Version 2.0

Date: November 29th, 2018

Northeastern University
College *of* Engineering

INFO 7290 DATA WAREHOUSING AND BUSINESS INTELLIGENCE

Presented By:

Akshat Bhatia

Kiran Putra Jagadeswar

Mayuri Salunke

Ruchika Sancheti

TABLE OF CONTENTS

INTRODUCTION 4

CMS HPSA AND LOW_INCOME ZIP CODE DATABASE 5

PRIMARY KEY 6

COLUMNS AND DATA ANALYSIS 6

MEDICARE PROVIDER UTILIZATION AND PAYMENT DATA: PHYSICIANS AND OTHER SUPPLIERS 7

FILE 1: MEDICARE PHYSICIAN AND OTHER SUPPLIERS PUBLIC USE FILE 8

PRIMARY KEY 8

COULMNS AND DATA ANALYSIS 8

SUMMARY FILES 10

OPEN PAYMENTS 11

FILE 1: OPEN PAYMENTS GENERAL 12

PRIMARY KEY 12

COLUMNS AND DATA ANALYSIS 12

FILE 2: OPEN PAYMENTS RESEARCH 12

PRIMARY KEY 12

COLUMNS AND DATA ANALYSIS 12

FILE 3: OPEN PAYMENT OWNERSHIP 13

PRIMARY KEY 13

COLUMNS AND DATA ANALYSIS 13

ZIP CODE DATA (INCOME) 14

FILE 1: ALL STATES WITH ADJUSTED GROSS INCOME 18

PRIMARY KEY DEFINITION 18

FILE 2: ALL STATES WITHOUT ADJUSTED GROSS INCOME 18

PRIMARY KEY DEFINITION 18

STATE FIPS 19

PRIMARY KEY 19

COLUMNS AND DATA ANALYSIS 19

COUNTY FIPS 19

PRIMARY KEY 19

COLUMNS AND DATA ANALYSIS 19

ZIP CODE REFERENCE 20
PRIMARY KEY: 20
COLUMNS AND DATA ANALYSIS: 20
ZIP / FIPS 22
FILE 1: ZIPCTYA 24
FILE 2: ZIPCTYB 24
PRIMARY KEY 24
COLUMNS AND DATA ANALYSIS 24
SSIS STAGING TASK 26
DATA MODEL 27
USE CASES: 29
CASE 1: 29
CASE 2: 34
CASE 3: 40
CASE 4: 43
REFERENCES AND LINKS 45

INTRODUCTION

The project comprises of a dataset build from 8 different file sources from different websites. The aim of the project is to combine these multiple data sources to create a reporting system that can address questions based on low income areas, HPSA open payments across different states in the United states.

The approach is to find the linkage between various datasets to paint a bigger picture and create summary tables that can quickly address the data needs of different users.

Talking about the Healthcare Professional Shortage Area dataset, it has the zip codes of all the cities in the United States with the 2010 census population of the areas with zip code designations.

The zip code designations are designed in a way that they have been categorized in 3 ways, i.e. Low-income areas, Healthcare professional shortage areas, and both areas. From these datasets we can always deduce the areas based on their zip codes and what approach needs to be taken to deal with the issues within these areas.

The dataset Medicare provider utility payments public utility file has very important bits of information. The information which we get from this dataset is National Provider Identification code, which is a unique number assigned to each provider in the united states.

The NPI number can be used in many ways by the health care providers, used in health plans, to coordinate health plan benefits between different providers and in the various electronic patient record systems. The State and county zip files can be used as a linkage between the various files to gain more insight on the Zip codes and their associated regions.

To further link these zip code with Zip+4 codes that enable faster delivery of mails at a zip+4 level the cross-reference Zip files come into picture.

The mailing and delivery system as created by USPS to achieve a faster and accurate mailing system the concepts of ZCTAs are used which can be found in the ZCTA.

These files combined are used to generate use cases and create Visualization to address adhoc data needs of the user.

CMS HPSA AND LOW_INCOME ZIP CODE DATABASE

This Dataset is used and maintained by the Centers for Medicare & Medicaid Services (CMS).

Medicare is type of health coverage provided by the Federal government for people who are 65years or older or have are severely disabled. Medicare is completely based on these two factors and does not take into consideration the economic standing of a person.

Medicaid is a type of health coverage program that is provided to the people with very low incomes.

CMS HPSA & Low – Income ZIP code database

The data gives us information on each Zipcode in the US and designates the Zip into three categories.

1. Low Income Area
2. Health Professional Shortage Area (HPSA)
3. Both Low Income Area and HPSA

The data is sourced from the Centers for Medicare & Medicaid Services (CMS). This dataset is said to be updated daily.

Data Cleaning:

- The original data had repeating ZIP records but with different population count. We did a SUM of the population count for ZIPs which had more than one record.
- After staging the data into the SQL Server Database, the data had four digit ZIPs, States which had Zipcodes starting with '0' was missing the first digit and hence there was only 4 digits. We appended the column to add a 0 in front of Zips which had only four digits.
- We also removed non-numeric characters in the population column so that it will be easier for us to aggregate the column.
- View of the Staging table*: [DWBIMARK].[dbo].[ND_HPSA1]*

	ZIP	City Location name	STATE	Population	FIPS Code	StCtyCode	County Name	Zip Code Designation
1	10002	New York	NY	81305	36061	33420	NEW YORK	Low Income Area/HPSA
2	10003	New York	NY	21039	36061	33420	NEW YORK	HPSA
3	10007	New York	NY	1189	36061	33420	NEW YORK	HPSA
4	10009	New York	NY	61455	36061	33420	NEW YORK	Low Income Area/HPSA
5	10012	New York	NY	6716	36061	33420	NEW YORK	HPSA
6	10013	New York	NY	28035	36061	33420	NEW YORK	Low Income Area/HPSA
7	10018	New York	NY	5402	36061	33420	NEW YORK	Low Income Area
8	01002	Amherst	MA	35361	25015	22080	HAMPSHIRE	Low Income Area
9	10024	New York	NY	5	36061	33420	NEW YORK	HPSA
10	10026	New York	NY	34313	36061	33420	NEW YORK	Low Income Area/HPSA
11	10027	New York	NY	60955	36061	33420	NEW YORK	Low Income Area/HPSA

PRIMARY KEY

No primary key column available in the data set

Generated column: ZIP_Income_ID

COLUMNS AND DATA ANALYSIS

Content: Database of HPSA and Low-Income ZIP Codes for Issuers Subject to the Alternate ECP Standard for the purposes of QHP Certification. This data is showing us if an area is low income area or Health Professional Shortage area or both. There are partial duplicates in multiple rows. The solution to that is to take max (2010 census population) CMS-Center for Medicare services HPSA-Health Professional Shortage Area

Available Columns:

a) ZIP code for the city
b) City/Location of the ZIP code
c) State
d) 2010 Census
e) FIPS code

Business Rules:

- Same city can have multiple ZIP codes
- Zip code designation can be either HPSA, or Low income, or both.
- A city may or may not have same FIPS code
- A city may or may not have same County Code
- Population is variable

MEDICARE PROVIDER UTILIZATION AND PAYMENT DATA: PHYSICIANS AND OTHER SUPPLIERS

This data basically provides information on the services provided by Medicare providers in the US. It has very detailed information on the Providers. The table has the description on the treatment/ service provided by the provider.

Each service has a code called a HCPCS code. HCPCS stands for Healthcare Common Procedure Coding System. These are basically billing codes used in the Medicare and Medicaid industry.

We also see that the might not be representative of a physician's entire practice. People who don't have coverage are also treated by physicians and this is not taken into account in the given data.

We also saw that most of the attribute data type used in the data is of the CHAR data type and column which can be aggregated have a NUM datatype, these columns are given below:

- line_srvc_cnt
- bene_day_srvc_cnt
- average_Medicare_allowed_amt
- stdev_Medicare_allowed_amt1
- average_submitted_chrg_amt
- average_Medicare_payment_amt
- stdev_Medicare_payment_amt1
- average_Medicare_standard_amt

The table has details on the transaction amounts the provider charges, submits and the payment amount. We see that some important columns that can be aggregated in the table are:

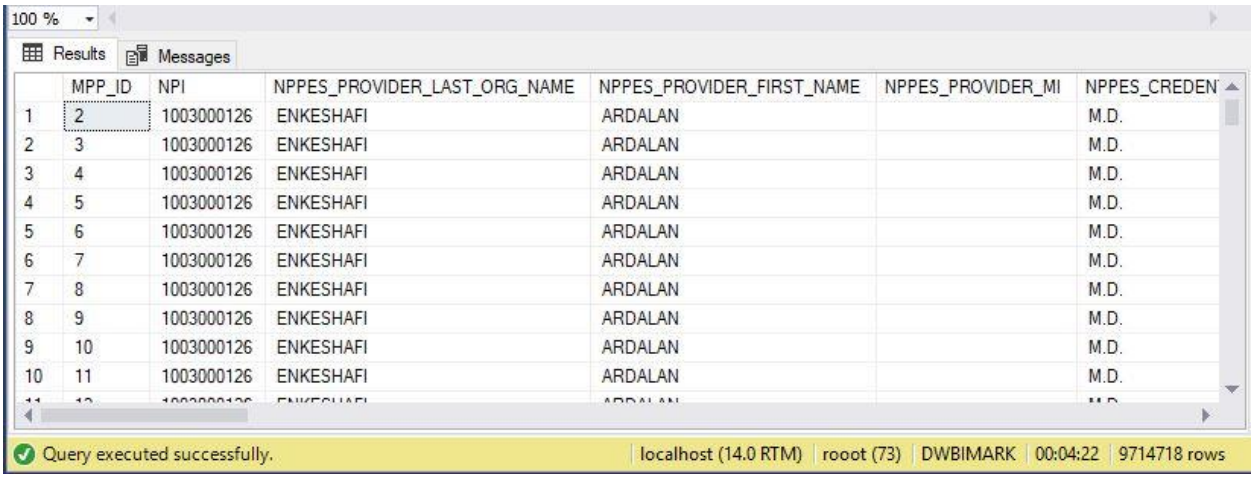
- average_Medicare_allowed_amt – Average of the Medicare allowed amount for the service; this figure is the sum of the amount Medicare pays, the deductible and coinsurance amounts that the beneficiary is responsible for paying, and any amounts that a third party is responsible for paying.

Medicare + deductible + Coinsurance beneficiary is responsible for paying + any amounts that a third party is responsible for paying.

- Average_submitted_chrg_amt – Average of the charges that the provider submitted for the service
- average_Medicare_payment_amt – Average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service.
- average_Medicare_standardized_amt – Average amount that Medicare paid after beneficiary deductible and coinsurance amounts have been deducted for the line item service and after standardization of the Medicare payment has been applied.

Data Cleaning and Staging:

- The data had Zipcodes with its extension, for the purpose of visualization, we extracted only the first 5 digits of the NPPES_PROVIDER_ZIP
- We created a surrogate primary key for this table to identify individual records and query them easily. We named it MPP_ID.
- *[DWBIMARK].[dbo].[MedicalProviderPaymentPUF]:*



FILE 1: MEDICARE PHYSICIAN AND OTHER SUPPLIERS PUBLIC USE FILE

PRIMARY KEY

No primary key column available in the data set

Generated PK =MPP_ID

COULMNS AND DATA ANALYSIS

Physician and Other Supplier Public Use File:

The dataset contains information on the procedures and services provided to the beneficiaries by physicians and other healthcare professionals. The data is based on the part B non-institutional claims submitted from 2012 to 2016.The provider details like name, credentials, gender, address are included from NPPES(National Plan & Provider Enumeration System) which assigns unique identifiers (NPI) to healthcare providers

The data has been aggregated to the following:

1. Provider NPI (National Provider Identifier)
2. HCPCS (Healthcare Common Procedure Coding System) Code
3. Place of Service(facility/non-facility)

Available Columns:

npi
nppes_provider_last_org_name
nppes_provider_first_name
nppes_provider_mi
nppes_credentials
nppes_provider_gender
nppes_entity_code
nppes_provider_street1
nppes_provider_street2
nppes_provider_city
nppes_provider_zip
nppes_provider_state
nppes_provider_country
provider_type
medicare_participation_indicator-Identifies whether provider participates in Medicare
place_of_service-Identifies if the service is within facility(F) or not (O)
hcpcs_code-Code used to identify the service offered by the provider
hcpcs_description-
hcpcs_drug_indicator-Identifies if the HCPCS code for the specific service is listed on Medicare Part B Drug Average Sales Price (ASP) File
line_srvc_cnt-Number of services provided

bene_unique_cnt-Count of distinct beneficiaries receiving the service
average_Medicare_allowed_amt-sum of the amount Medicare pays, the deductible and coinsurance amounts that the beneficiary is responsible for paying, and any amounts that a third party is responsible for paying.
average_submitted_chrg_amt – Average of the charges that the provider submitted for the service.
average_Medicare_payment_amt – Average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service
average_Medicare_standardized_amt – Average amount that Medicare paid after beneficiary deductible and coinsurance amounts have been deducted for the line item service and after standardization of the Medicare payment has been applied

Data limitations:

1. Entire data is based on the claims submitted and hence does not represent physician's entire practice.
2. It does not have information on the patients who do not have Medicare i.e patients covered under other federal programs.
3. The information does not indicate quality of service provided by individual physicians

SUMMARY FILES

- Medicare National HCPCS aggregate CY2016
- Medicare Physician and other supplier NPI aggregate
- Medicare State HCPCS aggregate CY2016

OPEN PAYMENTS

Open payments program promotes a more reliable healthcare system by making financial relationship between manufacturers and group purchasing organizations (GPOs) and health care providers (physicians and teaching hospitals) available to public.

- The files have unique Record_IDs which will be treated as its primary key.
- The Zip code column for the ownership data is different from the zipcode column in the general table. The former has an extension with its zip.

