Capstone Project 1 - Final Report

1. Introduction

The purpose of this Capstone Project is to analyze and visualize Physician payment data for the Medicare program in 2016, and related demographic data.

Background

The topic of healthcare costs is one that I am interested in for two major reasons: first, I worked as Data Analyst in a healthcare organization for almost three years, and I was able to see first hand how data is being used to improve quality metrics and to provide more efficient healthcare.

Second, as user of healthcare services and health insurance, I am often surprised by the increasingly high cost of medical care in the United States. Although the quality of care is undeniable, one often wonders if the prices charged to patients are reasonable, especially when one compares the prices charged in other countries which also have high-quality healthcare.

As stated by The Commonwealth Fund:

"Despite spending more on health care, Americans had poor health outcomes, including shorter life expectancy and greater prevalence of chronic conditions."[1]

The following chart with healthcare spending as a percentage of GDP for several OECD countries further illustrates this point:

[1] The Commonwealth Fund, "U.S. Health Care from a Global Perspective", https://www.commonwealthfund.org/publications/issue-briefs/2015/oct/us-health-care-global-perspective?redirect_source=/publications/issue-briefs/2015/oct/us-health-care-from-a-global-perspective

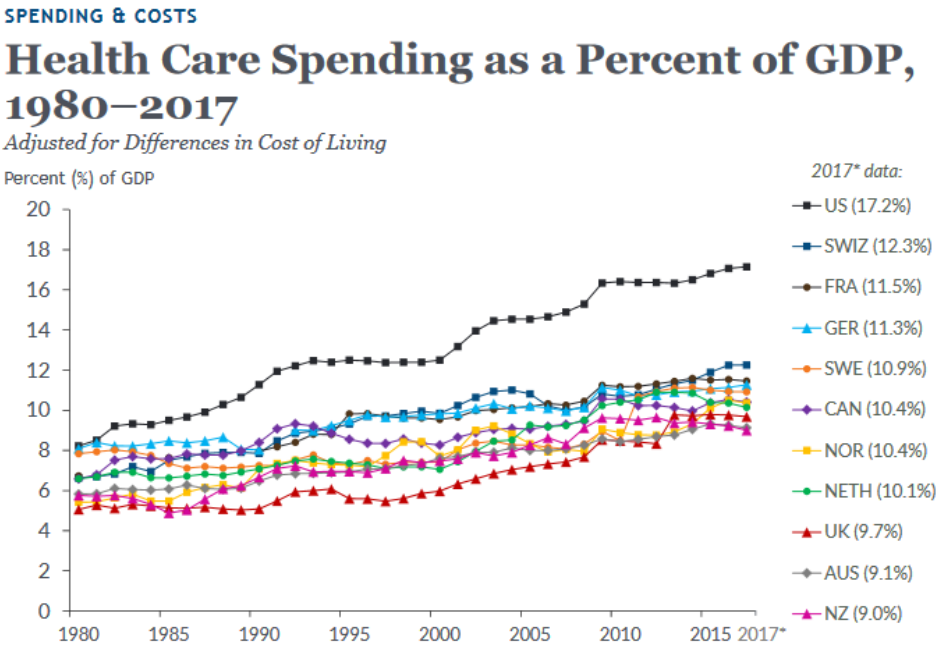


Figure 1 - Health Care Spending as GDP Percentage[2]

In light of this, I sought to analyze existing public data related to healthcare costs and charges as a starting point for a possible project on this subject.

The Centers for Medicare & Medicaid Services (CMS) is the federal agency in charge of administering Medicare, Medicaid, Children's Health Insurance Program (CHIP), and the Health Insurance Marketplace. The agency is a part of the Department of Health and Human Services (HHS).

As part of its mandate to "make our healthcare system more transparent, affordable, and accountable"[3] , CMS has begun making several data sets public based on the data that it collects from physicians and healthcare facilities from all over the United States. Amongst the many data sets which are made available by CMS, we can find data related to Medicare payments in several categories, such as data on Physician charges (i.e. doctor's appointments, medical procedures, etc.), Inpatient charges (i.e.: hospital admissions), Outpatient charges (i.e.: emergency services, outpatient surgery, etc.) and Part D charges (Medicare's drug prescription program).

Project Objectives

[2] The Commonwealth Fund, "Multinational Comparisons of Health Systems Data, 2018" - https://www.commonwealthfund.org/publications/publication/2018/dec/multinational-comparisons-health-systems-data-2018

[3] CMS, "New Medicare utilization and payment data available for medical equipment, supplies" - https://www.cms.gov/newsroom/press-releases/new-medicare-utilization-and-payment-data-available-medical-equipment-supplies

The purpose of this project will be to analyze and visualize Medicare charges data and try to answer some of the following questions:

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

The focus of this analysis will be on the Physician dataset, as it includes data on the most common medical costs (such as doctor's visits, medical exams, etc.) and it also a quite extensive dataset (around 1.7GB and over 9 million records). The analysis will focused on the 2016[4] year dataset, as it is the most current one available.

Other datasets to be used include the Geocoded National Provider Identification dataset[5], which will allow us to map the various providers to their geographical location, using FIPS codes. Additional datasets with demographic data (ie. Census data on a county level) maybe added as necessary.

It would be interesting to see if higher charges are related to better care, or if higher charges are related to the specific geographic and demographics of where the services are offered. It could be that physicians charge more in certain areas due to higher cost of living, or because there is less competition or other factors.

2. Data Wrangling

Overview of Main Datasets

---

[4] CMS Medicare Physician Data - https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier2016.html

[5] NPI Geographic data - https://www.naaccr.org/NPI/NPI_Geocode_Files_07_2015.zip

The data sets that I will be using for the capstone project are the following:

- Medicare_Provider_Util_Payment_PUF_CY2016.txt - Medicare Physician charges for 2016 (CSV file, size: 2.14 GB) -
  Link: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier2016.html

- NPI_Geocodes_CSV_2014_04.txt - National Provider Identification information for 2014 (CSV file, size: 563 MB)
  Link: https://www.naaccr.org/NPI/NPI_Geocode_Files_04_2014.zip

Both datasets provide clean data for the most, and they require only some minimal data cleaning efforts.

As the datasets that I am using are reasonably large, I used to SQL Developer to import the data into the database and perform the necessary data wrangling and cleanup.

Data Import Process

SQL Developer provides an efficient data import tool which, based on my previous experience as a data analyst, is faster than using a Python script to read data from a CSV file and importing it into the database.

Based on this, I used the “Data Import Wizard” to import both datasets into an Oracle database. During the process, I took great care to ensure that the columns would be of the appropriate data type and appropriate size. An example of this step is shown in the image below:

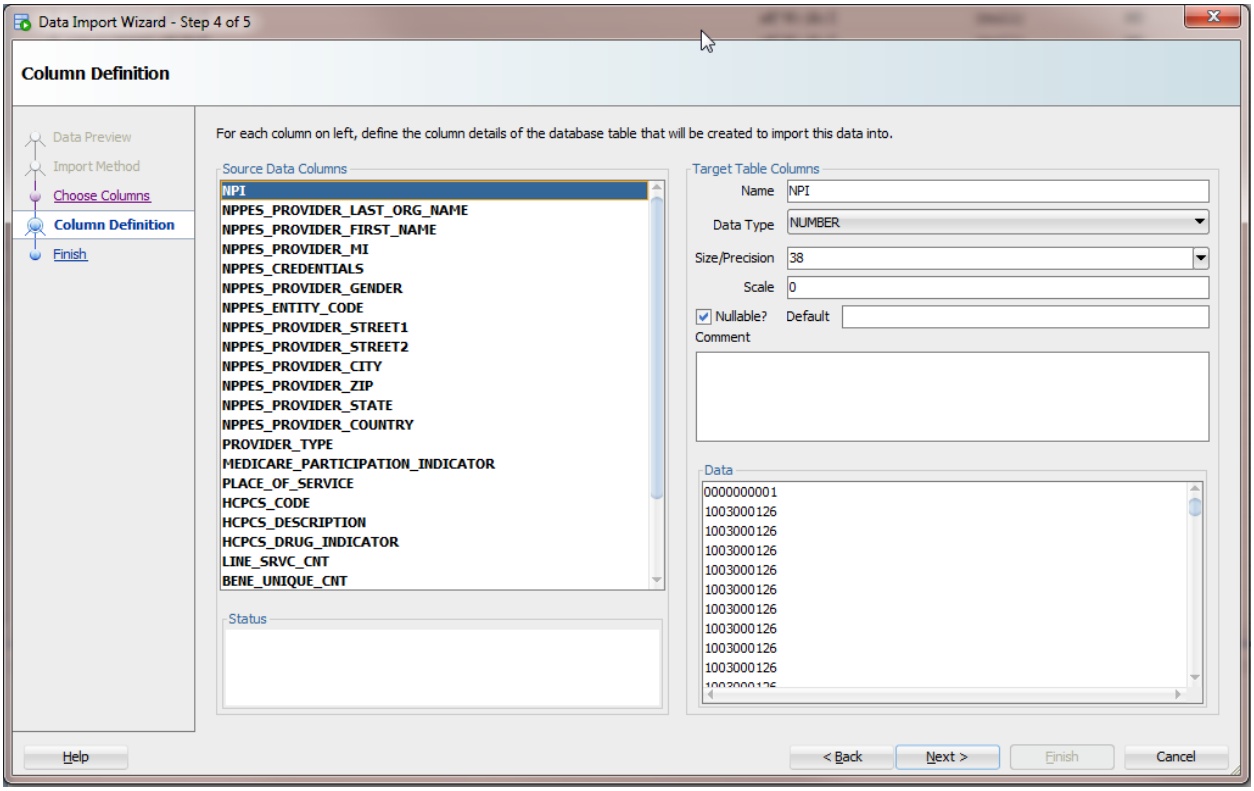
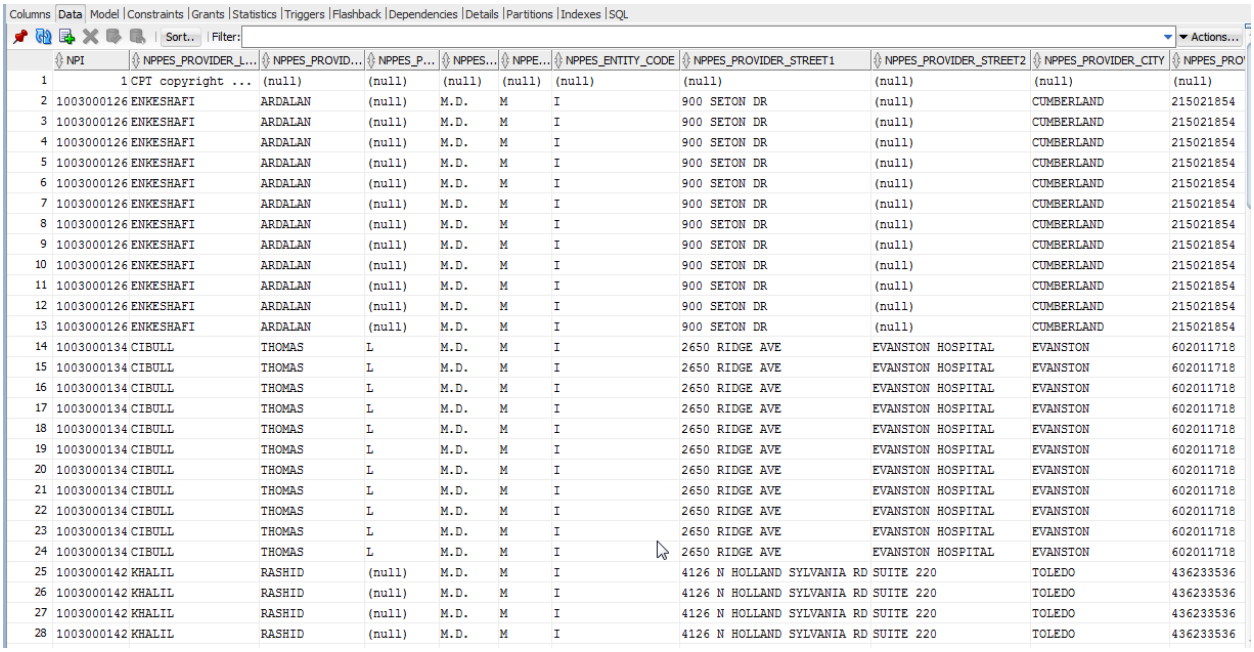


Figure 2 - Oracle SQL Developer Data Import Wizard

On the other hand, I did not worry so much about choosing appropriate sizes for the columns. Again, based on previous experiences, I preferred to use larger sizes for both VARCHAR columns and NUMBER columns in order to avoid any surprises with larger than expected values during the data import process, which would require me to repeat the process from the beginning.

Both datasets were successfully imported into the database, as exemplified below:



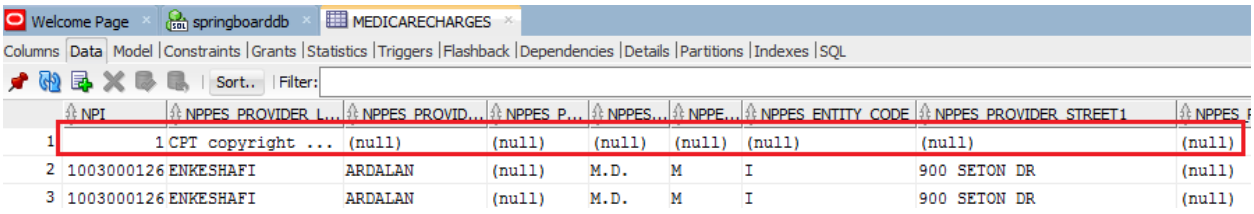
	NPI	NPPES_PROVIDER_L...	NPPES_PROVID...	NPPES_P...	NPPES...	NPPE...	NPPES_ENTITY_CODE	NPPES_PROVIDER_STREET1	NPPES_PROVIDER_STREET2	NPPES_PROVIDER_CITY	NPPES_PRO
1	1	CPT copyright ...	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)
2	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
3	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
4	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
5	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
6	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
7	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
8	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
9	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
10	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
11	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
12	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
13	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)	CUMBERLAND	215021854
14	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
15	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
16	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
17	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
18	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
19	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
20	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
21	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
22	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
23	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
24	1003000134	CIBULL	THOMAS	L	M.D.	M	I	2650 RIDGE AVE	EVANSTON HOSPITAL	EVANSTON	602011718
25	1003000142	KHALIL	RASHID	(null)	M.D.	M	I	4126 N HOLLAND SYLVANIA RD	SUITE 220	TOLEDO	436233536
26	1003000142	KHALIL	RASHID	(null)	M.D.	M	I	4126 N HOLLAND SYLVANIA RD	SUITE 220	TOLEDO	436233536
27	1003000142	KHALIL	RASHID	(null)	M.D.	M	I	4126 N HOLLAND SYLVANIA RD	SUITE 220	TOLEDO	436233536
28	1003000142	KHALIL	RASHID	(null)	M.D.	M	I	4126 N HOLLAND SYLVANIA RD	SUITE 220	TOLEDO	436233536

Figure 3 - Example of the data in the ‘MedicareCharges’ table

Data Cleaning Process

The data cleaning process involved mostly performing a visual inspection of data, by sorting it by ascending and descending values on various columns, in order to find any strange values.

Upon finishing the data import process, the very first row of the ‘MedicareCharges’ table contained an invalid row:



	NPI	NPPES_PROVIDER_L...	NPPES_PROVID...	NPPES_P...	NPPES...	NPPE...	NPPES_ENTITY_CODE	NPPES_PROVIDER_STREET1	NPPES_
1	1	CPT copyright ...	(null)	(null)	(null)	(null)	(null)	(null)	(null)
2	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)
3	1003000126	ENKESHAFI	ARDALAN	(null)	M.D.	M	I	900 SETON DR	(null)

Figure 4 - Erroneous data in ‘MedicareCharges’ table

This row was obviously a line with copyright information which got imported into the database from the CSV file.

SQL Developer provides a convenient way to delete unwanted data rows from the menu:

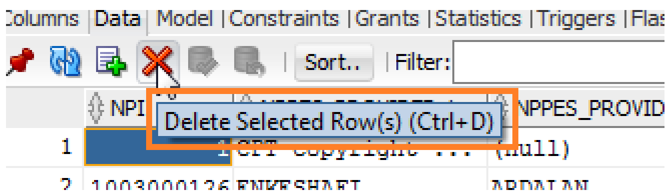
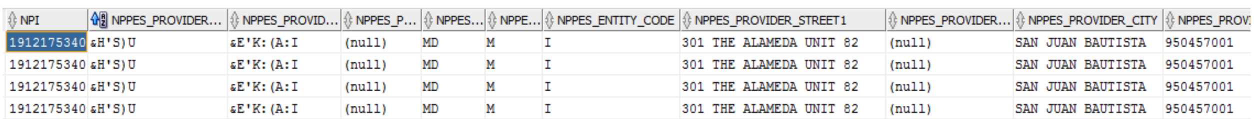


Figure 5 - SQL Developer Row Deletion

Some strange values were also found, after sorting by Provider name:



NPI	NPPES_PROVIDER...	NPPES_PROVID...	NPPES_P...	NPPES...	NPPE...	NPPES_ENTITY_CODE	NPPES_PROVIDER_STREET1	NPPES_PROVIDER...	NPPES_PROVIDER_CITY	NPPES_PROV
1912175340	&H'S)U	&E'K:(A:I	(null)	MD	M	I	301 THE ALAMEDA UNIT 82	(null)	SAN JUAN BAUTISTA	950457001
1912175340	&H'S)U	&E'K:(A:I	(null)	MD	M	I	301 THE ALAMEDA UNIT 82	(null)	SAN JUAN BAUTISTA	950457001
1912175340	&H'S)U	&E'K:(A:I	(null)	MD	M	I	301 THE ALAMEDA UNIT 82	(null)	SAN JUAN BAUTISTA	950457001
1912175340	&H'S)U	&E'K:(A:I	(null)	MD	M	I	301 THE ALAMEDA UNIT 82	(null)	SAN JUAN BAUTISTA	950457001

