US Healthcare Overview

Devin Nicholson, Linsey Stokes, Tyler Koizumi

Project Deliverables

- The group was assigned the task of exploring datasets related to healthcare
  - One dataset was to be extracted from the US Census
  - 6 datasets were explored
    - 2 from Census
- Project Management Plan
- ETL Documentation/Report
- Exploratory questions and visualizations answering those questions

Exploratory Questions

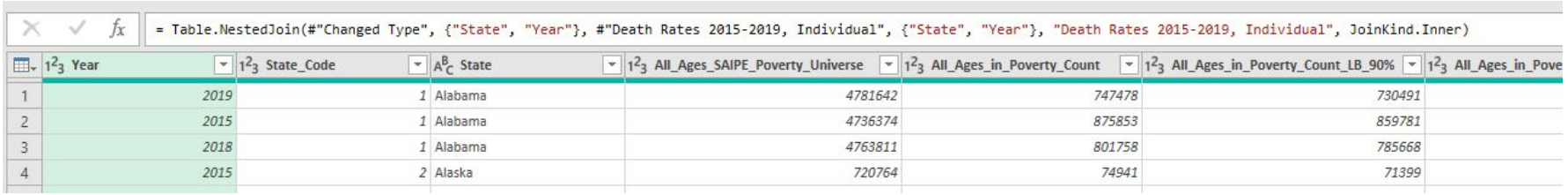
1. How many healthcare professionals are in each state?
2. Can we predict disease from its symptoms?
3. Is there a correlation between the amount of people uninsured and the amount hospitals by state?
4. Is there a correlation between poverty and insured people?
5. What states have the highest/lowest number of people uninsured/insured?
6. Find any correlation between poverty, number of hospitals, and mortality rates by state
7. Can we predict mortality rate from poverty status, insurance, and hospital availability

General

General Processes

This will cover the general ETL process for datasets/visualizations not regarding machine learning/predictions

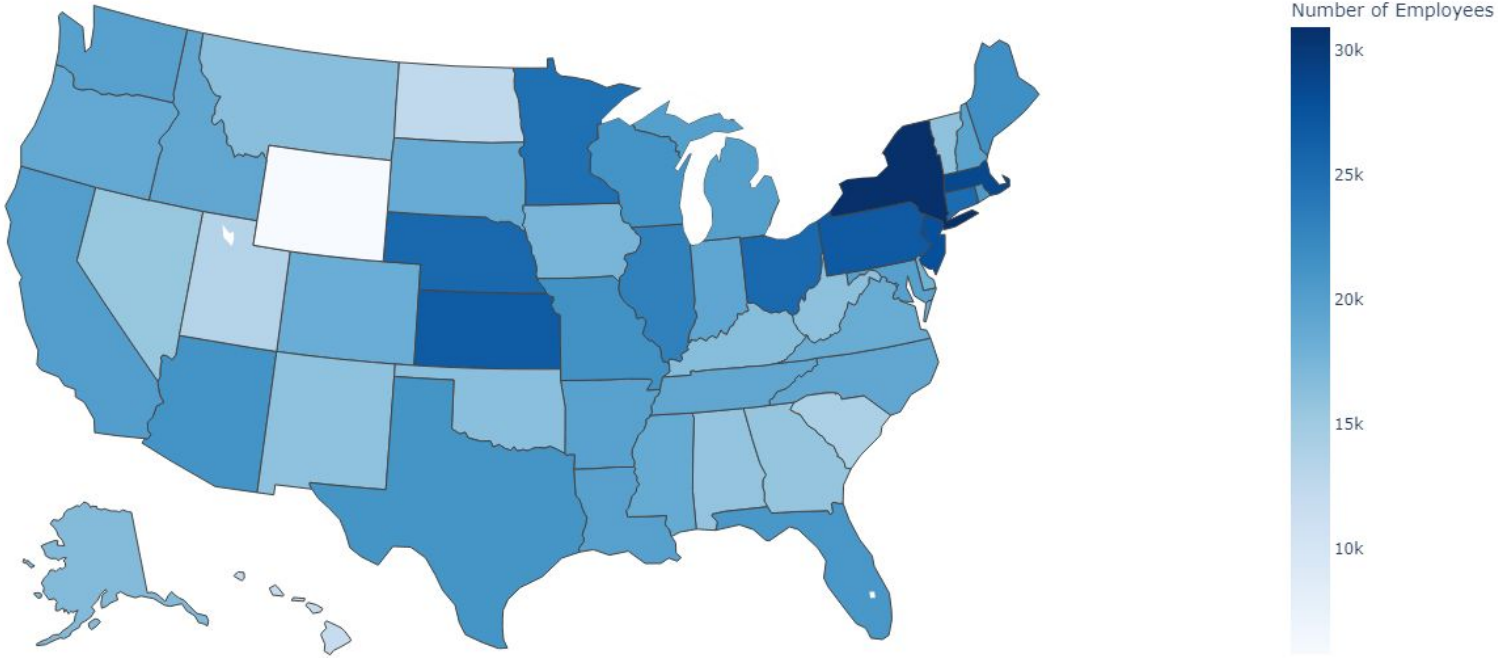
1. Dataset was retrieved as CSV file
2. Loaded into Excel Power Query
3. Data was cleaned for errors
    a. Values were removed
4. Data only pertaining to states were kept
5. The datasets pertaining to poverty, insurance, death rates, and hospitals were merged through state name using inner join function
6. If dataset contained multiple years, they were merged and averaged into one
    a. A copy was made beforehand so that visualizations could be made using unmerged datasets
    b. Done via creating pivot tables and saving data as CSV
7. Datasets were saved as CSV for easy python/pandas usage

