Writing sample 2

Patient-Diagnosis Analysis for Improvement of Treatment Quality

1. Introduction

Our group's interpretation of the medical data from Bon Secours involves a variety of patient information. Reading through the data, we found extensive information pertaining to patient demographics, providers, diagnosis, residence, and date of diagnosis or treatment. This detailed patient information can provide, through analysis, insight into possible trends within the medical system throughout the Richmond and Hampton-Roads regions.

We chose to investigate the inpatient cases in order to observe the behaviors incurring the most expensive treatments and care. The main goals of our investigation are to analyze specific patient diagnoses and the circumstances surrounding their admittance and discharge from the hospital and to understand the inpatient's impact on the healthcare and insurance systems. We chose these two objectives because a large issue facing today's healthcare industry is the rates of hospital admissions, visits, and procedures. We aim to make a breakthrough in evolving the way certain patients are treated within the hospital and post-treatment, in order to ensure an improvement in the quality of life and wellness.

Our group approached this project by defining the problem that is at hand by answering the following question: "how can we make additional interventions into treatments and care that would improve our interpretation of the patient's experience". We started by observing the data that was presented to us in the excel files while paying close attention to columns such as diagnoses, admittance/discharge, providers, location, and dates. Furthermore, we decided the best course of action was to build our project schema with five dimensions: patient, calendar, diagnoses, provider, and location. We attempted to identify influential variables that had a larger impact on whether a patient stayed an extensive amount of time in the hospital, was readmitted, or visited their primary care physician within the past couple of months.

Through our investigation, we were able to determine that a patient who visited their primary care multiple times was less likely to get readmitted or stay an extensive amount of time in the hospital. Although, patient visit rates change slightly with higher mortality rate diagnosis like Heart Failure and Chronic Obstructive Pulmonary Disease, which require the patient to be admitted to the hospital many times and often stay for long times too. We also found that areas with a lower population were more likely to face health adversity and often had a higher mortality rate and longer stays in the hospital, granted that they had access to said hospital. We also observed that the category of beneficiary type that made claims most often were those of the Retired Worker category. We can infer that they have more coverage for a multitude of reasons (possible reasons being they have more income from their work or they received post-retirement benefits). An unforeseen factor was that younger, retired workers were claiming more out of their insurance compared to their counterparts that were older and maybe did not have as good of coverage under the beneficiaries plan. This is the opposite of what we had initially expected as the healthcare industry is more geared towards elderly patients and their need for continuous care.

From these conclusions, we can say that targeted interventions towards low income and low population areas may equalize the level of treatment and care that is afforded to other patients. Additionally, we suggest that patients facing diseases and conditions that are linked with high mortality should be seeing their primary physician more often so that they may reduce their admission amounts and days spent in the hospital. This will lower the costs that patients face in the future and allow for hospitals and physicians to focus on more pertinent cases that require immediate attention or more attention.

