Hospital Quality of Care and Costs Across the U.S.

DS4A | Team #37

Lucy Du, Rennah Weng, Sofia Posada, Thao Vy Vuong, Yuqiu Dong (Jennifer)

Project Report

1) Introduction

1.1 Business Problem

For the past few decades, polls have shown that Americans are dissatisfied with their current health care system. There is no question the United States lags behind other countries in healthcare quality, yet holds the highest position in healthcare spending. In the U.S., per-capita spending from private sources (e.g. private health insurance premiums) at $4,092 per capita is higher than the OECD average. This private spending is more than five times higher than Canada, the second-highest spender, while on the lower end, in Sweden and Norway, private spending was less than $100 per capita. Given the evidence Americans are already spending more on healthcare than other first-world countries, it is clear this cost should be minimized as much as possible for a patient in the U.S.

Besides cost, healthcare also varies by accessibility and quality, especially for marginalized communities. The Affordable Care Act led to large coverage gains, but political debate threatens this legislation and the millions of Americans benefiting from it. Additionally, some groups still remain at higher risk of being uninsured, lacking access to care, and experiencing worse health outcomes. For example as of 2018, Hispanics are 19% likely to be uninsured compared to 7.5% for Whites, and individuals with incomes below poverty are 17.3% likely to lack coverage than individuals with incomes at 400% of the federal poverty level that are 4.3% likely. Since different hospitals across the U.S. can have different costs for performing the same procedure, patients should be directed to locations where they can receive both quality and affordable treatment, which is especially critical for lower socioeconomic and disadvantaged populations.

Addressing disparities in health care is essential to solving inequity, but letting these issues go unaddressed are also costly to society. Clear disparities exist and consequences of unaffordability result in fewer preventive services, poorer health outcomes, higher mortality and disability rates, and lower annual earnings because of sickness and disease. A 2018 analysis by Altarum Institute estimates that healthcare disparities amount to $93 billion in excess medical care costs later on, and $42 billion in lost productivity per year. Given the relevancy, widespread

prevalence, and urgency of this issue, we must address it through various avenues: new legislation or policy, structural hospital-healthcare-profit reform, and societal commitment to solving equity issues. On a patient level, we must also minimize an individual's spending, while maximizing care.

1.2 Business Impact

The goal of our project is to increase exposure and accessibility to medical resources and hospital data, thereby allowing patients to make informed decisions about their healthcare. We envision the healthcare system to shift its focus back to what it always should have been on — the patients. By empowering the everyday people, who are unfamiliar with the various choices of hospital resources out there, we hope that each and every individual and their family members can acquire the best and most cost-effective healthcare to suit their unique needs. The impact we plan to make is the ability to provide ample amounts of information upfront, before the patient receives care, to eliminate the guesswork and to decrease the readmittance of patients to different hospitals. This will save time and money for both patients and hospitals, and more importantly, reduces the chance of harm done on the patient by the delay of proper care.

Another impact we aim to have is by collecting racial data of hospitals' patients, we hope to investigate the disparity of healthcare in low-income communities and raise the awareness of potential healthcare inequality in order to encourage others to help make an impact on this social issue as well. We also want to quantify if race potentially affects the choice of hospitals and medical bills patients receive.

In order to realize the impact we envision, the project will be able to provide policy makers with insights about the current status of the healthcare system, as well as answer questions that patients typically have about their healthcare. Specifically, our research questions include:

- Are there disparities in cost and quality of healthcare in different geographic and socioeconomic areas?
- Which of the features have the most impact in predicting charges (cost of care)?
- Which of the features have the most impact in predicting length of stay (one indicator of quality of care)?
- Given a particular health condition, which hospital should you visit?

Our project will utilize both visualizations and modeling to answer the above questions. Specifically, we will implement linear regression and classification models to investigate the factors influencing cost and quality of care.

2) Data Analysis & Computation

a) Datasets + Data Wrangling & Cleaning:

a.1) Data Description

- We use the "***Hospital Inpatient Discharges in New York state***" dataset collected by the New York State Department of Health in 2017 for the first milestone. With time permitting, NYS Department of Health data for multiple years will be used for additional milestones (Please refer to the Milestone section).
- This dataset contains de-identified discharge level detail on patient characteristics, diagnoses, treatment, services, and charges. The de-identified data file does not contain data that is protected health information (PHI) under HIPAA. The health information is not individually identifiable; all data elements considered identifiable have been redacted.
- Details of the dataset:
    - 2.54M rows
    - 34 columns
    - 224 unique hospitals in NY state
    - Each row represents a patient

a.2) Data Wrangling process and methodology

This dataset has the following main issues:

- Missing values
- Incorrect data type
- Redundant features
- Lack of data on household income

a.2.1) Missing values

The following variables have missing values (in descending order of # of missing values, or NAs):

Variable	**% NAs**	**Reason for NAs**	**What have done**
Birth Weight	90%	Unknown	Leave it as it is. But suggest to remove because the number of NAs is too high
Payment Typology 3	74%	Maybe people don't use this, so there are no info	Leave it as it is. But suggest to remove because the number of NAs is too high. In addition, we

			can use ‘Payment Typology 1’: which has no NAs
Payment Typology 2	37%	Maybe people don’t use this, so there are no info	Leave it as it is. But suggest to remove because the number of NAs is high. In addition, we can use ‘Payment Typology 1’: which has no NAs
Zip Code - 3 digits	1.6%	Official explanation for documentation: - population size less than 20,000 - abortion records, or - cell size less than 10 on population classification strata. “OOS” are Out of State zip codes.	- Assume that patients in the same neighborhood will be likely to visit the same hospital, so we used the zip code of patients who also visit the same hospital to impute for NAs. - For those zip codes that have been redacted for confidentiality': We can't do anything about this. Remove the rows that have NAs (about 5150 rows ~) for modeling purpose only
Hospital Service Area	0.2%	Official explanation for documentation: data elements considered identifiable have been redacted	We can't do anything about this. Remove the rows that have NAs (about 5150 rows) for modeling purpose only
Hospital County	0.2%	Official explanation for documentation: data elements considered identifiable have been redacted	We can't do anything about this. Remove the rows that have NAs (about 5150 rows) for modeling purpose only
Operating Certificate Number	0.2%	Official explanation for documentation:: data elements considered identifiable have been redacted	We can't do anything about this. Remove the rows that have NAs (about 5150 rows) for modeling purpose only
Permanent Facility Id	0.2%	Official explanation for documentation:: data elements considered identifiable have been redacted	We can't do anything about this. Remove the rows that have NAs (about 5150 rows) for modeling purpose only

APR Severity of Illness Description	< 0.01%	Unknown	Drop rows with missing values for columns such as the APR Severity of Illness Description and APR Risk of Mortality since these have a relatively small number of missing values and thus will be safe to drop.
APR Risk of Mortality	< 0.01%	Unknown	Drop rows with missing values for columns such as the APR Severity of Illness Description and APR Risk of Mortality since these have a relatively small number of missing values and thus will be safe to drop.

a.2.2) Incorrect data type

Some variables have inappropriate data types, especially for modeling. The 2 main issues are:

- Numeric variables:
    - Issue: Variables which should have been numeric are strings because they contain special characters like "120 +" (longer than 120 days), or "UNKN" (stands for Unknown).
    - Solution: Approximate that longer than 120 days is 120 days, remove special characters from the string, coerce the UNKN to null values, and convert to numeric values.
- Categorical variables:
    - These variables are either in numeric or string type.
    - We do label encoding or one hot encoding
    - Prepare categorical variables for future modeling

Variable	**Old data type**	**New data type**
Length of Stay	object	numeric
Birth Weight	object	numeric
Operating Certificate Number	numeric	string
Permanent Facility Id	numeric	category
CCS Diagnosis Code: only for modeling purpose	Numeric (int)	category
CCS Procedure Code	Numeric (int)	category

APR DRG Code	Numeric (int)	category
APR MDC Code	Numeric (int)	category
APR Severity of Illness Code	Numeric (int)	category
APR Medical Surgical Description	Create new column to label encode them (to convert the variable into category) for modeling	category New name: ***apr_medical_surgical_code***
Age Group	Create new column to label encode them (to convert the variable into category) for modeling	category New name: ***age_group_code***
APR Risk of Mortality	Create new column to label encode them (to convert the variable into category) for modeling	category New name: ***apr_risk_mortality_code***
Payment Typology 1	Create new column to encode them (to convert the variable into category) for modeling	category New name: ***payment_typology_1_code***

a.2.3) Redundant features

- Remove the variable in the pair of variables that have the same meaning:
    - One of them uses words to describe, the other one uses numbers to represent the words.
    - For example, the pair "APR Severity of Illness Description" and "APR Severity of Illness Code": these two variables have the same meaning of representing the severity of illness.
    - Before removing, we checked if they have the same unique values to make sure they represent the same thing.
- Remove unnecessary features for modeling

a.2.4) Lack of data on household income

- Issue: Because we want to see if there is a disparity in cost and quality of care among different socioeconomic levels, household income is a great socioeconomic indicator, so we want to add data on household income to our dataset.

