Medicare Charges

Capstone Project 1 - Springboard Data Science Career Track

Agenda

1 - Introduction

2 - Initial Analysis

3 - In-Depth Analysis

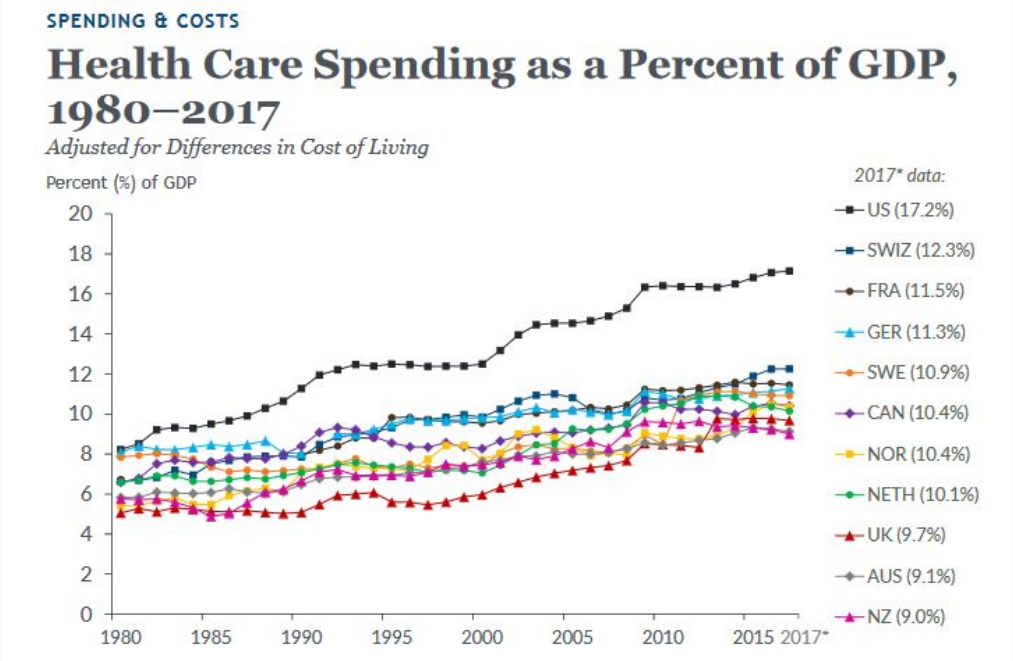
4 - Conclusions

Introduction

Background

"Despite spending more on health care, Americans had poor health outcomes, including shorter life expectancy and greater prevalence of chronic conditions."

The Commonwealth Fund



Project objectives

Analyze and visualize Medicare charges data and try to answer some of the following questions:

- What are the average and median costs for common medical procedures for Medicare patients across the country?
- Where outlier charges exist, where are they located?
- Are there common characteristics between the providers and/or geographic locations where such outliers exist?
- Are demographic factors also a factor to take into consideration in the areas where outlier charges exist?

Datasets used

Medicare Physician charges for 2016 (CSV file, size: 2.14 GB)

National Provider Identification information for 2014 (CSV file, size: 563 MB)

US County Demographic Data for 2016 (CSV file, size: 176 KB)

Initial Analysis

First steps

The initial analysis of the dataset focused on gathering an overall idea on the size and scope of the dataset.

