Venkata M M.

I. Research Question

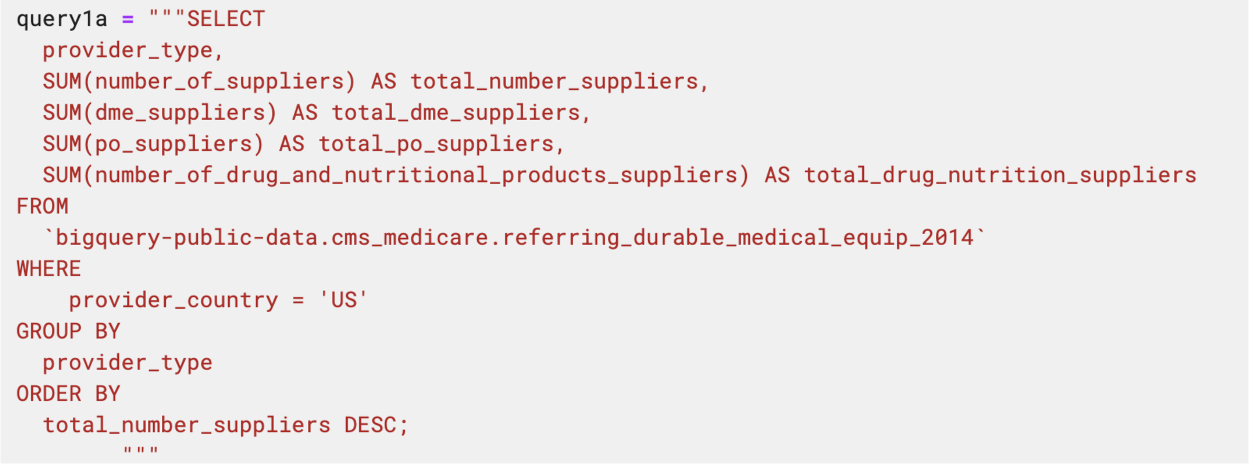
A. What are some trends that I notice in the medicare database? How can I use complex SQL methods to extract ideas from the dataset and try to tell a story with the data? How can I try to use the unique covariates and try to create new relationships using notable operators?

II. DataSet Description

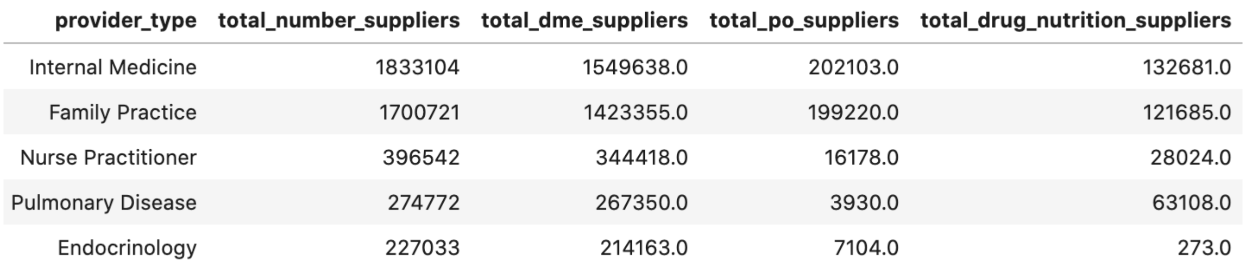
A. This dataset is accessible on Google Cloud using Big Query, from the CMMS. It has datasets regarding various types of medicare-related providers, services, drugs, costs, charges, etc. Since the data was not accessible on my local machine, I used JupyterLab with python and forked the notebook on Kaggle, allowing me to access the medicare database. It has a dozen datasets that span a couple years. I figured I can practice SQL queries and try to find some patterns. I can join datasets on a certain covariate since there are multiple datasets. Each dataset also has a different type of observation, which allowed me to make unique interpretations specific to that table.

