Research Questions

Andy Choens

11/16/22

Thank you for the opportunity to apply for a data analyst position with NYSTEC. I have used the document provided for this test as a starting point for this document. A theme repeated throughout this document is that I would prefer to refine requirements/understanding with stakeholders before attempting to understand this kind of a vague report. However, some reports must be completed quickly. In that case, the best approach is to interpret the requirements broadly and provide as much data/insight as possible in the time given to maximize the chances of successfully completing the report.

I implemented this solution using Quarto, some simple R, and SQL. The tables in this report are define by the in-line SQL and are formatted by Quarto/pandoc. This is a high quality open source method for producing reproducible reports, dashboards, etc.

# Research Question 1

The question, as defined by the test:

For payment period Jan 2016 through March 2016, identify the institutional claims by claim type (claim\_type\_cd) and Category of Service (COS\_CD) with the highest dollar paid amounts. Include the total number of claims and distinct number of members.

Note: Institutional claims are identified using the CLAIM\_CLASS\_CD. Include the Claim Type Cd and COS\_CD descriptions.

## Discussion

The use case for this question was not specified and the requirements are somewhat vague.

* What are the:
  + Assumptions:
    - Claims include a number of dates including service date, payment date, admit date, and discharge date. Based on the language “for payment period”, this analysis will use the payment date to filter the claims to claims paid for between Jan 2016 and March 2016.
    - The report requirements do not specify if the desired analysis is the highest total spend or the highest average spend per claim. However, the report does request the number of claims and members, thus the average can be derived if the total spend is included.
    - The report requirements are not clear if the report should stratify costs by both claim type AND category of service, or both separately. To avoid any possible confusion, this report includes all possible interpretations of that requirement. Future reports will be shorter as requirements are better understood.
    - Although the report identifies CLAIM\_TYPE\_CD and COS\_CD as columns to stratify by, this report uses CLAIM\_TYPE\_DESC and COS\_DESC, respectively, as they provide a more human friendly result.
  + Definitions:
    - The most important definition in this report is “highest dollar paid amounts”. This could be interpreted as total paid or highest average paid, or even the highest individual claim.
  + Questions:
    - I would like to better understand why this time frame was chosen.
    - I would like to better understand exactly what they mean by “highest dollar paid amounts”.
* What is the Expected Output: A table showing the claims, grouped by claim type and category of service in descending order of value. Because report requirements are vague, more than one answer will be provided. As requirements are refined, future reports will be shorter.
* What are our Filters/Criteria:
  + We are only interested in claims paid for in the first quarter of 2016.
* Analysis Approach: The report requirements are vague. My preference would be to first refine the report requirements with key stakeholders. When that is not feasible, the next best response is to identify a reasonable set of interpretations and provide all of them in the first report, with the expectation that future reports will refine the requirements and be more specific. This report could also be fulfilled in an Excel file, giving the recipient the flexibility to sort the data in several ways.

## SQL

This report provides three interpretations of the report requirements. Future reports will be provided as requirements are clarified.

### Results Grouped By Claim Type

The code/table below calculate the total paid and average paid per claim for each claim type. Results are sorted by descending total paid, although this could just as easily be sorted by the descending average paid per claim.

-- DuckDB does not allow the user to easily define variables.  
-- I prefer to define filter variables such as the date range  
-- at the TOP of the query, rather than in-line as done here.  
select  
 CLAIM\_TYPE\_DESC as 'Claim Type',  
 sum(PAID\_AMOUNT) as 'Total Paid',  
 count(distinct CLAIM\_TRANS\_ID) as 'Number of Claims',  
 round(avg(PAID\_AMOUNT),2) as 'Average Paid Per Claim',  
 count(distinct MBR\_ID) as'Number of Unique Members'  
from CLAIM\_TRANS  
where PMT\_DT between '2016-01-01' and '2016-03-31'  
group by CLAIM\_TYPE\_DESC  
order by sum(PAID\_AMOUNT) desc

