Analytics of Medicare Industry

APAN5310 SQL
Final Project Report
Team 1: Sarah Bortnem, Shengyao Li, Mingjun Ma, Leah Reinhard, Qianqian Wu

Fall 2019
School of Professional Studies
Columbia University

Introduction and Project Structure

Introduction

For our SQL final project, we were charged with the task of cleaning a messy dataset and transforming it into a live datasource that can support analytical applications. This project was meant to demonstrate the skills we learned throughout the semester and showcase our ability to support business decisions through data from non-uniform sources. The dataset we chose for this project was the Medicare Physician & Other Supplier NPI Aggregates data provided by the Centers for Medicare & Medicaid Services (CMS). (link provided below)

https://www.kaggle.com/cms/medicare-physician-other-supplier-npi-aggregates

Motivation & Research

The selection process for our dataset took a significant amount of time given the project requirements. We researched datasets on Kaggle, kdnuggets, and Quora, looking for an open, anonymized dataset to normalize and restructure into a relational form for our project. At first, we decided to choose a Kaggle dataset on US Traffic Fatalities. However, we later realized that this dataset was not available in csv format for downloading and we were set back at square one. After more research, we found a better dataset for our analysis: Medicare Physician & Other Supplier NPI Aggregates. We chose this dataset because it contains a significant amount of data (67 columns in the file) that could easily be split into more than 20 tables. Furthermore, this dataset is maintained by Socrata's API and Kaggle's API, meaning it is updated daily and allows for automated dashboard generation.

Objectives and Background

To provide an explanation for our objectives, it is necessary to explain background information on Medicare and its current status. Medicare is a federal health insurance program for people of all ages with disabilities. According to KFF's report "The Facts on Medicare Spending and Financing", Medicare spending was $605 billion in 2018 and made up 15 percent of the federal budget. Moreover, with the growth of enrollment due to a larger aging population and the increase in healthcare prices per person, the future spending of Medicare is predicted to keep growing. According to the report "Medicare Plan Finder: Usability Problems and Incomplete Information Create Challenges for Beneficiaries Comparing Coverage Options" issued by the General Accounting Office, there is a lack of information about provider networks for beneficiaries before making their coverage decisions. With a better understanding of payment data, government officials would be able to make wiser decisions regarding the refinement of current Medicare plans. Furthermore, beneficiaries would be able to make more educated decisions regarding their healthcare providers if given access to an organized Medicare database. This is our objective for our dataset, to create a database where providers and beneficiaries can have direct access to valuable information that will help them make educated decisions regarding coverage plans. We also want to provide an overview for government officials to observe

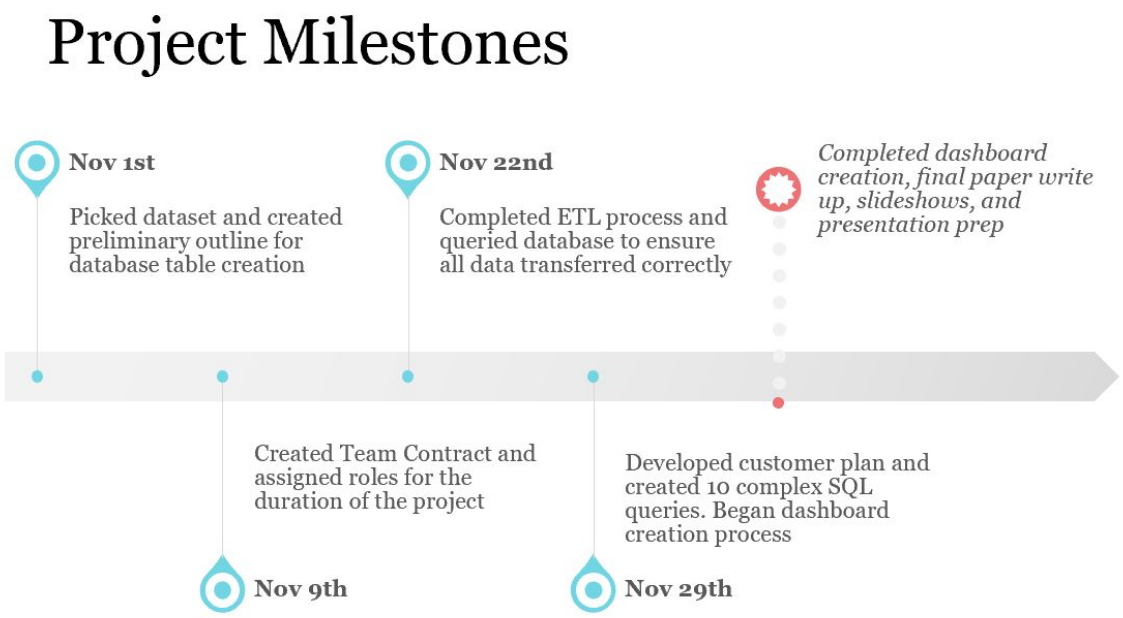
Medicare usage data to help them determine strategies to revamp health care policies. This leads us to discuss our proposed scenario.

Scenario and Benefits

The US government hired our project team as analysts to track trends in Medicare usage to derive valuable insights. This ranges from determining the amount of beneficiaries for each provider to collecting demographic information on Medicare users. This endeavor required our group of analysts to clean the extensive government data on Medicare physicians, organize it into readable tables, and create rules for new data input procedures. In terms of the data output, we needed to create restricted views for various external entities with access (such as insurance companies / beneficiaries) as well as internal users (such as managers / C-executives / government officials). We provided high-level overview through visualizations and interactive dashboards to allow users to make more well-rounded decisions. This research will help the US government inform the public of the success of Medicare, allow them to petition for appropriate changes to the bill based on findings, and provide them with a more organized Medicare data entry process for future analysis.

Team Contract

At the beginning of the project, we drew up a team contract to make sure that each team member would be held accountable for a certain aspect of the project. This helped us develop a rough outline for when deliverables should be completed and who was responsible for overseeing the project each week. Each member electronically signed this document knowing that it would be referenced by Professor Machairas while grading our final project submission. Below is a timeline of when the team completed the required project deliverables.



APAN5310 Project Team 1 Contract

This document represents an official agreement between the 5 members of Fall 2019 APAN5310 project Team 1 with members - **Sarah Bortnem, Shengyao Li, Mingjun Ma, Leah Reinhard and Qianqian Wu**.

Overall Goals

The following goals are expected to be accomplished throughout the duration of this project. Each point requires group discussion in great detail:

- Actively participate in team discussions and create outputs for each project task
- Finish project checkpoint 24 hours in advance (according to following due dates) for review
- Review deliverables as a team before checkpoint submission
- Allocate team member to submit final project deliverable at least 1 hour prior to due time
- Allocate team members to ensure project submission (for accountability)
- Rehearse in advance to guarantee high-quality final presentation

Team Meeting Ground Rules

- Meetings will begin and end on time
- Team members will come to the meetings (w. Both online/physical options)
- Agenda items for the next meeting will be discussed at the end of each meeting
- If a team member can not attend a meeting, he/she should inform team members in advance and catch up with meeting materials at a later time
- Each team member will take on a position of leadership for an aspect of the project, fairly distributing the workload
- Failure to comply with these ground rules will result in reallocation of next deliverable's responsibilities to ensure fair distribution of workload

