Exploring Advanced Healthcare Analytics through BigQuery and Looker Studio in Google Cloud

Rachel Quatroche and Kamryn Robertson
MATH 5210: Advanced Math & Stat Computing
December 15, 2023

Abstract

Conducting analysis of large datasets can be a complex and resource-hungry task, however, with so many cloud computing tools available today, it does not have to take an inordinate amount of time or stretch personal computing devices to their limits. To demonstrate the efficacy and use of these tools, we recreated an existing exploratory data analysis and k-means cluster model using BigQuery and Looker Studio through Google Cloud on the publicly available Medicare dataset. Additionally, we investigated some of the limitations on the free-use version of Google Cloud, identified impacts on data cleaning and analysis, and noted possible solutions to overcome gaps in available visualization techniques. We found that we were able to recreate the primary analysis exclusively in the cloud computing environment, without the use Python, identifying major data trends, and easily visualizing those trends in a meaningful way.

1 Background

In 1965, the United States created the federal health insurance program Medicare for people over 65 years old, regardless of their income status, medical history, or health status. Next, in 1972, Medicare was expanded to cover those under 65 with long-term disabilities. Currently, Medicare provides resources to over 60 million people that aid in medical care service payments that

include hospitalizations, physician visits, drug prescriptions, preventice services, nursing facility and home health care, and hospice care. Significantly large datasets are produced and maintained by the Centers for Medicare & Medicaid Services (CMS), a federal agency under the Department of Health and Human Services.[1] These datasets contain a wealth of information regarding Medicare beneficiaries, nursing support facilities, payments, costs, drug treatments, duration of stay for patients admitted to care facilities, and much more. The ability to efficiently analyze these large datasets plays a central role in creation of policy, federal budgeting requests, and individual state health mandates, among other things. In this analysis, we demonstrate some of the cloud computing techniques that can be used with the Medicare dataset to conduct exploratory data analysis, visualize the data in a meaningful way, and identify useful trends.

We begin by attempting to recreate the analysis conducted in the Kaggle "Deep Healthcare Analysis using BigQuery" by Shivam Bansal.[2] The original analysis used a BigQuery to conduct exploratory analysis in five areas, and Python with a BigQuery helper library to visualize this analysis and conduct k-means clustering. Out of the 1,826 Google Cloud services available, we exclusively used two of them, BigQuery and Looker Studio, to conduct a similar analysis. BigQuery is included in the free trial of Google Cloud through the sandbox interface and works with several integrated tools, like BigQuery ML and BI Engine, to allow users to analyze datasets. The main command-line tool is through SQL query, though client libraries that support Python and Java are also available. It includes basic visualizations (bar and line charts on simple queries), the ability to export query results, and resource expenditure information. Looker Studio is available in an extremely limited format with the trial version of google cloud, and we explored it using a paid version (though did not exceed the amount of free credits we had). It provides an intuitive code-free way to produce more advanced data visualizations than are available in BigQuery, and data from BigQuery can be easily accessed through Looker Studio. The report builder function ensures all graphs are interactable, well-organized, and collaborative. We discuss limitations of Looker Studio in **Section 4**.

2 Data Sources

The CMS_Medicare public dataset, provided by the Center for Medicare and Medicaid through GoogleCloud BigQuery, contains 29 tables of data from 2011 to 2014 with a variety of Medicare metrics. The dataset was selected for analysis for three primary purposes: high credibility, ease of

use due to hosted location within the Google Cloud environment, and the existence of a completed cloud computing basic analysis that could be used as a guideline for our own analysis. We utilized five subsets: Referring Durable Medical Equipment 2014, Inpatient Charges 2011, Home Health Agencies 2013, Nursing Facilities 2014, and Part D Prescriber 2014. As a representative example due to space constraints, we explain the Nursing Facilities 2014 set here.

The Nursing Facilities 2014 dataset contains 15,026 rows of data and 41 measures. The six inherent text measures include city, facility_name, provider_id, state, street_address, and zip_code, though we note that four of these can be easily converted to location/geo-data. The six-digit provider identification codes are unique for the skilled nursing facility that has made the Medicare claims. All other measures are numeric and include metrics such as number of beneficiaries by race, the percentage of beneficiaries who have a certain medical condition (Alzheimer's, depression, heart failure, etc.), and Medicare payment/charge information for the facility. Additional measures will be explained as necessary when used for analysis, however, most measure names are self-explanatory.

The Nursing Facilities 2014 dataset is relatively clean, with consistent text formatting for U.S. state abbreviation, city, postal zip code, valid percentage values, no obvious outliers or typos, though there are null values in many columns (see Treatment). After basic exploratory queries and visual review of the first 100 rows, the other subsets appear to have the same general data types, level of quality and consistency, and are sufficient for the purpose of this analysis.

3 Treatment

We ran a set of basic exploratory queries to ascertain the status of the data for each of the 5 subsets prior to analysis. Using a combination of the SELECT DISTINCT and SELECT COUNT SQL calls, we found no duplicate rows in any of the datasets. Next, using various groupings with the IS NULL call, we determined multiple columns contained null values. Normally, we would either filter these rows, interpolate values, or further explore to ensure it was not a data import problem. However, we desired to recreate an existing Kaggle analysis as close as possible, so elected to keep the values as they were, since that is what the original analyzer chose to do. Had problems arisen during our analysis, we could have revisited and made further adjustments.

Additionally, we acknowledge there are many more tools available in BigQuery through the SQL interface to clean and manipulate data. Options such as NORMALIZE, IFNULL, and

REGEXP are helpful for removing null data and fixing typos. We did not exercise as much control over this dataset during treatment as we might have liked for three reasons: first, it is a public dataset that we do not have full permissions to edit, secondly because we were operating on a free tier of Google Cloud and did not want to waste valuable queries, and lastly because of desire to stay true to the original analysis. Future analysis of these datasets may consider taking additional cleaning and treatment measures prior to analysis completion to validate our results.

4 Exploratory Data Analysis

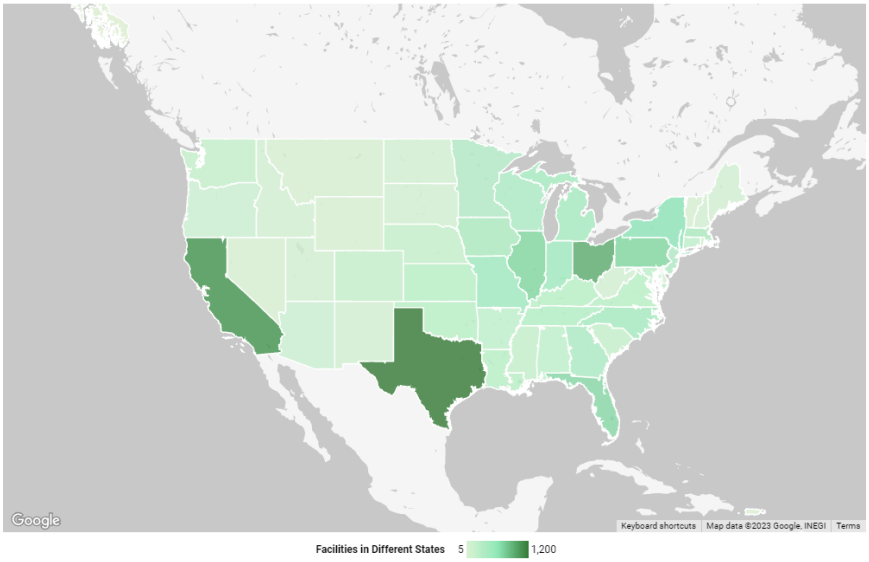
For this portion of our project, we analyzed the outcomes of a series of queries within five main focus areas in BigQuery to better understand the data. This was done utilizing the various samples of the Medicare dataset already partitioned in the cloud. Table outputs are truncated for space purposes. We then followed up using Looker Studio to recreate the original Python coded visualizations, or explored limitations of Google Cloud's free visualization tool when applicable.

4.1 Nursing Facilities By State

The number of nursing facilities in an area could possibly be an indicator of the quality of care. Utilizing the Nursing Facilities 2014 data subset, we wanted to determine which states had the highest and lowest number of nursing facilities. Using **Query 1** in Appendix 1 within BigQuery, we obtain the following truncated output:

	state	total_facilities
1	TX	1200
9	MI	424

This concurs with the original Kaggle analysis. In 2014, Texas had the highest number of nursing facilities with 1200, while Michigan had the lowest number of nursing facilities with 424. Next, we produce a United States map showing the number of Nursing Facilities in each state with Looker Studio, after converting the state column to a Country subdivision 1st level type. We do note some differences to the Kaggle graph, however it does not change the interpretation. Texas, California, and Ohio clearly have more facilities, while upper midwest states have the least number of facilities, which makes sense due to population density.



