# What Keeps Them Coming Back

Database Management Project Created Based on Bon Secours' Datasets

Course Number: BUAD 5272
Team Number: 12
Team Member: Stephen Blotkamp, Joshua Swerdlow, Lydia Zhao, Miranda Zhao

**PART 1. INTRODUCTION**

The datasets we are using are Medicare Shared Savings Program (MSSP) claims data for the Hampton Roads area from year 2016 to 2017. This claims data offers a comprehensive view of patient encounters for Medicare patients.

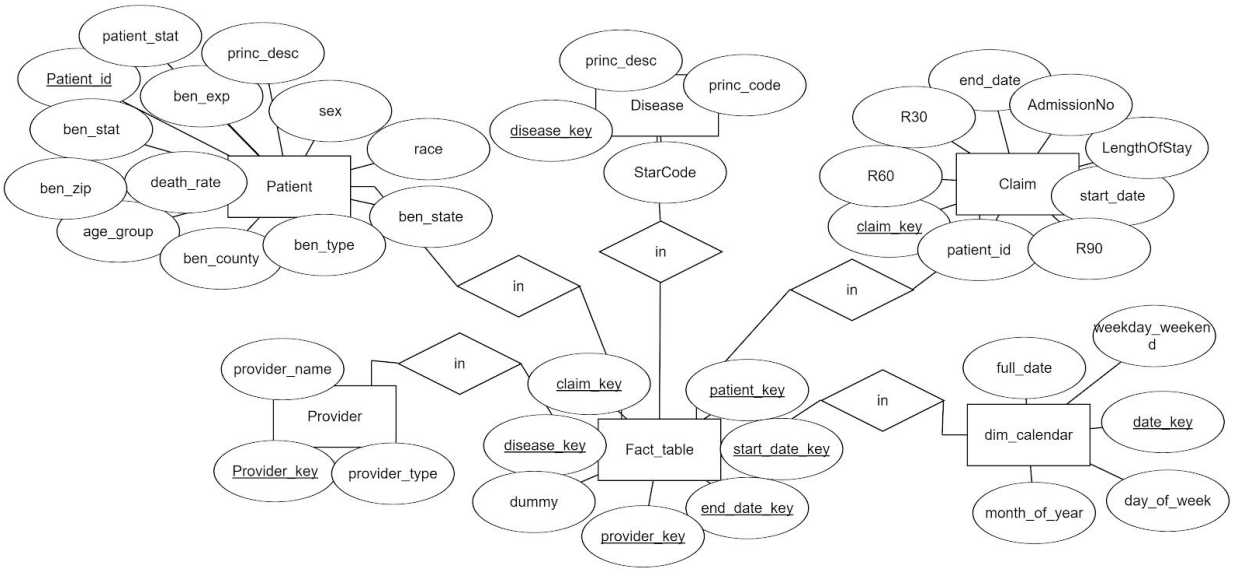
Our analysis based on this dataset focuses on two major questions:

1) Does the type of disease affect the readmission rate of the patients?
2) Does age/gender/patient status affect the admission rate for both 30-day period, 60-day period and 90-day period?