III. Methods

Below I queried the referring provider dataset, in which I wanted to explore the number of suppliers. It seems that there is a regular number of suppliers that is unrelated to the DME, PO, and drug/nutrition supplier sum count. I wanted to put focus on the US provider country because that is where most of the records are from. I wanted to group on the provider_type. Since there are so many types, I wanted to see which are the top entities in terms of count of suppliers.



```
query1a = """SELECT
  provider_type,
  SUM(number_of_suppliers) AS total_number_suppliers,
  SUM(dme_suppliers) AS total_dme_suppliers,
  SUM(po_suppliers) AS total_po_suppliers,
  SUM(number_of_drug_and_nutritional_products_suppliers) AS total_drug_nutrition_suppliers
FROM
  `bigquery-public-data.cms_medicare.referring_durable_medical_equip_2014`
WHERE
    provider_country = 'US'
GROUP BY
  provider_type
ORDER BY
  total_number_suppliers DESC;
        """
```



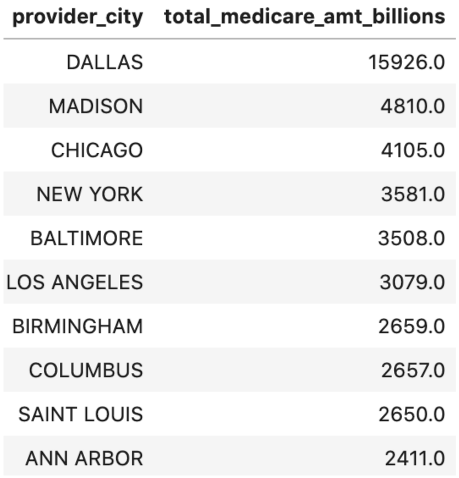
provider_type	total_number_suppliers	total_dme_suppliers	total_po_suppliers	total_drug_nutrition_suppliers
Internal Medicine	1833104	1549638.0	202103.0	132681.0
Family Practice	1700721	1423355.0	199220.0	121685.0
Nurse Practitioner	396542	344418.0	16178.0	28024.0
Pulmonary Disease	274772	267350.0	3930.0	63108.0
Endocrinology	227033	214163.0	7104.0	273.0

It looks like the top 5 provider types are from internal medicine, family practice, NP, pulmonary, and endocrinology departments as seen above. I am a bit surprised the emergency medicine is not here but it does make sense that internal medicine, which serves such a wide general audience of patients, is top with the total number of suppliers. It even leads in the other category counts as seen above.

I can analyze this table further by looking at the groupings of data by a US city. I can use the schema function to get to know more about the column descriptions and their relations with their neighboring covariates.

It seems that 'supplier_medicare_payment_amount' is the average amount that Medicare paid suppliers after deductible and coinsurance amounts have been deducted for the line item DMEPOS product/service. I can calculate the Total Medicare payment amounts by multiplying this number by the 'number_of_supplier_services'.