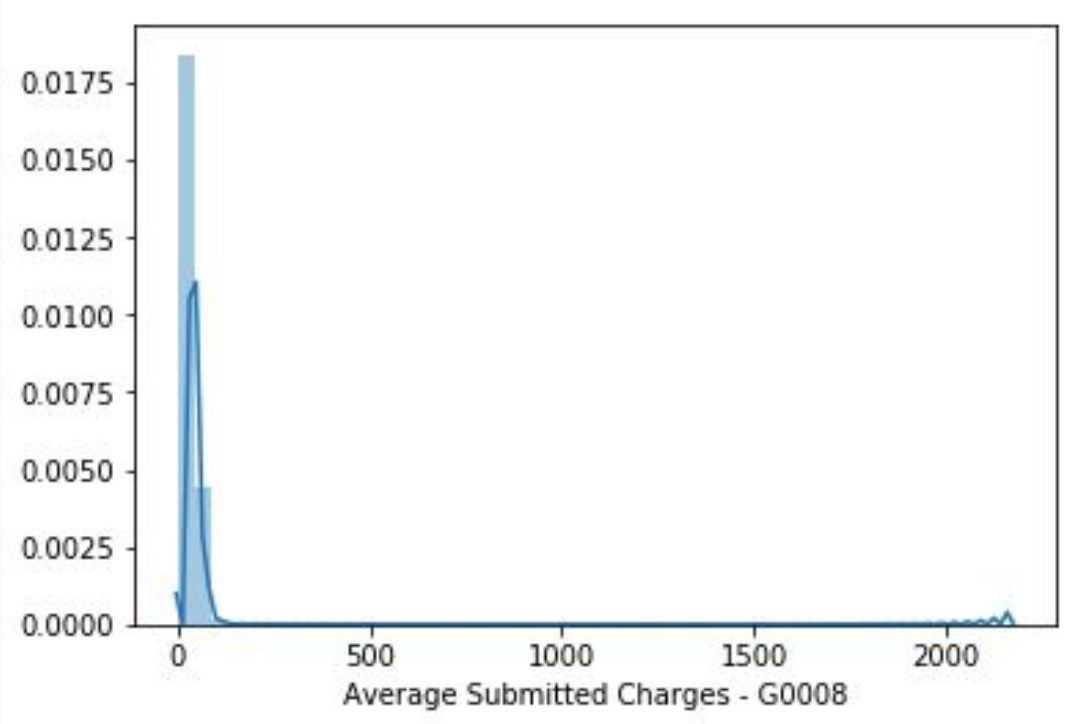
First, a general count of the most common medical procedures was made

	hcpcs_code	hcpcs_description	COUNT(*)
0	99213	Established patient office or other outpatient...	432107
1	99214	Established patient office or other outpatient...	396522
2	99204	New patient office or other outpatient visit, ...	172499
3	99203	New patient office or other outpatient visit, ...	169193
4	99232	Subsequent hospital inpatient care, typically ...	168455
5	G0008	Administration of influenza virus vaccine	140912
6	99212	Established patient office or other outpatient...	136629
7	99223	Initial hospital inpatient care, typically 70 ...	129439
8	99215	Established patient office or other outpatient...	114113
9	99233	Subsequent hospital inpatient care, typically ...	114086

First steps

After filtering for relevant medical procedures (office visits were left out), some EDA was performed on the distribution of average charges for common procedures

Example: Flu vaccination



Mass billing data

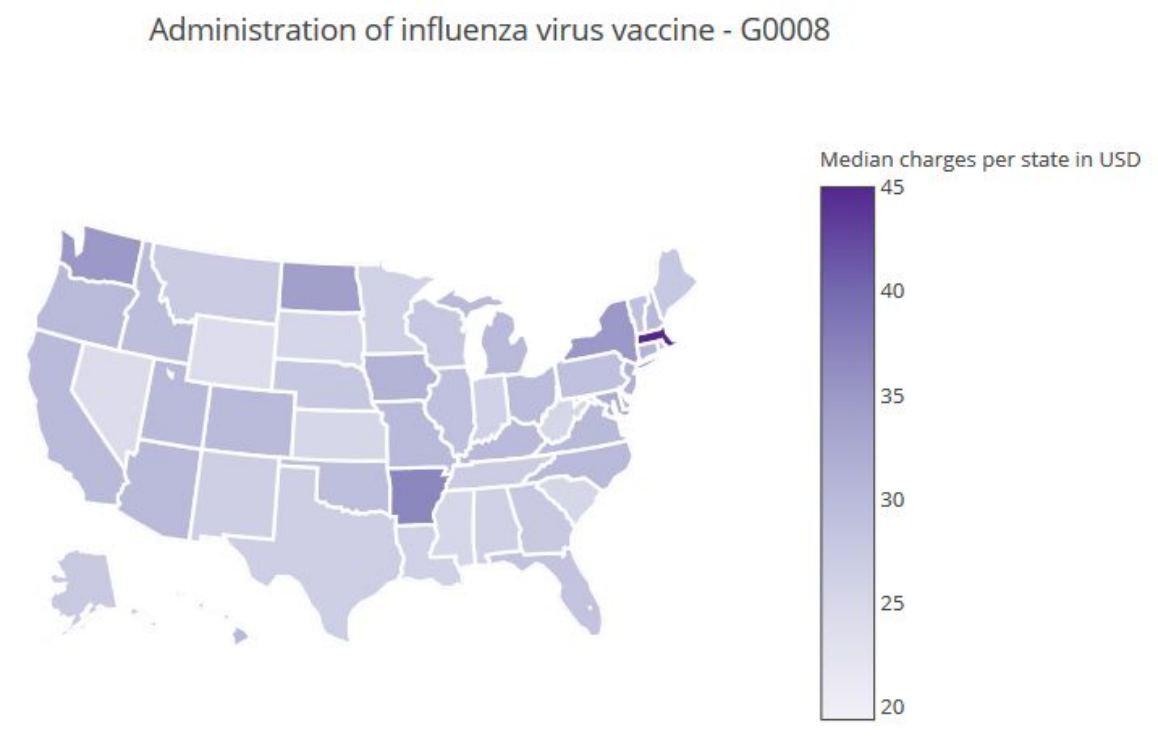
One of the reasons for such high charges, has to do with billing data which is submitted in bulk, as exemplified below:

	npi	nppes_provider_last_org_name	nppes_credentials	provider_type	average_submitted_chrg_amt
0	1790846335	CARLSON	MD	Family Practice	2516.1393132
1	1669789129	PRICE PHARMACIES INC	None	Mass Immunizer Roster Biller	2172.3846452
2	1790846335	CARLSON	MD	Family Practice	2082.1977273
3	1841373933	BOND'S DRUG STORE	None	Mass Immunizer Roster Biller	2030.4971613
4	1275840753	PRICE PHARMACIES INC	None	Mass Immunizer Roster Biller	1786.7261765
5	1891110615	HERITAGE VILLAGE PHARMACY INC	None	Mass Immunizer Roster Biller	960.58263158
6	1437165370	BORSOOK	M.D.	Family Practice	869.18518519
7	1124065461	SHOPKO STORES OPERATING CO LLC	None	Mass Immunizer Roster Biller	789.82818182
8	1518951359	BOUVIER PHARMACY INC	None	Mass Immunizer Roster Biller	652.34707207
9	1669789129	PRICE PHARMACIES INC	None	Mass Immunizer Roster Biller	594.77361446

Visualizing average charges

In the exploratory data analysis phase, some choropleth maps were made, in order to determine if there were regional differences in average charges.

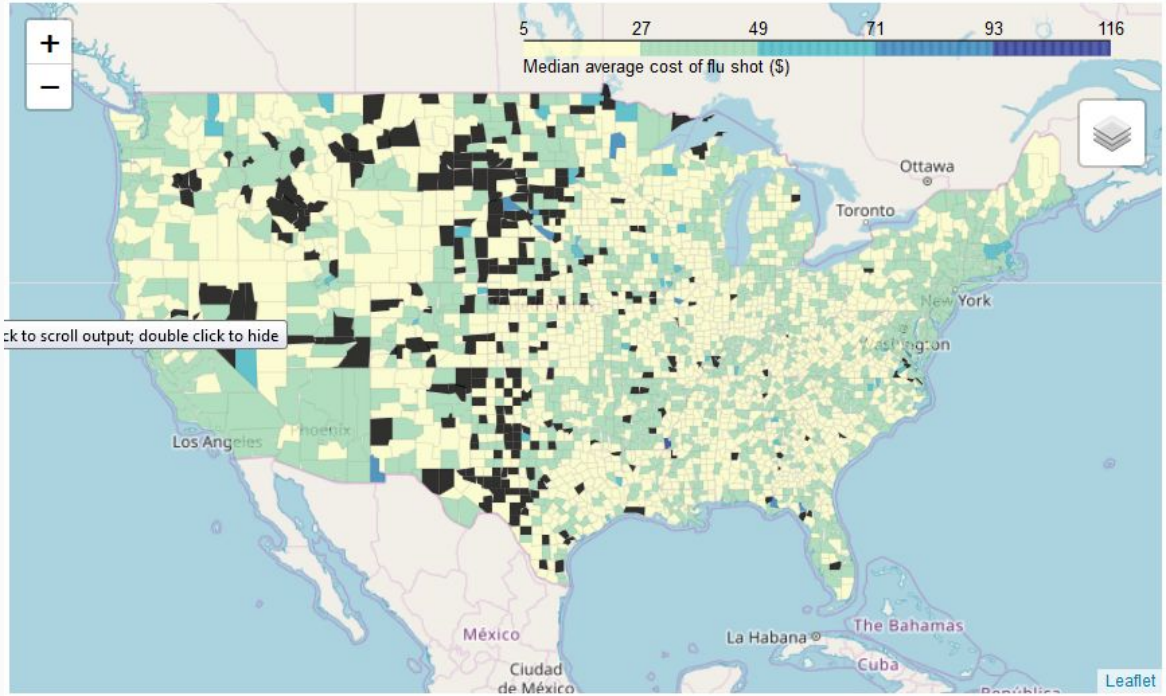
This map displays median charges at the state level for the Administration of Flu Vaccine procedure, where we began our analysis.



