College of William & Mary
MSBA

Section 2 Team 16 B
BUAD 5272 Database Management, Fall 2020

Examining Patient Readmission Rates in Conjunction with Multiple Health Care Outcomes to Find Novel Preventative Care Tactics

Roxy Mao, Alexander Giarracco

December 8th, 2020

Introduction

The United States has the largest health care expenditure as a percentage of GDP. While the powerful monetary pull of medicine does have its advantages, The U.S. leads the world's success rates in many elective and complex cancer surgeries; the healthcare system is positioned precariously with high drug costs and some of the highest readmission rates[1].

Our group used readmission rate as a starting block for our analysis; readmission rate is both an indicator of poor patient outcomes and unnecessary stress on health care. The data provided by Bon Secours was instrumental in gaining insight into the Hampton Roads and Richmond areas.

Virginia has exceptional levels of readmission. It nears the top ten states with the highest readmission percentage(). Before scuttling through the data, Virginia's high readmission rates blended one of our original theories on thoughts on leading causes. Obesity, high blood pressure, and other examples of pre-existing conditions were what shaped our original thesis. However, Virginia is not one of the most obese states, nor does it struggle with heart failure at a level than its southern neighbors.

After further sifting through the data, we began to select criteria useful for further analysis: benefit type, mortality rate, individual age, and varying diagnostic information. We constructed our two main aims, identifying a problem and deciding how best to intervene: better public benefits need to be put in place, so individuals visit primary care physicians instead of hospitals as a first and only resort. Additionally, many diseases people in the TideWater area were dying from are preventable, and people who did not have access to a Primary Care physician suffered a disproportionately low survival rate.

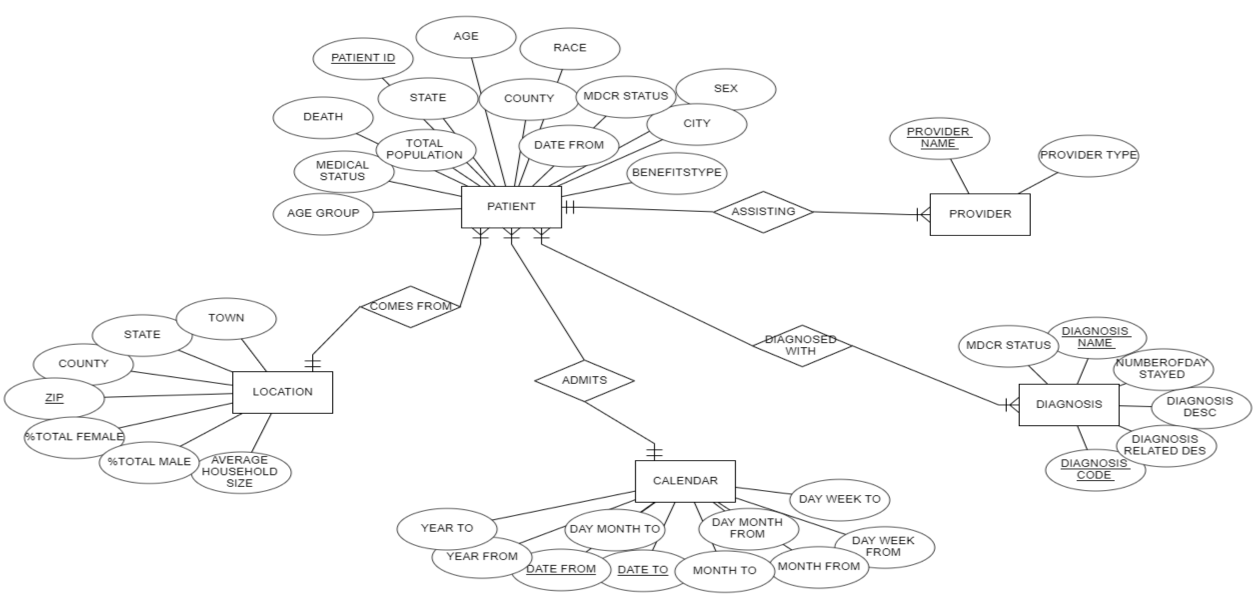
[1] Health Status

Modeling Approach

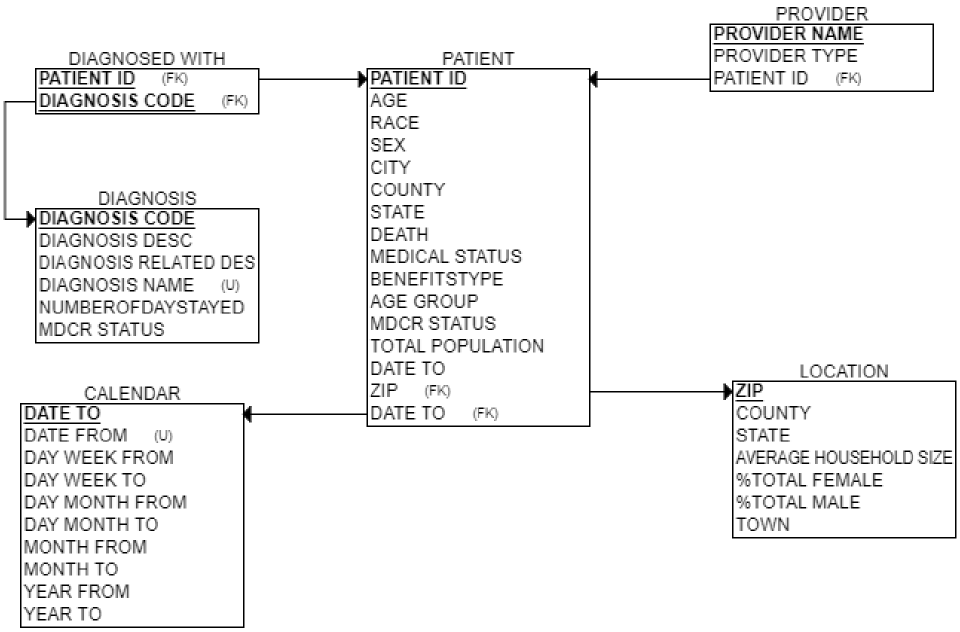
After exploring the analysis of the dimensions and the fact table during the ETL process in Alteryx, we came up with the Entity Relationship Diagram, Relational Schema as well as the Fact Table/ Dimensional Model with the data from the Medicare Shared Savings Program. Our group designed dimensions for the patient, location, diagnosis, provider, and calendar. These dimension tables are also designated from the basic information-gathering, which are who, what, when, where, and by whom. The five dimensions effectively answered our project questions regarding the patient readmission rate in conjunction with multiple health care outcomes.

