Exploring the Cost-Effectiveness of the United States Medicare Program

Final Project Report

Spring 2019 - BIG DATA ANALYTICS - (BYGB-7990-003)

Professor Yilu Zhou, Ph.D.

Group 15: Ila Srivastava, Katie Linh Cao, Lakshmi Sravya Katta, Tudor Tarina

1. **Executive Summary**
   1. **Problem:** Introduced in the United States since 1966, Medicare is a national social insurance program administered by the government to provide health insurance for Americans aged 65 and older who have paid into the system through the payroll tax1. The argument is that while the intention of the program is good, the system designed behind Medicare is not the most efficient. The goal of affordable medical care is being challenged by rapidly increasing costs and lack of readily available resources and access across all states.
   2. **Dataset:** The dataset is sourced from Kaggle website. This is a public dataset created by the Centers for Medicare & Medicaid Services. It summarizes the utilization and payments for procedures, services and prescription drugs provided to Medicare beneficiaries by hospitals, physicians and other suppliers.
   3. **Methodology (Tools and Algorithms):** This project performs an exploratory analysis in order to gain insight from the large Medicare database using a newly-introduced tool – Google BigQuery to query the information needed. Tableau is then the next step for visualization.
   4. **Key Findings:** Our key findings include:
      1. Coastal states, such as California, New York and Massachusetts, as well as the District of Columbia have the highest In-Patient costs per person. Kentucky, Oklahoma, Mississippi, Wyoming, the District of Columbia have the highest Out-Patient costs per person.
      2. Respiratory-related drugs cost the highest, when compared to drugs for other ailments.
      3. There is little variation in Total Average Medicare Payment for each drug over the years.
2. **Business Problem**

This project will provide an overview understanding of the current Medicare system from a cost perspective and how it spans across states and cities, down to drugs and medical provider levels. Findings will be beneficial for a variety of businesses and individuals, ranging from the policymakers, the medical provider, state legislators, patients, and so on. We have identified the three main goals for the project and will perform descriptive analytics based on these goals.

* 1. **Project Goal #1:** Explore and compare the average cost for inpatient and outpatient treatment in each state over the years.

**Rationale:** In-patient is a patient who stays in a hospital while under treatment and out-patient is who receives medical treatment without being admitted to a hospital. Comparing the average cost of the two types of patients will show interesting insight in the structure of the system in each state.

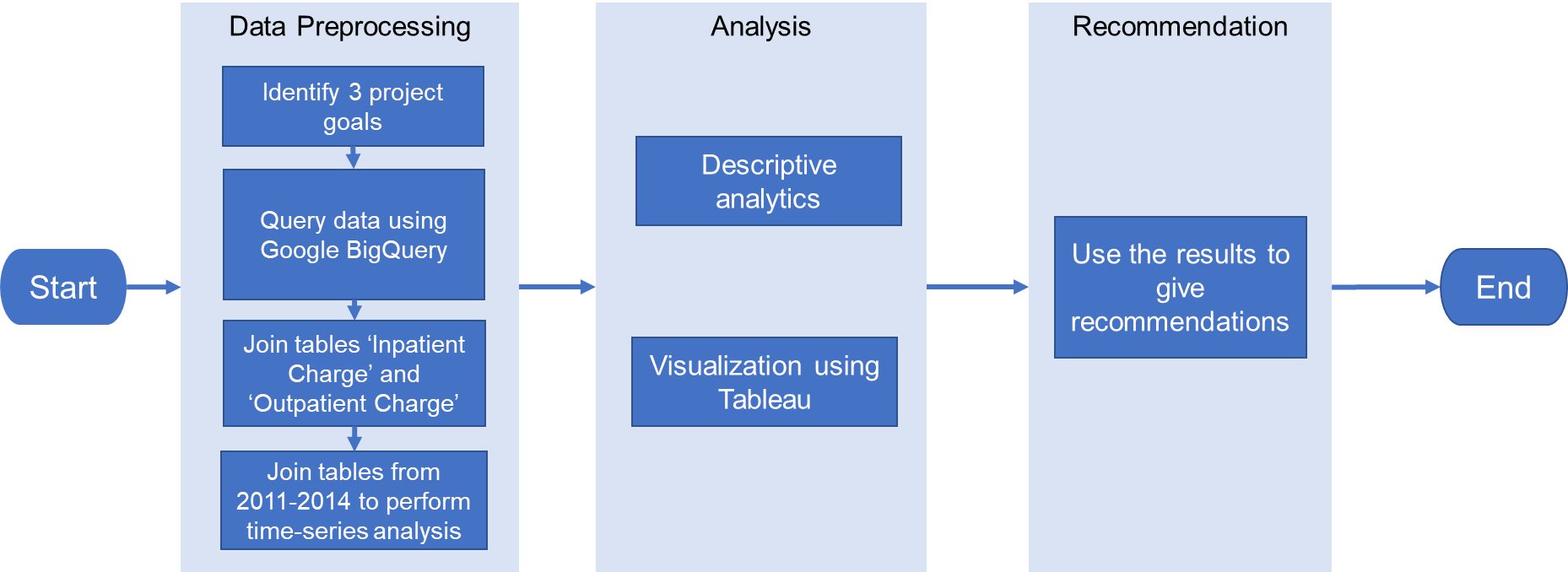
* 1. **Project Goal #2:** Explore the cities with the greatest number of cases for each diagnostic condition.