Figure 6 - Provider with erroneous data

Upon doing some quick research using the address information, we were able to get the correct name:

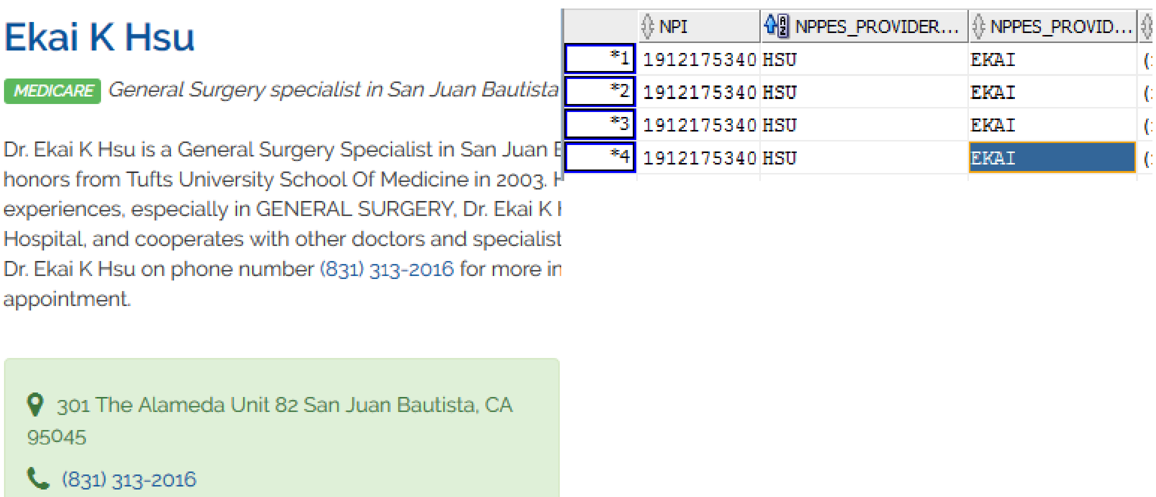


Figure 7 - Correcting provider information

Upon further inspection, I found that the dataset includes data for providers outside the United States:

NPPES_PROVIDER_STREET1	NPPES_PROVIDER...	NPPES_PROVIDER_CITY	NPPES_PROVIDER_ZIP	NPPES_PROVIDER_ST...	NPPES_PROVIDE...
755 YORK MILLS ROAD	APT NUMBER 1002	TORONTO	M3B 1X4	ZZ	CA
LANDSTUHL REGIONAL MEDI...	CMR 402	APO	09180	ZZ	DE
755 YORK MILLS ROAD	APT NUMBER 1002	TORONTO	M3B 1X4	ZZ	CA
515-30 ST NW	(null)	CALGARY	T2N2V4	ZZ	CA
515-30 ST NW	(null)	CALGARY	T2N2V4	ZZ	CA
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
NOFA TOWERS, APT 10A	BLOCK 2, COAS...	MAHBOULA	51008	ZZ	KW
CMR 414	BOX 1393	APO	09173	ZZ	DE
ASHTON ROAD	EMERGENCY DEP...	LANCASTER	LA1 4RP	ZZ	GB
ASHTON ROAD	EMERGENCY DEP...	LANCASTER	LA1 4RP	ZZ	GB
ASHTON ROAD	EMERGENCY DEP...	LANCASTER	LA1 4RP	ZZ	GB
LANDSTUHL REGIONAL MEDI...	CMR 402	APO	09180	ZZ	DE
LANDSTUHL REGIONAL MEDI...	CMR 402	APO	09180	ZZ	DE
LANDSTUHL REGIONAL MEDI...	CMR 402	APO	09180	ZZ	DE

Figure 8 - Provider data for other countries

Since our analysis will be focused only on the US, I decided to remove all the data for charges for providers outside of the US, using the following query:

DELETE FROM MEDICARECHARGES
WHERE NPPES_PROVIDER_COUNTRY !='US'

The query deleted 527 rows associated with charges in foreign countries. However, upon further review of the remaining data, I also verified that the dataset contained additional rows with ‘XX’ as an US State, as well as, data associated military addresses and other US territories. I decided to leave the data, as the NPI Geographical data file will probably have the correct state for these providers. Also, I left the data for the military addresses and US territories, as it may be useful at a later stage for comparison purposes.

The remaining columns of the ‘MedicareCharges’ had no issues with the data values.

With regards to the ‘NPIGeoData’ table, the issues revolved mostly around rows with null values which were of no use in the project. The purpose of this dataset is to provide a way to map providers to their appropriate location on a US county map, using FIPS codes. Given that some providers had null FIPS codes, these were deleted from the dataset, using the following query:

DELETE FROM NPIGEODATA
WHERE FIPS_ST IS NULL

The query deleted 309 rows containing null values.

The remaining columns of the ‘NPIGeoData’ table were not so relevant, therefore, our review of the data focused only on the NPI, FIPS County, FIPS State and State Abbreviation columns.

Merging the two datasets

As mentioned earlier, one of the goals of the project is to use both datasets for mapping purposes, therefore, it is necessary at some point to merge them together. In order to avoid further data processing in pandas which is more time-consuming, the merging of both datasets was also performed in SQL Developer, using the following query:

```
CREATE TABLE MEDICARECHARGESFINAL AS
SELECT MEDICARECHARGES.*, NPIGEODATA.FIPS_CO, NPIGEODATA.FIPS_ST, NPIGEODATA.ST_ABBR
FROM MEDICARECHARGES, NPIGEODATA
WHERE MEDICARECHARGES.NPI = NPIGEODATA.NPI
```

The above query creates a new table ‘MedicareChargesFinal’ which combines all the columns from the ‘MedicareCharges’ tabke and the columns from the ‘NPIGeoData’ table which contain the FIPS codes for State, County and the State Abbreviation.

This table has around 9.4 million rows, around 300,000 less than the original ‘MedicareCharges’ table. This has to do with the fact that the NPIGeoData does not have data for all of the providers. A quick inspection upon missing data is exemplified by the results of the following query:

```
SELECT DISTINCT NPIGEODATA.ST_ABBR, MEDICARECHARGES.NPPES_PROVIDER_STATE
FROM NPIGEODATA
RIGHT JOIN MEDICARECHARGES
ON NPIGEODATA.ST_ABBR = MEDICARECHARGES.NPPES_PROVIDER_STATE
ORDER BY NPIGEODATA.ST_ABBR, MEDICARECHARGES.NPPES_PROVIDER_STATE ASC
```

The results of this query showed the following:

	ST_ABBR	NPPES_PROVIDER_STATE
47	VA	VA
48	VI	VI
49	VT	VT
50	WA	WA
51	WI	WI
52	WV	WV
53	WY	WY
54	(null)	AA
55	(null)	AE
56	(null)	AP
57	(null)	AS
58	(null)	GU
59	(null)	MP
60	(null)	XX

Figure 9 - Distinct States of ‘MedicareCharges’ and ‘NPIGeoData’ tables

As we can see from the output of the query, the NPIGeoData table does not include data for providers with military address (AA, AE and AP state abbreviations), as well as, data for other territories (American Samoa, Guam and Mariana Islands). The XX state abbreviation is for addresses from the ‘MedicareCharges’ which have not been correctly identified, as described earlier. However, this will not be an issue, as we can use the State Abbreviations from the NPIGeoData in the merged table.

We now have a dataset that contains both the charges and the appropriate geographic data which will allows to generate visualization, such as choropleth maps

Data Wrangling Conclusions

Overall, both datasets were of very good quality, with little in terms of data errors and null values.

The large size of the ‘MedicareCharges’ dataset required that the cleaning operations be performed in the database directly, using SQL Developer. However, once relevant data for the project is extracted into pandas, if further cleaning is necessary, it will be performed in Python.

3. Data Story

Initial Analysis

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

As previously mentioned, the dataset is quite large, and with a total of 9,444,398 records. Given its large size, it was necessary to focus on particular aspects of the dataset.

First, a general count of the most common medical procedures was made. Medical procedures in the US are usually referred to according to their Current Procedural Terminology codes (commonly referred to as CPT codes). The table below contains a list of the top 10 most common procedures in the dataset, including their CPT code (referred to in the dataset as "hcspcs_code") and corresponding description:

	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

Figure 10 - Top 10 most common procedures

As it can be seen from the table above, the most common procedures are patient office visits, however, given the possible variance on this type of data (specialists usually charge more than general practitioners, variable office visit lengths), I decided to filter office visits out of the analysis.

The first procedure at the top of the above list after office visits is "G0008", which refers to the "Administration of influenza virus vaccine". This code is specific to Medicare patients[6] and is

[6] "Vaccine Admin. HCPCS Code range G0008-G0010" Link: https://coder.aapc.com/hcpcs-codes-range/139

what we would commonly call a ‘flu shot’. Based on this, I sought to explore a bit more any specific details regarding this particular CPT code.

