1. *Review the datasets and provide a commentary on their contents*

Here is an overview of the data:

Concept:

This table gives the general information about the concepts and their validity period, and the categories of each concept based on domain and vocabulary.

It includes the health care solutions (i.e. Ventilators, Drug Injection, etc). also, it categorizes each concept in different Domains and Vocabularies including:

Domain	Counts
Condition	234
Device	1373
Drug	9415
Measurement	368
Observation	1211
Place of Service	1
Procedure	3395
Provider Specialty	3
Grand Total	16000

Vocabulary	Counts
ATC	136
CPT4	3466
HCPCS	3115
ICD9CM	312
ICD9Proc	176
NDC	8790
NUCC	3
Place of Service	1
VA Class	1
Grand Total	16000

The other columns are validity start date and validity end date of each concept. “Invalid_reason” can be D (deleted row) or U (replaced by the updated row).

Conditions:

This table provides the conditions of each patient by presenting set of IDs. It also provides condition period for each patient. The below charts shows the distribution of condition periods for all the patients.

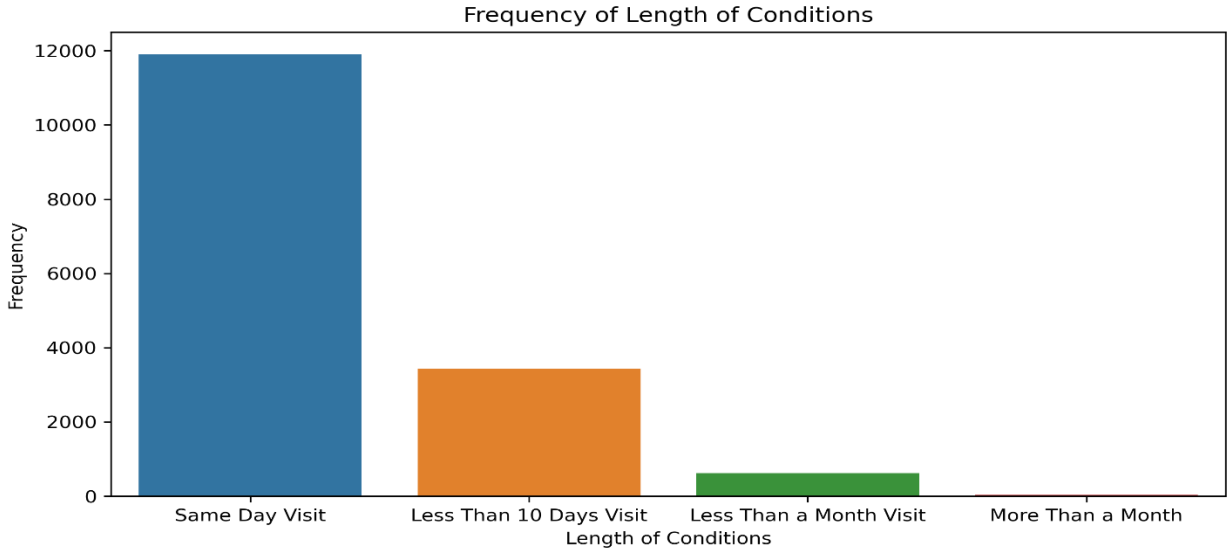


Chart shows that most cases had same day condition and visit of health care facility. Also, it indicates that the number of cases decreases constantly as the length of stay increases.