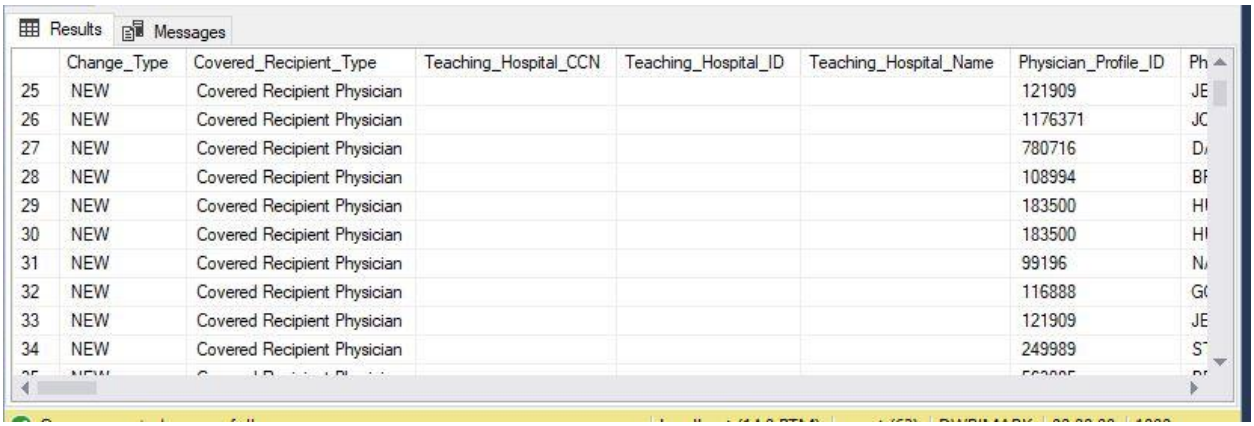
This data is sourced from the CMS which is a program created by the ACA. This data consists of transactions that the group purchasing organizations (GPO) and applicable manufacture make to physicians and teaching hospitals.

The data has information on the Physician's specialty and the payment amount between the payer and the recipient.

The three types of payments are:

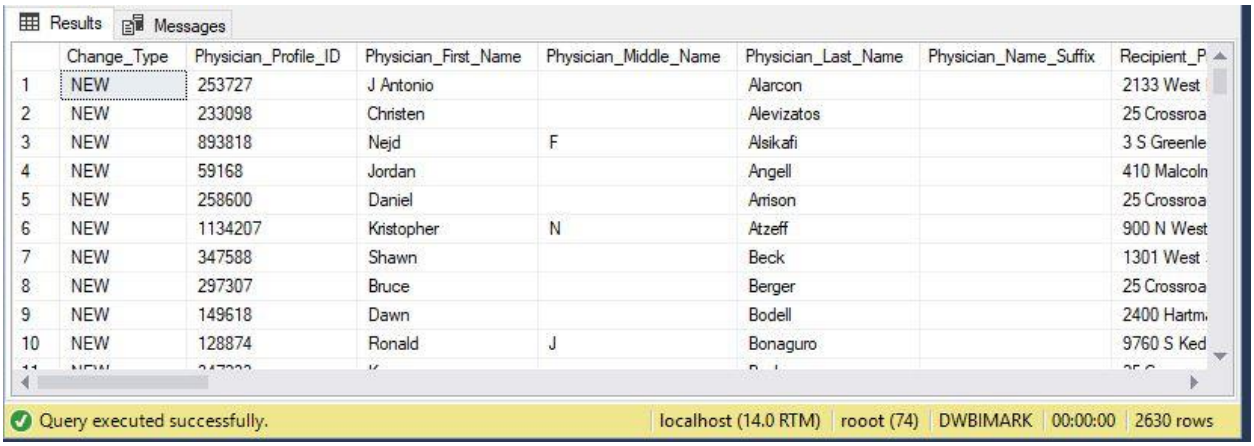
1. General Payment:
    a. General payment records provide the total value of general payments or other transfers of value to a particular recipient for a particular date.
    b. Each record includes identifying information for the applicable manufacturer or applicable GPO who made the payment, and identifying information for the recipient.
2. Physician Ownership: Physician ownership records provide information on physician ownership or investment interests in an applicable manufacturer or applicable GPO.
3. Research Payment: Research payment records provide the total value of a payment or other transfer of value made for research purposes to a particular recipient for a particular date.

[DWBIMARK].[dbo].[OP_General_Payments]:



	Change_Type	Covered_Recipient_Type	Teaching_Hospital_CCN	Teaching_Hospital_ID	Teaching_Hospital_Name	Physician_Profile_ID	Ph
25	NEW	Covered Recipient Physician				121909	JE
26	NEW	Covered Recipient Physician				1176371	JC
27	NEW	Covered Recipient Physician				780716	D/
28	NEW	Covered Recipient Physician				108994	BI
29	NEW	Covered Recipient Physician				183500	HI
30	NEW	Covered Recipient Physician				183500	HI
31	NEW	Covered Recipient Physician				99196	N/
32	NEW	Covered Recipient Physician				116888	G(
33	NEW	Covered Recipient Physician				121909	JE
34	NEW	Covered Recipient Physician				249989	S'

[DWBIMARK].[dbo].[OP_Owenership]:



	Change_Type	Physician_Profile_ID	Physician_First_Name	Physician_Middle_Name	Physician_Last_Name	Physician_Name_Suffix	Recipient_P
1	NEW	253727	J Antonio		Alarcon		2133 West
2	NEW	233098	Christen		Alevizatos		25 Crossroa
3	NEW	893818	Nejd	F	Alsikafi		3 S Greenle
4	NEW	59168	Jordan		Angell		410 Malcoln
5	NEW	258600	Daniel		Arrison		25 Crossroa
6	NEW	1134207	Kristopher	N	Atzeff		900 N West
7	NEW	347588	Shawn		Beck		1301 West
8	NEW	297307	Bruce		Berger		25 Crossroa
9	NEW	149618	Dawn		Bodell		2400 Hartm
10	NEW	128874	Ronald	J	Bonaguro		9760 S Ked

Query executed successfully. | localhost (14.0 RTM) | rooot (74) | DWBIMARK | 00:00:00 | 2630 rows

FILE 1: OPEN PAYMENTS GENERAL

PRIMARY KEY

RECORD_ID is the unique key in the data set

COLUMNS AND DATA ANALYSIS

General payment records provide the total value of general payments to a recipient for a date. Each record includes identifying information for the manufacturer or GPO who made the payment and identifying information for the recipient.

FILE 2: OPEN PAYMENTS RESEARCH

PRIMARY KEY

RECORD_ID is the unique key in the data set

COLUMNS AND DATA ANALYSIS

Research payment records provide the total value of a payment or other transfer of value made for research purposes to a particular recipient for a particular date. Each record includes identifying information for the manufacturer or GPO who

made the payment, as well as identifying information for the recipient. Information is also provided for up to five physician principal investigators associated with the payment.

FILE 3: OPEN PAYMENT OWNERSHIP

PRIMARY KEY

RECORD_ID is the unique key in the data set

COLUMNS AND DATA ANALYSIS

Physician ownership records provide information on physician ownership or investment interests in an applicable manufacturer or GPO.

ZIP CODE DATA (INCOME)

The data is based on administrative records of individual income tax returns (Forms 1040) from the Internal Revenue Service (IRS) Individual Master File (IMF) system

- State codes are based on the zip codes
- ZIP codes with less than 100 returns and identified as a single building or nonresidential ZIP code were categorized as **"other" (99999).**
- Income and tax items with less than 20 returns for a particular AGI class were combined with another AGI class within the same ZIP Code. Collapsed AGI classes are identified with a double asterisk (**).
- All number of returns variables have been rounded to the nearest 10.

This data has information on income of people living in specific zipcode regions and various other parameters related to people in that zipcode. The income range is given by an adjusted gross income (AGI_STUB) as follows:

- 1 = $1 under $25,000
- 2 = $25,000 under $50,000
- 3 = $50,000 under $75,000
- 4 = $75,000 under $100,000
- 5 = $100,000 under $200,000
- 6 = $200,000 or more

The attributes used in the data set can be used to analyze a lot of information on the Medicare spending trends in the United States.

Medicare in the US:

- There are two government-sponsored health insurance programs available to Americans. Medicare primarily covers adults 65 and over, while Medicaid covers low-income individuals and families. Medicare is funded by the federal government.
- Medicaid covers low-income individuals and families. Medicaid is jointly funded by the states, so eligibility for the program varies. Medicare eligibility, conversely, is standardized across the nation.

- Medicare is the federal health insurance program for Americans and permanent U.S. citizens 65 and over. Younger Americans with certain disabilities or illnesses, including Lou Gehrig's disease and terminal kidney failure, are also eligible for Medicare. However, Medicare is primarily thought of as a social health insurance program designed to help retired Americans pay their medical expenses. It's not free. Medicare is funded by taxpayer dollars and premiums paid by beneficiaries.

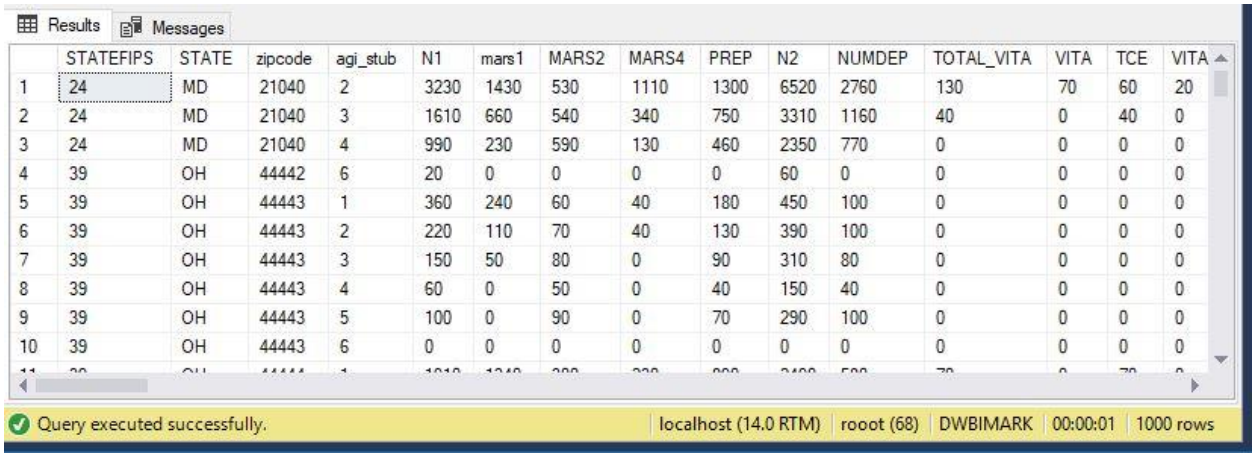
Income does not affect people's Medicare eligibility. Medicare is run by the Centers for Medicare and Medicaid Services (CMS), from which most of our data in this project is sourced from.

The columns that we thought would be useful for our analysis are:

- AGI_STUB - Size of adjusted gross income
- N1 - Number of returns
- N2 – Number of exceptions
- NUMDEP -Number of Dependents
- ELDERLY – Number of elderly returns
- A00100 - Adjust gross income
- N02650 -Number of returns with total income
- A02650-Total income amount
- N02300-Number of returns with unemployment compensation
- A02300-Unemployment compensation amount [6]
- N03270-Number of returns with Self-employed health insurance deduction
- A03270-Self-employed health insurance deduction amount
- N04800-Number of returns with taxable income
- A04800-Taxable income amount
- N07180-Number of returns with child and dependent care credit
- A07180-Child and dependent care credit amount
- N09750-Number of returns with health care individual responsibility payment
- A09750-Health care individual responsibility payment amount
- N85530-Number of returns with additional Medicare tax
- A85530-Additional Medicare tax amount

There were two data sets in the link. One of the data sets had the income data according to each AGI STUB and the other data had an aggregated data on the income in each zipcode excluding the AGI. It had one record per zip.
The data was staged as follows:

[DWBIMARK].[dbo].[All_agi]:



Results | Messages

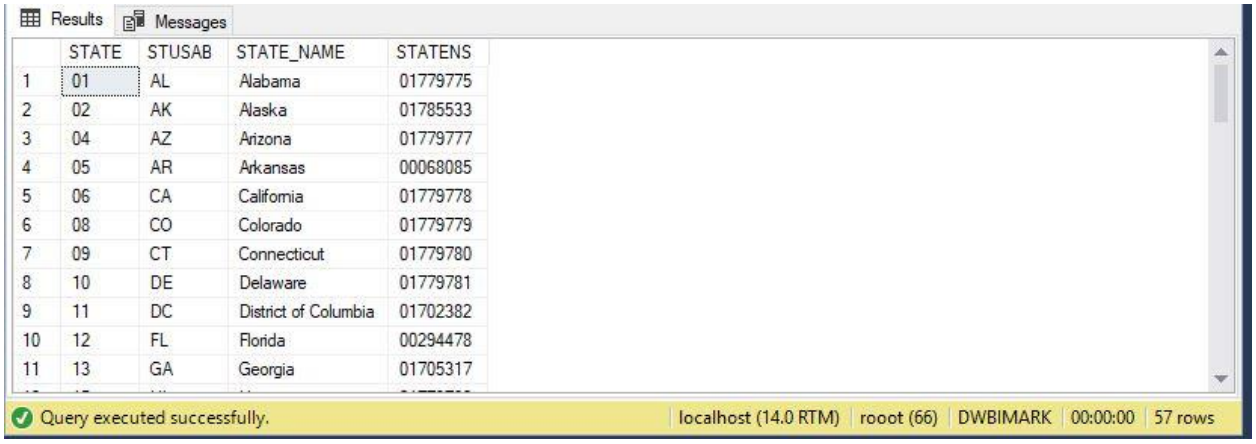
	STATEFIPS	STATE	zipcode	agi_stub	N1	mars1	MARS2	MARS4	PREP	N2	NUMDEP	TOTAL_VITA	VITA	TCE	VITA
1	24	MD	21040	2	3230	1430	530	1110	1300	6520	2760	130	70	60	20
2	24	MD	21040	3	1610	660	540	340	750	3310	1160	40	0	40	0
3	24	MD	21040	4	990	230	590	130	460	2350	770	0	0	0	0
4	39	OH	44442	6	20	0	0	0	0	60	0	0	0	0	0
5	39	OH	44443	1	360	240	60	40	180	450	100	0	0	0	0
6	39	OH	44443	2	220	110	70	40	130	390	100	0	0	0	0
7	39	OH	44443	3	150	50	80	0	90	310	80	0	0	0	0
8	39	OH	44443	4	60	0	50	0	40	150	40	0	0	0	0
9	39	OH	44443	5	100	0	90	0	70	290	100	0	0	0	0
10	39	OH	44443	6	0	0	0	0	0	0	0	0	0	0	0

Query executed successfully. | localhost (14.0 RTM) | rooot (68) | DWBIMARK | 00:00:01 | 1000 rows

State and County FIPS:The dataset contains codes to identify States, the District of Columbia, Puerto Rico, and the Insular Areas of the United States. The columns contain 2-digit state.The county data set consists of state keys as STATEFP and a COUNTYFP for a county in a state. A five digit code ([STATEFP] + [COUNTYFP]) called FIPS code is generated and is used in another table in the dataset. This column can be used to link the two tables.