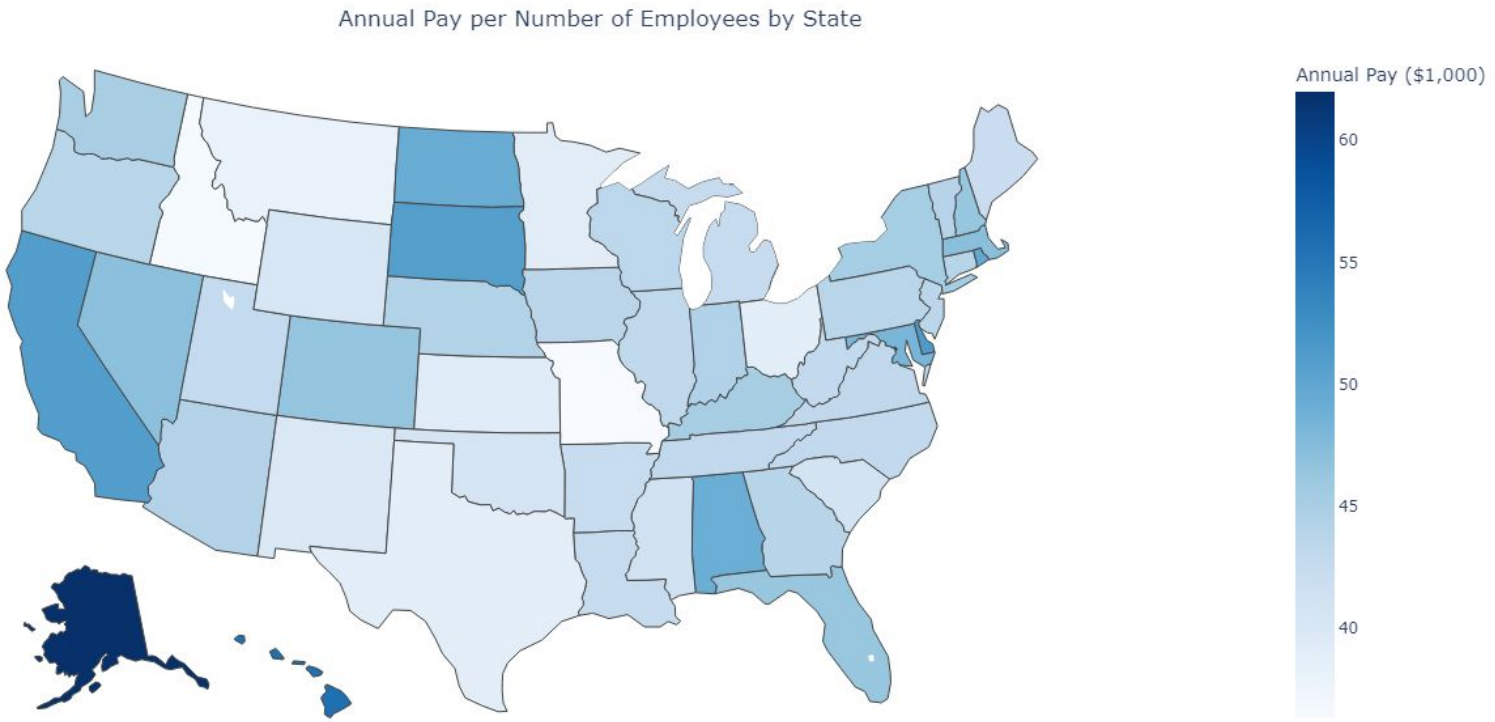


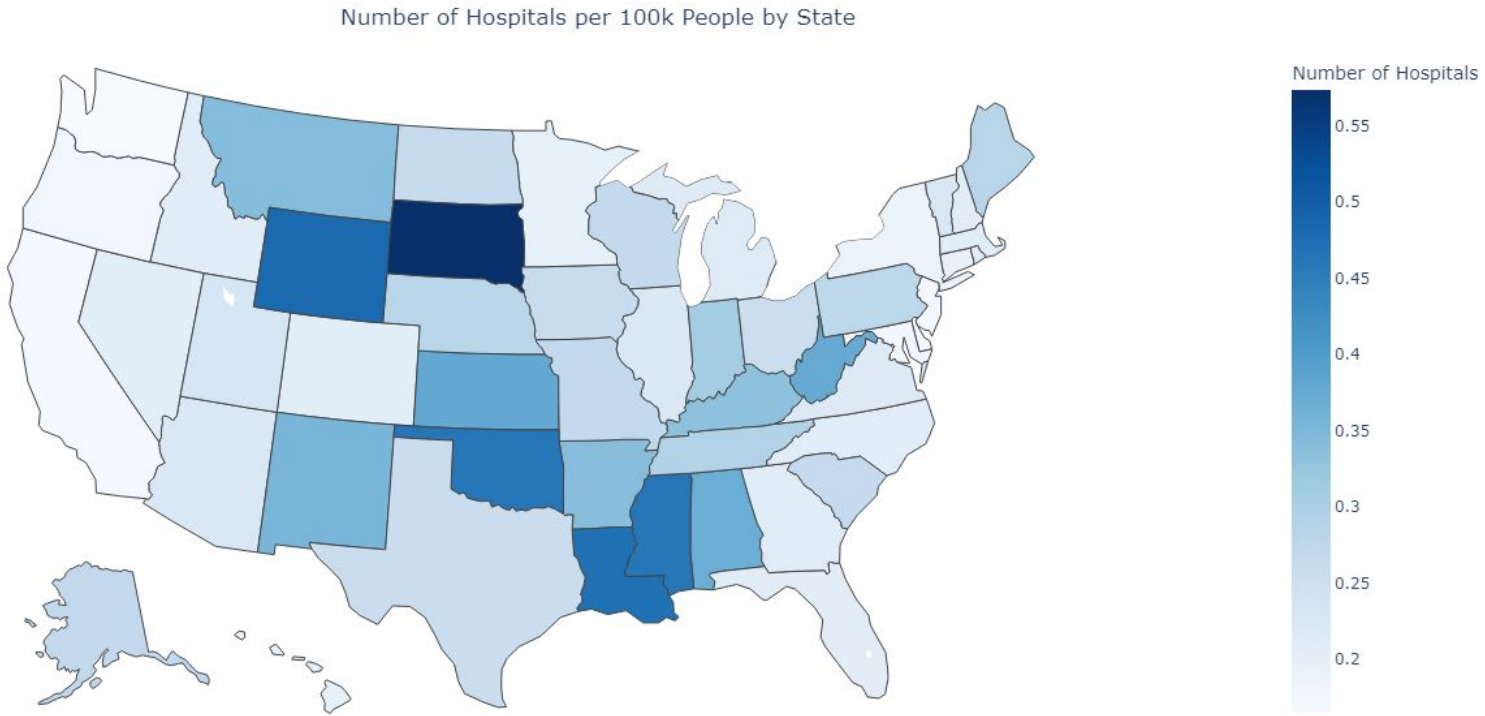
```
= Table.NestedJoin(#"Changed Type", {"State", "Year"}, #"Death Rates 2015-2019, Individual", {"State", "Year"}, "Death Rates 2015-2019, Individual", JoinKind.Inner)
```