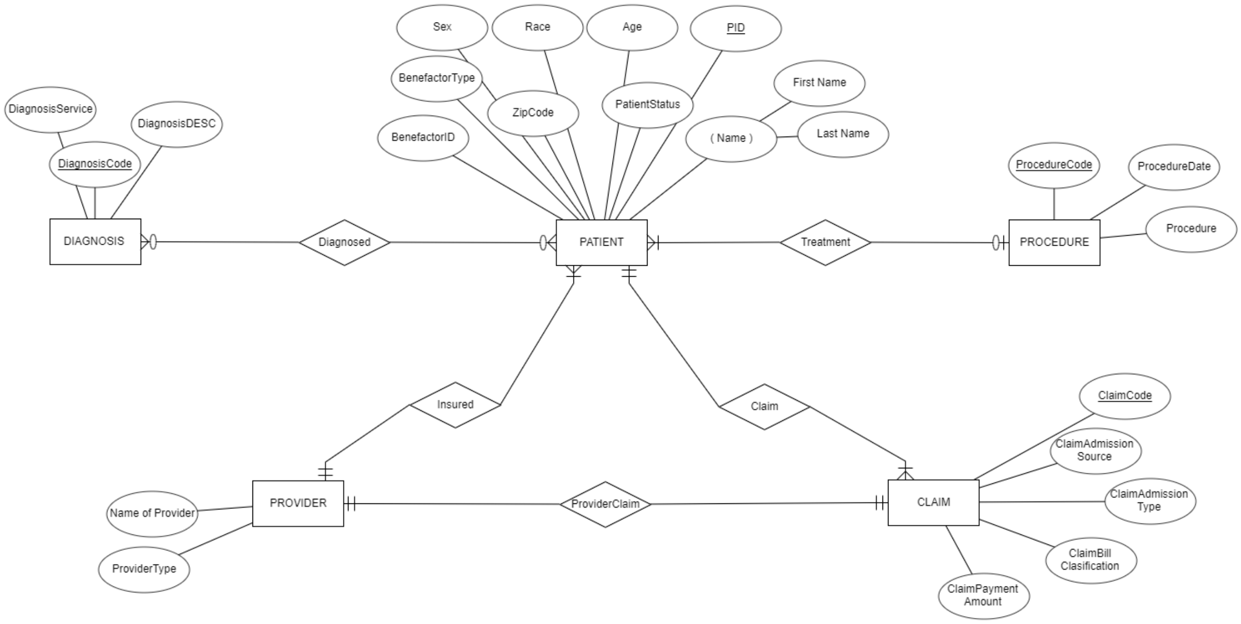
2. Modeling Approach

(written by Wenying Shen)

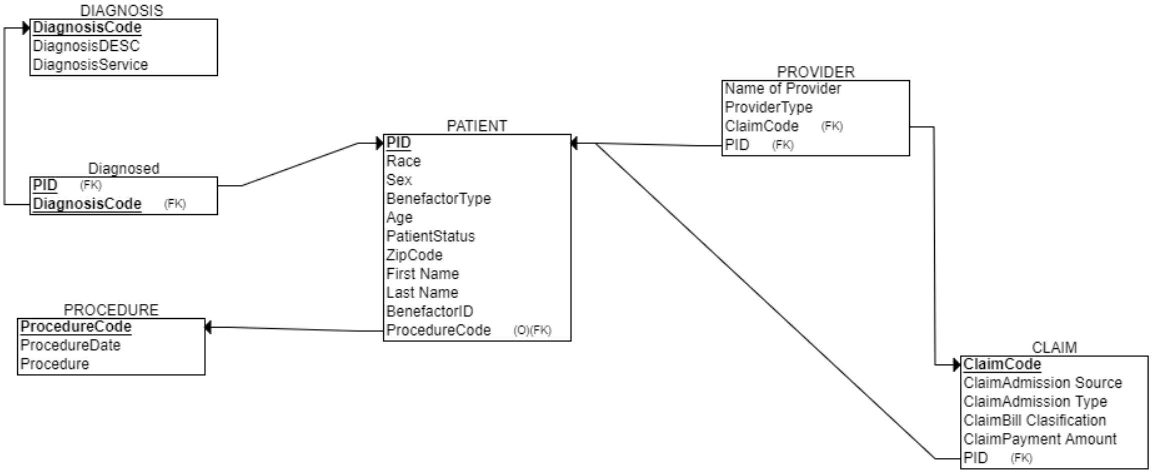
Through analyzing the provided data from Bon Secours, our group designated dimensions for location, patient, provider, diagnosis, and date. Breaking down the data into these five dimensions improved efficiency in running queries to answer questions we had about the medical information. The image below illustrates our group's Entity Relationship Diagram for the Bon Secours data.

Based on this ER diagram, we were able to identify the main attributes needed in order to support our previously discussed analysis. The relationship cardinalities aided our understanding of the complicated data correlation provided by the excel files and how to approach the data warehouse modeling.

Entity Relationship Diagram



<u>Relational Schema</u>



3. ETL Approach

Our group implemented the database with an initial selection of the observation variables that were of interest to our analysis, which included the patient, diagnosis, location, admission and discharge date, and the provider dimensions. The first step was to extract the data from the provided excel files into Alteryx in order to begin the manipulation and cleaning process. The second step was transforming the data by manipulating the different files into dimensions.

1. Patient Dimension: We created smart tiles for patient ID, population based on patient ID, the amount of time they stayed with each admission, and we utilized the most important attributes for each patient i.e age, sex, etc.

2. Location Dimension: We used the demographic analysis in Alteryx in order to rename the counties and cities based on the zip code provided. Next

3. Diagnosis Dimension: Identified the diagnosis name and service based on the diagnosis code provided from the high-level and low-level claim data.