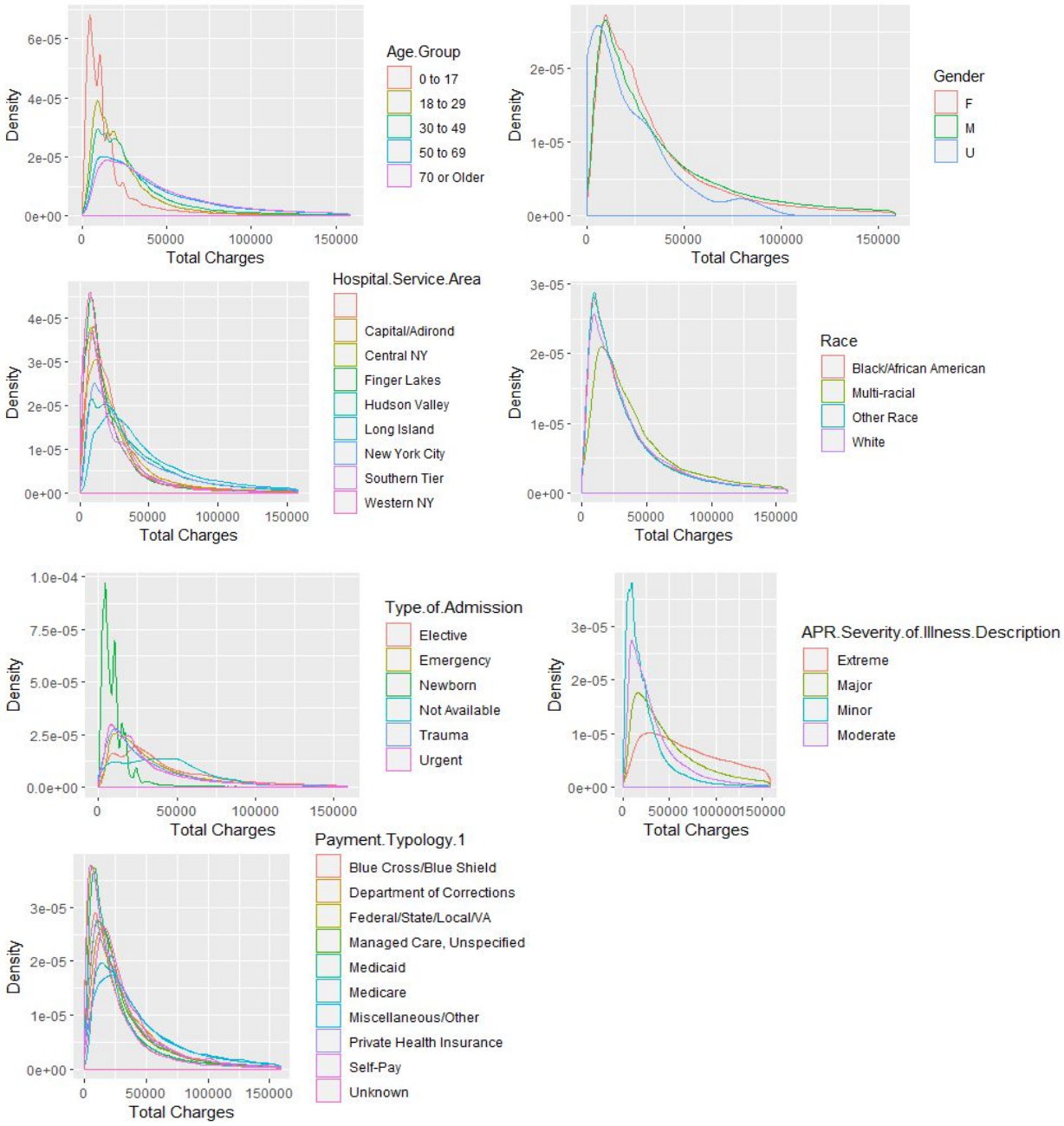
- Solution:
    - One possible solution is to look at the zip code. In NY, some zip codes are known to be higher income neighborhoods while others are known as low-income neighborhoods. Under the assumption that patients usually go to hospitals that are most nearby their houses, introduce another dataset(s) that has information on the average income for each zip code and map that to each patient's socioeconomic levels.
    - One source of data: Median household income by zip code throughout the US from the Institute of Social Research at the University of Michigan (they prepared the data from Census Bureau, time frame 2006 - 2010).
        - This dataset has the full 5-digit zip code, but the zip codes in our dataset only have the first 3 digits (for de-identifying purpose)
        - We extracted the first 3 digits of the income dataset, grouped by these 3 digits, and found the median income.
        - Then we merged with our patient dataset by the first 3 digits.

b) Exploratory Data Analysis:

Through visualizations, we want to explore descriptive and distributional statistics to better understand our data sample, and if insights can be drawn for our previous questions.

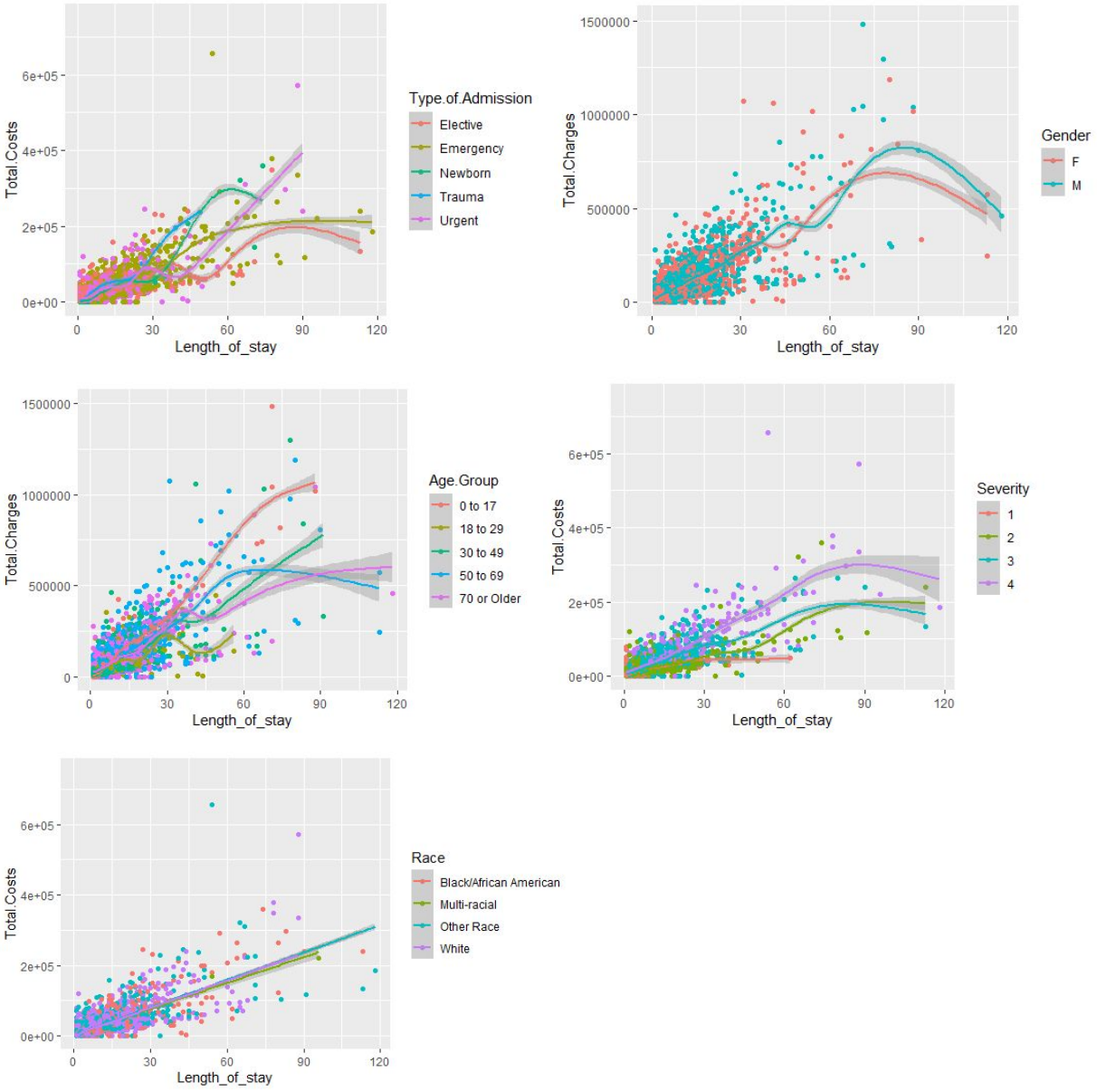
<u>Total Charges Distributions Per Demographic Group</u>

These visualizations illustrate the distributions of total charges for different age groups. For the age group, there is a large discrepancy between all categories. Total charges by gender and race are relatively similar. There are also differences in hospital service areas, particularly Long Island vs. Finger Lakes. Total charges by type of admission differ widely for newborn vs. all other types. By the severity of illness, each category varies in total charges distribution as well. By insurance, there are wide distributional differences, especially Medicare vs. Medicaid.



<u>Comparing Length of Stay to Total Charges by Demographic</u>

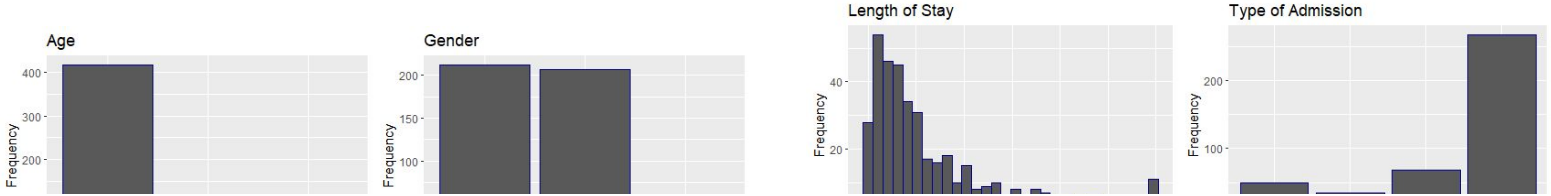
The length of stay of a patient should have significant impacts on the charges of their stay since the longer a patient stays the more resources and services the patient will use. However, other factors might impact the length of stay of a patient. The following graphs illustrate how different factors affect the total costs of a patient, even with similar lengths of stay.



