Version 2.0

Date: December 14th, 2018



DWBI Project journal

Presented By: Akshat Bhatia

Group 4

NUID: 001856253

INFO7290 DWBI

Northeastern University

This journal is written by Akshat Bhatia about the Data Warehousing and Business Intelligence project. This project so far has been a great learning experience for me. Data warehousing is not just warehousing but more than that. We, as a team were given multiple datasets from different websites about healthcare, postal codes, income areas, physicians and their field of practice. Everyone in the team had different approach towards the datasets and that’s what the project is all about. But here I am going to describe what my approach towards the datasets was initially when I saw them and what inferences I made earlier and how I look at the data sets now.

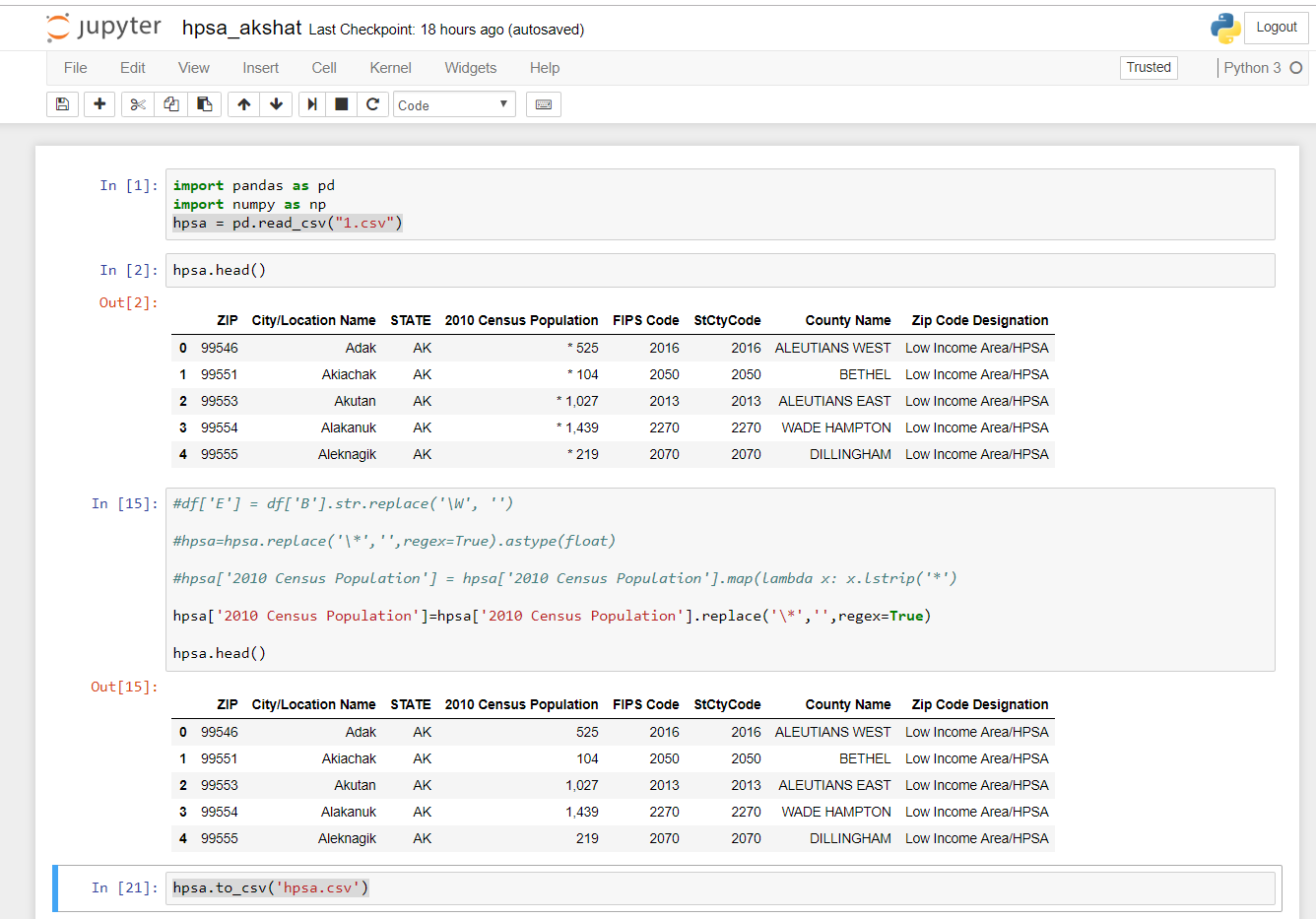
These datasets have many pivotal columns like ZIP code, State FIPS, County FIPS, Area population, Area categorization based on income zone and number of healthcare professionals available, Physician details, income details of residents, cost of getting treatments, tax information, tax return information, etc.

My primary aim was to find out the primary keys in all the tables so that when required in the future, this data can be added up to make sense and answer the project queries. The files were huge, and the data was not clean. We, as a team sat in meeting rooms, through different weeks to talk about how everyone pursues the datasets and what should be the right approach to integrate and clean the datasets.

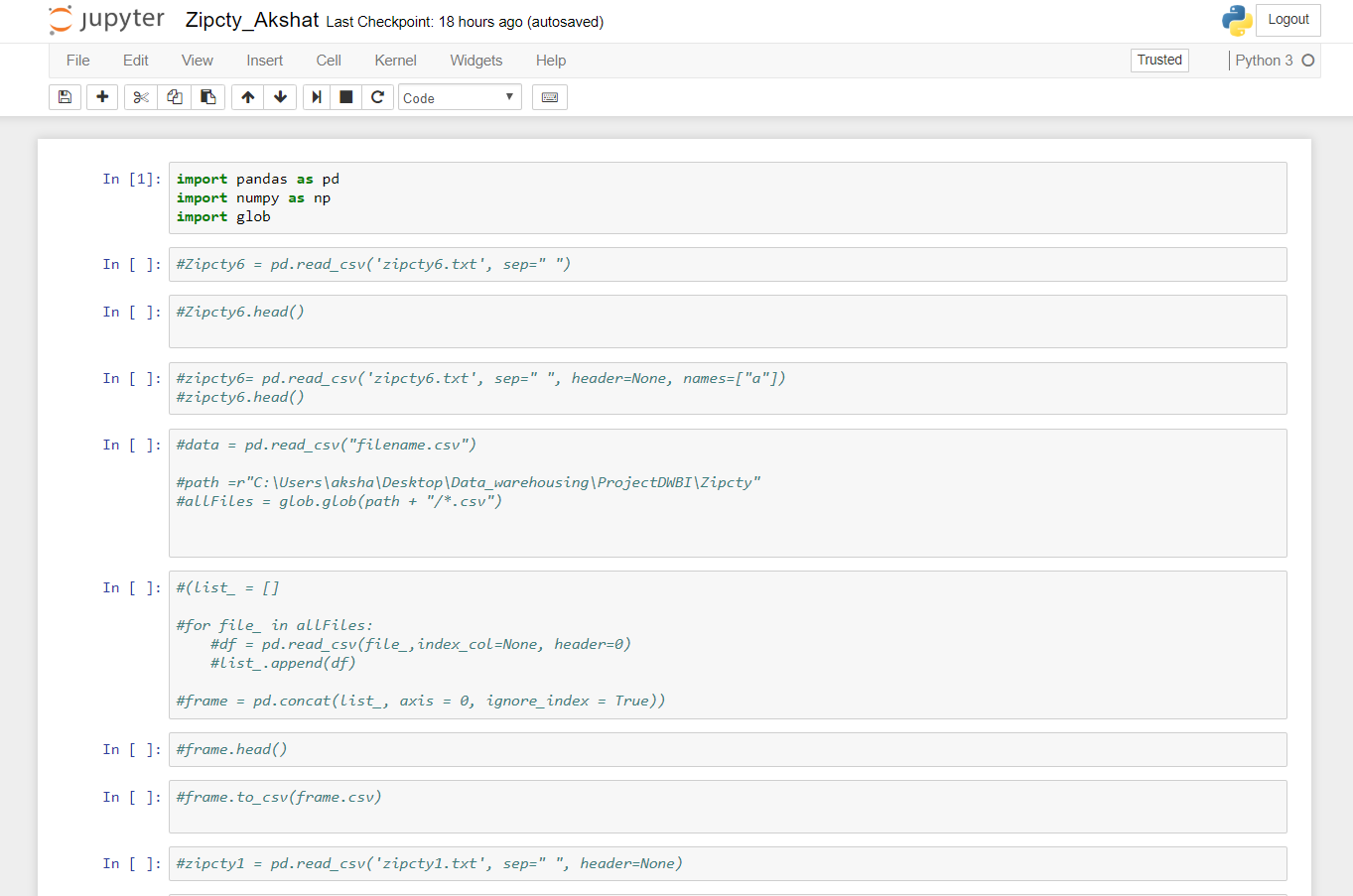
I as an individual, am good with visualizations using python and tableau. Even though we were taught to use different integration and cleaning tools in the class, I prefer using the skill sets which interested me. We had many files, like CMS HPSA low income zip code database, and USPS ZIPCITY datasets which needed cleaning and integration. I used python tool to clean the files and put them in a way that they made sense.