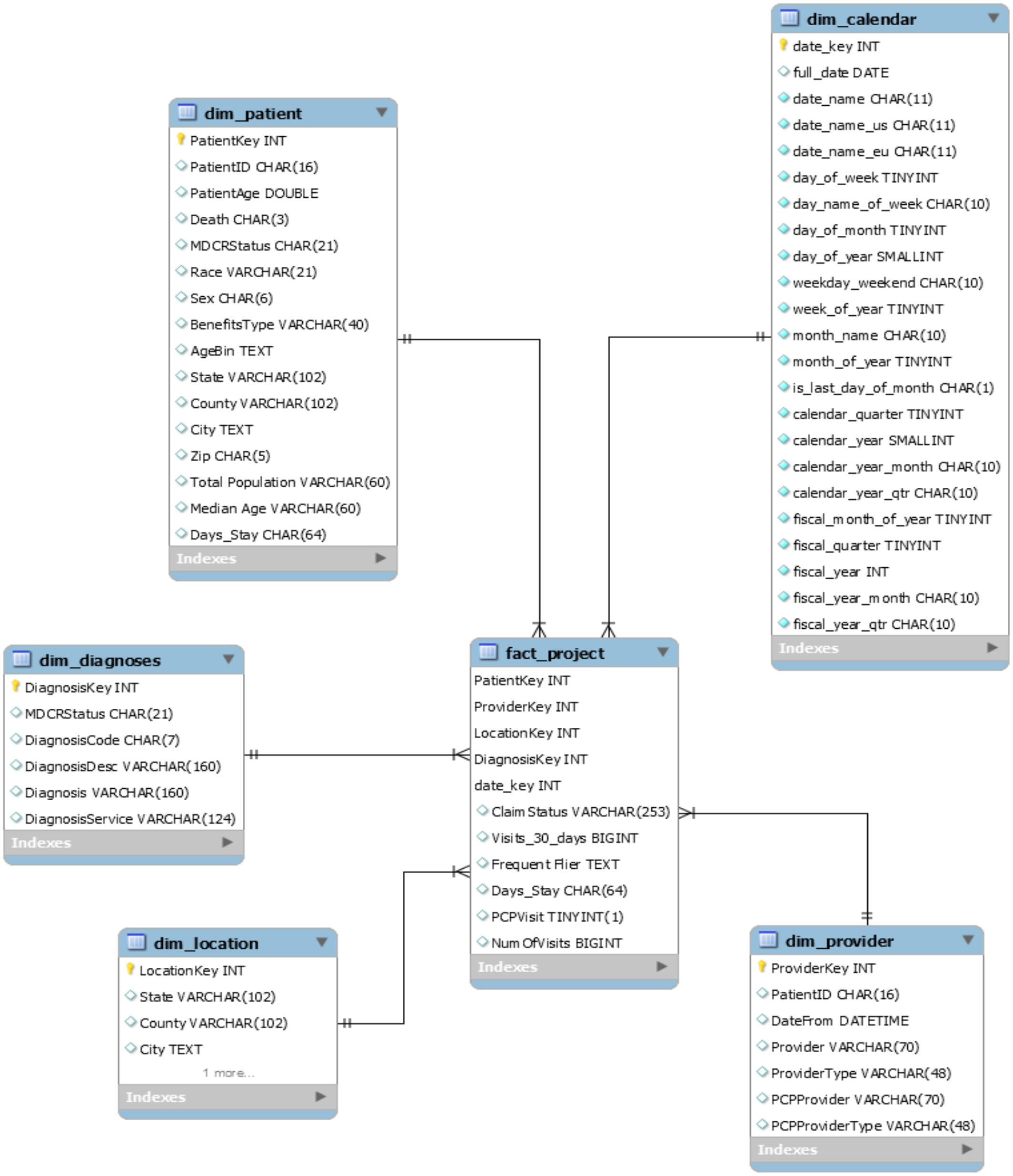
4. Provider Dimension: Identified the provider type and primary care provider based on the provider code supplied by the high-level and low-level claim data.

5. Calendar Dimension: Calculated the different date structures for the data and the difference in dates (date to minus date from).

Through the created dimensions we constructed the fact table by combining different keys together in order to make the primary keys for the latter table.

Third, by adding in the pre and post created SQL statements for each dimension, we were able to populate our project schema in MySQL. This allowed us to build queries in order to investigate what affects our goals in the project.

Star Schema



4. Queries

Query 1: Based on the different diagnosis names created during the ETL process, we wanted to investigate the average inpatient stay in the hospital compared to the average number of visits of these patients to their Primary Care Physician (PCP) within the last 6 months.