One note is that our graph does not allow for Alaska/Hawaii to be displayed "in-picture", is less customizable (legend position and title), and is a filled map instead of a choropleth. This is one limitation of the native free-use visualization tool; it lacks the variety and flexibility that Python can produce. However, it did speedily render and was intuitive to create. Additionally, a choropleth appears to be available through a partner visualization, but is not easy to use, requires additional API access and permissions, and may not be included in the free version.

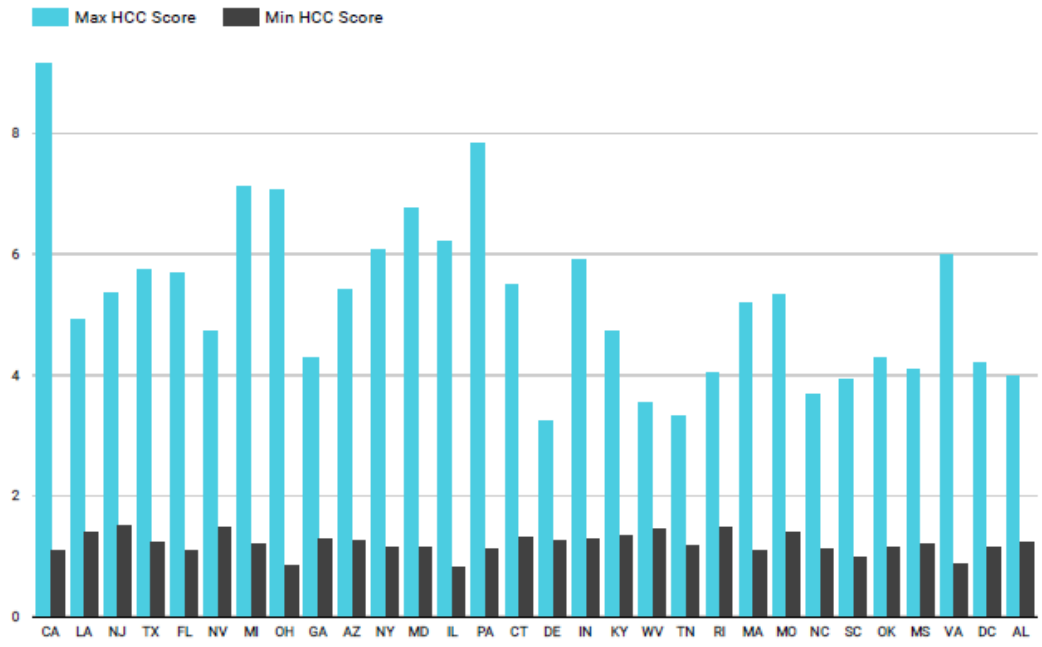
4.2 HCC Scores

A Hierarchical Condition Category (HCC) model is used to identify cost of treatment trends, assess risk for chronic diseases, and can be a predictor for beneficiary health. Using the Nursing Facilities 2014 subset, we attempted to determine which state and nursing facilities have high HCC scores. With **Query 2** in Appendix 1 within BigQuery, we obtain the following output:

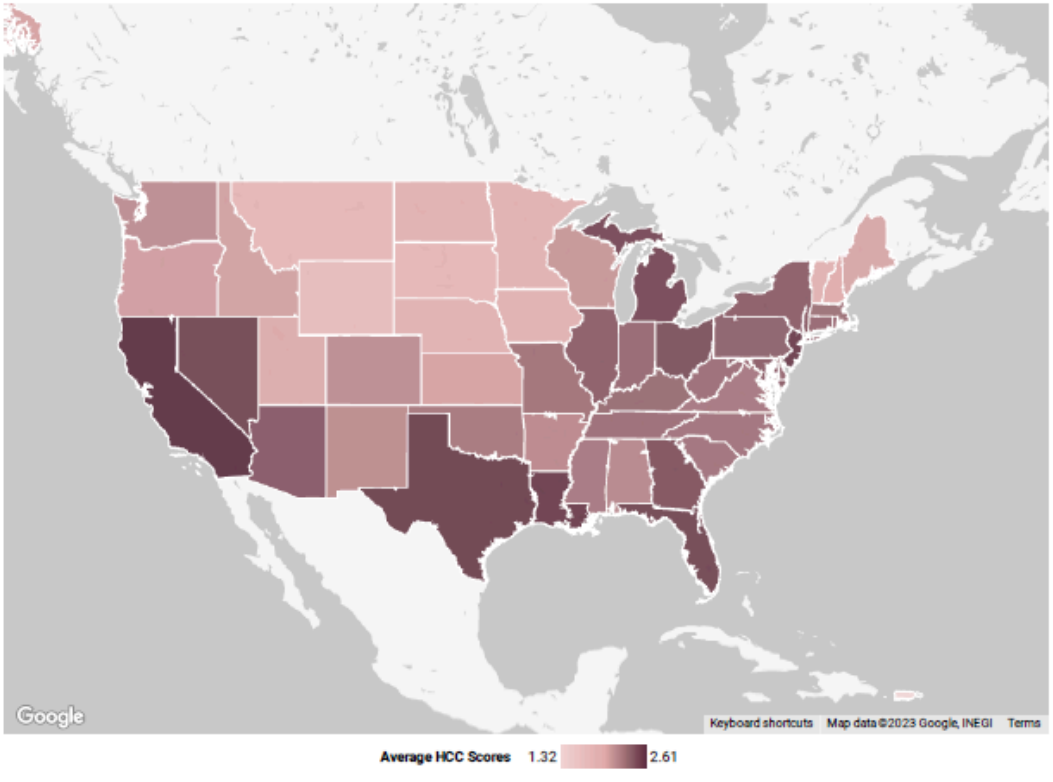
	state	avg_hcc_score	max_hcc_Score	min_hcc_score	total_facilities
1	CA	2.612878	9.15	1.11	1084
2	LA	2.557959	4.91	1.39	294

In 2014, California had the highest average HCC score of 2.61 and maximum HCC score of 9.15 spread across 1084 nursing facilities. This concurs with the Kaggle output, and correlates to California having the highest risk of chronic diseases and will likely mean higher healthcare

costs for Medicare recipients in the future. A Looker Studio visualization confirms and shows the minimum and maximum HCC scores of different states, truncated to 30 states:



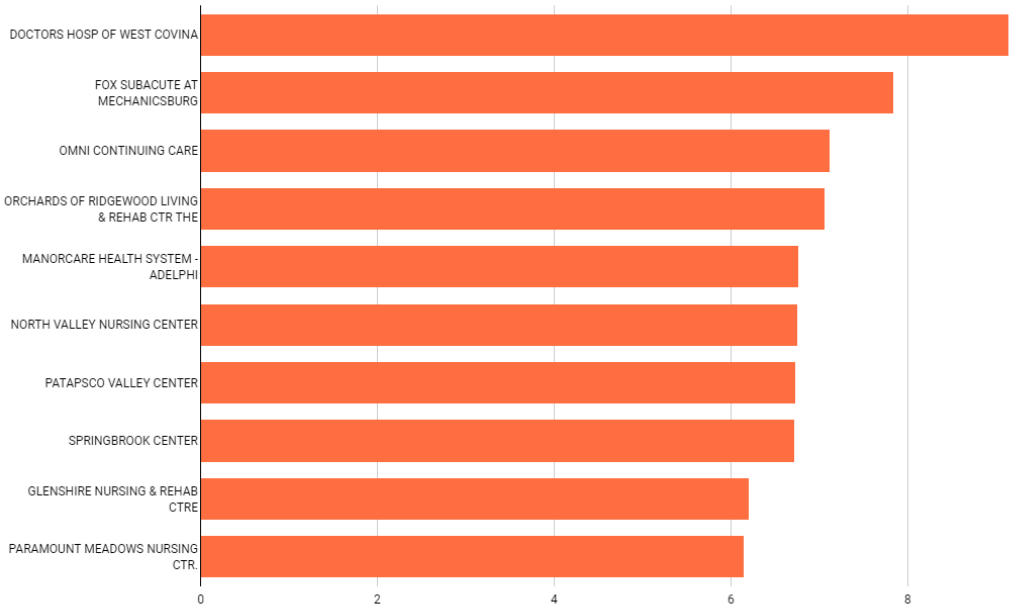
Similarly, we wished to visualize the average HCC scores for each state on a United States map, and created a Looker Studio filled map that clearly displays a gradient:



It is obvious with this style graph that California, Texas, Louisiana, Florida and New Jersey have the highest average HCC scores (darkest coloring), while the upper midwest states have the lowest (lightest coloring). Next, we identified the nursing facilities with high HCC scores using **Query 3** in Appendix 1 through BigQuery, with the following truncated results:

	facility_name	city	state	average_hcc_score
1	DOCTORS HOSP OF WEST COVINA	WEST COVINA	CA	9.15
2	FOX SUBACUTE AT MECHANICSBURG	MECHANICSBURG	PA	7.84