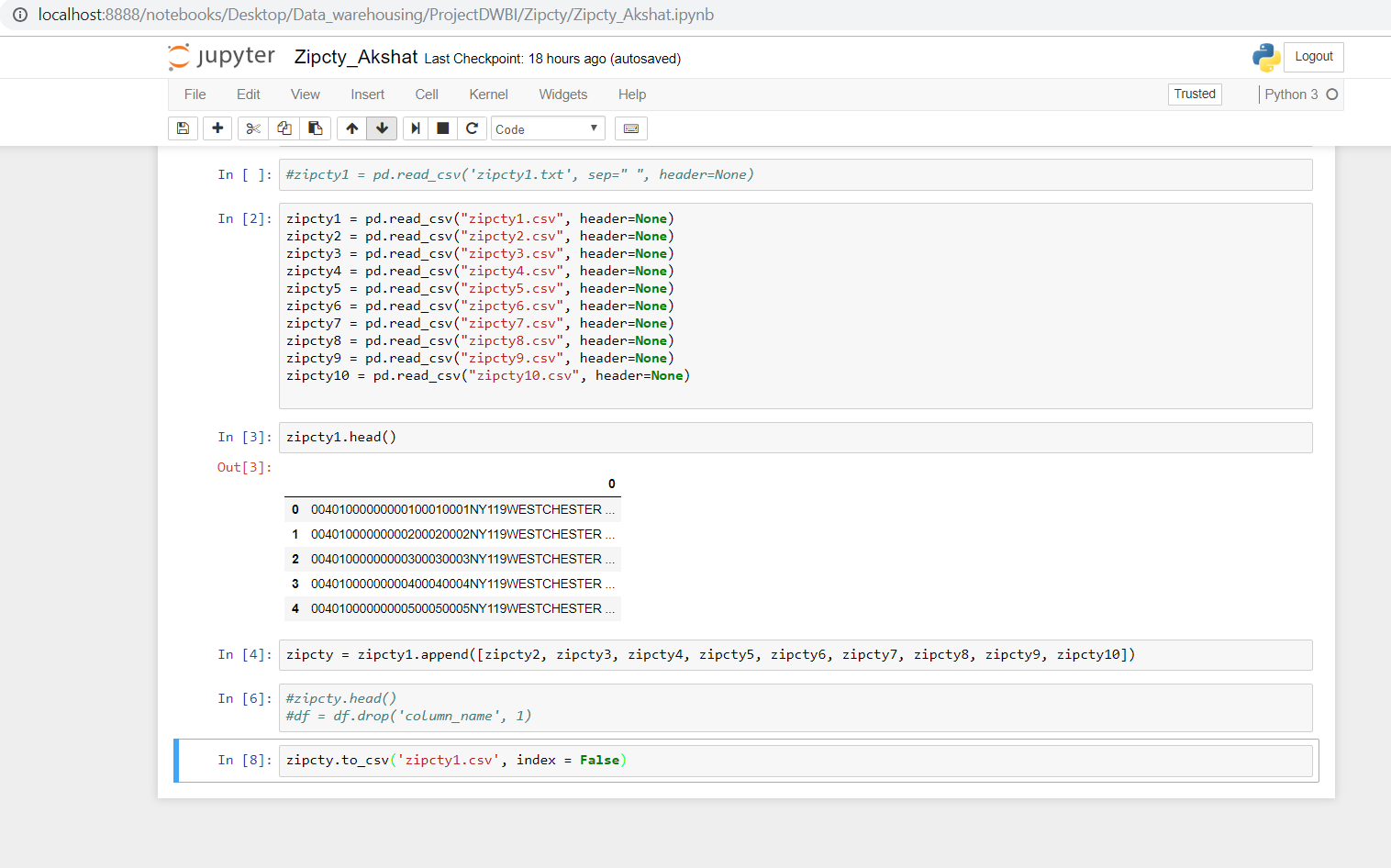
I would like to add some screenshots of my python codes which I used to clean my files.