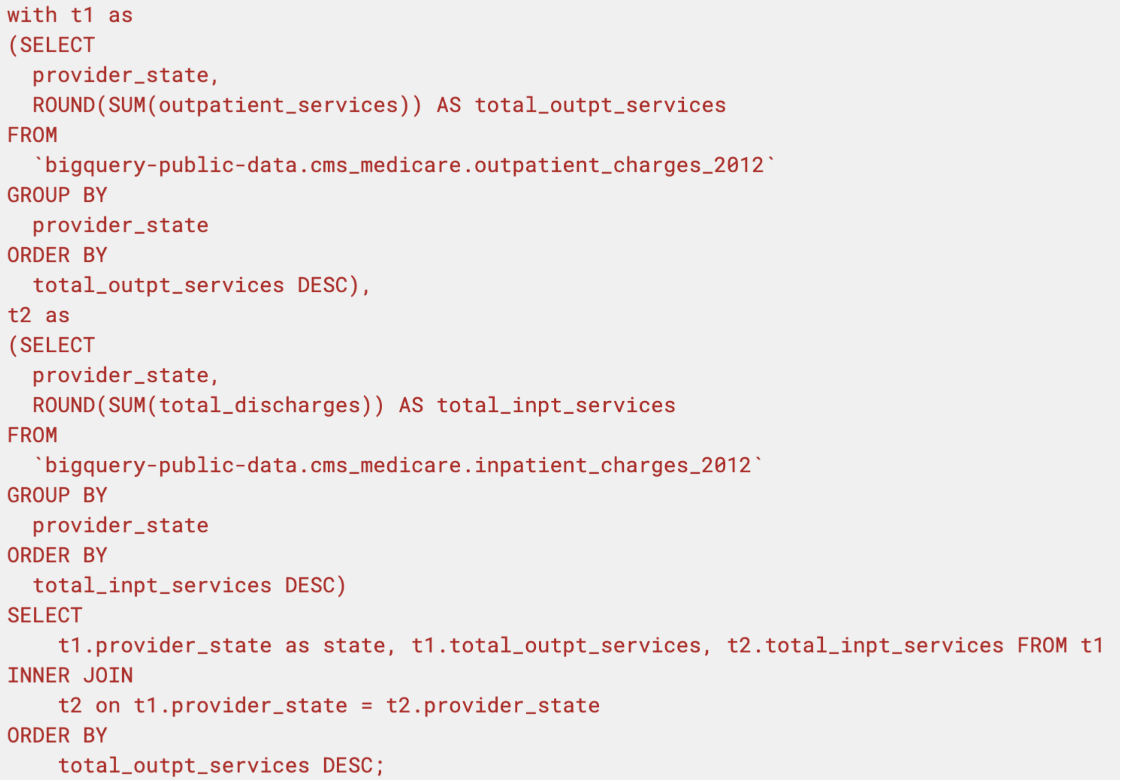
Performing this operation over a similar type of query from above gives me the following table.



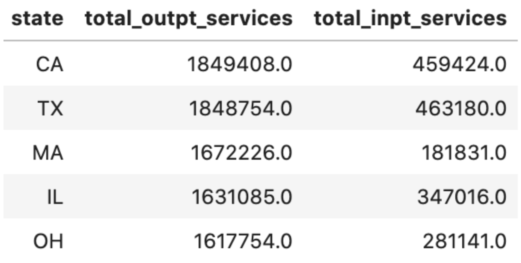
provider_city	total_medicare_amt_billions
DALLAS	15926.0
MADISON	4810.0
CHICAGO	4105.0
NEW YORK	3581.0
BALTIMORE	3508.0
LOS ANGELES	3079.0
BIRMINGHAM	2659.0
COLUMBUS	2657.0
SAINT LOUIS	2650.0
ANN ARBOR	2411.0

It seems that Dallas is leading all the cities by a larger margin for the total medicare payment amount. I thought it would be LA or NY as part of my predictions. It seems that there are two Madisons, one in Alabama and one in Wisconsin. The two different cities are being bunched together here, so I am going to group by state and other factors and explore the same relation. I was able to get a good picture of the top cities in the US, but accounting for all the duplicate cities seems like more of a problem. Looking at the table grouped by state, it looks like Texas, California, and Florida are the leading three states in the medicare payment amounts. Grouping over the provider_type gives me Nephrology at the number 1 spot with family, pulmonary, and internal medicine similar to our previous analysis.

The next two sets of dataset types I want to analyze the relationship between are the inpatient and outpatient services tables. There are a total of 5 tables per year 2011-2015 for each type. I first want to explore the dynamic of the schema during one year for example, and try to group the other datasets and try to provide a framework across the years.

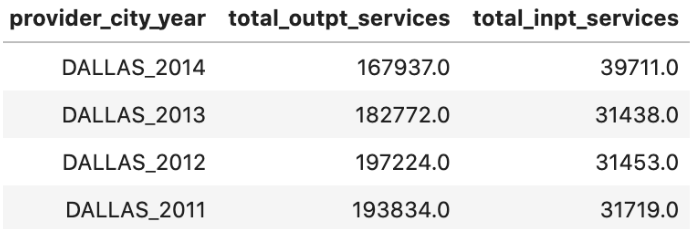


```
with t1 as
(SELECT
  provider_state,
  ROUND(SUM(outpatient_services)) AS total_outpt_services
FROM
  `bigquery-public-data.cms_medicare.outpatient_charges_2012`
GROUP BY
  provider_state
ORDER BY
  total_outpt_services DESC),
t2 as
(SELECT
  provider_state,
  ROUND(SUM(total_discharges)) AS total_inpt_services
FROM
  `bigquery-public-data.cms_medicare.inpatient_charges_2012`
GROUP BY
  provider_state
ORDER BY
  total_inpt_services DESC)
SELECT
    t1.provider_state as state, t1.total_outpt_services, t2.total_inpt_services FROM t1
INNER JOIN
    t2 on t1.provider_state = t2.provider_state
ORDER BY
    total_outpt_services DESC;
```