Below are details on the staging of these two datasets.

[DWBIMARK].[dbo].[State_FIPS]:

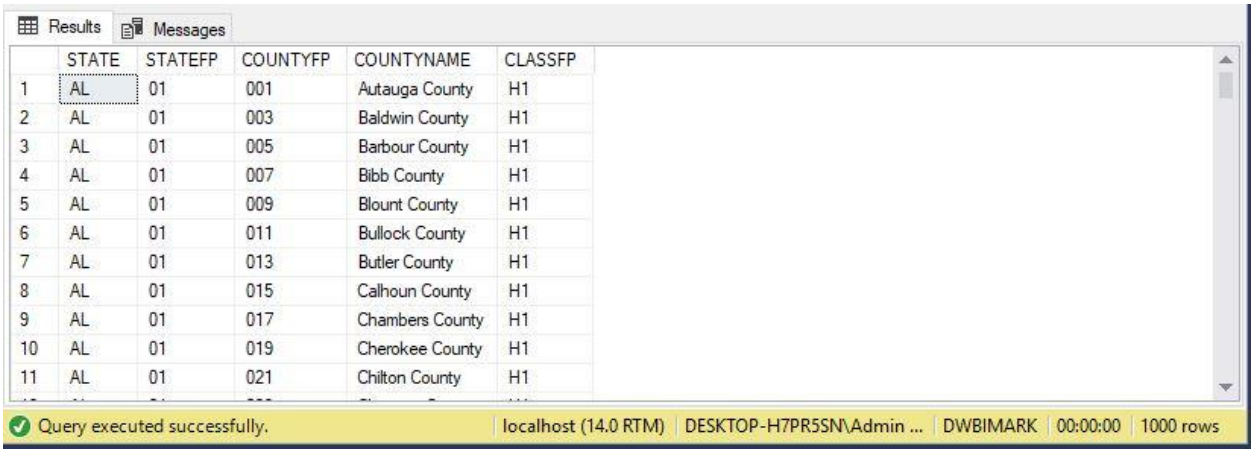


Results | Messages

	STATE	STUSAB	STATE_NAME	STATENS
1	01	AL	Alabama	01779775
2	02	AK	Alaska	01785533
3	04	AZ	Arizona	01779777
4	05	AR	Arkansas	00068085
5	06	CA	California	01779778
6	08	CO	Colorado	01779779
7	09	CT	Connecticut	01779780
8	10	DE	Delaware	01779781
9	11	DC	District of Columbia	01702382
10	12	FL	Florida	00294478
11	13	GA	Georgia	01705317

Query executed successfully. | localhost (14.0 RTM) | rooot (66) | DWBIMARK | 00:00:00 | 57 rows

[DWBIMARK].[dbo].[County_FIPS]:



Results | Messages

	STATE	STATEFP	COUNTYFP	COUNTYNAME	CLASSFP
1	AL	01	001	Autauga County	H1
2	AL	01	003	Baldwin County	H1
3	AL	01	005	Barbour County	H1
4	AL	01	007	Bibb County	H1
5	AL	01	009	Blount County	H1
6	AL	01	011	Bullock County	H1
7	AL	01	013	Butler County	H1
8	AL	01	015	Calhoun County	H1
9	AL	01	017	Chambers County	H1
10	AL	01	019	Cherokee County	H1
11	AL	01	021	Chilton County	H1

Query executed successfully. | localhost (14.0 RTM) | DESKTOP-H7PR5SN\Admin ... | DWBIMARK | 00:00:00 | 1000 rows

Excluded Data:

- Tax returns filed without a ZIP code and returns filed with a ZIP code that did not match the State code shown on the return
- Tax returns filed using Army Post Office (APO) and Fleet Post Office addresses, foreign addresses, and addresses in Puerto Rico, Guam, Virgin Islands, American Samoa, Marshall Islands, Northern Marianas, and Palau.
- items with less than 20 returns within a ZIP code.
- tax returns representing a specified percentage of the total of any particular cell

Available Columns:

STATEFIPS-The State Federal Information Processing System (FIPS) code

STATE

ZIPCODE

AGI_STUB-Size of adjusted gross income

N1-Number of returns

MARS1-Number of single returns

MARS2-Number of joint returns

MARS4-Number of head of household returns

PREP-Number of returns with paid preparer's signature

N2-Number of exemptions

NUMDEP-Number of dependents

TOTAL_VITA-Total number of volunteer prepared returns

FILE 1: ALL STATES WITH ADJUSTED GROSS INCOME

PRIMARY KEY DEFINITION

Combination of ZIP_CODE and AGI_STUB

FILE 2: ALL STATES WITHOUT ADJUSTED GROSS INCOME

PRIMARY KEY DEFINITION

Combination of ZIP_CODE and AGI_STUB

STATE FIPS

PRIMARY KEY

"State" column is unique to each record and can be used as a primary key

COLUMNS AND DATA ANALYSIS

State Fips:

The dataset contains codes to identify States, the District of Columbia, Puerto Rico, and the Insular Areas of the United States. The columns contain 2-digit state FIPS code, postal abbreviation and 8-digit Geographic Names Information Systems Identifier (GNISID)

Geographic Names Information System Identifier (GNISID) is the primary key

COUNTY FIPS

PRIMARY KEY

No primary key column available. A composite key of StateFp and CountyFp together is the new primary key

COLUMNS AND DATA ANALYSIS

The objective of state and county fips to improve the utilization of data resources of the Federal Government and avoid unnecessary duplications and incompatibilities in data processing.

County Fips:

The dataset contains state postal code, state fips, county fips, county name and fips class code

Available Columns:

1.State Describes the state postal Code

2.StateFp: State FIPS Code

3.CountyFp: County FIPS code

4.CountyName: Name of the county

5.ClassFp: Class FIPS Code

FIPS Class Codes Definitions:

H1: identifies an active county or statistically equivalent entity that does not qualify under subclass C7 or H6.

H4: identifies a legally defined inactive or non-functioning county or statistically equivalent entity that does not qualify under subclass H6.

H5: identifies census areas in Alaska, a statistical county equivalent entity.

H6: identifies a county or statistically equivalent entity that is a really coextensive or governmentally consolidated with an incorporated place, part of an incorporated place, or a consolidated city.

C7: identifies an incorporated place that is an independent city; that is, it also serves as a county equivalent because it is not part of any county, and a minor civil division (MCD) equivalent because it is not part of any MCD

ZIP CODE REFERENCE

PRIMARY KEY:

"ZIP" column is unique and can be used as a primary key column

COLUMNS AND DATA ANALYSIS:

This is a database on the US zip codes. The documentation states that it has only one entry per zip code which makes zip code the primary key column for the data.

Not all records contain county_fips.

(ZCTA) zip code tabulation area: US ZIP codes were created by the USPS for mail delivery. This means that they don't always cover a continuous geographical area. What's convenient for delivery purposes (e.g. delivering on one side of the street except for large businesses) doesn't always lend itself to mapping or statistics. Since zip codes are often used for these purposes, the US Census Bureau calculates the approximate boundaries of zip code areas. These areas are called Zip Code Tabulation Areas.

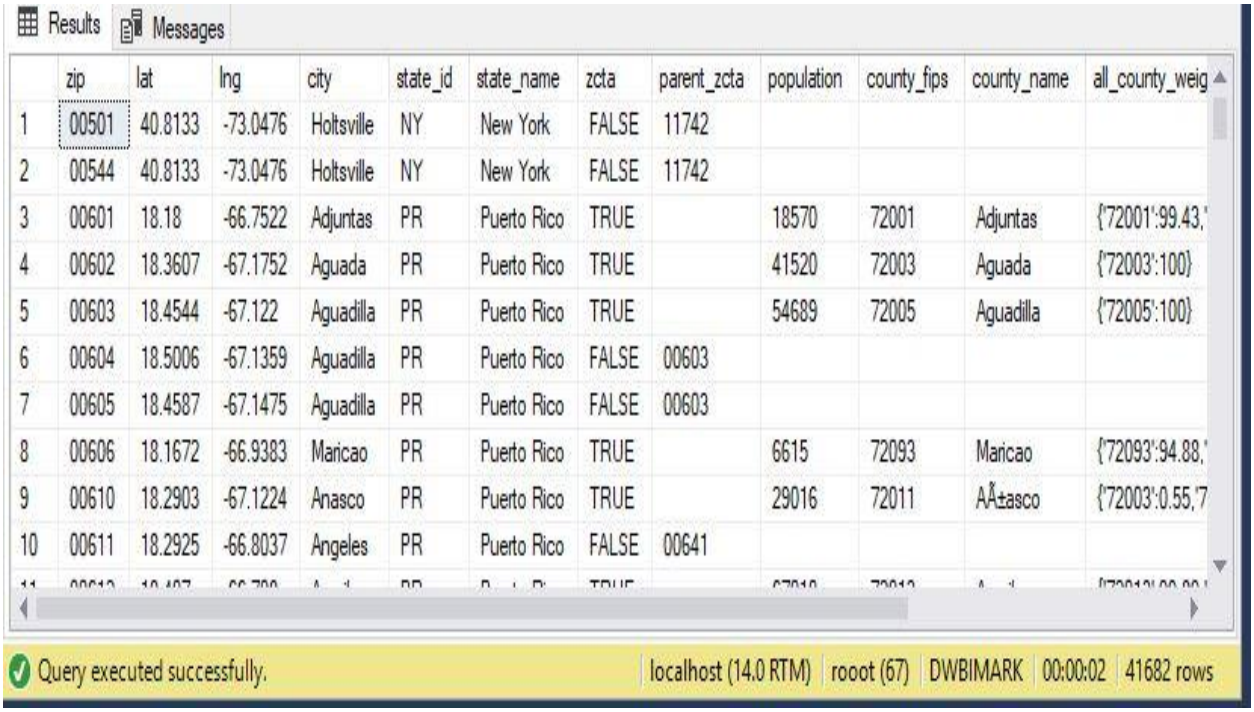
The Parent ZCTA column is filled only is ZCTA is false.
US Zipcode Reference:

For example, if the Zipcode of backbay is 02115 then this would be the parent ZCTA and under this ZIpcode there would be multiple child ZCTAs for this region.

This data can be used with the [MedicalProviderPayment1PUF] table to pull out information on providers in the military areas and create more insights on them.

The staging table for this data set is shown below:

[DWBIMARK].[dbo].[USZIPref]:



Results | Messages

	zip	lat	lng	city	state_id	state_name	zcta	parent_zcta	population	county_fips	county_name	all_county_weig
1	00501	40.8133	-73.0476	Holtsville	NY	New York	FALSE	11742				
2	00544	40.8133	-73.0476	Holtsville	NY	New York	FALSE	11742				
3	00601	18.18	-66.7522	Adjuntas	PR	Puerto Rico	TRUE		18570	72001	Adjuntas	{'72001':99.43,'
4	00602	18.3607	-67.1752	Aguada	PR	Puerto Rico	TRUE		41520	72003	Aguada	{'72003':100}
5	00603	18.4544	-67.122	Aguadilla	PR	Puerto Rico	TRUE		54689	72005	Aguadilla	{'72005':100}
6	00604	18.5006	-67.1359	Aguadilla	PR	Puerto Rico	FALSE	00603				
7	00605	18.4587	-67.1475	Aguadilla	PR	Puerto Rico	FALSE	00603				
8	00606	18.1672	-66.9383	Maricao	PR	Puerto Rico	TRUE		6615	72093	Maricao	{'72093':94.88,'
9	00610	18.2903	-67.1224	Anasco	PR	Puerto Rico	TRUE		29016	72011	AÃ±asco	{'72003':0.55,'7
10	00611	18.2925	-66.8037	Angeles	PR	Puerto Rico	FALSE	00641				