We see that the Doctors Hospital of West Covina has the highest HCC score, which we correlate with poor health. We can confirm these results with Looker Studio, where we used the drill down method with state, city, and facility name to produce in graph form:



We also retrieved the nursing facility with the lowest HCC score, correlating to the population being relatively healthy. Using **Query 4** in Appendix 1, we obtain the following:

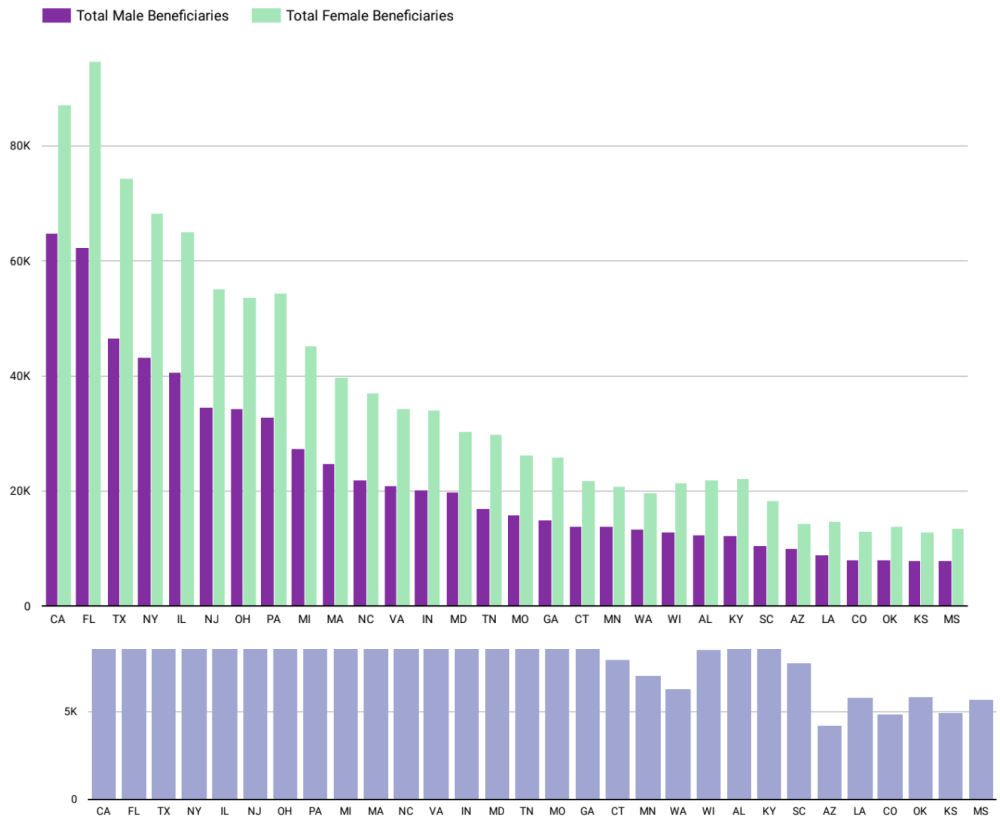
	facility_name	city	state	average_hcc_score
15024	POWDER RIVER MANOR	BROADUS	MT	0.79

15025	KFH - MALAMA 'OHANA NURSING AND REHAB CENTER	HONOLULU	HI	0.75

We gather that in the year 2014, KFH- Malama ‘Ohana Nursing and Rehab Center in Honolulu, HI had the most healthy patients on Medicare. If additional analysis were conducted, we might explore geographic similarities at the postal code granular level for HCC scores to identify the healthiest and unhealthiest regions, with HCC scores as a primary measure of health.

4.3 Beneficiary Gender

Still using the Nursing Facilities 2014 data subset, we next wanted to find the state(s) breakdown of male beneficiaries and female beneficiaries and determine which was greater. Verifying our output with **Query 5** from Appendix 1 in BigQuery and using Looker Studio for visualization, we obtain the below graph:



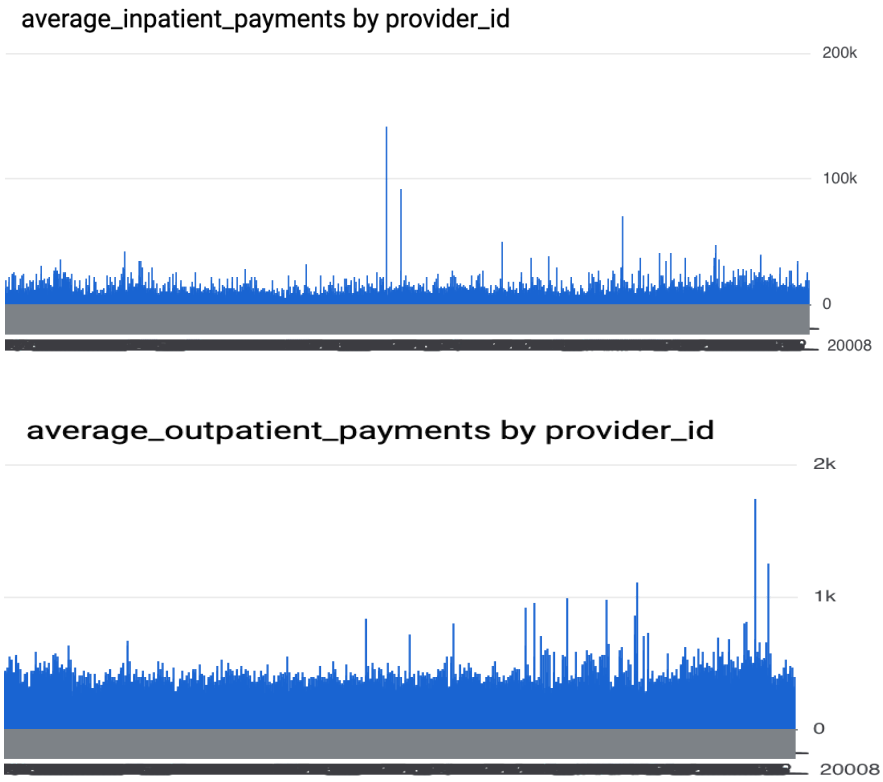
For every state, we find that there are more female beneficiaries than male beneficiaries. This correlates to the possibility that there exists a fixed quota for the number of beneficiaries in

total. However, Florida contained the highest number of female beneficiaries while California had the highest number of male beneficiaries. For supplementary purposes, we used Looker Studio to produce a graph showing the difference in total male and female beneficiaries in each state. Since this data is not an existing column in the dataset, we used Looker Studio's custom field function, which allows users to perform basic arithmetic functions, such as: max, min, count, sum, multiplication, and difference. This returned the following graph:

We recognize that Florida had the highest difference in the number of female vs male beneficiaries, which we saw in the first graph was approximately 30,000 more female beneficiaries. We also note that this could have been an easy way to identify if any state had more male than female beneficiaries, as it would have resulted in negative values.

4.3 Inpatient and Outpatient

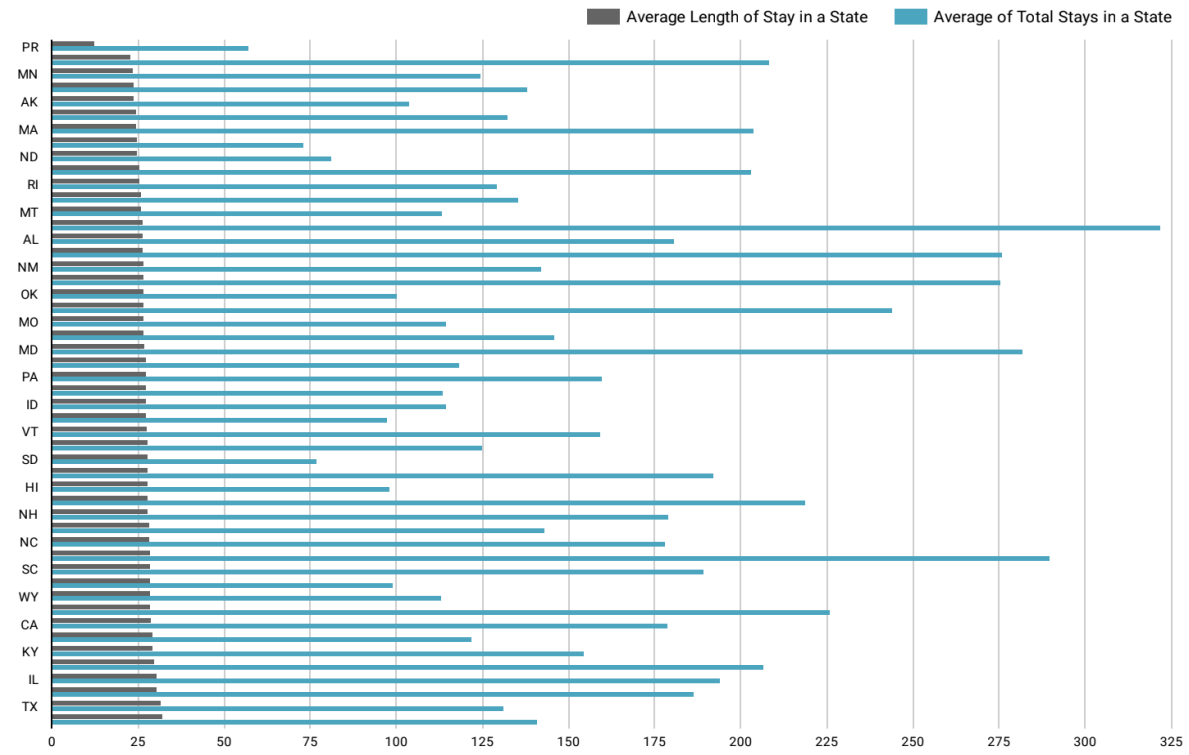
Healthcare providers are a vital source for determining the average treatment costs in relation to Medicare beneficiaries. Hospitals, physicians, and many other healthcare facilities that deliver medical services to Medicare beneficiaries can influence inpatient and outpatient costs by their billing practices, their quality of care, and even in negotiation with Medicare to increase reimbursement rates. Utilizing the inpatient and outpatient charges data subsets, we wanted to find the difference of average inpatient and outpatient payments. Using **Query 6** in Appendix 1 within BigQuery, we obtain the following basic bar chart visualizations:



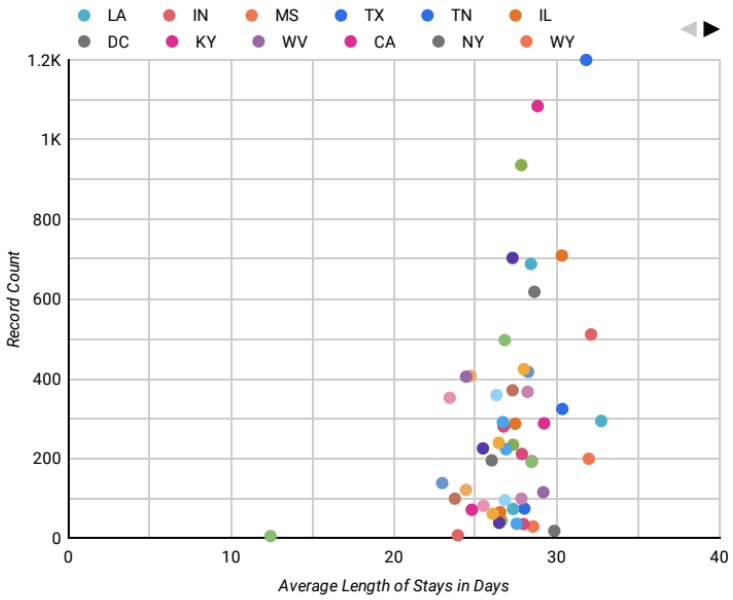
From our graphs we can see that the average inpatient payments were higher than outpatient payments (note the different y-axis scales). This can be an indicator that inpatient care costs more due to the level of care needed for patients, overhead costs to maintain facilities, and even the resources needed for the extensive care of Medicare beneficiaries. As an aside, we encountered our first two real limitations of Looker Studio when attempting to directly visualize a histogram of the difference in inpatient and outpatient payment data; first, Looker Studio does not have a native histogram available. A partner, Supermetrics Vis, purports to provide one, however, it requires a full paid version to access, and we were unable to get it to work anyway. Secondly, this information exists in two separate datasets; in order to get the raw data we need, both sets must be imported, then inner joined to merge, then the difference calculation must be performed. While not impossible in Looker Studio, the more complex analysis that must be completed to get the desired graph, the less likely you will be able to easily produce it in Looker.

4.4 Stay Days by State

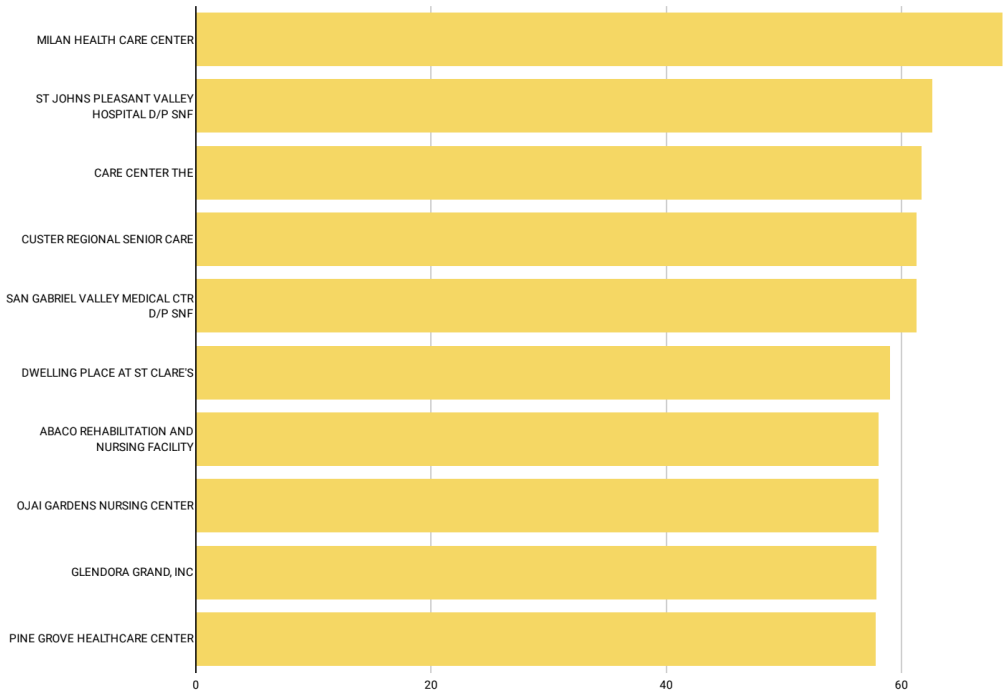
Analyzing the average length of stay in care facilities and the average total stays by state for Medicare beneficiaries can provide insights into healthcare utilization patterns, efficiency of care delivery, and potential cost implications. Utilizing the 2014 Nursing Facilities data subset, we found average total number stays of different facilities in comparison to average length of stays for each state. Using **Query 7** in Appendix 1 through BigQuery, and Looker Studio to verify, we show our results through the graph:



We note that Puerto Rico has the lowest average length of stay and that the rest of the results were consistent with the original Kaggle analysis, though we have truncated the view to only 30 states. However, we were unable to reproduce the double scatter plot from Kaggle, as LookerStudio does not provide significant customization to make that visualization. The double bar chart does visually display the same information, but is not preferred due to clutter. Another option would be a single scatter plot for each, however, we encountered an additional limitation: Looker Studio will not allow geographic or text/categorical data to be on the axis of a scatter plot. Therefore, we were forced to try to color code, but there are too many states for this to be useful. No additional information can be successfully gleaned from such a poor visualization, so we would seek alternative solutions to visualize this data outside of Looker Studio if necessary.



Next, we break down the results from these previous queries to determine which facilities had the highest average length of stay days. Using **Query 8** in Appendix 1 in BigQuery and LookerStudio to supplement, we show our results through the following graph:

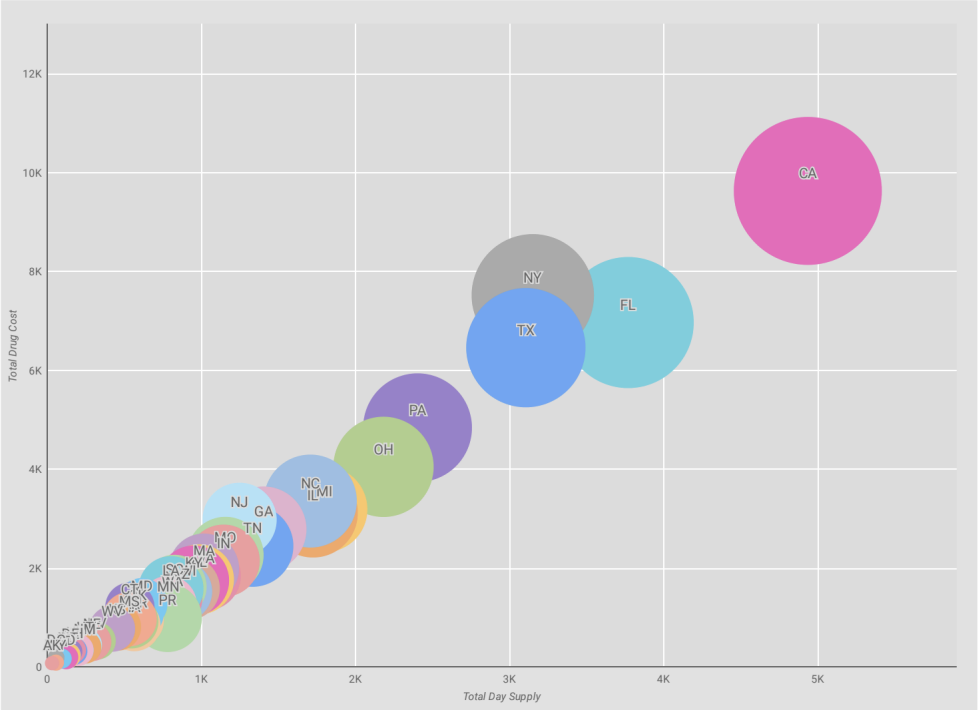


We conclude that Milan Health Care Center in Milan, TN had the highest average length of stats with 68.6. Analyzing the length of stay by facility can be beneficial in resource allocation and planning and again improve the quality of care for Medicare beneficiaries.

