Venkata M M.

Kaggle Healthcare EDA Dataset SQL Analysis on SnowFlake

I. Research Question

What are some trends that can be noticed throughout a health care dataset that gathers information about typical patient attributes? We hope to use intermediate to complex SQL queries on Snowflake. Some methods that are going to be used include analysis across time periods, clustering data into various groups, and understanding the limitations of the dataset.

II. DataSet Description

This dataset was obtained from Kaggle.

https://www.kaggle.com/code/hainescity/healthcare-dataset-eda

Each column provides specific information about the patient, their admission, and the healthcare services provided, making this dataset suitable for various data analysis and modeling tasks in the healthcare domain. Here's a brief explanation of each column in the dataset -

Name: This column represents the name of the patient associated with the healthcare record.

Age: The age of the patient at the time of admission, expressed in years.

Gender: Indicates the gender of the patient, either "Male" or "Female."

Blood Type: The patient's blood type, which can be one of the common blood types (e.g., "A+", "O-", etc.).

Medical Condition: This column specifies the primary medical condition or diagnosis associated with the patient, such as "Diabetes," "Hypertension," "Asthma," and more.

Date of Admission: The date on which the patient was admitted to the healthcare facility.

Doctor: The name of the doctor responsible for the patient's care during their admission.

Hospital: Identifies the healthcare facility or hospital where the patient was admitted.

Insurance Provider: This column indicates the patient's insurance provider, which can be one of several options, including "Aetna," "Blue Cross," "Cigna," "UnitedHealthcare," and "Medicare."

Billing Amount: The amount of money billed for the patient's healthcare services during their admission. This is expressed as a floating-point number.

Room Number: The room number where the patient was accommodated during their admission.

Admission Type: Specifies the type of admission, which can be "Emergency," "Elective," or "Urgent," reflecting the circumstances of the admission.

Discharge Date: The date on which the patient was discharged from the healthcare facility, based on the admission date and a random number of days within a realistic range.

Medication: Identifies a medication prescribed or administered to the patient during their admission. Examples include "Aspirin," "Ibuprofen," "Penicillin," "Paracetamol," and "Lipitor."

Test Results: Describes the results of a medical test conducted during the patient's admission. Possible values include "Normal," "Abnormal," or "Inconclusive," indicating the outcome of the test.

III. Methods and Discussion

In this project I will showcase my ability to make complex SQL queries on a health care dataset. This dataset was obtained from Kaggle. It seems that the dataset is not real or could not have been generated randomly, artificially accumulated names aside. I cannot speak so strongly towards real-world implications though I make them to a certain extent under the assumption that if this dataset were real, I am able to pick it apart with SQL to generate actionable interpretations.

I ultimately picked this dataset for this project because healthcare data, even downloadable artificial data that is able to show many covariates per patient as the observation was very difficult to find upon hours of searching. I did find several public healthcare datasets but many of them had each observation as a group of encounters rather than at the patient or encounter level.

I hope to continue to be on the look-out for these types of datasets that are either publicly released or are non-aggregated datasets without the source dataset. I also hope to find these raw datasets referring to multiple other raw datasets that can be joined on certain covariates. This way, I am able to show my ability to explore the logical and conceptual level schemas among relational databases, which is what excites me to be able to explore when working with real raw data.

So we have 10,000 patients with certain health conditions, associated healthcare providers, insurance, medication taken, blood type, etc.

```
with t as
(select doctor, count(1) as doctor_ct
    from eda
    group by doctor
    order by doctor_ct desc),
t2 as
(select *, dense_rank() over(order by doctor_ct desc) as doctor_rank
from t)
select * from t2 limit 10
```

We can see that the top doctor with the most visits is Dr. Michael Johnson with 7 visits. The next 7 observations comprise the 2nd place tied for 5 patient encounters.

