1. Describe a real-world organizational situation or issue in the Data Dictionary you chose, by doing the following:
   1. What characteristics, whether medical or lifestyle based, cause patients to be more likely to be readmitted after being discharged?
   2. By identifying groups that are more likely to be readmitted after being discharged, hospital management can better allocate resources to the patients who are readmittance risks and drive down the number of patients readmitted. As the data dictionary describes, there are regulations concerning the number and frequency of readmittances for each hospital with potential fines for hospitals that exceed these regulatory limits. By reducing the frequency of readmittance, the hospital can avoid fines, prevent monetary losses from over-providing care to lower risk patients, and ensure that groups that are under-served by the current allocation of resources receive proper care and attention.
   3. The data required for this question to be answered will include all diagnostic variables and variables that represent certain lifestyle decisions. Before we identify these variables, first we must identify a suitable index for our selected data. We have four options within the data set which are unique: CaseOrder, Customer\_id, Interaction, and UID. The CaseOrder variable simply represents the initial ordering of the data and provides us little information on the patient themselves. The Interaction and UID variables are more meaningful than the CaseOrder variable, but suffer from long, difficult values. The Customer\_id is a unique, meaningful variable that is short, easy to read, and easy to enter. For these reasons, we will select the Customer\_id variable as our index.

We will begin by discussing which lifestyle variables will provide meaningful information for the question posed. First, it may be useful to know where a person lives, as certain areas may be hotspots for readmittances, indicating that specific hospitals are under-serving their populations and require more funding, investigation, or some other remedy. We have many options for locations, but not all are equally useful or necessary for this investigation. We will begin with the State variable, providing us with a general overview of how quality of service varies regionally. We choose the State variable in favor of the City, County, Zip, or Lat and Lng variables because it provides a way to group patients without creating too many groups to analyze efficiently. The City and County variables contain thousands of unique entries, while there are only 52 categories within the State variable. The continuous nature of the Zip, Lat, and Lng variables does not allow them to depict the relationship as a categorical variable, so we will ignore these. Besides where someone lives, it will be useful for us to know some characteristics of this area as well. The Area and Population variables allow us some insight into the characteristics of the location in which a patient lives and provide an opportunity for subcategorization. For example, patients can be grouped by which state they live in and further divided by the area within that state, rural, urban, or suburban.

Continuing with lifestyle variables, it is pertinent for us to know more about the patient's personal life beyond where they live. These characteristics include how many children a patient has, represented by the Children variable, how old a patient is, represented by the Age variable, the patient's income, represented by the Income variable, the patient's marital status, represented by the Marital variable, the patient's gender, represented by the Gender variable, and the patient's relationship with soda, represented by the Soft\_drink variable. These variables paint a picture of the patient's life that will allow us to make meaningful observations regarding the chance a patient is readmitted.

The next group of variables we will consider involves the patient's medical history. The data dictionary provides us with many diagnoses that a patient may have, including: HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, and Asthma. In addition to these diagnoses, we will include observations of the patient's health upon admittance, namely the VitD\_levels and Complication\_risk variables.

Finally, some information about the patient's stay in the hospital could prove useful. These variables include the variables Doc\_visits, Initial\_admin, Services, and Initial\_days. The Full\_meals\_eaten variable could be useful here if combined with the Initial\_days variable to show how many meals the patient ate each day in the hospital.

Most importantly, we must also include the ReAdmis variable to test for relationships between the previously mentioned variables and whether a patient was readmitted.

1. Describe the data analysis by doing the following:
   1. We begin our exploration of the data by dividing the selected variables into data.frame objects depending on whether the variables are numerical or categorical. We identify this distinction by exploring the output of the summary function on the data.frame containing all the selected variables. We include the Customer\_id and ReAdmis variables in both newly created data.frame objects. The Customer\_id variable is used as an index and the ReAdmis variable is used to divide the populations of each variable into two samples, the patients that were readmitted and those that were not.

In order to answer the question presented in Part A.1, we will evaluate each variable for a difference between populations that were readmitted and those that were not. Beginning with our categorical variables, we will use a Chi-Squared test to determine whether the populations have any significant differences. The function we use to perform a Chi-Square test in R is chisq.test. An additional parameter of the chisq.test function is employed: simulate.p.values. This parameter allows the function to perform a Monte Carlo simulation, accounting for the inherent randomness of reality and providing more accurate results from the Chi-Squared test (Dizikes, 2010). This operation is performed for each variable, using the population of patients who were not readmitted compared with the population of patients who were by creating a table where the columns are the values "Yes" and "No" from the ReAdmis variable, the rows are the observations from the variable being examined, and the values are the counts of how many patients with the variable value of each row exist in each of the two populations. These results are printed for each variable with a for loop.

While a Chi-Squared test was useful for analyzing the categorical variables, the same test would prove useless with our numerical variables. Instead, we employ a two-sample T-test, where our first sample is patients who were readmitted, and the second sample is those patients who were not. We accomplish this test by calling the function t.test on each variable, divided into the two samples by the tilde (~) function. These results are printed for each variable with a for loop.

