HEALTHCARE DISPARITIES

PRIMARY CARE ACCESS ANALYSIS

Jacqueline Clinesmith, Tony Ferri, and Jenny Johnson

MSU Data Analytics Boot Camp, 2022

Healthcare Disparities
Primary Care Access Analysis

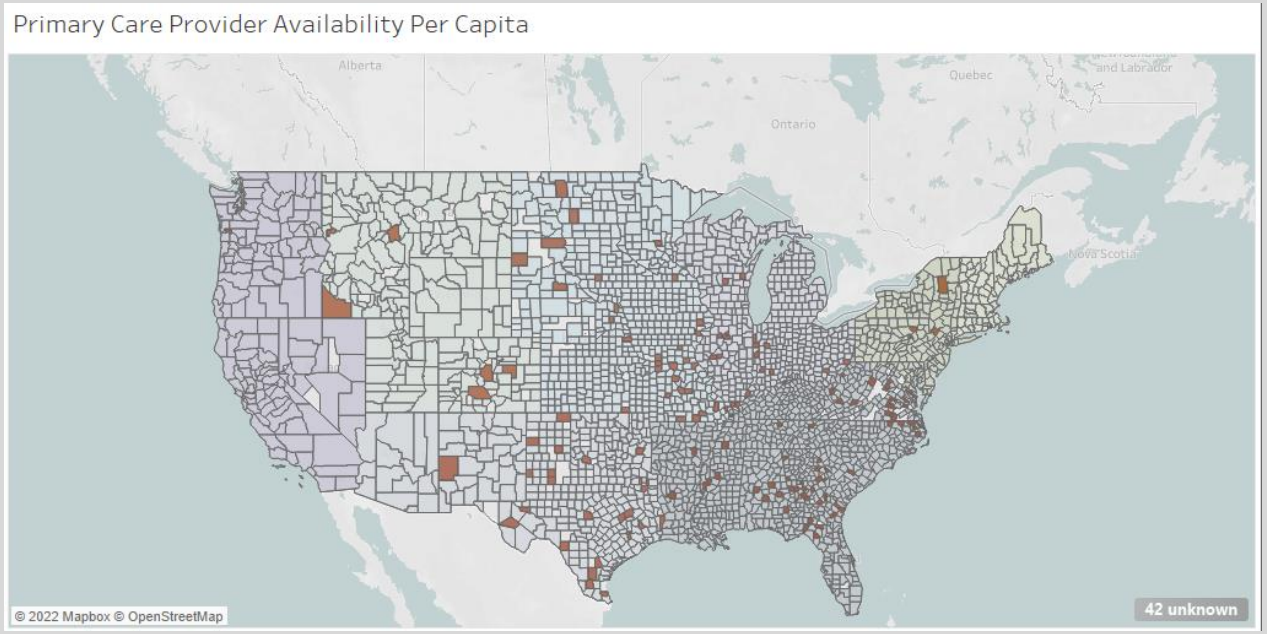
WHAT ARE WE TRYING TO DO?

DATA EXPLORATION

ANALYSIS RESULTS

WHAT ARE WE TRYING TO DO?

We're building a machine learning model to try and determine primary care physician availability