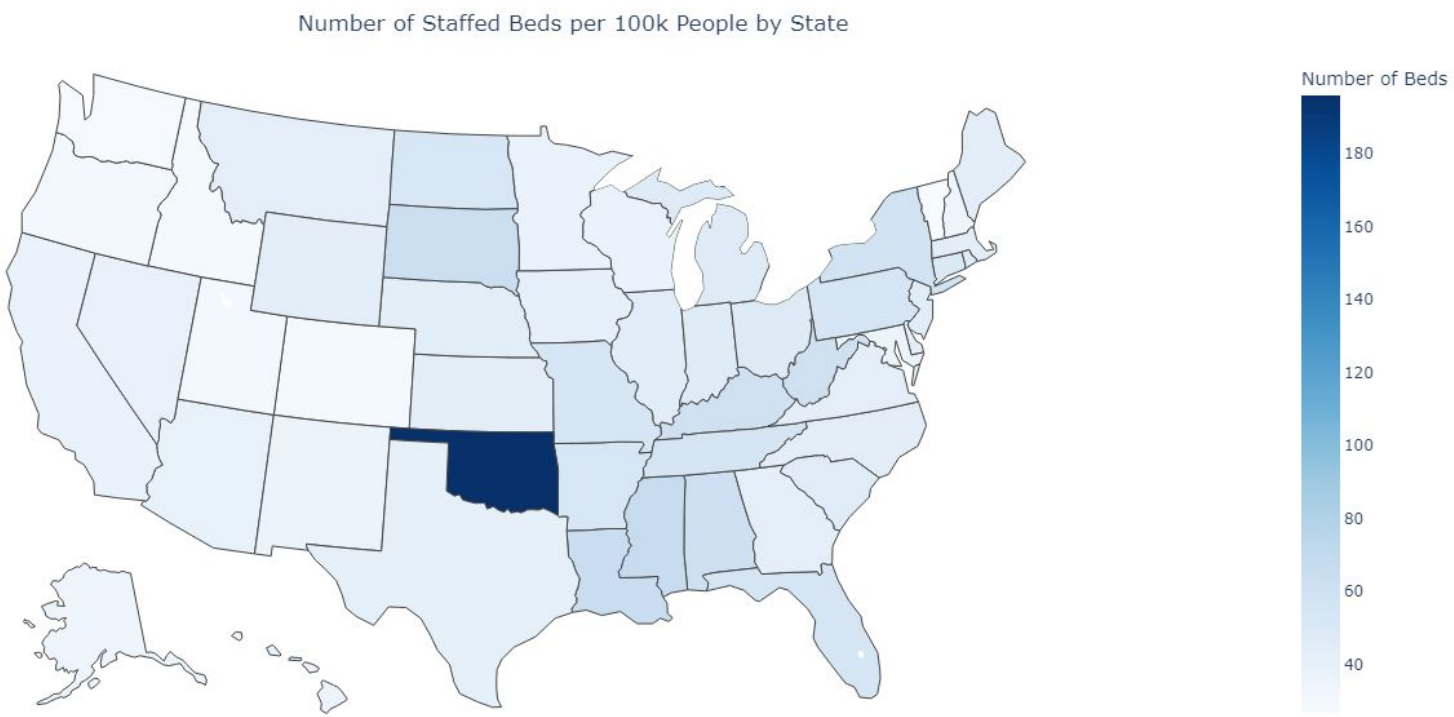
	Year	State_Code	State	All_Ages_SAIPE_Poverty_Universe	All_Ages_in_Poverty_Count	All_Ages_in_Poverty_Count_LB_90%	All_Ages_in_Pove
1	2019	1	Alabama	4781642	747478	730491	
2	2015	1	Alabama	4736374	875853	859781	
3	2018	1	Alabama	4763811	801758	785668	
4	2015	2	Alaska	720764	74941	71399	

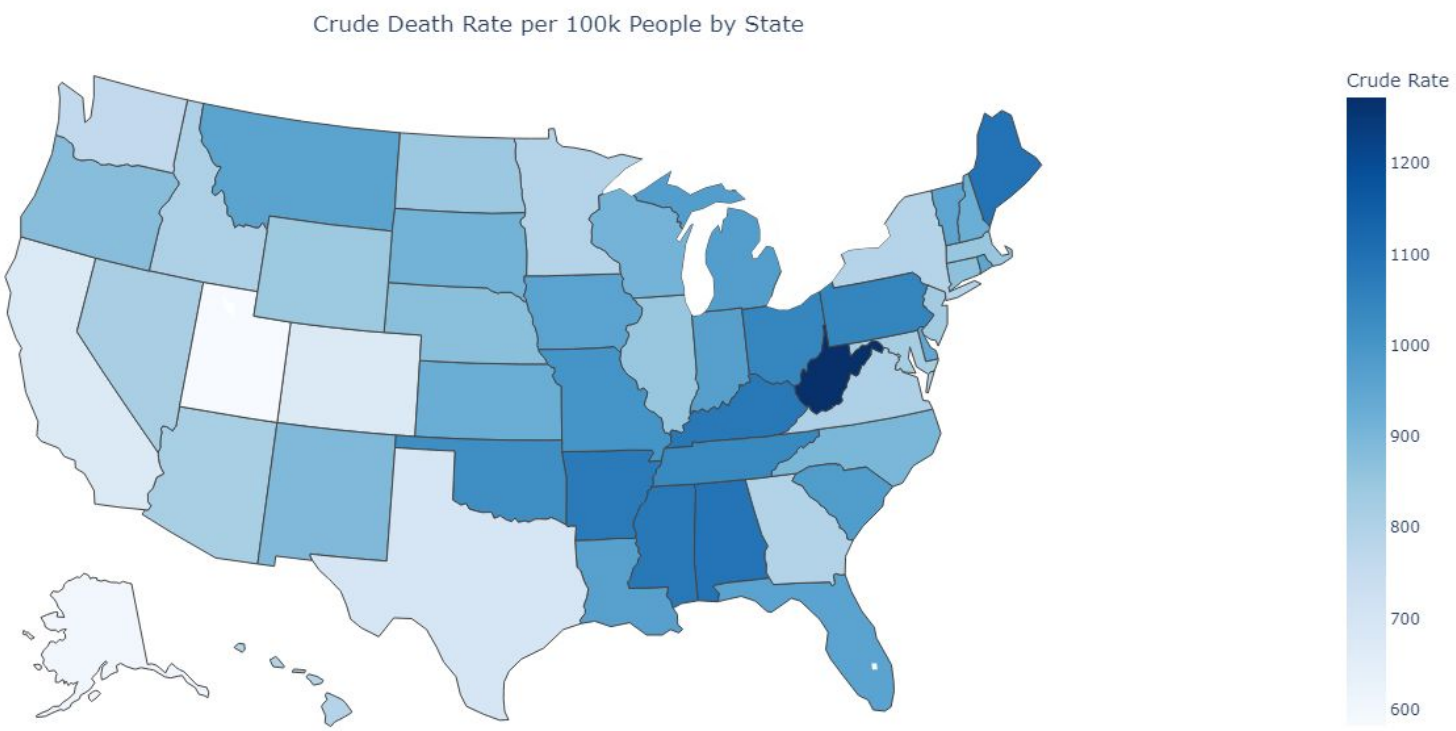
Number of Healthcare Employees per 100k People by State