**Rationale:** This could reveal differences between segments of the United States population.

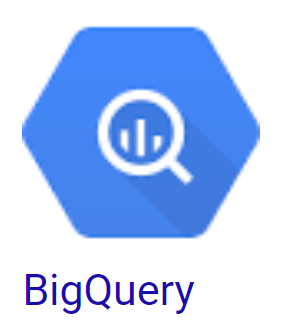
* 1. **Project Goal #3:** Explore the average payments for these conditions in these cities and how they compare to the national average.

**Rationale:** Understanding which diseases and conditions are prevalent in certain areas and how high the costs in these areas are in order to create policy to help bring the cost down.

1. **Dataset Description**
   1. **Data Source:** https://www.kaggle.com/cms/cms-Medicare
   2. **Data Format:** Tables via Google BigQuery
   3. **Data Size:** 13 GB
2. **System Design**

****

1. **Data Preprocessing**
   1. **Preprocessing:** Since the group is using Google BigQuery, we will be able to select and query only the necessary data and columns for each Project Goal, with minimal preprocessing.
   2. **Join** two tables ‘Inpatient Charge’ and ‘Outpatient Charge’ for comparison
   3. **Join** tables from different years (2011-2014) to perform time-series analysis
2. **Algorithm Development**
   1. **Google BigQuery**

****It is an [enterprise data warehouse](https://cloud.google.com/solutions/bigquery-data-warehouse) that solves this problem by enabling super-fast SQL queries using the processing

power of Google's infrastructure. You can control access to both the project and your data based on your business needs, such as giving others the ability to view or query your data. You can access BigQuery by using the [GCP Console](https://console.cloud.google.com/bigquery) or the [classic web UI](https://bigquery.cloud.google.com/), by using a [command-line tool](https://cloud.google.com/bigquery/docs/cli_tool), or by making calls to the [BigQuery REST API](https://cloud.google.com/bigquery/docs/reference/v2) using a variety of [client libraries](https://cloud.google.com/bigquery/docs/reference/libraries) such as Java, .NET, or Python. There are also a variety of [third-party tools](https://cloud.google.com/bigquery/third-party-tools) that you can use to interact with BigQuery, such as visualizing the data or loading the data.

