Bon Secours: A Study of Medicare Readmissions

Christine Malloy, Patrick Poehailos, Trace Zhang, Taylor Gilman

BUAD 5272: Database Management

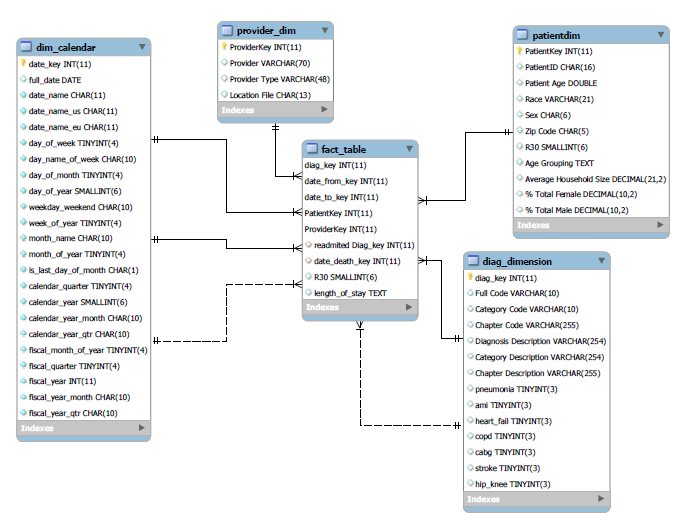
## Introduction

*Describe your interpretation of the data, the two main aims of your work, a summary of your approach and a description of your conclusions*

* The two main aims of our work were firstly to analyze how readmission varied by location and secondly, analyze how ethnicity had an impact on the length of stay/admission/readmission; this second question was later revised to analyzing how readmission varied amongst different diagnosis codes.
* Our approach was to combine the high level and low level data files. Although we believed we had done this correctly we soon found that our data was riddled with duplicates. After consideration we decided to remove the lower level data as it was not necessary in the assessment of our research questions, which focused solely on admissions into the hospital and not primary care information.
* After deciding which files we would use in the formation of our star schema we decided on including four dimensions, those being : Patient, Provider, Calendar, and Diagnosis.
* Our conclusions of the data showed that showed that for patients who were readmitted for stroke ended up having the longest length staying during their readmission compared to other diagnoses who had readmissions. We also saw from our conclusions that the highest densities of readmissions based on zip codes are not around the centers of the data sets (Richmond & Hampton Roads) but instead, in more rural areas. This may be because of economic status, proximity to primary care facilities, or other factors.

## Modeling Approach

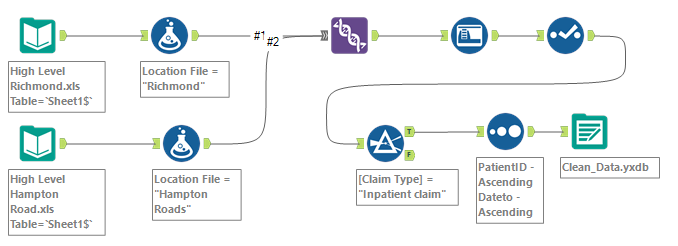
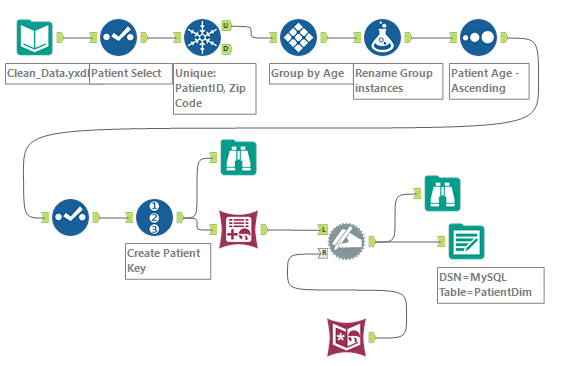
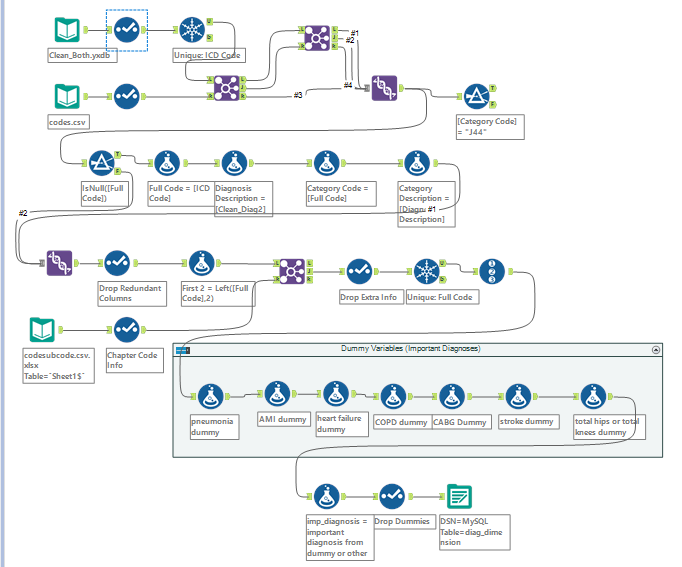
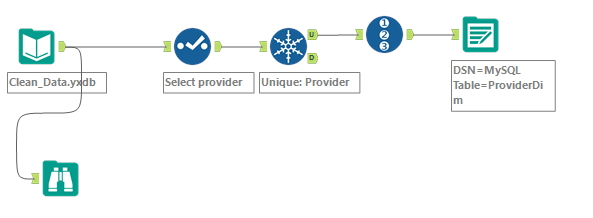
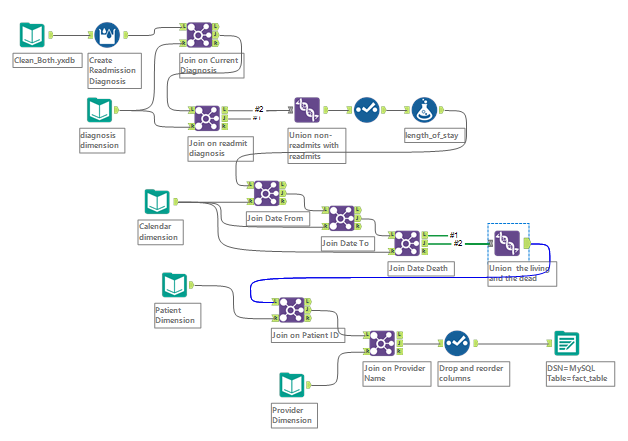
*Justification of your modeling approach. Include your ER and RS in this section.*