The code used to accomplish this analysis is provided below and in the accompanying R file.

library(dplyr)

library(ggplot2)

setwd("D:/Documents/WGU\_MSDA/OEM2")

data <- read.csv(file="medical\_clean.csv", stringsAsFactors = TRUE)

data\_sel <- select(data, Customer\_id, ReAdmis, State,

Area, Population, Children, Age, Income, Marital, Gender, Soft\_drink, HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, VitD\_levels, Complication\_risk, Doc\_visits, Initial\_admin, Services, Initial\_days, Full\_meals\_eaten)

summary(data\_sel)

data\_sel$Meals\_per\_day <- data\_sel$Full\_meals\_eaten /

data\_sel$Initial\_days

data\_sel\_cat <- select(data\_sel,Customer\_id, ReAdmis,

State, Area, Marital, Gender, Soft\_drink, HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, Complication\_risk, Initial\_admin, Services)

data\_sel\_num <- select(data\_sel, Customer\_id, ReAdmis,

Population, Children, Age, Income, VitD\_levels, Doc\_visits, Initial\_days, Meals\_per\_day)

for (i in 3:21) {

cont\_tbl <- table(data\_sel\_cat[,i],

data\_sel\_cat$ReAdmis)

cat(colnames(data\_sel\_cat)[i], "\n")

print(chisq.test(cont\_tbl, simulate.p.value = TRUE))

}

for (i in 3:10) {

cat(colnames(data\_sel\_num)[i], "\n")

print(t.test(data\_sel\_num[,i]~data\_sel\_num$ReAdmis))

}

* 1. The output of the Chi-Squared tests provides us with p-values for each variable that indicate how likely the difference of the population means is to occur if they were not correlated. Our null hypothesis is that there is no correlation between the values of each variable and the patient's readmission. We will conduct these tests with a confidence interval of 95%, meaning any p-value greater than 0.05 is interpreted as statistically insignificant. The output of the Chi-Squared tests shows that only the categorical variable Services seems to impact patient readmittance. The only other variable close to our desired p-value is Asthma, which has a p-value of approximately 0.09. The output from R has been provided below.

State

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 46.758, df = NA, p-value = 0.6512

Area

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.71331, df = NA, p-value = 0.7136

Marital

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 5.0852, df = NA, p-value = 0.2679

Gender

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 1.5858, df = NA, p-value = 0.4388

Soft\_drink

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.5933, df = NA, p-value = 0.4543

HighBlood

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.051531, df = NA, p-value = 0.8261

Stroke

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.0084356, df = NA, p-value = 0.9365

Overweight

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.73719, df = NA, p-value = 0.4063

Arthritis

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.58722, df = NA, p-value = 0.4488

Diabetes

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.093522, df = NA, p-value = 0.7651

Hyperlipidemia

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.18549, df = NA, p-value = 0.6642

BackPain

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 1.7723, df = NA, p-value = 0.1919

Anxiety

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.057901, df = NA, p-value = 0.8206

Allergic\_rhinitis

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.21629, df = NA, p-value = 0.6642

Reflux\_esophagitis

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.29396, df = NA, p-value = 0.6032

Asthma

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 2.9353, df = NA, p-value = 0.09445

Complication\_risk

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 0.15902, df = NA, p-value = 0.932

Initial\_admin

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 3.89, df = NA, p-value = 0.1339

Services

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: cont\_tbl

X-squared = 8.8926, df = NA, p-value = 0.03098

The output of our T-tests provides us with p-values for the probability that the mean values of the two samples, those that were readmitted and those that were not, could occur if readmission was not dependent on the variable being tested. Our null hypothesis is that the difference between the sample means for each sample is zero. Our significance level is 95%, meaning any p-value greater than 0.05 is interpreted as statistically insignificant. The output of these tests shows that the variables Population, Children, Initial\_days, and Meals\_per\_day are all statistically significant. Most notably, Meals\_per\_day and Initial\_days have p-values that are extremely close to zero, indicating that these variables have a significant effect on a patient's chance to be readmitted. The output from R has been provided below.

Population

Welch Two Sample t-test

data: data\_sel\_num[, i] by data\_sel\_num$ReAdmis

t = -1.9715, df = 7280.9, p-value = 0.0487

alternative hypothesis: true difference in means between group No and group Yes is not equal to 0

95 percent confidence interval:

-1229.243235 -3.515288

sample estimates:

mean in group No mean in group Yes

9739.104 10355.484

Children

Welch Two Sample t-test

data: data\_sel\_num[, i] by data\_sel\_num$ReAdmis

t = -2.3314, df = 7439.7, p-value = 0.01976

alternative hypothesis: true difference in means between group No and group Yes is not equal to 0

95 percent confidence interval:

-0.19445621 -0.01681343

sample estimates:

mean in group No mean in group Yes

2.058443 2.164077

Age

Welch Two Sample t-test

data: data\_sel\_num[, i] by data\_sel\_num$ReAdmis

t = -1.5837, df = 7700, p-value = 0.1133

alternative hypothesis: true difference in means between group No and group Yes is not equal to 0

95 percent confidence interval:

-1.5149612 0.1609549

sample estimates:

mean in group No mean in group Yes

53.26331 53.94031

Income

Welch Two Sample t-test

data: data\_sel\_num[, i] by data\_sel\_num$ReAdmis

t = 1.1527, df = 7709.3, p-value = 0.2491

alternative hypothesis: true difference in means between group No and group Yes is not equal to 0

95 percent confidence interval:

-476.8796 1838.3511

sample estimates:

mean in group No mean in group Yes

40740.26 40059.52

VitD\_levels

Welch Two Sample t-test

data: data\_sel\_num[, i] by data\_sel\_num$ReAdmis

t = -0.40716, df = 7599.4, p-value = 0.6839

alternative hypothesis: true difference in means between group No and group Yes is not equal to 0