Tasks and Assignments

Task	Description	Person(s) in Charge	Due Date for Team-wise Review
Project Checkpoint 2	Develop a Team Contract	Mingjun Ma	11:59PM, Nov 10th
	Finalize Project Dataset	Leah Reinhard	
Project Checkpoint 3	Submit the Database Schema (ER Diagram + SQL code)	Leah Reinhard & Sarah Bortnem	11:59PM, Nov 16th
Project Checkpoint 4	Submit the data plan for review (transform and enter the data to db system)	Qianqian Wu & Shengyao Li	11:59PM, Nov 23th
Project Checkpoint 5	Submit Customer Interaction Plan	Shengyao Li	11:59PM, Nov 30th
Project Report & Presentation	Presentation Slides (<10 slides)	Sarah Bortnem & Mingjun Ma	11:59PM, Dec 7th
	Final Report (10-20 pages, 1.2 spacing)	All Team Members(?)	11:59PM, Dec 8th

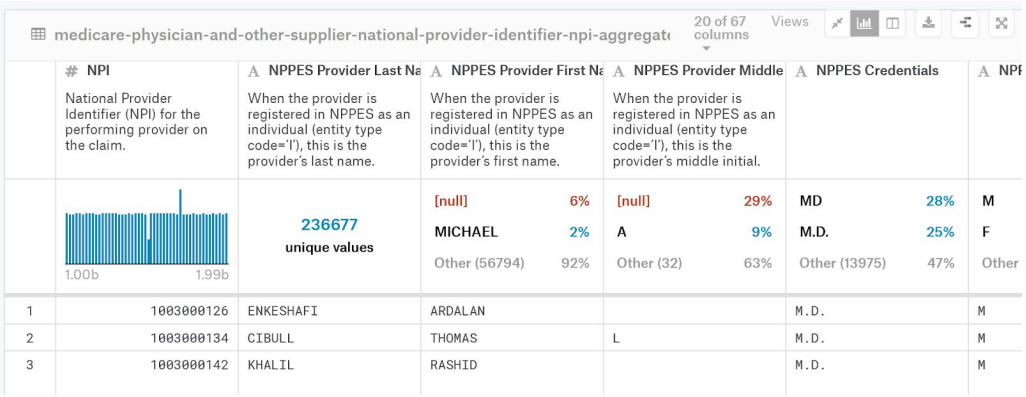
Database Description

Data Source

The dataset we chose for our project was provided by the Centers for Medicare & Medicaid Services (CMS), which is a federal agency in the US Department of Health and Human Services (HHS). It contains information on utilization, payments, and submitted charges by National Provider Identifier. This gives us insight into medicare utilization, payments, providers and beneficiaries. The dataset contains 67 columns and over a million rows, detailing information on all parties involved in Medicare transactions. Two important columns that helped us create primary keys for the database tables include:

- National Provider Identifier (NPI) : National Provider Identifier (NPI) is a unique ten-digit number for the performing provider on the claim issued by the CMS.
- Hcpcs Code: The Healthcare Common Procedure Coding System, a set of health care procedure codes based on the American Medical Association's Current Procedural Terminology (CPT).

Columns containing organization names and beneficiary names were also crucial for table creation. We chose this dataset since there was a significant amount of data that could be sorted through and placed into a logical format for database creation. We were able to efficiently clean the dataset, removing all redundant entries and unnecessary N/A values. We then normalized the dataset and reviewed columns to determine an organized table structure with around 16 tables. This led us to our next step which included table creation, proper foreign key referencing, and table constraints for smooth data integration (functions such as Cascade). An image sample of the dataset is provided below.



Normalization Plan

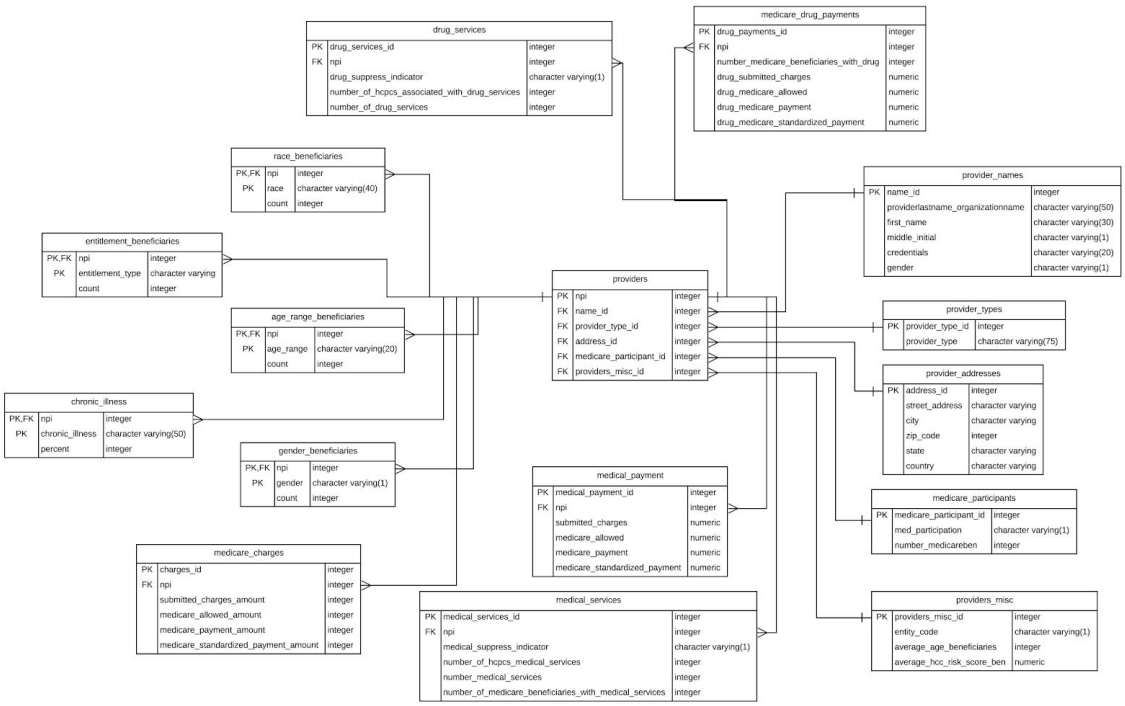
It was necessary to normalize the data to reduce data redundancy and improve data integrity. First, we checked to verify that all attributes were atomic so that the data could be properly queried. We then removed repeating attributes in the table by combining columns into one column with different row entries. For example, there were multiple columns that counted the number of beneficiaries belonging to a certain ethnic group. This means the table had six separate columns indicating race for each individual beneficiary. We decided to combine these columns into one column labeled "race", querying the table for columns that had numbers to indicate whether the beneficiary belonged to that racial group. Each entry showed up as a separate row with the corresponding information. We repeated this process with the columns containing different medical conditions, combining them into the "chronic_illness" attribute with its corresponding "percentage" attribute. The columns "gender", "entitlement_type", and "age_range" all had similar normalization requirements for table creation.

Once the preliminary normalization was complete, we began sorting through the columns to determine a logical layout for our database table creation. We determined that there were three main groups to divide the columns into: provider data, beneficiary data, and medicare process data. Once we sorted out these groups, it was easier to divide them further to create sixteen tables. Each table's attributes are listed below in the SQL code for table creation. When creating the tables, we determined that new attributes should be created to act as unique identifiers for certain table entries. For example, we created the column "address_id" in the provider_addresses table to have a primary key that would reference all the information in one row concerning addresses. We did this for several attributes, including "name_id", "provider_type_id", "providers_misc_id",

“provider_type_id”, “medicare_participant_id”, “charges_id”, “drug_payments_id”, “medical_payment_id”, “drug_services_id”, and “medical_services_id”. Once we formatted the layout for sixteen tables, we created the SQL code for table creation, pairing each attribute with the appropriate data type and character count. The GitHub repository for the SQL code is listed below.

https://github.com/sarahb511/SQL_FinalProject_TableCreation/blob/1.0/Final_Project_Table_Creation.sql

ER Diagram



Above is our final ER Diagram for the normalized tables. The link to the Lucidchart page can be found here. This star schema naturally formed for the database, as we developed the denormalized structure of the initial database.