**

* We felt that creating a star schema would be the easiest way to organize the data provided. For this reason we knew we needed to form dimensions around the:
  + Who - Patient
  + What - Diagnosis
  + When - Calendar
  + we also included the provider dimension for had wanted to dig deeper into questions regarding primary care information.

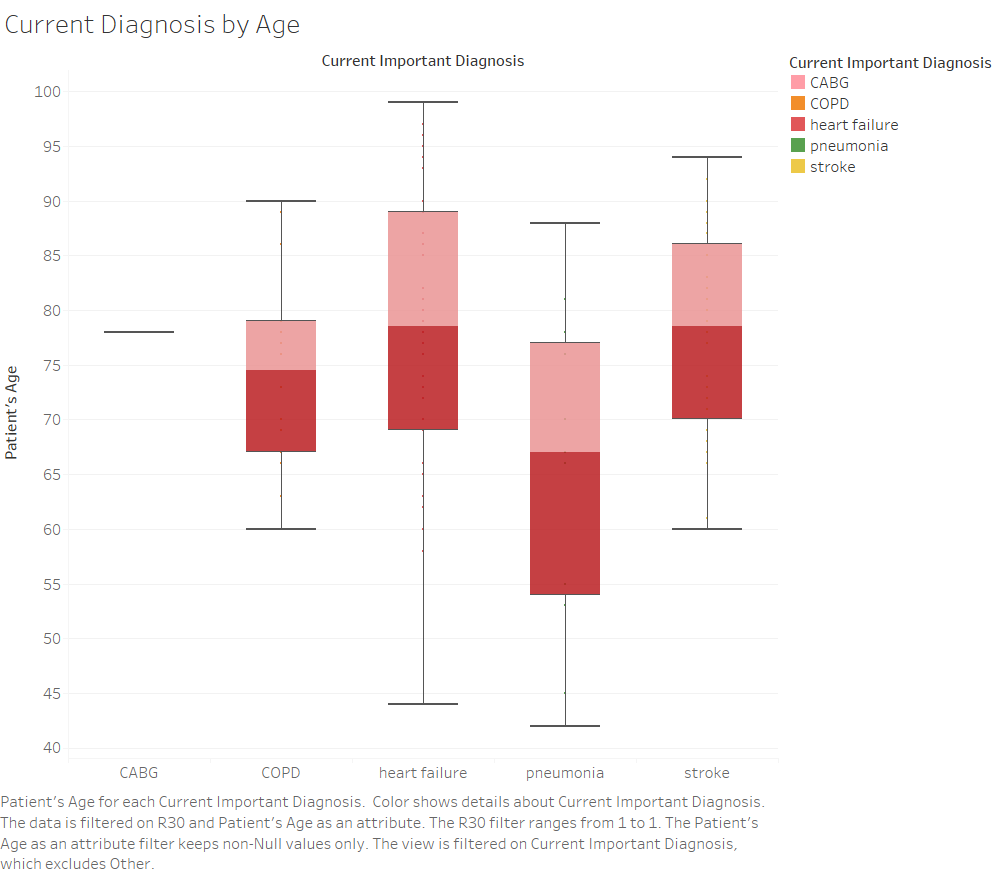
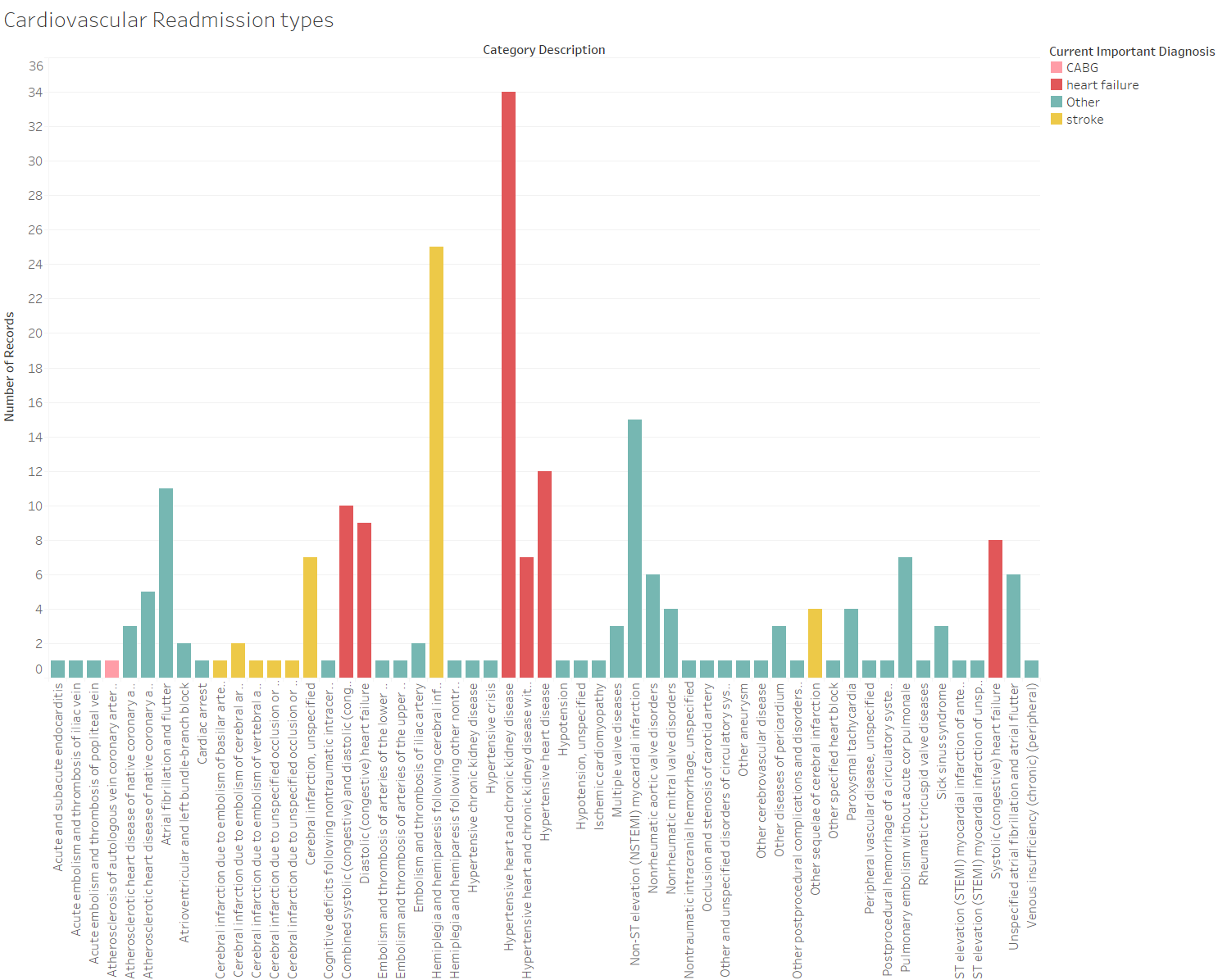
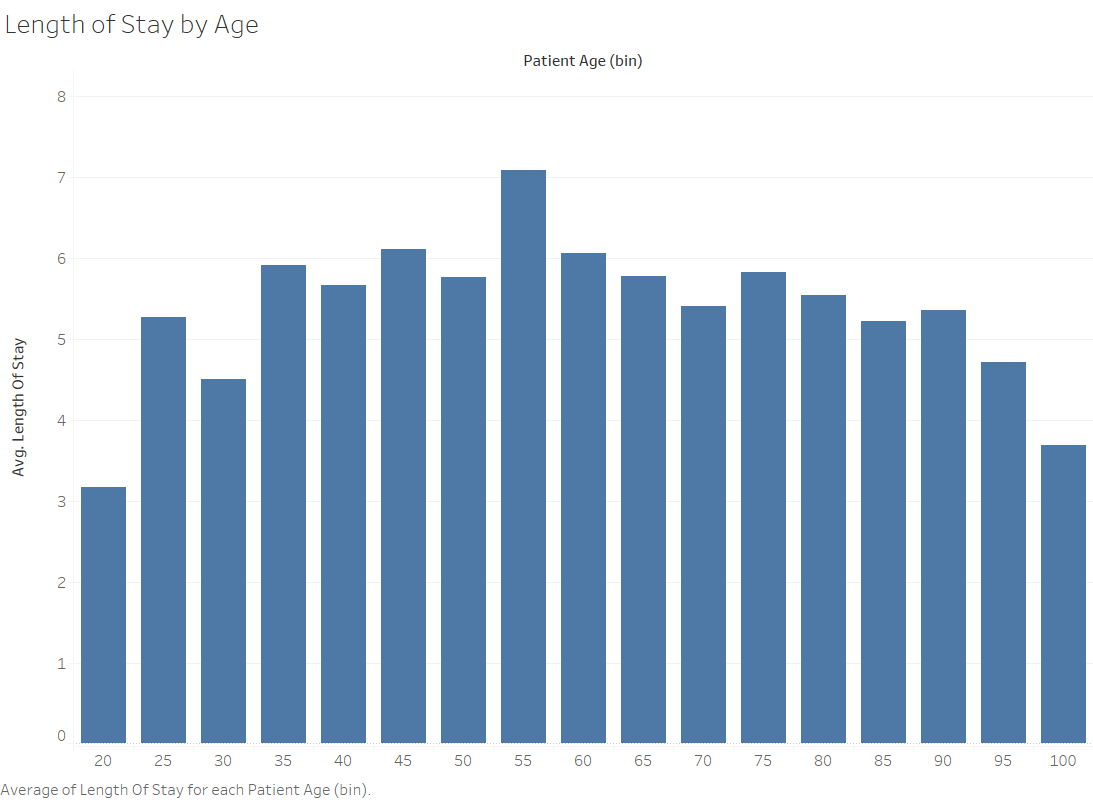
## ETL Approach

