**Methodology**

**Overview**

The data for this study come from a selection of U.S. government websites. They are not in any way proprietary and are easily found via an online search. This section will describe the sources of these data, the process used to clean and merge them together, and all other steps performed in order to prepare the data for analysis. We also review rates of missing data in order to inform us on possible biases inherent in our valid data.

For all data manipulation including cleaning, modelling and graphic production, we utilize R statistical software (R Core Team, 2017), which is open-source and available for free online. To make some of these tasks simpler, a number of R software packages are employed: *betareg* (Cribari-Neto & Zeileis, 2010), *car* (Fox & Weisberg, 2019), *descr* (Aquino et al., 2018), *knitr* (Xie, 2015), *MASS* (Venables & Ripley, 2002), *RColorBrewer* (Neuwirth, 2014), *realxl* (Wickham & Bryan, 2019), *sjPlot* (Lüdecke, 2020), and *tidyverse* (Wickham et al., 2019). To access this software, we rely on RStudio (RStudio Team, 2020), an integrated development environment designed for use with R.

**Data Sources**

The main source of data for the analyses below is Medicare’s Hospital Compare database (Centers for Medicare & Medicaid Services, 2019-2020). Medicare is a federal health insurance program in the United States providing coverage for individuals age 65 and older, certain young people with disabilities, and people with end-stage renal disease. In order for hospitals to register with Medicare and ultimately receive compensation for their services, hospitals are required to report certain performance statistics. These include speed and quality of services offered, complication rates of various surgeries, and levels of patient satisfaction. This helps Medicare officials provide quality assurance to patients they serve, as well as the general public.

It is among this trove of data that metrics on “Timely and Effective Care,” released quarterly, can be found. We will use data from the full 2018 calendar year exclusively. These data focus on performance within the emergency department, the area of a hospital responsible for providing care to patients arriving in need of immediate help. The data show how quickly and efficiently hospitals accomplish urgent tasks, and whether certain life-saving interventions are administered within the recommended time frame. We opt to use five variables from this database as our primary study outcome variables. Those variables are:

|  |  |
| --- | --- |
| **Response Variable** | **Definition** |
| *AdmitLOS* | Average (median) time from emergency department arrival to emergency department departure for patients admitted to the hospital as an inpatient |
| *WaitForBed* | Average (median) time from admit decision to time of departure from the emergency department for patients admitted to the hospital as an inpatient |
| *NonAdmitLOS* | Average (median) time from emergency department arrival to emergency department departure for all discharged patients |
| *MHLOS* | Average (median) time from emergency department arrival to emergency department departure for discharged psychiatric or other mental health patients |
| *LWBSrate* | Percentage of patients who leave the emergency department before being seen |

**TABLE 1: Response variable definitions**

Included in the same file as the variables above are several pieces of identifying information for each hospital (facility name, address, etc.) as well as of *ED.Volume*, a categorical variable showing the average volume of patients treated by each hospital’s emergency department each year. The four levels are: Low, Medium, High, and Very High. We then append that data with *Beds,* the total number of Medicare-certified beds at each hospital (Cecil G. Sheps Center for Health Services Research, 2019) and a useful proxy for the size of each hospital.

We then use data from Medicare’s 2018 Hospital Service Area (HSAF) file (Centers for Medicare & Medicaid Services, 2019) to merge the “Timely and Effective Care” data with demographic data form the U.S. Census and other sources. A version of the HSAF is released each year showing the home zip codes of every patient served by each hospital, as well as the number of patients from each zip code served that year. Analysis of the 2018 version shows that while large, urban hospitals usually serve patients from hundreds of different zip codes in one year, a small, rural hospital might treat patients from only a few dozen zip codes in close proximity to the facility. Using this file, we are able to merge performance metrics for each hospital with information on the demographics it serves.

To obtain the bulk of that information, we turn to the 2018 American Community Survey (United States Census Bureau, 2018). The United States Constitution requires that the federal government take a census of all persons living in the U.S. at least every ten years, a task currently delegated to the U.S. Census Bureau. In addition to producing a full count every ten years as required, the Census Bureau also administers an annual population study called the American Community Survey. The survey, sent to a representative sample of Americans, helps the Bureau produce estimates of population levels for each year in between census. Using the results of the 2018 ACS along with the 2018 HSAF file, we merge the performance metrics described previously with demographic information from the U.S. Census Bureau.

Next, we again utilize the HSAF to assign a *RuralScore* to each hospital. Every few years, the Federal Office of Rural Housing Policy publishes a list of zip codes that are to be considered “rural,” as opposed to urban, to assist federal agencies in resource allocation. Here we rely on the most recent version of that list (U.S. Office of Rural Health Policy, 2019) to give each hospital a score between 0 and 1 corresponding to the proportion of patients they serve who reside in a zip code found on that list. The higher the *RuralScore*, the more rural a hospital is considered to be.