ETL Process

We began the extraction process by uploading the file originally obtained from Kaggle. The data contained approximately a million rows, making loading the transformed data onto the server time-consuming. To save time, we decided to use a sample of 50,000 rows of the data to showcase the transformation process. We then renamed the data attributes to fit the schema we previously created during our normalization process. We started our data transformation process by searching for rows containing NA values, but found that none existed. We then sorted through the data and removed all duplicate values to create unique ids for each table row. As we

mentioned in the above normalization plan, we combined repetitive features that fell under a single category into one column for the attributes gender, race, chronic_illness, and entitlement. For each of these attributes, we added their corresponding count or percentage features into these created summarized tables. For example, we transformed the data for the age_range_beneficiaries table by taking each column with an age range and creating a dataframe for that column. We populated it with the NPI, set the age_range to the column's range, added a count, and removed the null values. This process was repeated for each age range. We then used the Concat function to merge all the age range columns into one. This created the age_range_beneficiaries table with three columns: NPI, age_range, and count. In this way, the primary key is a unique combination of NPI and age_range. The age_range_beneficiaries table was then populated and the cascade function added referencing another table. This process was repeated with the other tables containing repetitive features, leaving us with a normalized and populated database. Similar primary key structure was followed for gender (NPI and gender), race (NPI and race), chronic_illness and entitlement. This database was tested through SQL queries to ensure that the ETL process properly populated the tables. For code used in our ETL process, please refer to the following github page:

https://github.com/DDDaiii/SQL-APAN5310-Analytics-of-Medicare-Industry

Analytical Procedures to Build Dashboards

To showcase the various different ways our database can be used, we created 10 analytical procedures that created valuable insights for our clients. As we previously mentioned, our dataset is provided by the Centers for Medicare & Medicaid Services (CMS). This means that the C-suite executives for this dataset are members of a cabinet-level department of the US federal government, including the Secretary of Health and Human Services appointed by the President. The CMS also works in partnership with state governments and health insurance providers. Based on this information, we determined our database visuals should cater to C-suite executives with the highest clearance and government analysts, providing insights that could be shared with insurance providers and beneficiaries. Both dashboards will be built utilizing PostgreSQL queries and Metabase.

Provider Analysis

1. What providers have high drug submitted charges but low medical payments submitted?

- Returns NPI, provider type, drug_submitted_charges, drug_medicare_allowed, drug_medicare_payment, medical submitted_charges, medical medicare_allowed, and medical .medicare_payment. Only return values where drug_submitted_charges exceed $1,000,000 to narrow viewing points. Additionally, it will only return providers where the medical submitted_charges is less than 50% of the drug_submitted_charges.
- This will be useful for analysts to identify providers to further understand why their drug charges are so high and whether instances of medicare drug abuse is occurring.

```
SELECT p.npi, pt.provider_type, mdp.drug_submitted_charges,
```

```
        mdp.drug_medicare_allowed, mdp.drug_medicare_payment, mp.submitted_charges,
        mp.medicare_allowed, mp.medicare_payment
FROM medicare_drug_payments as mdp, medical_payment as mp, providers as p, provider_types
as pt
WHERE mdp.number_medicare_beneficiaries_with_drug != 0 AND
mdp.drug_submitted_charges > 1000000 AND mdp.npi = p.npi AND mp.npi = p.npi AND
p.provider_type_id = pt.provider_type_id AND mp.submitted_charges <
(mdp.drug_submitted_charges* 0.5)
ORDER BY mdp.drug_submitted_charges DESC,mp.submitted_charges ASC
```

2. Most popular provider type grouped by state.

- Return the state, provider type, and count of provider type per state. This can be utilized within a dashboard for C-suite executives -- in our case, government policymakers -- to understand what provider types should be prioritized for each state.

```
WITH state_provider_types (address_id, state, provider_type_id, provider_type) AS (
SELECT A.address_id, A.state, Pr.provider_type_id, P.provider_type
FROM provider_addresses A
JOIN  providers Pr
ON A.address_id = Pr.address_id
JOIN provider_types P
ON Pr.provider_type_id = P.provider_type_id)
SELECT state, provider_type, COUNT(*) AS amount
FROM state_provider_types
GROUP BY state, provider_type
ORDER BY amount DESC;
```

3. What is the average percentage of costs that medicare covers on medical bills by state and provider type?

- Government policymakers (C-suite) can use these metrics to build a baseline for business and coverage decisions at the state level

```
WITH charges_covered (medical_payment_id, npi, medicare_allowed, medicare_payment,
address_id, state, provider_type_id, provider_type) AS (
SELECT M.medical_payment_id, M.npi, M.medicare_allowed, M.medicare_payment,
A.address_id, A.state, Pr.provider_type_id, P.provider_type,
TRUNC((medicare_payment/ medicare_allowed) * 100) AS percent_covered
FROM medical_payment M
JOIN providers Pr
ON M.npi = Pr.npi
JOIN provider_addresses A
ON A.address_id = Pr.address_id
JOIN provider_types P
ON Pr.provider_type_id = P.provider_type_id)
SELECT state, provider_type, TRUNC(AVG(percent_covered)) AS avg_percent_covered
FROM charges_covered
WHERE percent_covered IS NOT NULL
AND medicare_allowed IS NOT NULL
GROUP BY state, provider_type
ORDER BY state, avg_percent_covered DESC;
```

4. What is the average gender count for the five most frequent provider types?

- This would be useful for the C-suite executives (policymakers within the US Dept of HHS) to build an understanding of the demographics for frequent provider types used under medicare.

```
SELECT  pt.provider_type as ProviderType, gb.gender as Gender, avg(gb.count) as GenderCount
FROM provider_types as pt, providers as p, gender_beneficiaries as gb
WHERE pt.provider_type_id = p.provider_type_id AND p.npi = gb.npi AND
        pt.provider_type IN (SELECT provider_type FROM (
                SELECT provider_type, COUNT(*) AS counted
                FROM   provider_types
                GROUP  BY provider_type
                ORDER  BY counted DESC, provider_type
                LIMIT  5) subquery)
GROUP BY ProviderType, Gender
ORDER BY ProviderType, Gender
```

Beneficiary Analysis

1. For California, list the names of the top 10 popular NPPES providers, where beneficiaries would go to when they have cancer, evaluated by the percentage of beneficiaries with cancer.

- The structure of this query could be used in the backend of a dashboard to provide analysts with a list of top used providers for a specific disease. This would be useful in building an understanding of how a beneficiary selects a new provider when a beneficiary is going through something as difficult as a diagnosis.

```
SELECT PN.provider_lastname_organizationname, PN.first_name, PA.city, PA.state,
PA.zip_code, CI.chronic_illness, CI.percent
FROM providers P
JOIN provider_names PN on P.name_id = PN.name_id
JOIN provider_addresses PA on P.address_id = PA.address_id
JOIN chronic_illness CI on P.npi = CI.npi
JOIN medical_services MS on P.npi = MS.npi
WHERE PA.state = 'CA'
AND CI.chronic_illness = 'Cancer'
ORDER BY CI.percent DESC
LIMIT 10
```

2. The best city to live nearby for beneficiaries older than 84 years old in each state with recurring heart failure problems when considering the shortest distance to the best Cardiac Surgery provider in their states, evaluated by the number of medical services provided.