95 percent confidence interval:

-0.09935129 0.06517780

sample estimates:

mean in group No mean in group Yes

17.95799 17.97508

Doc\_visits

Welch Two Sample t-test

data: data\_sel\_num[, i] by data\_sel\_num$ReAdmis

t = -0.024523, df = 7619, p-value = 0.9804

alternative hypothesis: true difference in means between group No and group Yes is not equal to 0

95 percent confidence interval:

-0.04314380 0.04207769

sample estimates:

mean in group No mean in group Yes

5.012004 5.012537

Initial\_days

Welch Two Sample t-test

data: data\_sel\_num[, i] by data\_sel\_num$ReAdmis

t = -203.01, df = 8109.5, p-value < 2.2e-16

alternative hypothesis: true difference in means between group No and group Yes is not equal to 0

95 percent confidence interval:

-46.89328 -45.99634

sample estimates:

mean in group No mean in group Yes

17.41470 63.85951

Meals\_per\_day

Welch Two Sample t-test

data: data\_sel\_num[, i] by data\_sel\_num$ReAdmis

t = 36.969, df = 6383, p-value < 2.2e-16

alternative hypothesis: true difference in means between group No and group Yes is not equal to 0

95 percent confidence interval:

0.1408447 0.1566180

sample estimates:

mean in group No mean in group Yes

0.16426493 0.01553362

* 1. Because our selection of variables includes both categorical and numeric variables, all compared to the categorical variable ReAdmis, one test does not suffice for analyzing these variables. The T-test works with numerical data and is adaptable to having two samples (Bruce et al., 2020, p. 111). Because the ReAdmis variable is a binary categorical variable, we can divide our dataset along this binary to create two samples. These samples can be evaluated effectively with a T-test. Alternatively, if the dependent variable had been a variable with more than two categories, an ANOVA test would have been ideal. As this was not the case, the T-test best fits for the analysis of our numerical data. It is important to discuss the inclusion of the Monte Carlo simulation in our implementation of the T-test. While statistics and probability are wonderful at working with deterministic outcomes, reality is often far too messy to be considered as such. The Monte Carlo simulation adds a small level of noise that accounts for other variables which may be exerting their influence and the random chance that is a reality of life. While this simulation makes the outcomes different each time, they differ only very slightly.

For comparing two categorical variables, the T-test is not applicable. The ANOVA test is performed with one numeric variable and one categorical variable. This leaves only the Chi-Squared test for us to explore. The Chi-Squared test works with count data rather than numerical inputs (Bruce et al., 2020, p.124). As we are comparing categorical variables with another categorical variable, ReAdmis, it is the test best suited to our needs.

By using both the T-test and the Chi-Squared test, we are able to analyze both the numeric and categorical variables in our selected data set.

1. We will be exploring the distributions of the numeric variables Age and Income, and the categorical variables Area and Initial\_admin with univariate statistics. In order to discover the distributions of our numeric variables, we will use the summary function in conjunction with box plots. To discover the distributions of our categorical variables, we will use the table function in along with histograms.

The summary of the Age variable shows that all patients are between the ages of 18 and 89. The mean of the ages is 53.51 years old, and the median is 53 years old. The summary of the Income variable shows that patients earn between $154.10 and $207,249.10 annually. The mean income is $40,490.50 and the median income is $33,768.40. The code used to obtain this output is included below and the resulting box plots are found in Part C.1.

The table function, when called on the Area variable, reveals that 3,369 patients live in rural areas, 3,328 patients live in suburban areas, and 3,303 patients live in urban areas. This distribution is nearly even, with each area representing roughly one-third of the sample. The table of Initial\_admin values shows that 2,504 patients were elective admissions, 5,060 patients were emergency admissions, and 2,436 patients were admitted for observation. Observation admittance and elective admittance are approximately equal, while emergency admittance is twice as likely as either alternative. The code used to obtain this output is included below and the resulting bar graphs are found in Part C.1.

summary(data\_sel$Age)

summary(data\_sel$Income)

boxplot(data\_sel$Age)

boxplot(data\_sel$Income)

table(data\_sel$Area)

table(data\_sel$Initial\_admin)