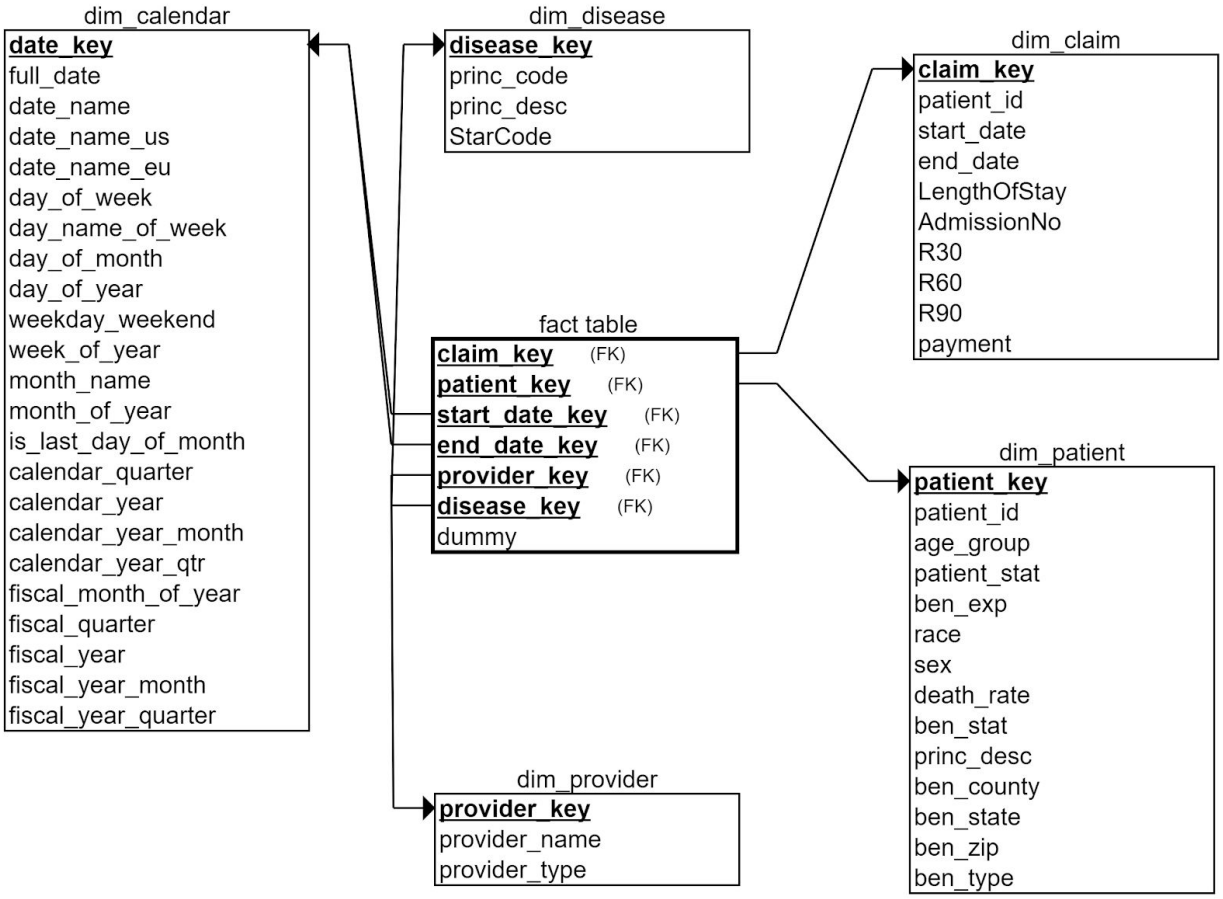
For this project, firstly, we start with modelling using ERD Plus; then, we employed Alteryx and MySQL workbench for our ETL process. Last, we mainly used Tableau to visualize and analyze the data. Based on our interpretations of data, we conclude that...