- Oftentimes when patients are dealing with serious or chronic diseases, relocation might be necessary to ensure the longevity of life. This query would be useful in the backend to build an understanding of what locations beneficiaries frequently relocate to, for government officials (our C-suite) to inform policy in these areas.

```
WITH Heart_Failure_surgery_each_state AS
(SELECT PA.city, PA.zip_code, PA.state, PT.provider_type, ARB.age_range, ARB.count,
MS.Number_of_medicare_beneficiaries_with_medical_services,
        ROW_NUMBER() OVER (PARTITION BY PA.state ORDER BY
```

```
        MS.Number_of_medicare_beneficiaries_with_medical_services DESC) AS Rank
    FROM providers P
    JOIN age_range_beneficiaries ARB ON P.npi = ARB.npi
    JOIN provider_addresses PA ON P.address_id = PA.address_id
    JOIN provider_types PT ON P.provider_type_id = PT.provider_type_id
    JOIN chronic_illness CI ON P.npi = CI.npi
    JOIN medical_services MS ON P.npi = MS.npi
    WHERE CI.chronic_illness = 'Heart Failure'
    AND ARB.age_range = 'Age Greater 84'
    AND ARB.count > 0
    AND PT.provider_type = 'Cardiac Surgery'
)
SELECT *
FROM Heart_Failure_surgery_each_state
WHERE Rank = 1
```

3. Explore illness type distribution based on provider state.

- This would provide analysts with an understanding of the frequent illnesses occurring for each provider state as a baseline for further analysis.

```
WITH illness_distribution (address_id, state, npi,chronic_illness) AS (
SELECT A.address_id, A.state, V.npi, C.chronic_illness
FROM provider_addresses A
JOIN  providers V
ON A.address_id =V.address_id
JOIN chronic_illness C
ON V.npi = C.npi)
SELECT state, chronic_illness,COUNT(*) AS amount
FROM illness_distribution
GROUP BY state, chronic_illness
ORDER BY amount DESC;
```

4. Explore illness type distribution based on the beneficiary age range

- This would provide analysts with an understanding of the frequent illnesses occurring for each beneficiary age range as a baseline for further analysis

```
WITH age_distribution (age_range, npi,chronic_illness) AS (
SELECT A.age_range, V.npi, C.chronic_illness
FROM age_range_beneficiaries A
JOIN  providers V
ON A.npi =V.npi
JOIN chronic_illness C
ON V.npi = C.npi)
SELECT age_range, chronic_illness,COUNT(*) AS amount
FROM age_distribution
GROUP BY age_range, chronic_illness
ORDER BY amount DESC;
```

5. Explore race distribution of beneficiaries of medicine based on the state location of the provider

- This would provide analysts with an understanding of the beneficiary demographics for each beneficiary race as a baseline for further analysis.

```
WITH race_distribution (address_id, state, npi,race) AS (
SELECT A.address_id, A.state, V.npi, R.race
FROM provider_addresses A
JOIN  providers V
ON A.address_id =V.address_id
JOIN race_beneficiaries R
ON V.npi = R.npi)
SELECT state, race,COUNT(*) AS amount
FROM race_distribution
GROUP BY state, race
ORDER BY amount DESC;
```

6. Explore the race distribution of beneficiaries that suffer from chronic diseases.

- This query would provide the C-suite audience with an understanding of what race demographics suffer from certain chronic illnesses. This could be used to adjust medicare payment plans accordingly.

```
WITH illness_race_types (npi, chronic_illness, race ) AS (
SELECT R.npi, R.race, C.chronic_illness
FROM chronic_illness C
JOIN  race_beneficiaries R
ON C.npi = R.npi)
SELECT chronic_illness, race,COUNT(*) AS amount
FROM illness_race_types
GROUP BY race, chronic_illness
ORDER BY amount DESC;
```

Implementation Plan

The dashboard for analysts allows views into baseline analytics upon which they can build further analysis. The dashboard also allows for query manipulations, so that analysts can alter the queries as they see fit. Visuals for the queries showcased above that include “analysts” within their descriptions are also included in the dashboard.

The dashboard for C-suite executives includes overviews and visualizations derived from the queries above that contain “C-suite” within their description. They will not be granted access to the database or query manipulation.

Future dashboards could also be built for other partners. It is important to understand the use of dashboards beyond the audience of analysts and business executives or officials. For example, Health Insurance Providers would benefit from the insights provided by our data. The structure

for their potential dashboard could be to view insights for their company, with understandings of US and regional statistics. It would be relatively similar to the structure of dashboards for state government officials.

Redundancy and Performance

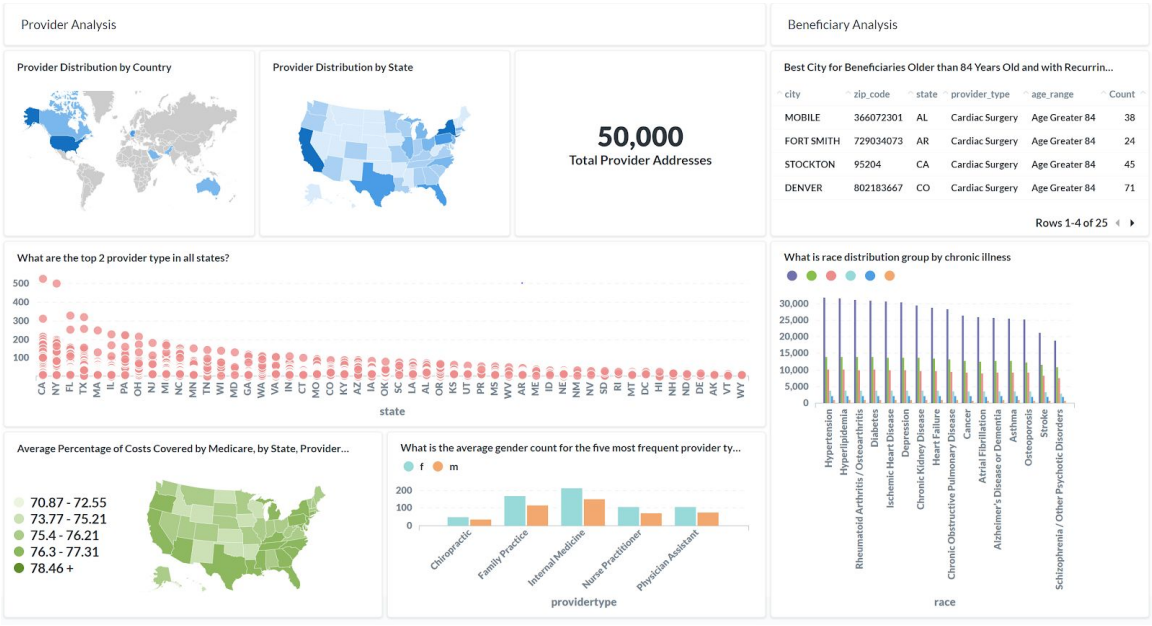
We selected partial observations and checked running speed, which does not take a significant amount of time. If there are any speed issues going forward, we could use web services such as cloud service to boost efficiency. For the current dataset, we used views to simplify many steps when analyzing the dataset.

For redundancy, we already broke the dataset into multiple tables and leveraged a normalization plan to make information more clear and precise. For example, what we did in the race_beneficiaries table and age_range_beneficiaries table, we combine original multiple columns into one column and set new variables to distinguish data types more easily.