Observations: We noted that people who have a Heart Failure condition at age 65-74 seem to have the longest period of stay and the lowest number of PCP visits compared to people with the same diagnosis, who are in the 75-84 age bin but have been frequenting their PCP more often. Therefore, the older patient ends up spending less time in the hospital. Here, we can assume that there is a negative correlation between PCP visits and length of stay in the hospital. Moreover, the older the patient is with "COPD" (even with a higher number of visits to their PCP) the patients still spend the most amount of time in the hospital on average.

Query 2: We wanted to investigate the disease that has the highest mortality between all the other diagnosis.

Observations: Based on the returned results, we can safely assume that patients who are prone to heart failure or have a hereditary history of heart failure in the family need to be visiting their PCP more often and, should be monitored more closely as they are the ones spending more time in the hospital or end up losing their lives to such conditions at a higher rate.

Query 3: Here, we wanted to determine the length of stay for patients considering the total population and the fact that they have been visiting their primary physician in the last 180 days.

Observations: The result displayed demonstrates that patients residing in low population areas visit their PCP the least amount of times in the previous 6 months, and stay the least amount of average days compared to other patients from other population tiles. This makes sense because they are unaware of their health issues and lower populated areas have less access to PCP care and patients usually have an overall lower income. The average and above-average populated areas follow the trend as they have more access to healthcare, as well as higher levels of income. On the other hand, Extremely low populated areas visit their PCP more often than the low populated areas. We suspect that this may be because there is a centralized PCP for certain towns/areas. The reason why extremely low areas of population have a lower average days stay in the hospital is because they most likely do not have immediate access to a hospital so they will be less likely to stay in one.

Query 4: The goal of this query is to count the average number of patient visits to their PCP in the 2016, 2017 calendar years based on benefitsType.

Observations: Based on the results, we can see that there is a decrease in all of the benefit types making claims from 2016 to 2017. In addition, we observed a general trend towards an increase in the number of patients visiting their PCP. This makes sense as patients are being more cautious with their health issues before the need for any sort of hospitalization, especially with the Retired workers (as prices for hospitalization are on the rise as well). This confirms our initial analysis that an increase in visits to PCP might help decrease the number of visits to the hospital, limiting the hospitalization length stay.

Query 5: We want to analyze the type of disease, average length of stay, PCPvisit, and hospital visits broken down by race to see if there is any correlation.

Observation: Following our result set, it seems that even though our data set includes more Whites than any other race, the average number of visits to both PCP and hospitals seems to be very similar. Also, we cannot assume or associate any of the diagnosed types to any race since there is no apparent trend to whether race influences any of these variables.

Query 6: We want to analyze all the different diagnoses from patients having their primary addresses in Virginia as well as other states and its effect on the average number of PCP visits and the length of stay in the hospital.

Observation: Interestingly enough, even though Virginia has the highest number of different diagnoses, the average days spent in the hospital for most diagnoses are considerably lower than in other states such as Ohio and Pennsylvania, which have more than double the Virginia state average. This may lead us to believe that Virginia has a better care system than other states comparatively.

Query 7: This query helps us list the patient’s providers, city, and state with the count of diagnoses that they receive.

Observation: Based on the results, we can observe that the “Bon Secour St. Mary's hospital” in Woodbridge, VA, has the highest number of repeated cases of “chronic kidney disease stage 3 (moderate)”. This suggests that there is an underlying issue facing the people who visit this hospital or that the hospital specializes in kidney disease treatment. We can target this area to see if there is a reason as to why patients are experiencing this kind of health concern. There may be a correlation between excessive smoking and drinking habits among these citizens. A possible intervention into diet and or health choices may be needed to decrease the dependency on hospital visits and lower the cost for the patients and the healthcare industry. When we queried our fact table to specifically observe the COPD diagnosis (the most average visits and the most average stays), we can conclude that the “Mary Washington Hospital” has 8 cases. As a further analysis we looked at the diagnosis that was of importance to us:

Query 8: This query helps us provide a deeper analysis on the provider when it comes to the diagnoses names set up during the ETL process.

Observation: This query helped us determine that Heart Failure has the most cases in both the “Calvary hospital” and “Waldo county general hospital”. This trend continues with observations into lower populated areas, and we can further conclude that the most affected areas are in the south of Virginia. There could be an underlying reason for this, an assumption would be that as these areas are less wealthy, they may face more health challenges compared to their more populated and wealthier northern counterparts. The basis for this claim is that we do not observe these trends in the northern cities and hospitals.