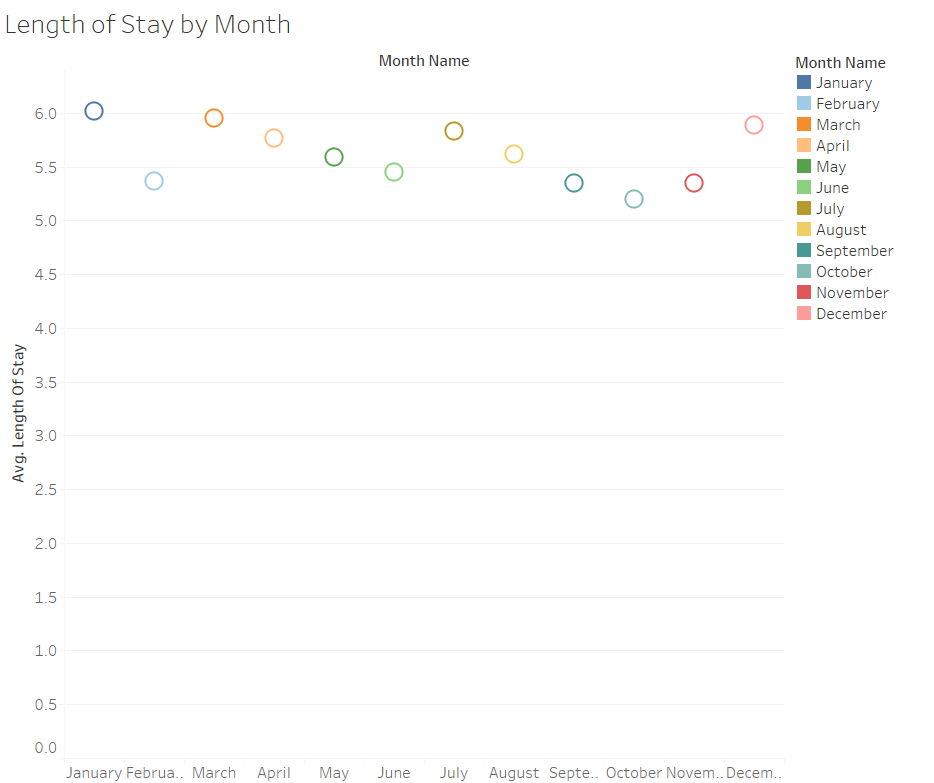
*Describe your database implementation and ETL approach. Use screenshots when necessary.*

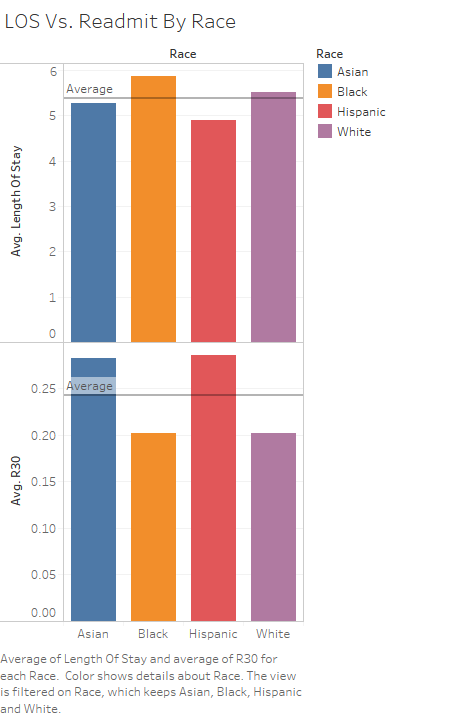
* 
* The cleaning process during this project included joining the two high level data files together filtering for inpatient claims.
* 
* After the initial cleansing process we started working on the patient dimension. In this dimension we grouped our patients into age bins and introduced census data to help fill in some of the nulls in the patient dimension by matching on zip codes.
* 
* Building the diagnosis dimension was challenging for because of necessity of outside ICD code files to ensure that our filtering process returned the appropriate types of diagnosis we were looking for. The steps in this process were as follows:
  + Join all of the unique codes from our clean file and introduce one of the ICD code files to join with the data to create a category code.
  + We then dropped redundant columns and extra info before introducing our second outside file to create chapter codes.
  + The last bit of work we did involved creating dummy columns to count of the actual occurrences of the diagnosis that were of interest before adding them into a single column and finishing our MySQL file for the the diagnosis dimension.
* Building the calendar dimension was as simple as running the code we were given in class and making modifications to the appropriate start and end dates.
* 
* Our final dimension, the provider dimension only included filtering for the unique providers, selecting the correct information to include in the dimension and giving it a primary key.
* 
* Our end fact table consisted of joining all of our dimensions together and creating a MySQL file to import and build the fact table to use in tableau.