1. Cleaning the CMS HPSA file:



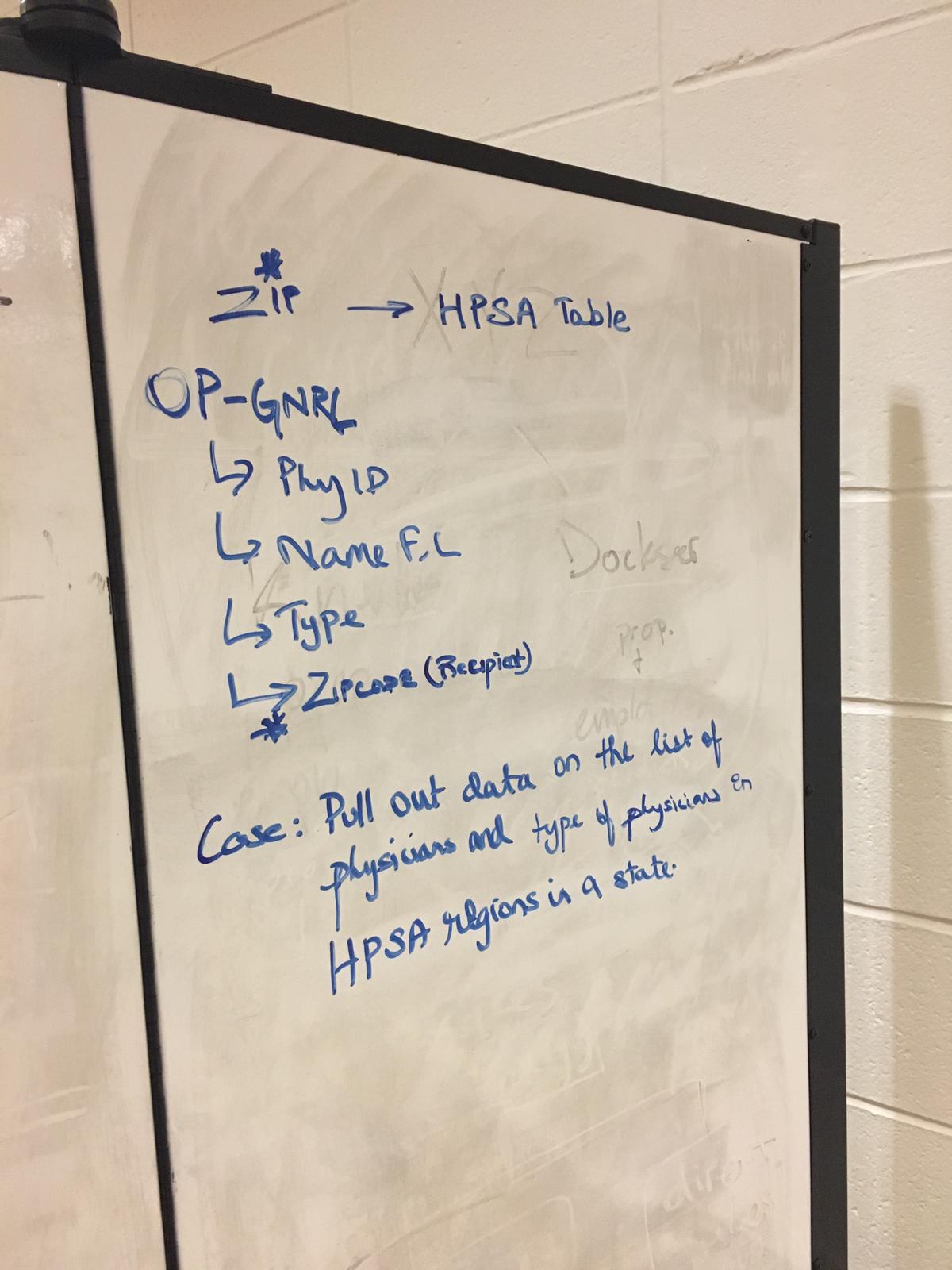
1. Cleaning the USPS ZIPCITY file and appending approximately **25 Million rows**