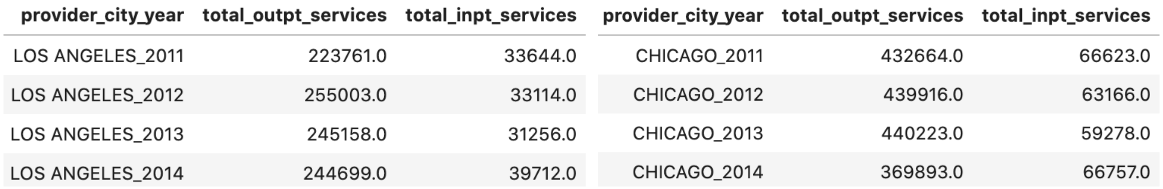
state	total_outpt_services	total_inpt_services
CA	1849408.0	459424.0
TX	1848754.0	463180.0
MA	1672226.0	181831.0
IL	1631085.0	347016.0
OH	1617754.0	281141.0

In the above query, I was able to use the join function to explore the covariates' relationships with each other. I can do a join on the provider state from inpatient and outpatient total discharges and sum over the count of the number of services. I wanted to do this for the city of Dallas, let's say, for all 5 years and compare over the 5 years what the inpatient vs outpatient difference is. I am going to use join again.



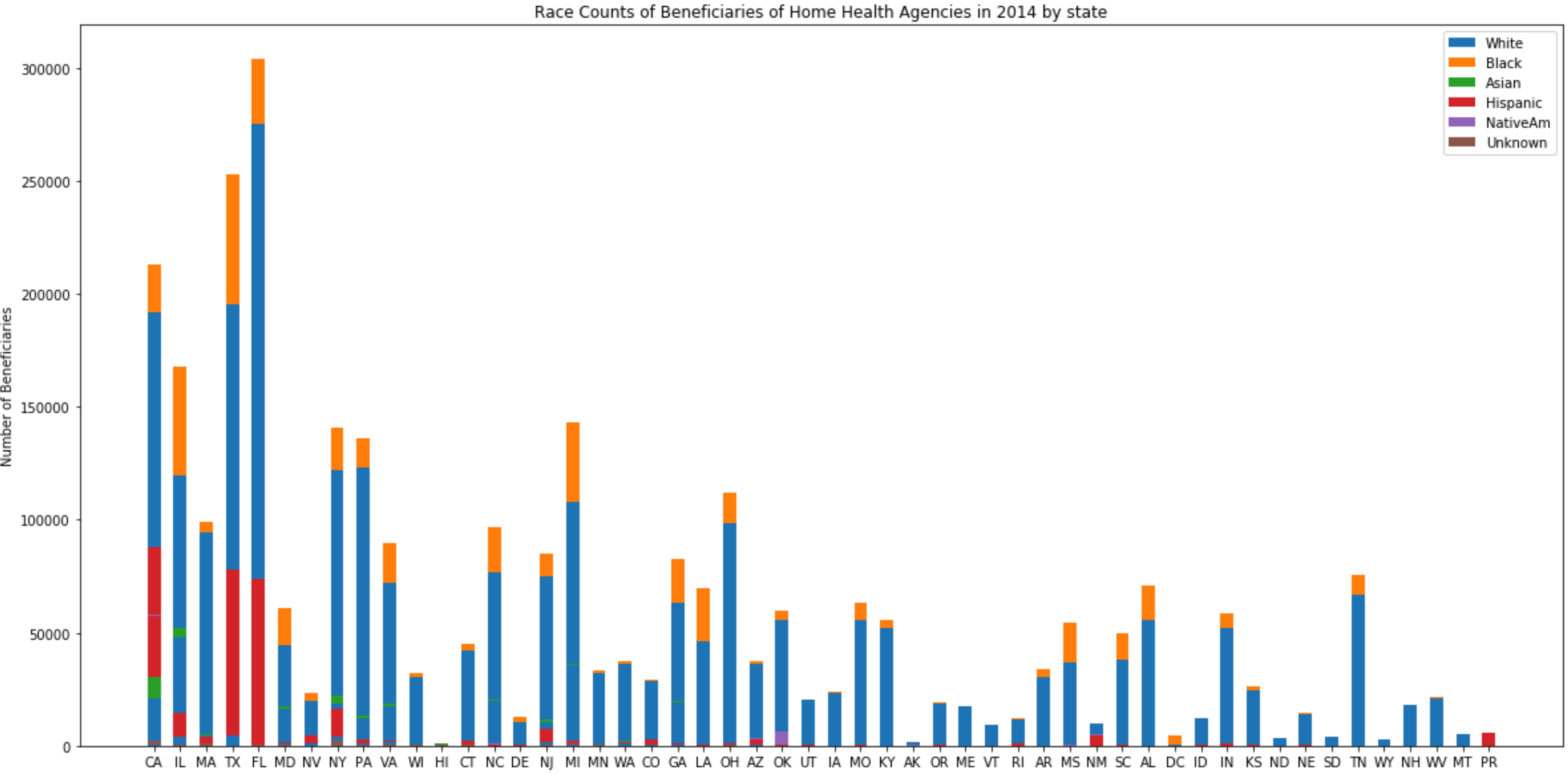
provider_city_year	total_outpt_services	total_inpt_services
DALLAS_2014	167937.0	39711.0
DALLAS_2013	182772.0	31438.0
DALLAS_2012	197224.0	31453.0
DALLAS_2011	193834.0	31719.0