**PART 2. MODELING APPROACH(Justification of your modeling approach. Include your ER and RS in this section.)**

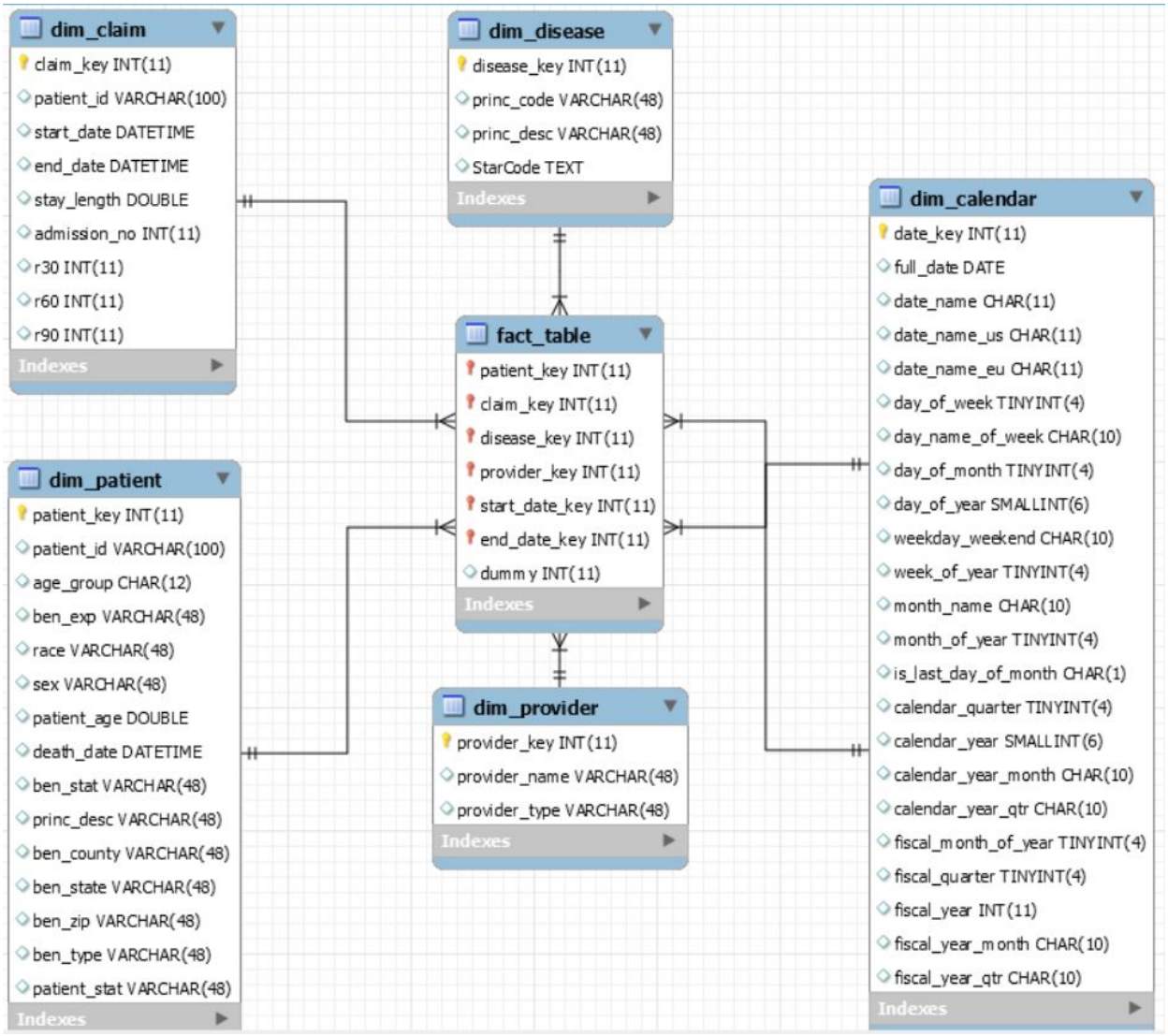
The graph below is the ER diagram that reflects our approach of how to categorize the data so that we can use them in a more organized way.




ER Diagram


dim_calendar
date_key
full_date
date_name
date_name_us
date_name_eu
day_of_week
day_name_of_week
day_of_month
day_of_year
weekday_weekend
week_of_year
month_name
month_of_year
is_last_day_of_month
calendar_quarter
calendar_year
calendar_year_month
calendar_year_qtr
fiscal_month_of_year
fiscal_quarter
fiscal_year
fiscal_year_month
fiscal_year_quarter
dim_disease
disease_key
princ_code
princ_desc
StarCode
dim_claim
claim_key
patient_id
start_date
end_date
LengthOfStay
AdmissionNo
R30
R60
R90
payment
fact table
claim_key (FK)
patient_key (FK)
start_date_key (FK)
end_date_key (FK)
provider_key (FK)
disease_key (FK)
dummy
dim_patient
patient_key
patient_id
age_group
patient_stat
ben_exp
race
sex
death_rate
ben_stat
princ_desc
ben_county
ben_state
ben_zip
ben_type
dim_provider
provider_key
provider_name
provider_type

Star Schema




Justification of our model selection approach

Our group chose dimensional modeling instead of third normal form tables as our modeling approach because we don't have a large number of dimensions to connect. We have also create a factless table from which we match keys with primary keys from other dimensions to fetch the information we want easily.

## PART 3. ETL APPROACH

### Datasource

The main dataset that we extracted the data from is the high level claims data for the Hampton Roads area. There are two reasons for why we decided to use this dataset. The first reason is that Hampton Roads is a smaller subset of the population than Richmond, but could be

extrapolated and representative of a larger population depending on the results of our study. So using data from only Hampton Roads would improve efficiency while facilitating a relatively comprehensive conclusion at the end. The second reason is that we are using data only from patient who had been admitted into the hospital and the High level claims data provide all information we need.