The first step was to collect all the submitted average charges by providers, as well as, the relevant geographic data, which included the FIPS county code and State of each provider, as illustrated below:

	npi	hcpcs_code	st_abbr	fips_co	avg_charges
0	1003000522	G0008	FL	127	73
1	1003001884	G0008	MI	049	30
2	1003002049	G0008	CA	001	31
3	1003002254	G0008	TN	179	19.4
4	1003006982	G0008	NY	085	15.025714286
5	1003004938	G0008	ME	001	22.493933333
6	1003006552	G0008	FL	105	25
7	1003007907	G0008	OR	051	27.25
8	1003008095	G0008	KY	235	45
9	1003009119	G0008	VA	107	40.155555556

Figure 11 - Example of relevant data to be used for analysis

Just a cursory look at the first 10 data rows, and it is visible the large variance in average charge values, ranging between 15 and 73 dollars. It was now time to do a distribution plot for the above data, so we could better understand the range of values for this particular procedures:

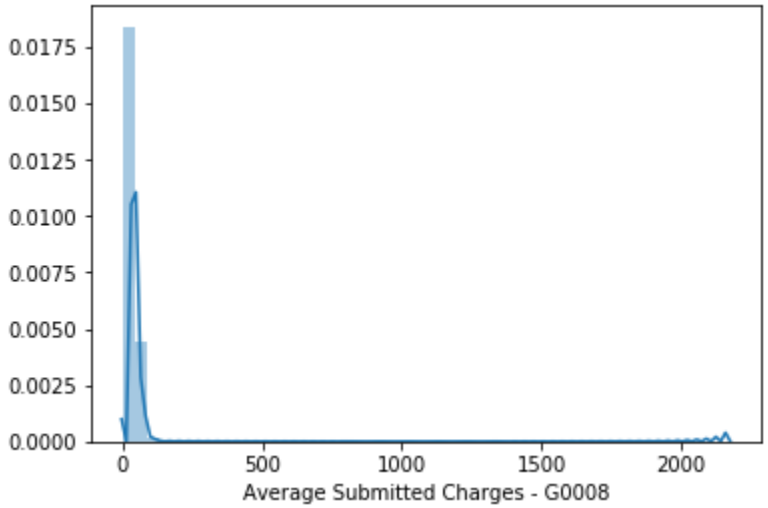


Figure 12 - Average submitted charges for administration of flu vaccine

. The distribution plot showed quite an usual range of values. It can be seen that the bulk of the charges are under 100 dollars, however, we also see some value as high as 2000 dollars. This warranted a closer inspection on the larger values, in order to understand what could be some possible reasons behind this huge discrepancy: The following table was based on a filter for submitted charges above 250 dollars and it confirmed the distribution plot above:

	npi	hcpcs_code	st_abbr	fips_co	avg_charges
0	1669789129	G0008	KS	035	2172.3846452
1	1790846335	G0008	MI	103	2082.1977273
2	1841373933	G0008	WV	107	2030.4971613
3	1275840753	G0008	KS	035	1786.7261765
4	1891110615	G0008	CT	009	960.58263158
5	1437165370	G0008	CA	037	869.18518519
6	1124065461	G0008	MO	079	789.82818182
7	1518951359	G0008	MA	017	652.34707207
8	1154418218	G0008	TN	149	563.05027397
9	1760576011	G0008	LA	015	447.27

Figure 13 - Top 10 highest average charges for administration of flu vaccine

Based on the above, some further investigation was warranted on these particular providers, and a possible explanation was found for some, but not all of the charges:

	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

Figure 14 - Provider information for top 10 highest average charges for administration of flu vaccine

As we can see from the table above, most of the very large submitted charges are from “Mass Immunizer Roster Biller”, meaning that these charges include charges for multiple procedures. However, we also see some M.D.’s who also have large values. It would be safe to assume that they too submitted their charges in bulk.

The next step of data analysis consisted in plotting the median submitted charges data on a choropleth map, in order to determine if there were any regional variables at play. Using the median would also help mitigate the extreme values in the distribution. The following plot was made using the plotly library:

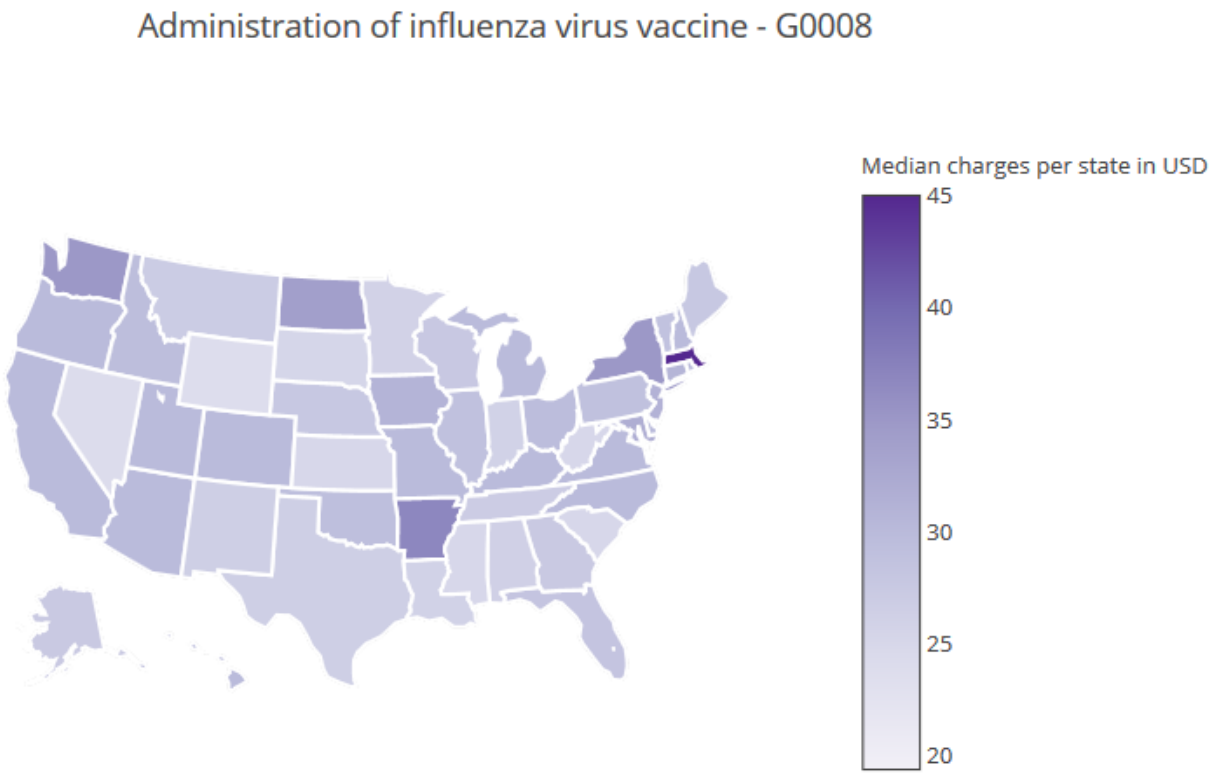


Figure 15 - Median average charges of administration of flu vaccine per state in USD

As we can see from the above plot, even using median values, a simple flu shot can cost 2.5 times more in certain states, when compared with the lowest median values.

Lastly, it would be relevant to get a more detailed idea of what are the median charges at the county level. The pandas “describe” function provided the following information about the medians of the average submitted charges, grouped by county:

count	**2922.000000**
mean	**27.934467**

std	**7.976592**
min	**5.000000**
25%	**24.000000**
50%	**26.241924**
75%	**30.000000**
max	**115.518614**

The median charges per county were plotted, using the “folium” library and the results were as follows:

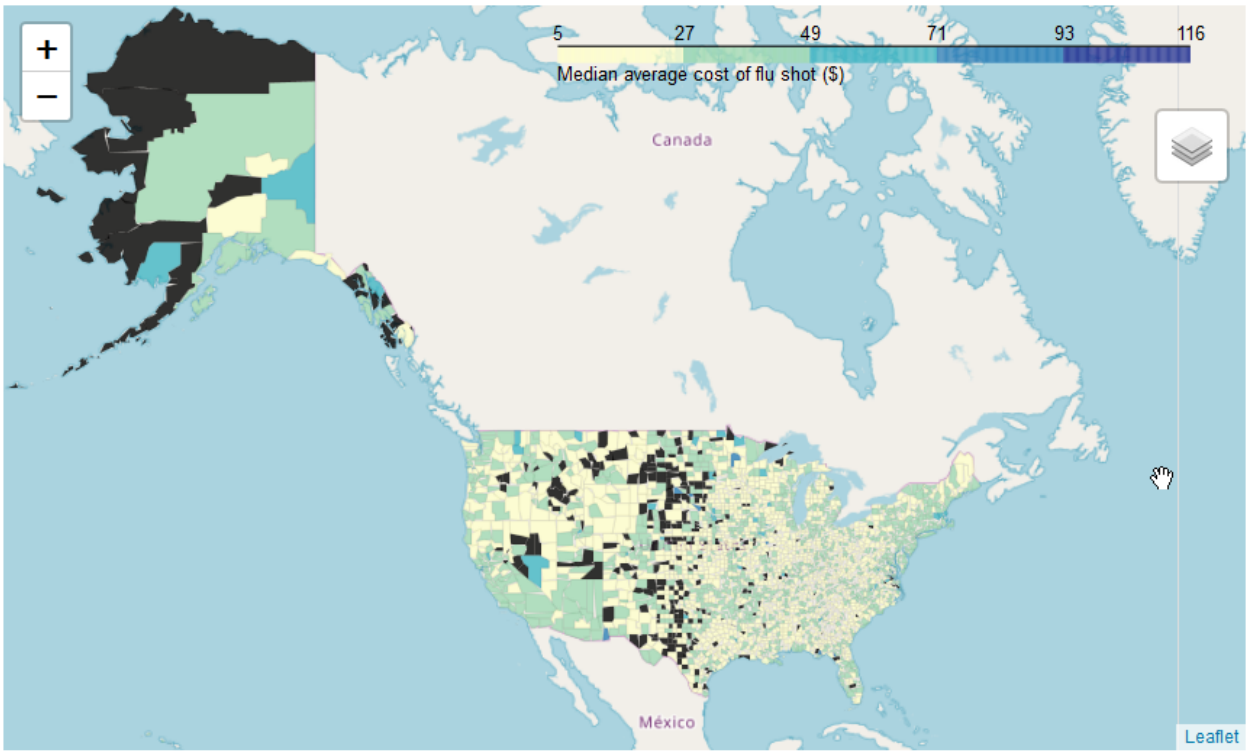


Figure 16 - Median charges map for administration of flu vaccination (Continental United States & Alaska)

A more detailed map of the Continental United States, is shown below:

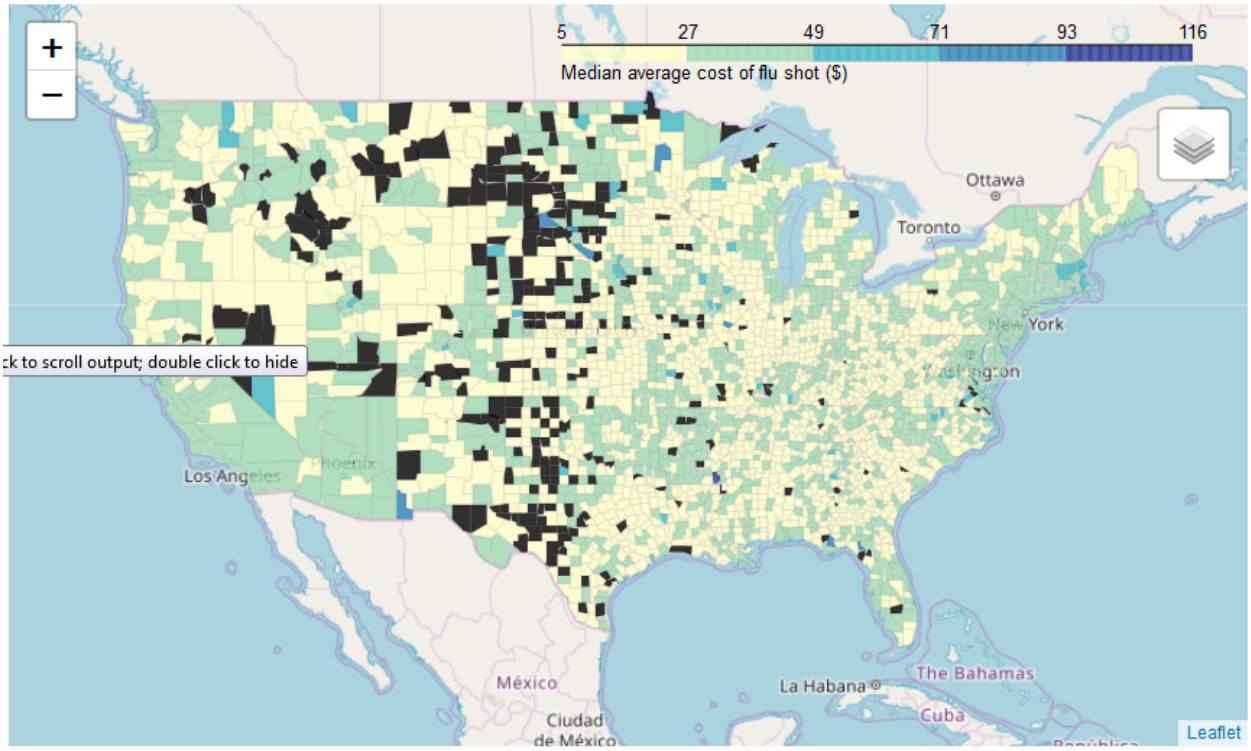


Figure 17 - Median charges choropleth map for administration of flu vaccination (Continental United States)

As it can be seen we see median charges above which are higher in some urban areas, which could be expected due to the higher cost of living. However, we also find extreme median values in what appear to be rural areas of several states.

First Analysis Conclusions

Based on this first analysis, further exploration is necessary to determine if there is any relationship between higher charges and certain regions of the country. This could be for a variety of reasons, such as, lack of competition, higher cost of living or other reasons. The more in-depth analysis on the next sections of the report will seek to possible answer some of these questions.

4. Inferential Statistics

Correlation analysis

After performing the initial analysis of the data which was described in the Data Story section, and discussing my initial results with my mentor, he suggested that I perform some correlation analysis on some of the relevant variables in the dataset.

One of the questions which I had lingering from the initial analysis was the fact that some of the high median values for avera charges appeared to be in isolated parts of the United States. This left me wondering if there was negative correlation between the number of providers in a county for a particular procedure and the the median charges, meaning that counties with less providers would have higher median values.

Based on this, I wrote some functions in python which would group both the median of the average submitted charges per county, as well as the total number of providers for each procedure according to its CPT code. A sample of the dataset is shown below:

FIPS	avg_charges	npi
01003	20.000000	7
01009	30.000000	2
01015	25.000000	9
01019	30.000000	1
01021	20.000000	1
01027	25.000000	1
01031	19.439999	2
01033	25.000000	5
01045	22.000000	1
01047	21.000000	1

Figure 18 - Average charges and number of providers per FIPS code sample data

After aggregating this data, I ran some correlation calculations over the entire dataset, which produced the following distribution plot for the top 160 most common procedures::

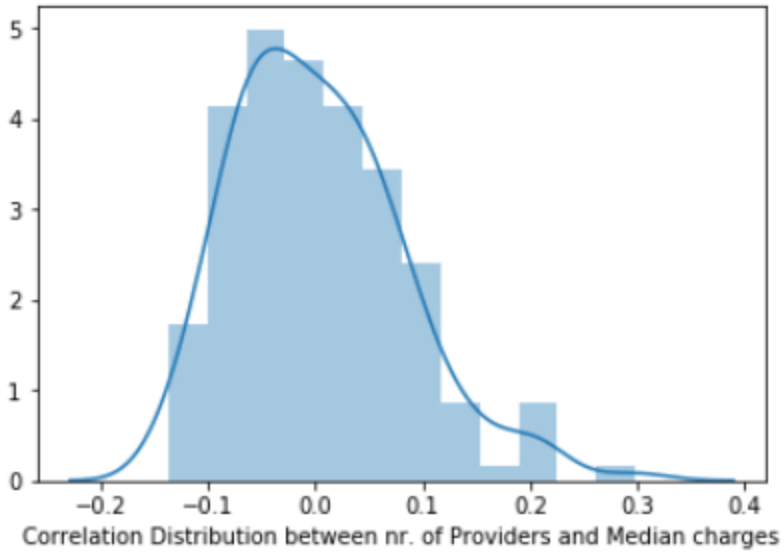


Figure 19 - Correlation Distribution between the number of providers and median charges

Based on the plot, there didn't seem to be any indications of a strong correlation between the number of providers and median charges per county for the top 160 most common procedures.

Given that relying only on the number of providers was not sufficient to give us any further insights which would explain the high median charges in some areas, I then proceeded to incorporate an additional dataset, the "2018 County Health Rankings National Data"[7]. Although the dataset was collected in 2018, the relevant data which I used was relative to 2016. I specifically chose Population, Median Income and Percentage of Population above 65 for all US counties. A sample of this data is shown below:

[7] County Health Rankings & Roadmaps - http://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation

	FIPS	State	County	Median_income	Population	Perc_over_65
0	01001	Alabama	Autauga	54487	55416	14.7
1	01003	Alabama	Baldwin	56460	208563	19.7
2	01005	Alabama	Barbour	32884	25965	18
3	01007	Alabama	Bibb	43079	22643	15.4
4	01009	Alabama	Blount	47213	57704	18
5	01011	Alabama	Bullock	34278	10362	16.3
6	01013	Alabama	Butler	35409	19998	19
7	01015	Alabama	Calhoun	41778	114611	16.9
8	01017	Alabama	Chambers	39530	33843	19.1
9	01019	Alabama	Cherokee	41456	25725	22

Figure 20 - Sample demographic data per county

With this new data in hand, I created another Jupyter notebook to specifically focus on this dataset and produce some more correlation plots between the median charges and these new variables.