So this is the output I received when I put together the same query as above, but for Dallas, as I wanted to explore this city more. I tried to join information from the 2015 inpatient and outpatient table but the cloud's connection was not as strong. Thus, the results only spanned from 2011 to 2014, which still gives me a good picture. It seems that for Dallas, the total number of outpatient services has been on the decline over time while the number of inpatient services has been around the same level relatively. I was able to group the data from multiple different datasets using a large, repetitive query, but it is able to give me a picture for Dallas. I can do the same for various cities or even group by a certain state using the same code.



provider_city_year	total_outpt_services	total_inpt_services
LOS ANGELES_2011	223761.0	33644.0
LOS ANGELES_2012	255003.0	33114.0
LOS ANGELES_2013	245158.0	31256.0
LOS ANGELES_2014	244699.0	39712.0

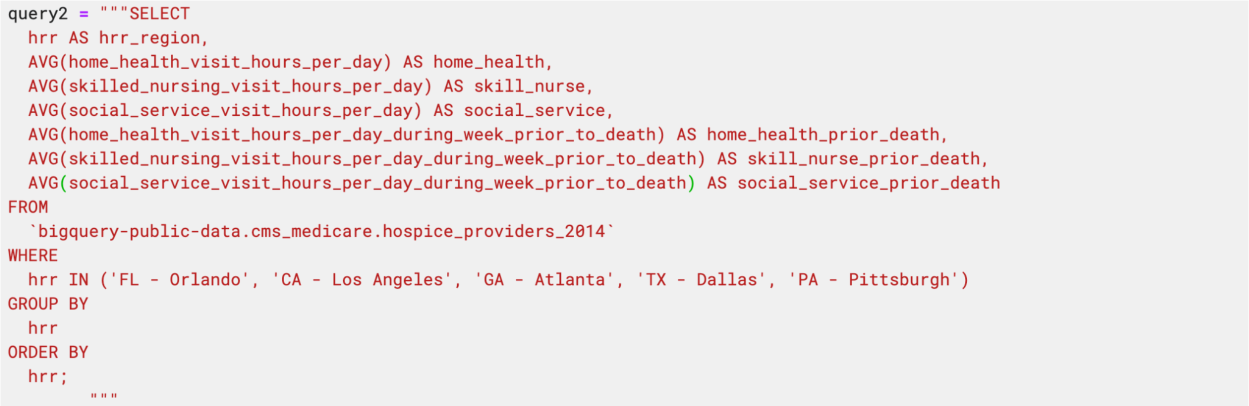
provider_city_year	total_outpt_services	total_inpt_services
CHICAGO_2011	432664.0	66623.0
CHICAGO_2012	439916.0	63166.0
CHICAGO_2013	440223.0	59278.0
CHICAGO_2014	369893.0	66757.0