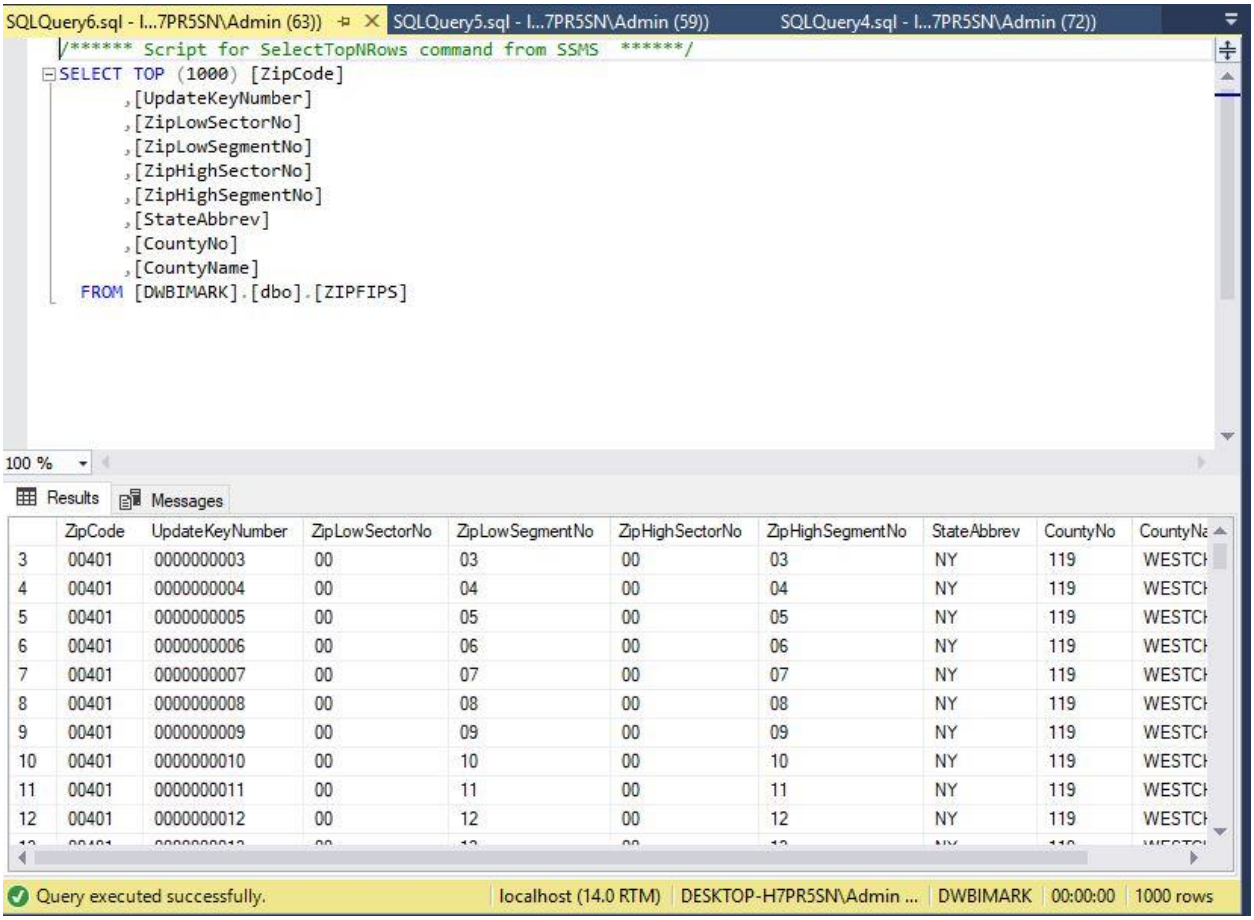
Query executed successfully. | localhost (14.0 RTM) | rooot (67) | DWBIMARK | 00:00:02 | 41682 rows

ZIP / FIPS

This data set initially consisted of ten separate files. The was not file delimited and was in a single field. The file was then studied and divided into columns and then the separate files were appended in a data frame in Python. The entire data set was brought out as a single FlatFile.

The data was then sent to SSIS where we used the Ragged right function to skim the data using its length and then put it into its respective columns. The columns were then named. The final staging table is shown below:

[DWBIMARK].[dbo].[ZIPFIPS]:



jupyter hpsa_akshat Last Checkpoint: 18 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Code

In [1]:
```
import pandas as pd
import numpy as np
hpsa = pd.read_csv("1.csv")
```

In [2]:
```
hpsa.head()
```

Out[2]:

	ZIP	City/Location Name	STATE	2010 Census Population	FIPS Code	StCtyCode	County Name	Zip Code Designation
0	99546	Adak	AK	* 525	2016	2016	ALEUTIANS WEST	Low Income Area/HPSA
1	99551	Akiachak	AK	* 104	2050	2050	BETHEL	Low Income Area/HPSA
2	99553	Akutan	AK	* 1,027	2013	2013	ALEUTIANS EAST	Low Income Area/HPSA
3	99554	Alakanuk	AK	* 1,439	2270	2270	WADE HAMPTON	Low Income Area/HPSA
4	99555	Aleknagik	AK	* 219	2070	2070	DILLINGHAM	Low Income Area/HPSA

In [15]:
```
#df['E'] = df['B'].str.replace('\W', '')

#hpsa=hpsa.replace('\*','',regex=True).astype(float)

#hpsa['2010 Census Population'] = hpsa['2010 Census Population'].map(lambda x: x.lstrip('*')

hpsa['2010 Census Population']=hpsa['2010 Census Population'].replace('\*','',regex=True)

hpsa.head()
```

Out[15]:

	ZIP	City/Location Name	STATE	2010 Census Population	FIPS Code	StCtyCode	County Name	Zip Code Designation
0	99546	Adak	AK	525	2016	2016	ALEUTIANS WEST	Low Income Area/HPSA
1	99551	Akiachak	AK	104	2050	2050	BETHEL	Low Income Area/HPSA
2	99553	Akutan	AK	1,027	2013	2013	ALEUTIANS EAST	Low Income Area/HPSA
3	99554	Alakanuk	AK	1,439	2270	2270	WADE HAMPTON	Low Income Area/HPSA
4	99555	Aleknagik	AK	219	2070	2070	DILLINGHAM	Low Income Area/HPSA

In [21]:
```
hpsa.to_csv('hpsa.csv')
```

jupyter Zipcty_Akshat Last Checkpoint: 18 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Code

In [1]:
```
import pandas as pd
import numpy as np
import glob
```

In [ ]:
```
#Zipcty6 = pd.read_csv('zipcty6.txt', sep=" ")
```

In [ ]:
```
#Zipcty6.head()
```

In [ ]:
```
#zipcty6= pd.read_csv('zipcty6.txt', sep=" ", header=None, names=["a"])
#zipcty6.head()
```

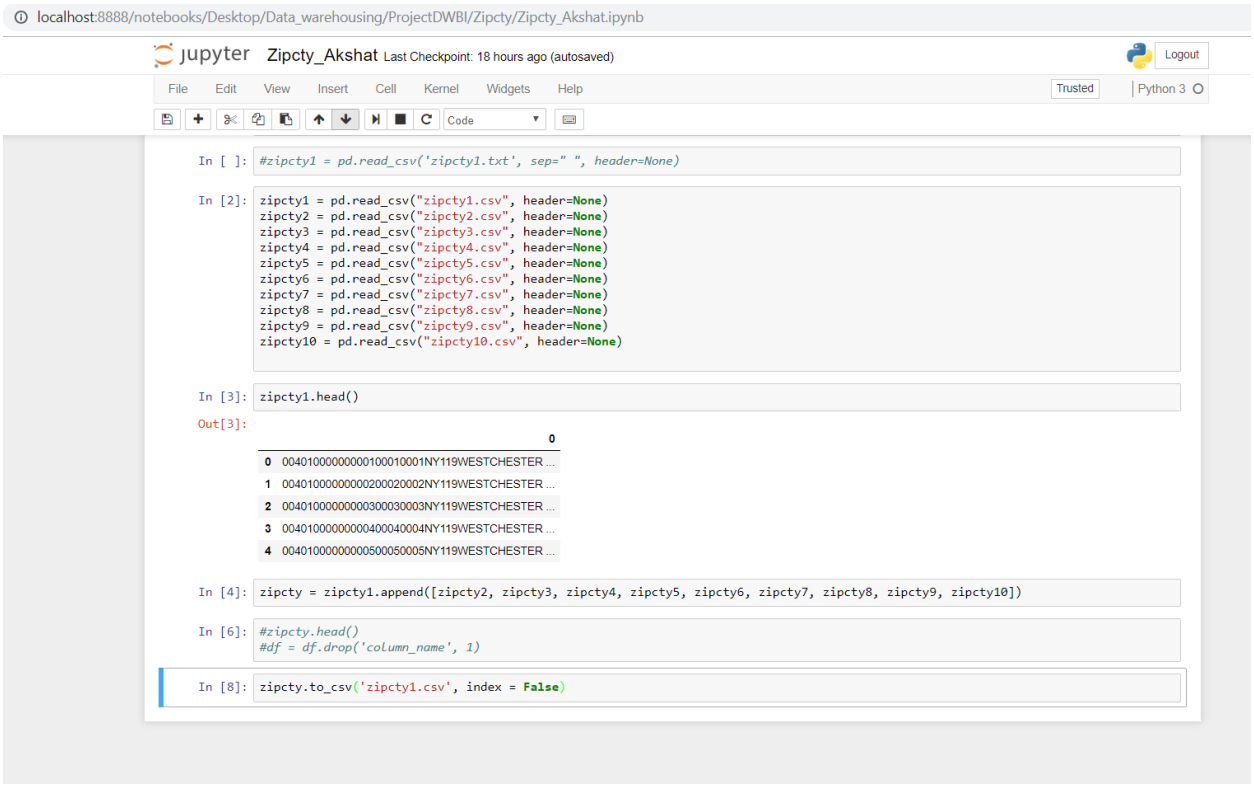
In [ ]:
```
#data = pd.read_csv("filename.csv")

#path =r"C:\Users\aksha\Desktop\Data_warehousing\ProjectDWBI\Zipcty"
#allFiles = glob.glob(path + "/*.csv")
```

In [ ]:
```
#(list_ = []

#for file_ in allFiles:
    #df = pd.read_csv(file_,index_col=None, header=0)
    #list_.append(df)

#frame = pd.concat(list_, axis = 0, ignore_index = True))
```

In [ ]:
```
#frame.head()
```

In [ ]:
```
#frame.to_csv(frame.csv)
```

In [ ]:
```
#zipcty1 = pd.read_csv('zipcty1.txt', sep=" ", header=None)
```



FILE 1: ZIPCTYA

The zipped folder contains files 1 to 5 comprises of zip codes in a serial manner

FILE 2: ZIPCTYB

The zipped folder contains files 6 to 10 comprises of zip codes in a serial manner

PRIMARY KEY

"UPDATE_KEY_NUMBER" is the unique column and will be used as a primary key column.

COLUMNS AND DATA ANALYSIS

ZIP codes are the regular five-digit codes that are used regularly. In order to facilitate faster delivery of mails a new system called then Zip+4 code was introduced.

They indicate delivery routes for mail delivery system.

The advantages of ZIP+4 codes are:

1. They require validation and hence ensure that the address is a real time address

2. Improve the speed of delivery sometimes up to two business days

Available Columns:

column 1: Zip code

column 2: Update key number

This field contains a number that uniquely identifies a record on the County Cross-Reference File.

column 3: Zip add on - LOW/HIGH

Zip add on number has the logical length 4 that comprises of a zip segment number and zip sector number.

Zip codes associated with non delivery area will have a blank zip sector number.

Zip add on HIGH no :ZIP ADD ON HIGH NO is the high-end ZIP ADD ON in a range of ZIP+4 Code ADD-ONs.

Zip add on LOW no :ZIP ADD ON HIGH NO is the low-end ZIP ADD ON in a range of ZIP+4 Code ADD-ONs.