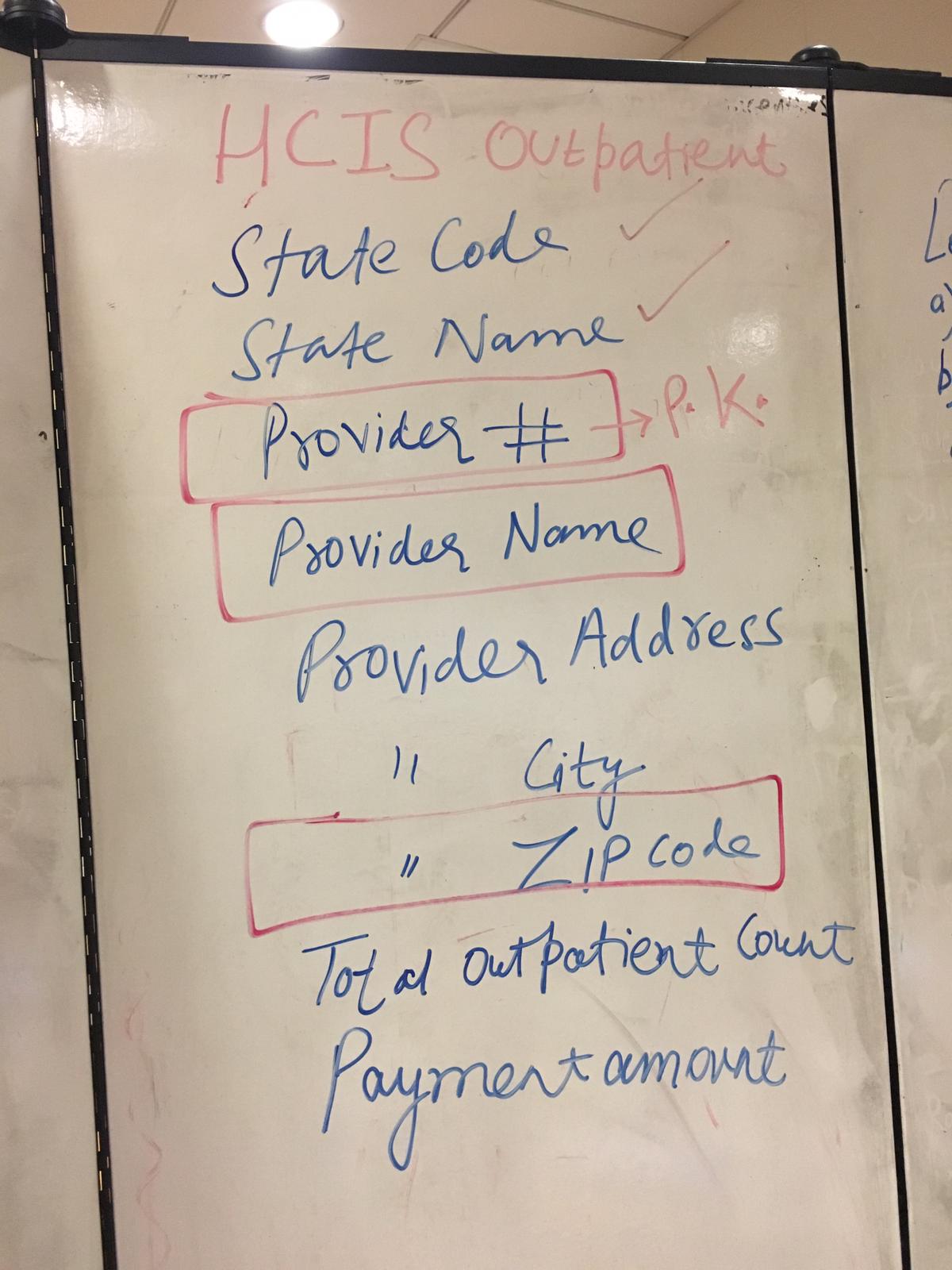


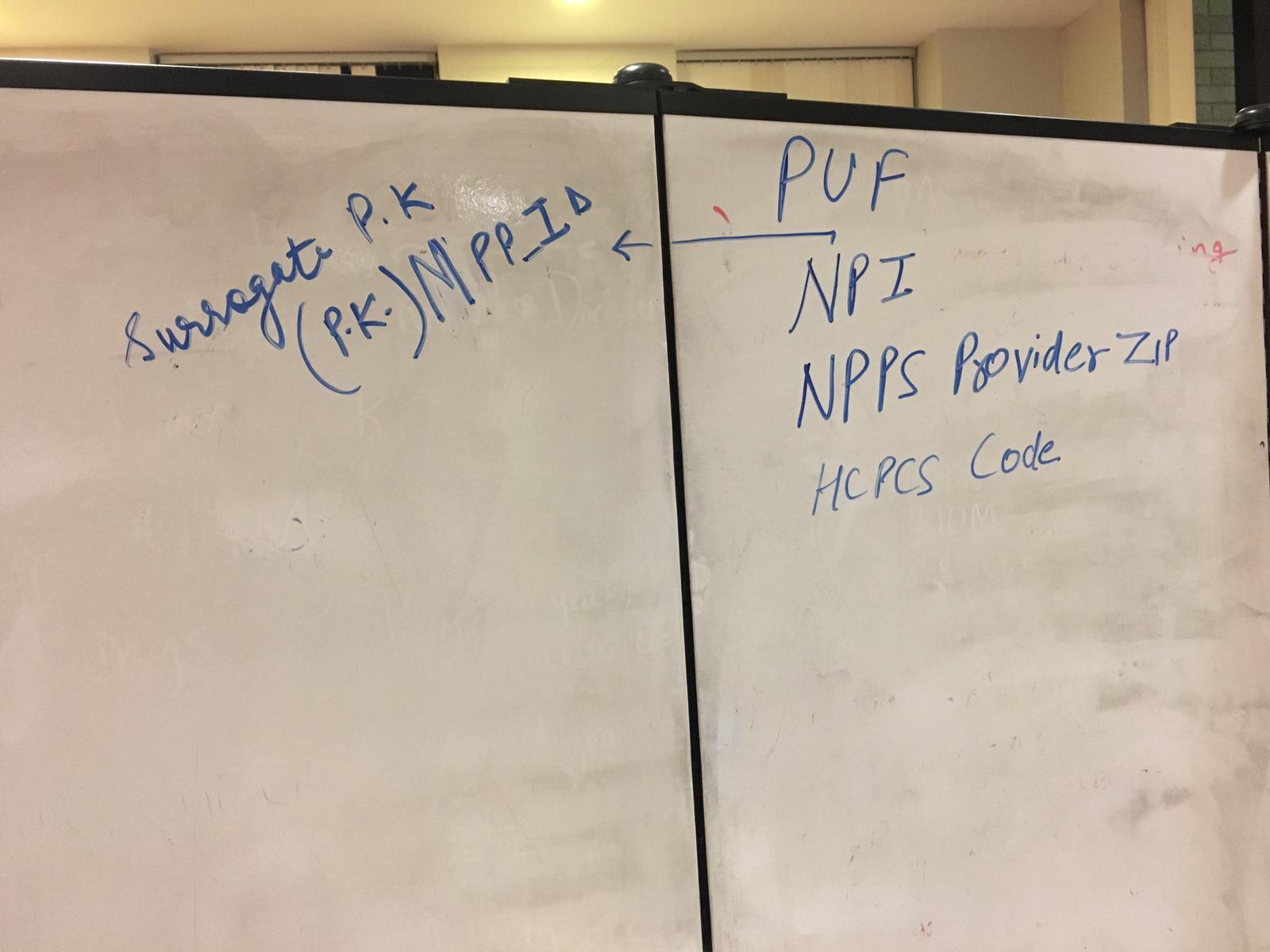


Once I was done cleaning files on my part using python then we all set as a group and figured out many ways to use up all the datasets and made plans using whiteboard in the library.

The meetings were planned with everyone in harmony. As I am working on a co-op, the meeting timings were usually in the evenings and a few times during the weekend as well. We spent approximately 2-3 hours per sitting working on this project so far.