Finally, we use data from the Kaiser Family Foundation (2020) to build a variable showing whether or not a hospital is located in state that had expanded Medicaid by 2018. Under the Affordable Care Act of 2010 (“Obamacare”), states were permitted to loosen Medicaid eligibility requirements if they so choose in order to register more low-income Americans. As of December 2018, 36 states and the District of Columbia had exercised their option to expand Medicaid in their state. It should be noted that between 2018 and 2020, three additional states chose to go forward with expansion.

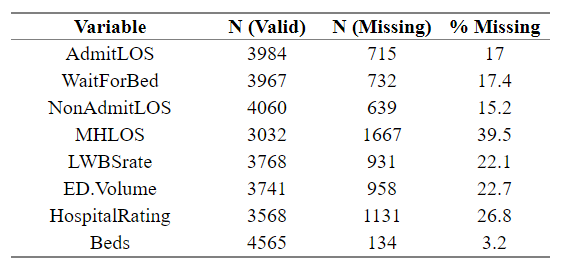
|  |  |
| --- | --- |
| **Independent Variable** | **Definition** |
| **Race/Ethnicity** | |
| *Asian* | Proportion of patients served who identify as Asian or Pacific Islander, either alone or in combination with another race/ethnicity |
| *Black* | Proportion of patients served who identify as Black or African American, either alone or in combination with another race/ethnicity |
| *Hispanic* | Proportion of patients served who identify as Hispanic or Latino, either alone or in combination with another race/ethnicity |
| *Native American* | Proportion of patients served who identify as Native American, either alone or in combination with another race/ethnicity |
| *White* | Proportion of patients served who identify as White or Caucasian, either alone or in combination with another race/ethnicity |
| **Other Demographics** |  |
| *Median Age* | Average (median) age of patients served, in years |
| *Rural Score* | Proportion of patients who reside in a zip code designated as rural by the Federal Office of Rural Housing Policy |
| *Sex Ratio* | Number of male patients served per 100 female patients served |
| ***Hospital-Level*** | |
| *Beds* | Total number of Medicare-certified beds |
| *ED Volume* | Categorical variable showing average emergency department volume (Low, Medium, High, Very High) |
| *Medicaid Expansion* | Is the hospital located in State that, by 2018, had expanded Medicaid under the Affordable Care Act of 2010? (Yes, No) |

**TABLE 2: Independent variable definitions**

**Treatment of Missing Data**

Because of varying reporting requirements, human error and other potential issues, a number of data points are missing from out dataset. It is important we examine patterns in the data that are missing to make sure that we do not introduce any significant bias by using only the data that are present for our study. We also discuss the decision to exclude certain data from the study.

The table below shows the rates of missing and valid data for eight of the variables to be used on our predictive models, including each of the five response variables. Not all variables are displayed, as there were no missing data among the demographic variables nor any of the hospital identifying information. The main reason that data are missing is that Medicare limits what data are made public in order to ensure the performance of small hospitals is not misrepresented. Unless hospitals reach a certain threshold for visits of a certain type, they are not required to report performance statistics for that visit type. This guarantees that the statistics that are reported are based on a sufficient sample size, and that a handful of negative patient outcomes occurring at random does not harm the reputation of an otherwise well-performing hospital.

  
**TABLE 3: Rates of Missing Data**

It is also important to note that data from U.S. territories have been removed. In addition to the 50 states plus D.C., Medicare also collects hospital performance metrics from U.S. territories including Puerto Rico, Virgin Islands, American Samoa, Guam, and Northern Mariana Islands. The issue with regard to our study is that the information that must legally be reported to Medicare is vastly different between the states and these territories, and in our data, information on emergency department wait times and length of stay is largely missing. We therefore define our study population as Medicare-registered hospitals in the 50 U.S. states plus the District of Columbia. We also exclude hospitals that did not report data on any of the five response variables in 2018 from dataset.

**Analysis**

**Overview**

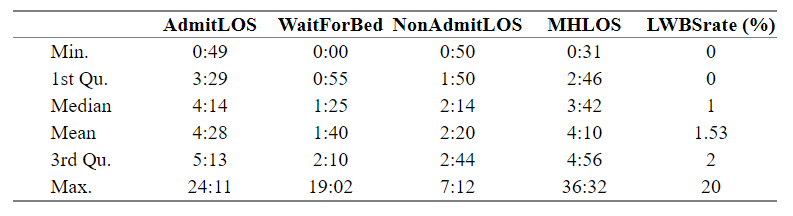
The final dataset of emergency department (ED) performance metrics contains N = 4216 observations from all 50 United States and the District of Columbia. Analysis will focus on the construction of predictive models for each of our five response variables: ED length of stay for admitted patients (*AdmitLOS*) , time spent in ED waiting for an inpatient bed (*WaitForBed*), ED length of stay for discharged patients (*NonAdmitLOS*), ED length of stay for mental health and substance use patients (*MHLOS*), and rate of patients leaving the ED without being seen (*LWBSrate*). To predict these outcomes, a selection of relevant independent variables is used.