4.5 Total Claims, Day Supply, and Drug Costs

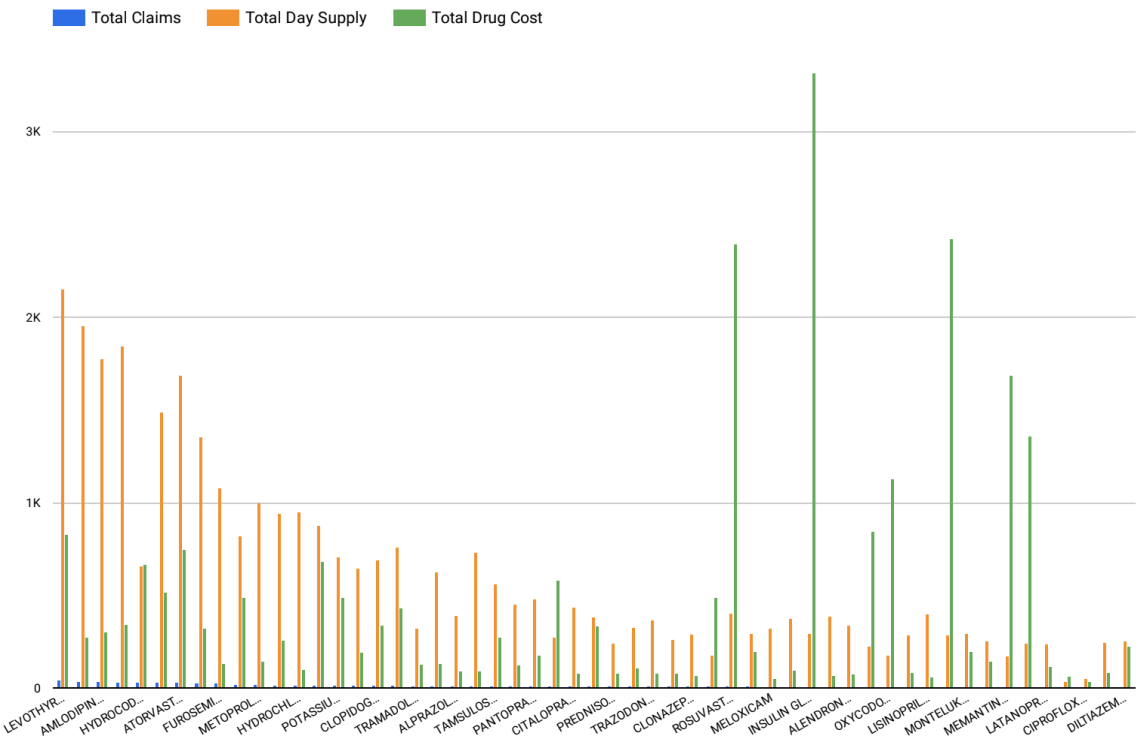
Referring back to the previous analysis, resource allocation and planning allows healthcare providers the ability to more effectively distribute medications to meet consumer demands. Through medication management, strategizing, and negotiating with pharmaceutical companies, Medicare maintains not only the quality of care of its beneficiaries but it also keeps cost control. Utilizing the 2014 Part D Prescriber table, we want to determine what state had the highest claims, day supply counts, and drug costs. Using **Query 9** in Appendix 1 in BigQuery and Looker Studio to supplement, with bubble size as claim count, we obtain the following output:

	state	total_claim_count_millions	total_day_supply_millions	total_drug_cost_millions
1	CA	116.0	4935.0	9634.0
2	FL	91.0	3770.0	6970.0



In 2014, we found that California had the highest number of claims with 116,000,000, highest day supply counts with 4,935,000,000, and highest drug costs with 9,634,000,000. This could be a direct influence of California's large population, diverse demographics, and California having the highest average HCC score. A gradient color scale is unavailable in Looker Studio.

Furthermore, using the 2014 prescriber data subset, we determine what drug had the highest frequency in claims, what drug cost the most to supply, and what drug was prescribed the most. Using **Query 10** in Appendix 1 within BigQuery and Looker Studio again to supplement, we show our results through the following graph:



From this graph, we conclude that Insulin had the highest total cost and Levothyroxine Sodium had the highest number of claims, as well as the highest per day supply. We also validate that Almodipine, Besylate, and Simvastatin had a higher total supply than their drug cost.

Expanding on the previous query, we break down the highest drug frequency in claims by each state. Using **Query 12** in Appendix 1 within BigQuery, we obtain the following output:

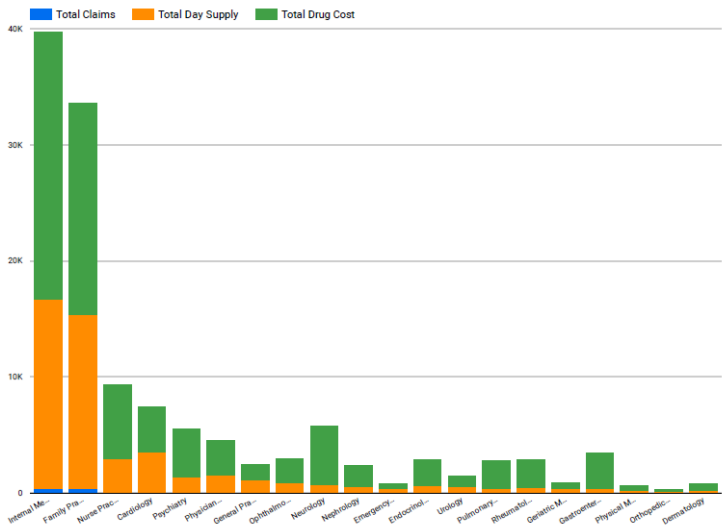
	state	drug_name	MaxClaimCount
1	AA	AMOXICILLIN	225
3	AL	HYDROCODONE/ACETAMINOPHEN	1214487
8	CA	LEVOTHYROXINE SODIUM	3845087

From the table we can see that in California, Levothyroxine Sodium was prescribed the most in all the claims of 2014. We also notice that from the results that a common pain reliever, Hydrocodone/Acetaminophen, is one of the most frequently prescribed drugs to Medicare recipients in general. We desired to produce a pie chart in Looker Studio to show how many states had a specific drug as the top prescribed drug, but this is a complex query that requires three merges, three re-aggregation calculations and a final tabulation. In this case, paid products such as the Vertex AI Workbench or Jupyter notebook tie-in would likely be our best solution, or we could use the same technique as the original analyzer (a bigquery library within a Python shell and visualization through Seaborn in Python).

Next, we were interested in looking at nursing specialties associated with highest claims. Using **Query 11** in Appendix 1 within BigQuery, we obtain the following output:

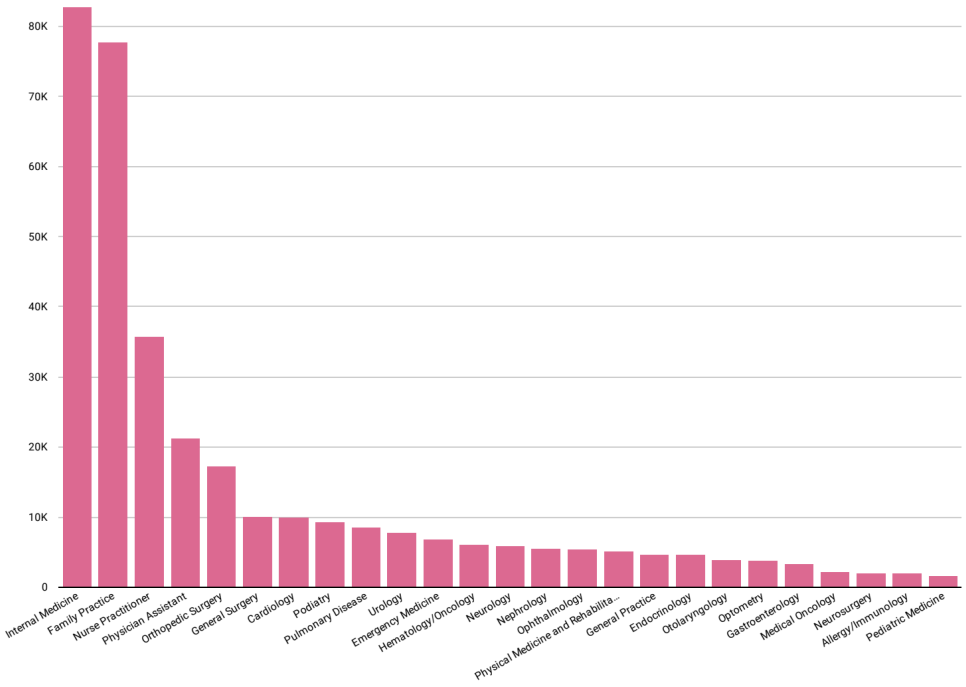
	specialty_description	total_claim_count_millions	total_day_supply_millions	total_drug_cost_millions
1	Internal medicine	386.0	16340.0	22983.0
2	Family practice	361.0	14967.0	18254.0