Visualizing average charges

A more detailed map was made using the fact that we have data at the county level.

This choropleth shows average charges for the same procedure but per county.



Some initial thoughts

After performing an initial analysis on the data, there was no clear indication of why charges were higher in certain parts of the country. The differences could be due to a wide variety of reasons.

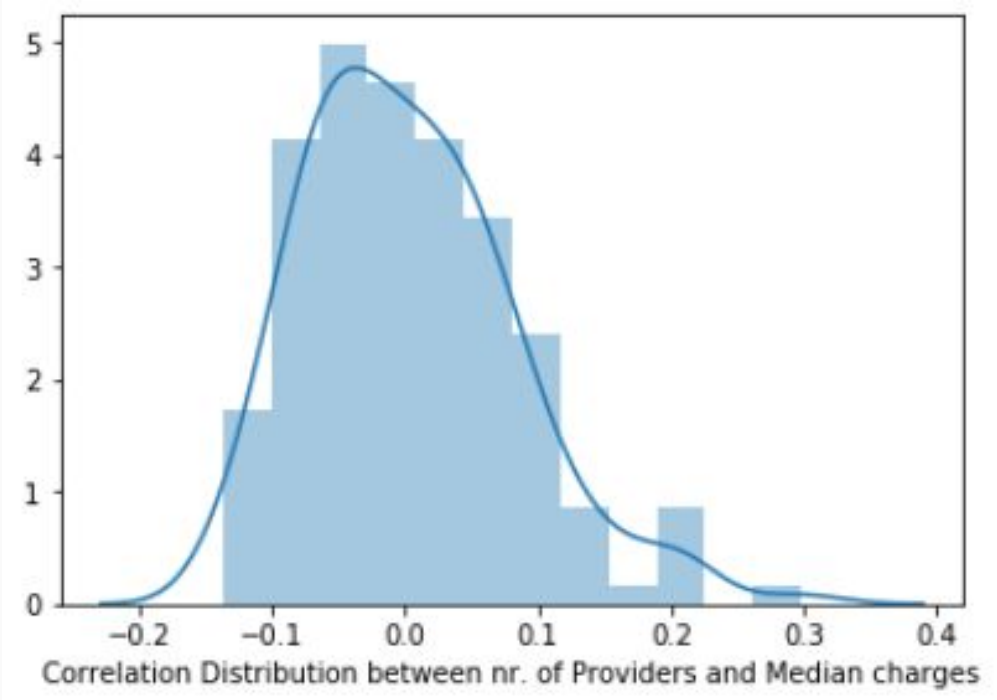
Also, it was necessary to find a way to filter out values which would belong to mass billing payments, so that these would not skew the rest of the data

In-depth Analysis

Correlation with nr. of providers

One of the first steps in performing in-depth analysis using inferential statistics was to perform correlation analysis between average charges for the procedures in the dataset and other variables.