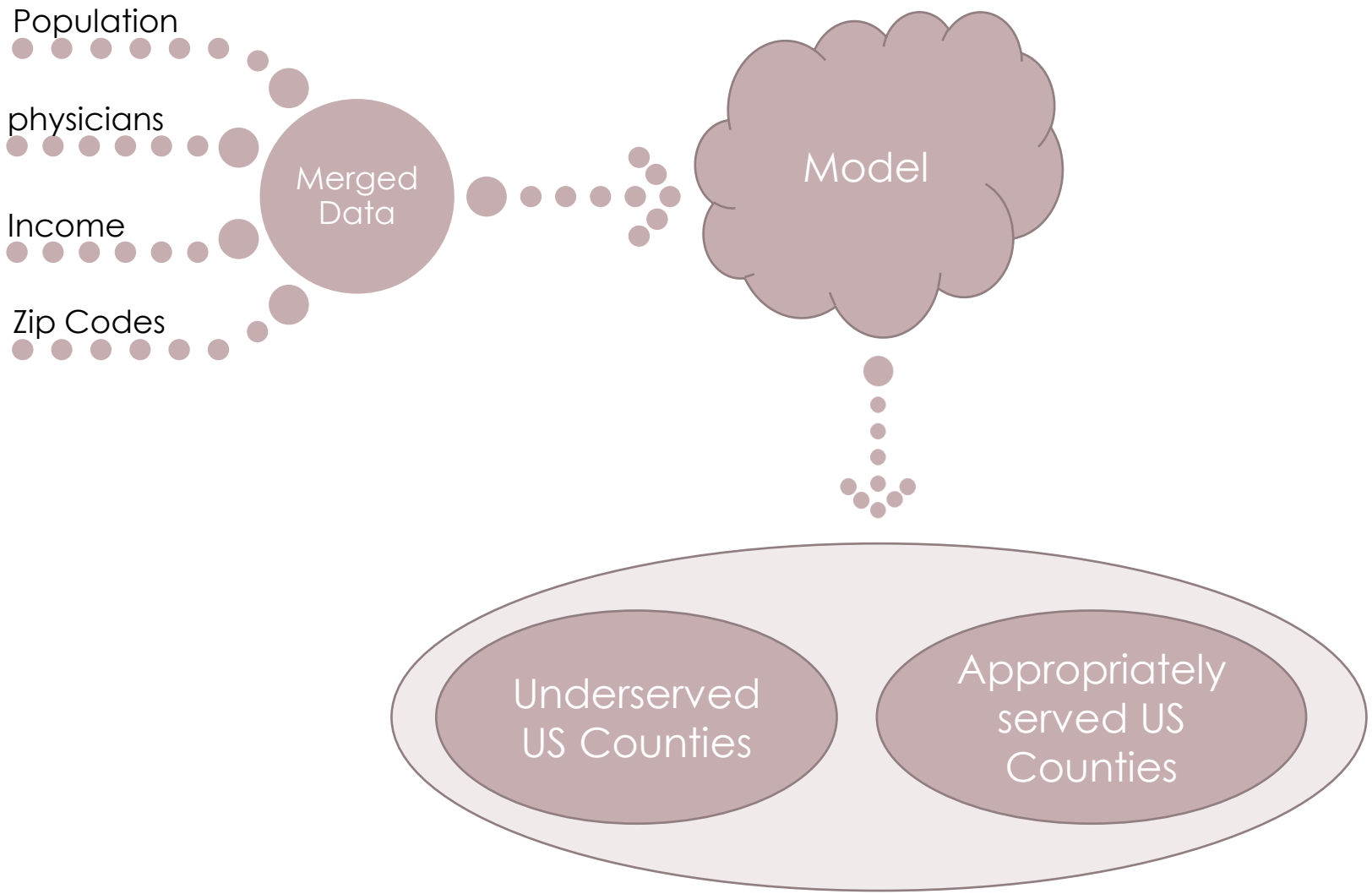
- We're hoping to determine:
  - Are there counties in the United States that will be underserved by primary care physicians?
  - Are there any other factors that impact availability, like income, region, or population?

Meaningful access to a primary care physicians can help reduce health care disparities.

Tools and Technologies Used

- Python
  - pandas
  - matplotlib
  - numpy
  - pathlib
  - collections
  - sklearn.metrics
- Tableau
- PostgreSQL
- Quick DBD