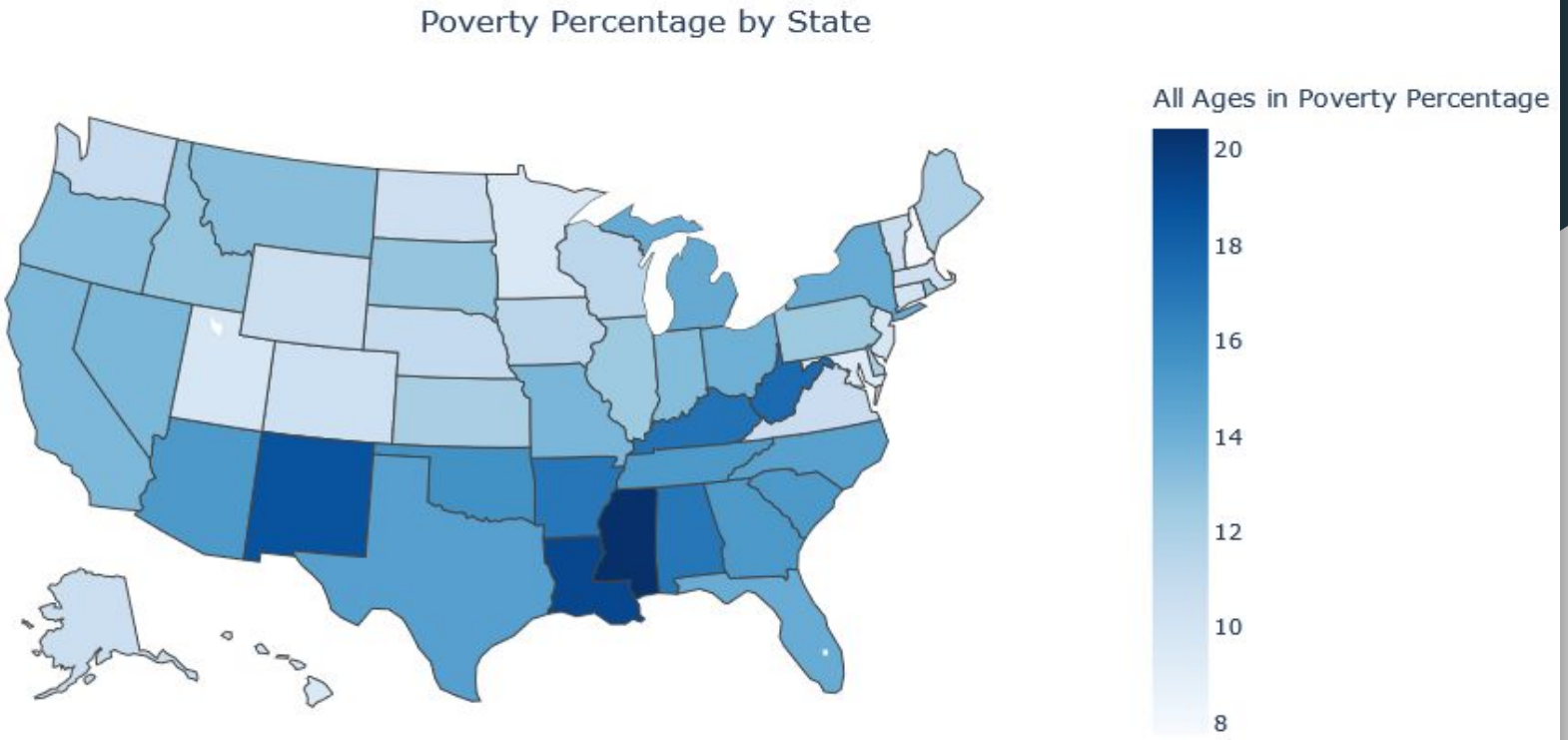



Annual Pay per Number of Employees by State
Annual Pay ($1,000)
60
55
50
45
40

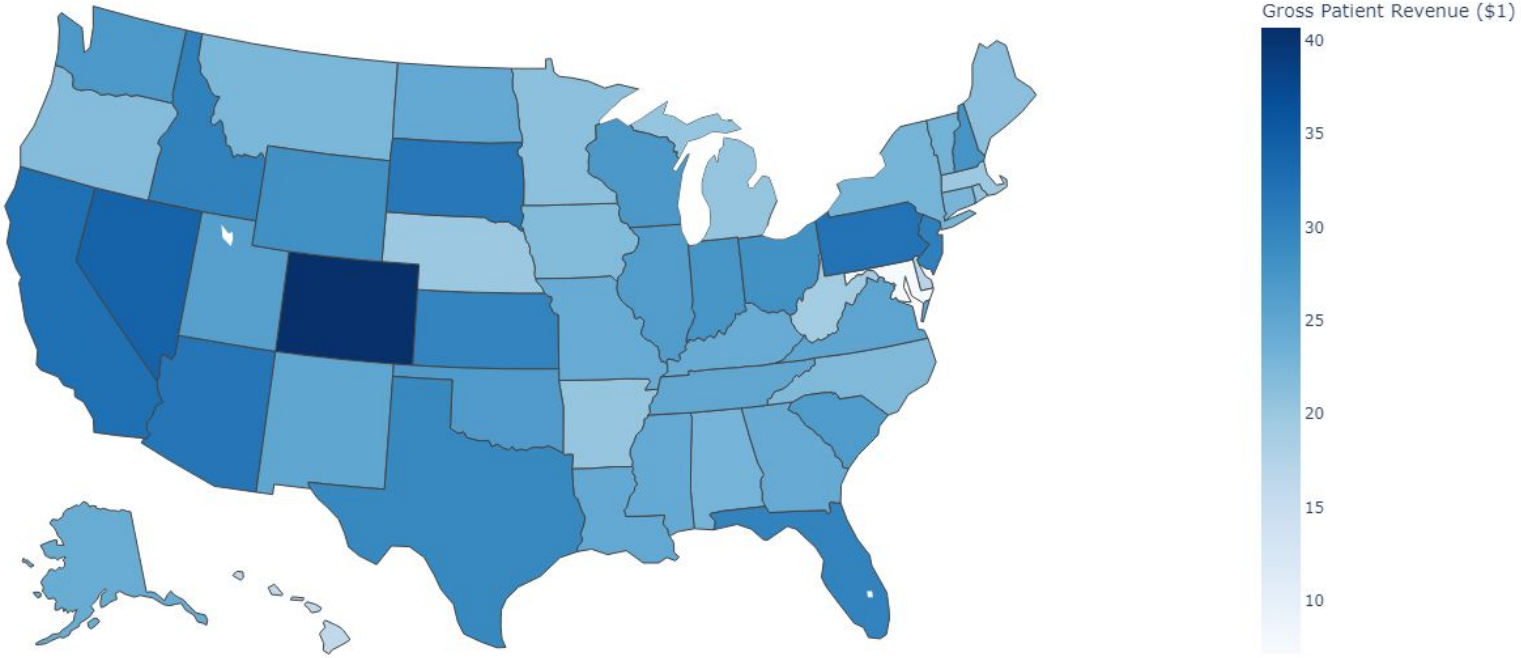

Number of Hospitals per 100k People by State
Number of Hospitals
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2


Number of Staffed Beds per 100k People by State
Number of Beds
180
160
140
120
100
80
60
40


Crude Death Rate per 100k People by State
Crude Rate
1200
1100
1000
900
800
700
600


Poverty Percentage by State
All Ages in Poverty Percentage
20
18
16
14
12
10
8

Gross Patient Revenue per Patient Day by State



Disease Prediction

Process

```
# Defining the decision tree algorithm
dtree = DecisionTreeClassifier()
dtree.fit(X_train,y_train)
print('Decision Tree Classifier Created')
# Predicting the values of test data
y_pred = dtree.predict(X_test)
```

1. Read Data
2. Melt Symptom columns (17) into columns for each symptom (132)
