D207: Exploratory Data Analysis

Lucas Benish

Western Governors University

# Describe a real-world organizational situation or issue in the Data Dictionary you chose, by doing the following:

## Provide **one** question that is relevant to your chosen data set. You will answer this question later in the task through an analysis of the cleaned data, using one of the following techniques: chi-square, t-test, or analysis of variance (ANOVA).

Hospital readmissions are known to be an issue within the health care system. My employer is looking to reduce their readmission rate to improve the quality of care that they provide and reduce budget constraints from fines. Therefore, the question that I am interested in addressing is as follows:

Is there a relationship between a patient’s level of medical complexity and the incidence of readmission?

## Explain how stakeholders in the organization could benefit from an analysis of the data.

If it is determined that there is a relationship between the levels of medical complexity and readmission to the hospital, the hospital may be able to decrease the readmission rates of specific patients given they can determine what level of medical complexity the patient falls into. Furthermore, the hospital can decrease the fines they pay by proactively addressing readmission rates with at-risk patients.

Additionally, if it is determined that there is a relationship between medical complexity and readmission status, there may be further analysis warranted to explore what other variables could be associated with the different levels of medical complexity, and therefore possibly with readmission incidence.

## Identify *all* of the data in your data set that are relevant to answering your question in part A1

The categories that are immediately relevant to answering this question are as follows:

Complication\_risk

ReAdmis

HighBlood (this was seen in PCA in D206 to have a large weighting in the 2nd PCA)

# Describe the data analysis by doing the following:

## Using one of the following techniques, write code (in either Python or R) to run the analysis of the data set:

## Chi-square

## T-test

## ANOVA

The code that I utilized prior to completion and for performing a Chi-squared test is as follows:

***Setting up the environment***

library(dplyr)

library(dtplyr)

library(naniar)

library(tidyverse)

library(visdat)

med\_clean <- read\_csv(‘C://Users//lgben//OneDrive//Desktop//MSDA//D207 – Exploratory Data Analysis//med\_clean\_data.csv’)

***Exploring relationship between Complication Risk and Readmission***

comprisk\_readmis\_df <- med\_clean %>%

group\_by(Complication\_risk) %>%

summarize(ReAdmis)

comprisk\_readmis <- med\_clean %>%

group\_by(Complication\_risk) %>%

summarize(ReAdmis) %>%

table()

***Visualizing relationship between Complication Risk and Readmission***

ggplot(comprisk\_readmis\_df, aes(Complication\_risk, fill=ReAdmis)) +

geom\_bar()

ggplot(comprisk\_readmis\_df, aes(Complication\_risk, fill=ReAdmis)) +

geom\_bar(position=’fill’)

***Performing Chi-Squared Test***

degfreedom\_comprisk\_readmis <- (ncol(comprisk\_readmis) – 1) \* (nrow(comprisk\_readmis) – 1)

comprisk\_readmis\_chisq <- chisq.test(comprisk\_readmis)

## Provide the output and the results of *any* calculations from the analysis you performed.

The output from the dataframe and table creation code was as follows:

A dataframe containing 2 columns and 10,000 rows, containing information regarding complication risk (column 1) and readmission (column 2).

A table with the following output:

Readmis

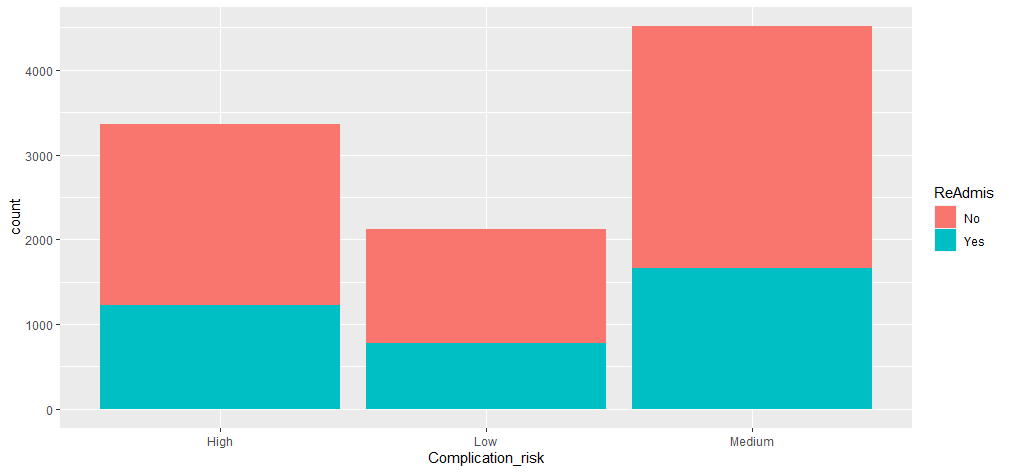
Complication\_risk No Yes

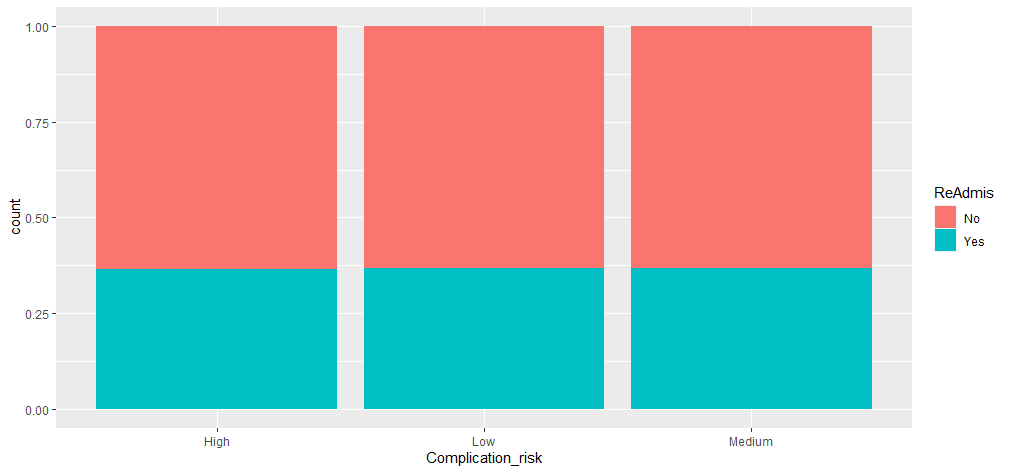
High 2135 1223

Low 1343 782

Medium 2853 1664

The outputs from my visualization code were as follows:





The output from the Chi-squared test is as follows:

Pearson’s Chi-squared test

data: comprisk\_readmis

X-squared = 0.15902, df=2, p-value = 0.9236

## Justify why you chose this analysis technique.

I decided to use a chi-squared analysis because the variables that I was interested in exploring were both categorical in nature. Complication risk has 3 levels to it: “Low,” “Medium,” and “High.” Readmission has 2 levels, simply a “Yes” or “No.” Because there are more than 2 values of interest, a t-test would not be appropriate. Furthermore, because the variables are not numeric, neither a t-test nor ANOVA would be appropriate.

# Identify the distribution of **two** continuous variables and **two** categorical variables using univariate statistics from your cleaned and prepared data.

The variables that I decided to explore for this portion of the task are Initial\_days, VitD\_levels, Initial\_admin, and Marital.

Initial\_days and VitD\_levels are both continuous variables, as they can have “partial values” that fall between whole numbers. These values can lie upon a theoretically infinite continuum. For these two variables, I believed that histograms would best represent them. Initial\_days was noted to be bimodal, with a left>right mode noted. VitD\_levels follows a relatively normal distribution.

The two categorical variables that I decided to explore are Initial\_admin and Marital. Initial\_admin describes what initially led to a patient’s hospitalization, while Marital describes the patient’s marital status. Initial\_admin has an obvious mode, accounting for almost half of the observations. Marital status was noted to be approximately evenly distributed across all the options.

## Represent your findings in Part C, visually as part of your submission

The code that I used to create the visualizations for my variables of interest from C is as follows:

***Univariate EDA***

*Continuous Variables*

ggplot(med\_clean, aes(Initial\_days)) +

geom\_histogram(binwidth = 2)

ggplot(med\_clean, aes(VitD\_levels)) +

geom\_histogram(binwidth = 0.5)