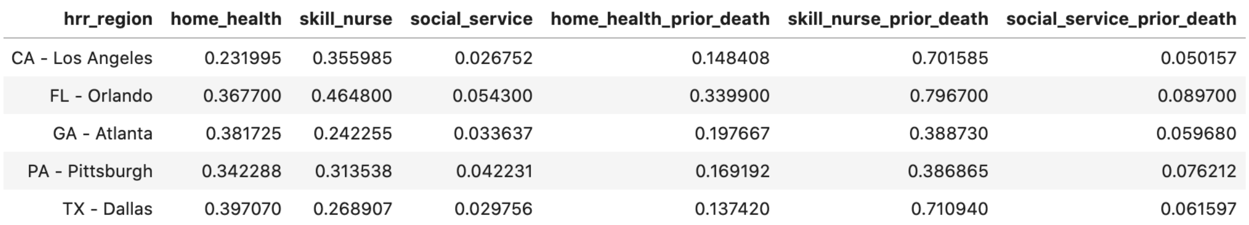
Now, I wanted to shift my attention to home health agency tables. They had a breakdown of race counts across the columns, so I wanted to query the results grouped by state. I thought it would be better to show the visual outcome of the query. I tried GGPLOT, Seaborn, and ultimately stuck with Matplotlib for a simple count of the races of the beneficiaries as part of the home health agency in 2014.

We can see that Florida, Texas, California, and Illinois are the clear leaders in terms of beneficiaries. It is also interesting to see just how much of the white race is dominating over this plot. The hispanic counts are very large in Texas and Florida compared to the other bars. Also, the Asian bar is significantly biggest in CA (most diverse) among other states.

Now, I went to explore the hospice providers dataset. I wanted to group by a different factor, and decided to go with the HRR region. Instead of grouping by state, I wanted to group by the referral region that was geographically bordered based upon the government maintenance costs, population demographics, urbanization, etc. Though there are more than 50 (counterparts to state), I wanted to first identify the top regions upon a metric. The metric I chose here was the total number of hospice care days. The top five regions that lead in this metric are FL - Orlando, CA - Los Angeles, GA - Atlanta, TX - Dallas, and PA - Pittsburgh.