Demographics for Top 5 Medical Conditions with Greatest Average Total Charges

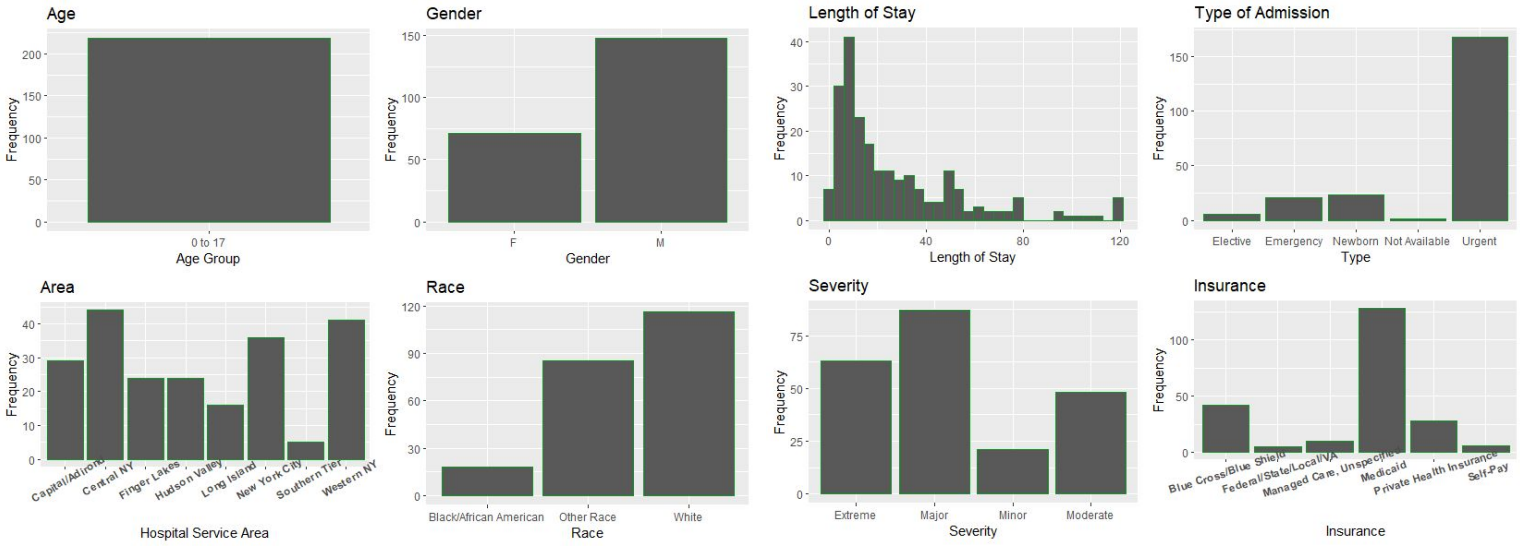
For this section, these data visualizations help reveal patterns on which cohorts are at risk of certain diseases. We focus on the top 5 most costly conditions, but ideally we would want to analyze each condition. By examining these descriptive statistics, we can determine which subpopulations are affected the most, and who would benefit the greatest from our modeling techniques with the highest cost savings.

Medical Condition: Short gestation, low birth weight, and fetal growth retardation (219)



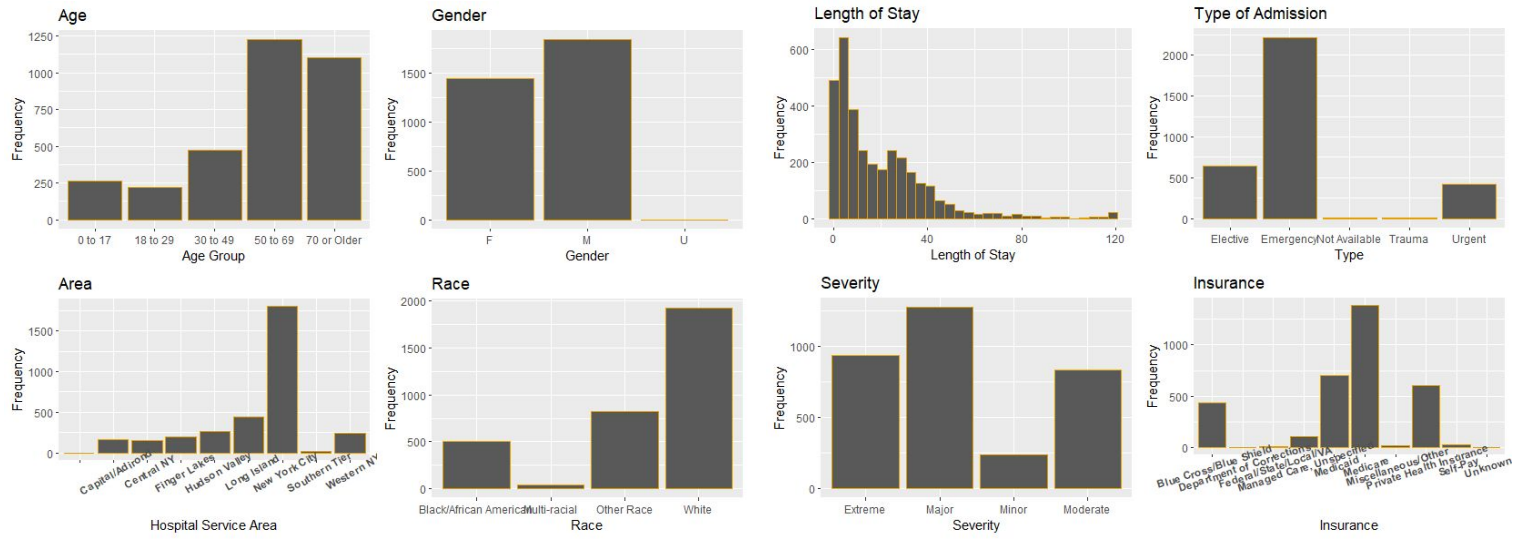
For this medical condition, it primarily affects the 0-17 age group with balanced gender, hospital service area, and race. Length of stay is primarily right skewed with urgent admission, but balanced severity of admission. Insurance used within this dataset is primarily Medicaid.

Medical Condition: Respiratory distress syndrome (221)



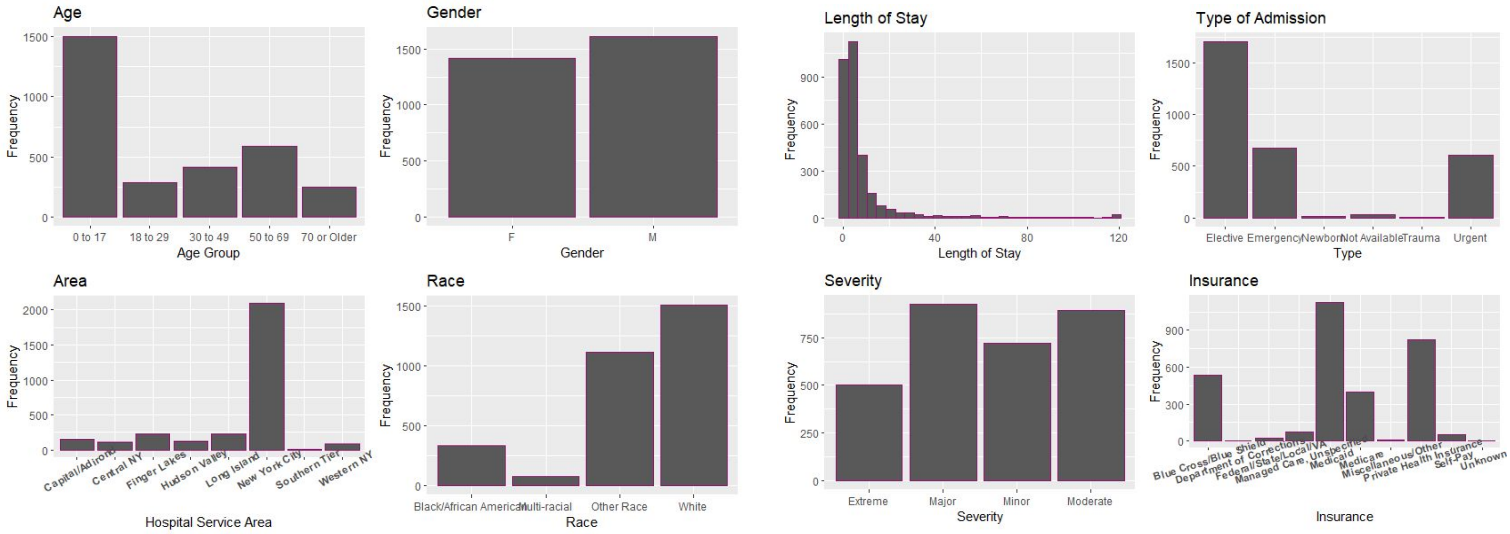
For this medical condition, it affects only the 0-17 age group, males more heavily than females, and whites more heavily than other races. The hospital service area is relatively balanced. Length of stay is right-skewed with mainly urgent admission, and major or extreme severity over minor or moderate. Insurance used within this dataset is primarily Medicaid.