**Extract**

From our datasource, we fully extracted the data and renamed all the fields we are going to use in our analysis later. Next, the data cleaning step is one of the most important as it ensures the quality of the data in the data warehouse. We first make sure there are no duplicated data in our dataset. Then, we added keys that are missing from the table we create in the ER diagram. Finally, we choose the attributed we want from the whole datasets and make them into their corresponding dimension tables.

**Transform**

At this step, essentially what we did is to transform the data from its original value to our target measurements. For example, since we want to know what is the time interval between the first admission and the next one, we calculated out the time difference using the “date from” and “date to” columns and created three new columns called “R30”, “R60” and “R90”, each representing the time interval during which a particular patient got re-admitted to the hospital.

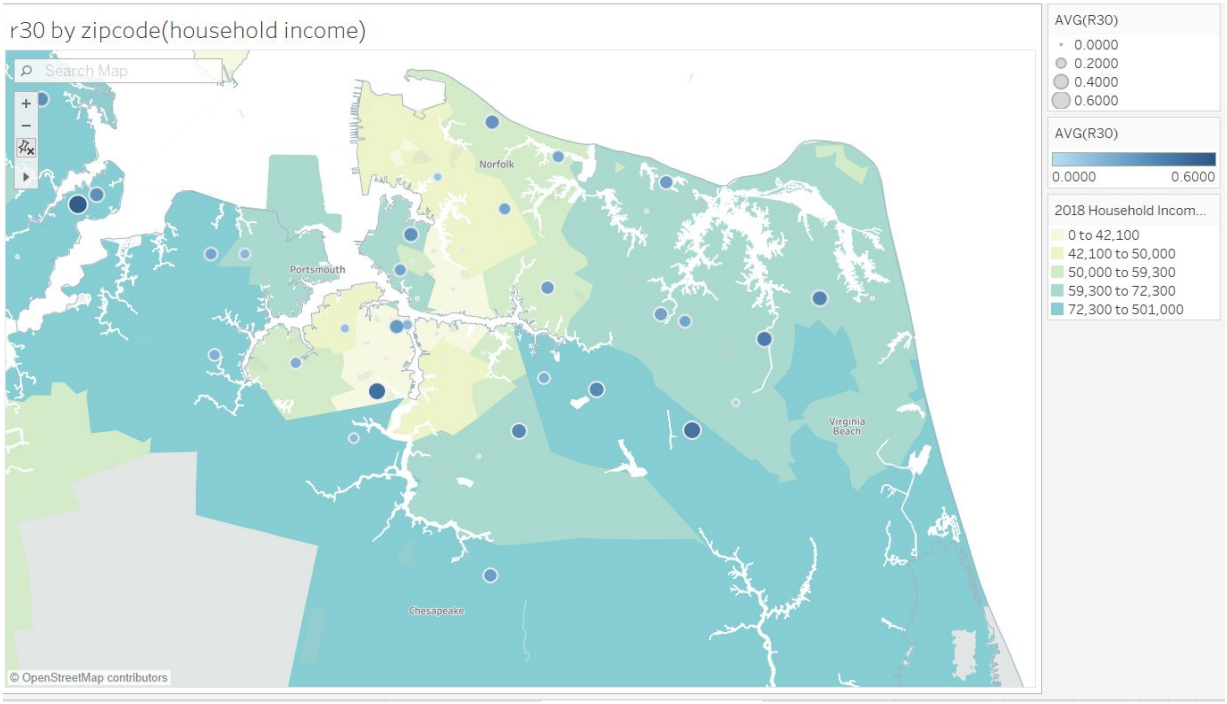
**Load**

Finally, we use MySQL workbench to create a database called “class_project” and populate the database using Alteryx. We also make sure that we added the foreign key constraints to MySQL as well as to Alteryx so that we ensure that all foreign keys in the fact table are present.

## PART 4. QUERIES AND VISUALIZATIONS

Our group replied on Tableau as our major data visualization tool. Here are ten visualizations that more or less, give us some important information about the data.

