Venkata M M.

CHHS COVID-19 Aggregated Dataset Exploration

I. Scientific Question

A. What are the geographical implications of the ration of the number of patient encounters to the ratio of available treatment? The objective is to describe a statewide space pertaining to the socioeconomic qualities that can be analyzed through SQL queries.

II. Dataset description

This dataset was obtained from the California Health and Human Services Open Data Portal where it shows the ER ratio of encounters that are categorized by ownership, geographic classification, and notable health-related conditions including but not limited to s heart and lung diseases, cancer, COVIDe-19, and diabetes. So each observation is an encounter

Covariate	Abbreviation	Data Type	Description
OSHPD_ID	oshpd_id	Number	OSHPD ID for the facility
Facility Name	FacilityName2	Plain Text	Name of the facility
County Name	CountyName	Plain Text	County the facility is located in
System	system	Plain Text	Hospital system the facility is a part of (if applicable)
Licensed Bed Size	LICENSED_BED_SIZE	Plain Text	Category(range) of ED treatment stations

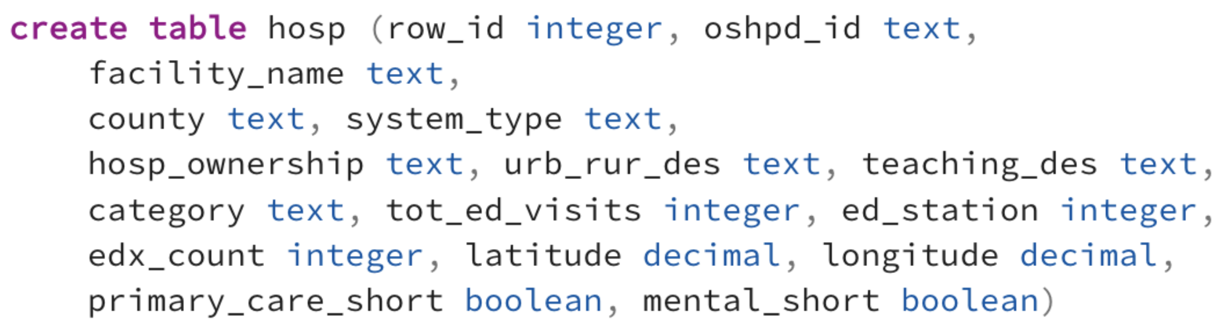
Hospital Ownership	HospitalOwnership	Plain Text	Facility ownership category
Urban Rural Designation	UrbanRuralDesi	Plain text	The area designation for the location of the facility
Teaching Designation	TEACHINGDesignation	Plain Text	Indicates if a facility is a teaching or non-teaching facilty
Category	Category	Plain Text	Health-relation condition
ED Encounters	Tot_ED_NmbVsts	Number	Total number of ED Encounters for the facility
ED Stations	EDStations	Number	Number of ED treatment stations
ED Burden	EDDXCount	Number	Number of ED visits for the specific category (health-related condition)
Latitude	LATITUDE	Number	Facilty latitude
Longitude	LONGITUDE	Number	Facility longitude
HPSA - Primary Care	PrimaryCareShortageArea	Plain Text	Indicates if a facility is in a Health Professional Shortage Area - Primary Care

HPSA - Mental Health	MentalHealthShortageArea	Plain Text	Indicates if a facility is in a Health Professional Shortage Area - Mental Health

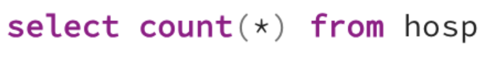
III. Methods and Discussion

I am going to run several SQL queries on PostGreSQL. I wanted to use this since MySQL is not as effective for large datasets.

First, I am going to create a table with the correct data type and import the .csv file into the interface so I can pass on some queries.

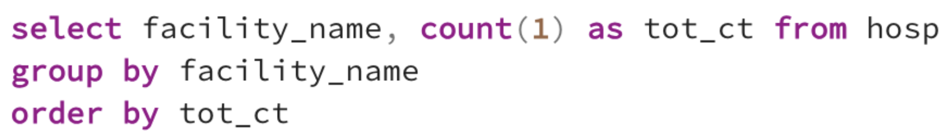


```
create table hosp (row_id integer, oshpd_id text,
    facility_name text,
    county text, system_type text,
    hosp_ownership text, urb_rur_des text, teaching_des text,
    category text, tot_ed_visits integer, ed_station integer,
    edx_count integer, latitude decimal, longitude decimal,
    primary_care_short boolean, mental_short boolean)
```



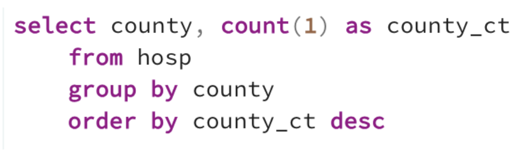
```
select count(*) from hosp
```

This gave me a dataset with 4265 rows and 16 columns, wherein the first column is the row_id. The bed size covariate was excluded because it resulted in problems to import the data.