Medical Condition: Leukemias (39)



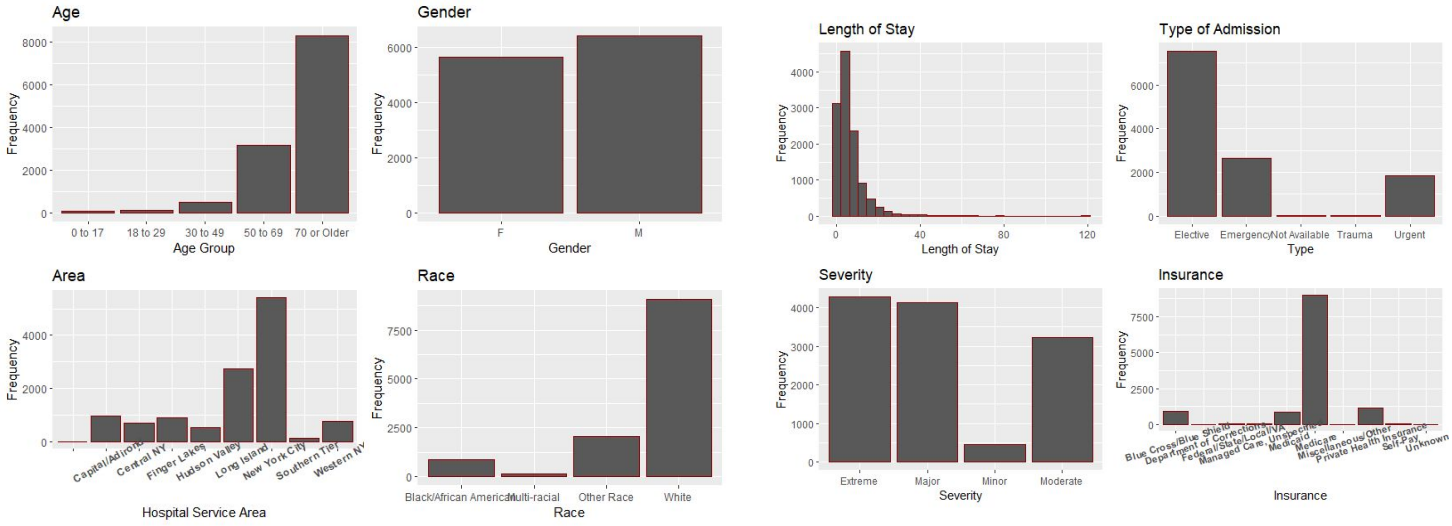
For this medical condition, it affects mainly older age groups like 50-69 or 70+, males more heavily than females, and whites more than other races. The hospital service area is mainly concentrated in New York City. The length of stay is rightly skewed with emergency as most common. The severity of extreme and major is more than minor or moderate. Insurance is primarily Medicare.

Medical Condition: Cardiac and circulatory congenital anomalies (213)



For this medical condition, it affects mainly 0-17, though it also affects older age groups as well. Gender affected is relatively balanced, and whites are more affected than other races. The hospital service area is mainly concentrated in New York City. The length of stay is rightly skewed with elective admission as most common. The severity is typically major and moderate. Insurance is primarily Medicaid.

Medical Condition: Heart valve disorders (96)

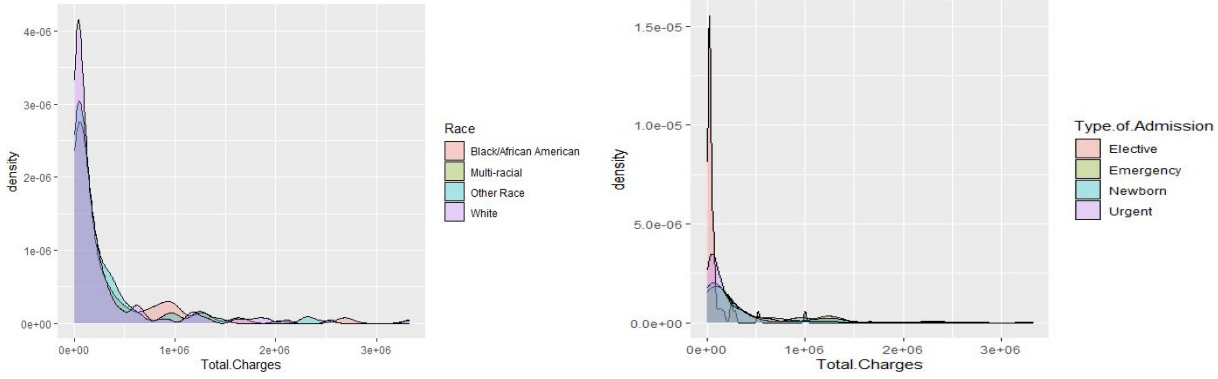


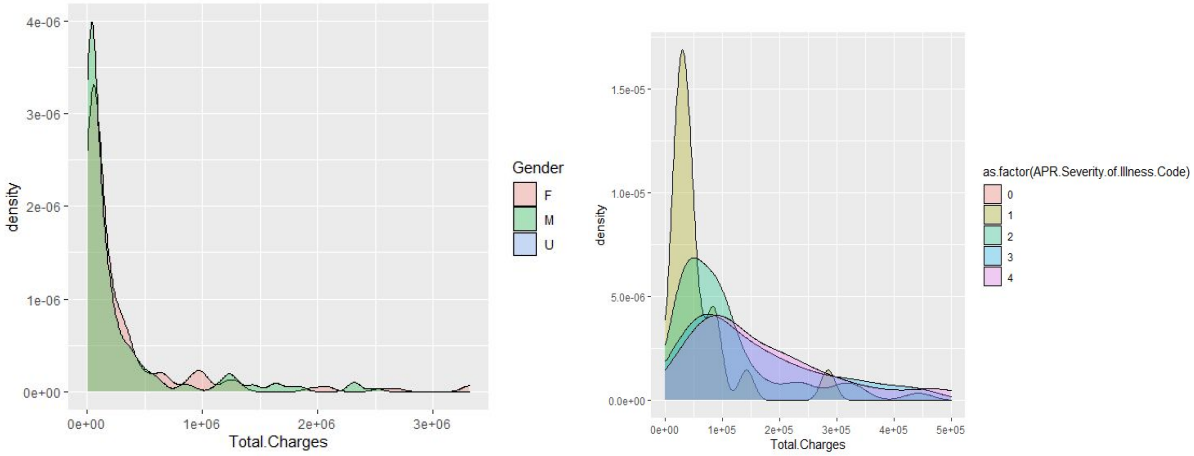
For this medical condition, it affects mainly 70+, though it also affects 50-69 as well. Gender affected is relatively balanced, and whites are more affected than other races. The hospital service area is mainly concentrated in New York City. The length of stay is rightly skewed with elective admission as most common. The severity is typically major and extreme. Insurance is primarily Medicaid.

<u>Comparing Distributions of Total Charges to the Top Costly Diseases</u>

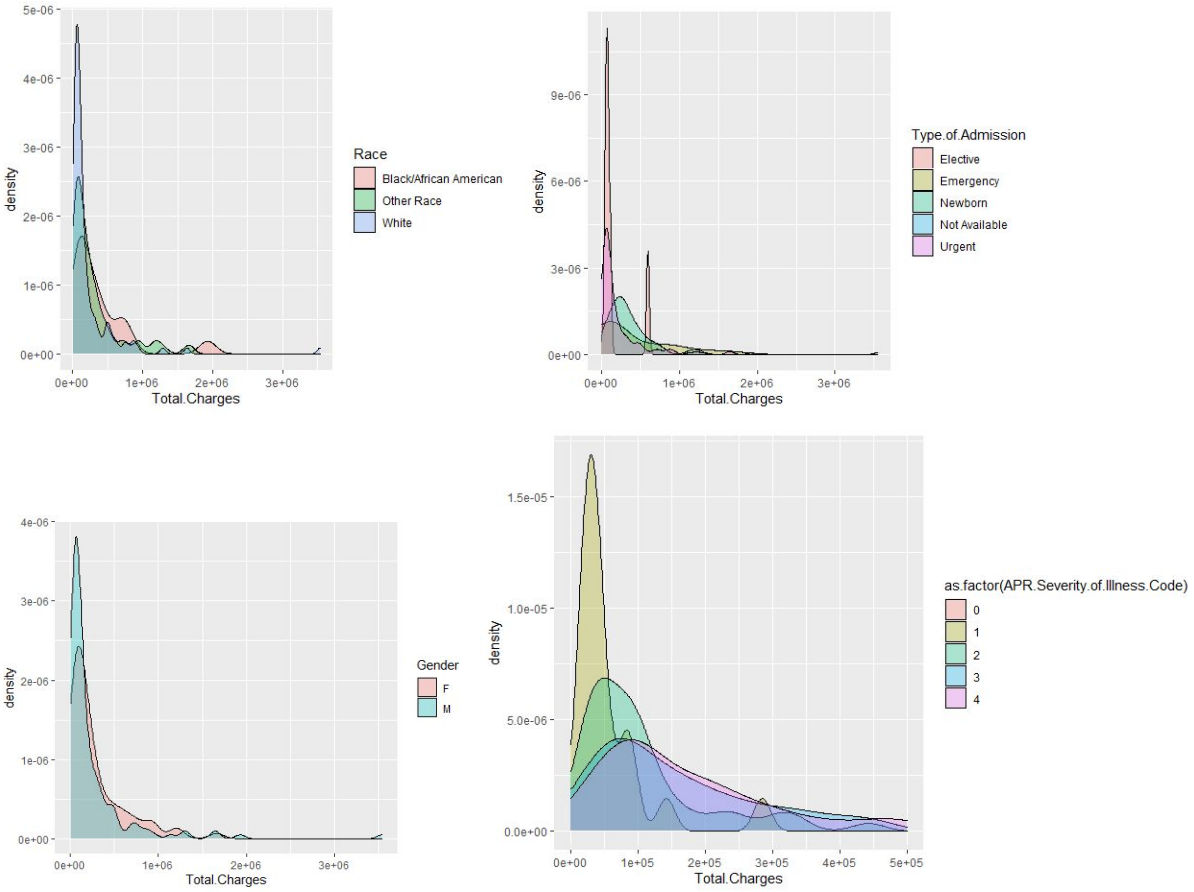
One of the main objectives of our project is to look at the disparity in the cost of healthcare services, even with similar diseases. The following graphs look at the 5 most costly diseases observed before and compare the distribution of total charges for those diseases.