Query 1:




From the visualization, we can see that the darker the shaded area, the higher the average household income. Additionally, darker points mean higher readmission rate in 30 days. And we can see that most of the darker points fall on the darker shaded area. So our interpretation is that, in richer areas where people have relatively high household income, people have better access to health care services; also, as most older people may retire at better-developed area, it's possible to see a concentration of retired people from this area.

Query 2:

r30 by patienttype

| ben_type | Avg. R30 | Avg. R60 |
|---|---|---|
| Widower | 0.6667 | 0.0000 |
| Surviving Divorced Wife | 0.3333 | 0.0833 |
| Widow | 0.3158 | 0.1128 |
| Divorced Wife | 0.2500 | 0.0000 |
| Retired Worker | 0.2402 | 0.0738 |
| Wife Of Retired Worker | 0.2245 | 0.0816 |
| Husband Of Retired Worker | 0.0000 | 0.0000 |

For people in different beneficiary types, we find that the people who lives alone tend to come back to hospital very often within 30 days. The reason is that people who live alone do not have people to take care of them, so they tend to seek help from hospital like widower/widow/ and divorced people.  However for retired workers, they usually have time and their partners to take care of themselves, so that they do not need to be readmitted as often as people living alone.

Query 3:

13. age - admission no & read days & R

| Age Group | Avg. Admission No | Avg. Read Days | Avg. R30 | Avg. R60 | Avg. R90 |
|---|---|---|---|---|---|
| 30-45 | 0.80 | 39.00 | 0.40 | 0.00 | 0.00 |
| 46-65 | 1.93 | 48.46 | 0.27 | 0.07 | 0.06 |
| 66-80 | 2.34 | 54.53 | 0.25 | 0.08 | 0.05 |
| 80+ | 1.66 | 54.10 | 0.21 | 0.07 | 0.07 |

People in age group of 30-45 have lowest average number of admissions. This is because the younger population has a better recovery mode.

People in age group of 46-80 tend to have a higher average number of admissions because the older they get, the more serious diseases they could get. They have a higher need to be readmitted to get recovered.

Query 4:

10. disease - admission no & read days & R

| ReformD.. | Avg. R30 | Avg. R60 | Avg. R90 |
|---|---|---|---|
| MIA | 0.3462 | 0.0385 | 0.0385 |
| HF | 0.3092 | 0.0672 | 0.0857 |
| COPD | 0.2526 | 0.1129 | 0.0554 |
| CABG | 0.1940 | 0.0662 | 0.0300 |
| Stroke | 0.1282 | 0.0577 | 0.0449 |

The readmitted rate is different depending on the type of disease a person was diagnosed with.

For AMI(Acute myocardial infarction ) and HF(Heart Failure), people who get these diseases tend to have a higher average R30 rate.

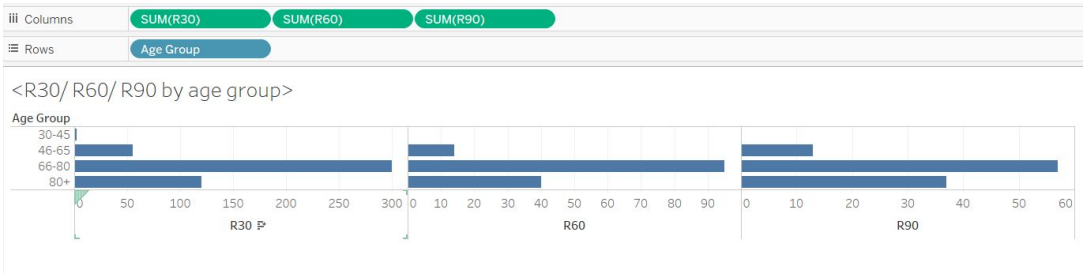
The R 30 for AMI is 0.34 and for HF is 0.30. Since we use binary variable to define R30, R 60, and R 90 variables, the number indicates the percentage of people admitted.
As an example, 34% of people who get AMI are readmitted within 30 days. This can have two interpretations, one is that everyone with AMI is more likely to be readmitted. However, this needs to be taking with a grain of salt as it could be possible there were some few outliers of intense cases that swayed the average as a whole.

We can get the conclusion that patients who get these diseases need higher priority.