*Categorical Variables*

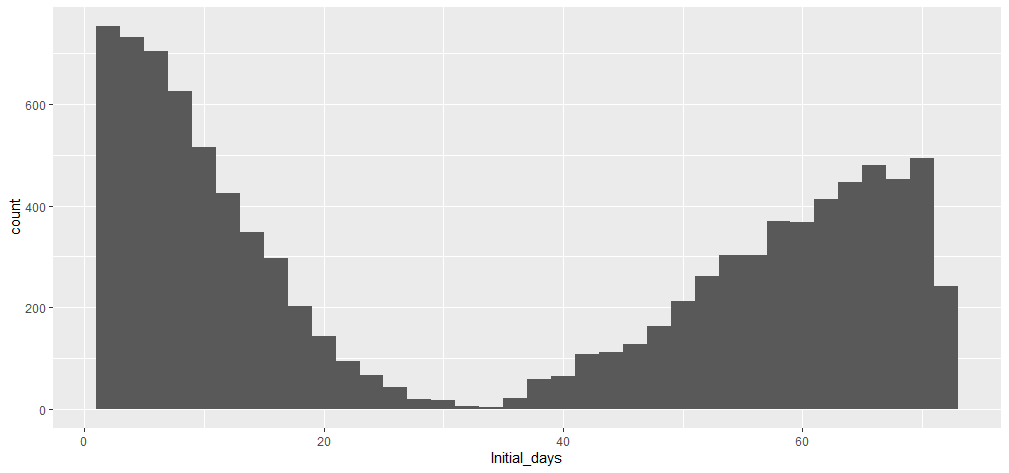
ggplot(med\_clean, aes(Initial\_admin)) +

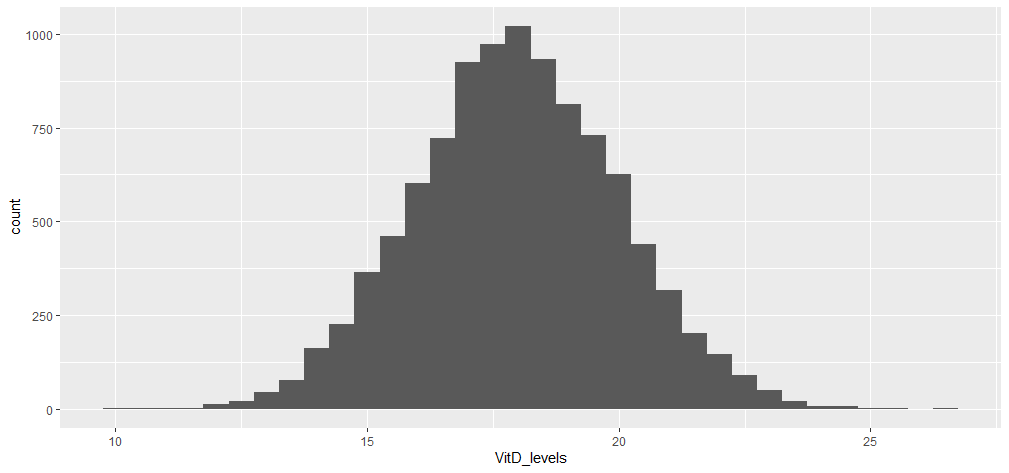
geom\_bar()