column 4: State abbreviation

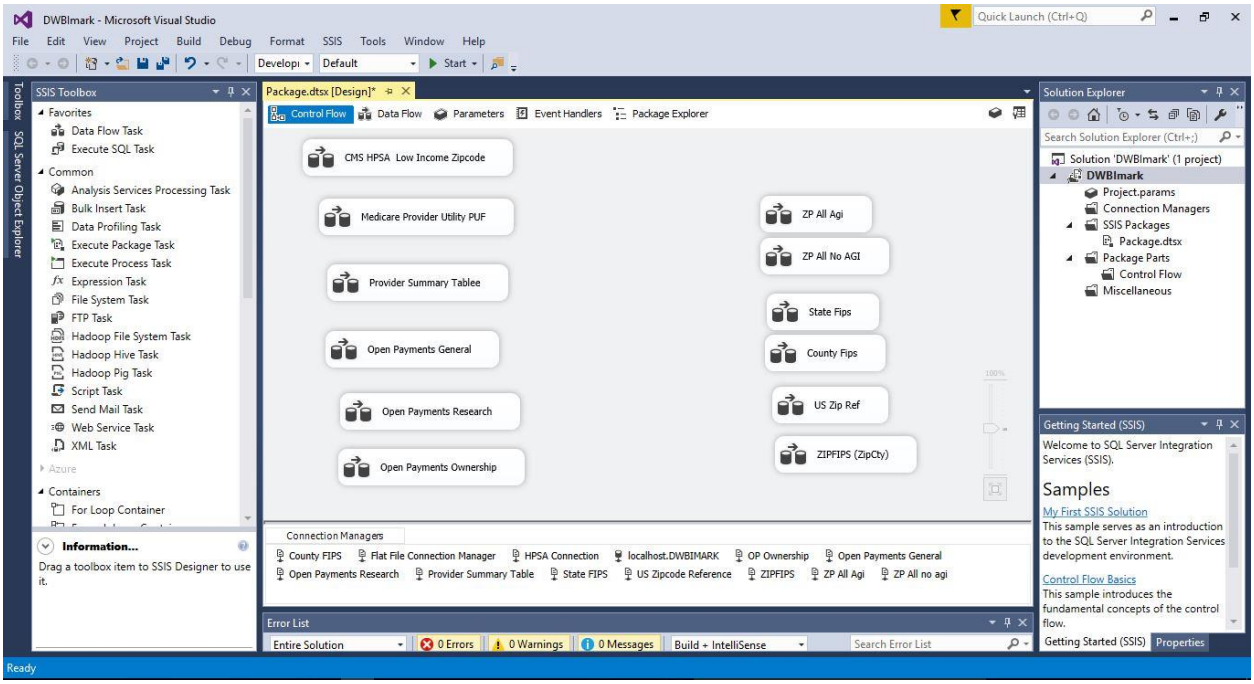
column 5: County Number

column 6: County Name

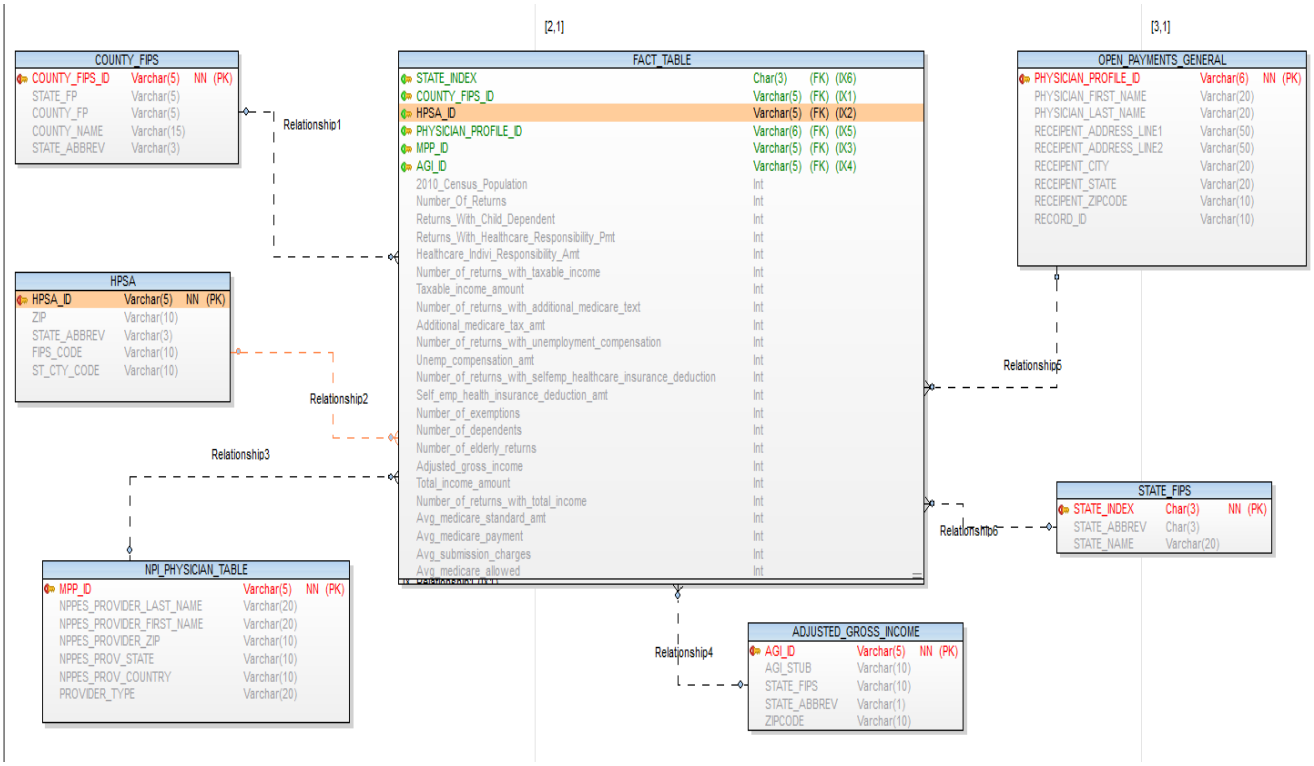
SSIS STAGING TASK

The entire staging process was implemented using SQL Server Integration Services , Notepad++, Excel and Python.

- SSIS staging tasks



DATA MODEL



1. **COUNTY_FIPS**: FIPS county code. The Federal Information Processing Standard is a five-digit Federal Information Processing Standards code which uniquely identified counties and county equivalents in the United States, certain U.S. possessions, and certain freely associated states. The Census Bureau decided that, based on decades of using the terminology FIPS to describe its codes, it would continue to use the FIPS name for its updated codes, where FIPS now stood for FIP "Specification", since there no longer existed an official FIP "Standard".
Keys Used: COUNTY_FIPS_ID: This is a surrogate key created for the table County_fips so that it can be joined with the fact table so that the AGI income data can be retrieved from fact table and use case can be answered.

2. **HPSA**: A Health Professional Shortage Area (**HPSA**) is a geographic area, population, or facility with a shortage of primary care, dental, or mental health providers and services.
Keys Used: HPSA table has the primary key as HPSA_ID which has been used to fetch data from the fact table as in the relationship of different zip-code designations like low income zone or health professional shortage areas with the adjusted gross income zone.

3. **NPI_PHYSICIAN_TABLE**: This table talks about the details of physicians and their treatments. This table has the NPI codes assigned to the healthcare providers so that they can be uniquely identified.
   Keys used: MPP_ID has been used as the primary key in this. This is a surrogate key which is used to fetch results from the fact table like population and income zones.

4. **Open_Payments_general**: **Open Payments** is a federal program that collects and makes information public about financial relationships between the health care industry, physicians, and teaching hospitals.
   Keys used: PHYSICIAN_PROFILE_ID: This key is sued to fetch the data about physician and his services and is linked with the fact table with the zip code geographical data so that the state based results can be fetched.
5. **ADJUSTED_GROSS_INCOME**: Adjusted gross income database is used to locate people based on their geographical locations and their income zone.
   Keys used: AGI_ID has been used here in this table as a primary key to fetch data from the fact table and to gain insights on returns and taxation system based on the geographical locations.

6. STATE_FPS: This table is providing all the state based information like state code, state abbreviations and state names.
   Keys used: State_index is being used as the primary key, which is a surrogate key apparently. This key is used to fetch data from the fact table like physician's profile and their treatments based on their geographical locations.

USE CASES:

CASE 1:

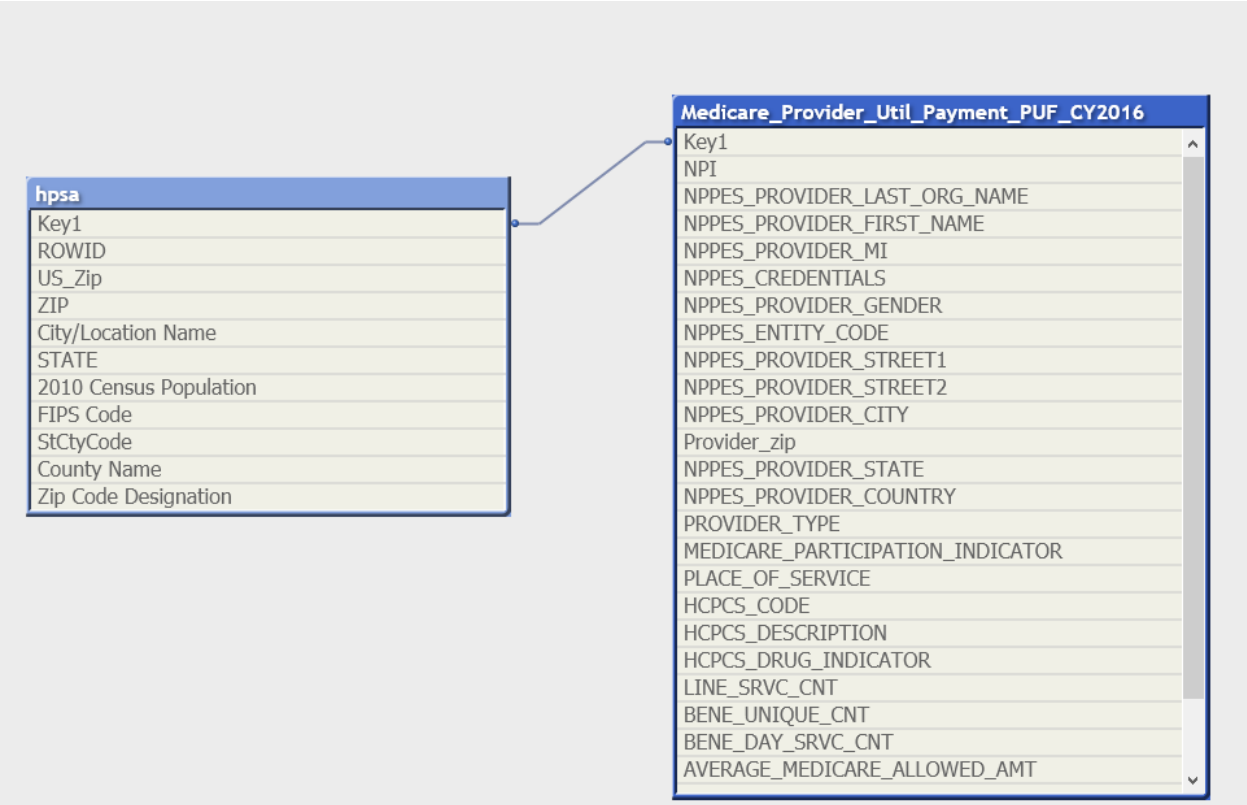
IDENTIFYING THE PROVIDERS/PROVIDER TYPES IN THE DIFFERENT ZIP CODE DESIGNATIONS SUCH AS LOW-INCOME AREAS AND HEALTH PROFESSIONAL SHORTAGE AREAS:

DATASETS USED:

1. **Medicare_Provider_Util_Payment_PUF_CY2016**
   SOURCE: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier2016.html
2. **CMS HPSA AND LOW_INCOME ZIP CODE DATABASE**
   SOURCE: https://www.kaggle.com/cms/cms-hpsa-low-income-zip-code-database/version/67

LINK:

Primary key for data set1 and 2: The zip code column can be used as a primary key for the table as it uniquely identifies the all the records in the table.



ASSUMPTION:

The following assumptions have been made:

a) **The data is not time variant and is from a single year 2016**
b) **The data set consists of updated data for all the zip codes and consists of records that have been reported by every single provider within those zip codes.**

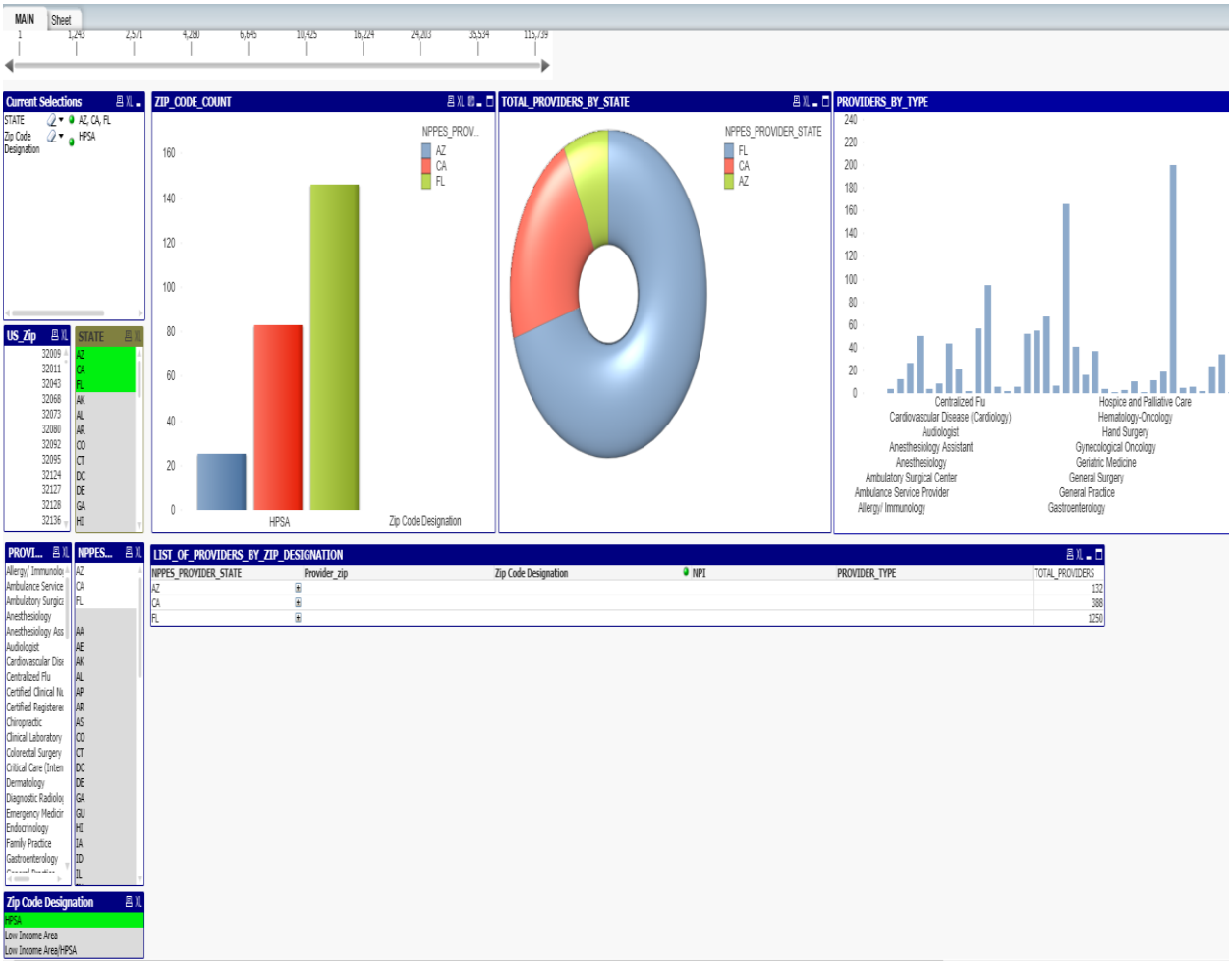
INFERENCES:

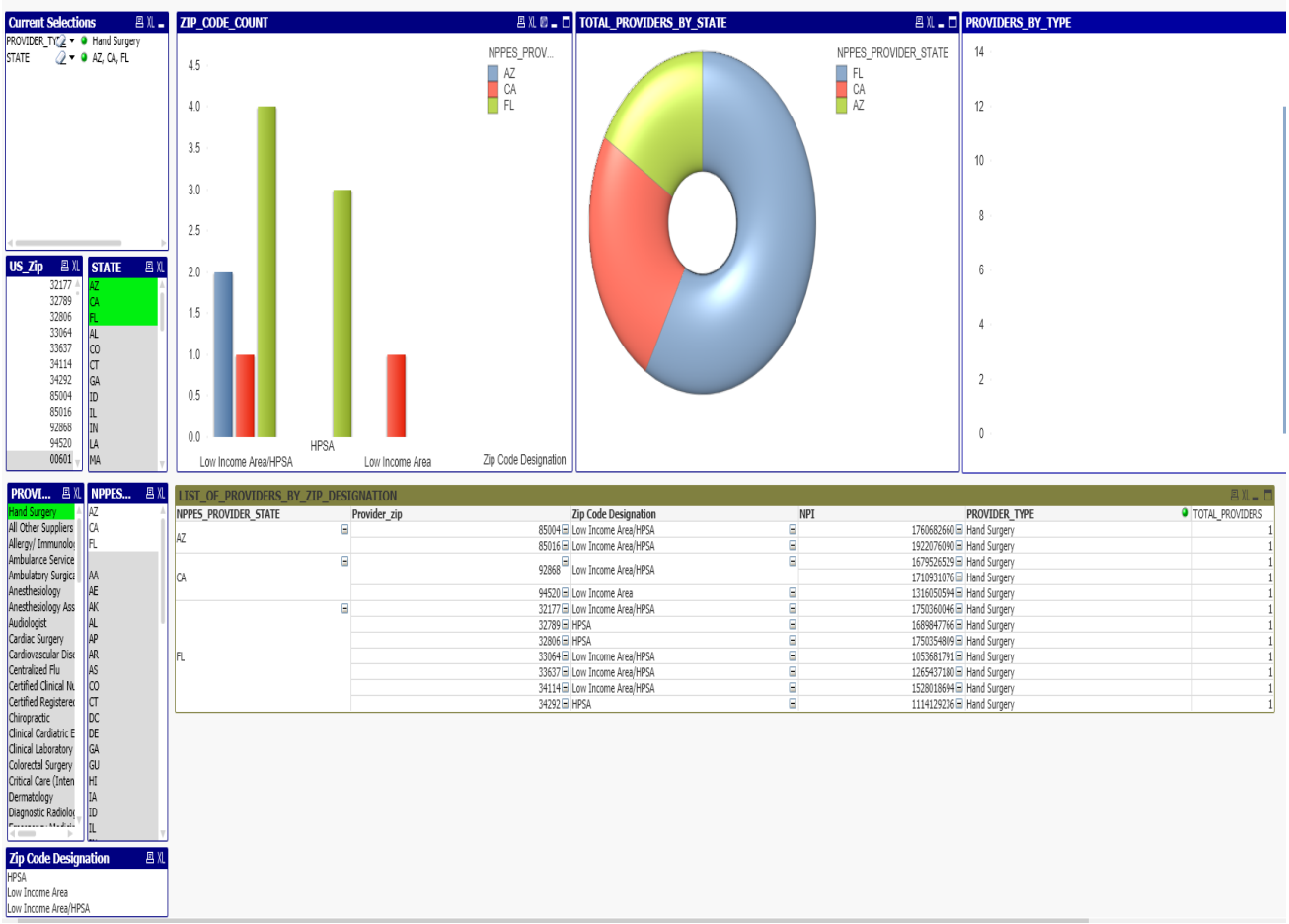
1. Consider the three states Arizona, California, Florida.
We can deduce the list of Zip codes in these areas that belong to

1) **Health professional Shortage area (HPSA)**
2) **Low income area**
3) **Low income and HPSA area**

2.Based on these search results you can further look at the type of physicians in each of the Zip code designations.

For example:
There is a total of 3609 Providers in the state of Arizona
Out of which Providers in HPSA Area: 388
Providers in Low income Area: 1834
Providers that are in low income HPSA Area: 1387
Based on these searches an individual type of provider can be in each of these Zip code designation types.
3.The state with the highest Shortage of Health Professional regions is Florida

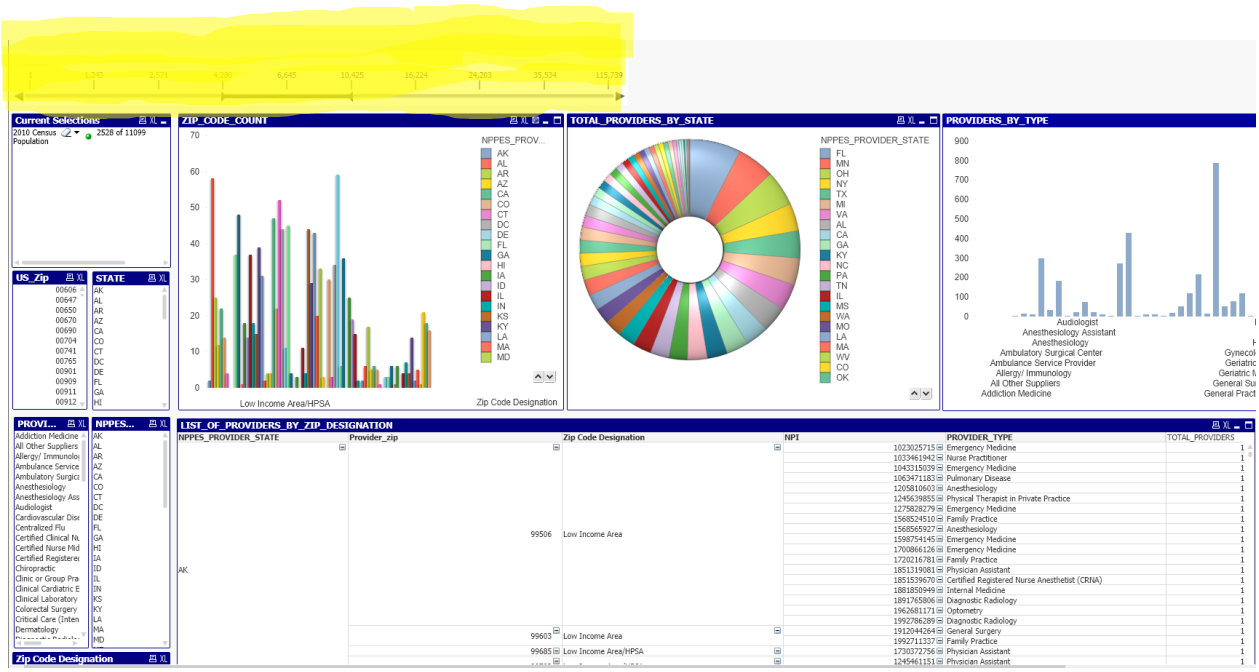

MAIN
Sheet
Current Selections
STATE
AZ, CA, FL
Zip Code Designation
HPSA
ZIP_CODE_COUNT
NPPES_PROV...
AZ
CA
FL
HPSA
Zip Code Designation
TOTAL_PROVIDERS_BY_STATE
NPPES_PROVIDER_STATE
FL
CA
AZ
PROVIDERS_BY_TYPE
Centralized Flu
Cardiovascular Disease (Cardiology)
Audiologist
Anesthesiology Assistant
Anesthesiology
Ambulatory Surgical Center
Ambulance Service Provider
Allergy/ Immunology
Hospice and Palliative Care
Hematology-Oncology
Hand Surgery
Gynecological Oncology
Geriatric Medicine
General Surgery
General Practice
Gastroenterology
US_Zip
STATE
Zip Code Designation
HPSA
Low Income Area
Low Income Area/HPSA
LIST_OF_PROVIDERS_BY_ZIP_DESIGNATION
NPPES_PROVIDER_STATE
Provider_zip
Zip Code Designation
NPI
PROVIDER_TYPE
TOTAL_PROVIDERS
AZ
CA
FL
132
388
1250



3. **For a provider type = Hand Surgery in these three states a list of all providers can be seen about and corresponding Zip code designations of the areas the providers belong to.**
4. **Population based searches:**

The use case allows you to select the range of population say the states between a population count of 10000 to 15000 only.

The highly populated states should have more number of providers and less HPSA areas.



FUTURE SCOPE:

We can further aggregate these zip codes by city and get a better view of how a state looks in terms of the healthcare shortage and what are the areas of improvement for each state.

A further analysis can be done on the states that have military presence and non-military regions of the states and the equivalent statistics

CASE 2:

This visualization gives an entire view and spread of the distribution of low Income, Health Professional Shortage Area and Low income/HPSA and comparing it with Republican and Democrat represented states

DATASETS USED:

CMS HPSA & Low-Income ZIP Code Database
https://www.kaggle.com/cms/cms-hpsa-low-income-zip-code-database/version/67#

State Fips
https://www.census.gov/geo/reference/ansi_statetables.html

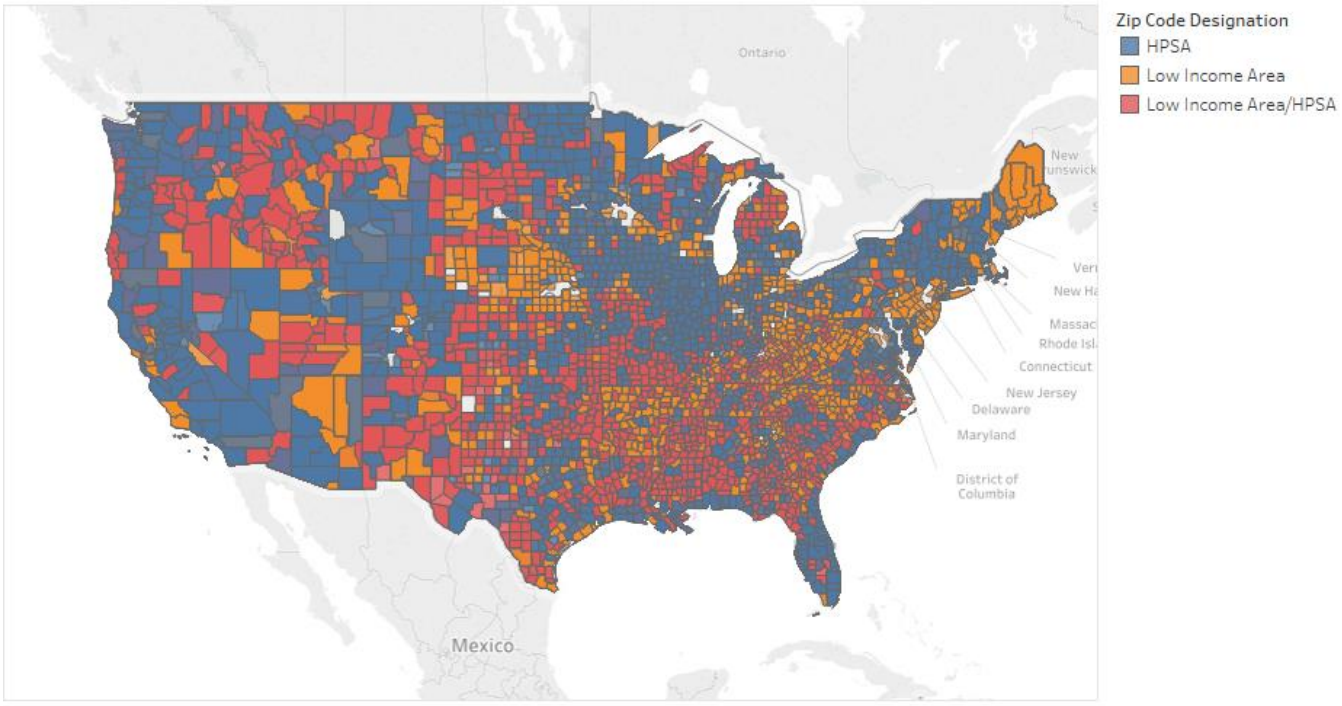
County Fips
https://www.census.gov/geo/reference/codes/cou.html

We can visually see that most of the South east states in the US come under the Low Income and HPS area, which means that area is categorized by the worst when compared to other states on the west coast and some of the state in the North east region, most of these states have only a Health Professional shortage situation.

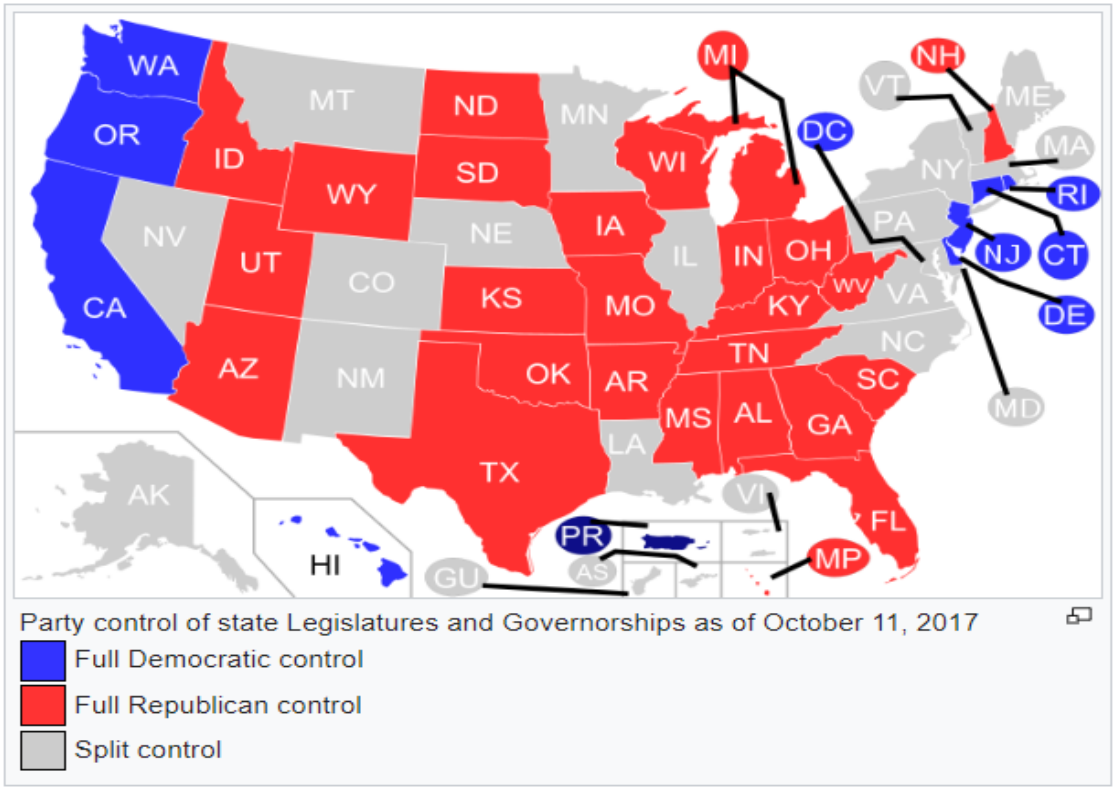
When we compare this with the states ruled by Democrats and Republicans, we can see that most of the states in the west coast and the North eastern regions are represented by Democrats. On the other hand, most the states which are categorized as Low-Income area and HPSA are mostly represented by republicans.

Our report and comparisons show that the democrats have been doing a better job in terms of raising the level of income in their states. However, it looks like the republicans must focus more on key issues such as improving the quality of healthcare, raising the minimum wage, increase the earned income tax credit for childless workers, invest in affordable, high-quality child care and early education, expand Medicaid, etc. to improve the quality of living in their states.