The diseases with lower average could therefore deal with somewhat less emphasis. The exception to this would be if these diseases had a higher death rate, which would be an

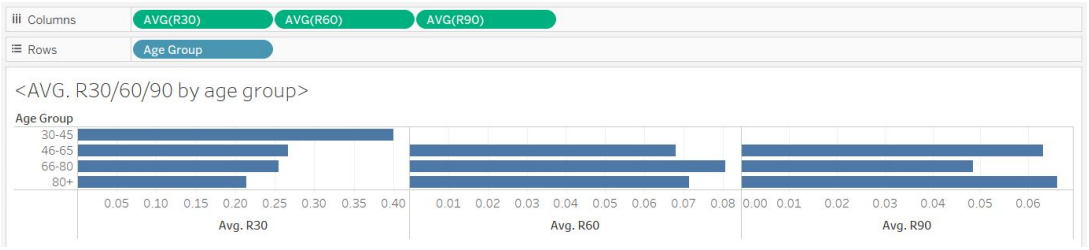
alternate hypothesis for why their readmit rates are low.  This would be an interesting question to ask in a future project, but we did not construct a binary variable to track whether or not a patient died in our model.

Query 5:



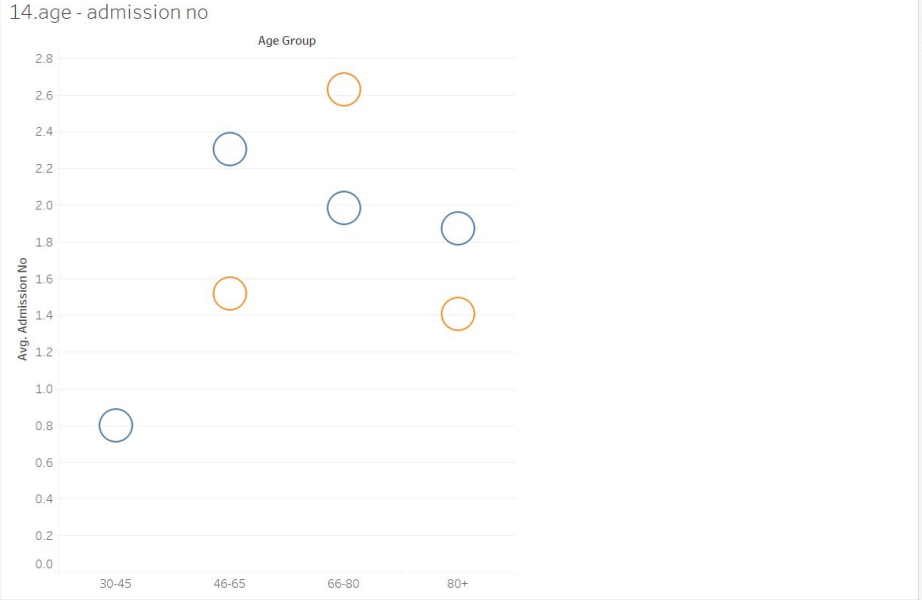

People in the 60-80 age range account for the majority of readmissions.

Query 6:




But their average rate of readmission is fairly similar to the age groups above and below them. This is perhaps indicative of the fact that the 66-80 age range simply accounts for more bodies in the system, but are not inherently different.

Query 7:




We can find women tend to have a lower admission numbers as age increases after 46. However men have the highest admission rate in age group 66-80.

Query 8:

2. disease- age(group by sex&race)

| | | ReformDiagno | | | | |
|---|---|---|---|---|---|---|
| Sex | Race | CABG | COPD | HF | MIA | Stroke |
| Female | Asian | | | | | 78.75 |
| | Black | 75.04 | 70.47 | 74.76 | 77.63 | 74.88 |
| | Other | 73.40 | 70.00 | | 75.00 | |
| | White | 77.43 | 74.18 | 79.06 | 78.43 | 78.43 |
| Male | Asian | 92.00 | | 71.00 | | |
| | Black | 76.27 | 69.67 | 72.28 | 69.11 | 79.40 |
| | Hispanic | 57.00 | | 78.00 | | |
| | Other | 71.29 | | 67.00 | 67.00 | |
| | Unknown | 70.00 | 70.00 | | | 68.00 |
| | White | 76.92 | 75.19 | 75.84 | 73.21 | 77.74 |

In this graph, we analyze the relationship between disease and age.