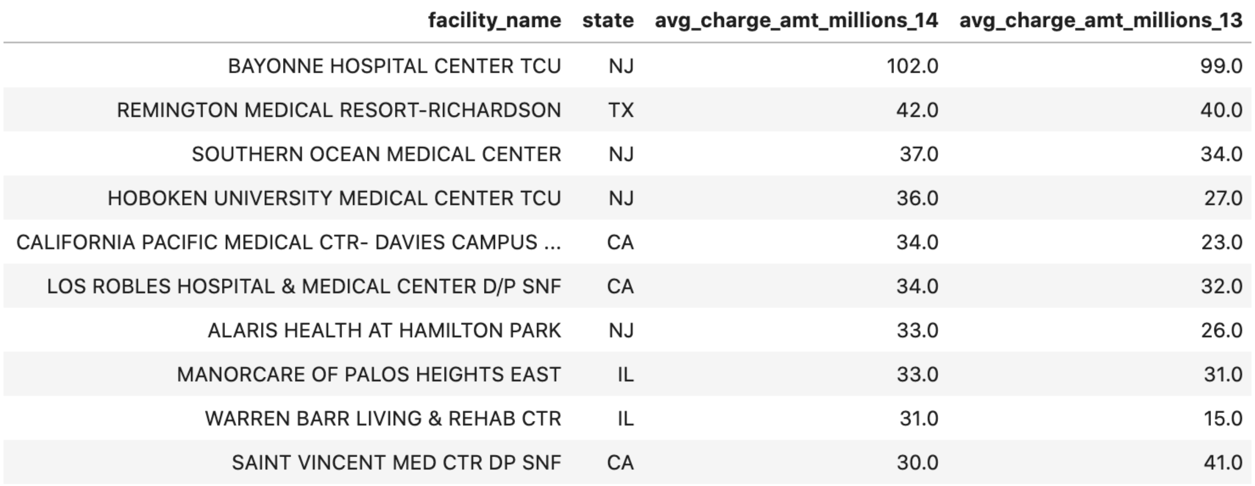
```
query2 = """SELECT
  hrr AS hrr_region,
  AVG(home_health_visit_hours_per_day) AS home_health,
  AVG(skilled_nursing_visit_hours_per_day) AS skill_nurse,
  AVG(social_service_visit_hours_per_day) AS social_service,
  AVG(home_health_visit_hours_per_day_during_week_prior_to_death) AS home_health_prior_death,
  AVG(skilled_nursing_visit_hours_per_day_during_week_prior_to_death) AS skill_nurse_prior_death,
  AVG(social_service_visit_hours_per_day_during_week_prior_to_death) AS social_service_prior_death
FROM
  `bigquery-public-data.cms_medicare.hospice_providers_2014`
WHERE
  hrr IN ('FL - Orlando', 'CA - Los Angeles', 'GA - Atlanta', 'TX - Dallas', 'PA - Pittsburgh')
GROUP BY
  hrr
ORDER BY
  hrr;
        """
```



hrr_region	home_health	skill_nurse	social_service	home_health_prior_death	skill_nurse_prior_death	social_service_prior_death
CA - Los Angeles	0.231995	0.355985	0.026752	0.148408	0.701585	0.050157
FL - Orlando	0.367700	0.464800	0.054300	0.339900	0.796700	0.089700
GA - Atlanta	0.381725	0.242255	0.033637	0.197667	0.388730	0.059680
PA - Pittsburgh	0.342288	0.313538	0.042231	0.169192	0.386865	0.076212
TX - Dallas	0.397070	0.268907	0.029756	0.137420	0.710940	0.061597

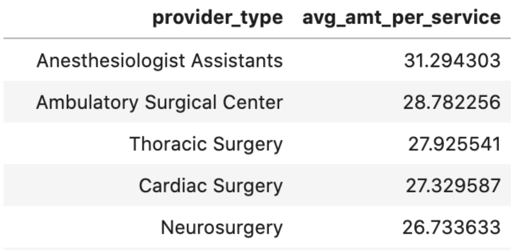
Just focusing on those five regions, I wanted to look at how the average number of visit hours from each health care provider changed from initial stages to the last stages of hospice care. Specifically, the average hours of visits per day from home health agencies, skilled nursing facilities, and social service providers. Looking across these time differences, it seems that the time spent on care decreased from all home health agencies. In this regard, Dallas, Atlanta, and Pittsburgh faced the largest decrease. The skilled nursing facility care increased in all regions and most prominently in Dallas, Orlando, and LA. Social service visiting hours have all increased but their involvement compared to our type of hospice care seems to be small. Home health and skilled nursing facilities' average hours of care seems to be similar in the initial stages of hospice care. It is also interesting that the five regions are spread out across the US. It notes highly populated areas across the country, adding to the credibility of it being representative.

Now, considering the skilled nursing facility datasets, I looked into the total average SNF charge amount for both of the years available 2013 and 2014. I wanted to group by facility_name this time instead, trying to see a new grouping.