The patient dimension contains all the information that is directly associated with an individual. PatientID, for example, can directly identify the individual as the person for whom the service or treatment is intended. Other variables such as zip code, city, state, age, sex, race, etc. are all identifiable information about a specific patient. The provider dimension holds the information about a provider, which in this case contains only the provider name and provider type; the provider name is also unique in the provider table. The location dimension contains detailed information like zip, town, county, and state. Because both the location and patient entities have information regarding zip, it is acted as a foreign key in the patient table. The calendar dimension consists of detailed time-related information, like day, month, and year. Last but not least, the diagnosis dimension contains two unique values, which are diagnosis name and diagnosis code. According to the relational schema down below, both PatientID and diagnosis code are foreign keys in the Patient table. The ER diagram, as well as the relational schema, exemplified the relationship between all the dimension tables and primary keys. Further explanations of each dimension table will be analyzed in the ETL models.

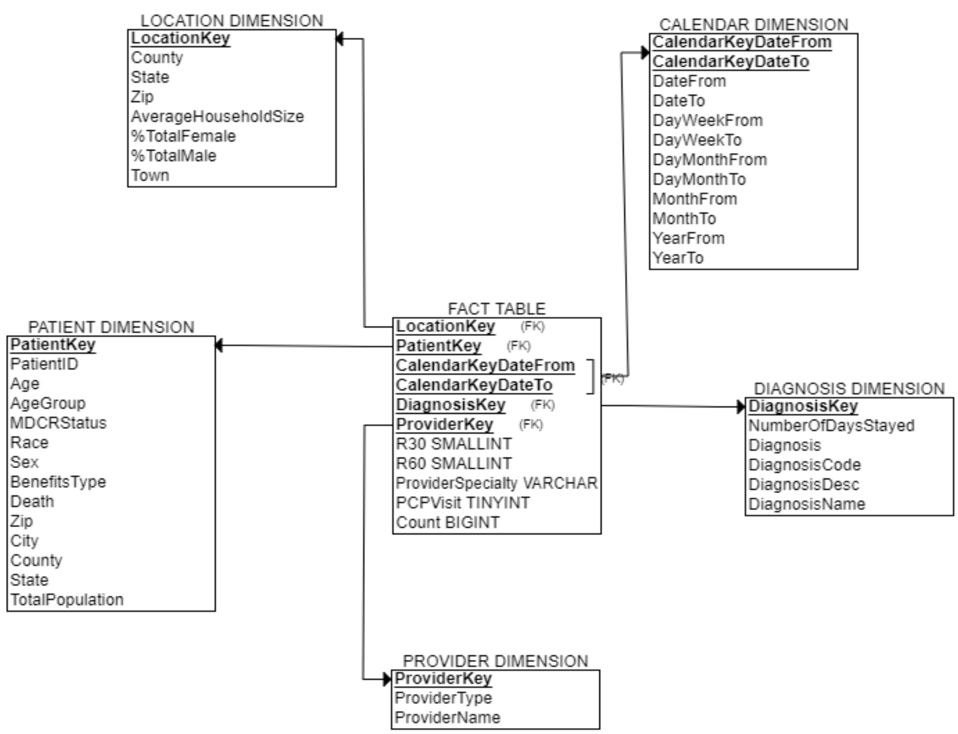
ER Diagram



Relational Schema



Fact Table/ Dimensional Model

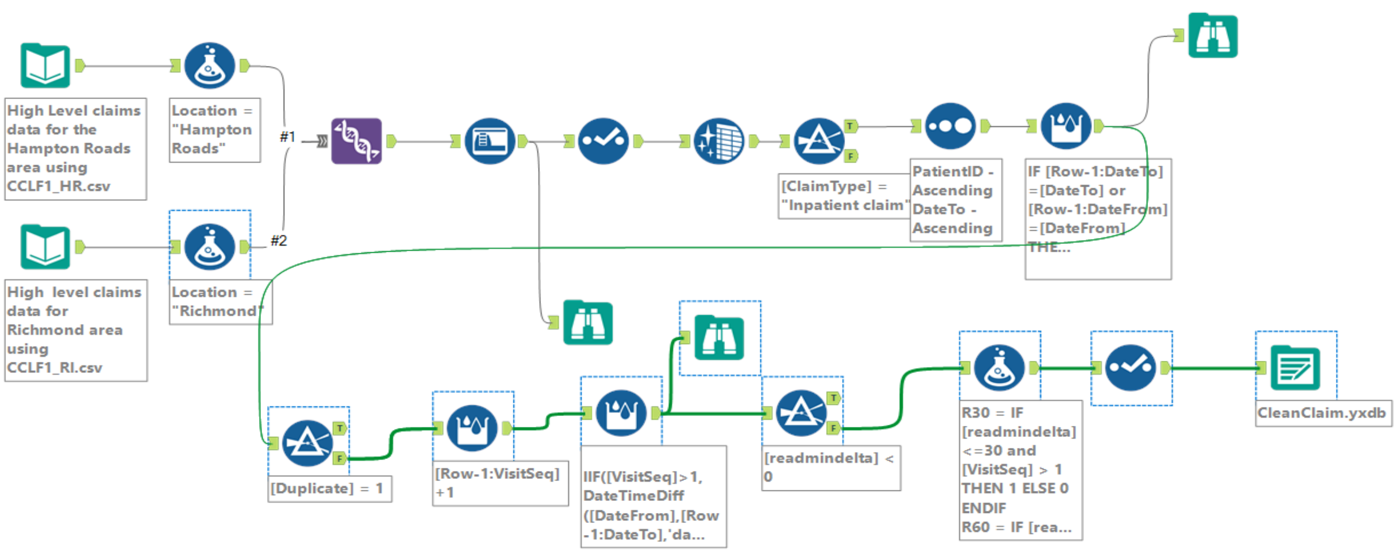


ETL Approach

Our group implemented the Extract Transform and Load (ETL) process to collect the patient data in multiple types of sources and extract the data from the previous form (Excel Sheet) to another database (Alteryx). Lastly, we loaded the data from Alteryx into the target database or data warehouse in MySQL. In our ETL approach, we have 5 different dimensions, which are patient, provider, diagnosis, location, and calendar. We separated the calendar dimension into two primary keys, which are "calendar from", and "calendar to".