We can see that  asian male tend to have a higher average age (92) when they get stroke. Black male tend to have a lower average age (74) when they get stoke.

This difference could related to eating habit, lifestyle habit, and genetic design. Just like Asian people have a higher incentive to have Lactose intolerance, Black people could have this issue in other disease like stroke. Additionally, this could support sociological studies which have indicated that oppressed minorities face higher rates of medical complications.  This issue has come up recently with the pregnancy complications faced by Serena Williams and Beyonce.

This approach can help us understand the disease phenomenon between age and gender.

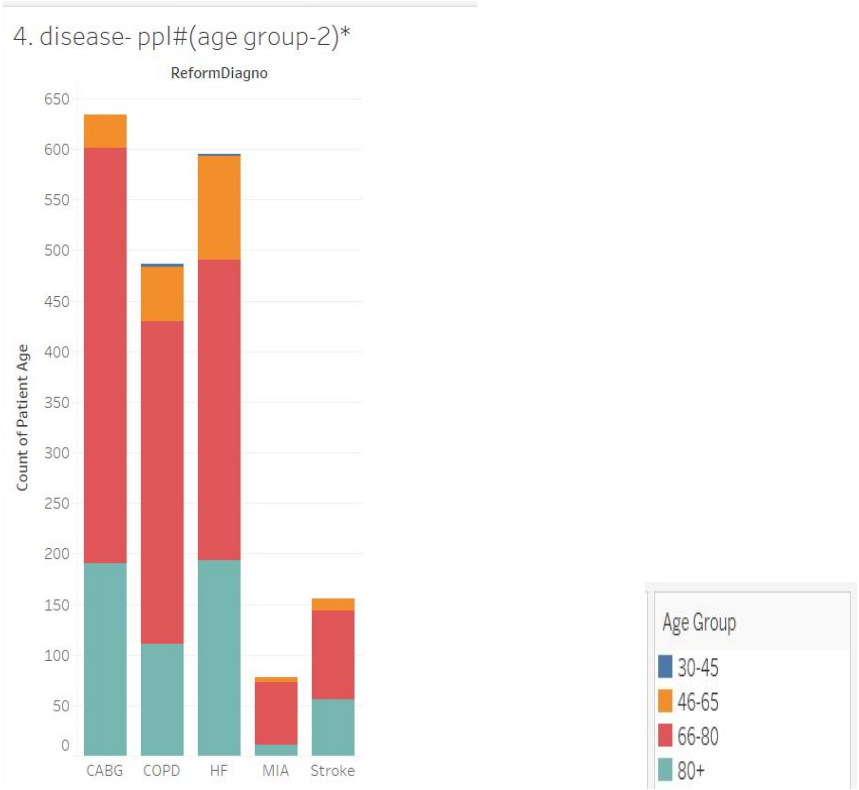
Query 9:

3. disease - ppl #(age group-1)

| | ReformDiagno | | | | |
|---|---|---|---|---|---|
| Age Group | CABG | COPD | HF | MIA | Stroke |
| 30-45 | | 3 | 2 | | |
| 46-65 | 33 | 54 | 102 | 5 | 12 |
| 66-80 | 411 | 319 | 298 | 62 | 88 |
| 80+ | 190 | 111 | 193 | 11 | 56 |

We can know the relationship between disease and the number of people in different groups. For example, we know for CABG, most people get it at age 66-80. This can let people notice that people tend to have this disease after age 65, so people need have a higher attention to monitor their health status and pay attention to visit doctor when they have symptom of this disease.

Query 10:



For people have these 5 diseases. Age group of 66-80 tend to have a higher incentive to get these diseases.  So people need to aware that these diseases have a higher chance to happen after age 65.

**PART 5. Conclusion and Reflection**
(Reflect on the assignment and describe lessons learned, how you might have approached the project differently)

**Lessons Learned:**

**1.Data Cleansing:**
We understand first hand healthcare data could be messy.Understanding and Cleaning Data is not that easy. Data cleansing part could include data renaming and data selection.