* 1. **When to use Google BigQuery?**

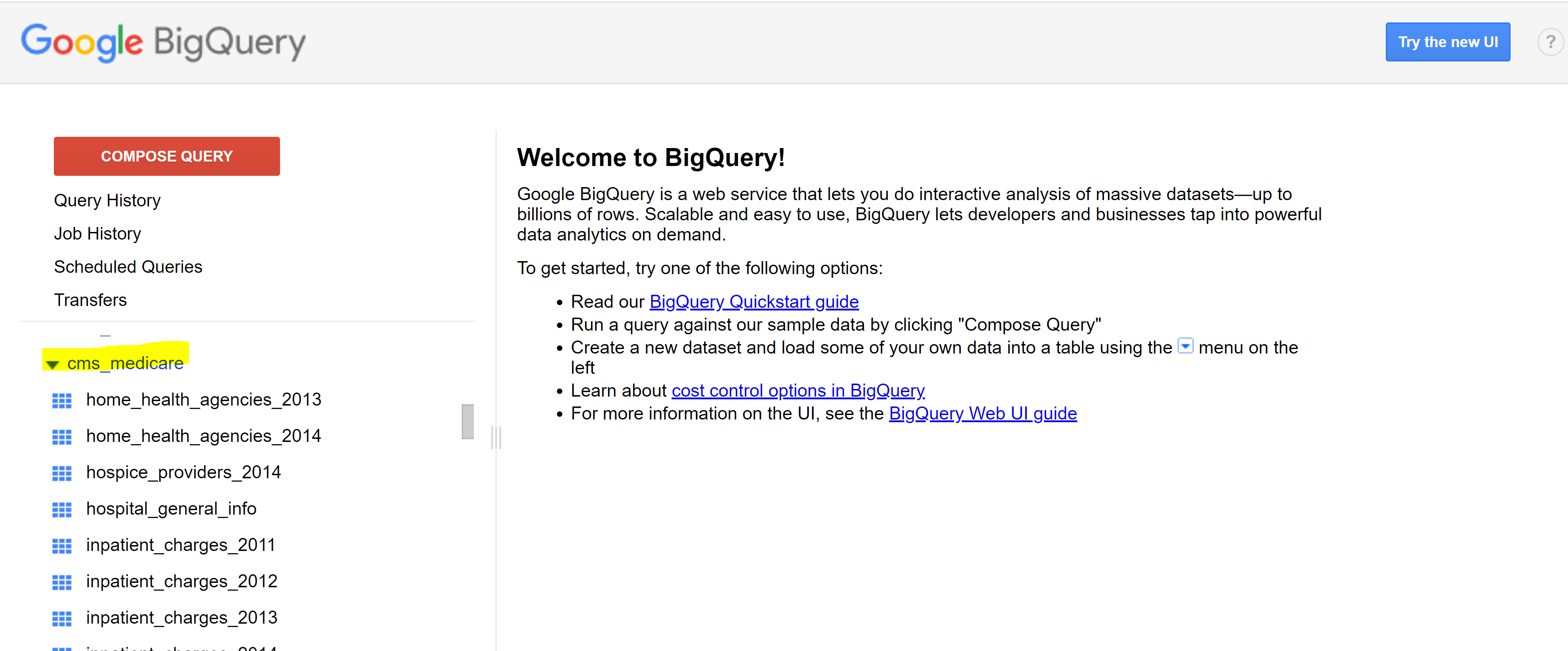
BigQuery typically comes at the end of the Big Data pipeline. It is not a replacement for existing technologies, but it complements them very well. Real-time streams representing sensor data, web server logs or social media graphs can be ingested into BigQuery to be queried in real time. After running the ETL jobs on traditional RDBMS, the resultant data set can be stored in BigQuery. Data can be ingested from the data sets stored in Google Cloud Storage, through direct file import or through streaming data.

Our study involves using Google BigQuery to explore the complete dataset to understand the cost-effectiveness of Medicare by analyzing the data generated from various queries. We used publicly available dataset in BigQuery known as *cms\_Medicare*. There are two user interfaces that can be used to access the public datasets:

1. The [BigQuery web UI](https://console.cloud.google.com/bigquery)
2. The [classic web UI](https://bigquery.cloud.google.com/)

We used classic web UI. The BigQuery-public-data project is automatically pinned to every project in both UIs. You can find the project in the navigation pane.

The size of the dataset is 13GB and comprises of 23 tables. The data summarizes the utilization and payments for procedures, services, and prescription drugs provided to Medicare beneficiaries by specific inpatient and outpatient hospitals, physicians, and other suppliers. The dataset includes the following data: Common inpatient and outpatient services All physician and other supplier procedures and services All Part D prescriptions. Providers determine what they will charge for items, services, and procedures provided to patients and these charges are the amount that providers bill for an item, service, or procedure.