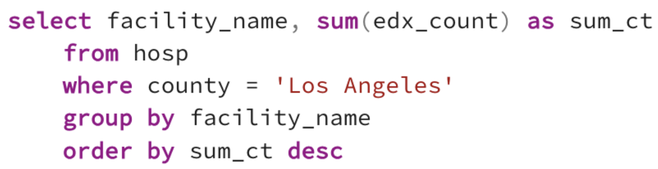
```
select facility_name, count(1) as tot_ct from hosp
group by facility_name
order by tot_ct
```

There are 251 facilities, with two facilities having 16 records and the rest have 17 records. There are 3925 non-teaching while the other 340 are teaching



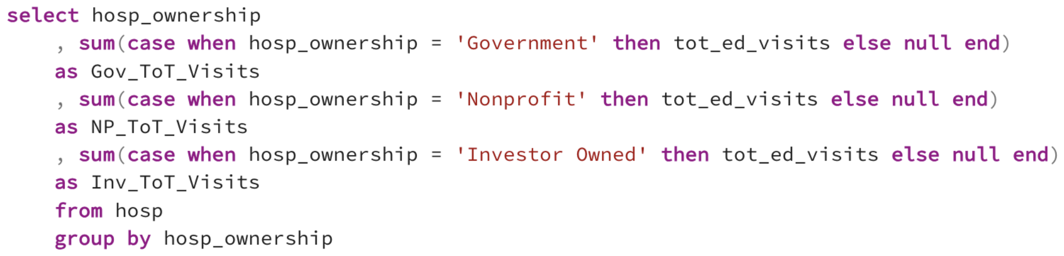
```sql
select county, count(1) as county_ct
    from hosp
    group by county
    order by county_ct desc
```

There are 52 counties. The top 5 counties with the most observations are LA, Orange, San Bernardino, Riverside, and Alameda counties. We can explore further more into these counties and maybe find some trends.



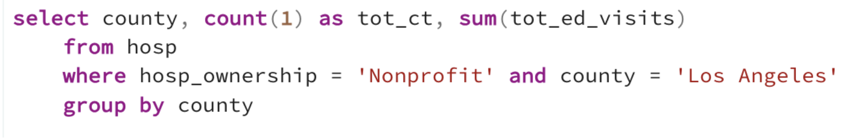
```sql
select facility_name, sum(edx_count) as sum_ct
    from hosp
    where county = 'Los Angeles'
    group by facility_name
    order by sum_ct desc
```

The top three hospitals with the most ED visits in the LA county are Antelope Valley Hospital, Kaiser in Downey, and Kaiser in West LA.

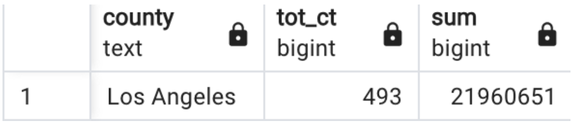


```sql
select hosp_ownership
    , sum(case when hosp_ownership = 'Government' then tot_ed_visits else null end)
    as Gov_ToT_Visits
    , sum(case when hosp_ownership = 'Nonprofit' then tot_ed_visits else null end)
    as NP_ToT_Visits
    , sum(case when hosp_ownership = 'Investor Owned' then tot_ed_visits else null end)
    as Inv_ToT_Visits
    from hosp
    group by hosp_ownership
```

We can see that nonprofits lead the ED burden with a ratio of 104M while investor owned is at 26M and government at 21M.



```sql
select county, count(1) as tot_ct, sum(tot_ed_visits)
    from hosp
    where hosp_ownership = 'Nonprofit' and county = 'Los Angeles'
    group by county
```



	county text 🔒	tot_ct bigint 🔒	sum bigint 🔒
1	Los Angeles	493	21960651

We can see that in LA county, the facility encounters lead the ED burden with 22M. We can see that the burden refers to the ratio of grouped encounters to the available ED stations.

The larger the ED burden, that means thare are more ED visits per available treatment station, easily could surpass the capacity at any given temporal time frame. These grouped observations and absence of time limited time series analytical observations that can be made. However, the measure of the burden does adequate to understanding trends that can apply to real-world decision making. This makes sense since modern LA is so heavily populated, it makes sense for it to carry a high emergency department burden.

Another limitation to this dataset is how scrambled the layout of the data collection is. Since each observation is unspecified, I am unable to devise accurate windows to be able to the data in a way allowing me to confidently extrapolate new trends.