3. Set target column and split train and testing values
4. Define DecisionTreeClassifier and fit training variables
5. Predict accuracy of X_test
6. Print visualizations such as confusion matrix and Tree map

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4
0	Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
1	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN
2	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	NaN
3	Fungal infection	itching	skin_rash	dischromic _patches	NaN
4	Fungal infection	itching	skin_rash	nodal_skin_eruptions	NaN

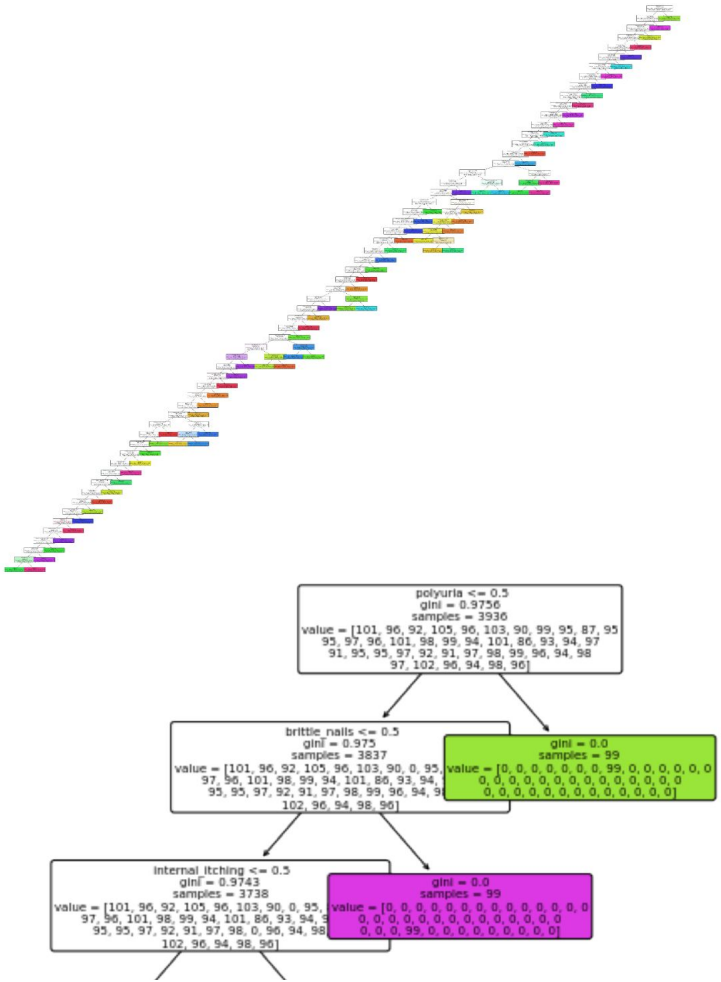
← Data Before Melt

Data After Melt -->

value index	label	abdominal_pain	abnormal_menstruation	acidity	acute_liver_failure
0	Fungal infection	0.0	0.0	0.0	0.0
1	Fungal infection	0.0	0.0	0.0	0.0
2	Fungal infection	0.0	0.0	0.0	0.0
3	Fungal infection	0.0	0.0	0.0	0.0
4	Fungal infection	0.0	0.0	0.0	0.0