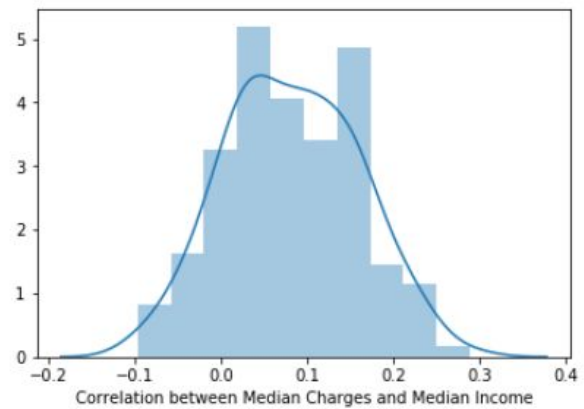
The first correlation was between the median values of the average charges and the number of providers per county. As demonstrated in the graph, there was no correlation between the two variables

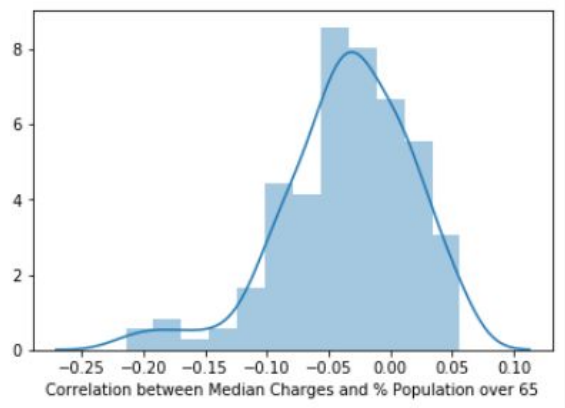


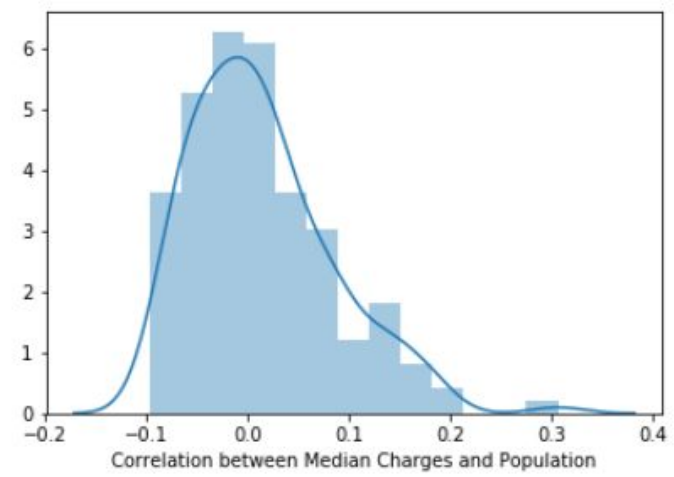
Correlation with demographic data

After incorporating some demographics data, some correlations were also performed between between the median values of the average charges and demographic variables.

Again, there was no strong correlation between the media values and the various demographic variables.







Normality test and percentiles

Normality tests were performed on all the procedures and none of them had a normal distribution.

Therefore, the focus of the analysis shifted into taking a closer look at the data in the highest percentiles, in order to determine if there was any relationship between high percentile values and other variables.

The analysis focused on the average charges data by procedure which was in the range of 75th and 90th percentile, in order to exclude outlier data points.

Matrix CPT vs FIPS

As part of the analysis, a matrix was developed which mapped all relevant procedures against all the counties (represented by FIPS codes). The values in the matrix represent the percentage of providers which charged prices in the 75-90th percentile range for each procedure and county combination. A sample of the data matrix is shown below:

FIPS	G0008	36415	G0009	90662	93000	90670	G0439	96372	71020
01001	4.761905	30.000000	NaN	NaN	NaN	NaN	NaN	5.263158	18.181818
01013	13.333333	11.111111	NaN	10.000000	22.222221	NaN	100.000000	NaN	40.000000
01017	27.272728	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
01019	10.000000	NaN	NaN	NaN	80.000000	NaN	NaN	NaN	NaN
01039	4.166667	NaN	20.000000	NaN	NaN	NaN	20.000000	8.695652	36.363636
01043	2.631579	2.173913	7.142857	NaN	15.000000	NaN	NaN	NaN	14.285714
01051	3.571429	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
01055	9.259259	NaN	NaN	NaN	26.470589	27.777779	20.000000	4.761905	13.043478
01059	7.692307	NaN	25.000000	NaN	33.333332	NaN	NaN	NaN	25.000000
01069	2.439024	NaN	10.000000	2.564103	8.064516	NaN	4.545455	2.197802	11.864407