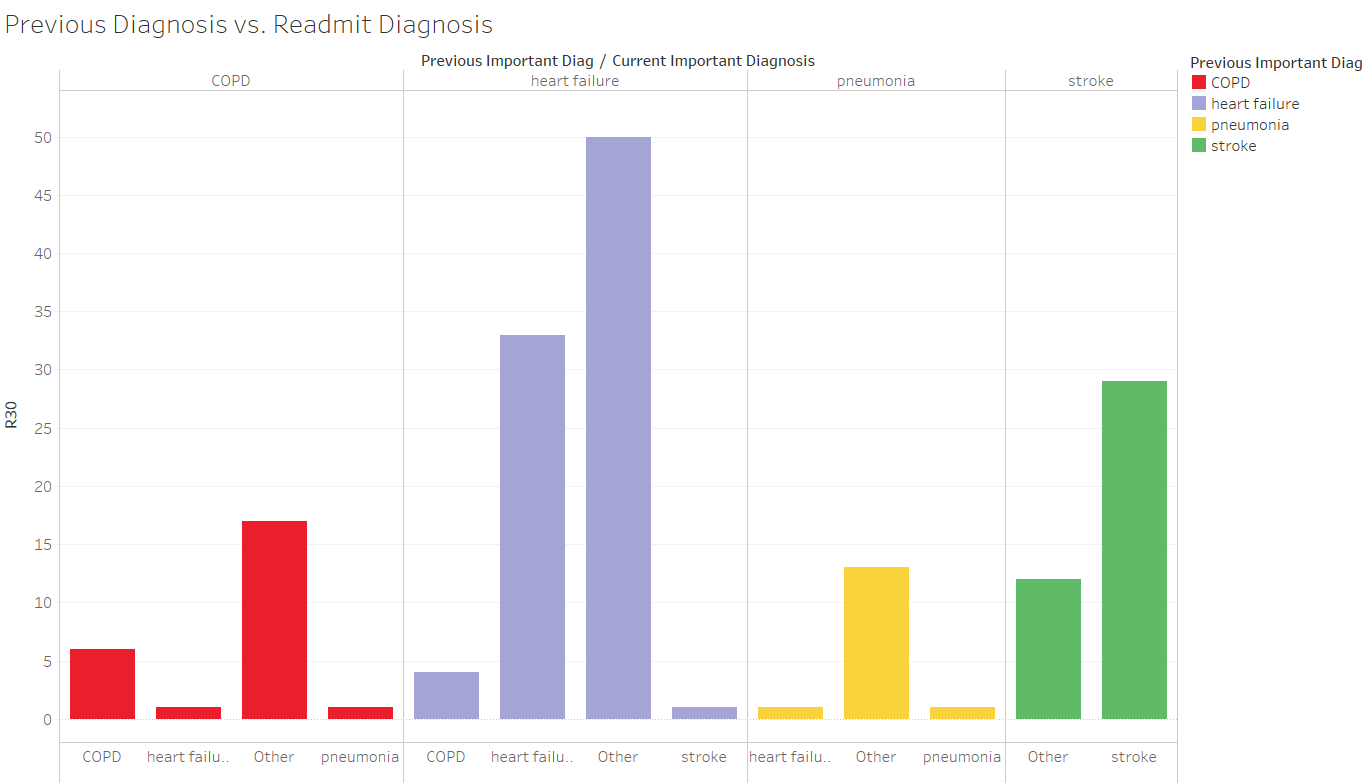
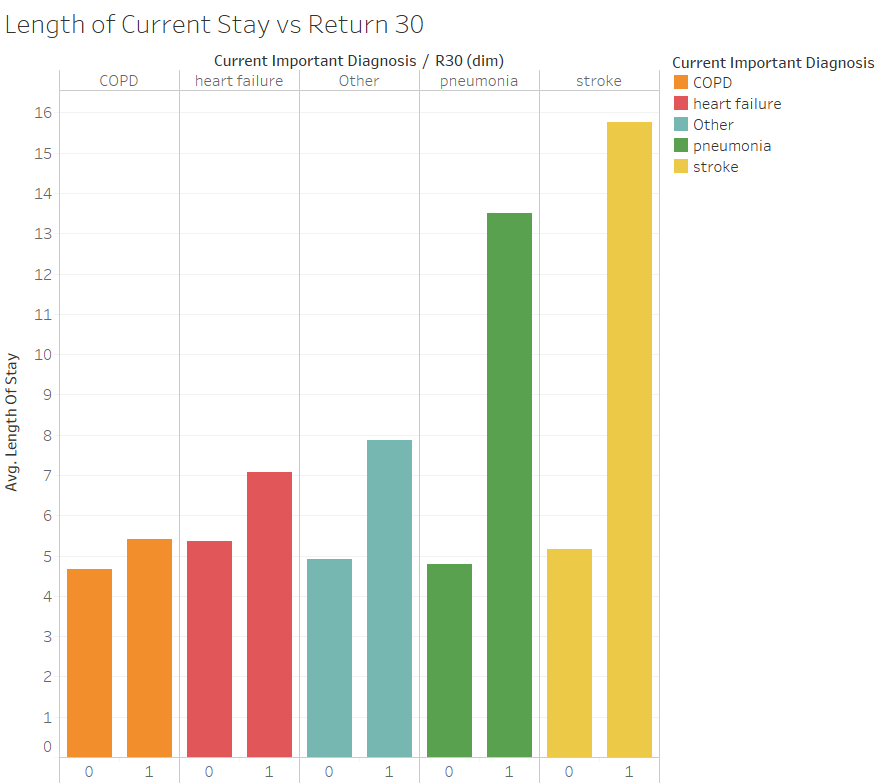
## Queries and Visualizations