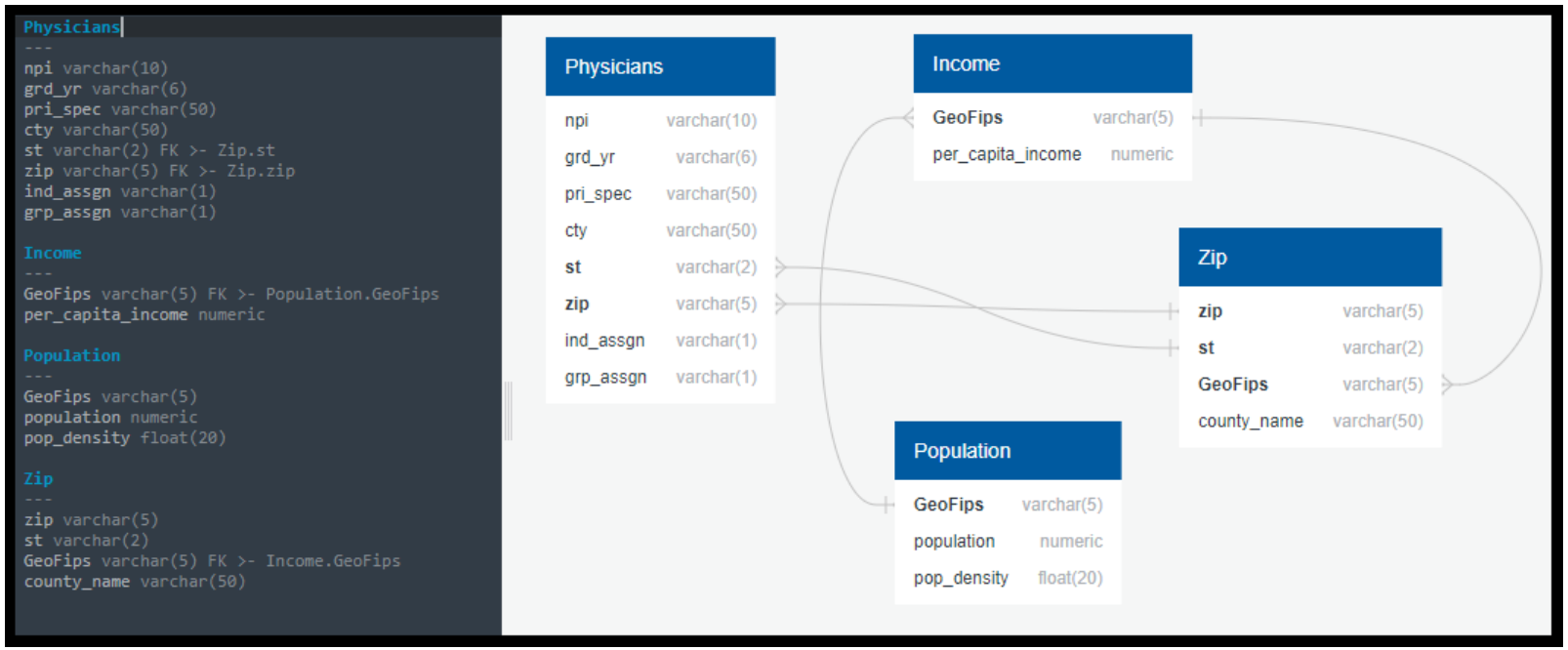
DATA EXPLORATION

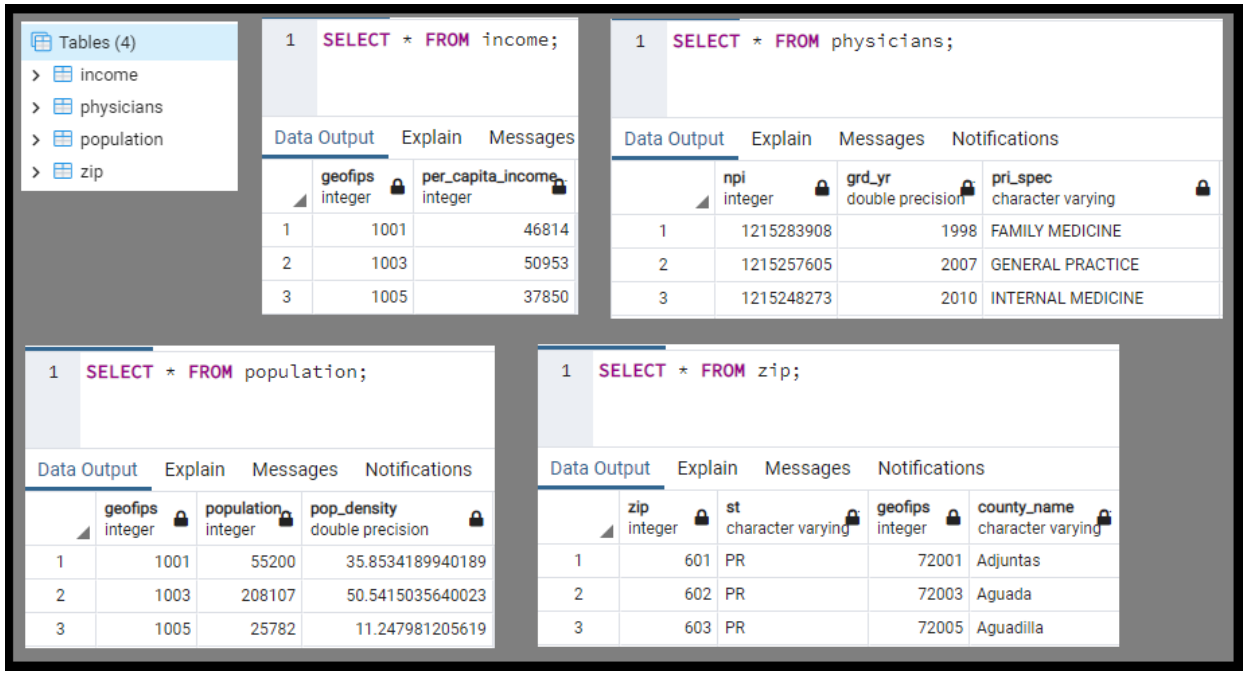


Data Exploration

- Data from official government sources was given preference due to accuracy and reliability
- Data preprocessing included:
  - Removing irrelevant information
  - Standardizing zip codes
  - Removing headers and footers
  - Standardizing column names
  - Transformed into bins
  - Calculating primary care physicians by county
  - Determine csv encoding1
- Defining key data:
  - Underserved county: A county with a Primary Care physician per capita that is less than or equal to 1 standard deviation from the mean. In this data set it was 0.488
  - Primary Care physician's primary specialty is listed as: family medicine, nurse practitioner, general practice, preventative medicine, emergency medicine, physician assistant, internal medicine, pediatric medicine, obstetrics / gynecology

Population provided by US Census Bureau
physicians provided by The Centers for Medicare and Medicaid Services
Income provided by US Bureau of Economic Analysis
Zip Codes provided by simplemaps





Database

- Cleaned data sources were loaded into a PostgreSQL database
- Cleaned data was used for Tableau and machine model