Spread of Low Income, HPSA and Low Income/HPSA areas in the country



Map based on Longitude (generated) and Latitude (generated). Color shows details about Zip Code Designation. Details are shown for State Name, ZIP and County Name. The view is filtered on Zip Code Designation, which keeps HPSA, Low Income Area and Low Income Area/HPSA.



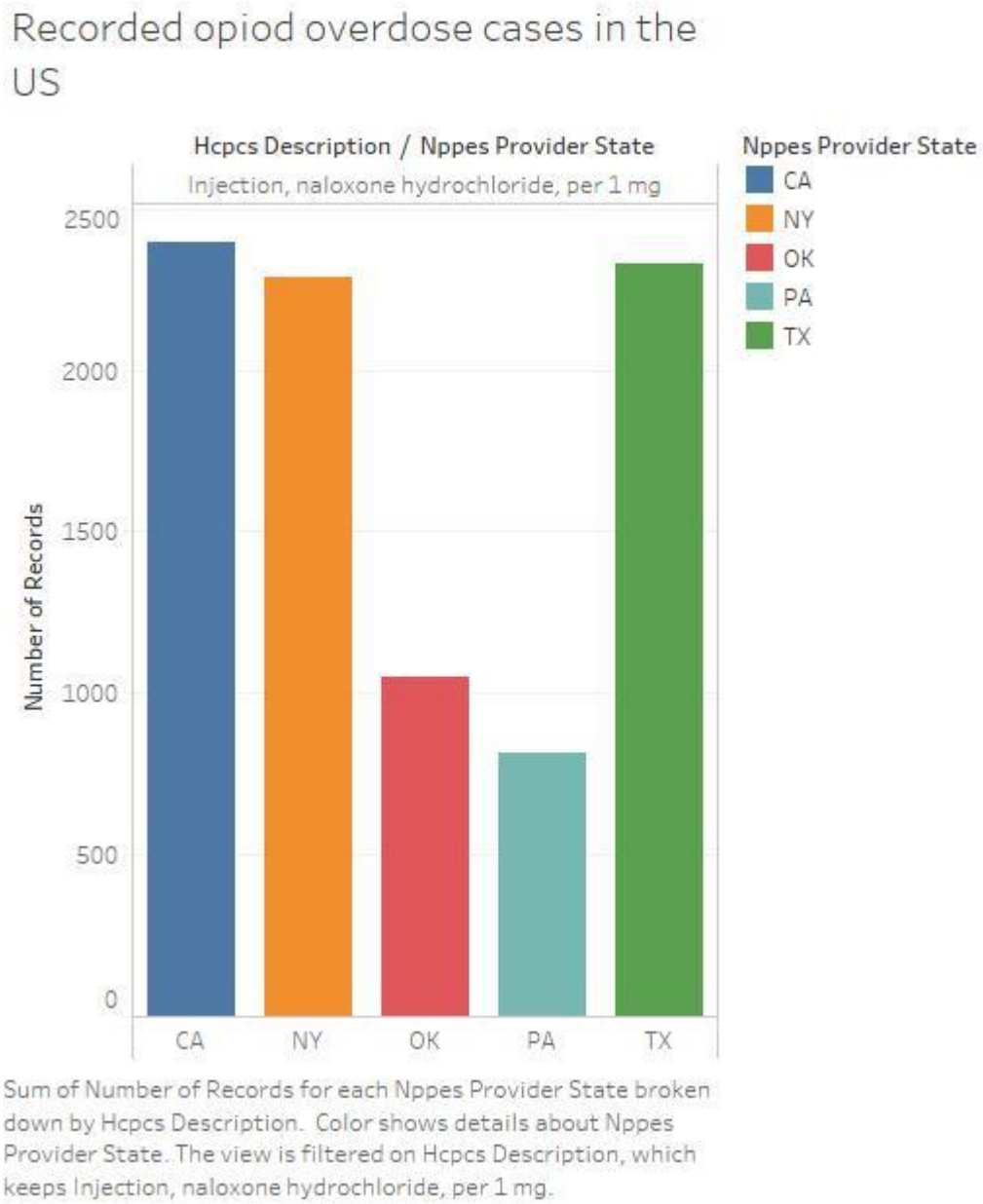
Reported Opioid overdose cases in US by state

Opioid overdose is one of the most common drug abuse in the country however, only these states have recorded cases of opioid overdose under the Medicare plan.

Naloxone is an anti-overdose drug given to individuals who have overdosed on Opioids. This treatment record was found through the 'HCPCS description' column in the data set. Hence, this can be related with opioids over dose cases.

DATASET LINKS:

The ' Medicare Provider Utilization and Payment Data: Physician and Other Supplier ' data set was linked with the ' STATE_FIPS ' table to bring out the following graph.



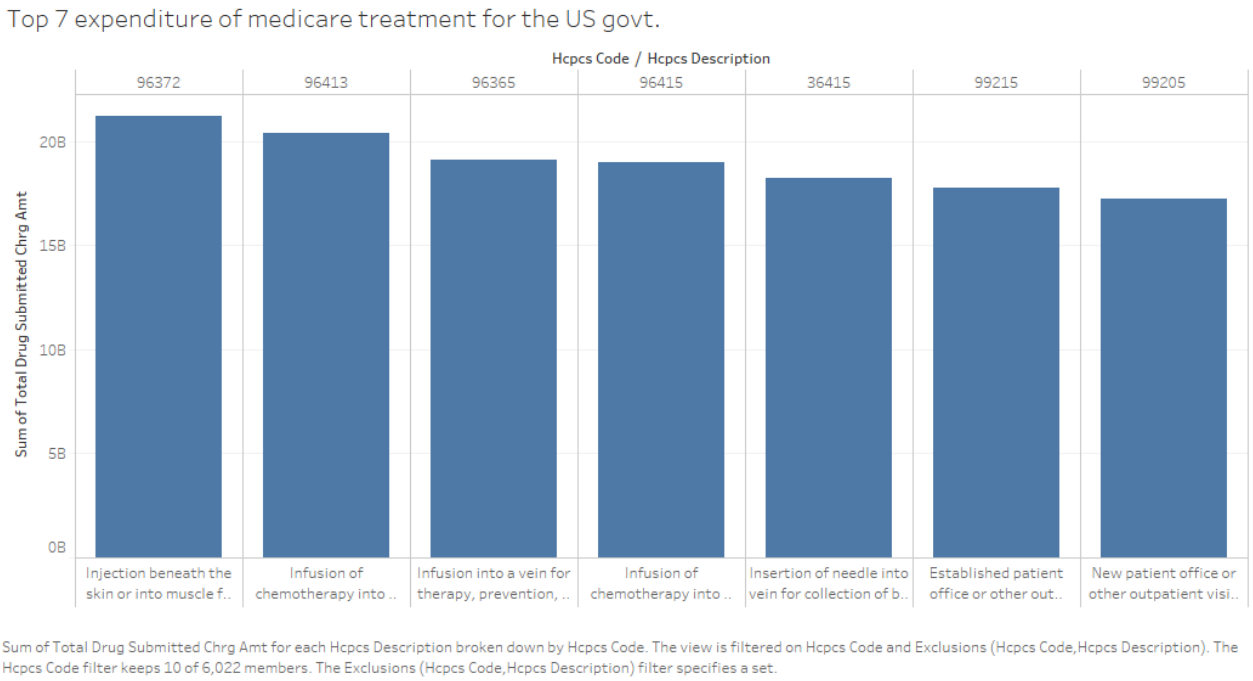
CASE: Top 7 government Medicare budget spending

This visualization shows us the top 7 areas in which the government's Medicare budget is being spent.

DATASETS Used: The Medical Provider Payment PUF dataset was used to bring out the visualization.

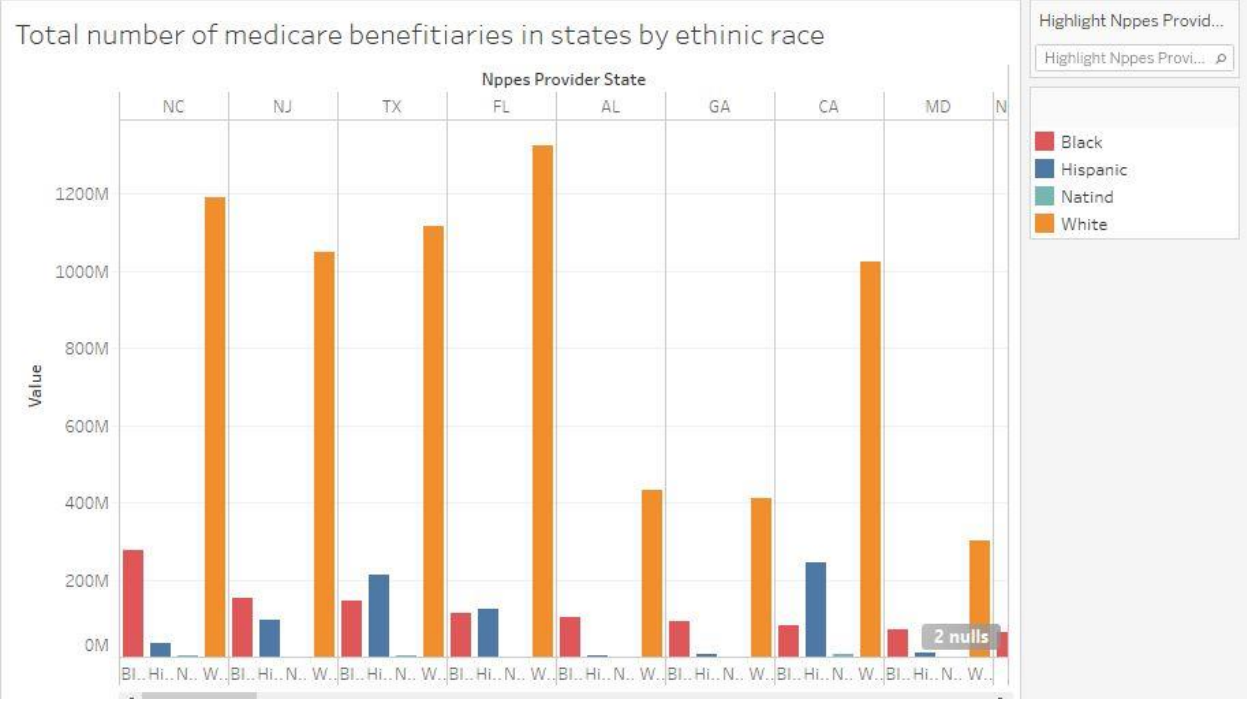
The 'Total Drug Submitted_Charg_amt' column was used with the HCPCS description and the HCPCS code.

INFERENCE: We see that more than $40B of the Medicare budget is spent on injections and chemotherapy infusion. HCPCS code 36415 (Collection of blood samples) uses up a lot of that budget too!



Govt. spends a major part of their Medicare expenditure on outpatient medical services and supplies which comes under the Medicare Part B as well.

A report on the total number of Medicare beneficiaries in States by Ethnic Race

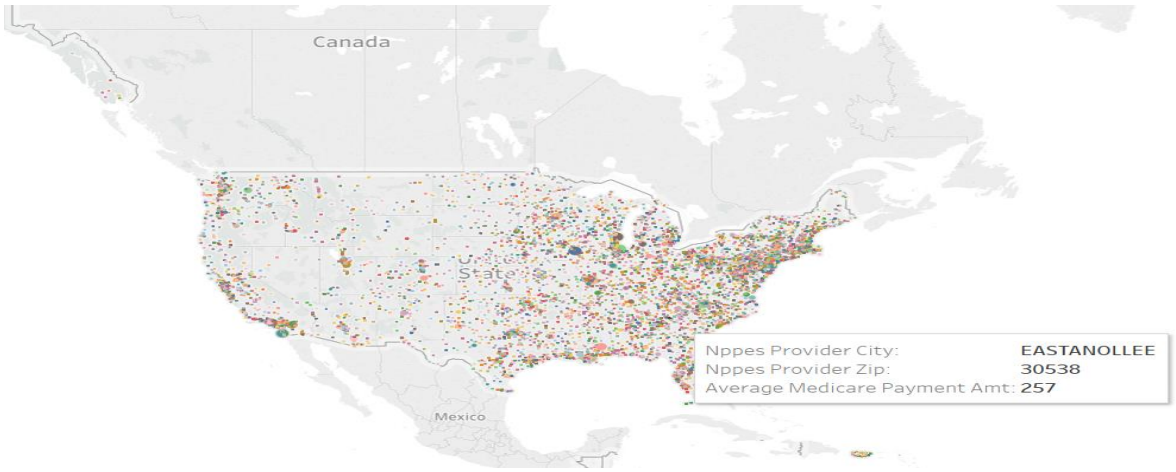


Overall, a major portion of the beneficiaries in most states are from the white ethnic population

CASE 3:

Visualizations created:

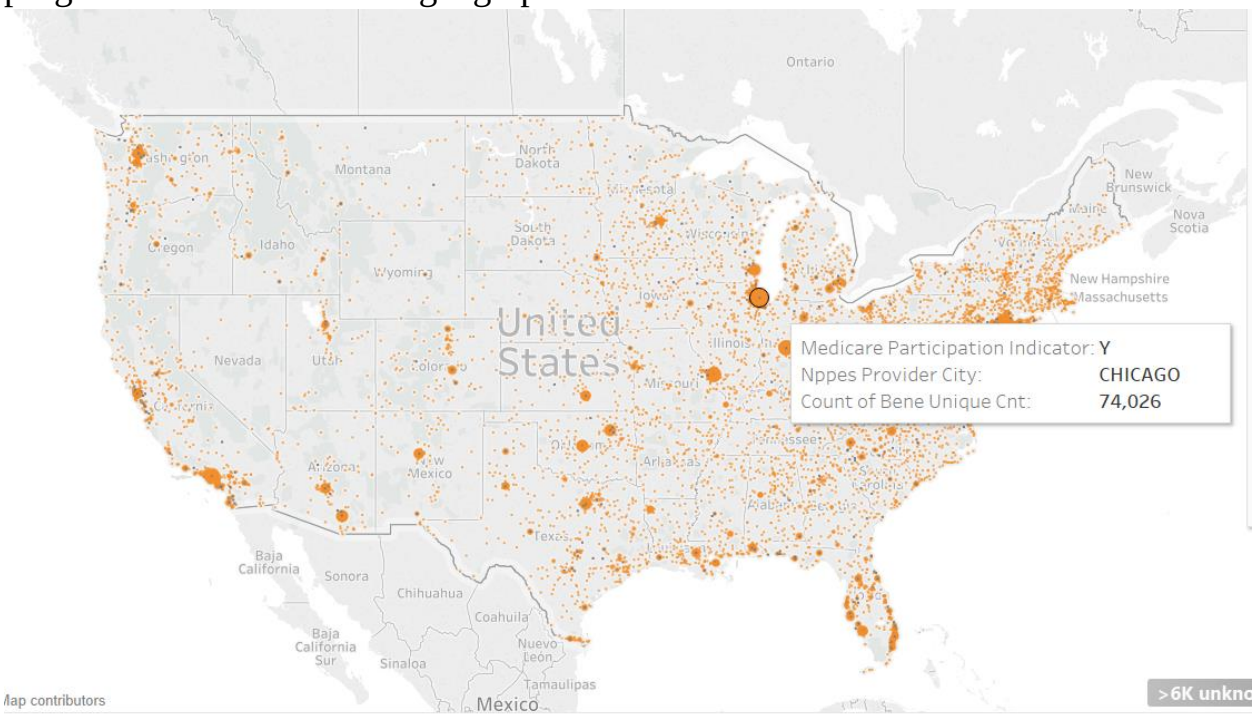
Visualization 2: The first visualization created here talks about the Medicare participation indicator. MPI is nothing but an indicator which talks about the organizations participating in medicare programs. **Medicare** is the federal health insurance **program** for People who are 65 or older. Certain younger people with disabilities. People with End-Stage Renal Disease (permanent kidney failure requiring dialysis or a transplant, sometimes called ESRD). In the areas where the Medicare Participation Indicator is TRUE, the cost of Medicare reduces drastically. This visualization shows the Average Medicare Payment amount in Dollars based on the geographical constraints like ZIP code and city name.