Dashboard Explanations

Dashboard for the C-suite (government policymakers)

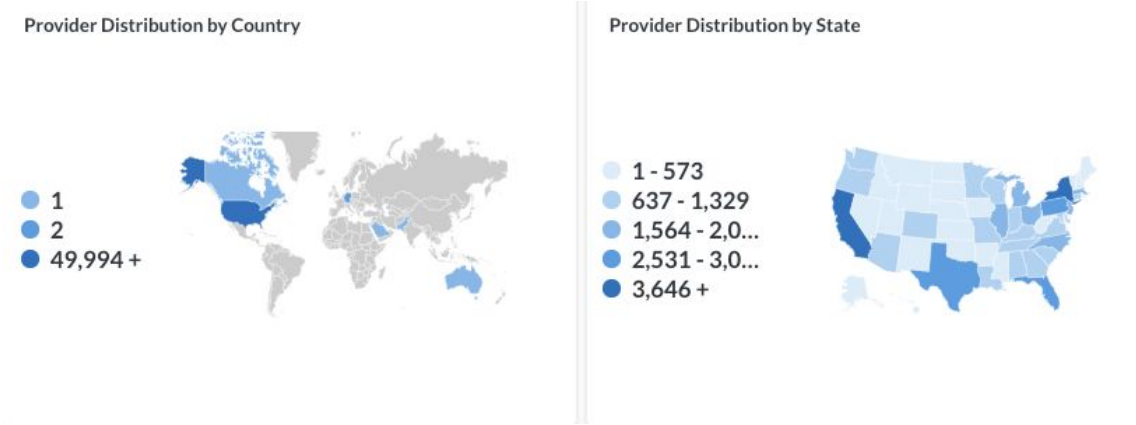
Below is a snapshot of the dashboard created for government policymakers in Metabase. As previously stated, this dashboard is interactive and provides policymakers with an understanding of the constituents impacted by medicare. The visualizations fall within two categories: Provider Analysis and Beneficiary Analysis.



Dashboard Link: http://f19server.apan5310.com:3201/public/dashboard/3329d04b-3677-4466-bac0-4816e979ec27

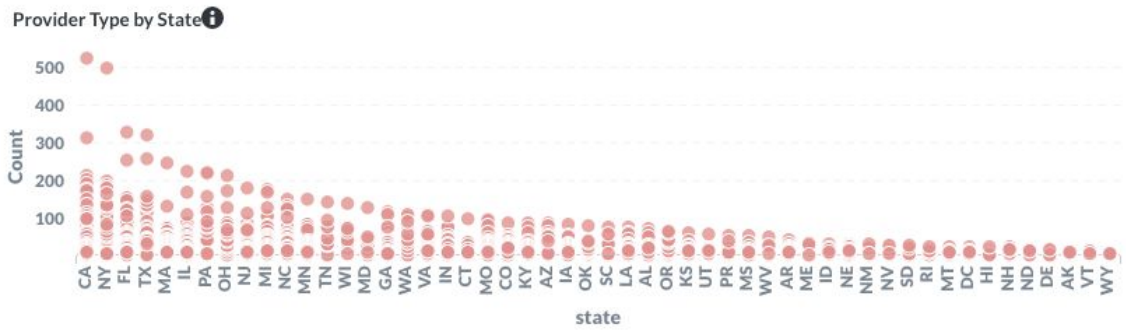
Provider Analysis

What is the geographical area that has the highest provider density?



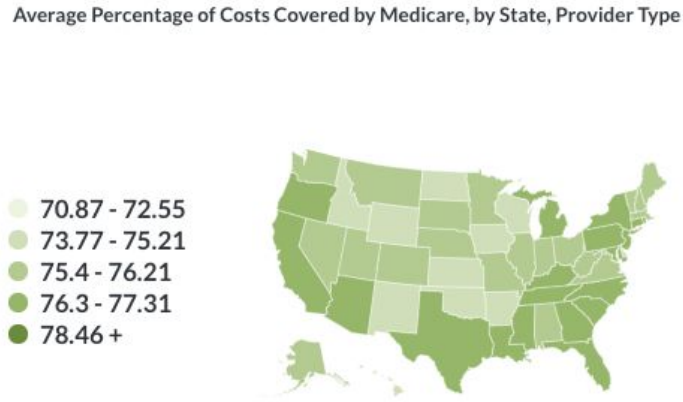
Though Medicare is funded by the US Government, there are provider addresses that exist external to the United States. We are focusing our understanding on US addresses. This map illustrates the geographical distribution of US medicare providers. The density, which is reflected by color shade, corresponds with the economic development, medicare popularity and population density. New York and California have over 3600 providers and followed by Texas, Florida and Pennsylvania.

What are the top two provider types in all states?



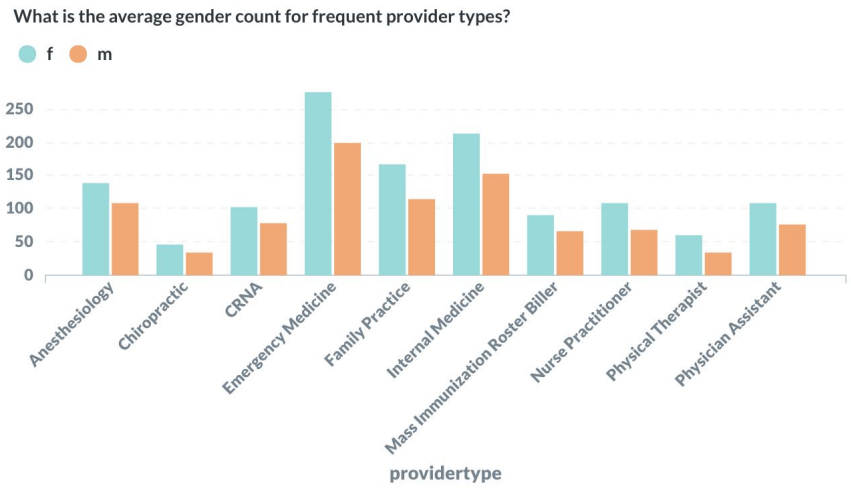
Provider types have varied significantly between states due to the difference in the constituency of the state, location of illnesses, and other factors. Though some states have a greater variety of provider types, the top two provide types in all states are internal medicine and family practice.

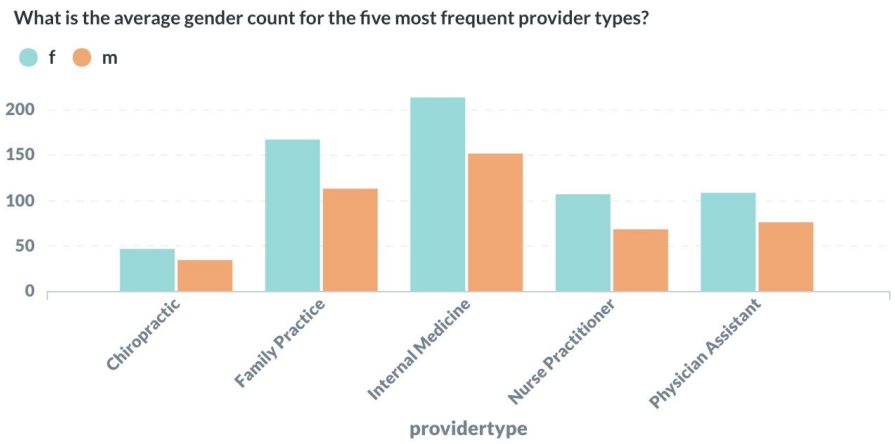
What are the top three states with the highest percentage of costs covered by Medicare and states with the lowest average percentage of costs covered by Medicare?



In this visualization, we took a closer look at the average percentage of costs covered by medicare in different states. For context, the average percentage of Medicare covered is calculated by (Medicare payment / Medicare Allowed)*100. The darker the color of the state is, the higher the average medicare coverage percentage is in that state. Many states have an average coverage percentage of over 78.46%, including many southern states, western states, and northeastern states. On the other hand, states in middle America have medium to low percentages for the average costs covered by Medicare. If we take a closer look at the data, Southern Carolina (SC), New Jersey (NJ) and Florida (FL) are the three states with the highest coverage percentage. On the contrary, Wyoming (74.55%), North Dakota (74.22%), and Hawaii(73.77%) are the three states ranked the lowest coverage percentage.

What is the average gender count for the five most frequent provider types?