The goal is to identify whether emergency departments that serve higher rates of patients from non-white races and ethnicities perform significantly worse than emergency departments serving largely white patients, as measured by the five response variables. While prior studies with this aim have focused largely on rudimentary approaches, particularly simple linear regression, this study uses a variety of Generalized Linear Models (GLMs), an iterative weighted regression technique where observations are assumed to be distributed according to some exponential family. For each response variable, an appropriate family of Generalized Linear Model is identified, and successively more complex models are built, with outliers removed when appropriate. Finally, model diagnostics are performed to ensure each model is an appropriate fit for the data. Results are then presented by response variable.

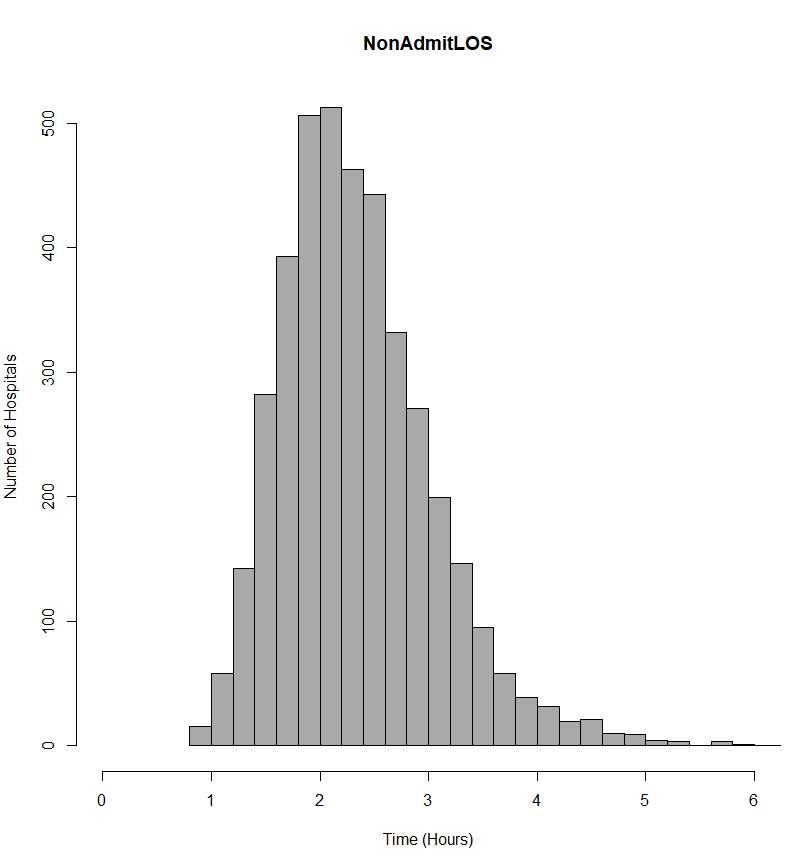
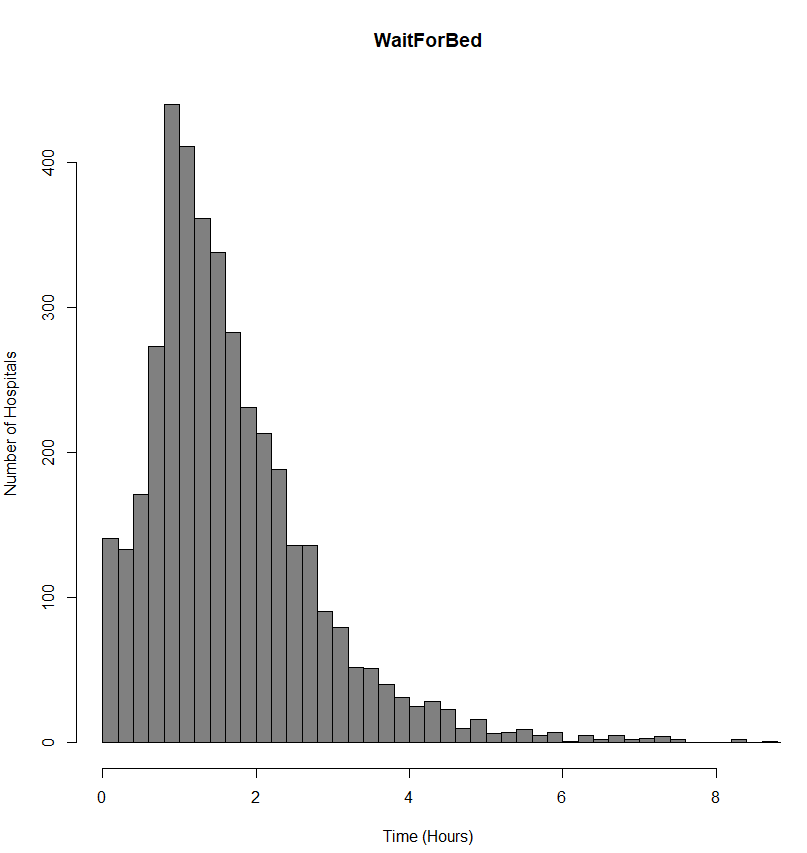
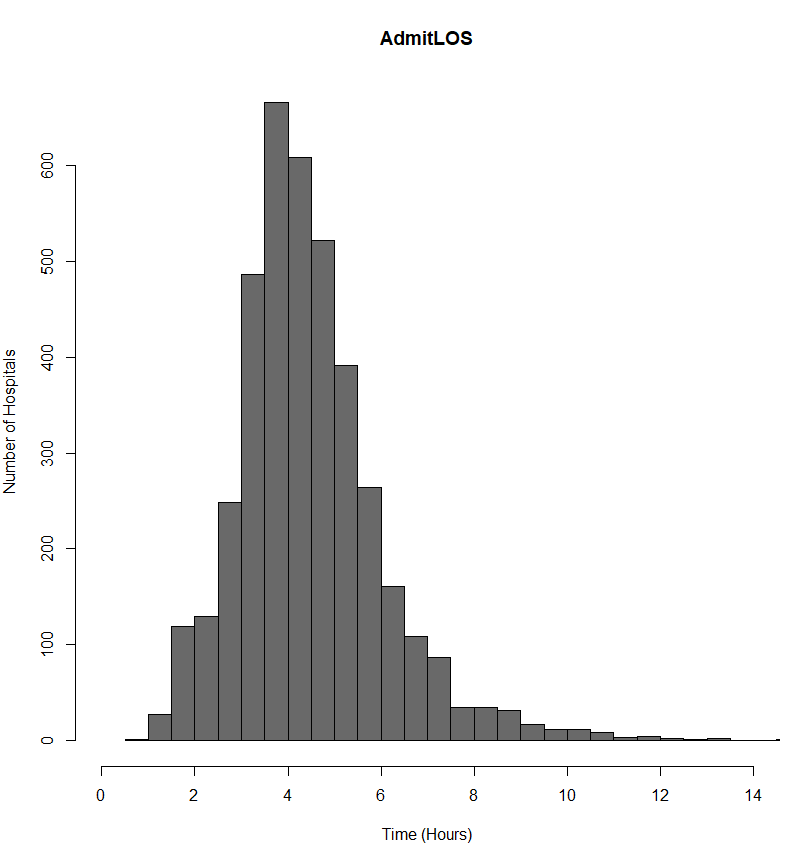
**Exploratory Data Analysis**