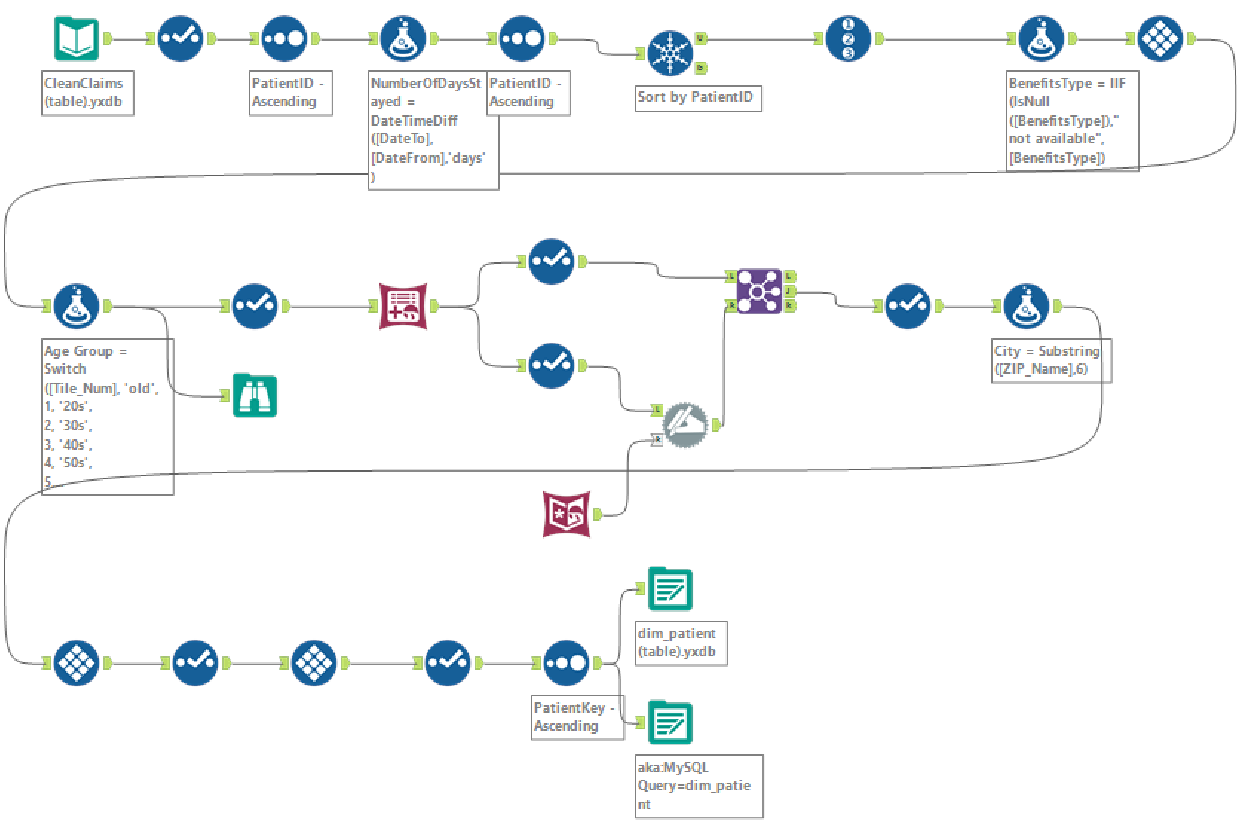
1) **<u>Filtering and Cleansing the High-Level Claim Data</u>**

We started off by cleaning the high-level data to "CleanClaim" including removing additional columns, filtering, renaming, and removing duplicates. The high-level data consists of both Hampton Roads and Richmond's high claim data and we filtered inpatient claim as the claim type. Our group was actively trying to analyze the variables and seeking for possible dimensions while cleansing the data; we decided to use a formula to create an R30 (Readmission 30) and R60 (Readmission 60) rate for our fact table later on.



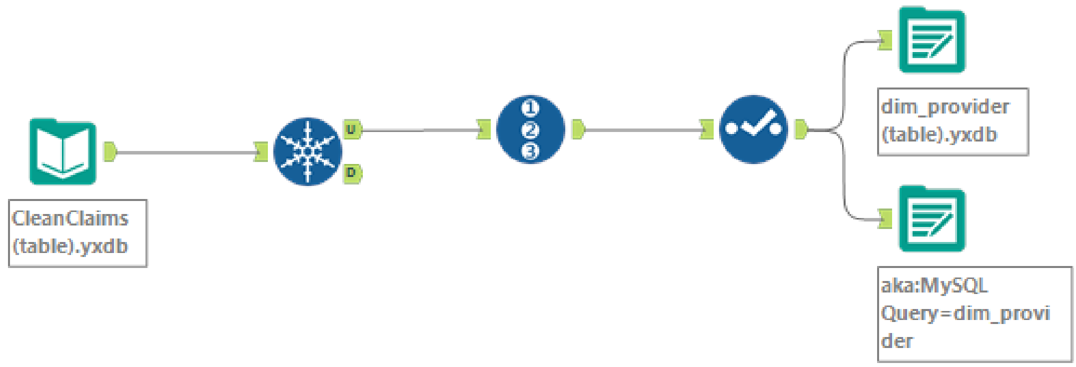
2) **Patient Dimension**

Next, we used the high-level clean claim data to create a patient dimension table sorted by PatientID. We generated an Age Group with an increment of 10 from the group of patients aging from 20 to 100 and over. The raw data from the Centers for Medicare and Medicaid Services (CMS) has null or false information for the city, county, zip, etc. In order to analyze complete and accurate data, we imported the US Census 2010 data from the demographic analysis tool in Alteryx to retrieve and filter the right information. We created smart tiles for Total Population and Med Age Group and ascended PatientKey at the end to select useful information for the patient dimension.



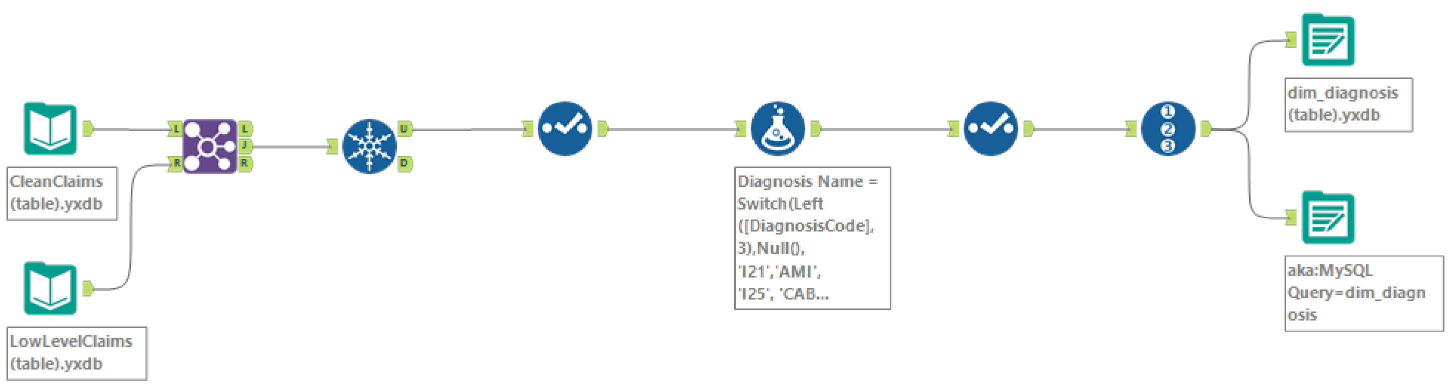
3) **Provider Dimension**

The provider dimension started off with the high-level clean claim data with a unique value of "provider". We identified provider (renamed to "provider name") and "provider type" and created a ProviderKey column in the provider dimension table.