9 records

| Claim Type | Total Paid | Number of Claims | Average Paid Per Claim | Number of Unique Members |
| --- | --- | --- | --- | --- |
| INPATIENT | 267952.25 | 2 | 11650.10 | 1 |
| MANAGED CARE CAPITATION | 21111.85 | 5 | 3015.98 | 3 |
| PRACTITIONER | 4869.33 | 10 | 105.86 | 4 |
| RESIDENTIAL HEALTH CARE (NURSING HOME) | 3700.57 | 7 | 119.37 | 1 |
| CLINIC | 2289.23 | 7 | 254.36 | 2 |
| SUPPLY MEDICAL (DME) | 1320.02 | 7 | 94.29 | 2 |
| HOME HEALTH AGENCY - PERSONAL CARE | 1049.56 | 4 | 209.91 | 2 |
| TRANSPORTATION | 632.58 | 5 | 79.07 | 3 |
| LABORATORY | 32.07 | 2 | 10.69 | 1 |

### Grouped By Category of Service

The code/table below calculate the total paid and average paid per claim for each category of service. Results are sorted by descending total paid, although this could just as easily be sorted by the descending average paid per claim.

select  
 COS\_DESC as 'Category of Service',  
 sum(PAID\_AMOUNT) as 'Total Paid',  
 count(distinct CLAIM\_TRANS\_ID) as 'Number of Claims',  
 round(avg(PAID\_AMOUNT),2) as 'Average Paid Per Claim',  
 count(distinct MBR\_ID) as'Number of Unique Members'  
from CLAIM\_TRANS  
where PMT\_DT between '2016-01-01' and '2016-03-31'  
group by COS\_DESC  
order by sum(PAID\_AMOUNT) desc

Displaying records 1 - 10

| Category of Service | Total Paid | Number of Claims | Average Paid Per Claim | Number of Unique Members |
| --- | --- | --- | --- | --- |
| HOS SVC: INPATIENT | 267952.25 | 2 | 11650.10 | 1 |
| PCPS: CAPITATION PROVIDER | 21111.85 | 5 | 3015.98 | 3 |
| D&TCS: DIAGNOSTIC & TREATMENT CENTER SVCS (FSTNDG) | 5071.32 | 10 | 144.89 | 2 |
| PHYS SVC: PHYSICIAN SERVICES | 4869.33 | 10 | 105.86 | 4 |
| MADS: MEDICAL APPLIANCE, EQUIPMENT, SUPPLY DEALER | 1320.02 | 7 | 94.29 | 2 |
| HOS SVC: HOSPITAL BASE OUTPATIENT SERVICES | 918.48 | 4 | 183.70 | 2 |
| HHAS: CASE MANAGEMENT SERVICES | 769.56 | 3 | 256.52 | 1 |
| TRANS: DAY TREATMENT | 380.48 | 1 | 380.48 | 1 |
| GROUPS: CHAIN PHARMACY (NOT PAYABLE, NOT ON BILL) | 280.00 | 1 | 140.00 | 1 |
| TRANS: AMBULANCE | 160.50 | 1 | 80.25 | 1 |

## Grouped by Claim Type and Category of Service

The code/table below calculate the total paid and average paid per claim for each claim type and category of service. Results are sorted by descending total paid, although this could just as easily be sorted by the descending average paid per claim.

select  
 CLAIM\_TYPE\_DESC as 'Claim Type',  
 COS\_DESC as 'Category of Service',  
 sum(PAID\_AMOUNT) as 'Total Paid',  
 count(distinct CLAIM\_TRANS\_ID) as 'Number of Claims',  
 round(avg(PAID\_AMOUNT),2) as 'Average Paid Per Claim',  
 count(distinct MBR\_ID) as'Number of Unique Members'  
from CLAIM\_TRANS  
where PMT\_DT between '2016-01-01' and '2016-03-31'  
group by CLAIM\_TYPE\_DESC, COS\_DESC  
order by sum(PAID\_AMOUNT) desc