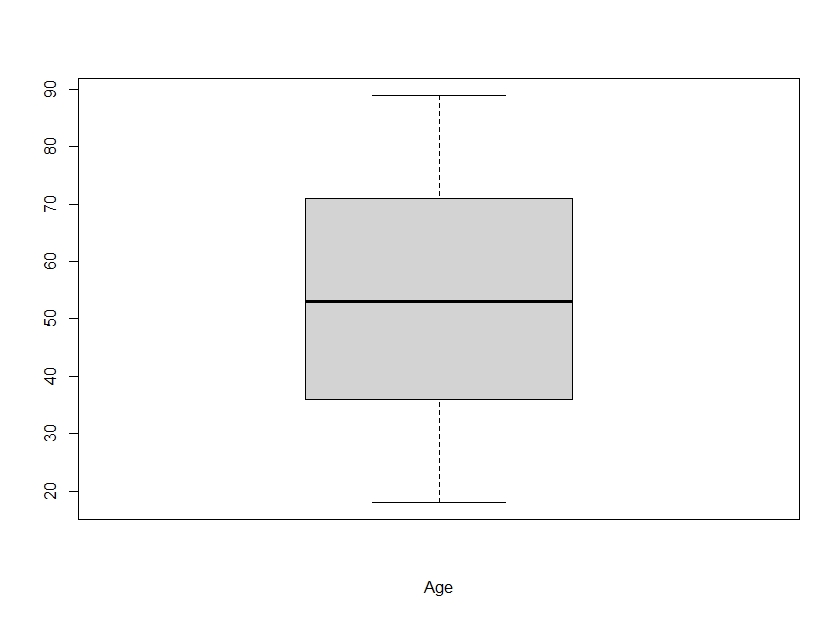
ggplot(data\_sel, aes(x=Area)) +

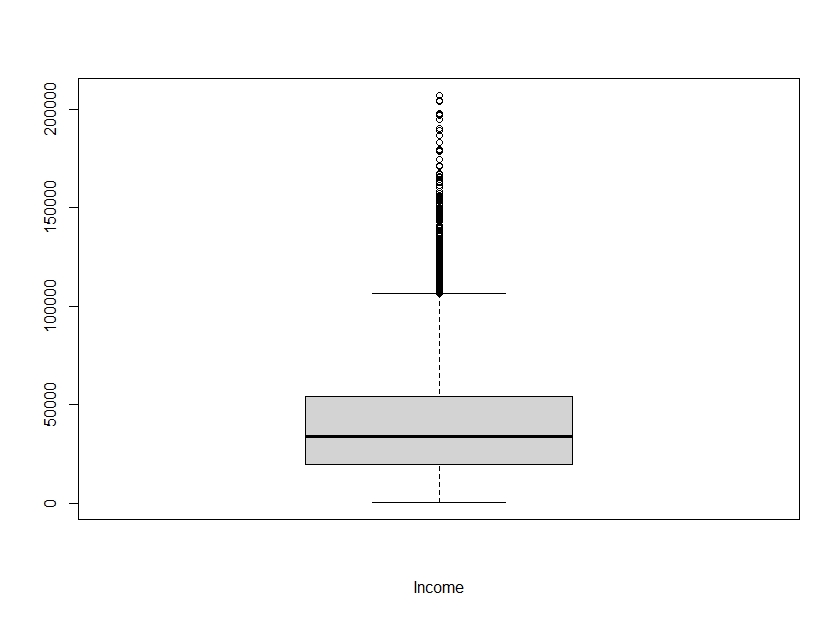
geom\_bar()

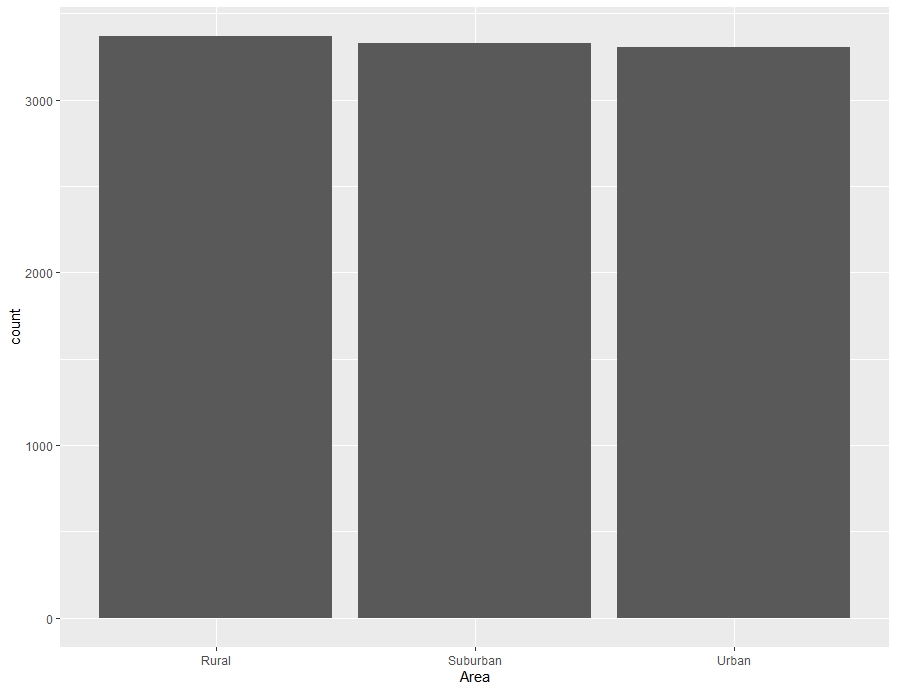
ggplot(data\_sel, aes(x=Initial\_admin)) +