With a similar query, I was able to note the number of patients with the top 5 insurance providers,

It seems the Cigna is in the lead with Blue Cross and Aetna trailing very closely. We can try to compare this to the billing amount comparison that was done on Tableau.

```
select hospital, avg(datediff(day, date_of_admission, discharge_date)) as los_emergency
    from eda
    where admission_type = 'Emergency'
    group by hospital
    order by los_emergency desc limit 10
```

Looking at this query, I was able to see that Pearson LLC is the leading hospital with length of stay. I performed this calculation by looking at the date difference between the admission and discharge date. The average of the emergency visits (admission type) was around 1 month whereas the shortest was a day. Trying to quantify this in decimal days would be a more viable option. But there was no time stamp available as part of the data.

```
select hospital, avg(billing_amount) as avg_bill_amt
    from eda
    group by hospital
    order by avg_bill_amt desc limit 5
```

The top 3 hospitals with the highest average billing amount were Arellano-Mahoney, Ellison-Johnson, Thompson, Carlson and Kim reaching around $50,000 per patient encounter.

	INSURANCE_PROVIDER	AVG_BILL_AMT	R_NK
1	Aetna	25837.92	1
2	Cigna	25656.95	2
3	Blue Cross	25652.49	3
4	UnitedHealthcare	25404.69	4
5	Medicare	25002.48	5

The top 5 insurance providers with the highest average billing amount are shown above. We can explore furthermore into this trend. By trying to partition the data over the insurance provider, we can try to identify some patterns. We can also analyze the clustering over years, which is another short number of clusters.

```
with cte as
    (select *, monthname(date_of_admission) as adm_month
    from eda)
select adm_month, avg(billing_amount) as month_avg
from cte
group by adm_month
order by month_avg desc
```

This query shows that highest billing amounts come from December, September, and July. This is interesting to note that the second half of the year, notably the end, is where the billing ends up as the highest. Applying a similar function but aggregation across the number of patient visits and stratifying on emergency admission type shows that the month of October, August, and November lead. So there is easily a notable correlation between the number of visits and the billing amount.

I am also going to explore the data clusters in the dataset. I tend to look for factor level covariates that can allow me to group column data row-wise and be able to see the data in a new lens. Since this dataset has a large amount of rows, it has good clustering health. The clustering keys are the factor level covariates that I can aggregate counts, patients, amount, etc over a small group of clusters less than 4-10 categories generally.

Some possible partitions I can make on the data are the insurance providers, medication type, medication condition, and test_results, to name a few.

When stratifying on the quarters, it can be noted that the 4th quarter, the end of the year, leads in the number of patient encounters. I used the quarter function. Though this data is pretty equal in terms of distribution, another limitation is that this data is already clustered and filtered prior to being published since the number of patient encounters, though different, still span to a quarter of 10,000 observations.

```
select medical_condition, round(avg(billing_amount),2) as new_b_amt
    from eda
    group by medical_condition
    order by new_b_amt desc limit 5
```

The top medical condition that is raking up the highest billing amount is diabetes. Though HTN, stroke, and obesity are correlated and a patient might have multiple conditions, the dataset column limits us to just one condition. So a limitation that could have improved our findings if there is ranking of the top medical conditions. Or, trying to categorize patients with the family of medical conditions

Now, exploring the age cohort distribution can also be helpful. I have placed the ages into arbitrary bins to see the patient encounter histogram.

```
 select gender,
    sum(case when age < 18 then 1 else 0 end) as "Under 18",
    sum(case when age between 18 and 34 then 1 else 0 end) as "18-34",
    sum(case when age between 34 and 48 then 1 else 0 end) as "34-48",
    sum(case when age between 48 and 62 then 1 else 0 end) as "48-62",
    sum(case when age between 62 and 75 then 1 else 0 end) as "62-75",
    sum(case when age > 75 then 1 else 0 end) as "Over 75"
from eda
group by gender
order by gender
```

	GENDER	Under 18	18-34	34-48	48-62	62-75	Over 75
1	Female	0	1250	1104	1127	1078	730
2	Male	0	1235	1063	1152	971	716