We see that internal medicine and family practice have the highest number of claims, which intuitively makes sense, as those are very common routine specialties for most illness and checkups. The original analysis included a word cloud to display this information graphically, which is a function that the old "Looker" supported, but is no longer a native tool. However, we were able to recreate a stacked column chart as a visualization:



The x-axis is sorted by highest to lowest claim count (left to right), and it is obvious to see that internal medicine and family practice far exceeds in every dimension when compared to other nursing specialties. They likely have the most visits, and are therefore prescribing more drugs, at a greater total cost, with the greatest number of claims, than any other specialty.

Finally, we attempt to answer the last exploratory question: what were the most common provider types in the United States during 2014? **Using Query 13** in Appendix 1 through BigQuery and Looker Studio to verify and visualize, we obtain the following:



Just as we intuitively suspected, there are more internal medicine and family practice facilities than any other provider type within the United States for this period, and there is a very large gap in the number of other facilities. This information could be used to direct funding to the maximum number of practices, identify care/access to care shortfalls, or inform policy makers in decisions regarding regulating these facilities.

5 K-Means Model

Finally, we conducted an advanced analysis of the 2014 Nursing Facilities dataset using k-means clustering. This model is available in BigQuery through the BigQuery ML, which allows paid users to use SQL queries to construct, execute, and modify machine learning models. We did

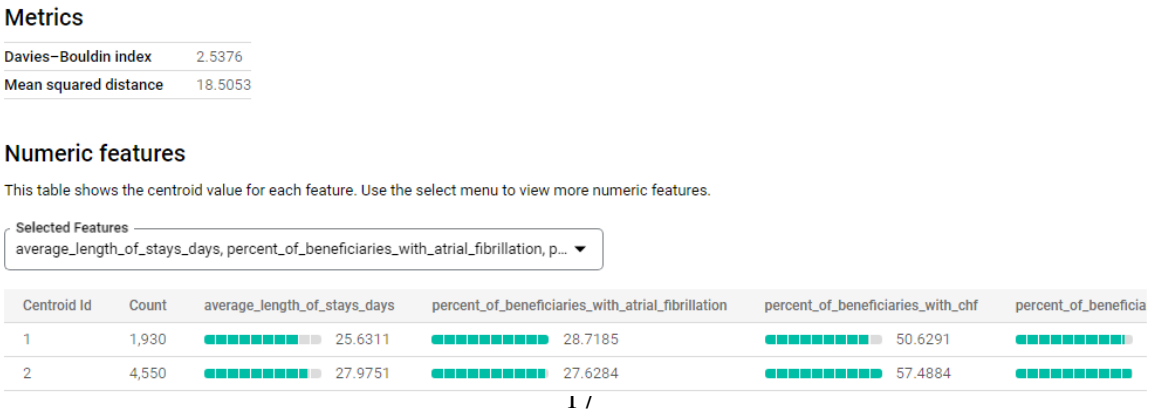
have to activate a paid subscription to complete this portion, however, total execution costs were under the free trial dollars provided. To better compare results, we attempted to recreate the model that the original analyzer used: four cluster, random start state, with 29 features, but without using sklearn's KMeans function, and instead using an existing tutorial to do it directly in BigQuery. [3] Additional minor data cleaning and processing was required prior to model creation, but was conducted directly within the SQL call. We did not omit NULL values, as BQML is designed to automatically handle null values and we desired to verify this feature.

5.1 Preprocessing and Model Run Statistics

The model preprocessing query can be seen in **Query 1** of Appendix 2. It eliminates all non-textual columns except the facility name in preparation for the model run. The model creation and run query can be seen in **Query 2** of Appendix 2. The 29 features include things like distinct number of beneficiaries per facility, percent of beneficiaries with certain diseases (like asthma, cancer, copd, depression, stroke, and schizophrenia), gender and race information, stay information, and payment totals to the facilities. The model run time was 44 seconds, which was incredibly quick compared to our experience running k-means in the Python environment. The breakdown was: 2 seconds validation, 6 seconds pre-processing, 32 seconds training, and 4 seconds evaluating. The model was completed in 4 training iterations.

5.2 Cluster Information Post-Run

Model information is available in BigQuery underneath the "Model Info" tab following the model run. The distance type confirms Euclidean, which is what we wanted, and random initialization. Four (4) clusters were identified with the following sizes for each specific cluster: Cluster 1 - 1930, Cluster 2 - 4550, Cluster 3 - 4009, Cluster 4 - 4537. The validation post-model confirms all 29 features were used, and even suggests the 10 most influential features, however, we desired to conduct our own interpretation and visualization. A sample of the information provided by BigQuery post-model run is seen below:



Metrics

Davies–Bouldin index	2.5376
Mean squared distance	18.5053

Numeric features

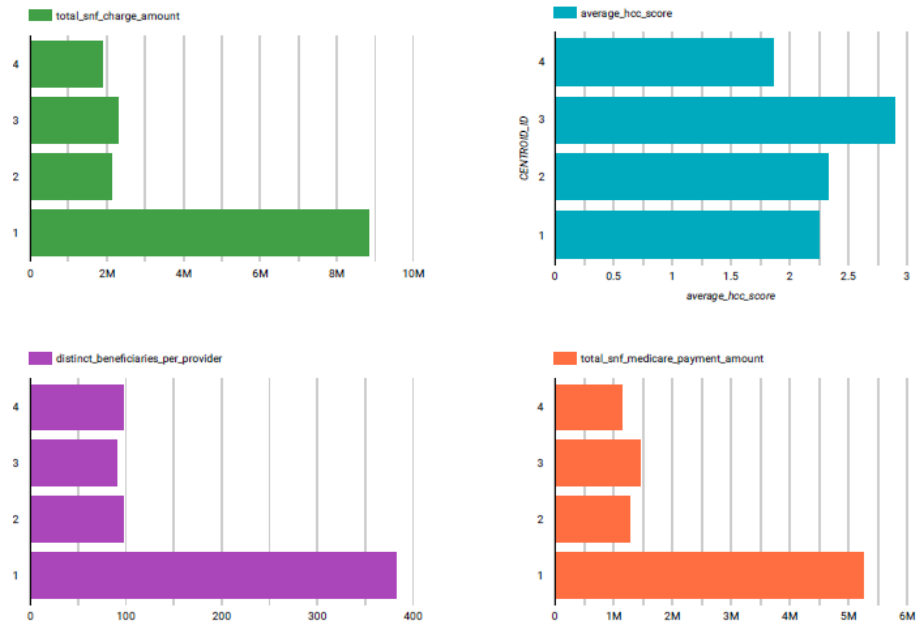
This table shows the centroid value for each feature. Use the select menu to view more numeric features.

Selected Features: average_length_of_stays_days, percent_of_beneficiaries_with_atrial_fibrillation, p...