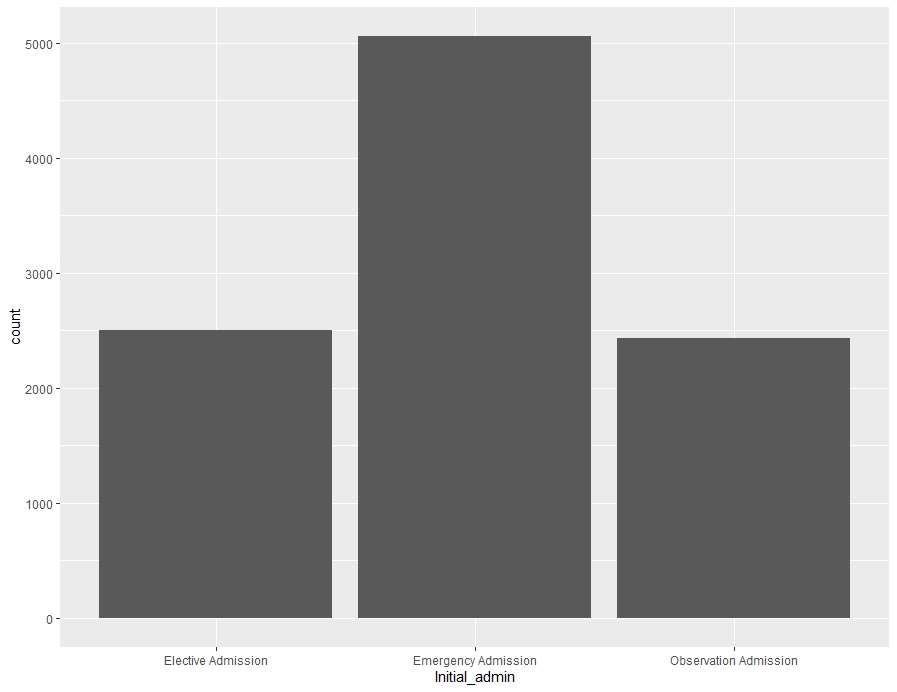
geom\_bar()

* 1. Represent your findings in Part C, visually as part of your submission.









1. Our bivariate analysis will explore the distributions of the numeric variable Initial\_days, the numeric variable Population, the categorical variable State, and the categorical variable Service when grouped by the binary categorical value ReAdmis. We will explore the distributions of the categorical variables and their corresponding ReAdmis values with stacked bar graphs. To better visualize the difference between groups, some of these variables will have a traditional bar chart and a bar chart with the position parameter set to "fill," showing us the proportion of readmitted patients in each level. We will examine the numeric variables using stacked histograms with corresponding filled histograms, where the position is set to "fill" when necessary to better visualize the proportions of readmitted patients.

The histogram of the Initial\_days variable grouped by ReAdmis shows a clear relationship between days spent in the hospital and chance of readmittance. There appear very few, possibly even zero, patients who were readmitted after spending less than 50 days in the hospital in their initial stay. Every patient who stayed 60 or more days was eventually readmitted.

The histogram of the population variable shows that a vast majority of patients live in an area with zero population, presumably excluding themselves, or a population of nearly zero. This conflicts with the distribution of the Area variable discussed in Part C as patient in non-rural areas, where populations are higher, account for two-thirds of all patients. Because of the extreme left-skew of this variable, a histogram with position set to "fill" is created to better visualize the proportions of readmitted patients. In this histogram, we see that a relationship between population and readmission may exist as patients that live in higher population areas are readmitted at a much higher rate. Because the populations that appear to be more likely to be readmitted are such a small portion of the data set, this correlation may be incidental.

In the bar graph of the State variable, we see that the states, or territories, Delaware, Washington D.C., and Rhode Island are severely under-represented. California, New York, Pennsylvania, and Texas are the most common states in the data set by a large margin. The distribution makes it difficult to visualize the proportions of readmitted patients, so another bar graph is created with position set to "fill." This bar graph shows that the states Connecticut, Washington D.C., Rhode Island, and South Carolina have the greatest proportion of readmitted patients. However, the variance of the distribution could be the source of this perceived increase in likelihood, as Washington D.C. and Rhode Island are among the least represented states and the highest proportions of readmitted patients in the data set.

Looking at the bar graph of the Services variable, we see that most patients received blood work, about a third of patients received intravenous care, a small portion received a CT scan, and a very small minority received an MRI. Another bar graph with position set to "fill" shows us that patients who received a CT scan or MRI were more likely than other patients to be readmitted, though only by a small margin. Because the groups that were least represented in the dataset were the most likely, the conclusion that patients receiving these services are more likely to be readmitted requires further investigation.

The code used to generate the histograms and bar charts is included below and the resulting graphs are provided in Part D.1.

ggplot(data\_sel, aes(fill=ReAdmis, x=Initial\_days)) +

geom\_histogram(bins=35)

ggplot(data\_sel, aes(fill=ReAdmis, x=Population)) +

geom\_histogram(position='fill', bins=12)

ggplot(data\_sel, aes(fill=ReAdmis, x=Population)) +

geom\_histogram(bins=12)

ggplot(data\_sel, aes(fill=ReAdmis, x=State)) +

geom\_bar(position='fill')

ggplot(data\_sel, aes(fill=ReAdmis, x=State)) +

geom\_bar()