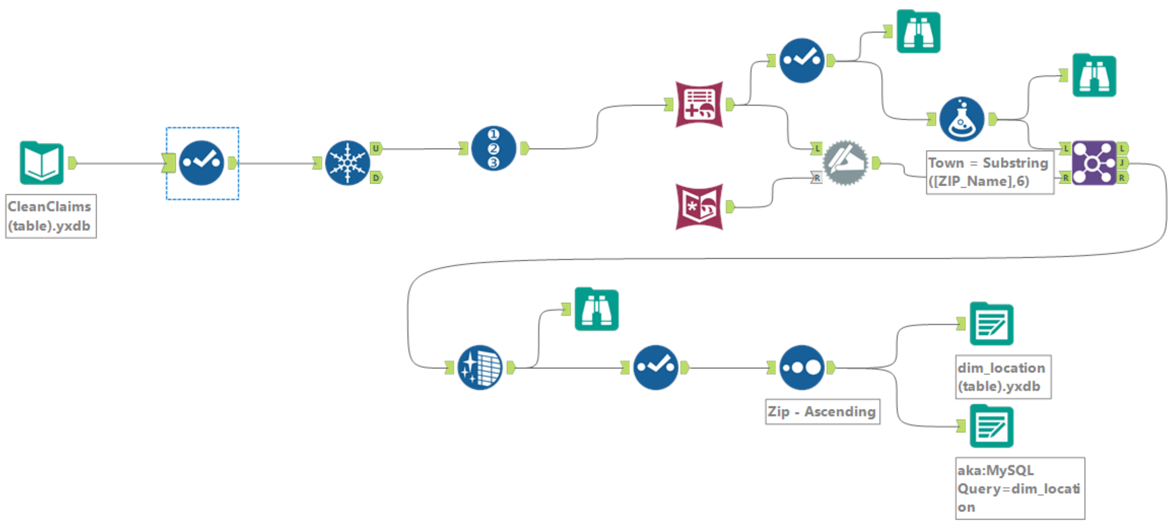
4) **Diagnosis Dimension**

We imported both the high-level clean claim data and low-level claim data for the diagnosis dimension. This dimension is joined by PatientID and contains unique values for "diagnosis code" and "diagnosis description". We created a formula for the diagnosis name and grouped it into five categories that include acute myocardial infarction (AMI), Coronary artery bypass grafting (CABG), Stroke, Chronic obstructive pulmonary disease (COPD), and Heart Failure (HF).



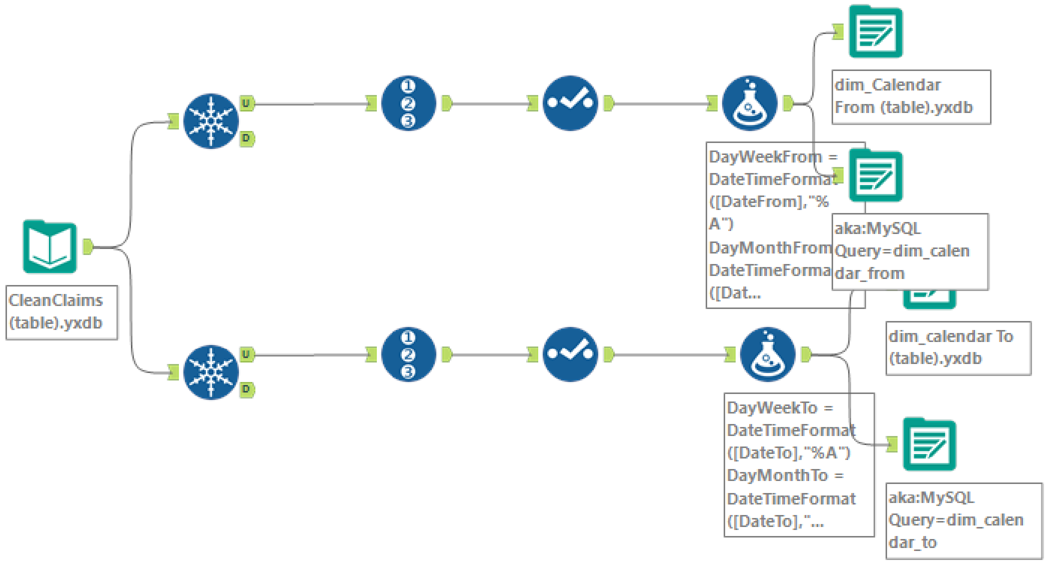
5) **<u>Location Dimension</u>**

Our location dimension is imported from the high-level clean claim data, and the unique value is being sorted by "Zip". Then, we used the demographic analysis tool in Alteryx and imported the data from the US Census 2010 in order to match the county, town, and zip code simultaneously without null values and duplicates. We retrieved the data from zip, average household size, % Total Female & Total Male in Total Population, and we sorted zip code in ascending order.