Query 9: In this query, we are analyzing the effect of benefits type holder and frequent flier levels (high, low, and no) while confirming the average day spent in the hospital.

Observations: When we changed the Frequent_Flier levels from “no” to “low” to “high”, we noticed a pretty interesting trend:

High → Retired workers in the age bin of 55-64 are more incentivized to go back to the hospital more often than any other benefit type holder mainly due to their existing coverage and suspected higher disposable income.

Low → Interestingly enough, the retired workers are also the highest claimers with fewer visits, but with more days spent in the hospital. This is probably due to a needed medical intervention.

No → Same as the two previous categories, it seems to us that the most claimants are the retired workers. They are more incentivized due to their coverage. We would expect more of a leveling off for the different types of beneficiaries as everyone becomes more equal in needed care as they grow older, however, this is not the case.

<u>Query 10:</u> This query describes the diagnosis name set up during the ETL process, the inpatient provider when the average stay is superior or equal to 5 days.

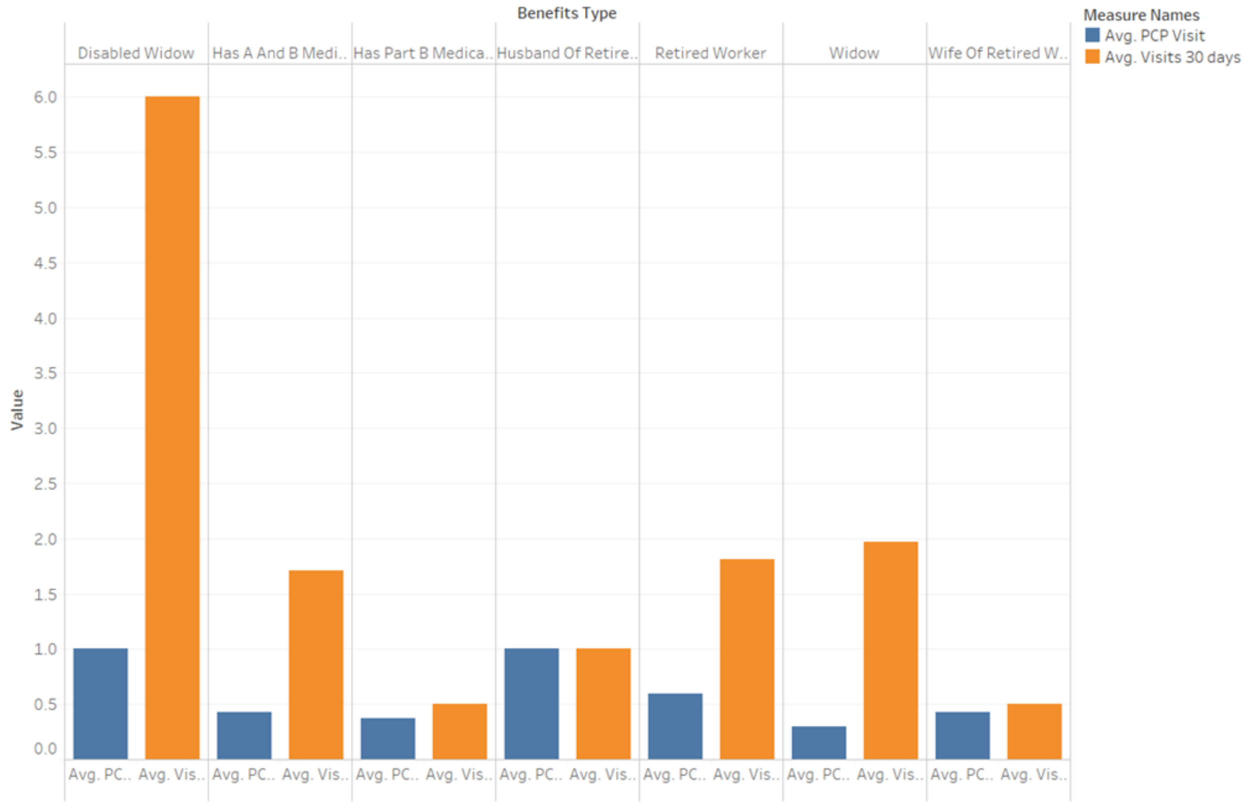
<u>Observations</u>: Based on the query, we can assume that hospitals with "Heart Failure" and "COPD" patients tend to keep patients for longer than other types of diagnoses. This may be because they have high mortality rates and the patients need extra care to stabilize them/make them live longer. Also, "MIA" has high numbers of days spent in the hospital. For example, if we observe the "MIA" days spent for different hospitals we can see that there is a drop in the amount of average days spent at the extreme of 21 to 6. This means that the care levels may be different. "COPD" cases are more likely to stay a similar amount of time in the hospital indicating that the diagnosis care is probably more predictable and serious. This proves to be true for "Heart Failure" as well. The pattern seems to indicate that chronic diseases require more visits to the hospital but may not need longer stays. Diseases that are more random in their onset may prove to have lower visits to the hospital but may end up having longer stays due to their unpredictability.