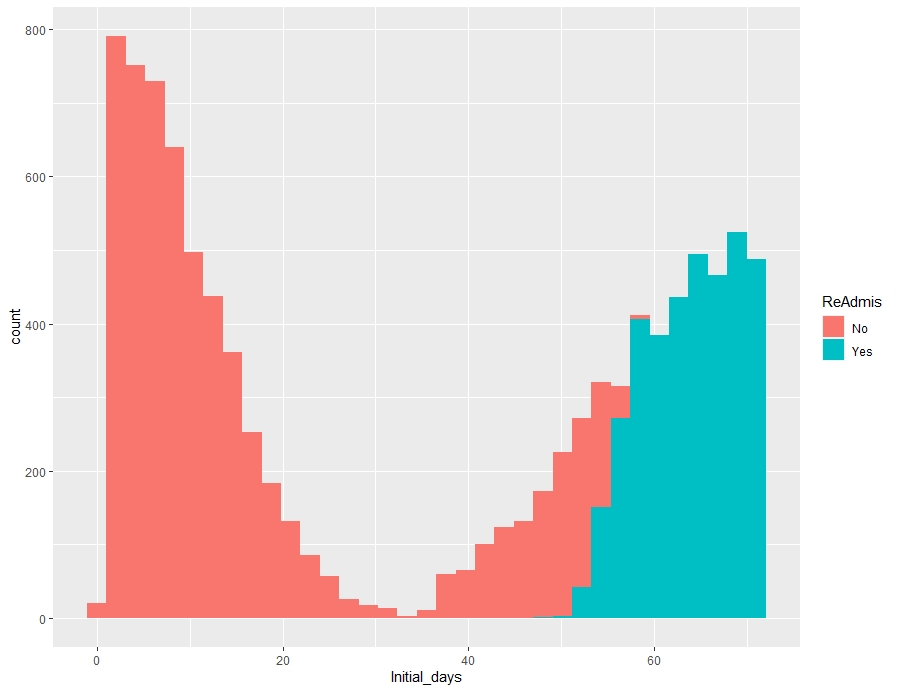
ggplot(data\_sel, aes(fill=ReAdmis, x=Services)) +

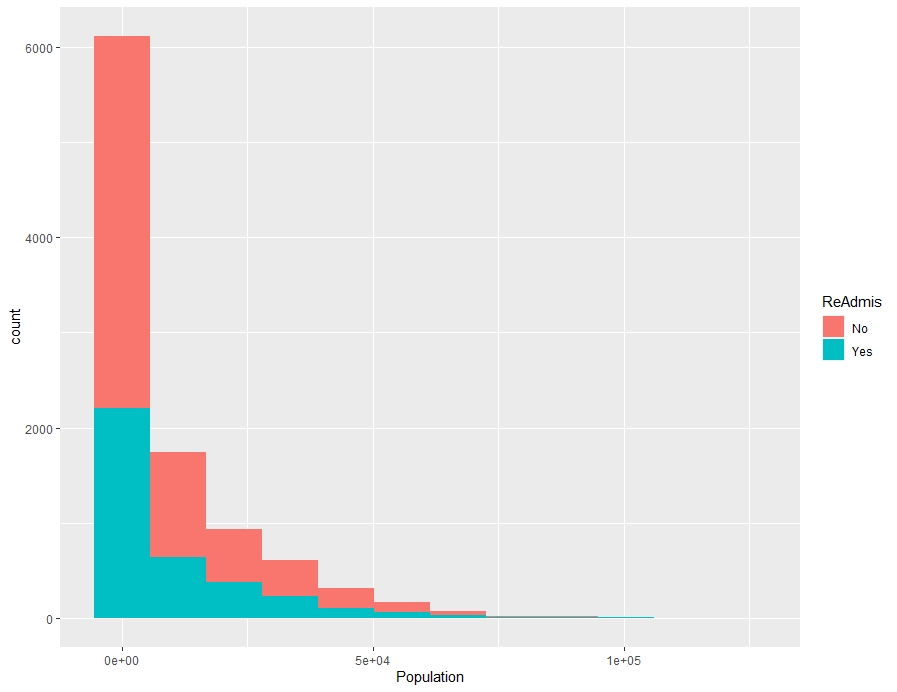
geom\_bar(position='fill')

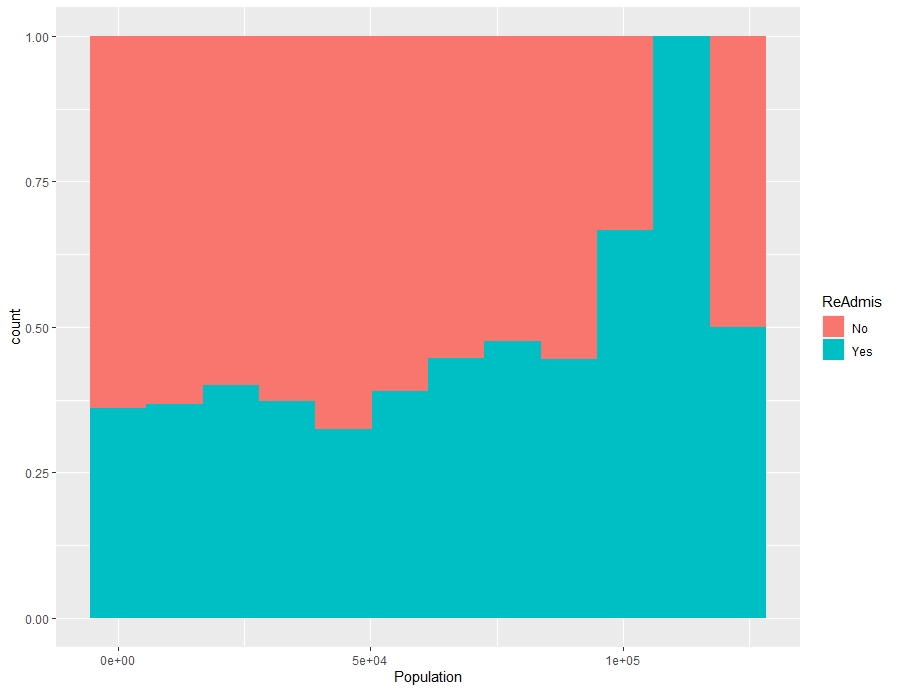
ggplot(data\_sel, aes(fill=ReAdmis, x=Services)) +