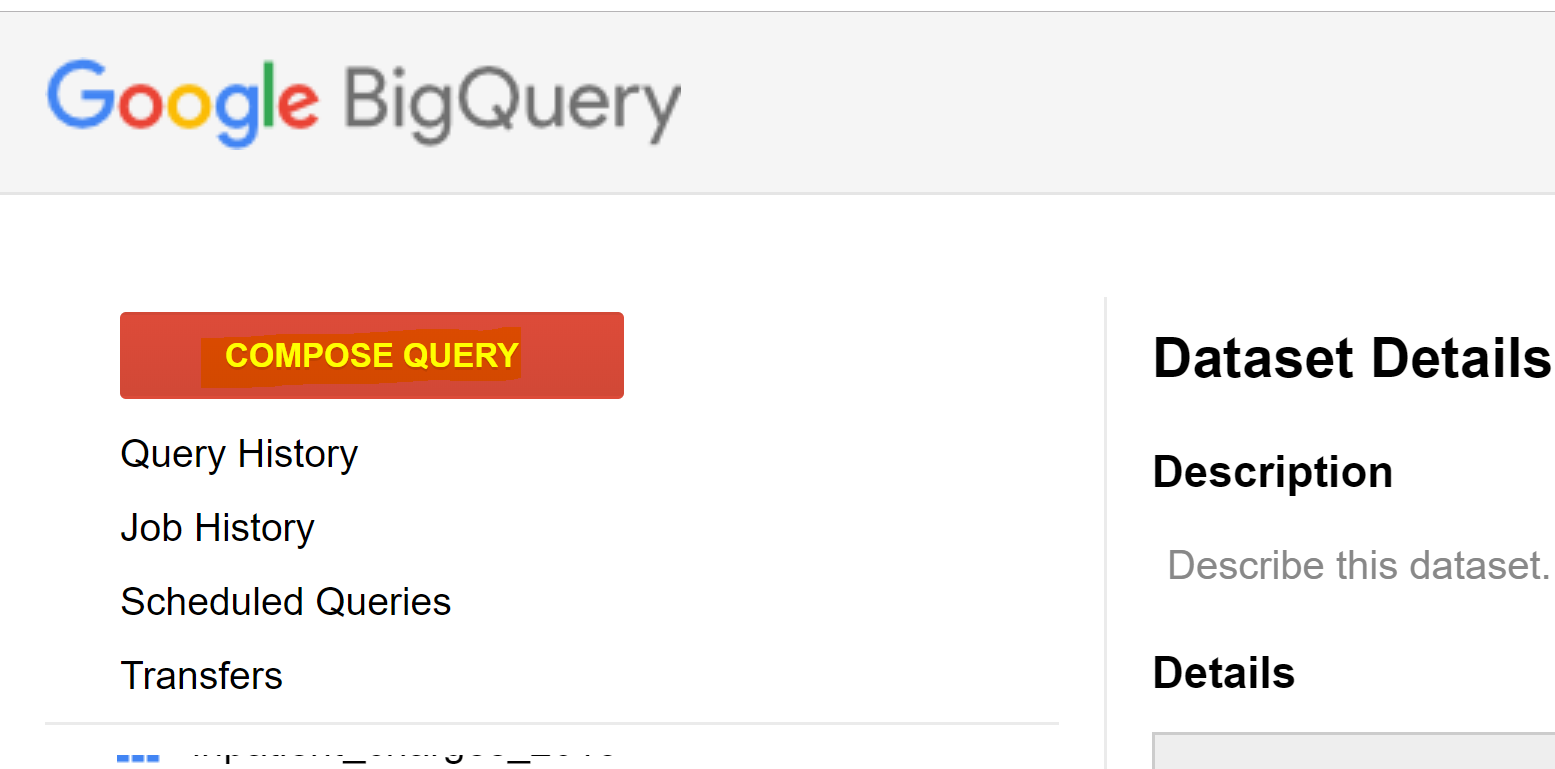
The goal of the project is to develop queries and generate output files which is further uploaded on visualization tool *Tableau* to perform descriptive analytics and come up with useful insights to bring improvements in the Medicare area. The project involved generating queries based on the following goals:

1. Variation in total average Medicare payment for each provider with specific drug for inpatients over year.
2. Top 50 provider names along with state and corresponding year with total Medicare submitted amount for each type of provider for that state and organization name.
3. What are the average payments for these conditions in these cities and how do they compare to the national average?
4. Variation in total average Medicare payment for each provider with specific drug for outpatients over year.
5. What is the total number of medications prescribed in each state?
6. What is the most prescribed medication in each state?
7. What is the average cost for inpatient and outpatient treatment in each City and State?
8. Ranking of conditions and comparison of average payments by city.

Let’s generate the data for one of our goals i.e. *What is the total number of medications prescribed in each state?*

On the top of left-hand side menu bar there’s a big red button *COMPOSE QUERY,* by clicking which you can write queries in the pop-up window. In the middle under your API Project you can check all your datasets and tables. On the bottom these are some sample large dataset provided by Google.

**1**



In the *New Query* window, type the following SQL query:

**2**

SELECT

nppes\_provider\_state AS state,

ROUND (SUM (total\_claim\_count) / 1e6) AS total\_claim\_count\_millions

FROM `bigquery-public-data.cms\_Medicare.part\_d\_prescriber\_2014`

GROUP BY

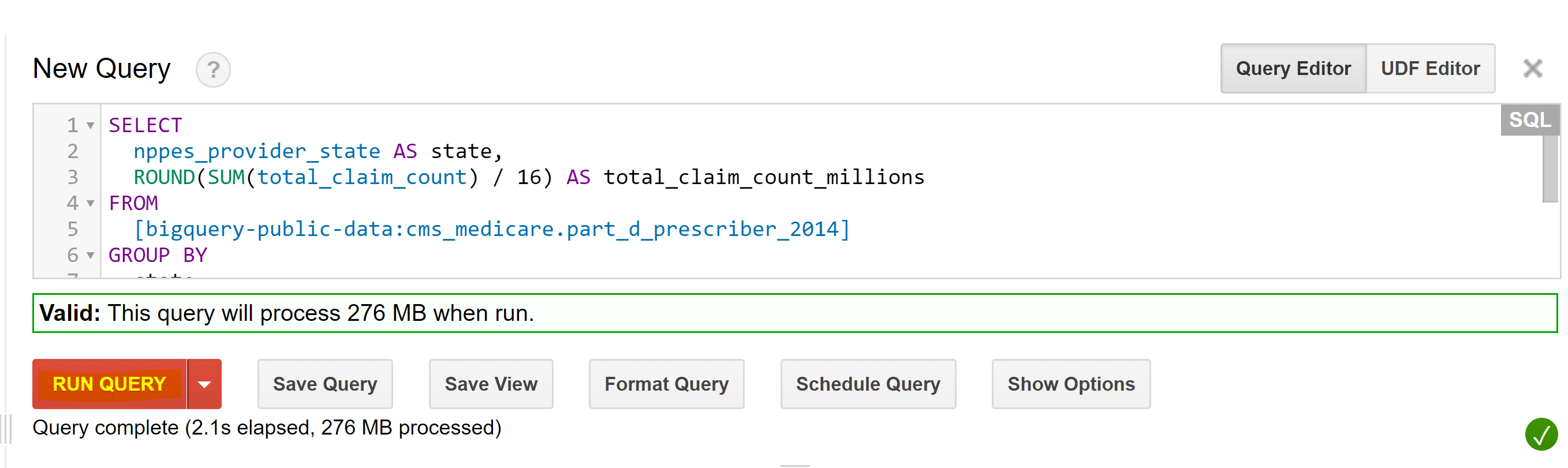
state

ORDER BY

total\_claim\_count\_millions DESC

LIMIT 5;

After writing the query click on *RUN QUERY***.** The console also displays the memory that will be consumed when running the query.



When the query is successfully executed the output is displayed below the *New Query* terminal window. By selecting on the options, you want to save your file in you can download your resultant data. We will select *Download as csv*.

**3**



The output is then uploaded to Tableau to perform descriptive analytics.

1. **Results and Evaluation**

Descriptive analytics will be the main component of this project, and it will be summarized in a series of ranked lists and visualization charts of cities and states with the greatest number of cases for each diagnostic condition. For example, we will visualize our data on the US geographical map, with colors and spikes representing different values.

* 1. Through descriptive analytics we should then be able to assess the current US Medicare program to provide policy recommendations to the US federal government and business recommendations to private insurance companies.
  2. We could also identify the major components of US Medicare spending, as well as the areas lacking funds.