6) **<u>Calendar Dimension (From & To)</u>**

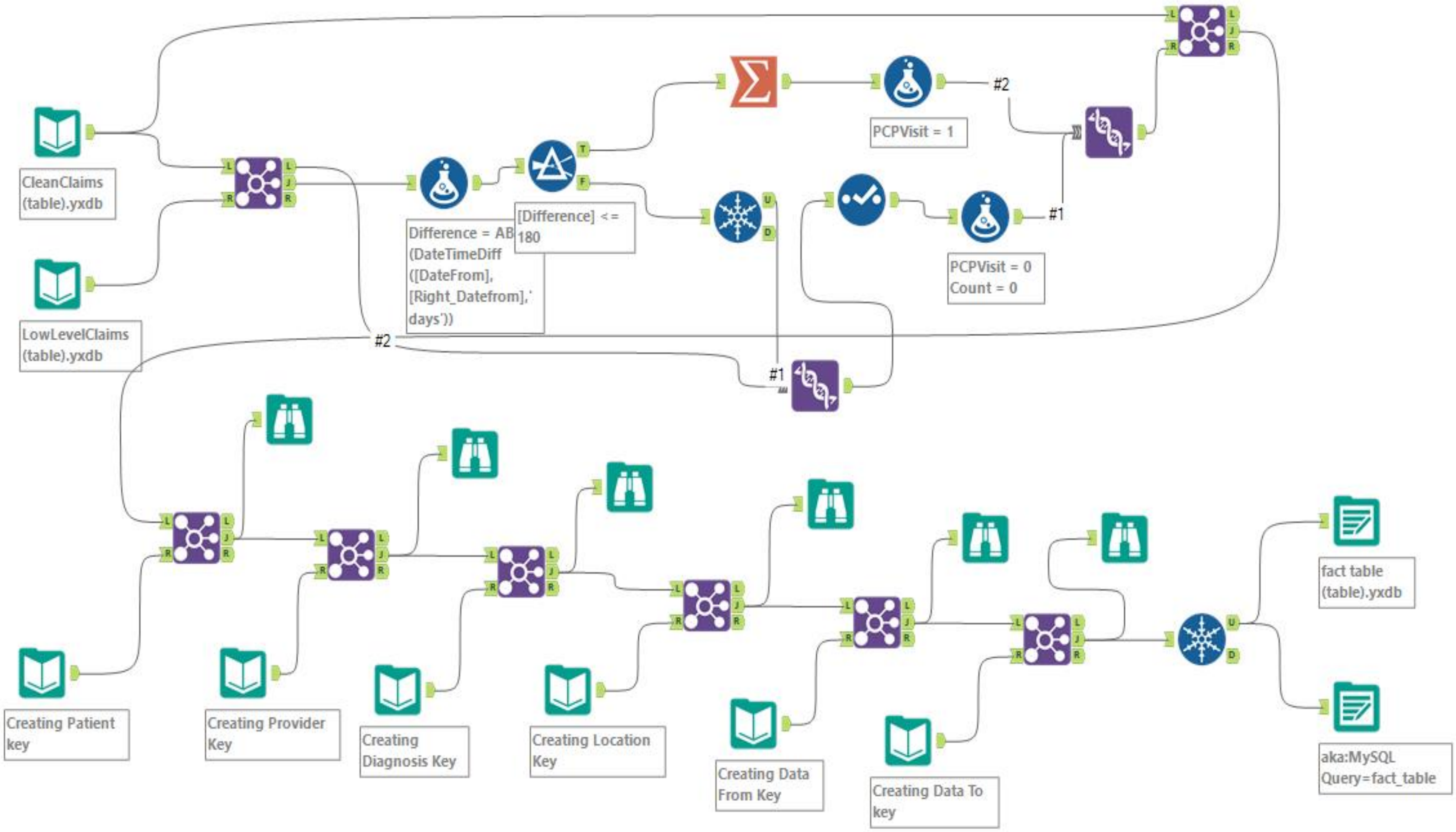
The calendar dimension table is a tricky one because we did not want to have two calendar keys (CalendarKeyDateFrom and CalendarKeyDateTo) at first. We attempted to use the calendar key directly imported from MySQL code, but the full date does not match the "date from" where the patient was first admitted into the hospital. We considered using a fuzzy match as well, but too much data was missing. We ended up with two calendar keys and decided to put "calendar data from" and "calendar date to" into two separate dimension tables.



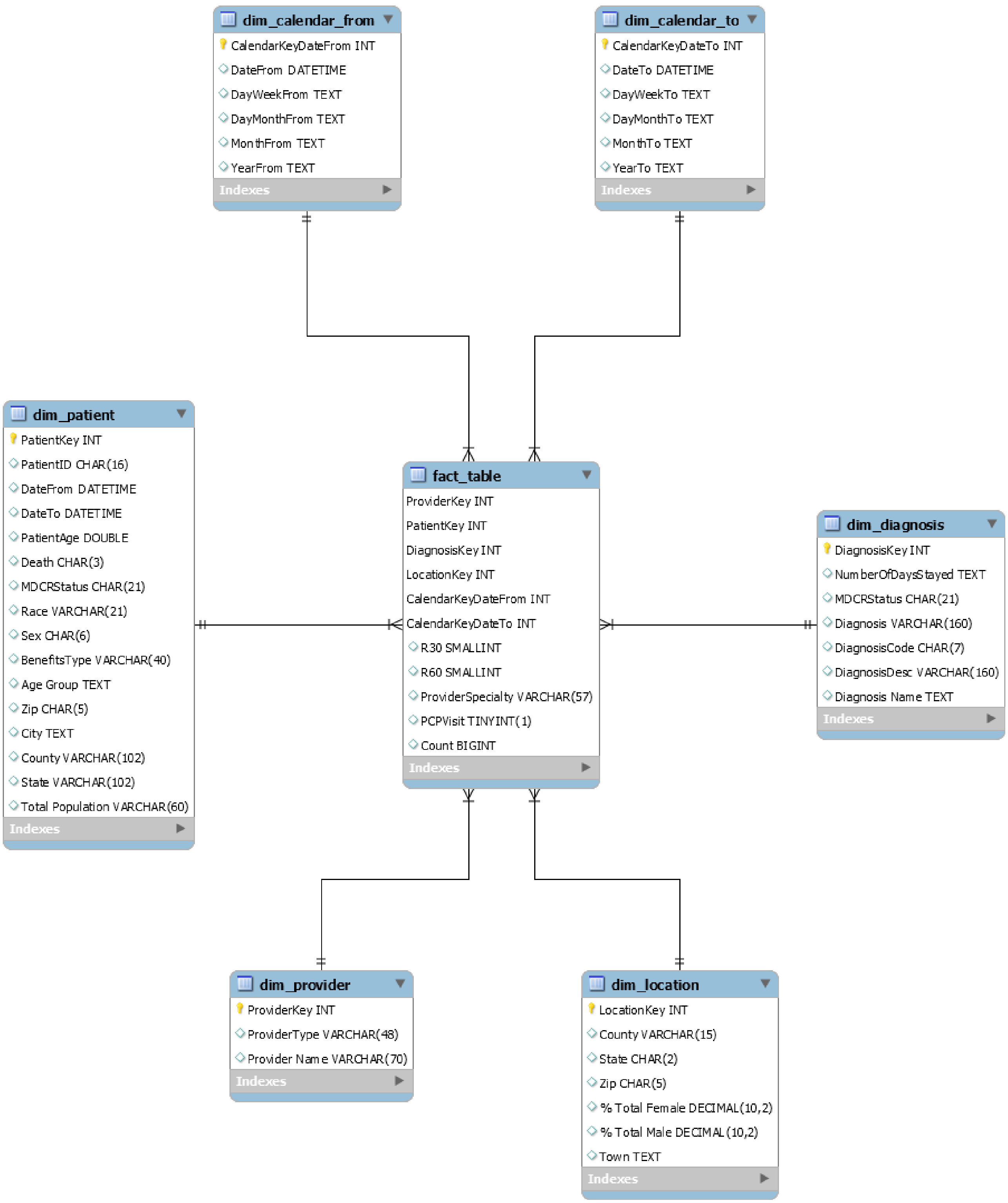
7) <u>Fact Table & Star Schema</u>

Our fact table was joined by PatientID with both the high-level clean claim and the low-level clean claim at the beginning. Then we connected the PCP fact table with six of our primary keys using PatientID, Date From, and Date To. The fact table contains surrogate keys to every single dimension and we set all the primary keys as unique in the last step. In addition, we added R30, R60, PCPvisit, ProviderSpecialty as well as Count in the fact table, and further explanation will be analyzed in the queries and visualizations section. Lastly, we loaded the data to MySQL and used the reverse engineer tool in MySQL to convert the fact table into a star schema.