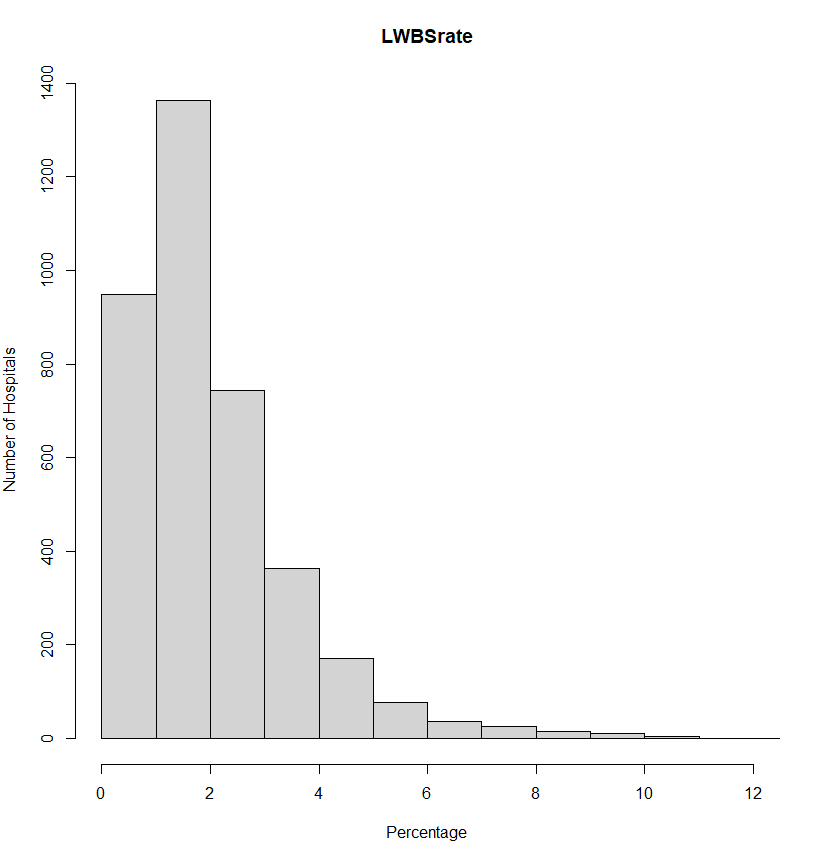
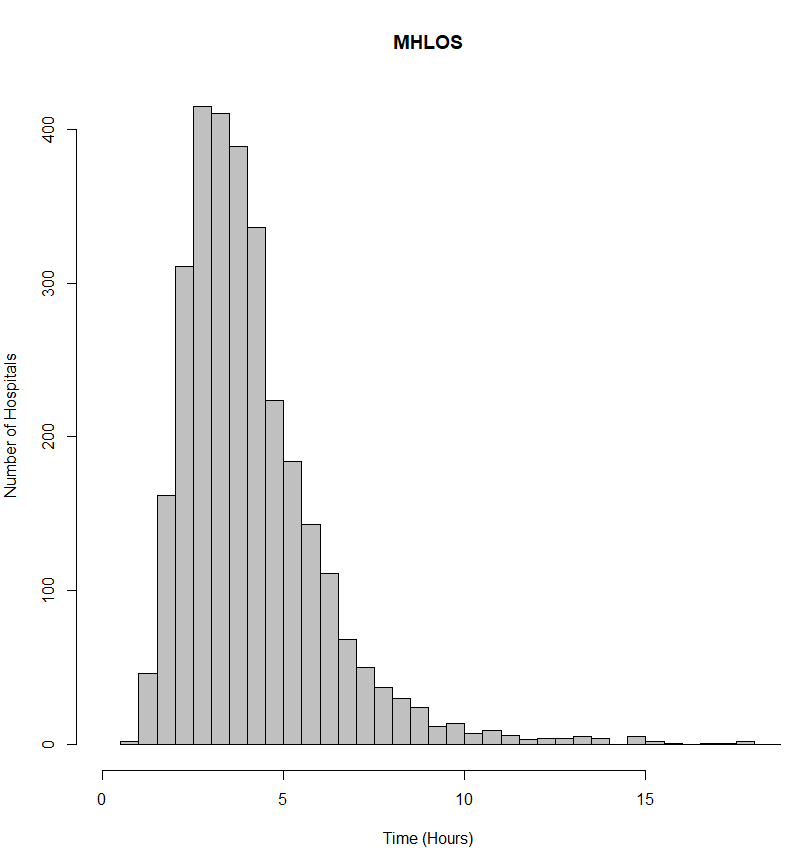
*Response Variables*