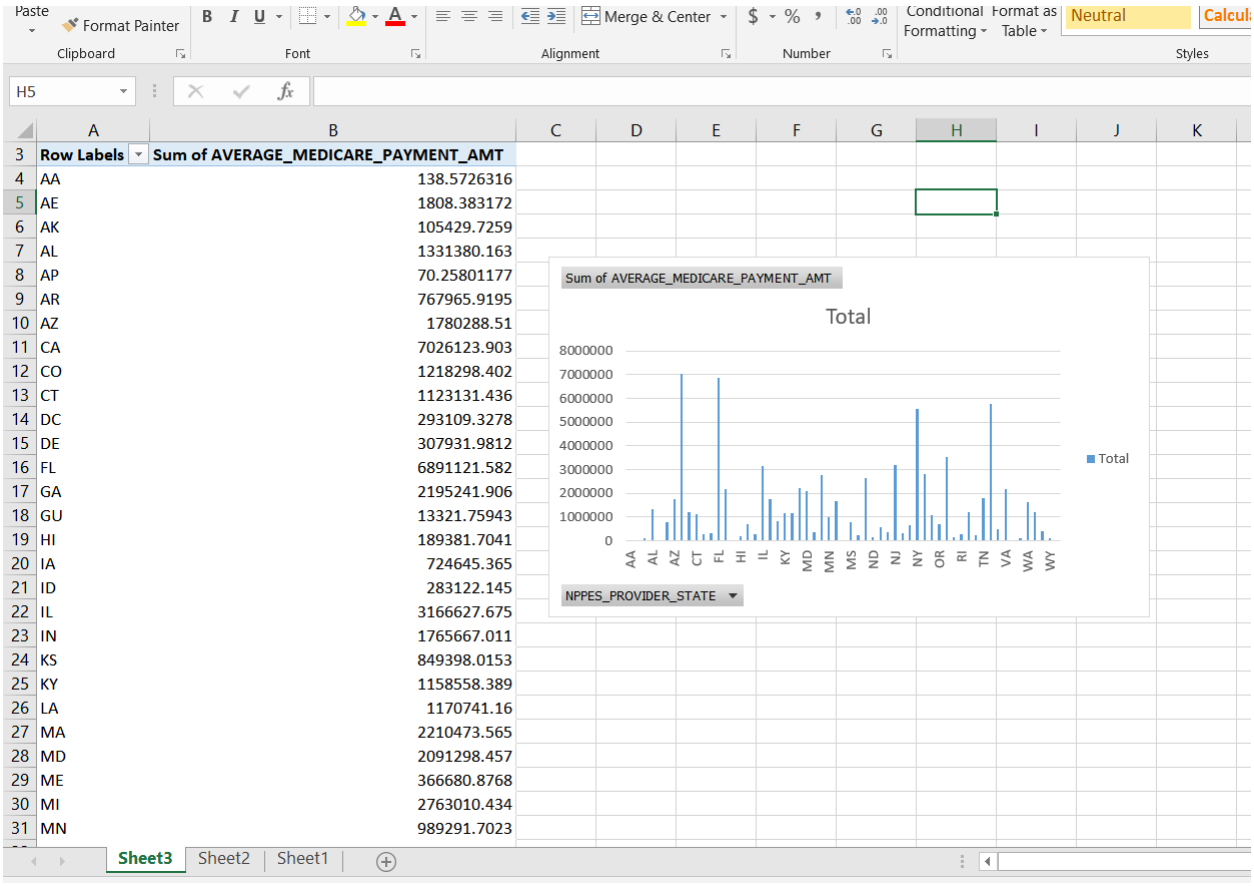
FUTURE SCOPE: The future scope of this visualization is that it can be used to plot a trend for the average Medicare payment amount every year under different ruling political parties.

Visualization 2*:* This visualization tells us about the Medicare participation indicator and the number of unique beneficiaries who utilized the Medicare

program based on different geographical locations.



FUTURE SCOPE: Future scope of this visualization is to keep an eye on the trend for number of unique beneficiaries based on different cities and under the Medicare Participation Indicator. This can be used to identify how the beneficiaries are benefitted under different ruling government



Row Labels	Sum of AVERAGE_MEDICARE_PAYMENT_AMT
AA	138.5726316
AE	1808.383172
AK	105429.7259
AL	1331380.163
AP	70.25801177
AR	767965.9195
AZ	1780288.51
CA	7026123.903
CO	1218298.402
CT	1123131.436
DC	293109.3278
DE	307931.9812
FL	6891121.582
GA	2195241.906
GU	13321.75943
HI	189381.7041
IA	724645.365
ID	283122.145
IL	3166627.675
IN	1765667.011
KS	849398.0153
KY	1158558.389
LA	1170741.16
MA	2210473.565
MD	2091298.457
ME	366680.8768
MI	2763010.434
MN	989291.7023

CASE 4:

IDENTIFICATION AND ANALYSIS OF STATES WITH HIGHEST RATE OF COLORECTAL CANCER

DATASETS USED: Low Income/HPSA and Medicare_PUF

LINKS:

https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html

https://www.kaggle.com/cms/cms-hpsa-low-income-zip-code-database/version/67#

ASSUMPTION: The recipient of the service and NPPES Provider are in the same state.

DESCRIPTION:

Based on HCPCS description, colorectal cancer cases were filtered. The number of beneficiaries receiving the colorectal cancer treatment were mapped as per provider's state(based on assumption that both provider and beneficiary belong to the same state). The analysis was based on number of beneficiaries / total population for that particular state. The states showed a certain trend which was then backed by the research on the commonality between the states.

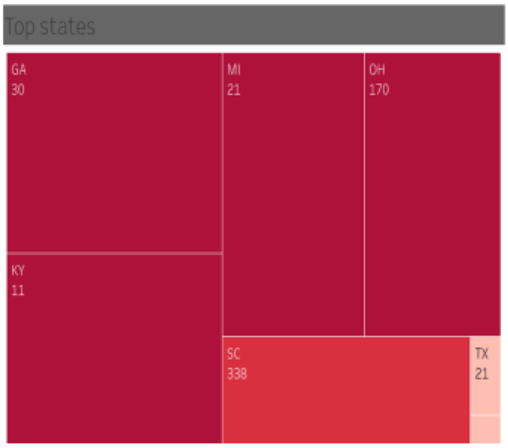
Primary key-[Zipcode] -Low Income/HPSA and [LEFT(NPPES_PROVIDER_ZIP,5) AS [N_ZIP]]- Medicare_PUF

INFERENCES: The database indicates higher colorectal cancer rate(% of total population) in eastern states as compared to the west.

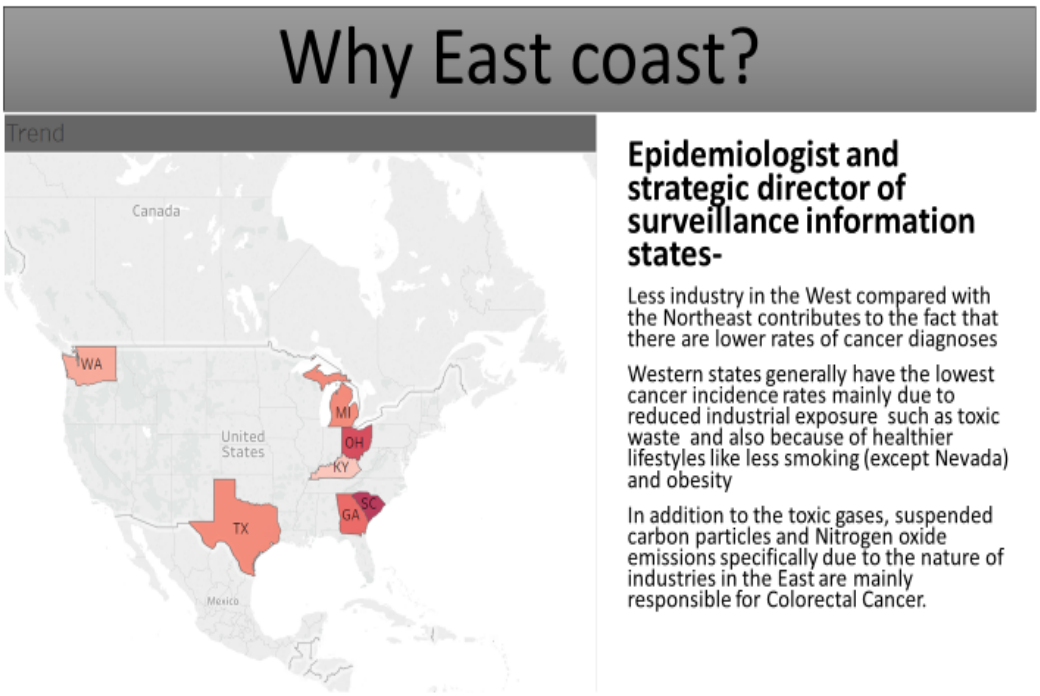
Eastern states have higher number of oil/gas/coal industries resulting in increased exposure to toxic gases. Air pollution is classified as carcinogenic to humans as researchers observed an association between some air pollutants and **mortality from colorectal cancer.**

Colorectal cancer in United States

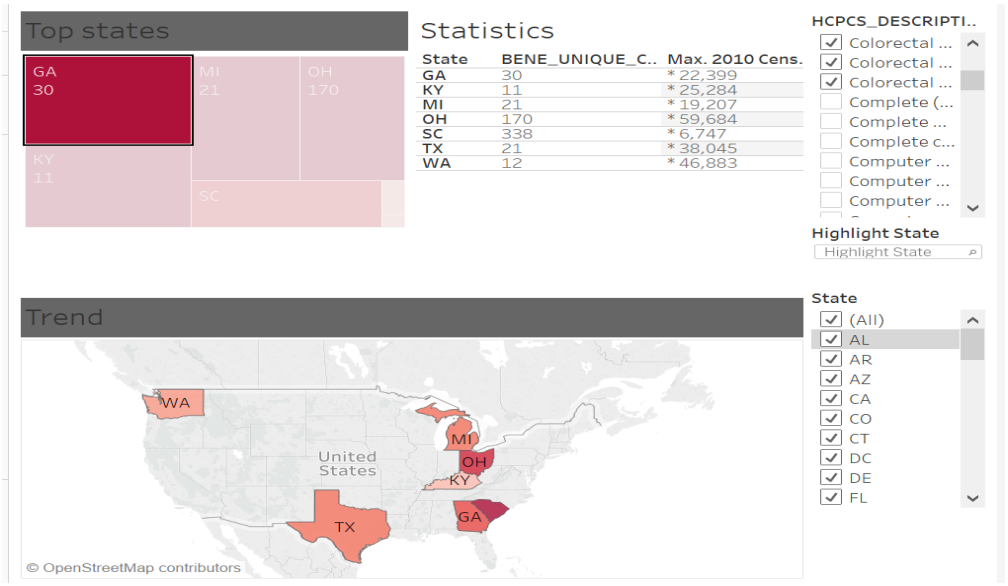
Colon and rectum make up the large intestine (or large bowel), which is part of the digestive system, also called the *gastrointestinal (GI) system*



State	Information
GA	Scherer coal-fired power plant near Macon is the number one producer of greenhouse gases in the United States, emitting 22.8 million metric tons of carbon dioxide
KT	Ranked worst in terms of toxic air pollution from coal-fired power plants
MI	Monroe coal plant registered total greenhouse gas emissions 17,850,341 CO2e with the EPA in 2010, making it the country's 6th worst polluter
OH	Ohio is home to the Gen J M Gavin coal plant in Cheshire, which is the 9th biggest polluter in the United States
SC	Electric sector ranked 8th in industrial toxic air pollution in 2010, emitting more than 14.6 million pounds of harmful chemicals



Dashboard



REFERENCES AND LINKS

https://www.verywellhealth.com/what-are-medicares-hcpcs-codes-2614952

https://www.policygenius.com/blog/a-state-by-state-guide-to-medicaid/

https://www.policygenius.com/blog/a-state-by-state-guide-to-medicaid/

https://www.policygenius.com/medicare/medicare-part-b-plans/