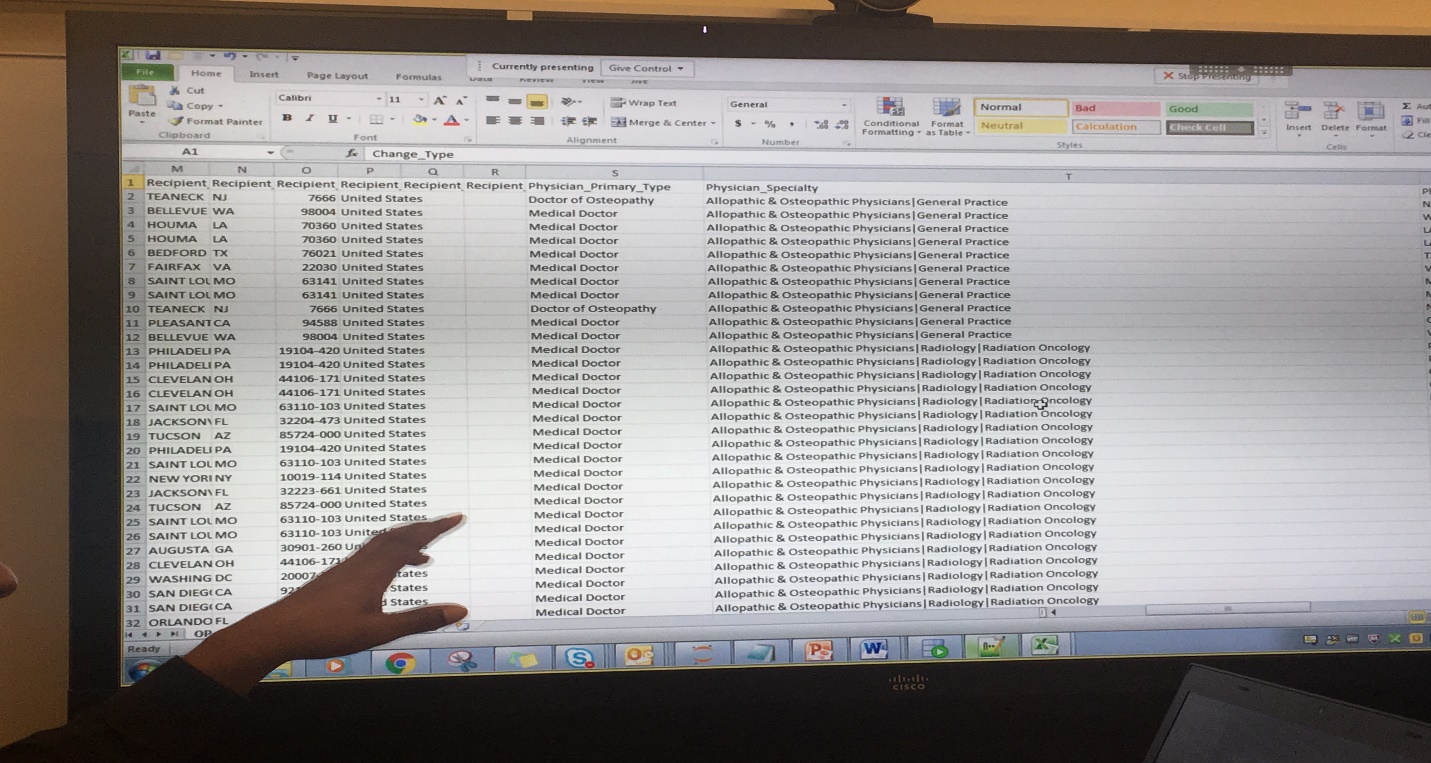


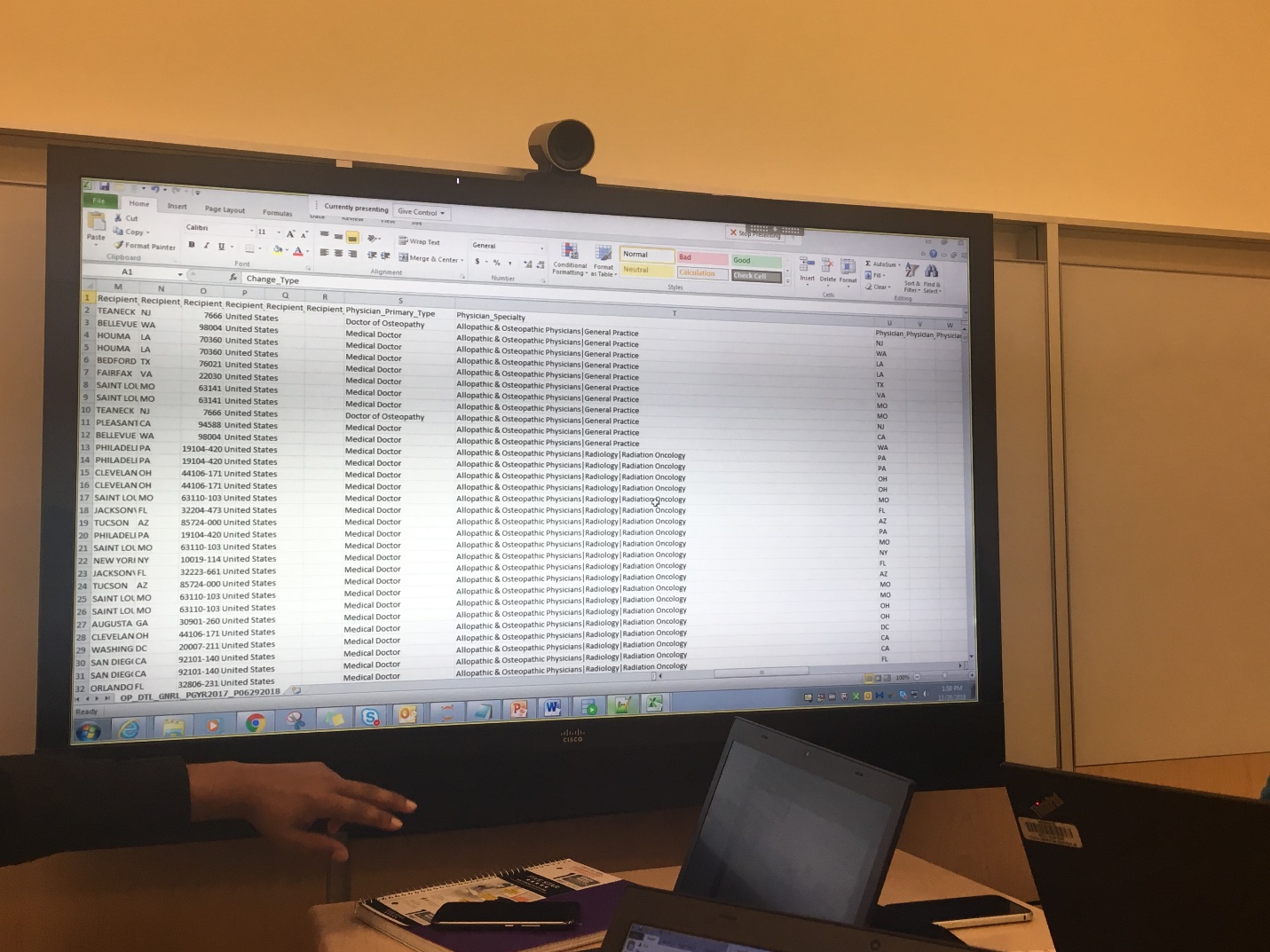
We, as a team used a different machine to make a warehouse where we uploaded the files in the staging table. All the files were uploaded in the staging table successfully. We met a couple of times in the library and utilized the resources like TV display to analyze the datasets, white boards to make tables and to identify primary keys and to understand data as well.

**ACTION PLAN:**

Now I plan to work on this dataset as an individual by bringing up more visualization using Tableau. I bring multiple skillsets to the project team table and so do my teammates. It has been a great project experience with the team so far.

I have made a few execution plans to take the project ahead and have discussed these plans with my team as well. From the coming week, we are going to work on our summary tables. I will be using Tableau dashboard to make the summary table. By using the Tableau dashboard, I can identify dimensions and measures in the datasets that we have been assigned. I am more interested in fetching the information by joining geographical pin codes and the healthcare situation of those areas and how these people are a part of the system in paying taxes and applying for returns. I think I have gain more knowledge by working on this project and my lookout towards keys have changed a lot. It’s about creating a great data warehouse by creating the lookup tables and then joining the primary keys as per our requirement to get the answers to the questions we want from the warehouse.





This is what we, as a team think will help us in understanding the links better among the different datasets.

Approach here is to get all the files with zip Codes so that all other props can be used in relationship with the geography.

The columns used here are the columns which we think should make sense in the near future as the project progresses and give us the answers we want from them.

1. Zip fips: Required Columns

* Zip code
* State abbreviation
* County name
* Zip low sector number
* Zip low segment number

1. Zip\_all\_no\_agi

* State fips
* State
* Zip code
* A02650- Total income amount
* A00200 salaries and wages amount
* A00700 State and local income tax refunds amount
* A02300 Unemployment compensation amount
* Elderly -Number of elderly returns
* A11901- Tax due at time of filing amount

1. All\_agi

* State fips
* State
* Zip code
* A02650- Total income amount
* A00200 salaries and wages amount
* A00700 State and local income tax refunds amount
* A02300 Unemployment compensation amount
* Elderly -Number of elderly returns
* A11901- Tax due at time of filing amount
* Student loan interest deduction

1. US\_Zip\_ref:

* Zip
* Zcta
* Population

1. HPSA

* ZIP
* Zip code designation
* State
* City name

1. Open Payments general payments

* Physician profile ID
* Physician specialty
* Physician first Name
* Physician Last Name
* Physician primary type
* Physician Specialty
* Physician address

1. PUF

* NPI
* Provider type
* HCPCS code
* HCPCS description
* NPPES first name
* NPPES last name

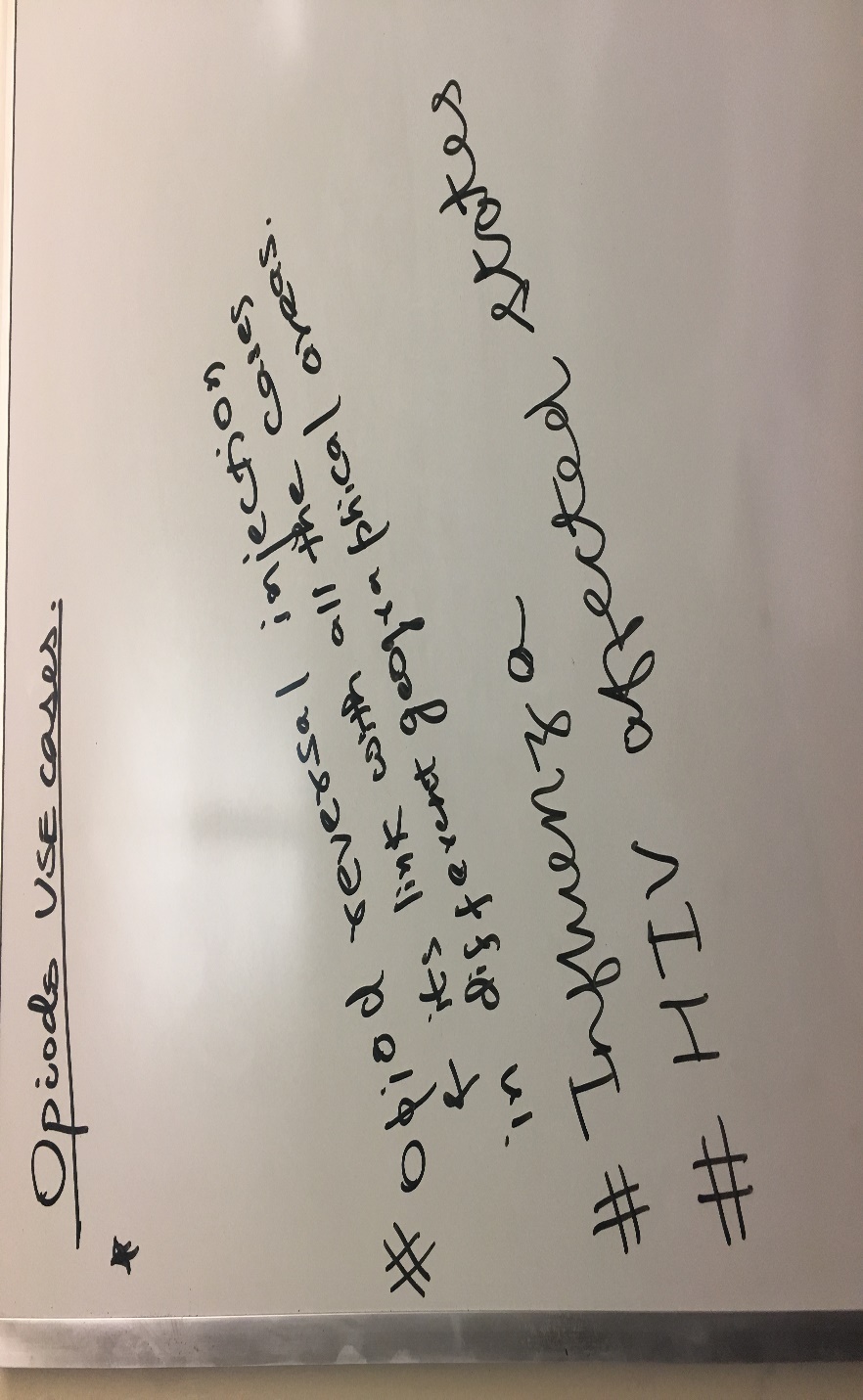
This is my journal of looking at the project datasets halfway through the project presentation finals. I will continue to update this journal as we progress with the project and develop more understanding.