Summaries of the five response variables are presented here. Across the United States, the average patient who ultimately is admitted to the hospital spends four and a half hours in the emergency department. That stay includes, on average, a one hour and forty minute wait between the time of the physician decision to admit the patient to the hospital and the time they depart the emergency department for their inpatient bed. For dischanrged patients, the average emergency department length of stay is two hours and twenty minutes—approximately half that of admitted patients. For discharged patients who arrive at the emergency department with a condition related to mental health or substance use, the average length of stay balloons to over four hours. Finally, an average of 1.53% of all emergency department users leave the department without being seen. Summaries by state can be found in the Appendix.

  
 **Table 6 – Summary Statistics for Response Variables**

Histograms of the repsonse variables are also presented, in order to show the shape and skewness of each, and give an idea of which models might be appropriate later on. All distributions of five variables are unimodal, approximately bell-shaped, and right-skewed. For *AdmitLOS* and *MHLOS*, the bulk of the hospital averages fall between two and six hours. For *NonAdmitLOS*, most discharged patients leave after an emergency department stay of one to four hours. With *WaitForBed*, most patients experience between zero and three hours of boarding time. Finally, *LWBSrate*, whose values in the original data were already rounded to the nearest 1%, typically clocks in at between zero three percent.

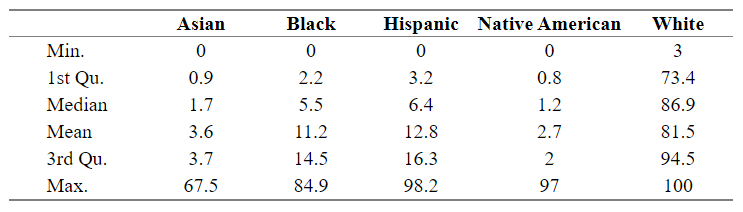


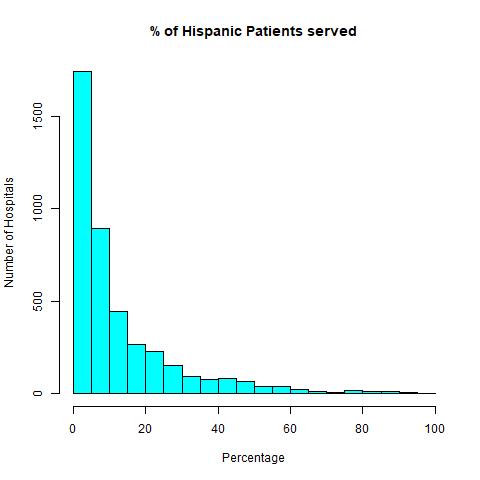
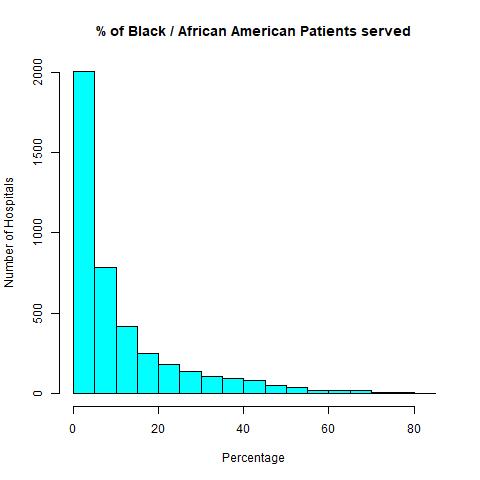
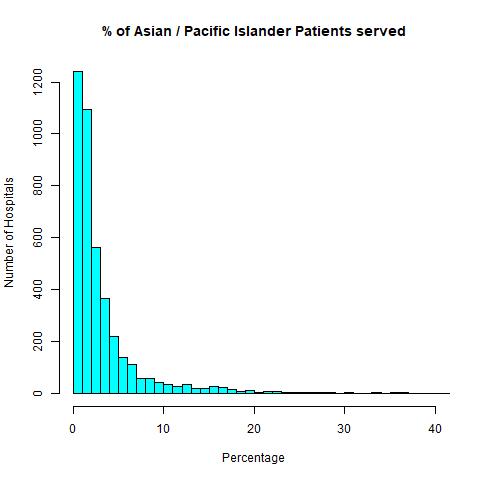
  
**FIGURE 3: Histograms of response variables**

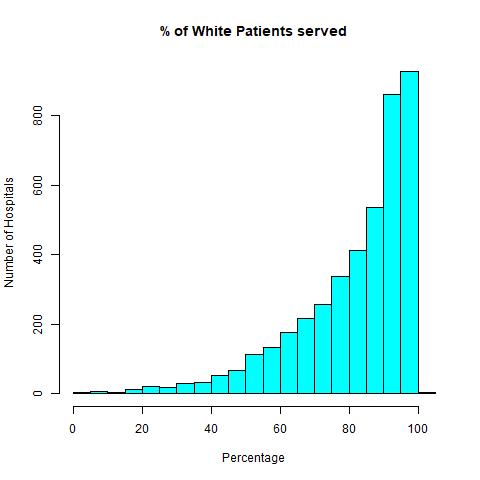
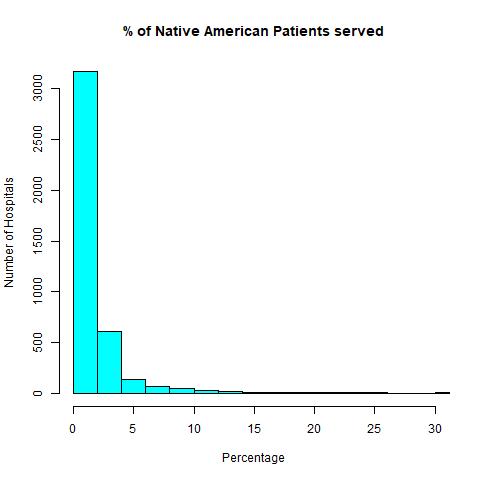
*Race/Ethnicity Variables*

The five race/ethnicity variables in the dataset are summarized here. While data pertaining to less prevalent races and ethnicities are available from the U.S. Census Bureau, because a vast majority of individuals living in the United States identifies with one or more of these groups, the decision is made to limit the scope of this study to these larger race/ethnicity categories.