Medical Condition: Short gestation, low birth weight, and fetal growth retardation (219)

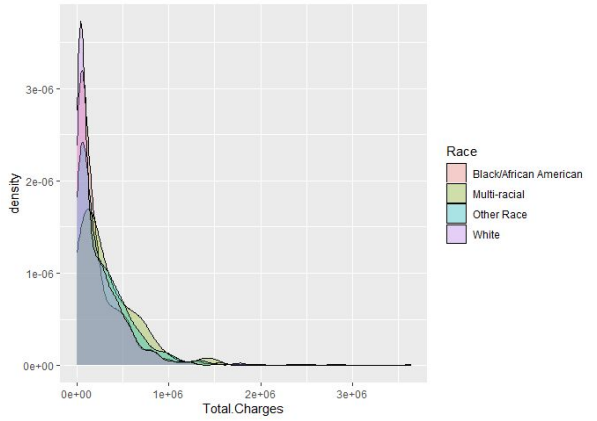


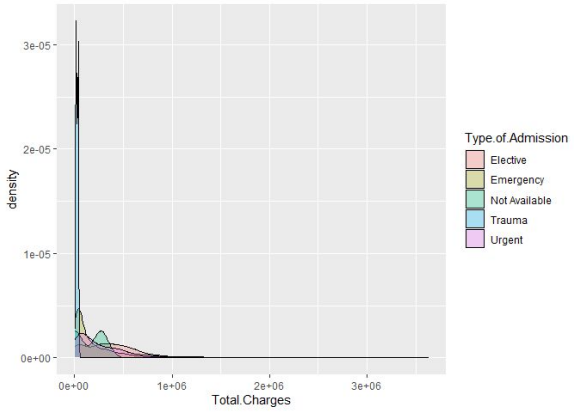


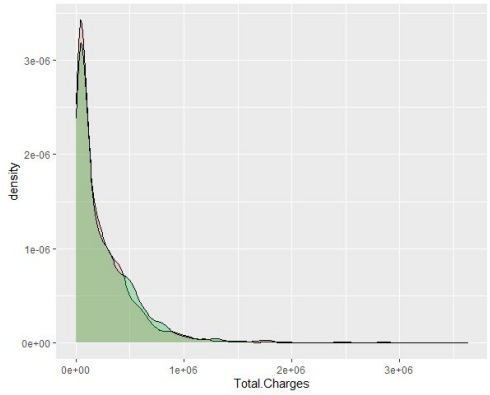
Medical Condition: Respiratory distress syndrome (221)

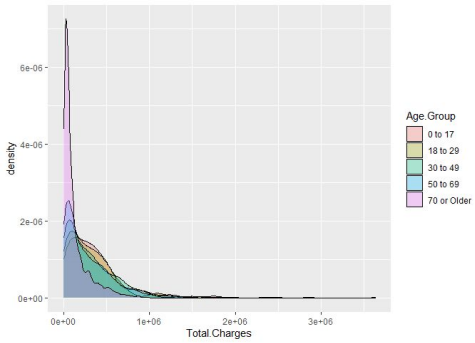


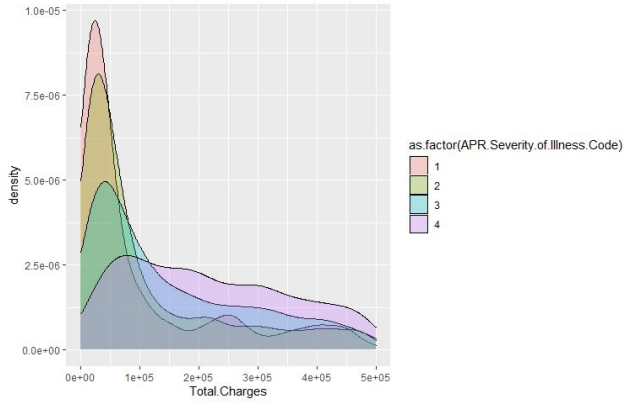
Medical Condition: Leukemias (39)