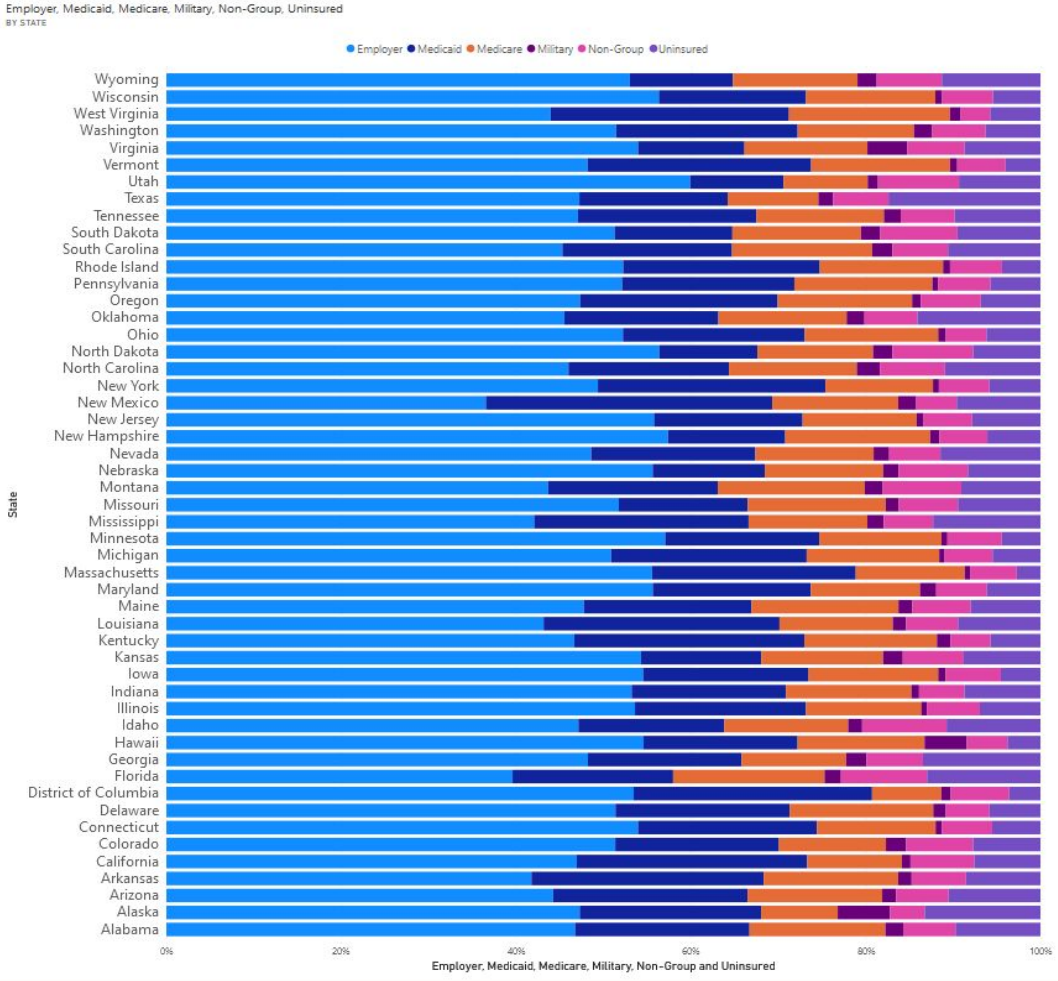
Outcome

- The accuracy of the predictions was found to be 100%
- The decision tree generated is to the right
- Each decision node shows the possible symptoms and branches into two parts
    - A disease that can be predicted from the symptom
    - Another symptom that can be used in junction with the preceding symptom(s) to predict a disease
- The tree goes down until all the possible diseases in the dataset are exhausted