5. Visualizations

(written by Wenying Shen)

Graph 1:

Average PCP visits Vs Average 30 Day Hospital Visit per Benefits Type



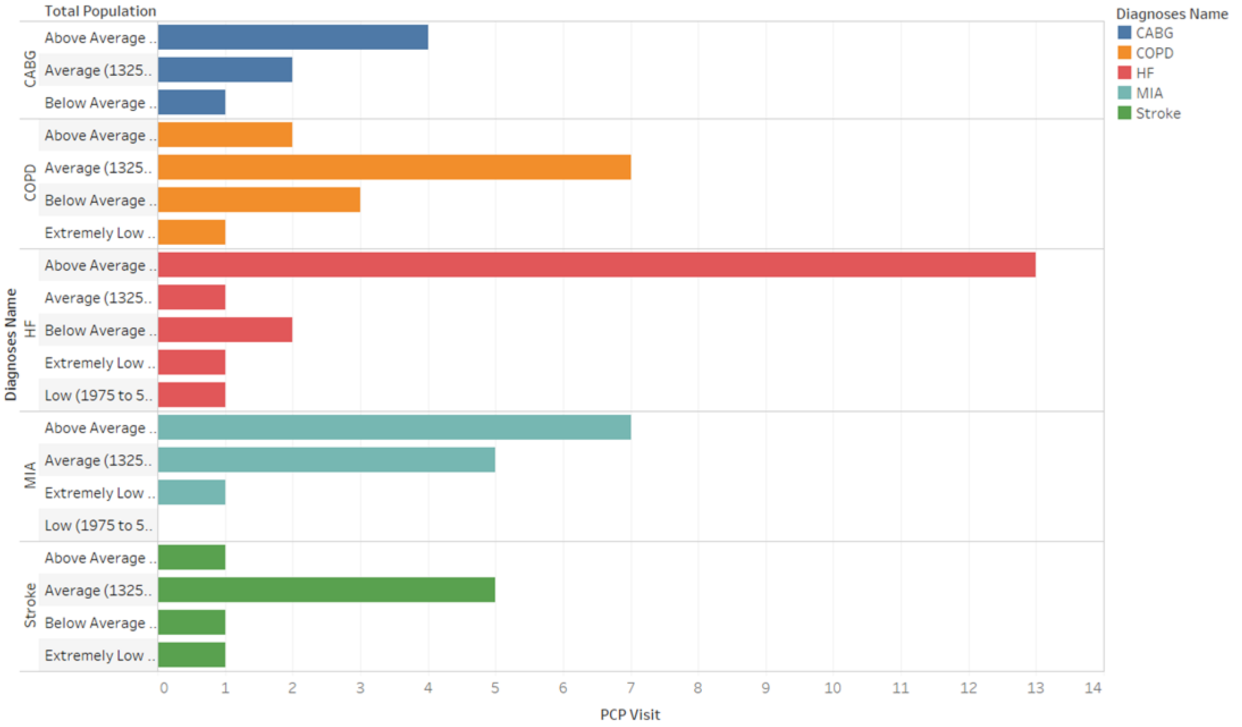
Avg. PCP Visit and Avg. Visits 30 days for each Benefits Type. Color shows details about Avg. PCP Visit and Avg. Visits 30 days. The data is filtered on Days Stay, which has multiple members selected. The view is filtered on Benefits Type, which excludes and Surviving Divorced Wife.