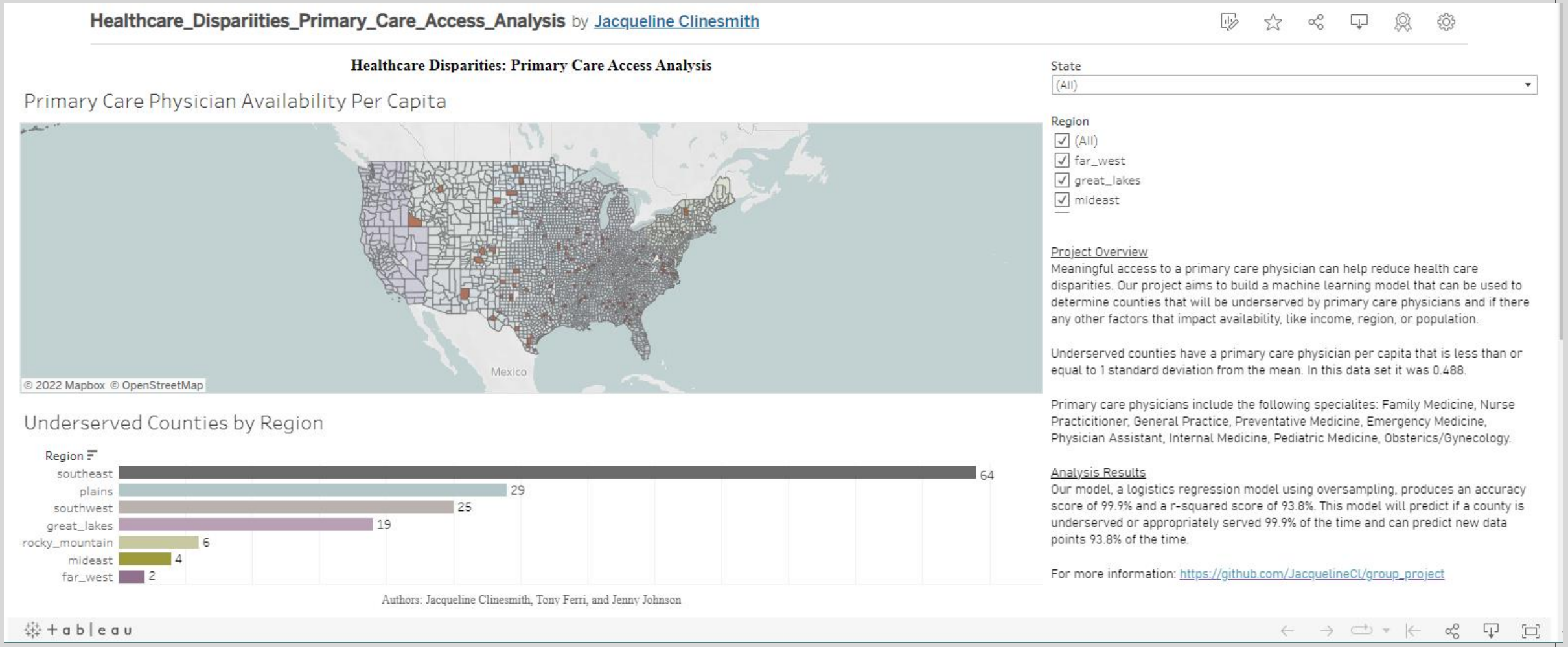
ANALYSIS RESULTS

- Out of 3,092 counties, 148 are underserved, or ~5% of all counties, while 2881 counties, or ~95% of counties are appropriately served.
- Due to the data being split 19:1, the model must include oversampling and the removal of extraneous data.
- Before the logistic regression model included oversampling, the accuracy score was 96.2%, however, the r-squared score was -.077%, meaning the model was an exceptionally poor fit and could not predict new data points with accuracy.
- Hyper tuning using SMOTE (Synthetic Minority Oversampling Technique) random oversampling and SMOTEENN (oversampling using SMOTE and cleaning using Edited Nearest Neighbor) using combination sampling does not improve the accuracy score or r-squared score of the model (both remain the same)
- Undersampling using random undersampling or cluster centroids undersampling lowers the accuracy score and drastically lowers the r-squared score, meaning the model is no longer a good fit and can no longer accurately predict new data. For this reason, undersampling the data should be avoided.

Analysis Results

- Logistical Regression
- Accuracy score: 99.9%
- R-squared score: 93.8%

A Tableau Dashboard was created to visualize the data and allow for easy filtering.



Factors the model found impactful

```
# List the features sorted in descending order by feature importance
feature_names = X.columns
sorted(zip(random_forest.feature_importances_, feature_names), reverse=True)
```

```
[(0.3579915653747901, 'pcp_count'),
 (0.17517299037918277, 'population'),
 (0.06466228298342844, 'GeoFips'),
 (0.05129256564212864, 'pop_density_lvl'),
 (0.012369037184989565, 'region_Southeast'),
 (0.007735871933347174, 'region_Rocky Mountain'),
 (0.007471946511600442, 'region_Plains'),
 (0.006049277187512302, 'region_Far West'),