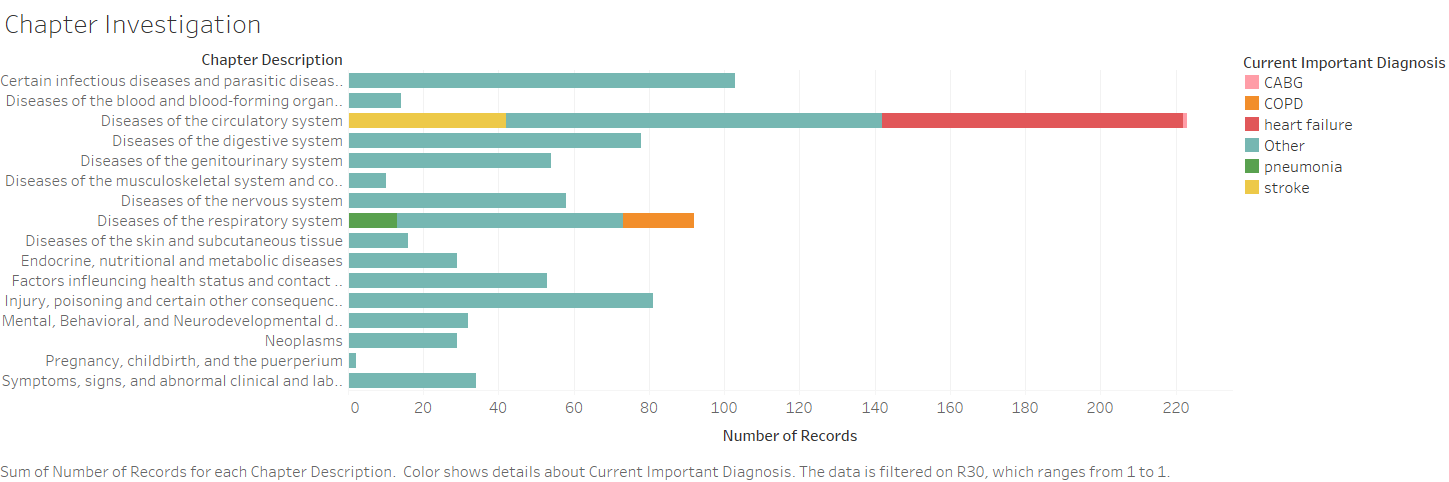
*Describe your queries, visualizations and results. No need to provide the full code of the queries, you will be submitting theses separately.*