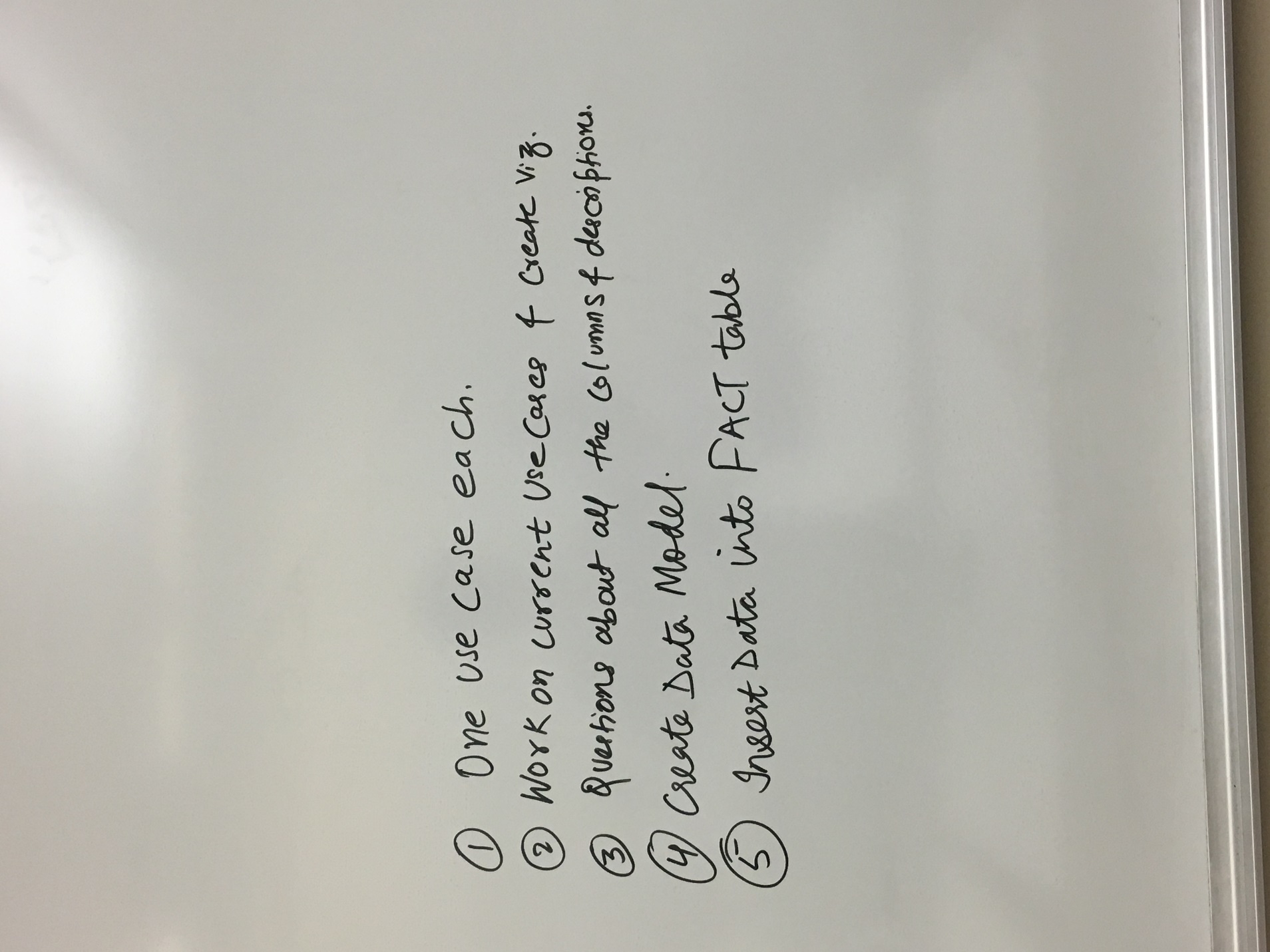
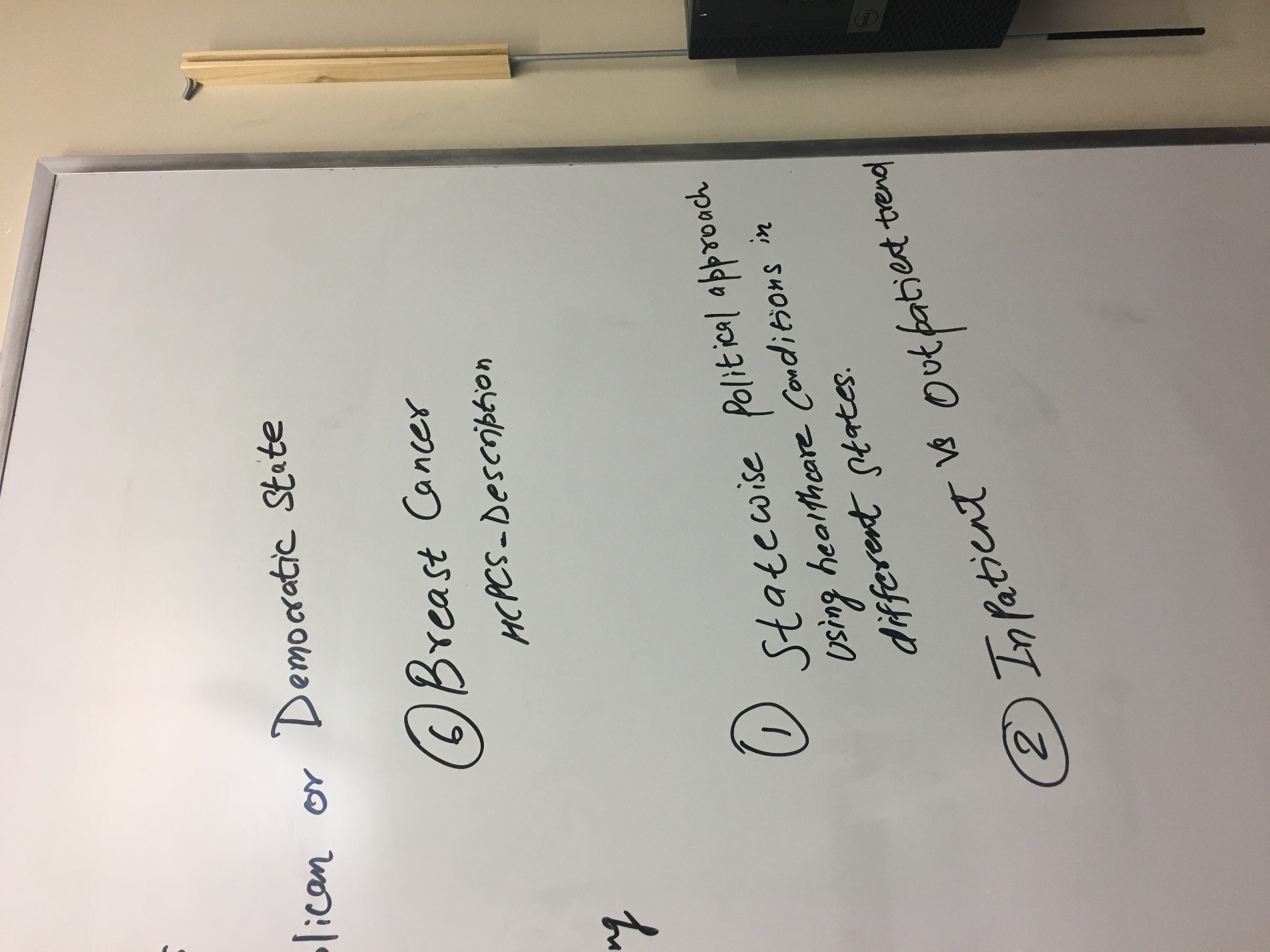
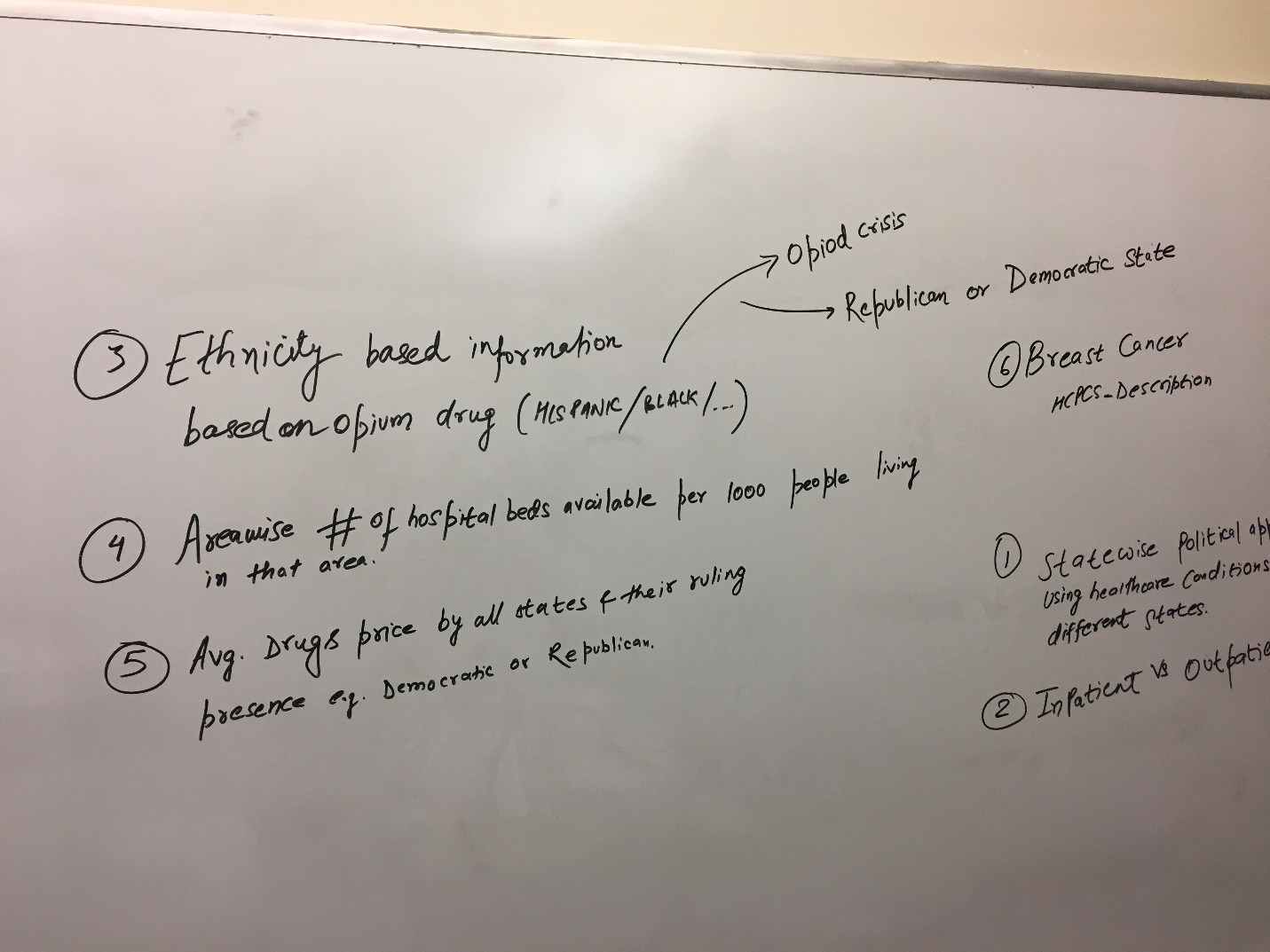
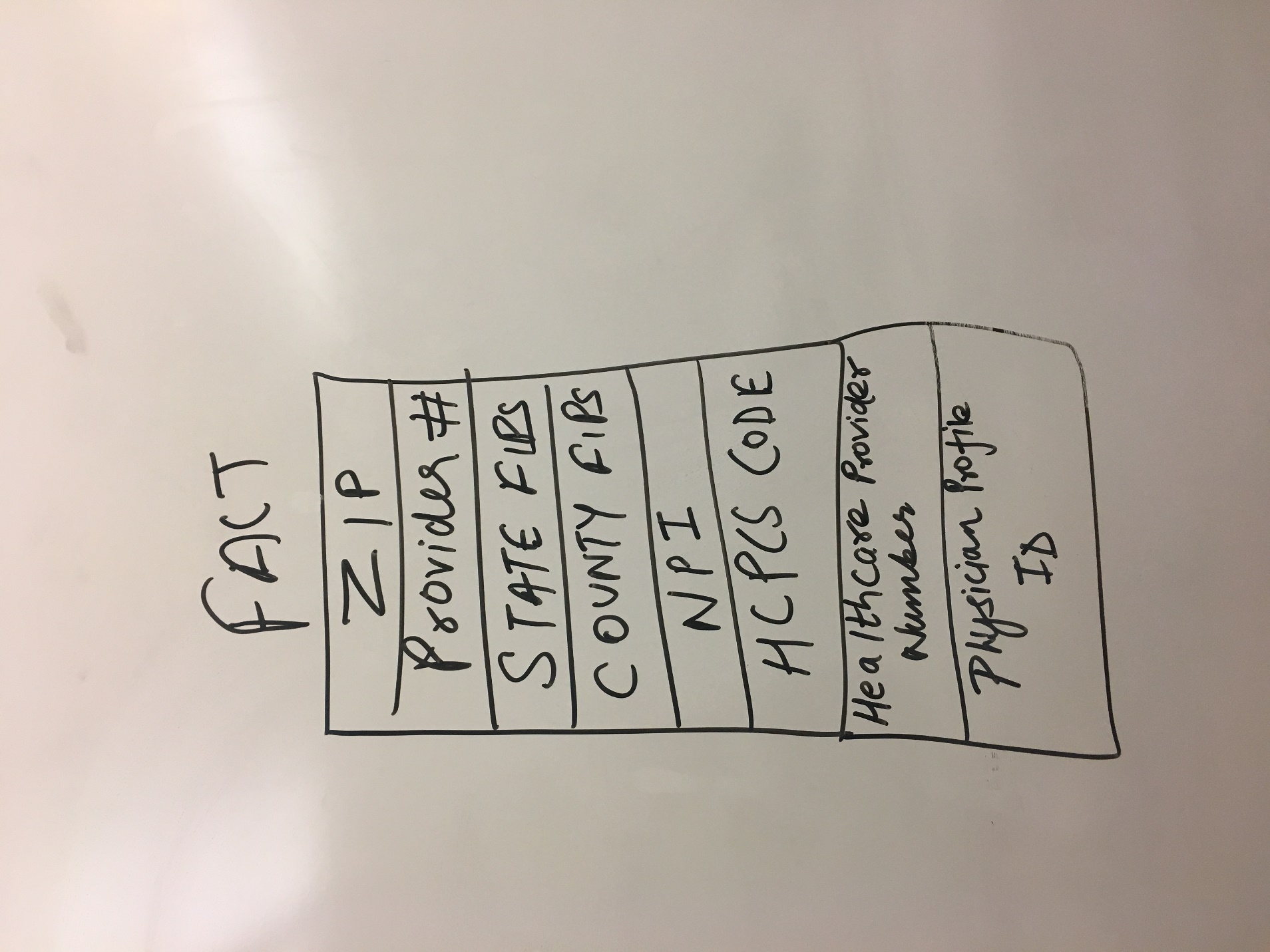
Part 2: Part 2 brings in more than part 1. Now We were thinking about the visualizations and our use cases implementations. Progressing further, all of us sat in the law school and utilized white board to write punch list and came up with score cards.