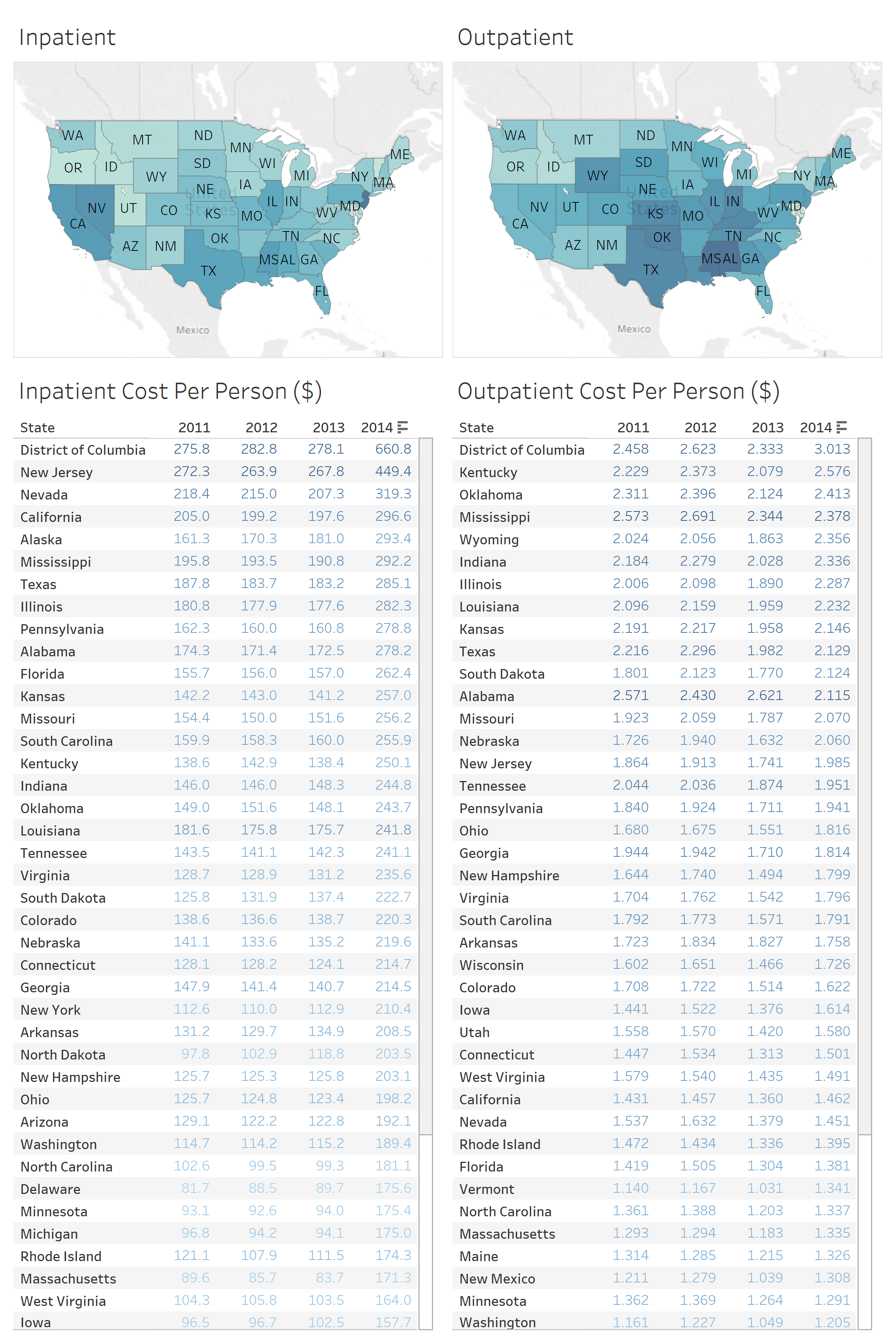


Figure 1: Tableau Chart of the Overview Medicare In-Patient versus Out-Patient Cost per Person across States

1. Taking the 65 and older population dataset from 2011 to 2014 from the US Census website, we standardized the cost of in-patient and out-patient per person and compared across states. We found that for in-patient cost, DC, New Jersey, Nevada, California have the highest in-patient cost per person throughout the years. For out-patient cost, DC, Kentucky, Oklahoma, Mississippi, Wyoming have the highest out-patient cost per person throughout the years. We also saw a jump in 2014 in-patient cost compared to previous years. This raises further research question into the factors that cause this phenomenon.
2. This finding is noteworthy because the structure of out-patient cost differs from in-patient cost. States in the Midwest seem to incur higher out-patient cost per population. This shows that the current Medicare system in these states might not be as effective compare to other states, even though the cost of living there is relatively lower.

**Query 1:** *Variation in total average Medicare payment for each provider with specific drug for inpatients over year*

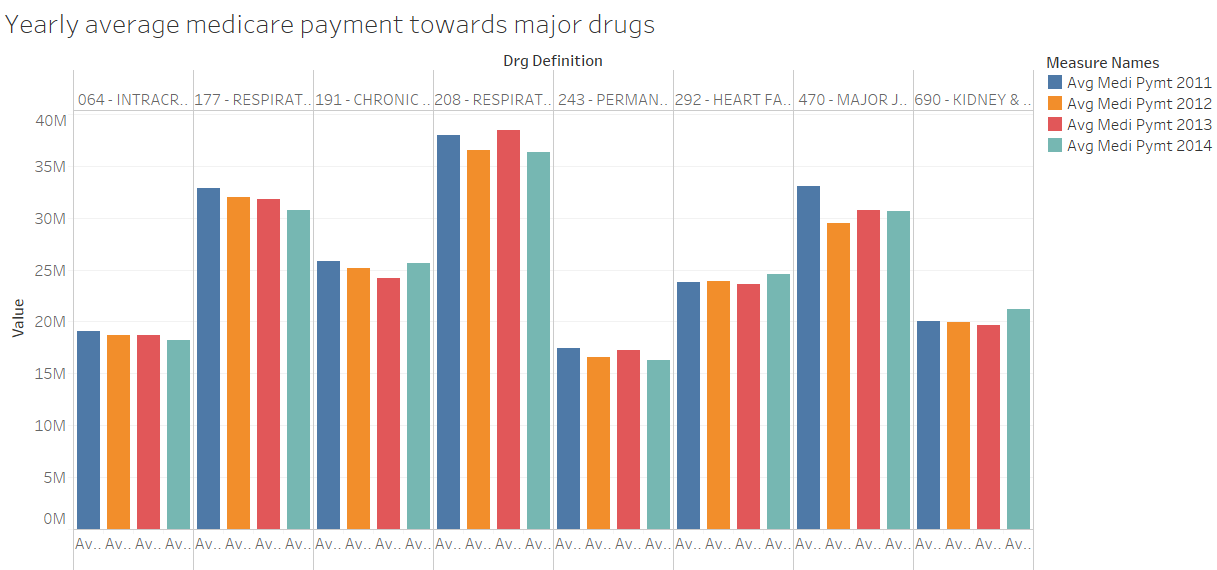


Figure 2: Tableau Chart of Yearly Average Medicare Payment Towards Major Drugs

1. We can infer that the yearly variation for average Medicare payment for drugs mentioned above is minor. It can be noticed that Respiratory drugs cost the highest in comparison to drugs for other ailments.

