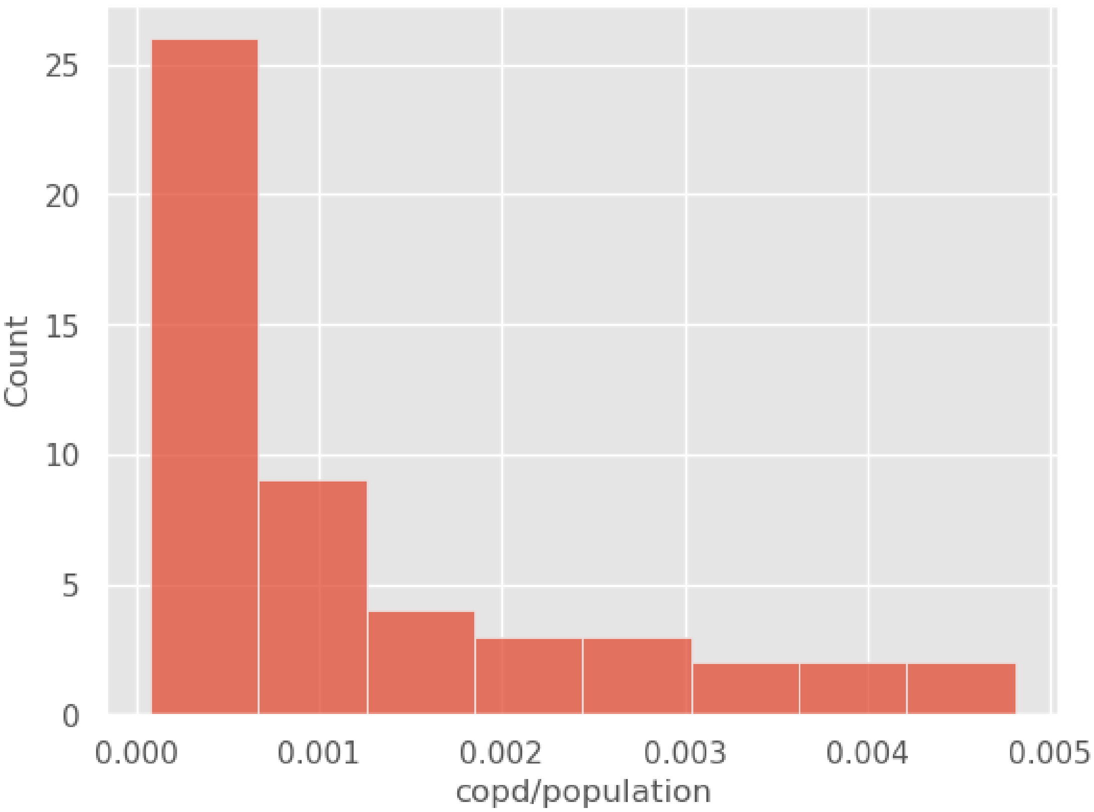
```
[45]: #imported csv file matching state names with abbreviations
      states = pd.read_csv("states.csv")
```

```
[46]: #got csv file on state names with abbreviations to better merge other datasets
      states = states.rename(columns={"Abbreviation": "STATE"})
      combined_df = pd.merge(states, copd_bystate, on="STATE", how="inner")
      combined_df = pd.merge(combined_df, state_populations, on='State', how='inner')
      combined_df.head()
```

```
[46]:         State STATE  NUM OF COPD CASES      2023
      0     Alabama    AL               2808   5108468
      1      Alaska    AK               2808    733406
      2     Arizona    AZ               2976   7431344
      3    Arkansas    AR               2892   3067732
      4  California    CA               2892  38965193
```

```
[47]: #histogram of copd/population to use for empirical bayes
      combined_df['copd/population'] = combined_df['NUM OF COPD CASES'] /␣
       ↪combined_df['2023']
      sns.histplot(combined_df, x='copd/population')
```

```
[47]: <Axes: xlabel='copd/population', ylabel='Count'>
```



Describe any trends you observe, and any relationships you may want to follow up on.

In this visualization, we observe that many of the states have exactly 2808 rows of data, which leads us to want to follow up on why we see that number of rows for each state. We also want to follow up on whether each row shown is 1 case of COPD or 1 person diagnosed with COPD. After figuring out how to differentiate the data so that we are able to see number of COPD cases per state accurately, we want to later compare that number to the population to better understand what the COPD rate per state is. Despite many states being capped at 2808, there are other states that have drastically fewer rows, such as Guam (GU), Puerto Rico (PR), and the Virgin Islands (VI), which are all US territories. We would want to compare how these territories consistently show less range of ‘cases’ than the other US states.

Explain how your visualizations should be relevant to your research questions: either by motivating the question, or suggesting a potential answer. You must explain why they are relevant.

These visualizations are relevant to our research question: Can we fit a Bayesian hierarchical model to the distributions of Chronic Obstructive Pulmonary Disease (COPD) by state? (Bayesian hierarchical modeling) because we want to create a prior based on state COPD count/state population. The first quantitative visualization is to first help us see how each state compares in individual COPD cases. The second visualization shows how each state compares on the population level. If we assume each row in the COPD dataset represents an individual and divide that by the state populations, then we will can use Empirical Bayes to find parameters of a Beta distribution that make the histogram above as likely as possible. However, we currently are still unsure what each row of the COPD dataset exactly means, so we need further research to be able to get the histogram we want.

```
[ ]:
```