The first correlation plot which I made was between median charges and median income for the top 160 most common procedures, as done previously:

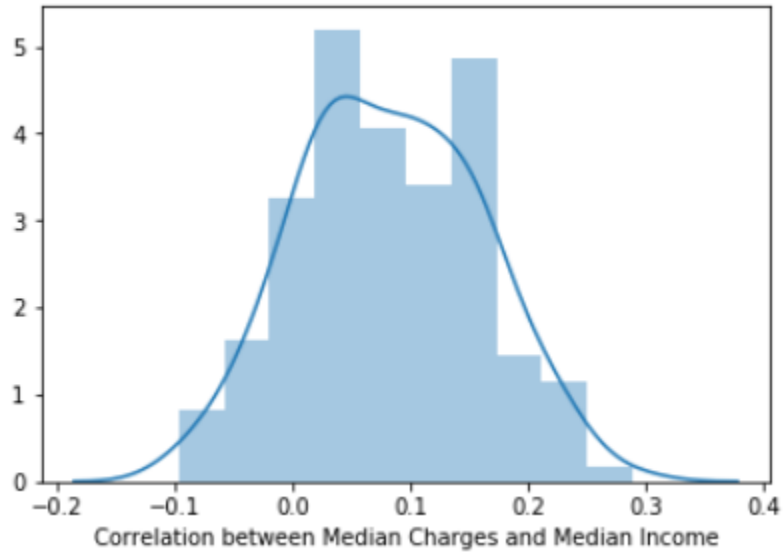


Figure 21 - Plot of correlation between median charges and median income per county

As it can be seen, the correlation plot seems to be evenly distributed around a correlation value 0.1, however the low correlation values are not enough to warrant a strong relationship between higher income and higher median charges.

The following plot correlates Median Charges and Population:

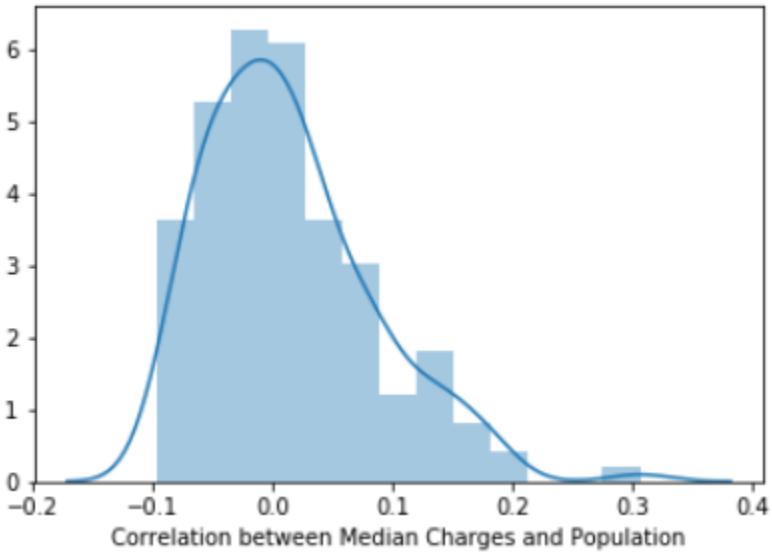


Figure 22 - Plot of correlation between median charges and population size per county

This correlation plot is slightly skewed to the right, however most of the correlation values are still centered around 0, showing again not much of a relationship between the two variables.

Lastly, one additional plot was made for the correlation between median charges and the percentage of population over 65 years of age:

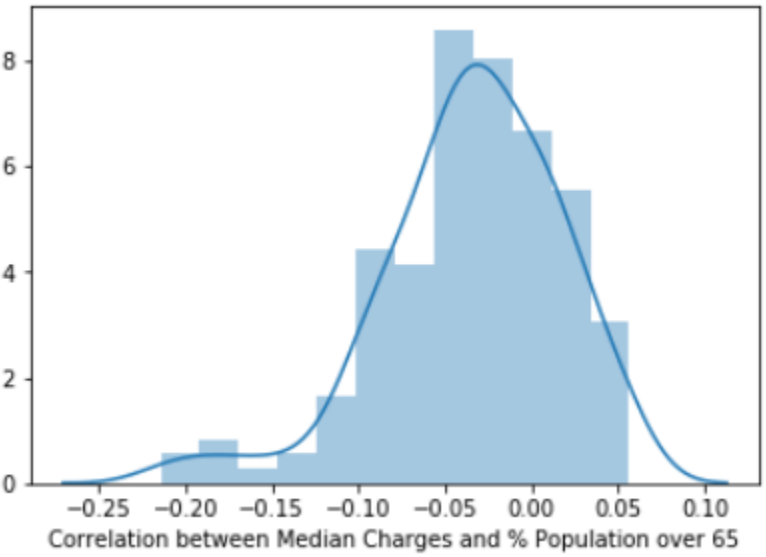


Figure 23 - Plot of correlation between median charges and percentage of population over age 65 per county

This distribution plot is slight left-skewed, where most of the values show a slight negative correlation, but again, not strong enough to warrant a definite conclusion.

Normality Test

As a last step in this stage of the project, my mentor suggested running a normality test on average charges data per procedure. The test was performed using the “normaltest” pandas function and it showed that none of the 160 procedures had a normal distribution.

Based on this, it was determined that it was necessary to look at the charges data into more detail, namely, in terms of specific percentiles, in order to gather more insights. This analysis will be discussed ahead in the report.

5. In-depth Analysis

Continuation of Inferential Analysis

The nature of the data being utilized in the project did not warrant the use of Machine Learning techniques, as the type of analysis being performed is focused on trying to understand possible relationships, not on predicting future values or trying to infer any sort of classification patterns in the data.

Given these constraints in the project, my in-depth analysis focused on further analysing the charges data in further detail using visualizations and inferential statistics, as done in the initial stages of the project.

The following step in the analysis of the charges data was to create a table with descriptive statistics of average charges by CPT code, with a particular focus on the various percentiles, as illustrated below:

	count	mean	std	min	10%	20%	25%	50%	75%	80%	90%	max
G0008	140912.0	33.228168	19.967112	0.010000	19.4	22.150320	23.000000	30.000000	40.000000	43.461697	53.333332	2172.384766
36415	93570.0	14.862174	9.786969	0.010000	5.5	8.000000	9.000000	14.000000	20.000000	20.000000	25.000000	874.090027
G0009	89307.0	35.801098	24.382719	0.010000	18.0	20.399092	23.043973	31.285715	44.629398	46.348903	57.094544	2057.633789
90662	85351.0	50.535107	18.505285	0.010000	40.0	40.590000	40.917183	42.328163	54.585850	60.000000	75.000000	500.000000
93000	83106.0	64.042839	33.132668	11.545403	33.0	40.000000	44.000000	59.000000	75.000000	80.000000	99.042194	525.859192

Figure 24 - Sample percentile data by CPT code

I decided to focus on the values in the range between 10% and 90%, and the range between 20% and 80%. By focusing on these percentile ranges, we can filter the majority of the outlier values, both those which are too small, as well as, those which can be associated with bulk submissions, as we saw in our initial analysis.

Using the data outlined above, a first distribution plot was made of the ratio of the 90% percentile over 10% percentile across our reference medical procedures:

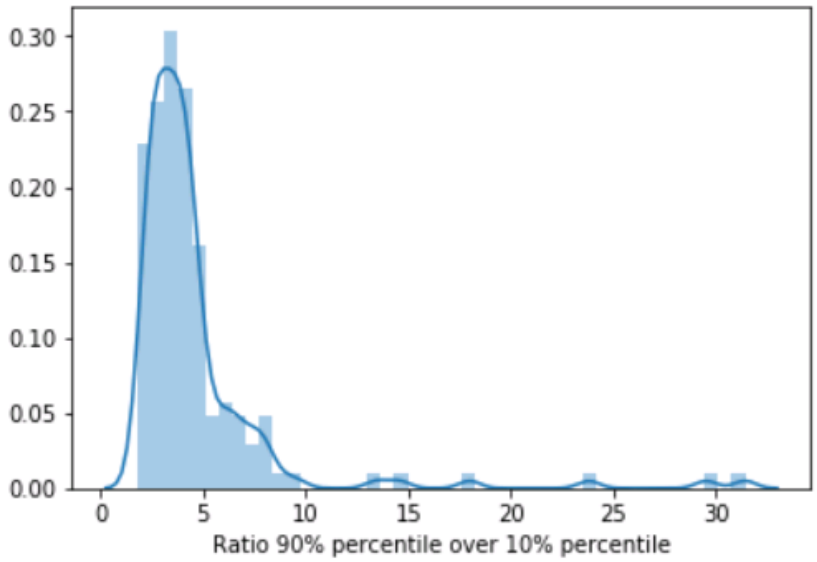


Figure 25 - Plot of average charge ratios of 90% percentile over 10% percentile

As the plot demonstrates, there is a substantial number of procedures which show a ratio which is at least five times or above. Even when filtering out for possible outliers, it is quite interesting to find such a range of values for medical procedures across the country.