Fact Table - Alteryx



Star Schema



dim_calendar_from
CalendarKeyDateFrom INT
DateFrom DATETIME
DayWeekFrom TEXT
DayMonthFrom TEXT
MonthFrom TEXT
YearFrom TEXT
Indexes
dim_calendar_to
CalendarKeyDateTo INT
DateTo DATETIME
DayWeekTo TEXT
DayMonthTo TEXT
MonthTo TEXT
YearTo TEXT
Indexes
dim_patient
PatientKey INT
PatientID CHAR(16)
DateFrom DATETIME
DateTo DATETIME
PatientAge DOUBLE
Death CHAR(3)
MDCRStatus CHAR(21)
Race VARCHAR(21)
Sex CHAR(6)
BenefitsType VARCHAR(40)
Age Group TEXT
Zip CHAR(5)
City TEXT
County VARCHAR(102)
State VARCHAR(102)
Total Population VARCHAR(60)
Indexes
fact_table
ProviderKey INT
PatientKey INT
DiagnosisKey INT
LocationKey INT
CalendarKeyDateFrom INT
CalendarKeyDateTo INT
R30 SMALLINT
R60 SMALLINT
ProviderSpecialty VARCHAR(57)
PCPVisit TINYINT(1)
Count BIGINT
Indexes
dim_diagnosis
DiagnosisKey INT
NumberOfDaysStayed TEXT
MDCRStatus CHAR(21)
Diagnosis VARCHAR(160)
DiagnosisCode CHAR(7)
DiagnosisDesc VARCHAR(160)
Diagnosis Name TEXT
Indexes
dim_provider
ProviderKey INT
ProviderType VARCHAR(48)
Provider Name VARCHAR(70)
Indexes
dim_location
LocationKey INT
County VARCHAR(15)
State CHAR(2)
Zip CHAR(5)
% Total Female DECIMAL(10,2)
% Total Male DECIMAL(10,2)
Town TEXT
Indexes

Queries and visualizations

Query 1 and 2

Name of Diagnosis	Readmission Times in 30 days	Readmission Rate 30
HF	241	0.2263
AMI	303	0.2045
COPD	268	0.1413
CABG	45	0.1301
Stroke	54	0.1095

Name of Diagnosis	Readmission Times in 60 days	Readmission Rate 60 Days	Number of Days Stayed
HF	353	0.3315	5.25
COPD	458	0.2414	4.96
AMI	344	0.2321	4.81
CABG	45	0.1301	6.25
Stroke	61	0.1237	3.87

Explanation 1 &2:

From earlier, we discussed the surprisingly high readmission rate in Virginia and wanted to compare that to our preconceived notions. It turns out that we were right. Diseases that often stem from pre-existing conditions (such as heart failure, COPD, and AMI) are correlated with an uptick in readmission rate for 30 days, significantly if the time increment changed to 60 days. Meaning, while stressing checkups with a PCP within the first 30 days is essential, it is even more vital to follow up in the second month. A possible implementation of this is more frequent phone calls and reminders from a physician's assistant.

Query 3

Age Group	Readmission Rate 30 Days	Readmission Rate 60 Days	Population in Age Bin	AVG(b.NumberOfDaysStayed)
40s	0.2511	0.3638	940	5.246808510638298
50s	0.1988	0.2836	2133	4.9976558837318334
80s	0.1829	0.2634	10122	4.760225251926497
70s	0.1747	0.2205	11943	4.66767143933685
90s	0.1677	0.2359	3722	5.009403546480387
60s	0.1546	0.2387	5881	4.639177010712464
30s	0.0280	0.0327	214	5.4672897196261685

Explanation 3:

The results of query three were arguably one of the more unexpected queries we did. While the 40s/50s age bin is a noticeably smaller sample, the high readmission rates for 40-year-olds and 50-year-olds aren't intuitive. Our group hypothesized that medical conditions that land 40-year-old and 50 years olds in the hospital were more serious, explained by the more extended average day stayed. More severe infections are more likely to be readmitted, as shown in the previous two queries.

Query 4

Diagnosis Name	Average Patient Age	Total Deaths	Total Cases	Morbidity Rate	Readmission Rate 30 Days	Readmission Rate 60 Days
HF	79.83	588	1065	0.5521	0.2263	0.3315
Stroke	77.29	106	493	0.2150	0.1095	0.1237
COPD	75.25	347	1897	0.1829	0.1413	0.2414
AMI	77.51	188	1482	0.1269	0.2045	0.2321
CABG	74.04	7	346	0.0202	0.1301	0.1301

Explanation 4

Stroke data from query 4 helps examine different solutions to reduce readmission rates. Despite having a high morbidity rate, strokes have a comparatively low readmission average - on par with hospitals' general readmission rate. The American Stroke Association cited that this is likely to happen to patients going back to monitored environments where urinary tract infection incidents are less likely.

Query 5

Provider Name	Readmission Rate 30 days	Readmission Rate 60 days	Average Age	Sample of Patients
SOUTHSIDE REGIONAL MEDICAL CENTER	0.2522	0.2611	63.26	226
SENTARA NORFOLK GENERAL HOSPITAL	0.2391	0.2877	70.22	1213
CJW MEDICAL CENTER	0.2127	0.2265	72.05	2102
MEDICAL COLLEGE OF VIRGINIA HOSPITALS	0.1942	0.3033	63.05	788
BON SECOURS ST. FRANCIS MEDICAL CENTER	0.1911	0.2536	76.01	2973
BON SECOURS MEMORIAL REGIONAL MEDICAL...	0.1763	0.2585	78.81	3238
HENRICO DOCTORS' HOSPITAL	0.1702	0.1790	81.93	1939
SENTARA OBICI HOSPITAL	0.1547	0.2960	76.95	1713
RIVERSIDE TAPPAHANNOCK HOSPITAL INC	0.1437	0.1437	77.53	174
SENTARA LEIGH HOSPITAL	0.1427	0.2133	73.88	2039
BON SECOURS MARYVIEW MEDICAL CENTER	0.1230	0.1707	76.32	4504
RAPPAHANNOCK GENERAL HOSPITAL	0.1137	0.2298	78.24	853
BON SECOURS ST MARY'S HOSPITAL	0.1035	0.1779	80.15	4262
SENTARA VIRGINIA BEACH GENERAL HOSPITAL	0.0808	0.1394	75.15	990
BON SECOURS DEPAUL MEDICAL CENTER, INC.	0.0782	0.1513	72.48	1381

Explanation 5

For query 5, it was essential to check that there were no hospitals with alarming patient track records. While Southside Regional Medical Center did have a high readmission rate, the smaller number of patients could have skewed the information.