Displaying records 1 - 10

| Claim Type | Category of Service | Total Paid | Number of Claims | Average Paid Per Claim | Number of Unique Members |
| --- | --- | --- | --- | --- | --- |
| INPATIENT | HOS SVC: INPATIENT | 267952.25 | 2 | 11650.10 | 1 |
| MANAGED CARE CAPITATION | PCPS: CAPITATION PROVIDER | 21111.85 | 5 | 3015.98 | 3 |
| PRACTITIONER | PHYS SVC: PHYSICIAN SERVICES | 4869.33 | 10 | 105.86 | 4 |
| RESIDENTIAL HEALTH CARE (NURSING HOME) | D&TCS: DIAGNOSTIC & TREATMENT CENTER SVCS (FSTNDG) | 3700.57 | 7 | 119.37 | 1 |
| CLINIC | D&TCS: DIAGNOSTIC & TREATMENT CENTER SVCS (FSTNDG) | 1370.75 | 4 | 342.69 | 2 |
| SUPPLY MEDICAL (DME) | MADS: MEDICAL APPLIANCE, EQUIPMENT, SUPPLY DEALER | 1320.02 | 7 | 94.29 | 2 |
| CLINIC | HOS SVC: HOSPITAL BASE OUTPATIENT SERVICES | 918.48 | 4 | 183.70 | 2 |
| HOME HEALTH AGENCY - PERSONAL CARE | HHAS: CASE MANAGEMENT SERVICES | 769.56 | 3 | 256.52 | 1 |
| TRANSPORTATION | TRANS: DAY TREATMENT | 380.48 | 1 | 380.48 | 1 |
| HOME HEALTH AGENCY - PERSONAL CARE | GROUPS: CHAIN PHARMACY (NOT PAYABLE, NOT ON BILL) | 280.00 | 1 | 140.00 | 1 |

# Research Question 2:

The question, as defined by the test:

For Research Question 2, you are being asked to determine what the question is based on the SQL below. Please fill in the Question and all the bullets below. This is a ‘reverse engineering’ exercise.

## SQL

This was provided by the test. I was not able to run this query with the information I have in the provided data set.

I don’t like the way this query is structured. The join clause includes filtering criteria not needed for the actual join. It would be easy to skim this query for the where clause and not notice the filter criteria defined in the join clause. In a small query like this, it is a minor detail. In a longer/more complex query it can easily lead to misunderstandings and misuse of a previously written query.

SELECT  
 CLAIM\_TABLE.MBR\_ID,  
 CLAIM\_TABLE.BILL\_NPI,  
 PROV\_TABLE.NPI\_NAME,  
 COUNT (DISTINCT CLAIM\_TABLE.ADMIT\_DATE)  
FROM  
 CLAIM\_TABLE  
JOIN  
 PROV\_TABLE  
 ON CLAIM\_TABLE.BILL\_NPI = PROV\_TABLE.NPI  
WHERE  
 CLAIM\_TABLE.ADMIT\_DATE BETWEEN TO\_DATE('01-JUL-2015','DD-MON-YYYY') AND TO\_DATE('30-JUN-2016','DD-MON-YYYY')  
 AND  
 CLAIM\_TABLE.CLAIM\_TYPE\_DESC = 'INPATIENT'  
 AND   
 CLAIM\_TABLE.MBR\_AGE\_ON\_DOS BETWEEN 18 AND 64  
GROUP BY CLAIM\_TABLE.MBR\_ID, CLAIM\_TABLE.BILL\_NPI, PROV\_TABLE.NPI\_NAME  
HAVING  
 COUNT(DISTINCT CLAIM\_TABLE.ADMIT\_DATE) > 3;

## Discussion

* What are the:
  + Questions: The question here appears to be a list of patients with more than three inpatient admissions at the same hospital between July 2015 and June 2016.
  + Assumptions:
    - For inpatient records, BILL\_NPI is a hospital. For example NPI 1548374549 is New York Presbyterian in Brooklyn.
  + Definitions:
    - This report does not use the standard New York State fiscal year as I remember it. Without understanding the use case of the report, I cannot comment further, but I would want to better understand why that date range was chosen. And the reason may be very good.
    - This report is based on the begin date of the inpatient stay, not the discharge date, which is not how DOH OQPS counts episodes of inpatient care. But, there may be a good reason for this and/or a standard I am not familiar with.
* What is the Expected Output:
  + The output of this query will be a list of patients, hospitals and the count of inpatient stays the member has had at that hospital in the given time frame.
* What are our Filters/Criteria:
  + Inpatient claims only.
  + Inpatient episode must have started between the beginning of July, 2015 and the end of June, 2016.
  + Only includes adults, 18 to 64. This excludes most Medicare members, but not all.
  + Only includes patients with more than three inpatient stays at the same hospital.
* Analysis Approach:
  + I would first want to better understand the use case of this report. There are several things which stick out to me, but I don’t want to judge them without first knowing why this report is wanted and how it would be used.