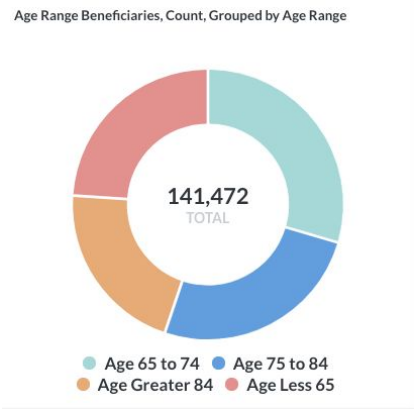




Based on the chart above, we noticed that for females, Medicine (276.03), Internal Medicine (213.81), Family Practice (166.9), Anesthesiology (138.66), and Nurse Practitioner (107.83) are the top five most frequent provider types. For males, Emergency Medicine (199.39), Internal Medicine (152.03), Family Practice (114.01), Anesthesiology (108.24), and CRNA (78.68) are the top five most frequent provider types. If looking into the five most frequent provider types without gender separation, Chiropractic, Family Practice, Internal Medicine, Nurse Practitioner, and Physician Assistant are the top five frequent provider types.

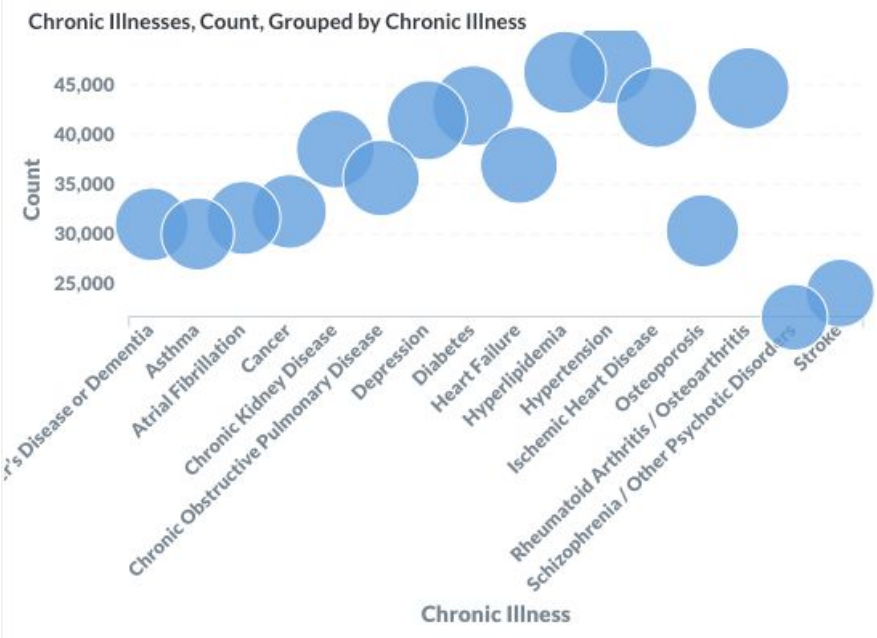
Beneficiary Analysis

What beneficiary age range holds the largest population?



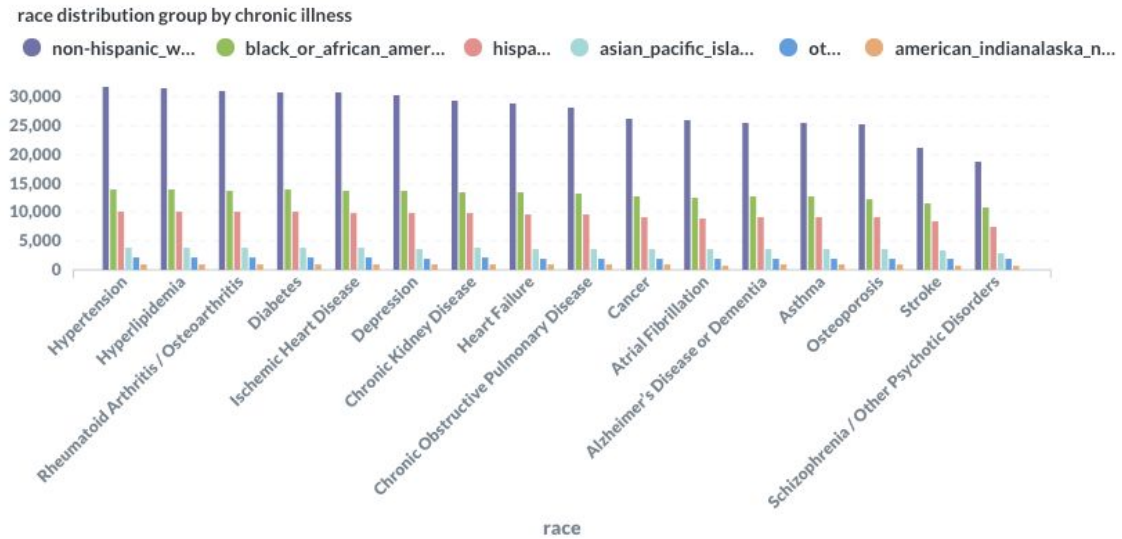
In this chart, we aim to provide the age range distribution among our beneficiaries. The result indicates an aging population of beneficiaries with four age groups: less than 65, 65 - 74, 75 - 84, and greater than 84. Among all four groups, age 65-74 took account of 29,6% of the total 141,472 records, followed by age 75-84 (25.6%), age less than 65 (23.9%), and age greater than 84 (20.9%).

What are the three highest occurring and lowest occurring Chronic Illnesses?



This chart shows how the frequency of chronic illnesses is distributed among beneficiaries. Most chronic illnesses have more than 30,000 occurrences within the beneficiary population. In particular, Hypertension (47,336), Hyperlipidemia(46,358), and Rheumatoid Arthritis / Osteoarthritis(44,730) ranked the top three chronic illnesses. On the contrary, Schizophrenia / Other Psychotic Disorders (21,579), Stroke (24,030), and Asthma (29,994) ranked the lowest three chronic illness appearances. It would be important for policymakers to understand which diseases are frequent within the people they directly impact.

What is the distribution of race for different chronic illnesses?



This chart presents illness distribution in ethnic races. All chronic illnesses follow the same order of frequency; the order of count by race from highest to lowest occurs in the following order: non-Hispanic White, Black/African American, Hispanic, Asian/Pacific Islander, Other, then American Indian/Alaska Native. Besides the differences in races, this result also could be potentially affected by the amount of population for each race. For example, non-Hispanic white is the majority race in the US with more than 70% of total US population.

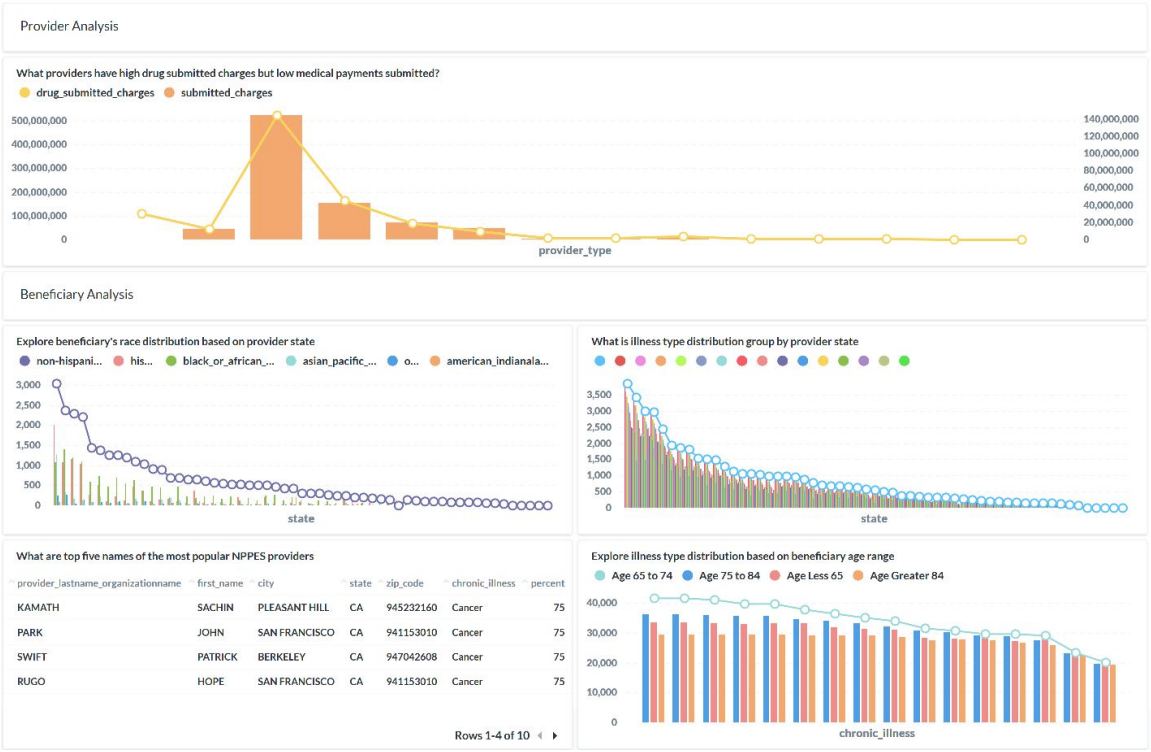
What are the cities with the high heart-related diseases patients density?