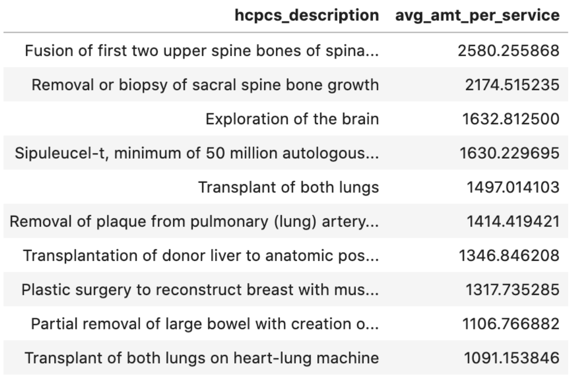
facility_name	state	avg_charge_amt_millions_14	avg_charge_amt_millions_13
BAYONNE HOSPITAL CENTER TCU	NJ	102.0	99.0
REMINGTON MEDICAL RESORT-RICHARDSON	TX	42.0	40.0
SOUTHERN OCEAN MEDICAL CENTER	NJ	37.0	34.0
HOBOKEN UNIVERSITY MEDICAL CENTER TCU	NJ	36.0	27.0
CALIFORNIA PACIFIC MEDICAL CTR- DAVIES CAMPUS ...	CA	34.0	23.0
LOS ROBLES HOSPITAL & MEDICAL CENTER D/P SNF	CA	34.0	32.0
ALARIS HEALTH AT HAMILTON PARK	NJ	33.0	26.0
MANORCARE OF PALOS HEIGHTS EAST	IL	33.0	31.0
WARREN BARR LIVING & REHAB CTR	IL	31.0	15.0
SAINT VINCENT MED CTR DP SNF	CA	30.0	41.0

In the above graphic, I joined both of the tables on the facility_name, with the charge amount being aggregated for both years matched on the facility name. It can be seen that in both years, Bayonne Hospital in NJ exceeds all other nursing facilities in the total average charges. We can also see that the states leading are CA, TX, NJ, and IL. It is interesting to see that New Jersey is taking up such a high charge since we have already seen the other states being leaders before. The scale was changed to millions so that the money amount can be compared easier. Warren Barr faced the largest increase proportionally while Saint Vincent did not see improvement from the year prior. The gap between Bayonne Hospital and the second place is very large.



provider_type	avg_amt_per_service
Anesthesiologist Assistants	31.294303
Ambulatory Surgical Center	28.782256
Thoracic Surgery	27.925541
Cardiac Surgery	27.329587
Neurosurgery	26.733633

Now looking in 2013 at the groupings by provider_type, I wanted to see the average charge amount and how it changes for the total services provided for that type of provider. It seems that Anesthesiologist assistants submit the most charges per service, while other surgery personnel take up the next few spots. Overall, it looks like surgery departments charge the most per service. There is CRNA on the 6th position but the others are all surgery-related.



hcpcs_description	avg_amt_per_service
Fusion of first two upper spine bones of spina...	2580.255868
Removal or biopsy of sacral spine bone growth	2174.515235
Exploration of the brain	1632.812500
Sipuleucel-t, minimum of 50 million autologous...	1630.229695
Transplant of both lungs	1497.014103
Removal of plaque from pulmonary (lung) artery...	1414.419421
Transplantation of donor liver to anatomic pos...	1346.846208
Plastic surgery to reconstruct breast with mus...	1317.735285
Partial removal of large bowel with creation o...	1106.766882
Transplant of both lungs on heart-lung machine	1091.153846

Venkata M M.

In 2015, I wanted to look at the description of the operations that the providers were involved in. I took the average charge amount over the accumulated average number of services. It looks like spine-related surgeries take up the top 2 spots. Then Brain exploration, prostate cancer therapy, and lung transplant follow next. Then it is the artery plaque removal, liver transplant, and breast reconstruction. It is interesting to see that spine surgeries are taking in the most charges, I thought it would be heart or brain surgery. This changed over the years. In 2012, prostate therapy was #1, in 2013 it was liver suture, and in 2014 it was lung transplant. So there is not any one HCPCS category that has been dominating at the top position.

Looking over the male vs female breakdown, it seems that the male providers have been taking a proportionally higher charge amount for 2012-2015. However, this metric is not as accurate to compare because there are a lot more records with no gender accurately identified. So, the male vs female comparison doesn’t hold much water here as a representative sample.

This table is organized by NPI so we can also gauge metrics to the NPI level, but I felt this was not as useful since it is such a large geographical span, that grouping broader will be better.