Observation and Recommendation: Here we are comparing the average number of times a patient has visited their primary care physician in the last 180 days, and the average number of days spent in the hospital based on the patients’ type of beneficiary. We can observe that “Disabled Widow” type visits their primary physician about once every 180 days but their respective average days spent in the hospital is at 6 days which is about 300% higher than the next highest average days spent in the hospital (Widows). We may need to further investigate as to why widows of all types have such a long average stay in the hospital. A possible explanation may be that they do not have anyone to take care of them if they were to be discharged, so they must rely on the hospitals to make sure they make a full recovery. We could possibly implement services to help them recover at home so that they take up fewer beds in the hospital and incur lower costs from their treatment.

Looking at the two medicare type beneficiary categories, we can see that they visit their primary practitioner at about the same rate. However, patients with both A and B medicare stay in the hospital on average 2.5 times longer than patients who only have part B medicare. We can attribute this to the fact that AB medicare patients have better coverage than their B only counterparts holder, which gives them an advantage when using their socially funded medical care. We can conclude that we can improve this discrepancy between the two groups by possibly expanding medicare coverage for the part B patients so they can receive a similar quality and length of treatment.

Graph 2:

PCP Visits for Diagnosis By Population In Virginia

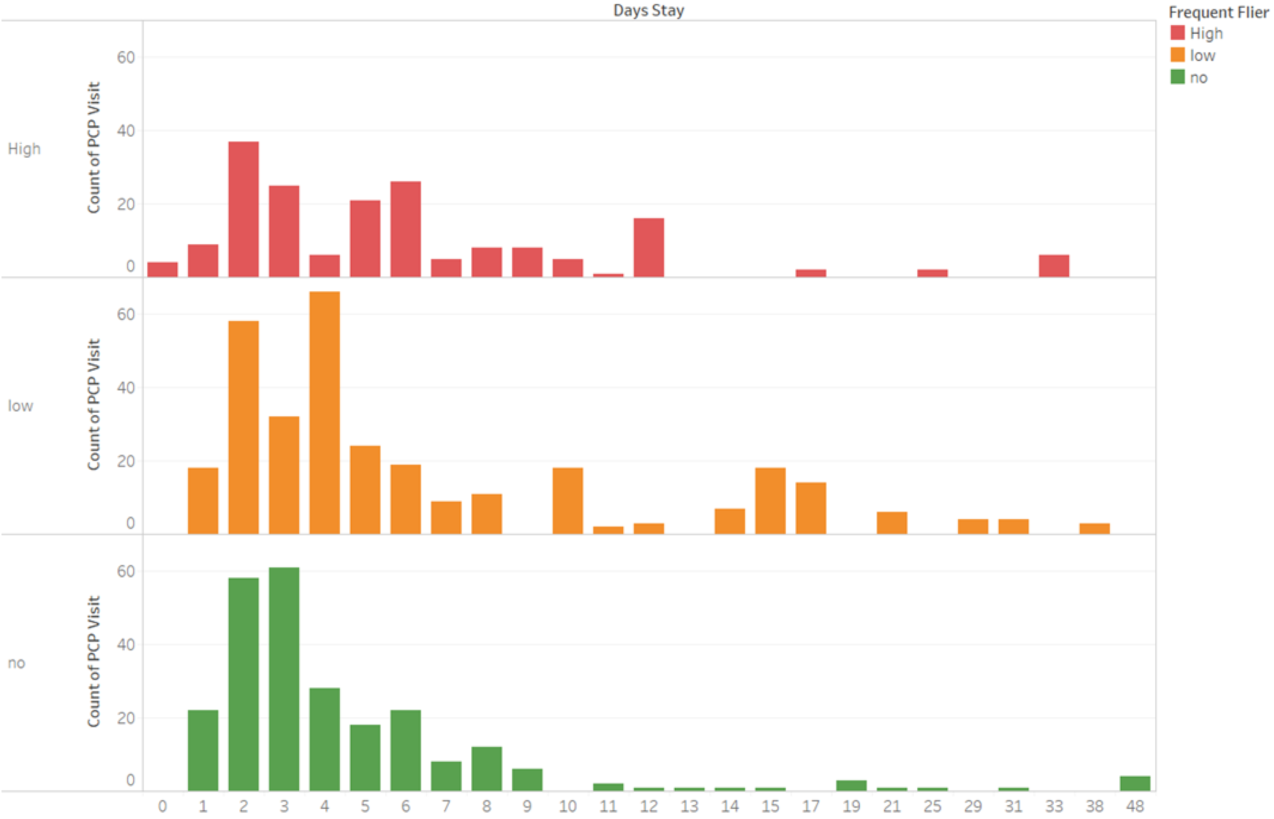


Sum of PCP Visit for each Total Population broken down by Diagnoses Name. Color shows details about Diagnoses Name. The data is filtered on [State]= "Virginia", which keeps True. The view is filtered on Diagnoses Name, which keeps CABG, COPD, HF, MIA and Stroke.