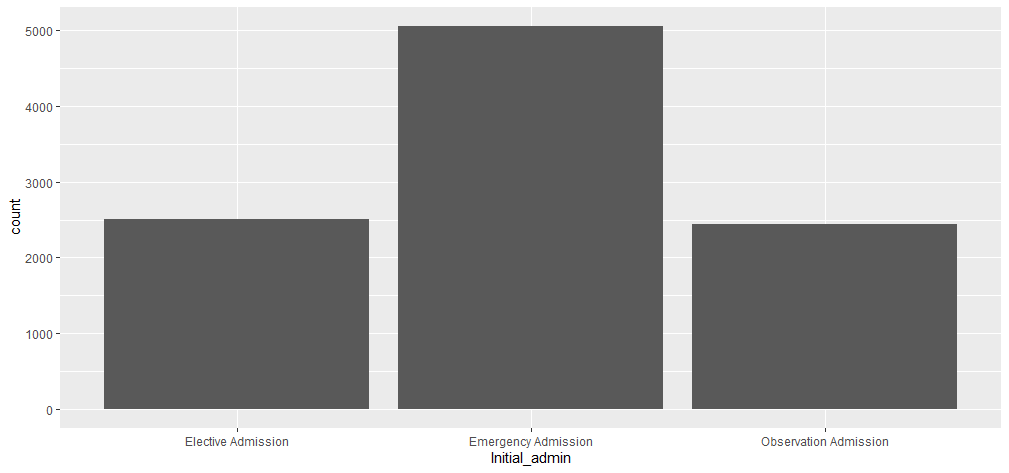
ggplot(med\_clean, aes(Marital)) +

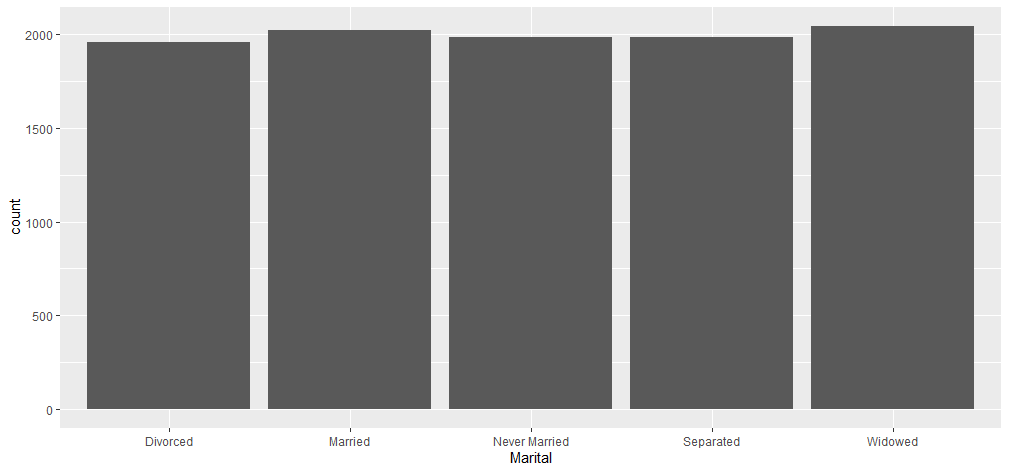
geom\_bar()

The following visual representations, in order, are Initial\_days and VitD\_levels histograms, and Initial\_admin and Marital bar charts.









# Identify the distribution of **two** continuous variables and **two** categorical variables using bivariate statistics from your cleaned and prepared data.

The two relationships that I selected for exploring continuous variables were the relationship between Age and Income, and the relationship between Initial\_days and TotalCharge. I was curious to see if there was any relationship within our sample between a patient’s age and their reported income. I also wanted to explore this same curiosity with a patient’s initial length of stay and the amount that they are charged for such stay.

It can be seen in the upcoming visuals that within our sample there is no relationship between age and income, and the correlation of these two variables was calculated to be -0.01222814. However, there was a strong, positive correlation between Initial\_days and TotalCharge. The correlation between these two variables was calculated to be 0.9876403, which is nearly a 1:1 relationship.

The categorical variables that I am interested in visualizing the relationship between are similar in nature to what I previously performed my Chi-squared test on. First, I wanted to visualize the relationship between Complication\_risk and HighBlood (pressure). High blood pressure is known to be an important cardiac risk factor and could therefore affect a patient’s risk of complications. The second pair of categorical variables that I wanted to explore were BackPain and Overweight. This is not exactly pertinent to the original question of Complication\_risk and ReAdmission, however with my background in physical rehabilitation, it is a relationship I would like to explore.

## Represent your findings in part D, visually as part of your submission

The code that I utilized to produce the scatter plots of my continuous variables and perform the correlation calculations were as follows:

***Bivariate EDA***

*Continuous Variables*

ggplot(med\_clean, aes(x=Income, y=Age)) +

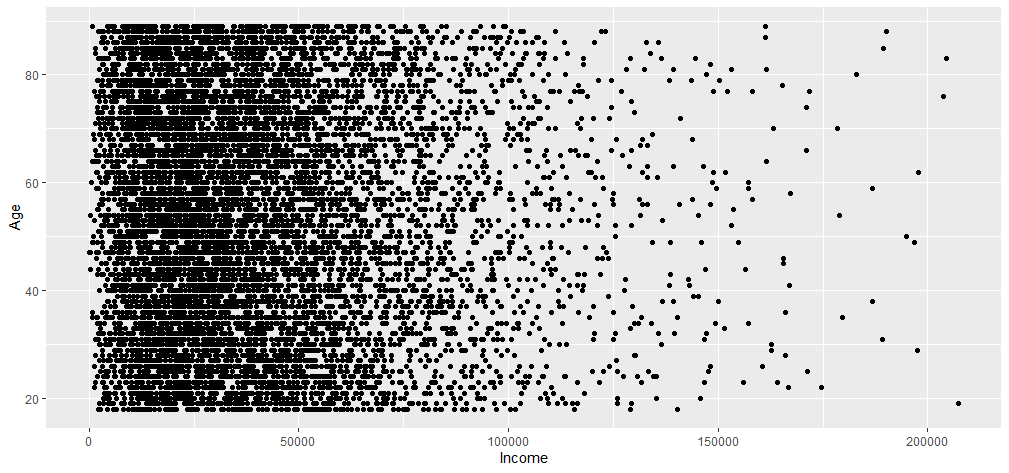
geom\_point()