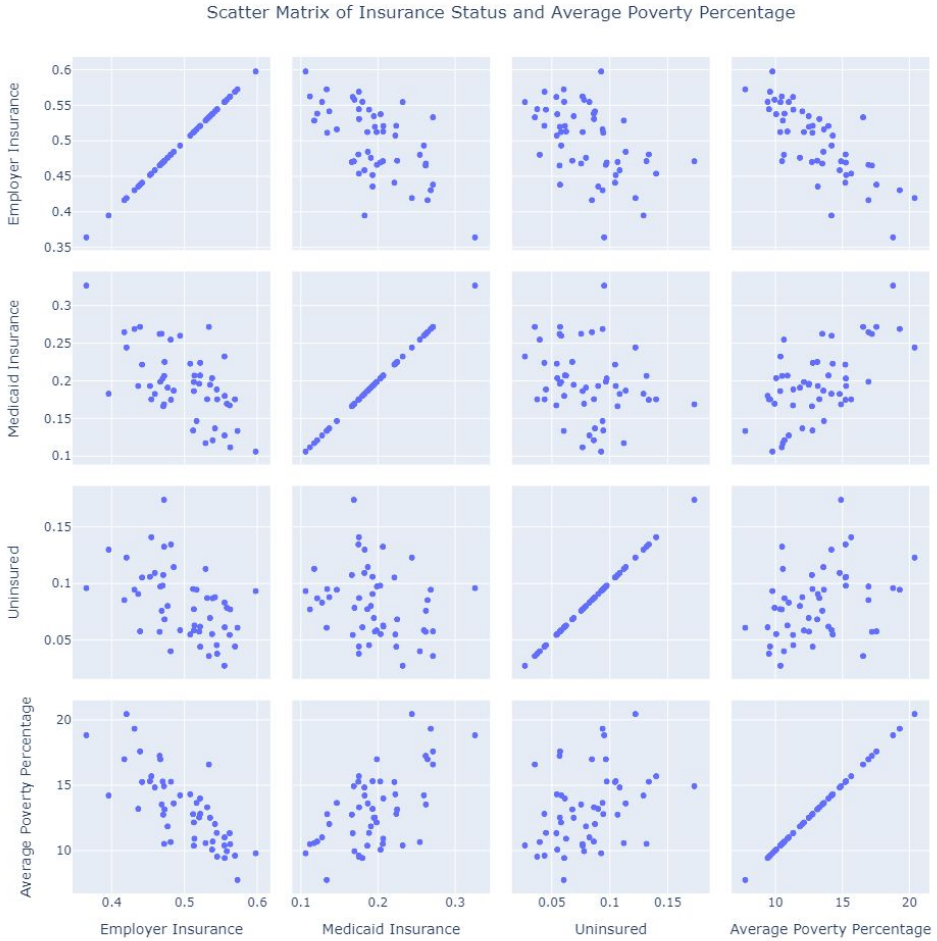
Insurance

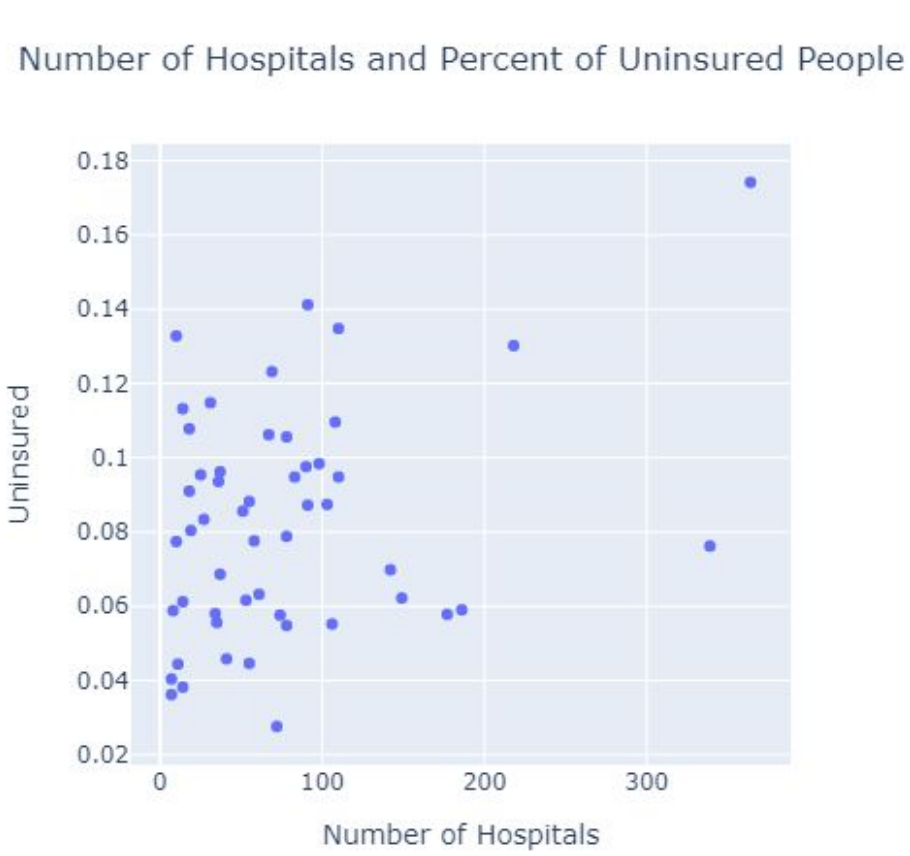
- Breakdown of types of insurance broken down by state, sorted in alphabetical order
    - Categories of insurance are as follows:
        - Employer
        - Non-Group
        - Medicaid
        - Medicare
        - Military
        - Noninsured
- Texas has the largest uninsured percentage
- Massachusetts has the smallest uninsured percentage



Strongest Indicators:

- Employer insurance has a strong negative correlation with the average poverty percentage (-.78)
- Medicaid insurance has a moderately positive correlation with the average poverty percentage (.62)
- Medicaid insurance has a moderately negative correlation with Employer insurance (-.61)
- Being Uninsured has a weak negative correlation with Employer insurance (-.49)




Number of Hospitals and Percent of Uninsured People
Uninsured
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
100
200
300
Number of Hospitals