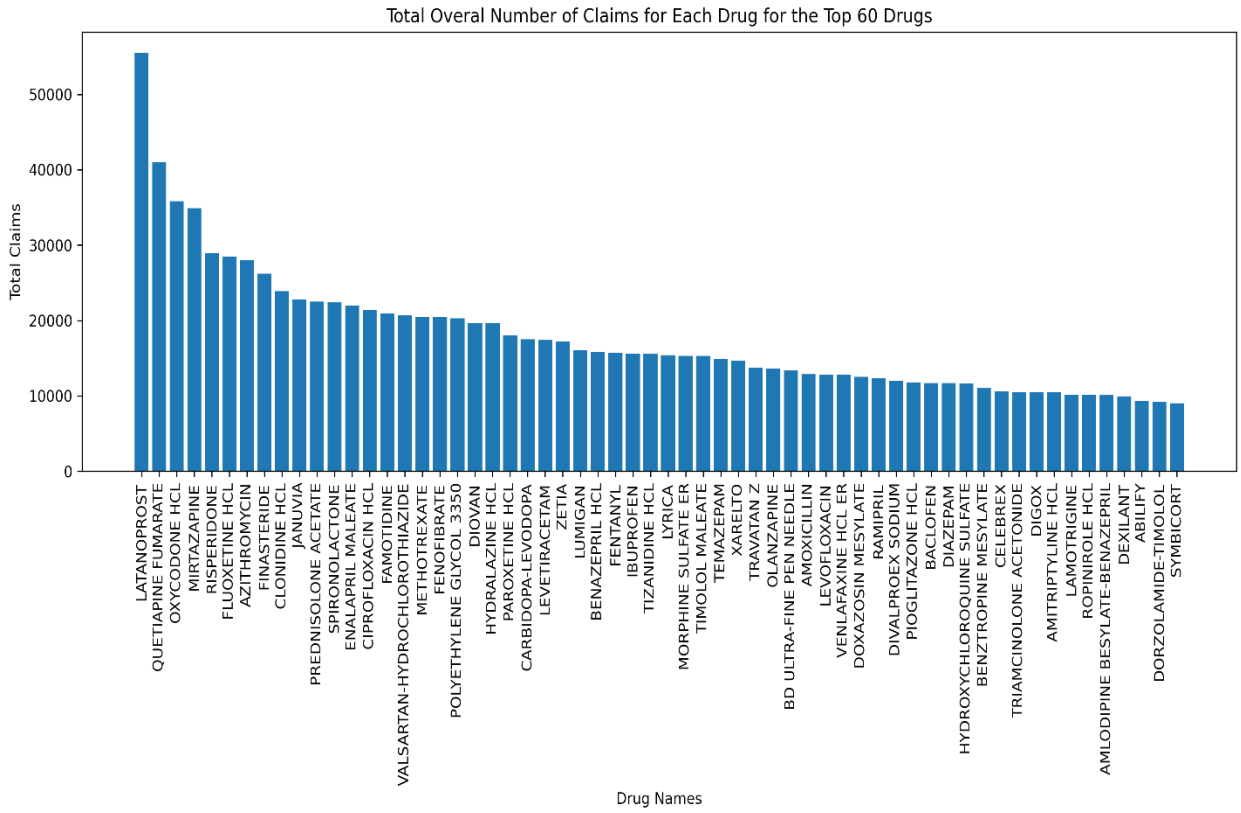
Sample_Part_D:

This table gives information about the providers. It gives information about city and state of provider, specialty, the list of drugs they used, bene count (the total # of part D beneficiaries with at least one claim drug), total claim count, total_drug_cost, etc.

Also, besides “total_claim_count”, data includes “total_claim_count_ge65”, and same for the cost columns, etc.

“total_claim_count_ge65” means total claims in people above 65 years old. Based on my research in the data, “total_claim_count” includes the counts in “total_claim_count_ge65” as well. That is the same thing about “total_drug_cost” which includes “total_drug_cost_ge65” too. Same case is correct about “total_30_day_fill_count” and “total_30_day_fill_count_ge65”. This data file provides more information like total supply per day and for people over 65 years old as well (and same scenario as above).

Chart below shows the top 60 most frequently used drugs in all the facilities by specialists.



2. ***Generate a query that shows the total claim counts, total cost spent for each NPI and drug***

Python:

```
sample_d.groupby(['npi','drug_name'])['total_claim_count','total_drug_cost'].sum()
```

SQL:

```
SELECT sum(total_claim_count), sum(total_drug_cost)

FROM sample_d

GROUP BY npi, drug_name;
```

3. ***Generate a query that shows the average cost of medications by specialty, and only show the top 3 medications with the highest average***

Python:

```
sample_d.groupby('specialty_description').total_drug_cost.mean().sort_values(ascending=False).head(3)
```

SQL:

```
SELECT AVG(total_drug_cost)
FROM sample_d
GROUP BY specialty_description
ORDER BY 1 DESC
LIMIT 3;
```

4. ***Generate a query that provides patient identifier and the description of their condition***

Python:

```
pd.merge(concept,conditions,left_on='concept_id',right_on='condition_source_concept_id')[['person_id','concept_name']]
```

SQL:

```
SELECT cond.person_id, conc.concept_name
FROM concept conc
JOIN conditions cond
ON conc.concept_id = cond.condition_source_concept_id;
```

The codes with the results are available in:

https://github.com/fradmehr/Patient_Data/blob/main/Patient%20Data.ipynb