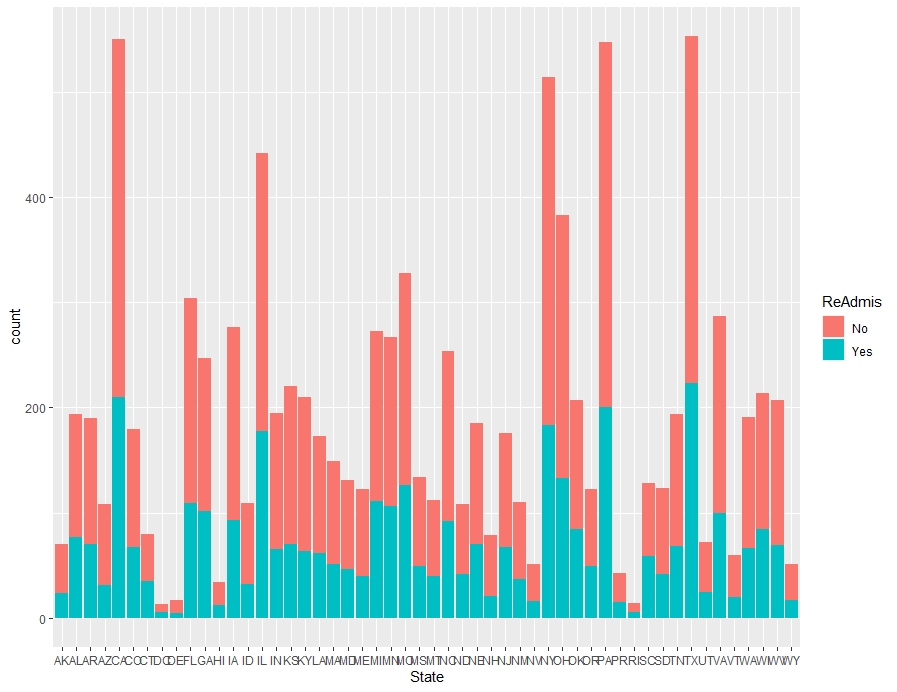
geom\_bar()

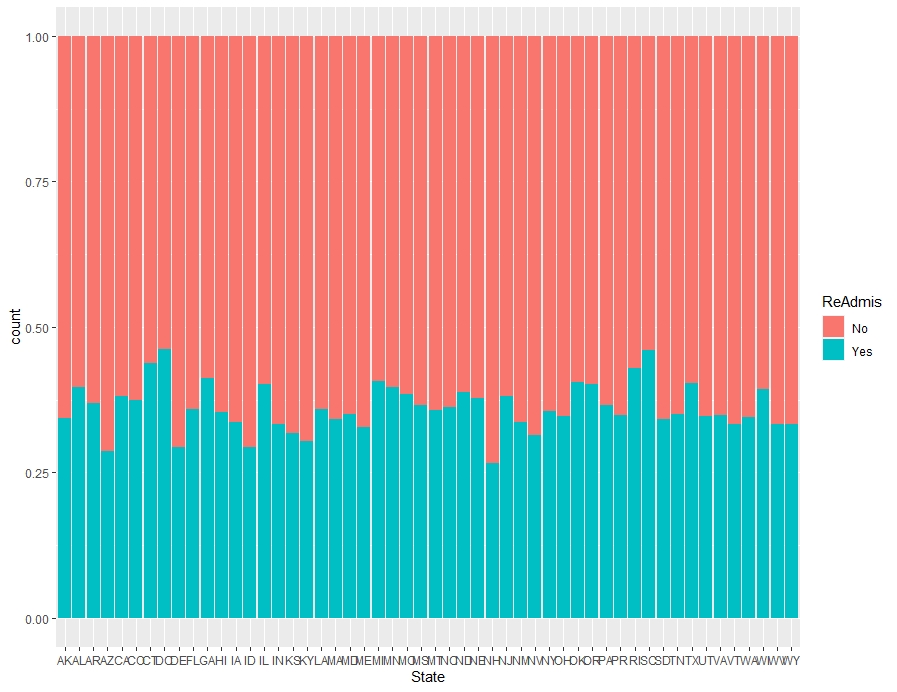
* 1. Represent your findings in Part D, visually as part of your submission.