As shown in Table 3, the typical (median) U.S. hospital serves a population that is 1.7% Asian or Pacific Islander, 5.5% Black or African-American, 6.4% Hispanic or Latino, 1.2% Native American, and 86.9% White. The notion of a “typical” hospital, however, may be misleading. While most U.S. hospitals serve high proportions of white patients, the large gaps between the “3rd Qu.” and “Max” values for each of the four non-white variables suggest that a small number of hospitals serve large proportions of non-white patients. This trend is confirmed by the histograms shown in Figure 5, where each of the non-white variables is heavily right skewed while proportion of white patients served is left skewed.

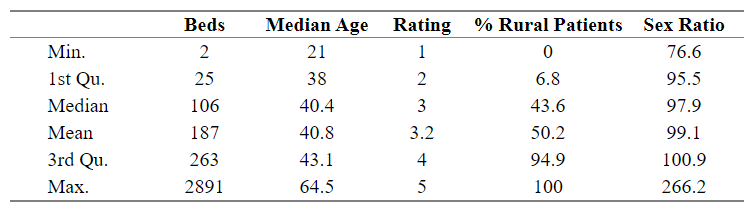
 **Table 3 – Summary Statistics for Race/Ethnicity Variables**

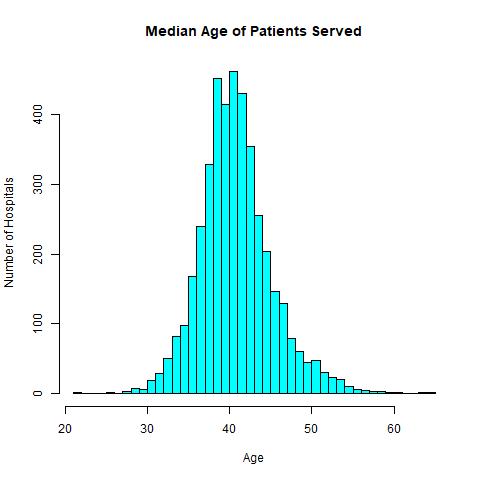
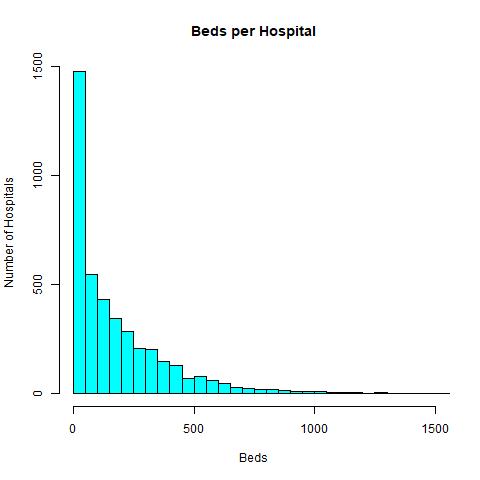


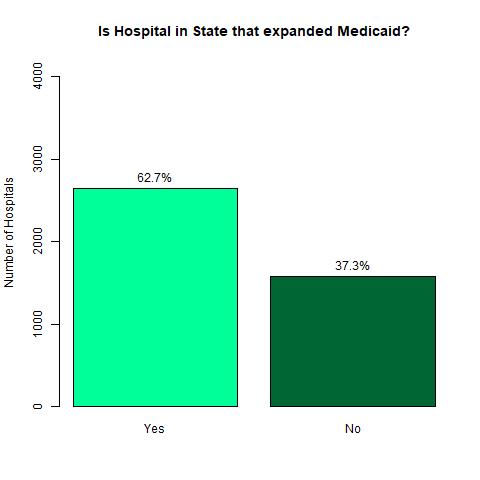
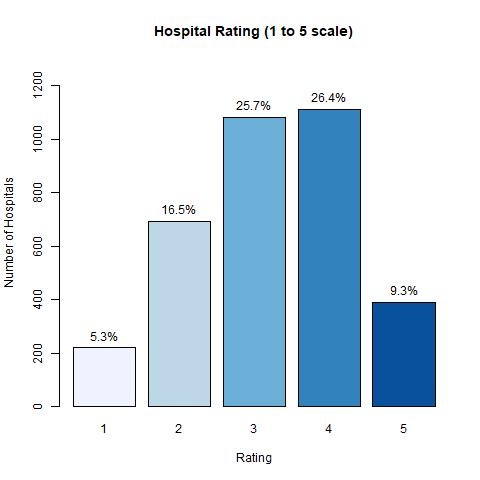
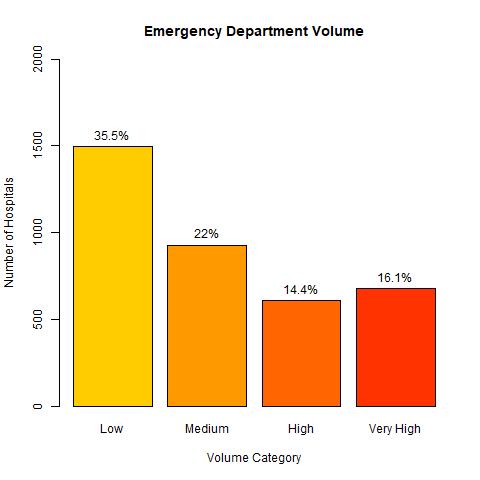
  