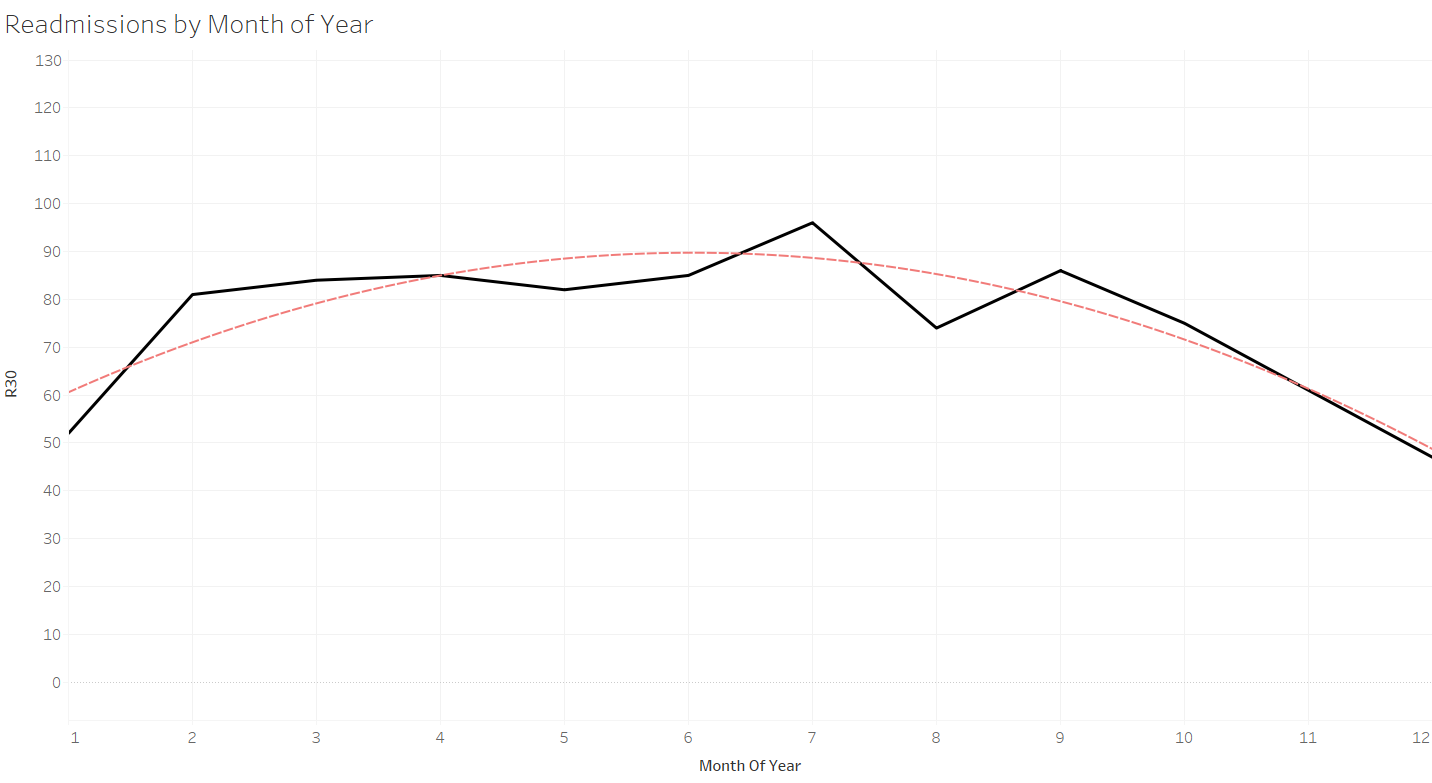
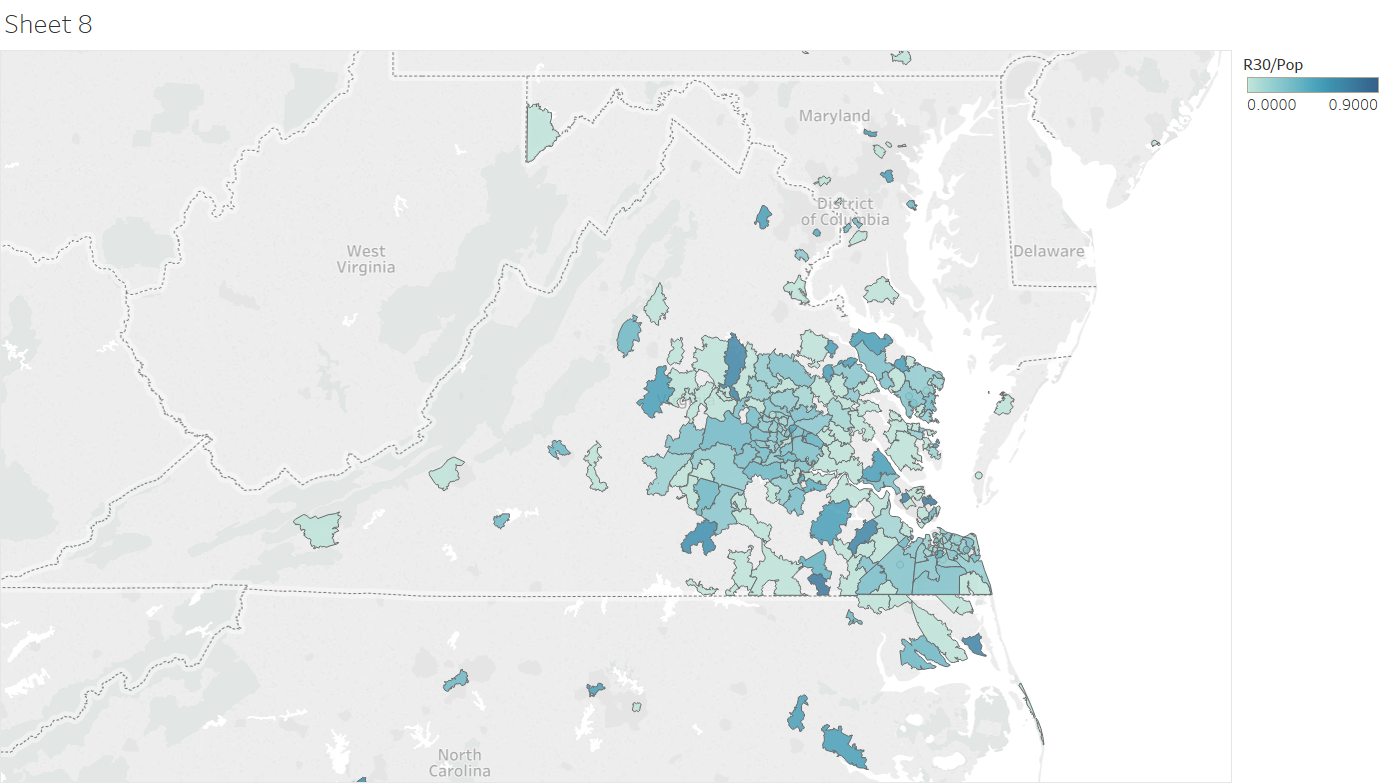
1. 
   * For patients readmitted with our key diagnoses, this shows the box-plot distribution of their ages. Notably, heart failure patients tend to be more widely varied, pneumonia patients tend to be younger on average, and 1 patient was readmitted with CABG.
2. 
   * For diagnoses in the Chapter Code referring to cardiovascular issues, here is the count of readmissions. Importantly, besides our key diagnoses there are notable readmissions in myocardial infarction (heart attack) and heart flutters.
3. 
   * This shows the length of stay based on age (bins of 5 years). The peak is around 50-60 years of age, which may be because of financial situation, ability to recover, or something else.