Centroid Id	Count	average_length_of_stays_days	percent_of_beneficiaries_with_atrial_fibrillation	percent_of_beneficiaries_with_chf	percent_of_beneficia
1	1,930	25.6311	28.7185	50.6291	
2	4,550	27.9751	27.6284	57.4884	

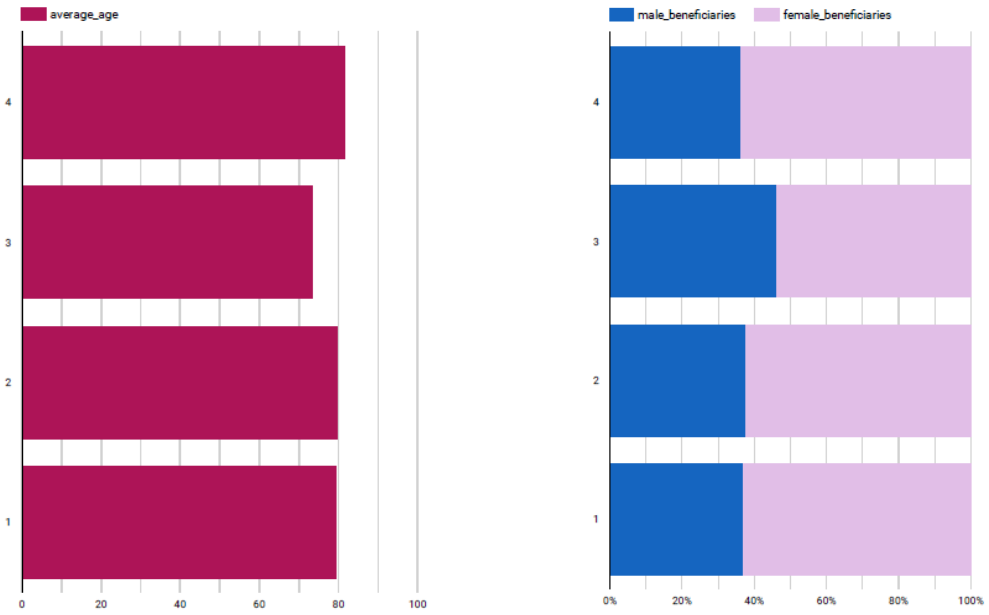
5.3 Cluster Visualization and Interpretation

We exported the data post-model creation to Looker Studio and first attempted to recreate visualizations from the original analysis that display key features for each cluster, as seen below:

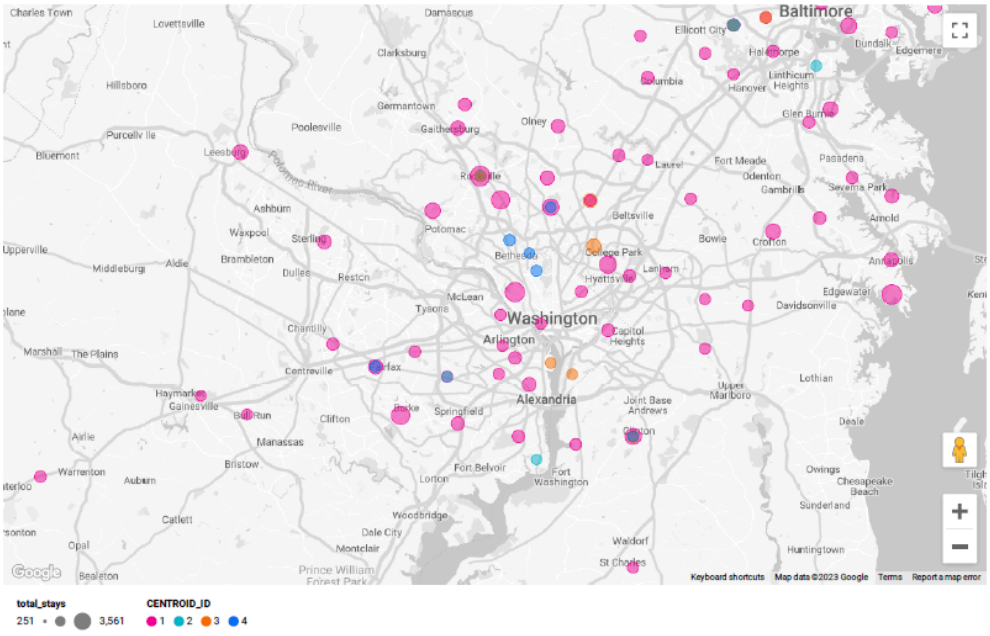


We are aware that, due to the nature of the k-means algorithm, we cannot directly replicate the original analyzer's clusters, however, we did obtain similar results. Specifically, we note that cluster 1 facilities have the highest charge amount (top-left), average HCC scores are close with cluster 4 having the lowest and cluster 3 having the highest (top-right), cluster 1 nursing facilities have the highest number of distinct beneficiaries (bottom-left), and cluster 1 facilities also have the highest payment amounts (bottom-right). The original analysis included histogram and distribution data regarding the races of beneficiaries in each cluster. Since we cannot natively use histograms within Looker Studio, we elected to explore 3 additional metrics instead: age, gender, and geographic location of the clusters.

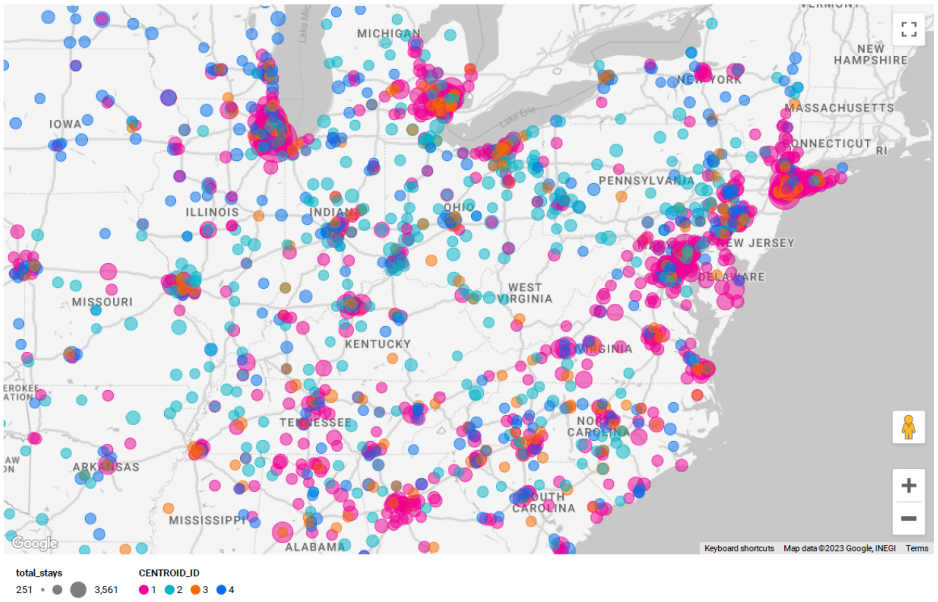
First, we explored the age and gender of each cluster, seen below. These graphs show that the gender and age breakdowns are relatively close, though cluster 3 has more men than the other clusters. Interestingly, cluster 3 also has the lowest average age. We know that men have a slightly lower life-expectancy than women, so it makes sense that the average age of beneficiaries might be related to the gender disparity between the clusters. Additional cluster analysis would be necessary before we confirmed that supposition.



Next, we used Looker Studio to look at local area providers, within the Metropolitan D.C. area. To accomplish this, we used a bubble map, with the color as the cluster id number and the size of the bubble as the total number of stays at that nursing facility, as seen below:



These results were surprising to us, because we did not expect so many of one cluster in one area. It looks like most of the facilities in the DC area are part of cluster 1, with some minor outliers. Additionally, we can see some facilities that appear to be co-located, but in different clusters, such as in Fairfax, VA. We then expanded the graphical area to the East Coast:



We note high concentrations of cluster 1 facilities in metropolitan areas, including Detroit, Chicago, Atlanta, and New York City. It is likely that since metropolitan areas have more people, there are just more facilities there, giving a false appearance that cluster 1 nursing facilities are likely to be in metro areas. We do note that it seems likely that cluster 1 components are larger facilities, taking into account all of the visual evidence (beneficiaries per provider, payment amount totals, and the geographic placement of these facilities on the map). We do not identify distinguishing features for the other clusters in this initial analysis of the k-means model. Additional analysis would be necessary to further understand each of the cluster's defining characteristics, but it does not seem to be that geography is one of them.

6 Conclusion

Using Google Cloud's BigQuery and Looker Studio, we were able to conduct sufficient exploratory data analysis on the large Medicare datasets to answer interesting questions about the data and identify trends. We were further able to use the machine learning tool to k-means cluster nursing facility data into four groupings based on 29 features and explore them in a meaningful way. The analysis was much faster and easier with these cloud computing resources and the efficiency in querying and visualizing demonstrates the effectiveness of using cloud computing to analyze big data for policy and decision makers. Though each of these tools has its limitations, they are scalable, intuitive, and invaluable to better analyze big data.

A Appendices

Appendix 1 contains the text of SQL queries used within BigQuery for exploratory data analysis.

Appendix 2 contains the text of SQL queries used within BigQuery for k-means clustering.

References

[1] Published: Feb 13, 2019. “An Overview of Medicare.” KFF, 13 June 2023, www.kff.org/medicare/issue-brief/an-overview-of-medicare/.

[2] Bansal, Shivam. “Deep Healthcare Analysis Using BigQuery.” Kaggle, Kaggle, 6 Oct. 2018, www.kaggle.com/code/shivamb/deep-healthcare-analysis-using-bigquery/notebook.