Query 6

ESRD Status	Avg. PCP Visits Previous 6 Months	Readmission Rate 30 days	Readmission Rate 60 days
Aged without ESRD	0.3833	0.1715	0.2339
Disabled without ESRD	0.3878	0.1552	0.2326
Disabled with ESRD	0.6120	0.3417	0.5000
Aged with ESRD	0.5143	0.2532	0.4315
ESRD only	0.1731	0.1218	0.1731

Explanation 6

We used the information we had gathered about stroke patients' low readmission rate and tried to see if chronic disease sufferers would be more or less likely to go to the hospital considering they would due to more /less potential complications. The data indicate that people who suffer from chronic illnesses like ESRD are no more likely to go to a PCP or get admitted into the hospital unless they have a conflating factor. The possible takeaway, complications to people with preexisting conditions could be compounded heavier and could lead to higher readmission into the hospital than other individuals with the same circumstances.

Query 7

Readmission Rate 30 Days	Readmission Rate 60 Days	Total Population	Avg. PCP Visits Previous 6 Months	Morbidity Rate
0.2333	0.3072	Below Average (5100 to 13250)	0.4941	0.2187
0.1898	0.2556	Above Average (34500 to 90000)	0.4111	0.2405
0.1655	0.2292	Average (13250 to 34500)	0.3710	0.2350
0.1627	0.2384	Low (1975 to 5100)	0.3542	0.2796
0.0795	0.1843	Extremely Low (Below 1975)	0.3593	0.1659

Explanation 7

Going back to one of our original ideas, we thought a potential healthcare desert could cause a higher hospital readmission rate while simultaneously seeing a decrease in the average appointments with PCPs. We selected the morbidity rate in this query to further assess potential readmission. It turns out the hypothesis was only partially right. Yes, PCP visits did fall, but so did readmission rates to the hospital. We are currently unsure why the lower populated area had lower (or around the same) readmission to hospitals. It is possible that a lack of a PCP does not indicate a person is more likely to be admitted to the hospital; however, the rest of the data collected directly conflicts with that notion.

Query 8

Benefits 1-Yes 0-No	Morbidity Rate	Average Patient Age	Avg. PCP Visits Previous 6 Months
1	0.1517	52.82	0.4548
0	0.1982	58.30	0.6037

Explanation 8

Query 8 is the most central to our recommendation. We used two SELECT CASE WHENs in conjunction with a WITH JOIN query to create the binary data points for morbidity and benefits. If our data are correct, it indicates an apparent discrepancy in inpatient morbidities and access to primary care physicians. To run the query, we decided the group needed to be working-age when they would be off parents' health insurance, likely relying on employer provided health insurance or not having any at all. Our recommendation is to provide a safety net for the people without access to insurance in this age demographic to avoid heart failure, AMI, or other preventable conditions. Skirting around diseases like that will save the person a lot of medical hardship, but it could also reduce hospital readmission down the line and shrink the enormous health care cost.

Query 9

MDCRStatus	Morbidity Rate
Aged with ESRD	0.3000
ESRD only	0.2727
Aged without ESRD	0.1817
Disabled without ESRD	0.1037
Disabled with ESRD	0.0357

Explanation 9

This code was performed in reference to query 6 and examines if Age and ESRD do correlate with morbidity rate, unsurprisingly, they are linked.

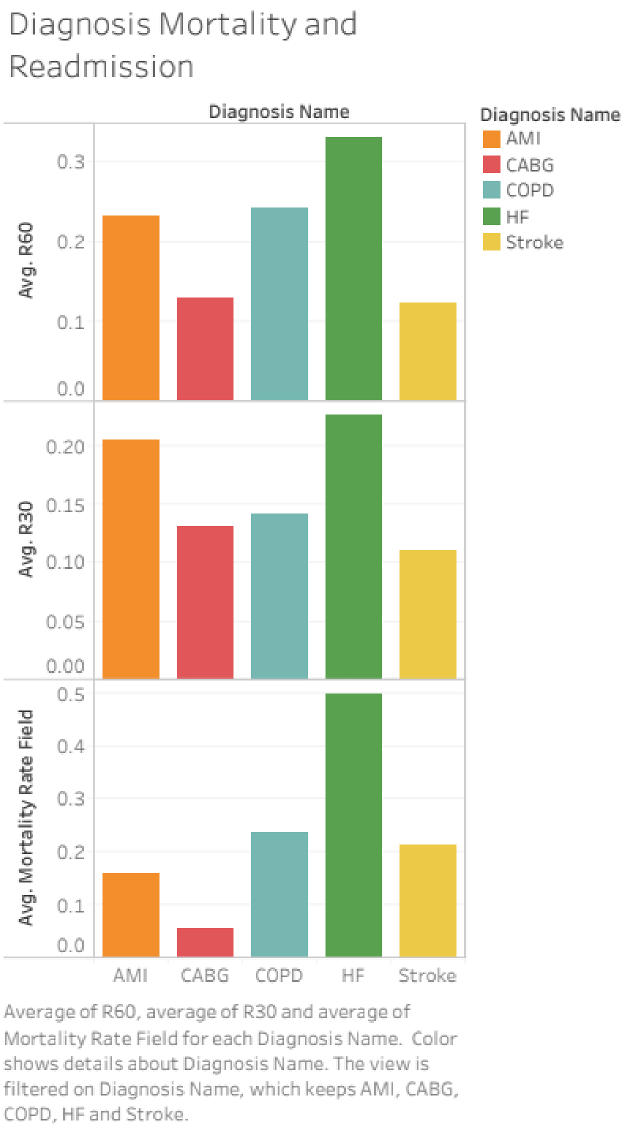
Query 10

Diagnosis Name	Avg. PCP Visits Previous 6 Months	Total Claims for Provider
COPD	0.5018	1897
HF	0.4657	1065
AMI	0.4177	1482
CABG	0.3757	346
Stroke	0.2677	493