**Query 2:** *Top 50 provider names along with state and corresponding year with total Medicare submitted amount for each type of provider for that state and organization name*

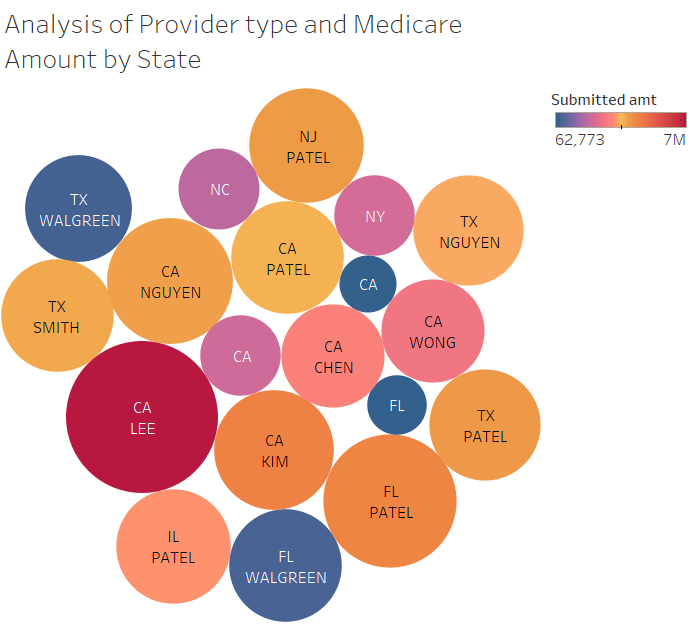


Figure 3: Tableau Chart of the Analysis of Provider Type and Medicare Amount by US State

1. Most of top providers belong to bigger states like California, Texas, Florida in terms of total Medicare amount submitted. Largest provider is Lee from California whose submitted amount reaches $7M.

**Query 3:** *What are the average payments for these conditions in these cities and how do they compare to the national average?*

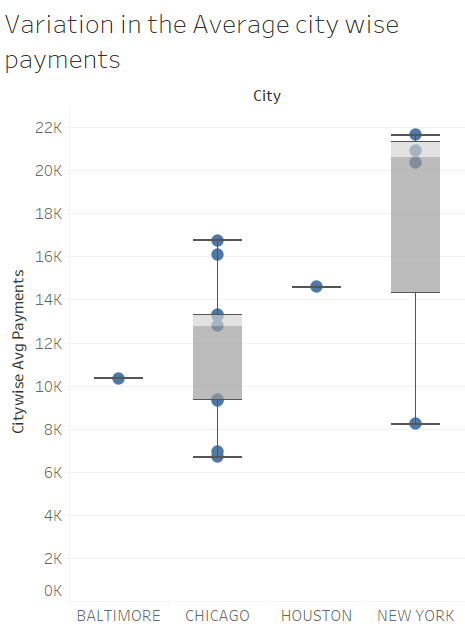


Figure 4: Tableau Chart of the Variation in the Average City-wise Payments

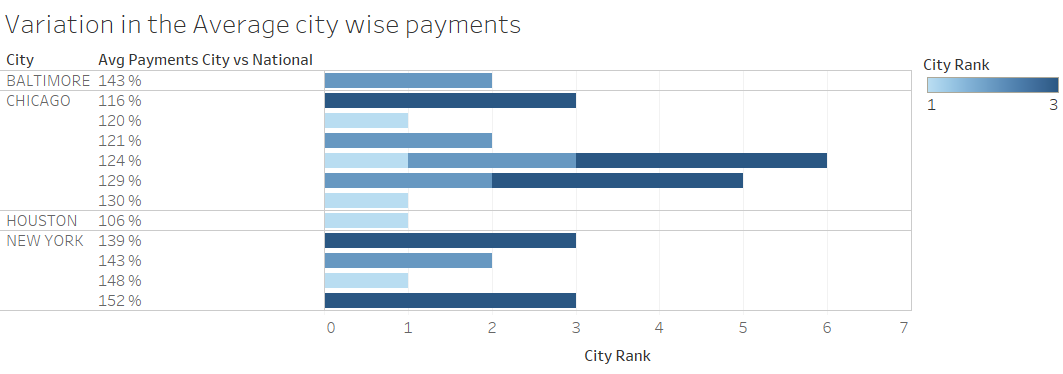


Figure 4A: Tableau Chart of the Variation in the Average City-wise Payments

1. New York has the highest average payments and sees a lot of variation in terms of payments for its residents. When compared to national average, some New York residents pay 152% of the national average payment.