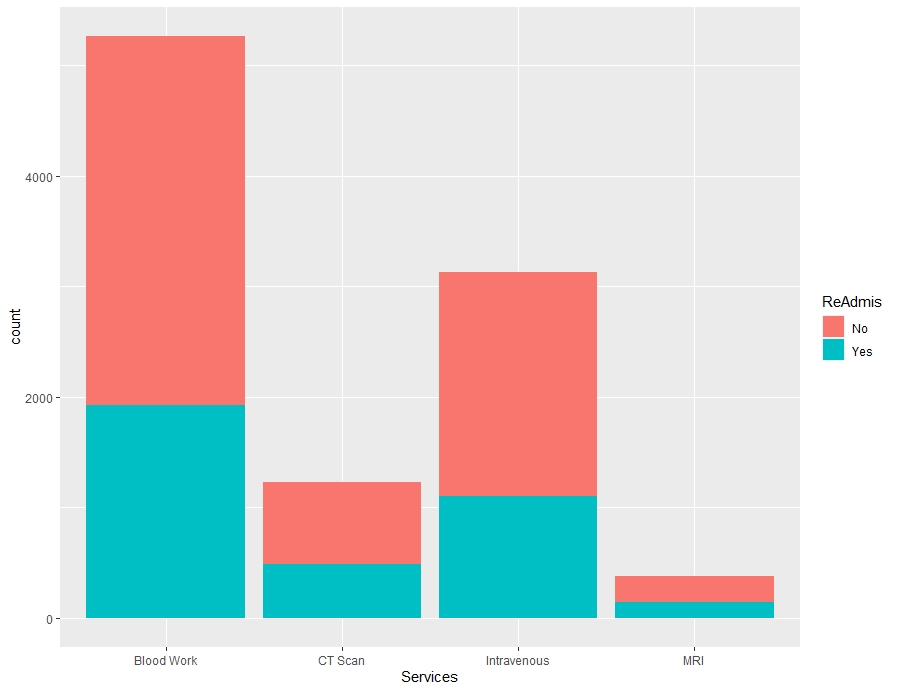


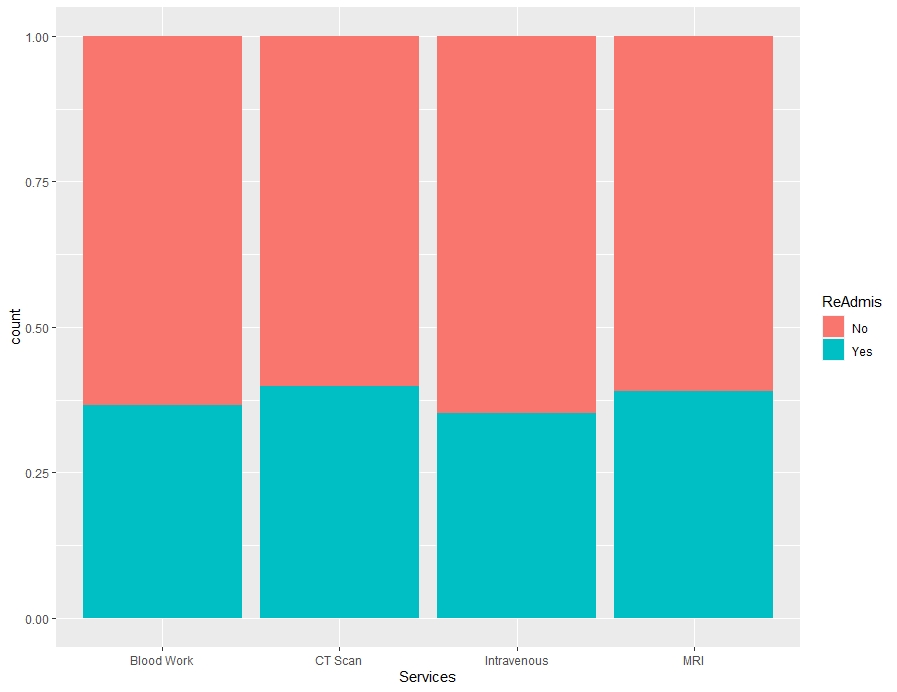












1. Summarize the implications of your data analysis by doing the following:
   1. The hypothesis tests reveal that many of the selected variables have only marginal correlations to the likelihood that a patient is readmitted. The variables determined to have a significant impact on patient readmittance are Population, Children, Initial\_days, Meals\_per\_day, and Services.
   2. Our analysis is most limited by the nature of the data itself. Many variables are not well distributed, leading to a lower confidence in the relationships they shed light on. For example, the Population variable appears to have a strong correlation with readmittance when viewed in the bar graph where position is set to "fill"; however, when the "fill" parameter is removed, we see that very few patients live in an area with a population of more than 25,000. Because there are so few data points above this value, any relationship observed above this level is subject to scrutiny and is not actionable.

Additionally, correlation does not imply causation. Some variables in this data set do indicate some form of correlation, especially the Initial\_days variable; however, we cannot be certain that this correlation is anything more than relationship. It is possible that patients who are readmitted and stay for a long period on their initial visit do not return because of the length of their stay. For example, patients with a serious illness or traumatic injury may be more likely to require longer stays in the hospital and be more likely to be readmitted. The causal variable in this example is not the length of the initial stay but the illness or injury; length of stay and readmittance are simply symptoms of this outside cause. Without the ability to know absolutely everything, it is difficult to ascertain causal relationships from correlation alone.

* 1. It is my recommendation that hospitals focus on ensuring that patients with long initial stays be granted higher priority in their diagnosis and treatment. This could come in the form of faster care, achieved through better technology, more personnel, or better trained personnel, or in the form of better care, achieved through more thorough diagnostic tests, longer observation periods post care, or more advanced treatment plans. Likely, a combination of these efforts will be required to address the needs of all patients.

1. <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=9a8b66b8-3008-4801-85d0-ad8e000a0f96>
2. No outside sources were used to generate code for this project and all code was written personally by the student.
3. Bibliography

Bruce, P., Bruce, A. G., & Gedeck, P. (2020). *Practical statistics for data Scientists: 50+ essential concepts using R and Python*. O'Reilly.

Dizikes, P. (2010, May 17). *Explained: Monte Carlo simulations*. MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2010/exp-monte-carlo-0517.