city	zip_code	state	provider_type	age_range	计数	number_medical_services	r.
MOBILE	366072301	AL	Cardiac Surgery	Age Greater 84	38	3,999	1
FORT SMITH	729034073	AR	Cardiac Surgery	Age Greater 84	24	549	1
STOCKTON	95204	CA	Cardiac Surgery	Age Greater 84	45	794	1
DENVER	802183667	CO	Cardiac Surgery	Age Greater 84	71	1,491	1
NEWARK	197180002	DE	Cardiac Surgery	Age Greater 84	12	308	1
JACKSONVILLE	322044735	FL	Cardiac Surgery	Age Greater 84	27	820	1
KANKAKEE	60901	IL	Cardiac Surgery	Age Greater 84	357	7,518	1
MERRILLVILLE	464106693	IN	Cardiac Surgery	Age Greater 84	57	732	1
WICHITA	672262019	KS	Cardiac Surgery	Age Greater 84	36	781	1
LONDON	40741	KY	Cardiac Surgery	Age Greater 84	15	798	1
NEW ORLEANS	701212429	LA	Cardiac Surgery	Age Greater 84	16		1
LANHAM	207061104	MD	Cardiac Surgery	Age Greater 84	12	755	1
SAINT CLAIR SHO...	480801181	MI	Cardiac Surgery	Age Greater 84	35	1,853	1
GULFPORT	39503	MS	Cardiac Surgery	Age Greater 84	86	5,066	1
GREENSBORO	274011230	NC	Cardiac Surgery	Age Greater 84	13	310	1
OMAHA	681312137	NE	Cardiac Surgery	Age Greater 84	11	453	1
MONTCLAIR	070422629	NJ	Cardiac Surgery	Age Greater 84	269	4,005	1
NEW YORK	100654870	NY	Cardiac Surgery	Age Greater 84	48	1,024	1
CINCINNATI	452192906	OH	Cardiac Surgery	Age Greater 84	22	411	1
OKLAHOMA CITY	731124430	OK	Cardiac Surgery	Age Greater 84	15	546	1
SOMERSET	155012271	PA	Cardiac Surgery	Age Greater 84	144	1,414	1
CLARKSVILLE	370434523	TN	Cardiac Surgery	Age Greater 84	14		1
HOUSTON	770302607	TX	Cardiac Surgery	Age Greater 84	32	425	1
ARLINGTON	222053610	VA	Cardiac Surgery	Age Greater 84	28	451	1

As the table shows above, all cardiac patients are under the Age Greater than 84 group and distributed in large metropolitan cities within each state, where superior medical resources are located.

Dashboard for Analysts

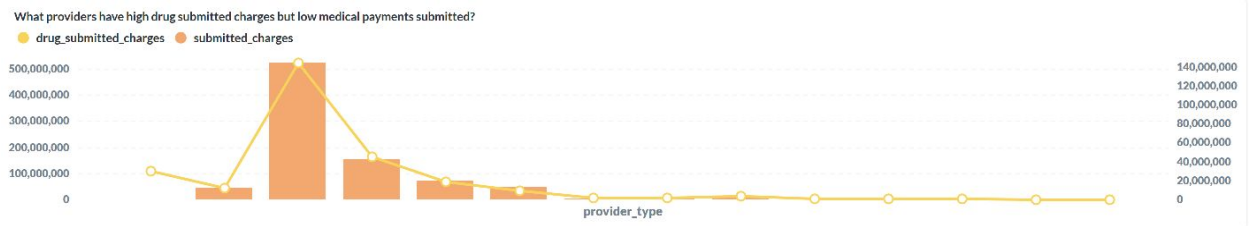
Below is a snapshot of the dashboard created for government analysts in Metabase. As previously stated, this dashboard is interactive and provides analysts with a broad overview of demographic data and payment information concerning medicare. The visualizations below fall within two categories: Provider Analysis and Beneficiary Analysis.



Dashboard Link: http://f19server.apan5310.com:3201/public/dashboard/c20cd3d5-bea7-43c1-a2e4-b64444a7e2a8

Provider Analysis

What providers have high drug submitted charges but also high medical payments submitted?



This metric is useful, as it provides an understanding of which areas of practice have incredibly high drug submitted charges and low medical submitted charges. This could potentially be indicative of abuse of medicare for drugs. However, there are also provider types that have an understandably higher proportion of drug submitted charges such as Hematology/Oncology. It should be noted that the maximum value on the index for drug submitted charges is $500,000,000 while the maximum value on the index for medical submitted charges (submitted_charges on the plot) is $40,000,000. The discrepancy between the two is the most impactful for business value. The query used to build this visualization provides a list of all providers where the

drug_submitted_charges is over double the medical submitted_charges. This list could be used by analysts to focus their search for fraud or abuse by Medicare patients and providers.

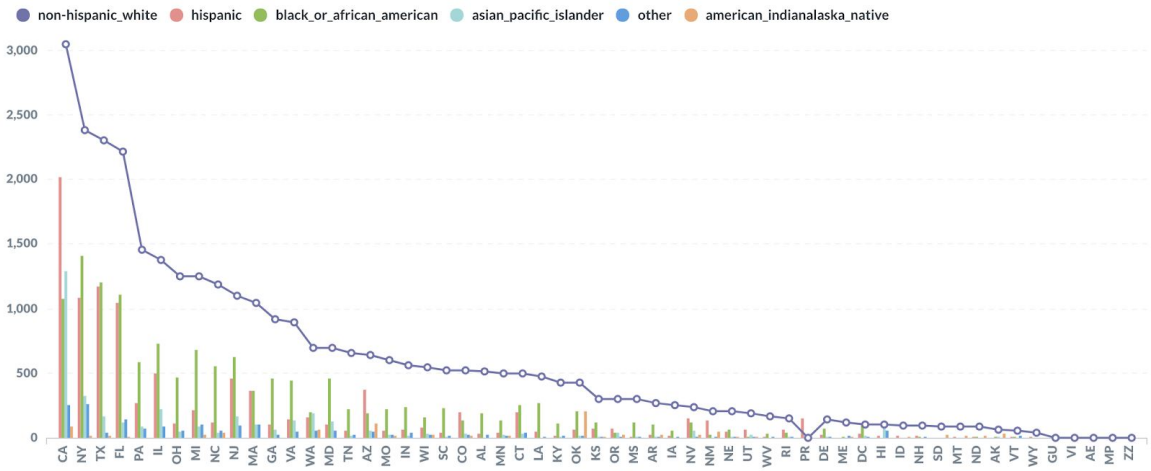
What are the popular NPPES providers for cancer patients in California?