1. 
   * Average length of stay for each month of the year. There appears to be a trend downwards in the spring and again in the summer, bottoming out in June and October.

1. 
   * Though we don’t have the breadth of data needed to create a query of true value we still attempted to create a query to look at our beginning research question of how readmission and length of stay differ by ethnicity. The graph does so interesting trends among Black and White ethnicities having less average readmissions in 30 days, and longer average lengths of stay. The converse effects were seen in Asian and Hispanic ethnicities. We do understand that these results should be taken with a grain of salt because the averages are skewed due to lack of data on ethnicities other than White.

1. 
   * This depicts the type of readmissions diagnosis based on the initial diagnosis. For initial diagnosis of heart failure, a large portion of those admitted were readmitted for heart failure, as well as other illnesses. In particular, those who are initially diagnosed with stroke and were readmitted were largely readmitted due to stroke and corresponding diagnoses.
2. 
   * This visualization plots the average length of stay against our key diagnoses and whether this was a readmission of that diagnosis or an initial diagnosis. The key takeaway from this visualization is that the return length-of-stay for stroke diagnosis is 3x longer than the initial stay. This is much more significant than any of the other individual diagnosis. As such, there is likely a higher cost associated with an increased length-of-stay.

1. 
   * This plots the number of readmissions (the graph is filtered by R30 = 0) by the different chapter types as set in the ICD. As we can see, cardiovascular and respiratory issues make up a large amount of our readmissions, as well as certain infectious diseases (A00-B99).

1. 
   * This is a time series graph showing the sum of R(30) grouped by each month of the year (by number, 1 = Jan , 12 = Dec). Additionally, we plotted a quadratic regression against it to show the general shape. This graph shows that readmissions are higher in the summer months.
2. 
   * This is a map showing density of R30’s by zip code. It is interesting to note that the highest densities are not around the centers of the data sets (Richmond & Hampton Roads) but instead, in more rural areas. This may be because of economic status, proximity to primary care facilities, or other factors.

## Conclusions

First and foremost, the most time consuming and difficult part of the project was cleaning the data. This affirms that what we have been told so far throughout the program is true. Much of our time as data analysts will be spent manipulating data so that it is in a useable format. Most notably, in this project, there were issues in the initial datasets with ICD code mismatches. In short, this problem was remedied by acquiring a more robust list of ICD codes online to be able to validate correct codes and fix issues.

On the flip side, because of the research questions we established, determining dimensions needed for our star schema/fact table was relatively simple. Since we focused on R30’s and high-level information, choosing Provider, Diagnosis, Patient, and Date dimensions was straightforward. If we did it all over again, we would work to combine high level and detailed data differently, in order to avoid duplicates. Thankfully it was brought to our attention that there were duplicates being created and thus, we were able to restructure our cleansing process to not combine the datasets.

Some highlights of the project included our Tableau queries and unique ICD codes. Tableau, as we have come to know, is a fantastic tool for visualizing relationships between different fields of a dataset. The ease of using it made our querying process much smoother. In fact, if we were curious about a certain relationship in the data, it was as simple as drag and drop to confirm or deny our suspicions. The funniest (or maybe most unfortunate) ICD code is  V97.33XD, which stands for “sucked into a jet engine, subsequent encounter”. If we had not been able to work on this Bon Secours project, we would have never known that fun fact!

Moving forward, we see two possible future avenues for analysis. First, the data showed interesting relationships. Two examples: first, higher readmission density was not centered in urban areas (i.e. Richmond, Virginia Beach) and second, strokes led to readmission lengths of stay nearly 3x longer than the initial admission. Stats like these make us wonder: why is that the case? Regarding readmission density, our hypothesis is that demographic characteristics of the geographic areas are the reason for the trend. Specifically, it is possible that in lower income, more geographically remote areas, hospitals are some patient’s method of primary care. We do not know that this is the case, but being able to gather additional information to answer this question would be interesting. Regarding stroke readmission length of stay: how can it be explained? Is it the treating practitioner’s plan that is prolonging the stay? Is it the severity of the issue? Looking outside of the scope of strokes alone, gathering more data on treatment plans, amenities available, etc, may be the first step towards answering this question.

Overall, this project opened our eyes to the numerous possibilities present when a large dataset with interesting information is available for manipulation and querying. The level of nuance and detail in the data we analyzed was apparent, and is obviously just the tip of the iceberg for providing a business with answers to its questions.