Explanation 10

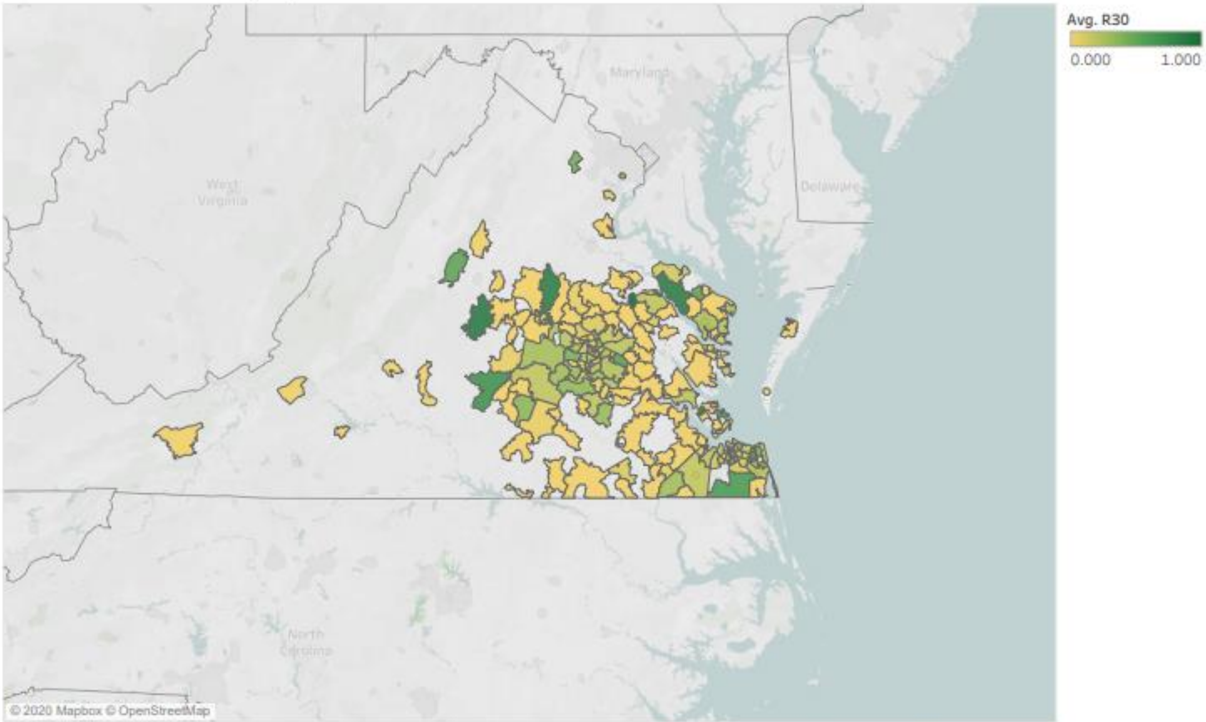
Query 10 was an examination of the amount of claims for each provider, a reference to what medical care centers were seeing most frequently.

Tableau Visualization



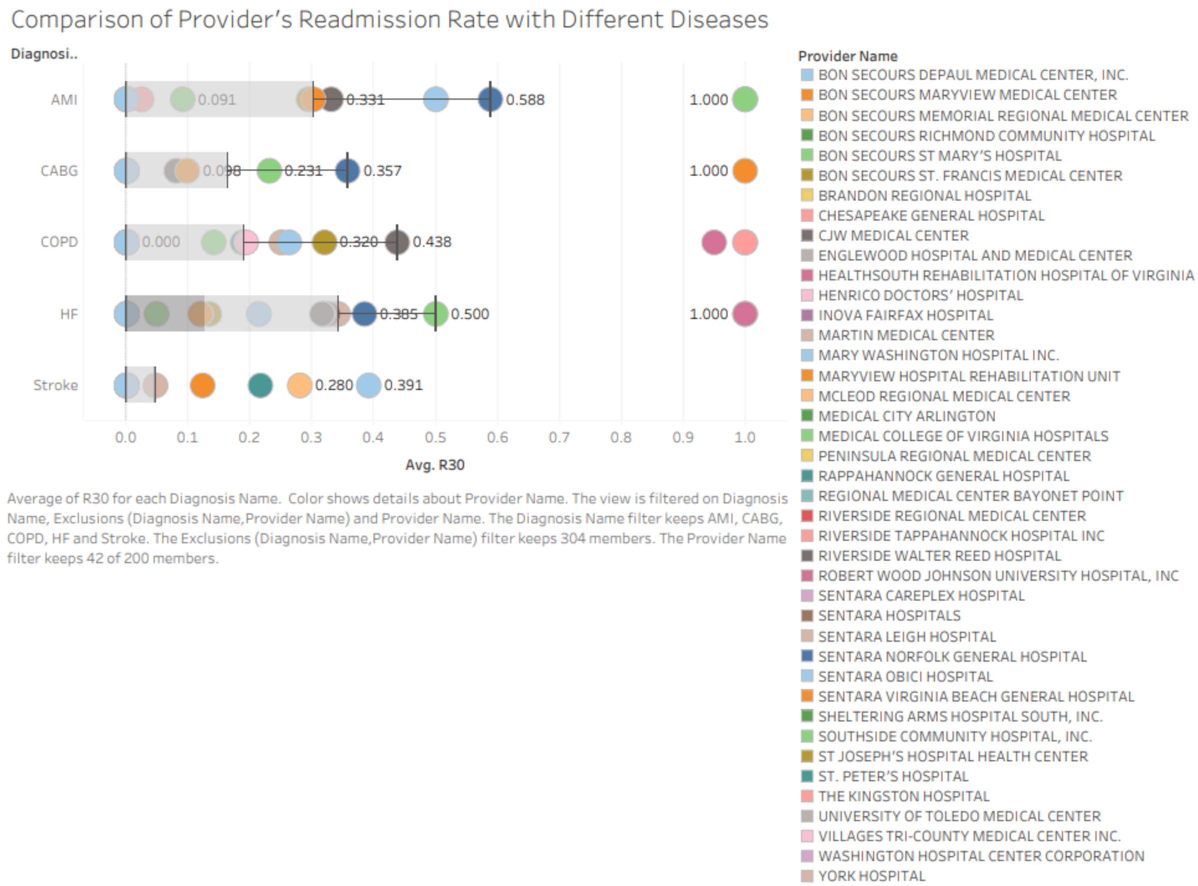
The graph is depicting the modality rate, average 30-day readmission rate, as well as the average 60-day readmission rate of the 5 diseases listed above. It is clearly shown that heart failure has the highest readmission rate and the mortality rate, followed by AMI. CABF in general has the lowest mortality rate, and has relatively low readmission rate as well.

Readmission Rates by Zip Code



Map based on Longitude (generated) and Latitude (generated). Color shows average of R30. Details are shown for Zip (dim location). The data is filtered on State (dim location), which keeps VA. The view is filtered on Zip (dim location), which keeps 324 of 324 members.

This is a map from Virginia that indicates the patients' 30-day readmission rate by zip code. The darker color reflects higher readmission rate that can go as high as 100%, and vice versa. Compared to the previous graph regarding the provider's 30 days readmission rate with different diseases, this graph can clearly demonstrate the location of the providers based on the longitude and the latitude of the region. This information is helpful to sort out the location of the provider, where patients can compare apples to apples and decide which providers to go to.



This graph depicts the 30-day readmission rate with different diseases. AMI has the highest 30-day readmission rate compared to the other 4 diseases listed, followed by HF (Heart Failures). Stoke in general has the lowest readmission rate within the first 30 days among different providers. As shown on the graph, there are several providers where the 30-day readmission rate for AMI, CABG, COPD, and HF are 100%. This indicates our assumption from the introduction that despite having some of the best health care professionals operate successful surgeries, many of the recovery processes involved are known to be quite expensive because of the high readmission rates.