We all took our individual parts of data visualizations and started working on them.

We met 4-5 times and every time we met, we discussed our project and graded individual performances. By doing this we came to know each other’s importance in the whole project.

I will attach images from our meetings for the project collaboration:

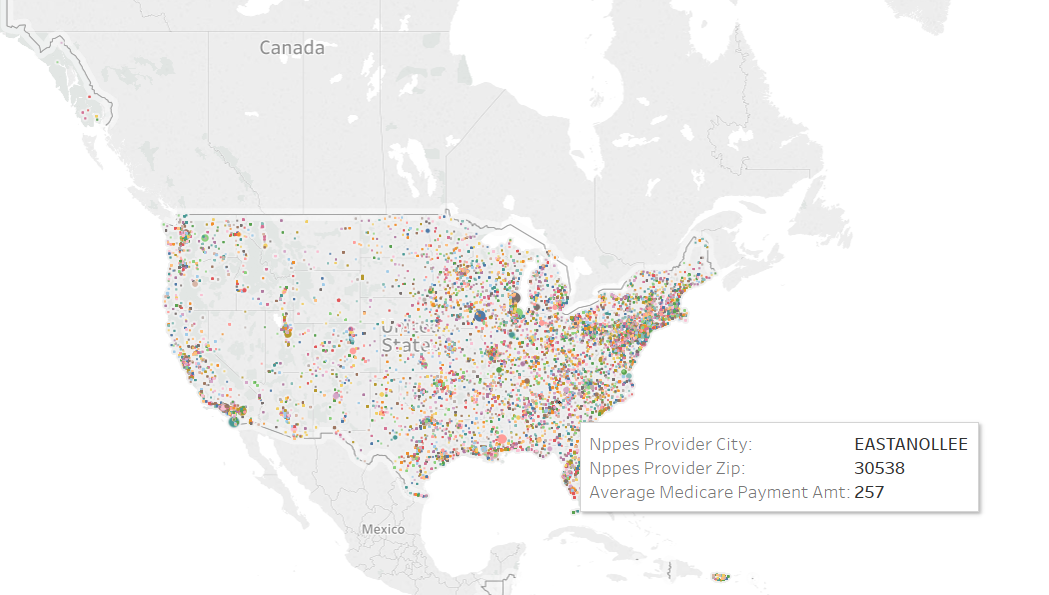


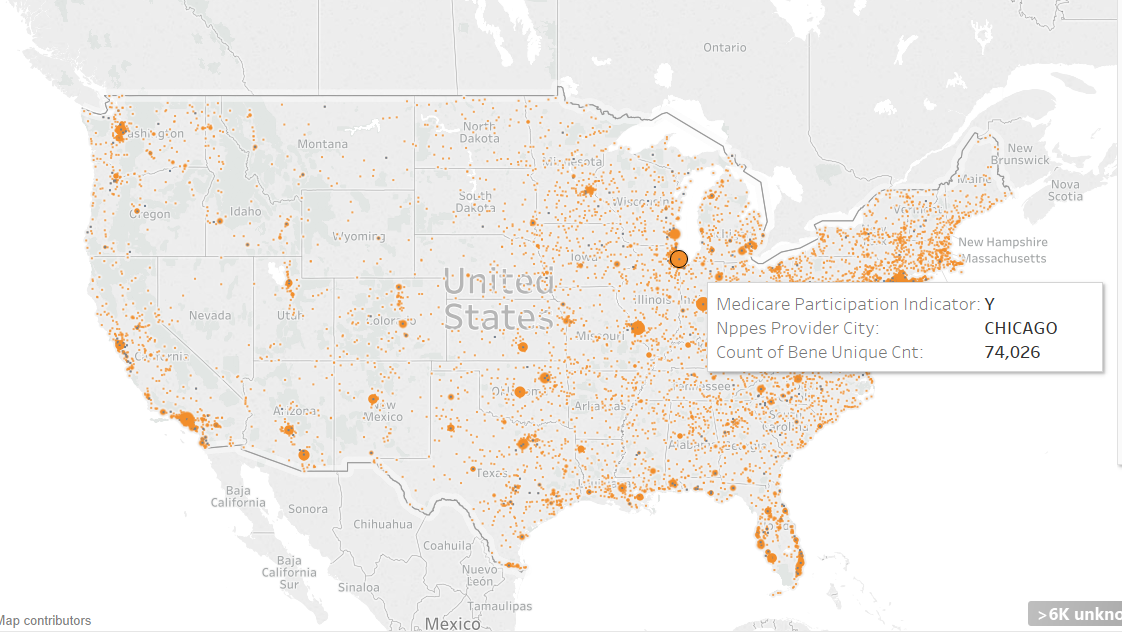


Vizualization:

Visualizations created:

The first visualization created here talks about the Medicare participation indicator. MPI is nothing but an indicator which talks about the organizations participating in medicare programs. **Medicare** is the federal health insurance **program** for People who are 65 or older. Certain younger people with disabilities. People with End-Stage Renal Disease (permanent kidney failure requiring dialysis or a transplant, sometimes called ESRD). In the areas where the Medicare Participation Indicator is TRUE, the cost of Medicare reduces drastically. This visualization shows the Average Medicare Payment amount in Dollars based on the geographical constraints like ZIP code and city name.



This visualization tells us about the medicare participation indicator and the number of unique beneficiaries who utilized the Medicare program based on different geographical locations.

I also did a pivot table analysis for the location wise Medicare cost throughout the US.