- Data Renaming:
  It is not easy to understand the data logic behind the scenery. First, we need to figure out the label name. For example, we need to rename the “CCLF1_PRNCPL_DGNS_CD” to “Principal Diagnosis Code”.
- Data Selection:
  Second, too much data makes the Tableau processing part too slow. In that case, we only select the patientients under the hospital in Claim Bill Facility Type.

**2. Data Processing**:
Take disease code processing as an example. We learned how to find targeted problem, use multiple softwares to process data, and apply decision into Alteryx system.
Take Disease Code Searching Method as an example.

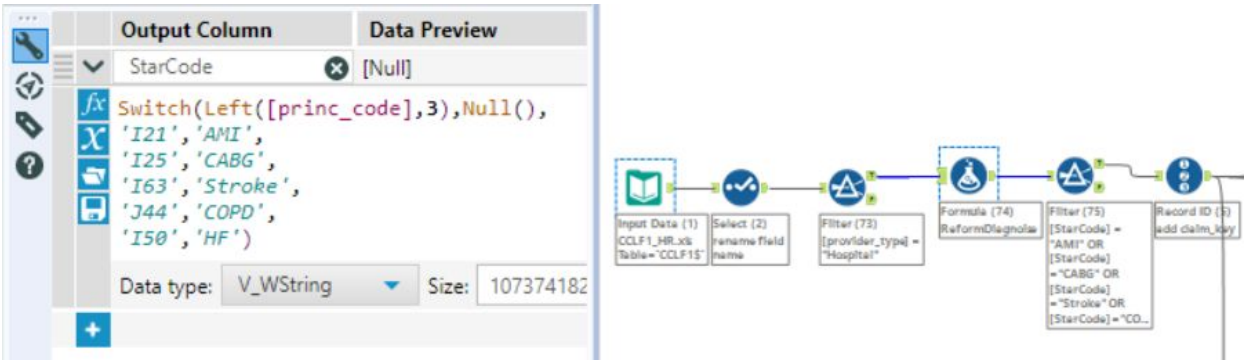
- Targeted Problem: The problem we are seeking to solve is to find out the data related to six specific diseases. At the very beginning, we found that it is hard to find the disease code for each specific disease online. So we choose to use filter in Excel to search the key word of each disease in the following methodology.

| The analysis you will conduct will revolve around hospital readmissions, in particular for the following diagnoses | | |
|---|---|---|
| Requirement | Filter Condition | Filter by font color to find patientients in each category |
| | | CCLF1 |
| Acute myocardial infarction (AMI), | contains myocardial infarction | 100 |
| Heart Failure, | contains heart failure, does not contain without | 1083 |
| Pneumonia, | contains pneumonia, | 237 |
| Chronic obstructive pulmonary disease (COPD), | contains chronic obstructive pulmonary disease | 634 |
| Coronary artery bypass grafting (CABG), | contains coronary artery bypass | 43 |
| Stroke and Total Hips/Total Knees. | contains stroke; contains hip;contains knee | 1;617;1068 |

- Use multiple sources to solve: And then we put them in Excel to find that the common result is that each disease starting from a specific code.

| myocardial infarction | search code begin with | I21 |
|---|---|---|
| CCLF1_PRI | CCLF1_PRNCPL_DGNS | |
| I214 | Non-ST elevation (NSTEMI) myocardial infarction | |
| I2109 | ST elevation (STEMI) myocardial infarction involving other coronary artery of anterior wall | |
| I2119 | ST elevation (STEMI) myocardial infarction involving other coronary artery of inferior wall | |
| I213 | ST elevation (STEMI) myocardial infarction of unspecified site | |
| I2102 | ST elevation (STEMI) myocardial infarction involving left anterior descending coronary artery | |
| I2129 | ST elevation (STEMI) myocardial infarction involving other sites | |

- Alteryx Application: we apply decision in Alteryx to processing disease searching method at last .




**Strategy Next Time**

It would be great if we can fixing technical issues like unexpected issues of connecting Alteryx on different laptops, clean ERD before Altery, and consulte to professor about the data processing methodology at first.

We may also use dimensional modelling more skillful, because it is easier to retrieve information and generate reports and easier to identify and describe when/where/who and what of the business process. We will also have confidence to use factless fact table which can be easy to use, capture information from each dimension.