provider_lastname_organizationname	first_name	city	state	zip_code	chronic_illness	percent
KAMATH	SACHIN	PLEASANT HILL	CA	945232160	Cancer	75
PARK	JOHN	SAN FRANCISCO	CA	941153010	Cancer	75
SWIFT	PATRICK	BERKELEY	CA	947042608	Cancer	75
RUGO	HOPE	SAN FRANCISCO	CA	941153010	Cancer	75
BOTNICK	MARC	BURBANK	CA	915054809	Cancer	75
BERKOWITZ	ARNOLD	MURRIETA	CA	925635561	Cancer	75
GOLDSMITH	BRIAN	CAMERON PARK	CA	956828237	Cancer	75
HILL	DENNIS	OAKLAND	CA	946093480	Cancer	75
LANG	JULIE	LOS ANGELES	CA	900890112	Cancer	74
PHILBEN	VICKI	REDDING	CA	960021800	Cancer	73

Since California is the state with the highest prevalence of illness, we decided to analyze what providers were most popular in that state. This way, analysts can determine which providers need the greatest amount of funds allocated for their beneficiaries. Furthermore, analysts can use this information to provide reports to beneficiaries regarding the most popular provider locations for certain illnesses.

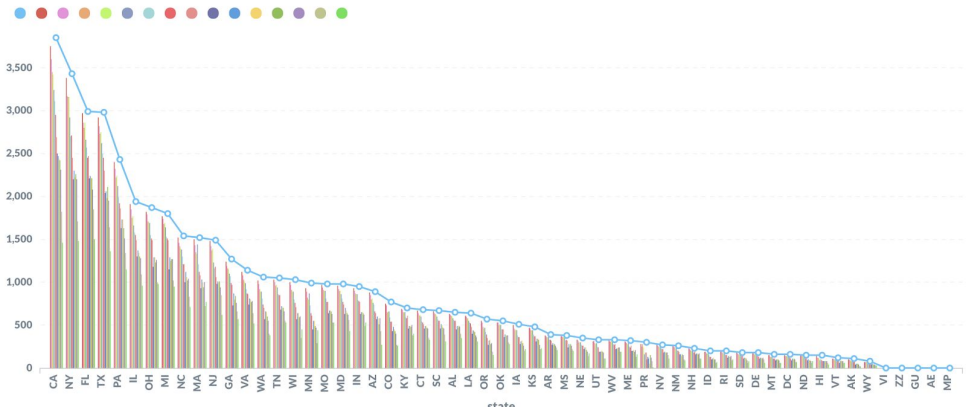
Beneficiary Analysis

What is the race distribution based on the provider state?



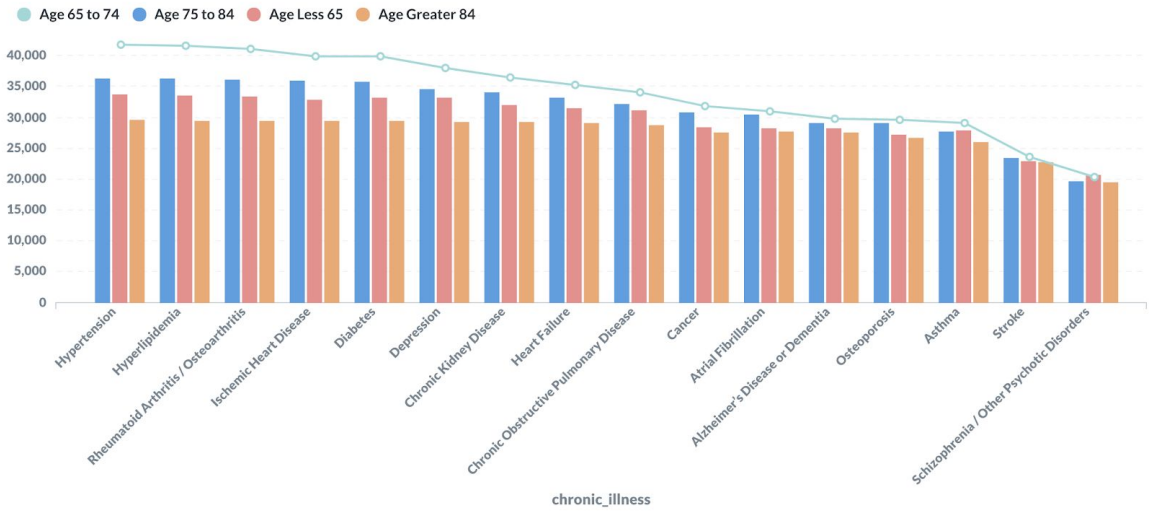
Based on the graph above, we can tell that the non-Hispanic white demographic has the highest distribution in most states. However, other races are distributed differently in each provider state. For example, Hispanic people are ranked as the second-largest race demographic in California, while African American is the second-largest in New York. This visual is useful for analysts to have a general understanding of population demographics, especially if there are patterns found in the data concerning medicare expenditures related to racial demographics or location. This way, they can pass the information on to policymakers to help allocate Medicare funds where needed.

What are the three states with the highest illness occurrence?



If we group chronic illnesses by provider states, the resulting three states with the highest illness rates are California, New York, and Florida. These results are skewed based on differences in state population, economic developments, and local dietary habits among states. However, these statistics provide insight into the geographical areas where Medicare coverage is most needed.

What is the illness type distribution based on beneficiary age range?



This graph illustrates illness type distribution based on different beneficiary age ranges. The age group 65-74 has the highest occurrence for all illnesses listed, followed by age group 75-84, ages less than 65 and ages greater than 84. However, the Age less than 65 group has a higher occurrence of Asthma and Schizophrenia/other Psychotic disorders than the age group 75-84. Analysts can use this information to help C-suite execs determine what age demographics need the most funds based on chronic illness distribution.

Benefits of the above dashboards

These dashboards represent the high-level insights into our analytical results from the raw data we obtained. Our client, the US government, will be able to digest these insights and transform them into actions for future medicare policymaking. For example, provider and beneficiary analysis will be a great reference for the government's medicare resource allocation.

Conclusion

Our goal was to create a comprehensive database along with dashboards providing high-level visualizations for government officials and analysts. We achieved these goals by cleaning the extensive dataset, structuring a logical table format, creating the database, conducting extensive ETL, and developing 10 complex SQL queries to derive valuable insights from the data to assist with dashboard creation. Creating a RDMS allowed analysts to conduct more complex queries and give a more detailed account of Medicare usage. The ETL process made it possible to give unique identifiers for specific attributes, simplifying the structural logic of the data. Analysis of our database provided our clients with valuable information regarding beneficiaries, providers, and overall Medicare expenditures. Overall, this project gave our clients a better understanding of Medicare data, allowing government officials to make more informed decisions regarding Medicare policy.

References

Centers for Medicare & Medicaid Services. (2019, April 15). Medicare Physician & Other Supplier NPI Aggregates. Retrieved from https://www.kaggle.com/cms/medicare-physician-other-supplier-npi-aggregates.

@jcubanski, J. C. F., Neuman, T., & Freed, M. (2019, August 20). The Facts on Medicare Spending and Financing. Retrieved from https://www.kff.org/medicare/issue-brief/the-facts-on-medicare-spending-and-financing/.

Government Watchdog Report Outlines Problems with Current Medicare Plan Finder As New Version Awaits Debut. (2019, August 8). Retrieved from https://www.medicareadvocacy.org/government-watchdog-report-outlines-problems-with-current-medicare-plan-finder-as-new-version-awaits-debut/.