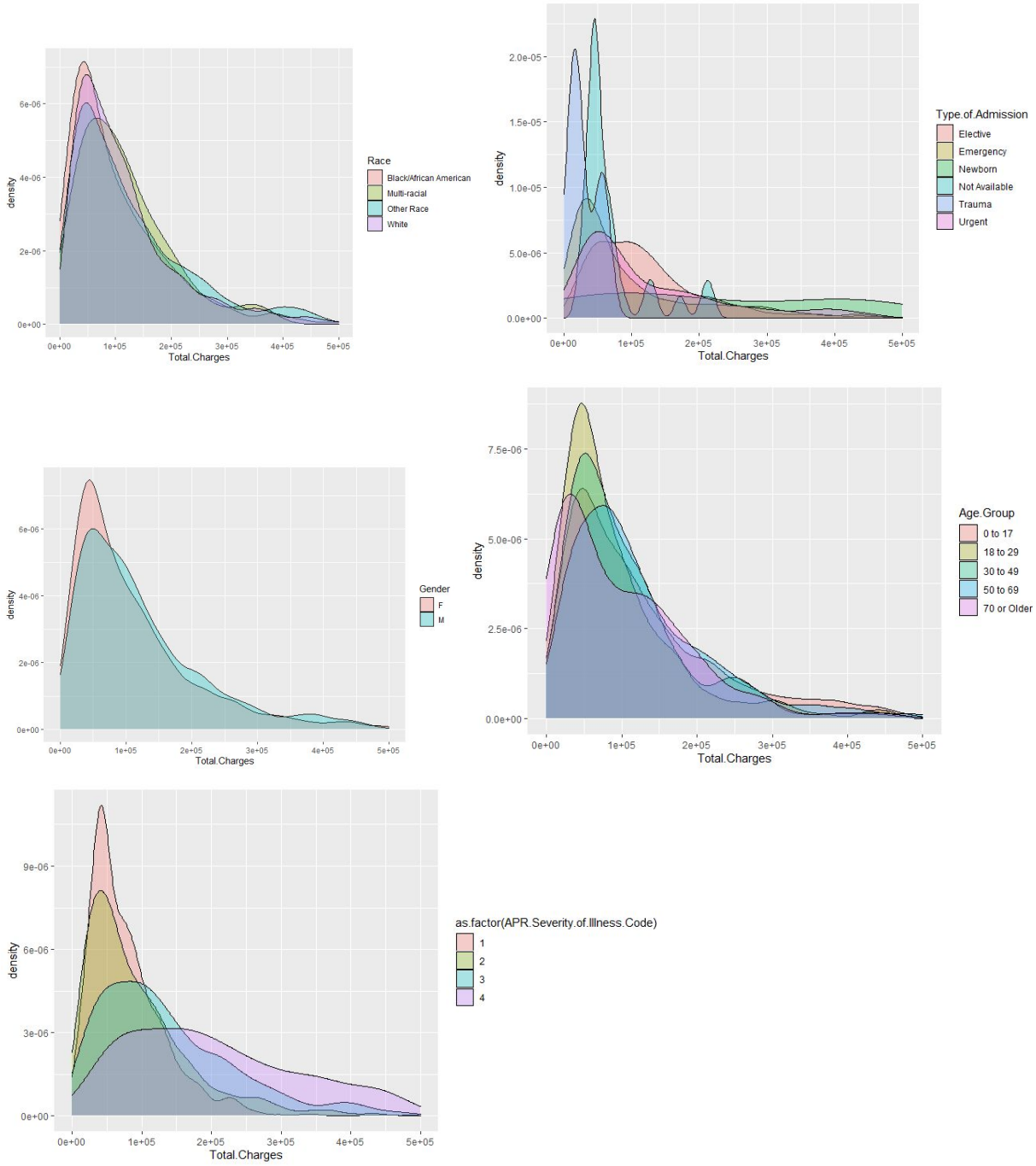






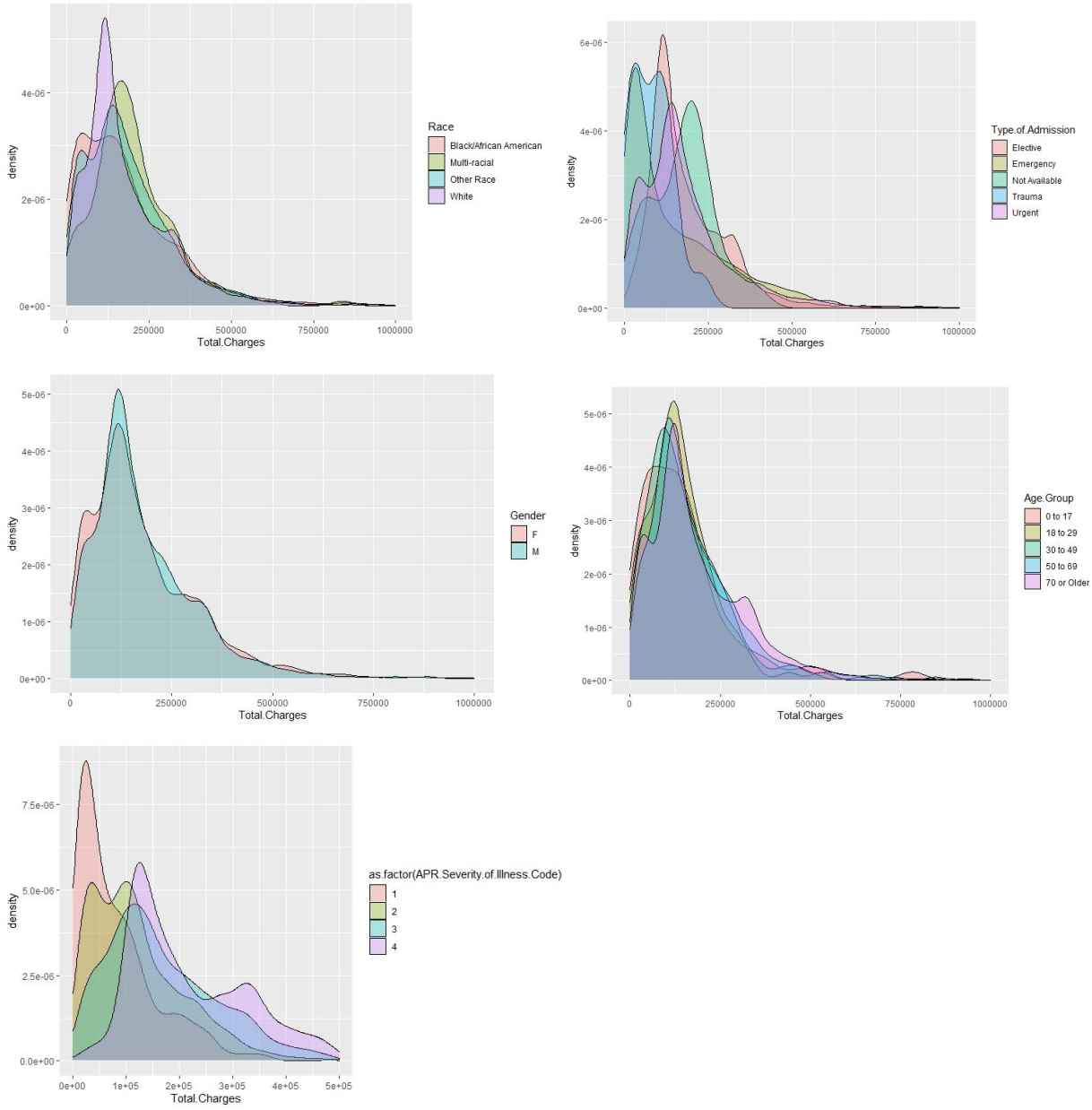


Medical Condition: Cardiac and circulatory congenital anomalies (213)



**These graphs are cut off to observe the differences in distribution better.*

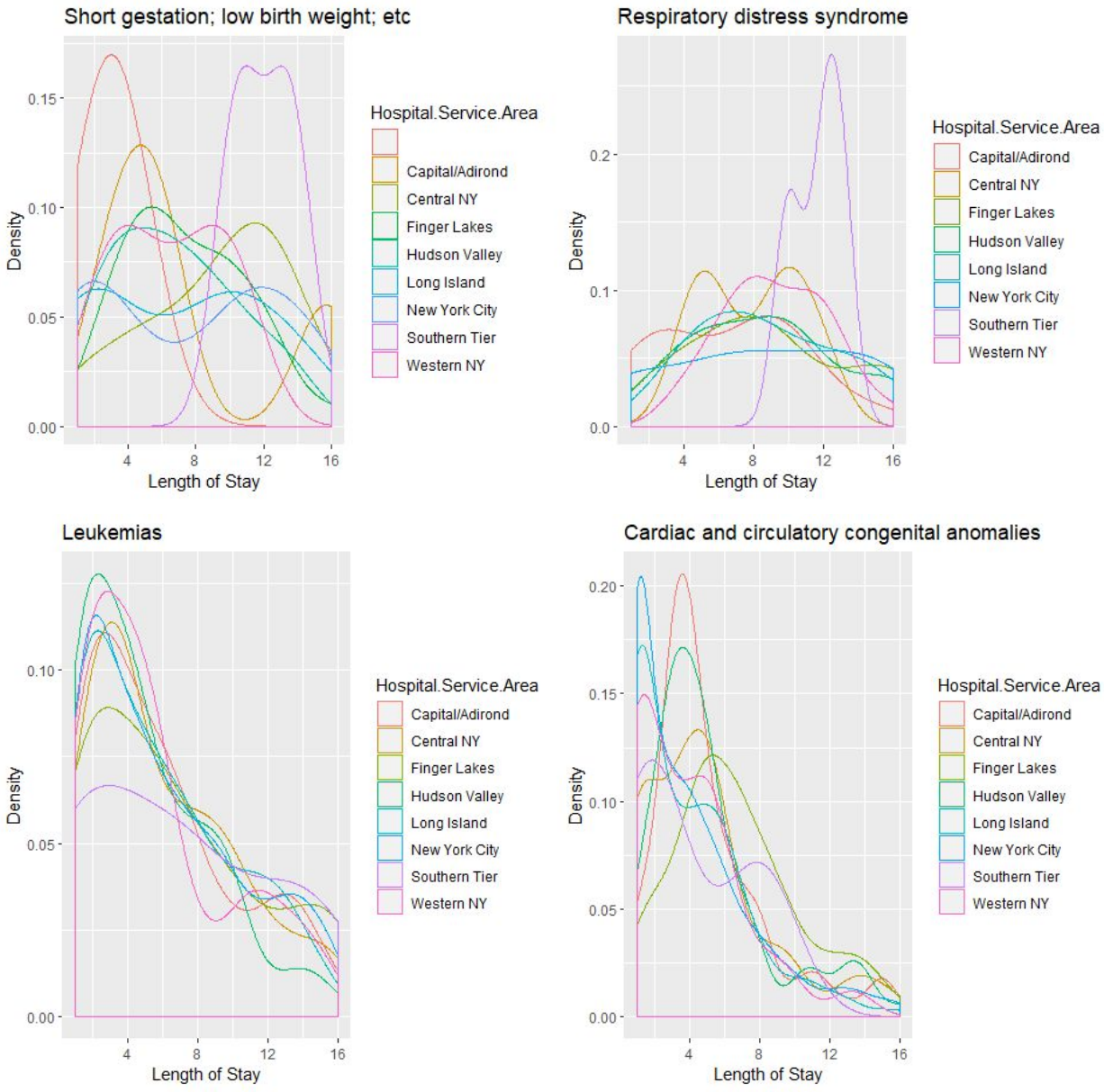
Medical Condition: Heart valve disorders (96)

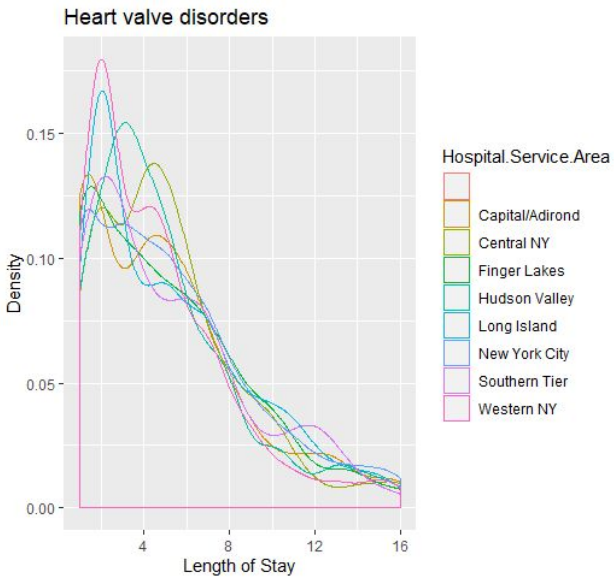


<u>Comparing Length of Stay for Top Costly Diseases</u>

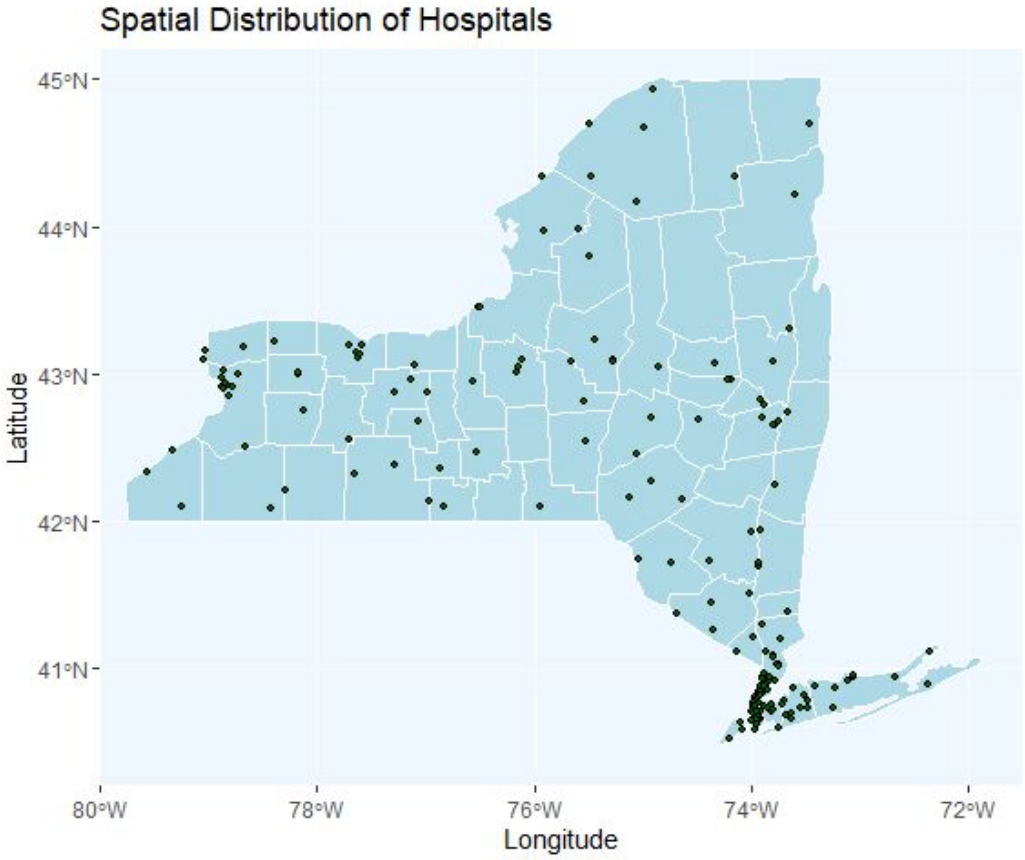
One of the project's objectives is to examine the disparity in the quality of healthcare services, even with similar diseases. With our dataset, length of stay acts as a proxy for quality of care. Length of stay has historically been used as indicators for hospital performance, hospital efficiency, quality of care. For example, if hospitals attempt to lower costs by prematurely discharging patients, length of stay significantly lower than expected might be indicative of poor care. Conversely, if poor quality of care causes complications, it would tend to extend length of stay. Under these assumptions, both longer and shorter than expected length of stays are indicative of poor quality care.