 (0.005499866584241567, 'region_Southwest'),
 (0.005145980420641363, 'region_Mideast'),
 (0.004572959904752935, 'region_Great Lakes'),
 (0.0035399121199545646, 'county_Jefferson'),
 (0.0033221910336043666, 'county_Cass'),
 (0.003265422770929177, 'region_New England'),
 (0.003247423431408616, 'county_Calumet'),
 (0.00265458958228761, 'county_Harris'),
 (0.002597147769073972, 'county_Upshur'),
 (0.002569810669515998, 'county_Benton'),
 (0.002501874480003684, 'county_Macon'),
 (0.0023562743356298054, 'county_Hart'),
 (0.002311117184697795, 'county_Posey'),
 (0.002090548717674469, 'county_Greene'),
 (0.002017866945358578, 'county_Monroe'),
 (0.0020119236074613196, 'county_Lincoln'),
```

Questions

APPENDIX

Data Sources

Source	Information Used	Location
Income data from US Bureau of Economic Analysis	GeoFips and 2020 from bea_income_2020.csv	https://apps.bea.gov/iTable/iTable.cfm?reqid=70&step=30&isuri=1&major_area=4&area=xx&year=2020&tableid=20&category=720&area_type=4&year_end=-1&classification=non-industry&state=xx&statistic=3&yearbegin=-1&unit_of_measure=levels
Region data from US Bureau of Economic Analysis	Regions by State from website	https://www.bea.gov/news/2015/gross-domestic-product-state-advance-2014-and-revised-1997-2013/regional-maps
Population data from US Census Bureau	GEOID, B01001_001E, and B01001_calc_PopDensity from population_census.csv	https://covid19.census.gov/datasets/average-household-size-and-population-density-county/explore?location=15.251650%2C0.315550%2C3.67&showTable=true
Zip code data from simplemaps Basic Download	zip, state_id, county_fips, and county_name from uszips.csv	https://simplemaps.com/data/us-zips
physician data from The Centers for Medicare and Medicaid Services	NPI, grd_yr, pri_spec, city, st, zip, ind_assgn, and grp_assgn from physician_data.csv	https://data.cms.gov/physician-data/dataset/mj5m-pzi6