This shows that there are no patients under 18. There is a relatively equal distribution of patients from 18-75 as seen in the cohorts above. Another limitation in the dataset here is that there are no patients under 18. This would span the pediatric sector, which is entirely discarded as part of this dataset. So our conclusions cannot be extrapolated to this department. Looking now across male and female, we can see that the variance is once again small. The count values are relatively close amongst each other in a given category, distanced out by 100-200 units. Overall, the male and female count were extracted equally under the assumption this data was extracted from a larger dataset or generated nonrandomly, or the study was conducted with this separation count in mind.

```
select medical_condition,
 sum(case when test_results = 'Inconclusive' then 1 else 0 end) as Inc_ct,
 sum(case when test_results = 'Normal' then 1 else 0 end) as Norm_ct,
 sum(case when test_results = 'Abnormal' then 1 else 0 end) as Abn_ct,
 from eda
 group by medical_condition
 order by medical_condition
```

Using the above query, I was able to pivot test_results count over the medical condition category. I was able to view the distribution of the health condition groups over the three types of results. It can be seen that the variance among the group counts is small.

	MEDICAL_CONDITION	INC_CT	NORM_CT	ABN_CT
1	Arthritis	553	542	555
2	Asthma	551	534	623
3	Cancer	556	570	577
4	Diabetes	542	544	537
5	Hypertension	554	532	602
6	Obesity	521	545	562

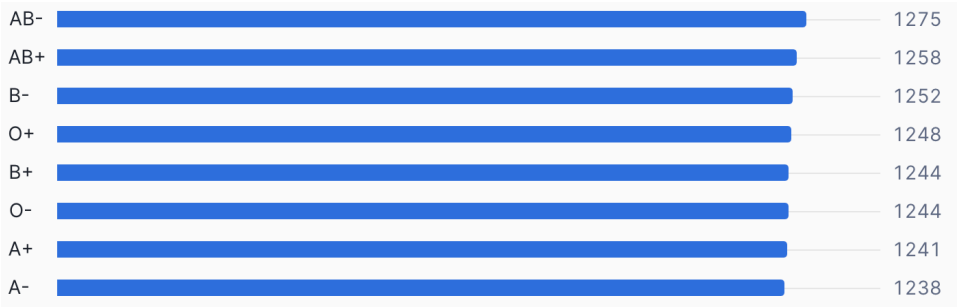
We can see that any one test result does not over dominate any one condition. However, the argument can be made that abnormal counts tend to have less homoscedasticity. It can be interpreted that having an abnormal test result for a patient having major medical conditions tends to vary a bit more. Test results being ineffective curiously is evenly grouped. So the number of times a doctor is unsure about the diagnosis does not depend so much on the type of major condition.

	INSURANCE_PROVIDER	ELECTIVE_AMT	URGENT_AMT	EMERGENCY_AMT
1	Aetna	17185694.85	17322360.30	17813739.39
2	Blue Cross	17856641.97	18258618.05	16010598.83
3	Cigna	17353470.87	18310270.42	16676430.35
4	Medicare	15844987.77	16762069.37	15522717.92
5	UnitedHealthcare	15700526.06	17379868.35	17170073.51

A similar query was applied onto the insurance provider across the admission type. Instead of returning the counts, I summed over the billing amount. I wanted to explore the relationships between the urgency of the visit to how much the visit's billing amount comes to, and I wanted to see this comparison over the different top insurance companies. It can be seen that for those choosing medicare tend to get billed an overall less amount comparatively across other insurance providers. UHC however could not match the other 3 private health insurance companies in the elective department. We can make an interpretation that when UHC patients

choose to schedule an appointment instead of waiting until an emergency despite the urgency, they are rewarded for their personal preventative health care by being charged less overall. This is not the pattern with the other 4 insurance companies, since the total billing amount is less varied in the other four. I tried to format the numbers into a dollar amount with commas but the format() function is not supported onto Snowflake.

This is a bucketed chart from Snowflake. I could not make a pie chart, which would have been the ideal visual in this case.



Since I don't have further information regarding the room numbers, I did not want to explore their relation with the other covariates. If there was a geographical advantage to the room numbers that make a certain room exceptional in the realm of treatment which can affect the billing amount, number of visits stratified over insurance companies for example, I may be able to produce associated interpretations.