Observation and Recommendation: This graph clearly demonstrates the diagnoses that had the most primary practitioner visits. We can see that "HF" or heart failure had the highest amounts of visits, especially amongst the above average populated areas. This is somewhat expected as patients with potential HF are more prone on visiting their primary practitioner and take the matter more seriously. Interestingly enough, patients suffering from strokes are amongst the lowest in terms of PCP visits within the last 6 months. This may be due to the fact that strokes have a more sudden onset and are less predictable, so patients would not be visiting their primary care physician as much prior to their initial stroke. Further analysis could be done to observe how much more frequently stroke patients visit their primary care physician after their initial stroke. Furthermore, we can infer that as areas get more populated, it is more common for patients to go see their primary practitioner on average. We can conclude that more outreach for the areas with less population to encourage patients to visit their primary practitioner which can lower the overall cost of hospitalizations in the long run.

Graph 3:

Number of Day Stayed compared to PCP visit By Level of Hospital Revisit



Count of PCP Visit for each Days Stay broken down by Frequent Flier. Color shows details about Frequent Flier.

Observation and Recommendation: This graph demonstrates 3 levels of patients that we considered in our analysis, a high, low, and a non frequent flier. A frequent flier is a patient that is readmitted into a hospital more than once in the last 30 days, with high being more than 2 visits, and a low frequent flier being 2 visits, and no being less than 2 visits. We can observe that on average, high frequent fliers tend to spend fewer days in the hospital even though they have the lowest rate of primary practitioner visits within the last 6 months. This could be attributed to the fact that they may be suffering from a chronic illness that is not deadly, but requires constant attention. Patients who are low Frequent fliers tend to spend more days in the hospital compared to the other 2 categories. Our conclusion is that even if we consider high frequent fliers a high cost, we actually should be focusing more on the low frequent flier who seems to be spending longer days in the hospital and consequently accumulating a higher claim cost. One possible intervention could be that we develop a manner that predicts who is at risk for these potential longer stays in the hospital based on the observations that are afforded to us so we can capture the health issues at an earlier time frame as to avoid the prolonged stay.

6. Conclusion

(written by Wenying Shen)

Before this project, we did not imagine that we could have faced a large database and be able to wrangle and transform it in a way that could help us make a report and run queries out of it. Although we completed the project in time, we might have approached this project differently by getting a better understanding of the metadata even before starting the ETL job. Trying to understand basic relationships and metrics before creating the data warehouse so that we could tailor our project towards things we deemed more interesting. Making sense of the granularity of the given files would have allowed for better wrangling and insights into the relationships between the different variables of interest such as the diagnoses, patient's interaction with both their PCP and providers. We may have found more information by consulting outside sources or databases that others have already done to reinforce our findings and conclusions.

Based on our findings we recommend that patients be more proactive about their health and that there be more of a focus on curtailing disease that may cause long and frequent hospital stays so that the cost and burden on the healthcare system be lowered. This is important because the cost of healthcare in the United States is only continuing to rise and so are negative health attributes e.g. obesity, diabetes, cancer, HF, and COPD, etc. If there is more of a social safety net for preventing these kinds of things we would see an overall improvement in the quality of life of citizens and an improvement in the quality of treatment for patients because hospitals will be less overwhelmed and more able to dedicate their time to patients who are in need of serious care. Also, it is advised for the health care workers to start incentivizing and encouraging all sorts of patients to start seeing their primary practitioner at least one time every 6 months as it turned out to have a positive correlation with an overall shorter hospital stay.

Moreover, future improvements might be made by observing multiple factors, because in a real-life situation those factors may have a negative correlation or positive correlation on the length of inpatient stay. This will be important in determining the type of care patients receive and how they should go about their future regarding their health and proactive actions they can take in order to lessen their dependence on the healthcare industry.

In conclusion, this project helped us realize the endless inferences that data wrangling, implementation, and analysis would have based on very few variables. Being able to seek the root causes of diseases and make informed decisions based on other unexpected variables was an exciting task and it lends itself to business analytics wholeheartedly by being able to see things that maybe were not so obvious at first glance.