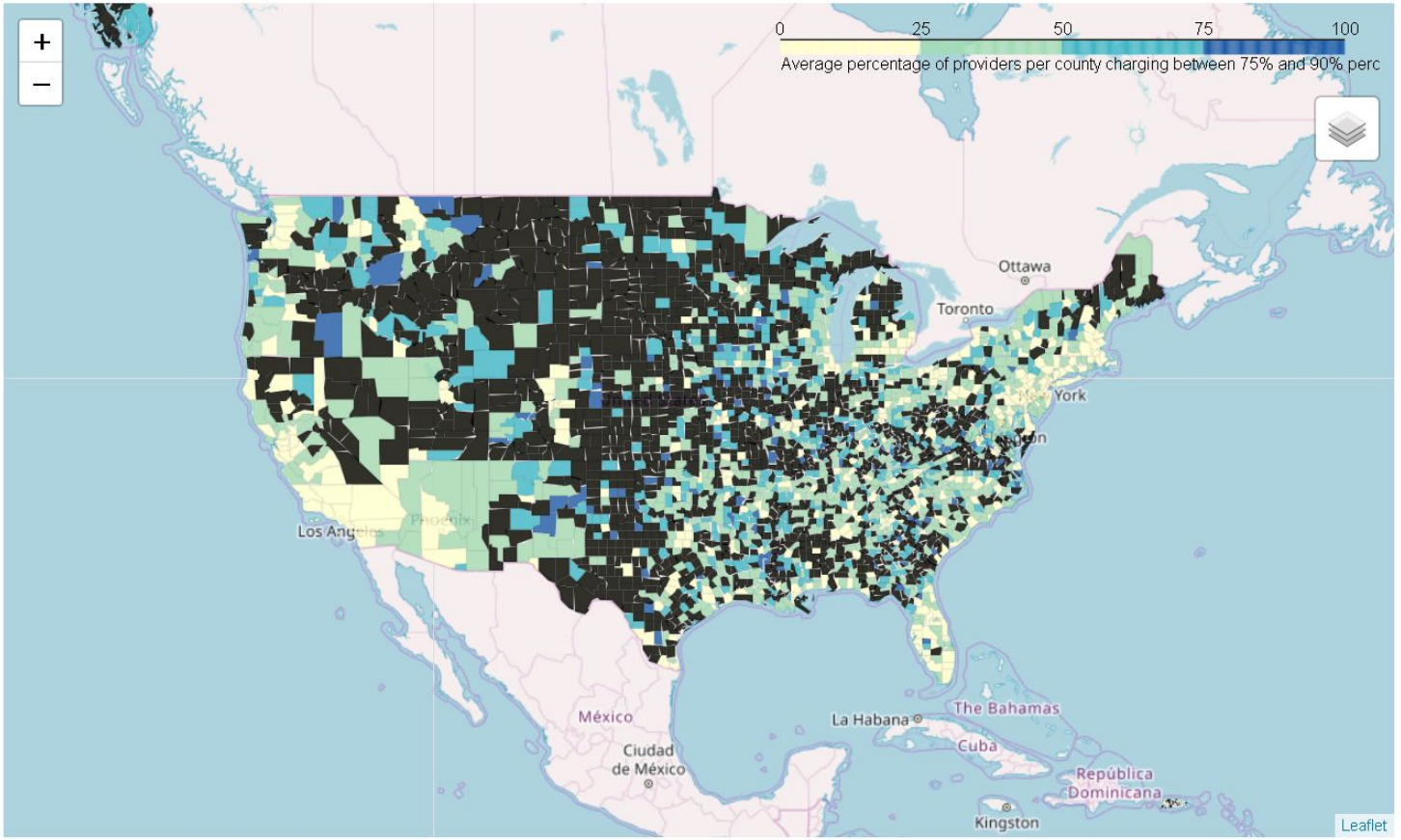
Analysis of matrix data

Given the high number of medical procedures in the data, it was decided to calculate the mean percentage by county.

Upon filtering the data by those counties with a mean value above 50%, there were 525 out of 1572 counties, which have more than 50% of their providers charging average prices between the 75% and the 90% percentile range.

These average values per county were mapped, as shown in next slide

Choropleth map for avg. percentage of providers with avg. charges between 75th and 90th percentiles



Visualization of matrix data

The choropleth maps of the matrix data seem to indicate that there is a higher percentage of providers charging higher than average prices mostly in more remote areas.