**Figure 5 – Race/Ethnicity Variables**

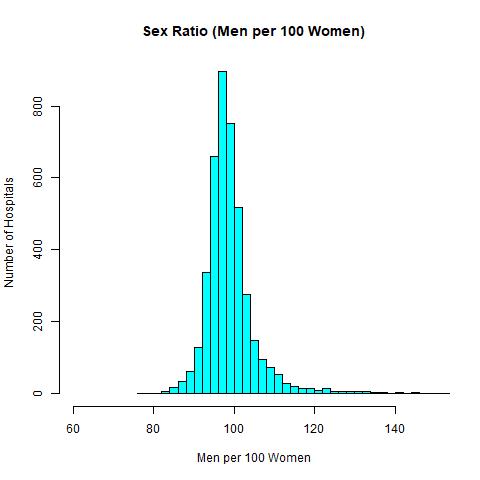
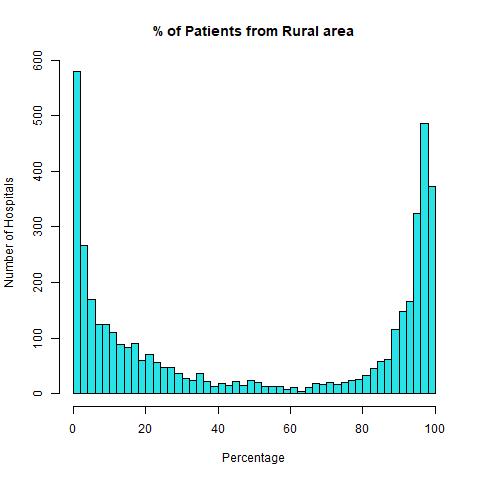
*Other Independent Variables*

The seven remaining independent variables are summarized below. It is here that we see just how widely U.S. hospitals vary in terms of their size, location, and the population they serve. We begin with *Beds*, where average number of Medicare-certified beds across U.S. hospitals is 187, but where the totals range wildly, from the 2-bed Johnson County Community Hospital in Mountain City, Tennessee to the 2891-bed Advent Health Hospital in Orlando, Florida. The histogram forms an exponential shape, with a heavy right skew. Following *Beds* is *Median Age*, the median age of patients served at U.S. hospitals, which appears to the normally distributed and has an average of 40.8 years old. *Rating* appears next, where U.S. hospitals average a score of 3.2 on Medicare’s 1 to 5 rating scale. While *Rating* can take on only five possible values, because its values are ordinal and form an approximately normal distribution in Figure 7, the decision is made to treat it as a numeric variable as opposed to a categorical one for the purposes of analysis. Next is *RuralScore*, where the “typical” (median) hospital serves a 42% rural population. However, as its histogram shows, the data are bimodal and are heavily skewed to the right and left, away from the average, meaning very few hospitals likely fit the “typical” profile. *Sex Ratio*, like *Median Age*, is approximately normally distributed, with the average hospital serving 99.1 men per 100 women. Finally, two of the variables, *Medicaid Expansion* and *ED Volume*, are categorical and so are excluded from Table 5, but can still be found in Figure 7. Their charts show that 62.7% of U.S. hospitals are located in a state that has expanded Medicaid while 37.3% are not, and that two-thirds of hospitals are categorized as having either Low or Medium emergency department volume.

  
**Table 5 – Summary Statistics for Other Independent Variables**







**Figure 7 – Frequency Plots for Other Independent Variables**

**Assessment of Viable Models**

(Gamma vs Gaussian vs Poisson) based on shape of histogram, whether data are integer counts, and whether zero is included.

Text

More Text:

|  |  |
| --- | --- |
| **Response Variable** | **Model Selected** |
| *AdmitLOS* | Gamma GLM with Log Link |
| *WaitForBed* | Negative Binomial GLM |
| *NonAdmitLOS* | Gamma GLM with Identity Link |
| *MHLOS* | Inverse Gaussian GLM with Log Link |
| *LWBSrate* | Beta GLM with Log Link |

Some more text

**Model Diagnostics**

Make sure assumptions are met before presenting results

Describe outlier removal process

Cook’s Distance > 0.5 is concerning

**Model Structure**

Describe decision to have successively more complex models (Models 1, 2, 3, 4)

**Results**

Text and more text