Mortality

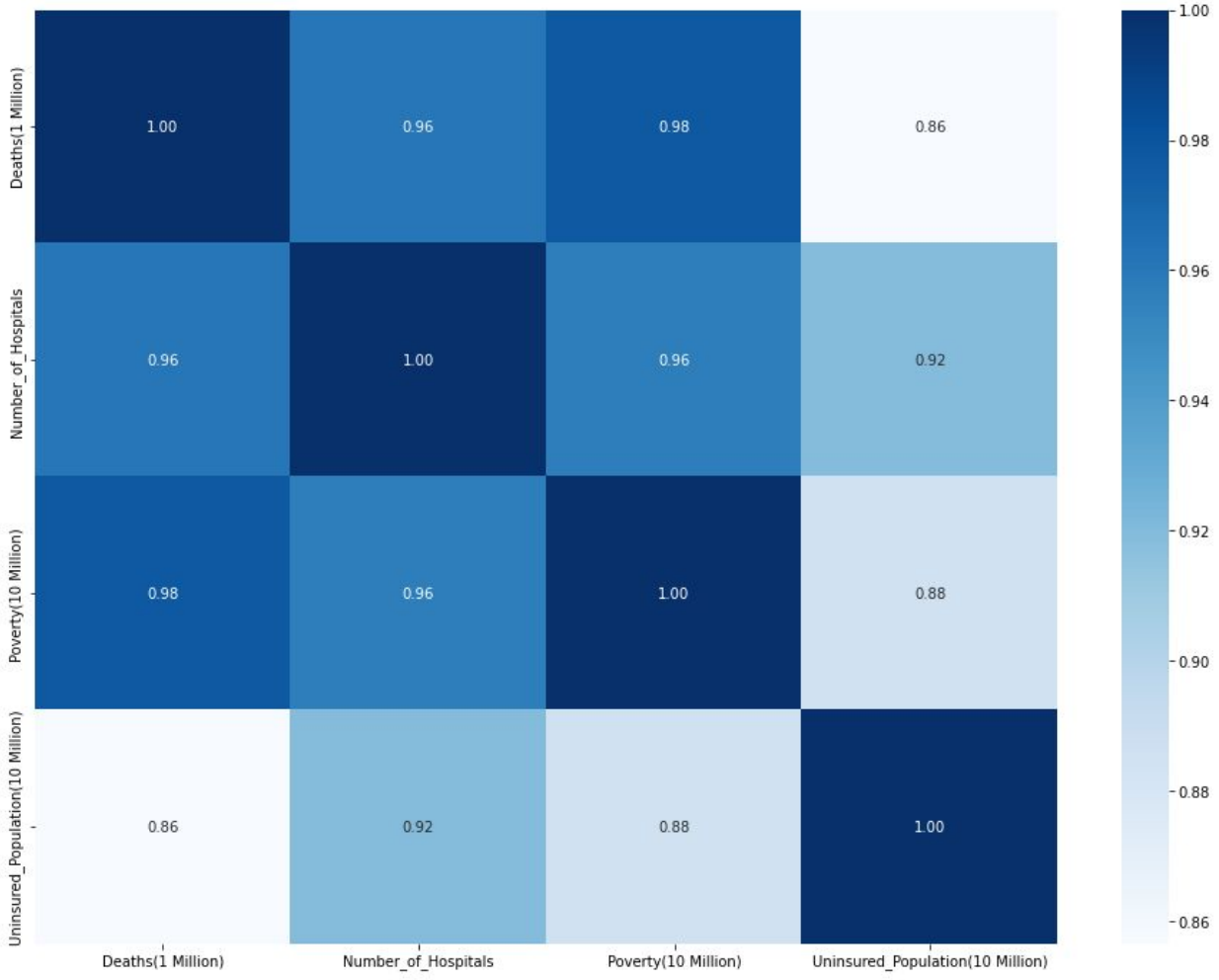
Process

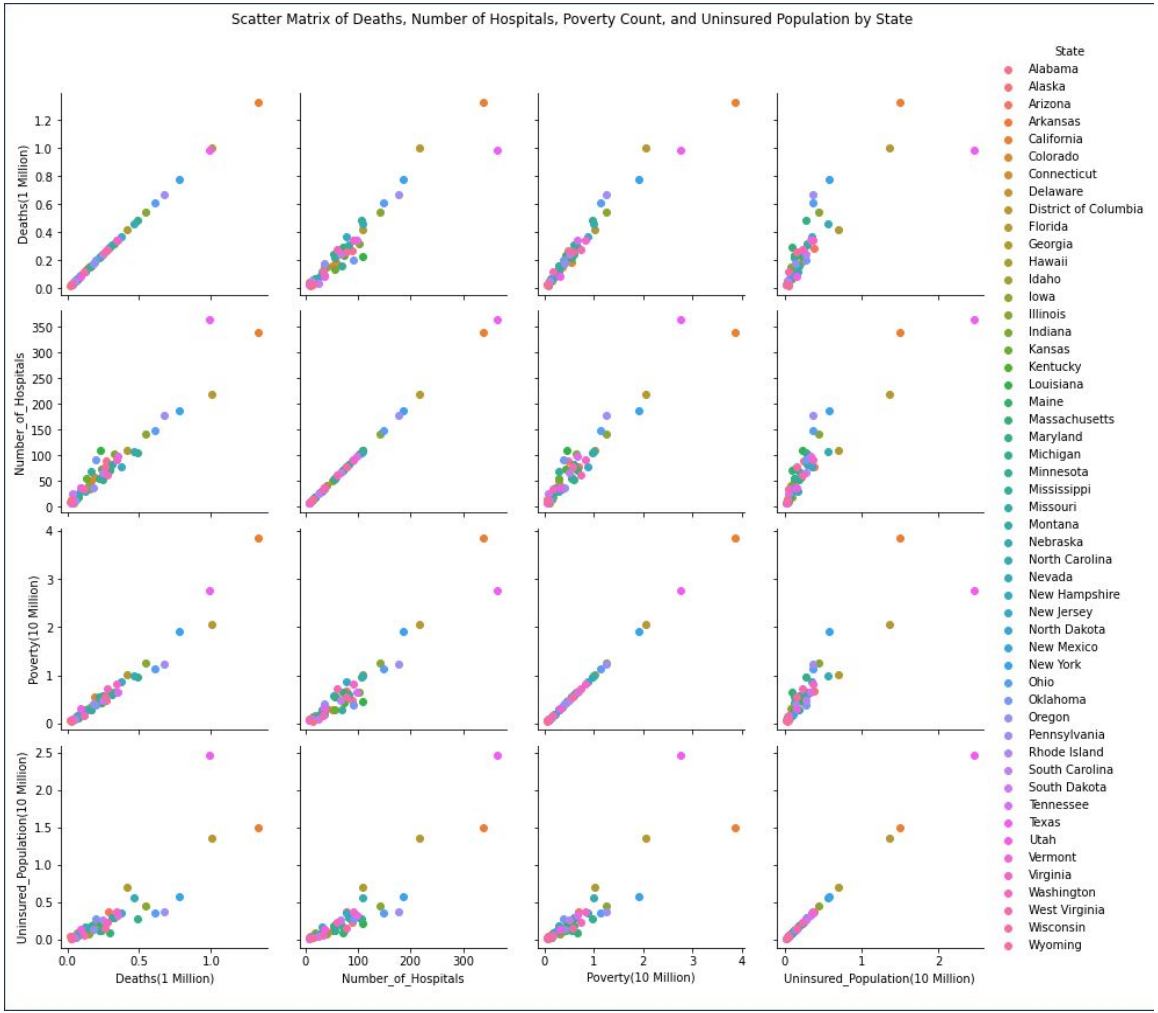
- Read in data
- Check correlation
- Drop columns
- Created training and Test data
- Built the model
- Tested the model

```
y = df_corr.pop('Deaths(1 Million)')
X = df_corr
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
```

```
df_corr.head()
✓ 0.1s
```

	State	Deaths	Number_of_Hospitals	Poverty	Uninsured_Population
0	Alabama	266073	90	4755140.2	2380276
1	Alaska	22287	10	719958.4	489913
2	Arizona	288220	78	6885729.4	3719841
3	Arkansas	161185	51	2911377.8	1284202
4	California	1328284	339	38651294.2	15011457


Deaths(1 Million)
Number_of_Hospitals
Poverty(10 Million)
Uninsured_Population(10 Million)
1.00
0.96
0.98
0.86
0.96
1.00
0.96
0.92
0.98
0.96
1.00
0.88
0.86
0.92
0.88
1.00
Deaths(1 Million)
Number_of_Hospitals
Poverty(10 Million)
Uninsured_Population(10 Million)
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86


Scatter Matrix of Deaths, Number of Hospitals, Poverty Count, and Uninsured Population by State
Deaths(1 Million)
Number_of_Hospitals
Poverty(10 Million)
Uninsured_Population(10 Million)
State
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Iowa
Illinois
Indiana
Kansas
Kentucky
Louisiana
Maine
Massachusetts
Maryland
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
North Carolina
Nevada
New Hampshire
New Jersey
North Dakota
New Mexico
New York
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

Outcome

```
X_string = ""
count = 0
for i, coef in enumerate(model.coef_):
    X_string += f"( {round(coef, 3)} X{i} )+\n"
equation = f'Equation is y = {model.intercept_} + {X_string}'
print(equation)
```

✓ 0.2s

```
Equation is y = 16672.74060862948 + ( 1638.483 X0 )+
( 0.027 X1 )+
( -0.011 X2 )+
```

```
model.score(X_test, y_test)
```

✓ 0.2s

```
0.9563103563671331
```

y_test	predictions	Difference
486960	421936.843	65023.157
288220	288221.588	-1.5884593
263633	278871.558	-15238.558
50441	68443.6593	-18002.659
93242	121158.805	-27916.805
90372	141663.508	-51291.508
62209	70074.8522	-7865.8522
24459	51194.3492	-26735.349
32027	49531.2188	-17504.219
83709	102122.106	-18413.106
161185	164272.072	-3087.072
121739	127570.816	-5831.8159
281499	286253.226	-4754.2256

Questions?

References

1. Small Area Income and Poverty Estimates (SAIPE) - US Census Bureau
2. Hospital Statistics by State
3. Disease Symptom Prediction
4. Underlying Cause of Death, 1999-2019 Request
5. Health Insurance Coverage of the Total Population
6. https://www.census.gov/data/developers/data-sets/abs.2019.html

Thank You