The following graphs look at the 5 most costly diseases observed before and compare the distribution of length of stay for those diseases per hospital service area. The distributions vary widely, especially for “short gestation; low birth weight; and fetal growth retardation” and “respiratory distress syndrome.” There is evidence to suggest that Southern Tier length of stays may indicate poor quality of care for “respiratory distress syndrome.”



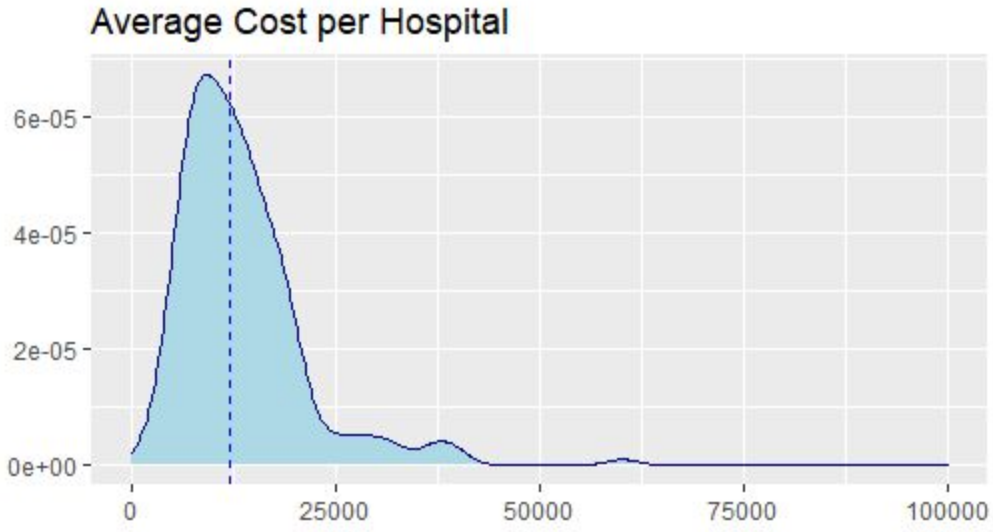


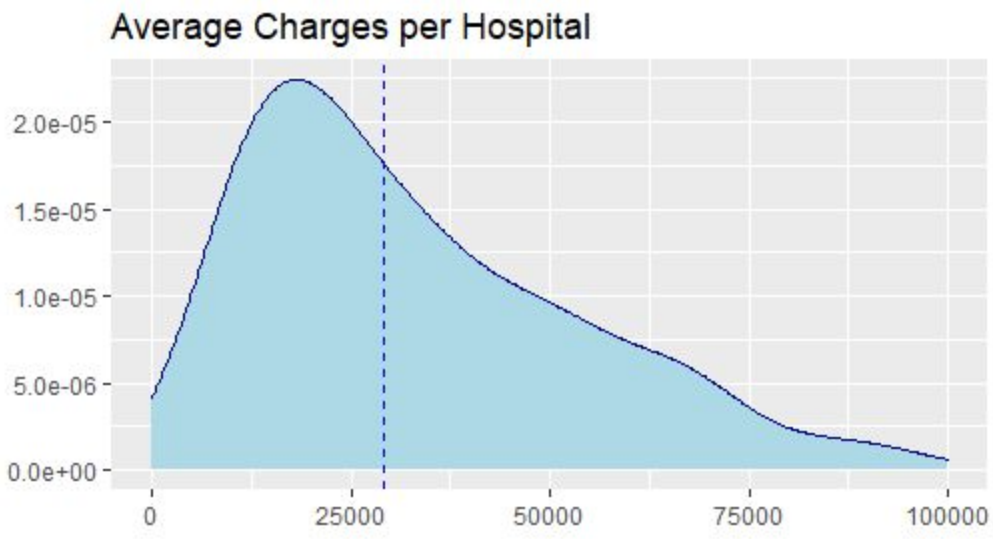
Geographical Distribution of Hospitals in the Dataset



Distribution of Average Cost and Charge per Hospital

The cost of a visit to the hospital can vary depending on the facility that gave the treatment. We looked at the average cost and charge per hospital and plotted the distribution, to visualize how hospitals might differ in their cost for healthcare. While the distribution of average cost was a tighter distribution, the average charges were a lot more spread out.





c) Statistical Analysis & Machine Learning:

For the first milestone, the two questions we are investigating are:

- Which of the features have the most impact in predicting charges (cost of care)?
- Which of the features have the most impact in predicting length of stay (one indicator of quality of care)?

We built different models and compared their performance to choose the optimal model (highest accuracy score and return results that are scientifically proven).

We treat the target variables (charges, length of stay) as both categorical and continuous variables.

Prepare data for modeling and conduct hypothesis testing:

1. ***Preprocessing data***

- "Length of Stay" and "age_group_code" was left as a number since it is an ordinal numeric feature.
- The numeric data types were binned based on their distributions and changed to categories.
- The nominal numeric types were changed to categorical types since there was no inherent meaning in the numeric data type.
- Dummy variables were used for categorical types.
- Drop unnecessary columns: Discharge Year was dropped since we are only looking at a single year. Payment Typology 2, Payment Typology 3, Abortion Edit Indicator, and Birth Weight were dropped since there were too many null values (>30%). Total Costs was dropped since it would impact Total Charges too much to be useful for our analysis. However, we did do 1 run with Total Costs as an independent variable, and the results were that a high cost results in high charges, which matches what we would expect. Hospital_service_area_code since we already had facility_id, i.e. we want to analyze on hospital level, so we drop area level to avoid too many categories. Patient zip code was also dropped to avoid too many categories.
- Reduce the number of categories of too-many-level categorical variables (ex. CCS Diagnosis Code has more than 200 levels):
    - Since APR MDC code is the umbrella for CCS Diagnosis Code (APR MDC has 25 categories, while CCS Diagnosis divides to smaller branches from these 25 categories), we only used APR MDC code to avoid too-many-level categorical variables and multicollinearity.
    - Patient Disposition: we further grouped this variable into major categories according to the documentation from the Joint Commission National Quality Measures, reducing the number of categories from 20 to 7 groups.
    - CCS Procedure Description: we further grouped this variable into major categories according to the documentation from the The Healthcare Cost and Utilization Project , reducing the number of categories from 222 to 16 groups.
    - Group all the categories that have less than 0.02% of the data to the "Other" group

2. ***Hypothesis Testing***

- Used Chi square test to detect highly correlated predictor variables → remove
- Used Kruskal-Wallis H-test to preliminarily investigate which factors are correlated with the response variable (before that, also used QQPlot and Shapiro-Wilk Test to determine the normality of the data)

1. Decision Tree Model for Total Charges

The reason we chose to investigate the data with a decision tree model was because a decision tree allows us to trace the different splits that lead to the final classification. What splits first contributes the most to the reduction in uncertainty and increase in purity, which gives us a better sense of what features are driving the prediction. A decision tree also gives a lot of control over how to design the branches for each feature.

Features used for the independent variables:

1.	Permanent Facility Id	category	dummy
2.	Gender	object	dummy
3.	Race	object	dummy
4.	Length of Stay	int64	original numeric - kept number
5.	Type of Admission	object	dummy
6.	APR MDC Description	object	dummy
7.	APR Severity of Illness Code	category	dummy
8.	APR Medical Surgical Description	object	dummy
9.	Payment Typology 1	object	dummy
10.	Emergency Department Indicator	object	dummy
11.	Total Charges	float64	dropped
12.	age_group_code	int64	original numeric - kept number
13.	apr_risk_mortality_code	category	dummy
14.	patient_discharge_group	category	dummy
15.	ccs_procedure_major_group_code	category	dummy
16.	total_charges_code	category	Binned into 3 based on distribution, dummy

Target Variable:

- "Total Charges" was binned into 3 buckets by distribution and dummied
- Target were high "Total Charges" values -> "total_charges_code_3"

- We decided to use Total Charges as the dependent variable, as it would be more relevant at the patient level. Total Charges is the total for everything charged, but not necessarily paid by the patient (can be covered by insurance). Total Costs is what it costs the hospital for providing care and service.