The plot of the ratio of the 80% percentile over 20% percentile is shown below:

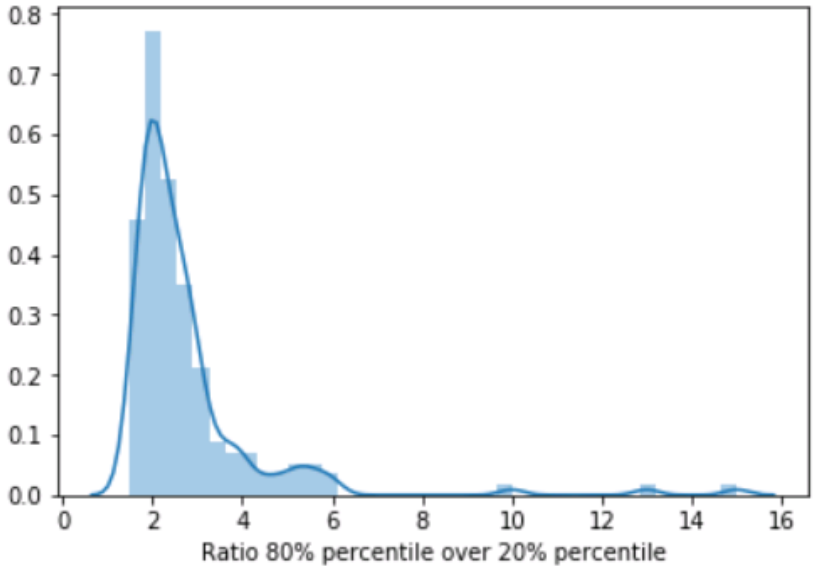


Figure 26 - Plot of average charge ratios of 80% percentile over 20% percentile

This plot provides further confirmation that even when looking at the narrower section of the data, the higher end of the charges can be between two and six times higher than the lower end of the range.

As done previously, I decided to take again a closer look at the provider ratio by county, but now focusing exclusively on the data between the 75% and 90%, in order to see if there were any particular areas where the vast majority of providers charges higher prices on average.
This resulted in a matrix of between CPT code columns and county FIPS code rows, as illustrated 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

Figure 27 - Sample data from FIPS x CPT matrix with percentage of providers with average charges between 75th and 90th percentile range

The values in the matrix are the percentage of providers in a particular county which have average charges which are between the 75th and 90th percentile for each particular procedure. “Nan” values mean that there aren’t providers for that particular procedure in that particular county.

As it can be seen in the matrix sample data below, there are indeed situations where in particular counties and for particular procedures, there is a ratio of 100%, meaning that there either very few providers and thus they charge above average prices, or providers in that area charge prices consistently above the average:

FIPS	G0008	36415	G0009	90662	93000	90670	G0439	96372	71020
05013	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
30095	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
19159	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
48131	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
37177	100.000000	NaN	100.000000	NaN	NaN	NaN	NaN	NaN	NaN
17059	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
19089	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
54105	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
20107	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
48169	100.000000	NaN	100.000000	NaN	NaN	NaN	NaN	NaN	NaN
19177	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
02195	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
42113	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	100.000000	NaN
48275	100.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	100.000000
21219	100.000000	NaN	100.000000	100.000000	NaN	NaN	NaN	NaN	NaN

Figure 28 - Sample data from FIPS x CPT matrix with percentage of providers with average charges between 75th and 90th percentile range (in descending order)

Also, from the sample data shown above, it would be safe to infer that in certain locations where there aren’t that many providers (as indicated by the “Nan” values), there can be a propensity for above average prices.

Given the high number of medical procedures in the data, I 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.

I decided again to map these values in some choropleth maps. The first map includes Alaska and it somewhat confirms that in more remote areas, the average charges are higher:

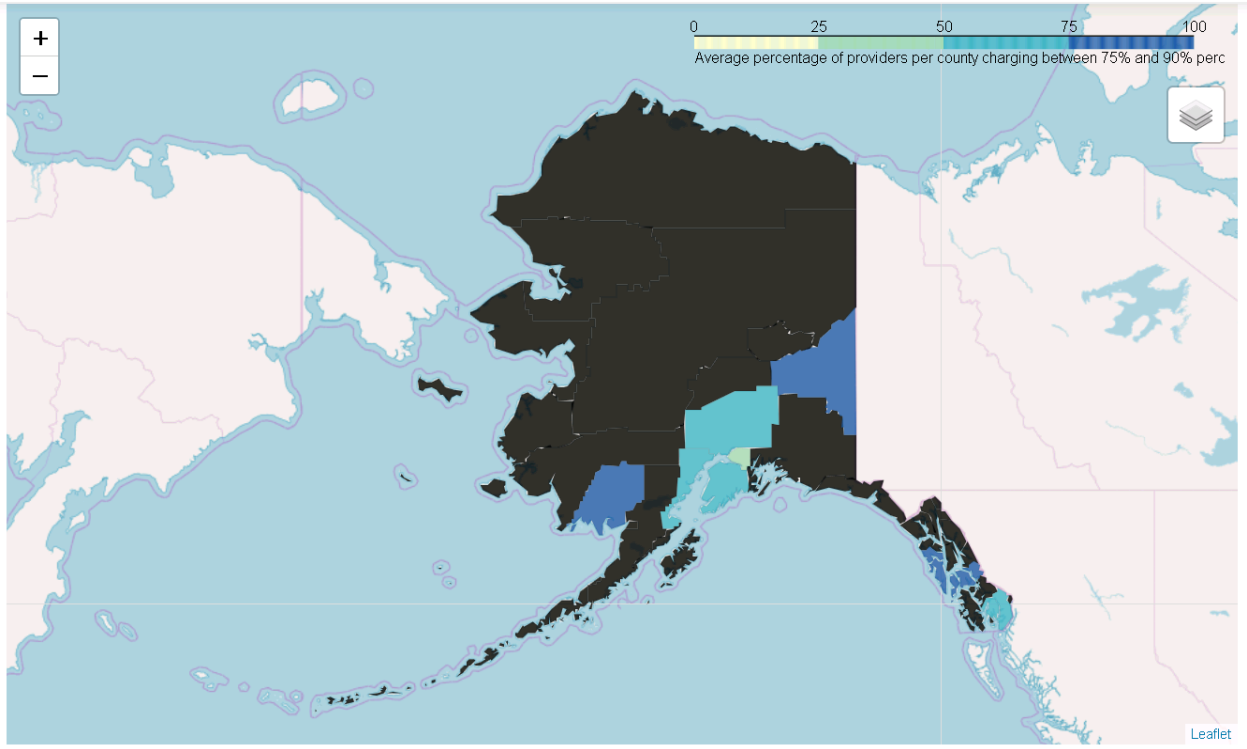


Figure 29 - Choropleth map for average percentage of providers with average charges between 75th and 90th percentile (Alaska)

The second map, includes the Continental United States and it also confirms the same assumption, as the counties with higher percentage of providers charging above average appear to be located mostly in more rural areas, whereas some of the larger metro areas (i.e. New York, Los Angeles, Miami) have a lower percentage:

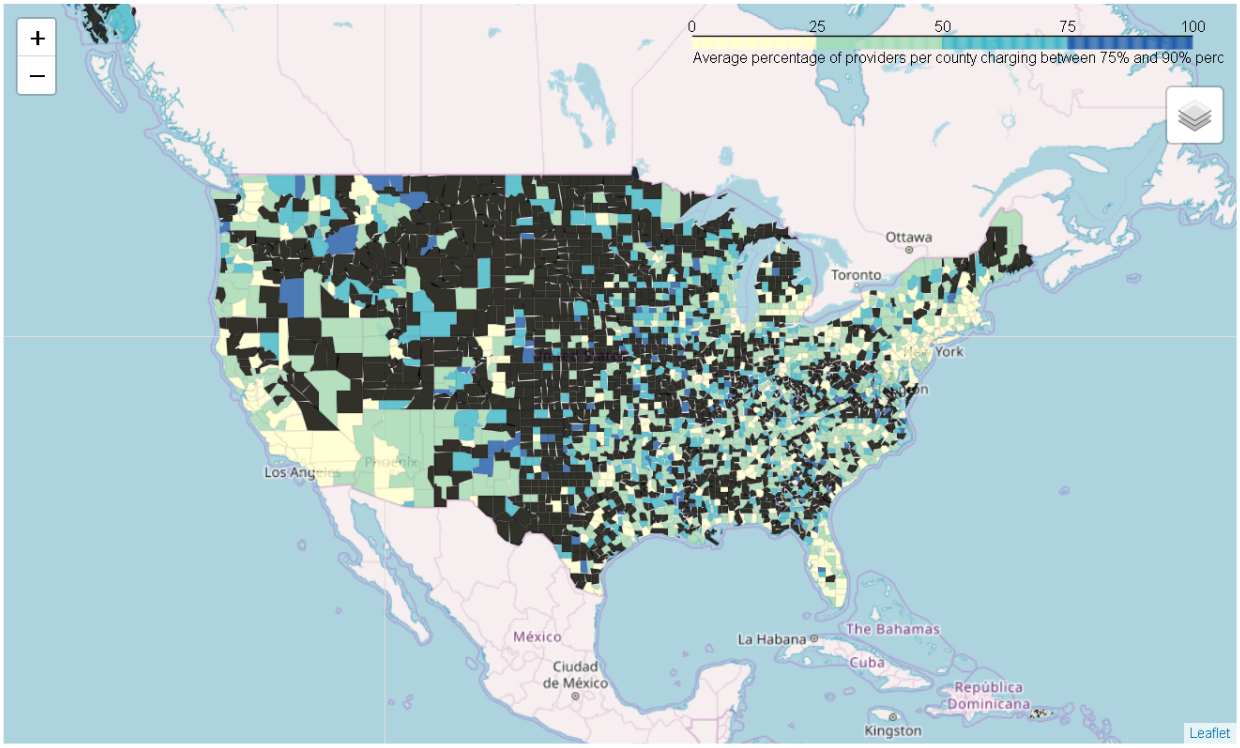


Figure 30 - Choropleth map for average percentage of providers with average charges between 75th and 90th percentile (Continental United States)

Lastly, a few more distribution plots were made between the percentage of providers in a county charging above average prices and the demographic variables we utilized earlier. The results were as follows:

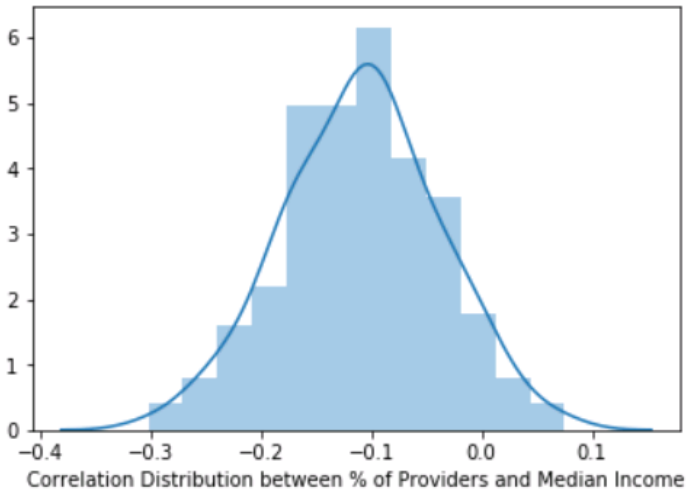


Figure 31 - Plot of correlation between percentage of providers with average charges between 75th and 90th percentile and median income per county

The correlation is almost evenly distributed, although mostly in the small range of negative correlation values. However, the fact that we are dealing with average values, may lend some

confirmation to the fact there is some weight to an inverse relationship between these two factors. This can be due to the fact that, if our assumption is correct, that more rural areas have less providers and also lower median income.

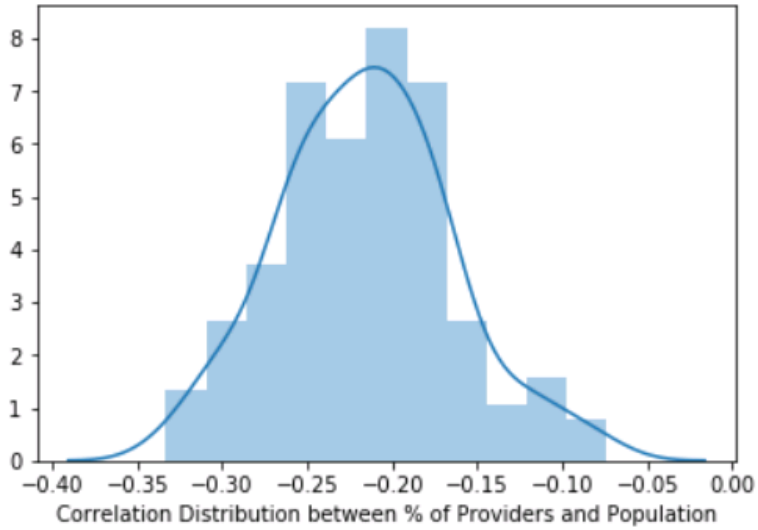


Figure 32 - Plot of correlation between percentage of providers with average charges between 75th and 90th percentile and population per county

The distribution plot for the correlation between the percentage of providers and population size is exclusively in negative territory, showing that there is also an inverse relationship, between higher charges and population size, as we have been presuming based on the available data. These correlation values confirm what we had seen earlier in our matrix of procedures and counties, where a substantial number of counties with above average prices are in areas with few providers and, as one would expected, areas with smaller population sizes.

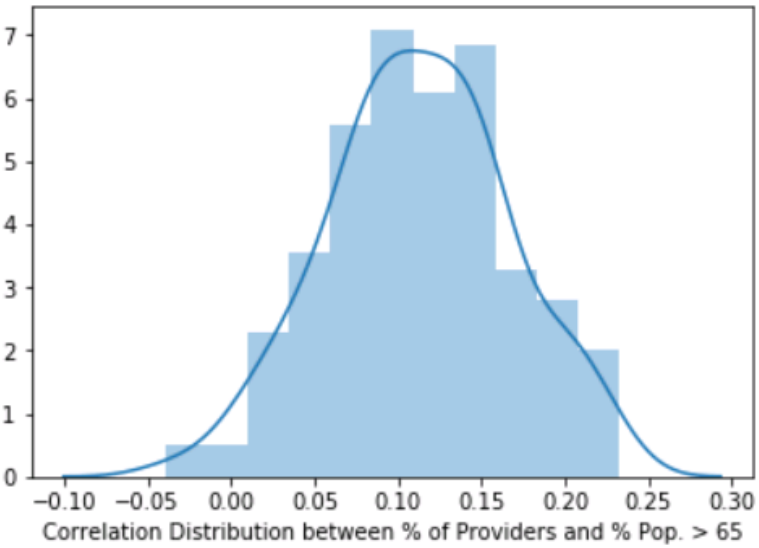


Figure 33 - Plot of correlation between percentage of providers with average charges between 75th and 90th percentile and population per county

Lastly, the correlation between the percentage of providers charging above average prices and the percentage of population over 65 show some positive correlation, meaning that area with retirement age population may show slightly higher charges. However, this factor would have to be analyzed in further detail, as older population can be concentrated in certain areas (i.e. Florida) or can also be spread out in more rural areas.

The last analysis to be performed was to calculate the correlation between the mean percentage of provider per county with above average charges and these same 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

Again, out of the three demographic values, there seems to be a moderate negative correlation between the mean percentage and the population, which indeed confirms our assumptions that less populated areas have higher average charges.

Conclusions & Future work

The Medicare charges dataset was quite interesting and, given its large size in terms of records and data variables, it provided many possible avenues of exploration.

The focus of this capstone project was to try to determine if there was a possible explanation for the variability in submitted charges by various providers for standard medical procedures.

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. Additionally, the fact that the dataset includes data for bulk charges, it introduced some noise which made it more difficult to come to a definite conclusion.

Despite these limitations, and based on the analysis performed, 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 submitted charges than more populated parts of the country.

However, even filtering for extreme values, as it was performed using the 10% and 90% percentile ranges, there is still quite a discrepancy between average charge values at the lower end and those at the higher end. As a consumer of healthcare, and after performing this analysis, I do not find a clear reason for why such discrepancies exist, other than there is not much regulation regarding healthcare costs and providers are pretty much free to charge whatever they like. Also, a lack of competition in rural areas can explain part of the issue, however, this factor should not be used a reason to do so.

One other cause of concern when analysing healthcare charges data is also the possibility of fraud. Unfortunately, fraud with Medicare and other government programs in not uncommon, and making this sort of analysis can be relevant to help spot anomalies and unusual situations in the data. Just as an example, one of the largest Medicare fraud cases in the United States is now in its initial court proceedings[8], involving a total of around $1 billion dollars, while recently, Walgreens settled a fraud case in the amount of $296 million dollars[9]

Lastly, this dataset provided many avenues of exploration and, for future work, it would be interesting to analyse average charges based on the qualifications of the providers in order to understand if there is any sort of relationship between the two variables.

---

[8] Opening statements made in $1B Florida Medicare fraud case - https://wtop.com/national/2019/02/jury-selection-starts-in-1b-florida-medicare-fraud-case/

[9] Walgreens Agrees to $296M Settlement in Healthcare Fraud Cases - https://healthpayerintelligence.com/news/walgreens-agrees-to-296m-settlement-in-healthcare-fraud-cases