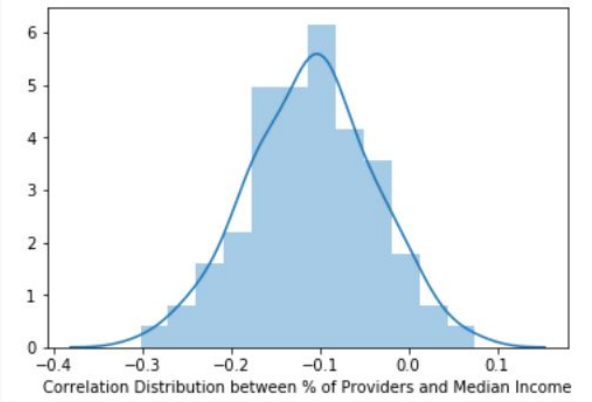
This would seem to indicate that there could be some correlation between more remote areas and higher than average prices.

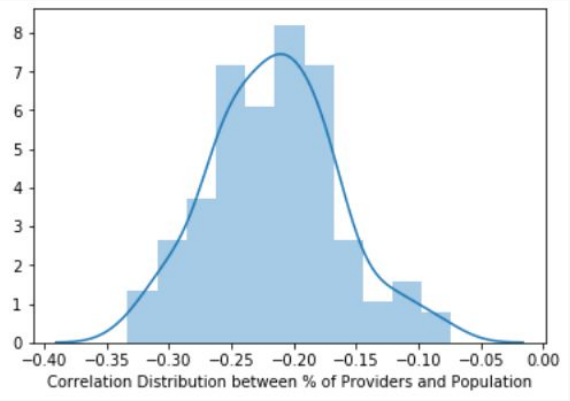
One last analysis was performed in order to determine if such a relationship existed.

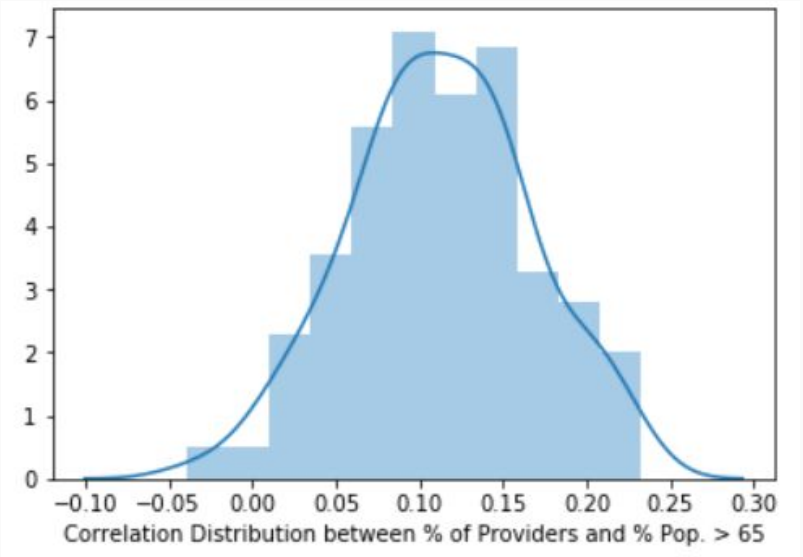
Correlation with demographic data

Again, some correlations were also performed between between the median values of the average charges and demographic variables.

However, as now the analysis focused on the values in the 75-90th percentile range, our results were somewhat different, as we found some moderate correlations between the above average charges and demgraphic variables, as illustrated by the distribution plots.







Correlation values

The last analysis to be performed was to calculate the correlation between the mean percentage of provider per county with above average charges and demographic variables. The results were as follows:

- Correlation between Mean % and Median Income: -0.275
- Correlation between Mean % and Population: -0.383
- Correlation between Mean % and % Population over 65: 0.264

The correlation values obtained indicate that there is a moderate inverse correlation between above average charges and population and median income, and a moderate correlation with percentage of population over 65.

Conclusions

Conclusions

The Medicare charges dataset was quite interesting and it provided many possible avenues of exploration.

The dataset, however, had its limitations, as the charges data provided are average charges, which in itself can somewhat distort the true range of values.

Despite these limitations, it can be concluded that there is a moderate negative correlation between average charges and population size, meaning that rural areas of the country may have higher charges than more populated parts of the country.