Results:

1. Summary:
    - By examining each level of the decision tree, we were able to trace the features that led up to the high charge prediction. The top predictors of high charges were:

Split	Branches
First	- Length of Stay <= 5.50 (Length of Stay > 5.50)
Second	- APR Medical Surgical Description_Surgical <= 0.50 (APR Medical Surgical Description_Surgical > 0.50) - APR Medical Surgical Description_Medical <= 0.50 (APR Medical Surgical Description_Medical > 0.50)
Third	- Length of Stay <= 3.50 - APR MDC Description_Pregnancy, Childbirth and the Puerperium <= 0.50 - Length of Stay <= 10.50 - Length of Stay <= 9.50
Fourth	- APR MDC Description_Diseases and Disorders of the Nervous System <= 0.50 - Length of Stay <= 4.50 - Length of Stay <= 2.50 - Permanent Facility Id_1463.0 <= 0.50 - ccs_procedure_major_group_code_7 <= 0.50 - Length of Stay <= 17.50 - APR MDC Description_Mental Diseases and Disorders <= 0.50 - apr_risk_mortality_code_1 <= 0.50
Fifth	- ccs_procedure_major_group_code_7 <= 0.50 - Permanent Facility Id_1463.0 <= 0.50 - Permanent Facility Id_1139.0 <= 0.50 - APR MDC Description_Diseases and Disorders of the Circulatory System <= 0.50 - Permanent Facility Id_1447.0 <= 0.50

	- Length of Stay <= 4.50 - Length of Stay <= 3.50 - Length of Stay <= 7.50 - APR MDC Description_Diseases and Disorders of the Circulatory System <= 0.50 - Permanent Facility Id_1178.0 <= 0.50 - Length of Stay <= 21.50 - Permanent Facility Id_541.0 <= 0.50 - Permanent Facility Id_3975.0 <= 0.50 - ccs_procedure_major_group_code_0 <= 0.50 - Length of Stay <= 36.50

2. Evaluation Metrics:
    - Predictive accuracy (with 80% training set and 20% testing set): 90.63%
    - The precision TP/(TP+FP) and recall TP/(TP+FN) were 0.8381% and 0.7754%
    - Adjusted pruning parameters with max_depth = 20 to prevent overfitting

3. Analysis:
    1. Top predictor was “Length of Stay”. This makes sense because longer stay would mean using more resources and requiring more care so charges would accumulate.
    2. The second top predictors were “APR Medical Surgical Description_Surgical” and “APR Medical Surgical Description_Medical”. This makes sense because surgery would cost more than other types of medical care, since a surgical procedure would require resources such as the operating room, recovery room, anesthesia (APR-DRGs Methodology Overview).
    3. Other predictors: “APR MDC Description_Pregnancy, Childbirth and the Puerperium”, “APR MDC Description_Diseases and Disorders of the Nervous System”, “Permanent Facility Id_1463.0”, “ccs_procedure_major_group_code_7”, “APR MDC Description_Mental Diseases and Disorders”, “apr_risk_mortality_code_1”
    4. The feature “apr_risk_mortality_code_1” corresponds to minor mortality risk. Compared with the visualization for APR Severity of Illness Description, there was a skew for minor mortality risk to be low charge so it makes sense that it is a determining factor of charge outcome in the decision tree [Total Charges Distributions Per Demographic Group].
    5. This tree is n-ary but most branches are binary.
    6. For the dummied variables, the nodes correspond to the split on if the value was the case or not. For example, instead of splitting on animal: cat, dog, or hamster, the nodes split on animal_cat: yes or no. This lends itself to being more specific to the feature value for predicting high or low charge.

4. Next Steps:
    - Decision tree for what variables affect length of stay
    - Other models
    - Further exploration of replacing the zip codes with median income of each cluster of zip codes would be interesting
    - Do more expensive hospitals (charge more) have higher internal costs
    - Cost and charge difference disparities across demographics

2. Linear regression for Total Charges

Features used for the independent variables: The same with the Decision Tree model.
Preprocessing: Similar with the Decision Tree, except for leaving the Length of Stay as a continuous variable.

Advantage of the model:

- Quantify the effect of each predictor on the dependent variable at the same time take confounding variables into account
- Linear equations have an easy to understand interpretation on a modular level
- Use to predict the outcome of a patient

Disadvantage when applying on this dataset

- Very long run-time

Results:

2.1 Summary:

- By examining the standardized coefficients of the linear regression, we saw that the following predictors have the most influence on the total charges

Variable	**Description**
- Length of Stay	Positive correlated
- Type of Admission	Newborn (-), Emergency (-): Negative correlated compared with Type of Admission_elective, newborn has a lower charge
- Permanent Facility ID	1139 (+), 1463 (+), 1464 (+), 511 (+), 1458 (+), 541 (+), 1446 (+), 1169 (+), 1175 (-) compared with Permanent Facility ID 1
- ccs_procedure_major_group_code	7 (+), 14 (+): Positive correlated compared

	with non procedure
- APR MDC Description	Diseases and Disorders of the Circulatory System (+), Newborns and Other Neonates with Conditions Originating in the Perinatal Period (+), Mental Diseases and Disorders (-), Diseases and Disorders of the Musculoskeletal System and Conn Tissue (+) Diseases and Disorders of the Nervous System (+) Compared with (Alcohol/Drug Use & Alcohol/Drug Induced Organic Mental Disorders)

2.2 Evaluation Metrics:

- Adjusted R-Squared - Test set: 0.595
- Adjusted R-Squared - Train set: 0.613

2.3 Analysis:

- Similarities with the Decision Tree:
  - Top predictor was "Length of Stay". This confirms the result in the Decision Tree Model
  - Having surgery is also a good predictor for higher charge, confirming the result from the Decision Tree.
  - Permanent Facility ID and APR MDC Description are also important predictors
- An important predictor that this Linear Regression found is Type of Admission

3. Principal Component Analysis

3.1 Description

- Reduce dimensions for the dataset, which speed up the run-time
- Able to extract the most significant predictors

Summary Results:

Top 20 predictors for Total Charges

	Variable
1	age_group_code
2	Payment Typology 1_Medicare
3	apr_risk_mortality_code
4	APR MDC Description_Newborns and Other Neonates with Conditions Originating in the Perinatal Period
5	Type of Admission_Newborn
6	APR Severity of Illness Code
7	Type of Admission_Emergency
8	Payment Typology 1_Medicaid
9	APR MDC Description_Pregnancy, Childbirth and the Puerperium
10	ccs_procedure_major_group_code_13
11	ccs_procedure_major_group_code_11
12	APR MDC Description_Diseases and Disorders of the Circulatory System
13	Race_Other Race
14	Race_White
15	ccs_procedure_major_group_code_7
16	APR MDC Description_Infectious and Parasitic Diseases, Systemic or Unspecified Sites
17	Length of Stay
18	Payment Typology 1_Private Health Insurance
19	APR MDC Description_Diseases and Disorders of the Respiratory System
20	Permanent Facility Id_3376.0

Next Steps:

1. Check linear regression model assumptions
2. Try mixed effect linear regression model

3) Conclusions and Future Work:

Preliminary analysis demonstrated cost differences across hospitals where distributional differences between patient care cost for the hospital and patient care cost for the patient

supports the hypothesis that some hospitals are overcharging for care. Other visualizations also show differences in total charges and length of stay depending on patient demographics, hospital service area, and admission severity, even when controlling for a patient's medical condition. Knowing this, it is in the patient's best interest to identify which hospitals will be the most cost-effective for their care.

Through our modeling analysis, we are able to conclude which factors are most predictive in the cost of a patient's hospital care. In our decision tree model, the most important factors include length of stay, an indicator for surgical or non-surgical care, diagnosis descriptions, and risk of mortality. In our regression model, the most important factors include length of stay, type of admission, facility ID, diagnosis description, and procedure code. Through PCA, the most important factors include age, payment type, risk of mortality, diagnosis description, and severity of illness. Other notable features include race, length of stay, and type of admission.

There are several similarities when comparing our model results. In each analysis, we find length of stay and diagnosis description to be powerful predictors. Having or not having a surgical procedure is prevalent in the decision tree, and variations of this variable in the form of procedure code is also in the regression model. Related attributes like risk of mortality, type of admission, and severity of illness are also common among analyses. With similarities in results among all three analyses, we conclude the aforementioned factors to hold predictive power in determining a patient's cost of hospital care.

Additionally, our project aimed to investigate the cost disparities between different population groups. Through PCA, we begin to answer the question of racial cost disparity since we see race being in the top 20 predictors for total charges. To explore cost disparities for different socioeconomic groups, we would need to incorporate the household income data mentioned in the Data Wrangling section above. This is one manner our work can be expanded.

One major consideration of patient care is finding the balance between quality and cost of care. We began preliminary analysis on quality of care with length of stay as a proxy, which has historically been used as indicators for hospital performance, hospital efficiency, and quality of care. Hospitals attempting to lower costs by prematurely discharging patients would have lower length of stay, and hospitals with high rates of complications would have higher length of stay. At both extremes, quality of care is low.

Because of this complicated relationship between length of stay and quality of care, we believe length of stay may be a poor stand alone indicator, so our modeling analyses does not attempt to predict quality, rather focusing solely on cost instead. Therefore, our work should be expanded to include other data that incorporate quality metrics such as timely and effective care, attitude of medical providers, and ease in accessing services. One example source for data on quality of care is The Centers for Medicare and Medicaid Services Hospital Compare database.

Other possibilities in which our analysis could be expanded is examining data earlier than our 2017 data for time series analysis to determine if predictive cost indicators, or cost of care in general, are stable through time. Our analysis also only focused on New York state, so further

analysis could compare if predictive cost indicators are the same across states. Furthermore, after incorporating quality of care metrics, a recommender system could be created to directly answer a patient's question of which hospital to visit given a health condition and their personal priorities.

References

1. https://www.fah.org/blog/words-matter-defining-hospital-charges-costs-and-payments-and-the-numbers-t
2. https://www.hcup-us.ahrq.gov/db/nation/nis/APR-DRGsV20MethodologyOverviewandBibliography.pdf
3. https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp#download
4. https://manual.jointcommission.org/releases/TJC2015B/TableOfContentsTJC.html
5. https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/
6. https://www.kaggle.com/mihirjhaveri/inpatient-analysis-and-predicating-length-of-stay
7. https://medium.com/gett-engineering/handling-rare-categorical-values-in-pandas-d1e3f17475f0
8. https://towardsdatascience.com/categorical-feature-selection-via-chi-square-fc558b09de43
9. https://codereview.stackexchange.com/questions/96761/chi-square-independence-test-for-two-pandas-df-columns
10. https://stackoverflow.com/questions/51632900/pandas-apply-kruskal-wallis-to-numeric-columns