[3] Google. (n.d.). *Create a K-means model to cluster London bicycle hires dataset | Bigquery | google cloud*. Google. https://cloud.google.com/bigquery/docs/kmeans-tutorial

Appendix 1

Query 1. Which State had the highest number of nursing facilities? Which State had the lowest number of nursing facilities?

```
SELECT
 state,
 COUNT(DISTINCT provider_id) AS num_facilities
FROM
 `healthcare-analytics-407516.Medicare.nursing_facilities_2014`
GROUP BY
 state
ORDER BY
 num_facilities DESC
LIMIT 2;
```

Query 2. Which state and nursing facilities had poor HCC risk adjustment scores?

```
SELECT
 state,
 AVG(average_hcc_score) as avg_hcc_score,
 MAX(average_hcc_score) as max_hcc_score,
 MIN(average_hcc_score) as min_hcc_score,
 COUNT(average_hcc_score) as total_facilities
FROM
 `healthcare-analytics-407516.Medicare.nursing_facilities_2014`
GROUP BY
 state
ORDER BY
 avg_hcc_score DESC;
```

Query 3. Which nursing facilities had high HCC risk adjustment scores?

```
SELECT
 facility_name,
```

```
 city,
 state,
 average_hcc_score
FROM
 `healthcare-analytics-407516.Medicare.nursing_facilities_2014`
ORDER BY
 average_hcc_score DESC;
```

Query 4. Which nursing facilities had low HCC risk adjustment scores?

```
SELECT
 facility_name,
 city,
 state,
 average_hcc_score
FROM
 `healthcare-analytics-407516.Medicare.nursing_facilities_2014`
ORDER BY
 average_hcc_score DESC
LIMIT 5;
```

Query 5. What state(s) sum of male beneficiaries was greater than the sum of female beneficiaries?

```
SELECT
 state,
 SUM(male_beneficiaries) AS male_ben,
 SUM(female_beneficiaries) AS female_ben
FROM
 `healthcare-analytics-407516.Medicare.nursing_facilities_2014`
GROUP BY
 state
ORDER BY
 male_ben DESC;
```

Query 6. What is the difference in average inpatient payment and average outpatient payments?

```
-- Query for outpatient charges
SELECT
 provider_id,
 AVG(average_total_payments) AS average_outpatient_payments
FROM
`healthcare-analytics-407516.Medicare.outpatient_charges_2014`
GROUP BY
 provider_id;

-- Query for inpatient charges
SELECT
provider_id,
AVG(average_total_payments) AS average_inpatient_payments
FROM
`healthcare-analytics-407516.Medicare.inpatient_charges_2014`
GROUP BY
 Provider_id;
```

Query 7. What were the states average total number stays of different facilities and average length of stays in different facilities?

```
SELECT
 state,
 AVG(average_length_of_stays_days) AS average_length_of_stays_days,
 AVG(total_stays) AS total_stays
FROM
 `healthcare-analytics-407516.Medicare.nursing_facilities_2014`
GROUP BY
 state
ORDER BY
 average_length_of_stays_days DESC;
```

Query 8. What facilities had the highest average length of stay days?

```sql
SELECT
 facility_name, city, state, average_length_of_stays_days
FROM
 `healthcare-analytics-407516.Medicare.nursing_facilities_2014`
ORDER BY
 `average_length_of_stays_days` DESC;
```

Query 9. What state has the highest claims, day supply, and drug costs?

```sql
SELECT
nppes_provider_state AS state,
ROUND(SUM(total_claim_count) / 1e6) AS total_claim_count_millions,
ROUND(SUM(total_day_supply) / 1e6) AS total_day_supply_millions,
ROUND(SUM(total_drug_cost) / 1e6) AS total_drug_cost_millions
FROM
 `healthcare-analytics-407516.Medicare.part_d_prescriber_2014`
GROUP BY
state
ORDER BY
total_claim_count_millions DESC;
```

Query 10. Which drugs were most prescribed in the United States in 2014?

```sql
SELECT
 generic_name AS drug_name,
ROUND(SUM(total_claim_count) / 1e6) AS total_claim_count_millions,
ROUND(SUM(total_day_supply) / 1e6) AS total_day_supply_millions,
ROUND(SUM(total_drug_cost) / 1e6) AS total_drug_cost_millions
FROM
 `healthcare-analytics-407516.Medicare.part_d_prescriber_2014`
GROUP BY
drug_name
ORDER BY
total_claim_count_millions DESC;
```

Query 11. What nursing specialities were associated with the highest claims?

```sql
SELECT
 specialty_description AS specialty_description,
 ROUND(SUM(total_claim_count) / 1e6) AS total_claim_count_millions,
 ROUND(SUM(total_day_supply) / 1e6) AS total_day_supply_millions,
 ROUND(SUM(total_drug_cost) / 1e6) AS total_drug_cost_millions
FROM
 `healthcare-analytics-407516.Medicare.part_d_prescriber_2014`
GROUP BY
 specialty_description
ORDER BY
 total_claim_count_millions DESC;
```

Query 12. Which drugs were most prescribed in each state in 2014?

```sql
WITH StateMaxClaims AS (
 SELECT
   nppes_provider_state AS state,
   generic_name AS drug_name,
   SUM(total_claim_count) AS total_claim_count_millions
 FROM
   `healthcare-analytics-407516.Medicare.part_d_prescriber_2014`
 GROUP BY state, drug_name
 ORDER BY
   total_claim_count_millions DESC
)

SELECT
 A.state,
 B.drug_name,
 A.MaxClaimCount
FROM (
 SELECT
   state,
   MAX(total_claim_count_millions) AS MaxClaimCount
 FROM StateMaxClaims
 GROUP BY state
) A
```

```
INNER JOIN StateMaxClaims B ON A.MaxClaimCount = B.total_claim_count_millions;
```

Query 13. What were the most common provider types in the United States?

```
SELECT
 provider_type,
 COUNT(provider_type) AS count
FROM
 `healthcare-analytics-407516.Medicare.referring_durable_medical_equip_2014`
GROUP BY
 provider_type
ORDER BY
 count DESC;
```

Appendix 2

Query 1: KMeans Preprocessing

```
SELECT provider_id, facility_name, city, state, total_stays,
distinct_beneficiaries_per_provider,
       average_length_of_stays_days, total_snf_charge_amount,
       total_snf_medicare_allowed_amount,
       total_snf_medicare_payment_amount,
       total_snf_medicare_standard_payment_amount, average_age,
       male_beneficiaries, female_beneficiaries, nondual_beneficiaries,
       dual_beneficiaries, white_beneficiaries, black_beneficiaries,
average_hcc_score,
       percent_of_beneficiaries_with_atrial_fibrillation,
       percent_of_beneficiaries_with_asthma,
       percent_of_beneficiaries_with_cancer,
       percent_of_beneficiaries_with_chf,
       percent_of_beneficiaries_with_chronic_kidney_disease,
       percent_of_beneficiaries_with_copd,
       percent_of_beneficiaries_with_depression,
       percent_of_beneficiaries_with_diabetes,
       percent_of_beneficiaries_with_hyperlipidemia,
       percent_of_beneficiaries_with_ihd,
       percent_of_beneficiaries_with_osteoporosis,
       percent_of_beneficiaries_with_ra_oa,
       percent_of_beneficiaries_with_schizophrenia,
       percent_of_beneficiaries_with_stroke FROM
  `bigquery-public-data.cms_medicare.nursing_facilities_2014`
```

Query 2: KMeans Model Run

```
CREATE OR REPLACE MODEL `kmeans_medicare.clusters`
OPTIONS(model_type='kmeans', num_clusters=4) AS SELECT total_stays,
distinct_beneficiaries_per_provider,
       average_length_of_stays_days, total_snf_charge_amount,
       total_snf_medicare_allowed_amount,
       total_snf_medicare_payment_amount,
       total_snf_medicare_standard_payment_amount, average_age,
       male_beneficiaries, female_beneficiaries, nondual_beneficiaries,
```

```
       dual_beneficiaries, white_beneficiaries, black_beneficiaries,
average_hcc_score,
       percent_of_beneficiaries_with_atrial_fibrillation,
       percent_of_beneficiaries_with_asthma,
       percent_of_beneficiaries_with_cancer,
       percent_of_beneficiaries_with_chf,
       percent_of_beneficiaries_with_chronic_kidney_disease,
       percent_of_beneficiaries_with_copd,
       percent_of_beneficiaries_with_depression,
       percent_of_beneficiaries_with_diabetes,
       percent_of_beneficiaries_with_hyperlipidemia,
       percent_of_beneficiaries_with_ihd,
       percent_of_beneficiaries_with_osteoporosis,
       percent_of_beneficiaries_with_ra_oa,
       percent_of_beneficiaries_with_schizophrenia,
       percent_of_beneficiaries_with_stroke FROM
  `bigquery-public-data.cms_medicare.nursing_facilities_2014`
```

Query 3: Cluster Information

```
SELECT * EXCEPT(nearest_centroids_distance) FROM ML.PREDICT(MODEL
`m5210finalproject.kmeans_medicare.clusters`, (SELECT * FROM
`bigquery-public-data.cms_medicare.nursing_facilities_2014`))
```