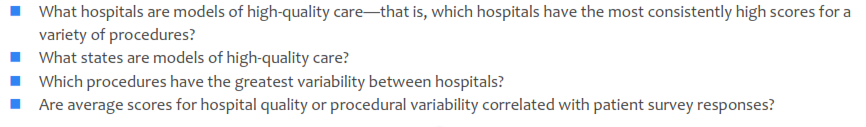
I addressed these questions with a very simple methodology because I did not have enough familiarity with the metrics in the dataset to do more sophisticated approaches (e.g. join Readmission Death with Timely and Effective care metrics in a sensible way).

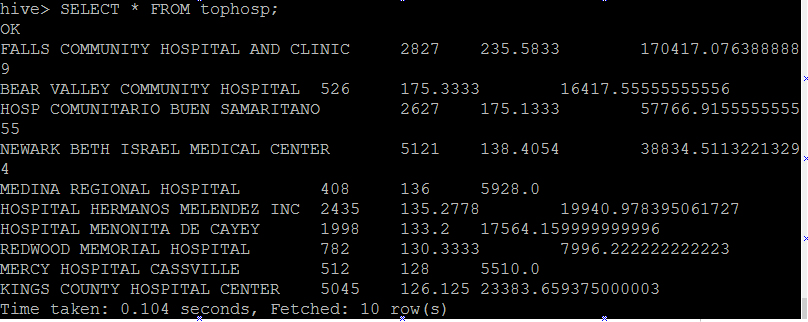
In each computation, I took the highest value for the dimension of interest (e.g. average or variance), and took the top 10 after sorting them in descending order. The dataset had many outlier issues such as entries without data, hospitals with few samples, and so on—this makes me uncertain in my conclusions because of the number of subjective filters and transforms applied to accommodate the datasets’ imperfections.

~~Lastly, I didn’t successfully compute the correlation between quality and survey responses because Hive was corrupt on my runs and data loaded to tables would disappear / be altered between queries. I have not been able to debug as to why, but it wasted a lot of my damn time and I can’t find any documentation on it anywhere. I wrote the query to compute correlation and easily conclude whether a correlation exists or not.~~

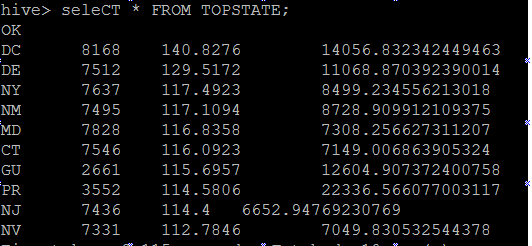
I finally was able to get my query to run for the correlation between hospital quality and survey responses using the Pearson correlation. The value was about -0.24 which corresponds to a weak negative correlation to survey responses for care quality. This is a bit surprising because I would expect a hospital with a better score to solicit better survey responses. Determining the reason behind this would be interesting but would require understanding the datasets in much greater detail. This could be influence by many issues such as survey bias at good hospitals because patients are used to high quality care and have a higher bar than poor hospitals where patients are used to lower quality care, which is simply one possible hypothesis.



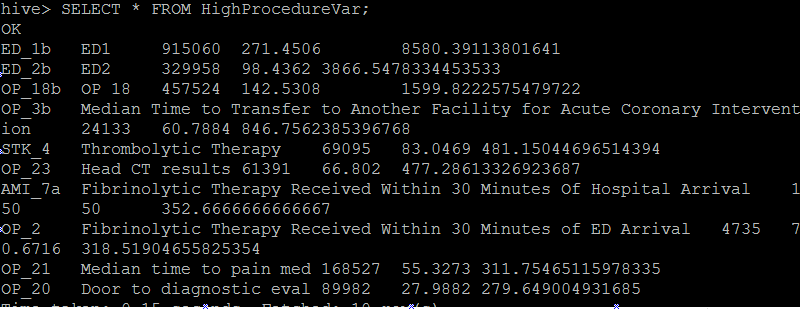
TOP HOSPITAL (NAME, SUM, AVG, VAR): choose highest average



TOP STATE (STATE, SUM, AVG, VAR): choose highest average



HIGHEST VARIATION IN PROCEDURE QUALITY (MEASURE NAME, MEASURE ID, SUM, AVG, VAR): choose highest variance



CORRELATION between quality and survey