**Query 4:** *Variation in total average Medicare payment for each provider with specific drug for outpatients over year*

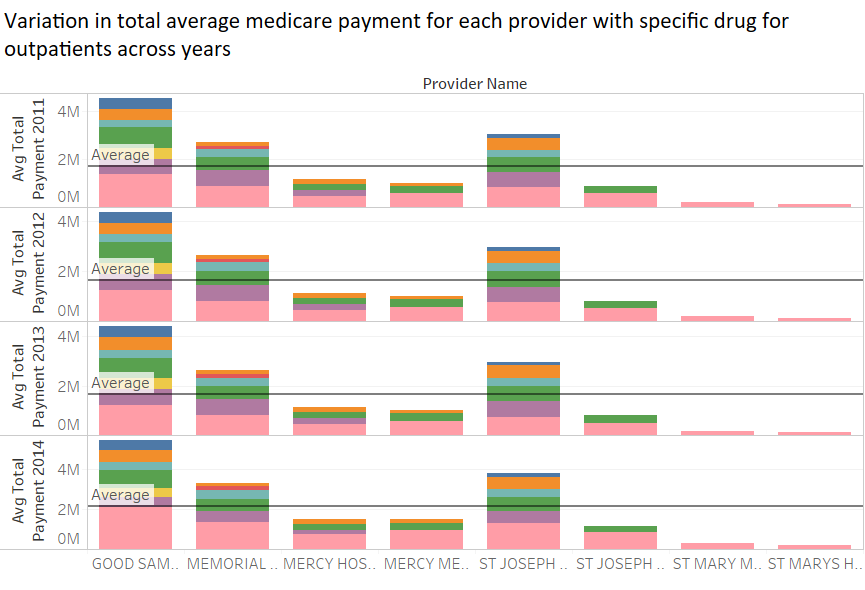
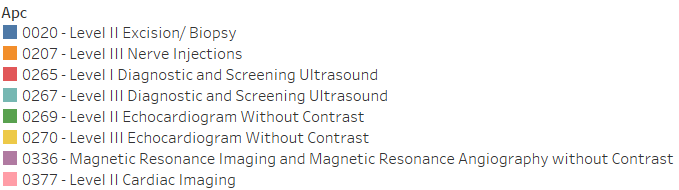


Figure 5: Tableau Chart of Variation in Total Average Medicare Payment for Each Provider with Specific Drug for Out-Patients Across Years



*\*APCs or "Ambulatory Payment Classifications" are the government's method of paying facilities for outpatient services for the Medicare program.*

1. We can see that payment cost for Cardiac Imaging is the highest for all the providers followed by Level II echocardiogram and magnetic resonance imaging. Providers Good Samaritan and St Joseph have the highest Medicare payments. There is not much variation in Medicare payments to providers for each APC across the years.
2. **Conclusion and Lessons Learned**
   1. Based on the descriptive analytics of the Medicare dataset, we can conclude the following:
      1. 2014 saw a jump in In-Patient Cost per Person in all US States, compared to the three previous years.
      2. Kentucky, Oklahoma, Mississippi, Wyoming have a relatively high Out-Patient Cost per Person, showing an ineffective system even though the cost of living is low compared to other US States.
      3. There is little variation in Total Average Medicare Payment for each drug over the years.
      4. Respiratory-related drugs cost the highest, when compared to drugs for other ailments.
      5. Coastal states, such as California, New York and Massachusetts, as well as the District of Columbia have the highest In-Patient costs per person. This is partly due to higher costs of living in those States.
      6. Most of the top Providers belong to bigger states like California, Texas and Florida, in terms of Total Medicare Amount Submitted.
      7. New York City has the highest average payments and witnesses a greater level of variation in terms of Payments for its residents.
      8. Cardiac Imaging is the highest cost-incurred in Medicare payments.
   2. Based on our project, our major take-aways include the following:
      1. Use population as a control variable when dealing with national issues
      2. When tackling national issues, like healthcare, there could be several external variables affecting the dataset
      3. Gain a solid understanding of the subject at hand before processing and querying the dataset

**References**

[1] - https://www.healthmarkets.com/resources/Medicare/Medicare-vs-medicaid-infographic/

[2] - https://www.kaggle.com/cms/cms-Medicare

[3] - https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html