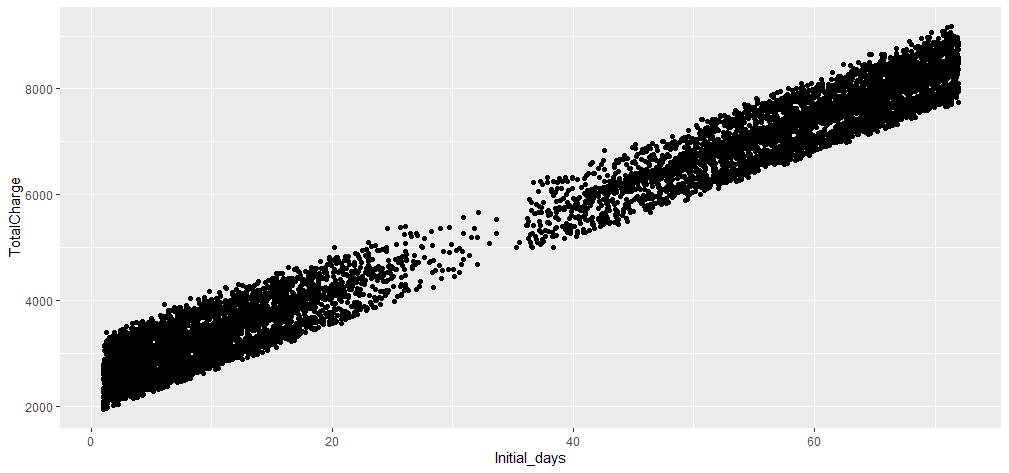
cor(x=med\_clean$Income, y=med\_clean$Income)

ggplot(med\_clean, aes(x=Initial\_days, y=TotalCharge)) +

geom\_point()

cor(x=med\_clean$Initial\_days, y=med\_clean$TotalCharge)





The code that I utilized to create the bar graphs for the categorical variables is as follows. I created graphs with and without ‘fill’ to demonstrate the differences/similarities in the values of each group, as well as the proportions.

*Categorical Variables*

ggplot(med\_clean, aes(Complication\_risk, fill=HighBlood)) +

geom\_bar(position = ‘fill’)

ggplot(med\_clean, aes(Complication\_risk, fill=HighBlood)) +

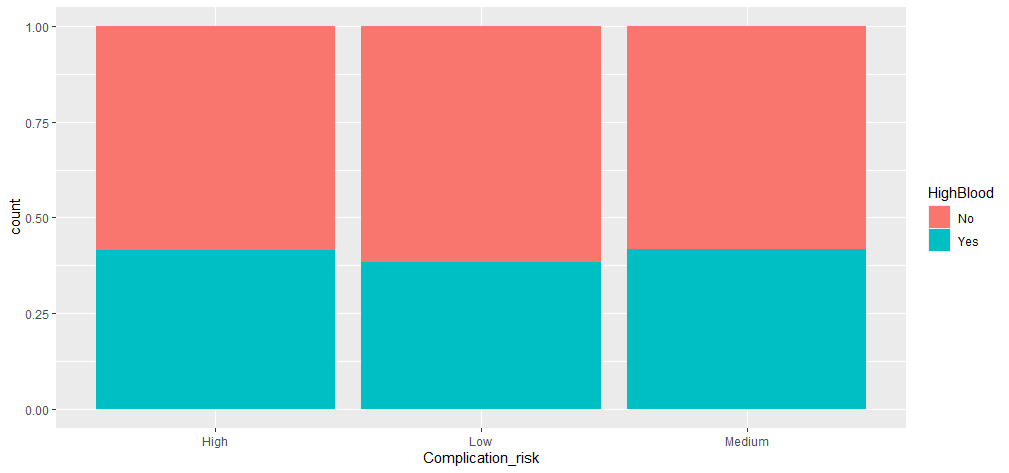
geom\_bar()

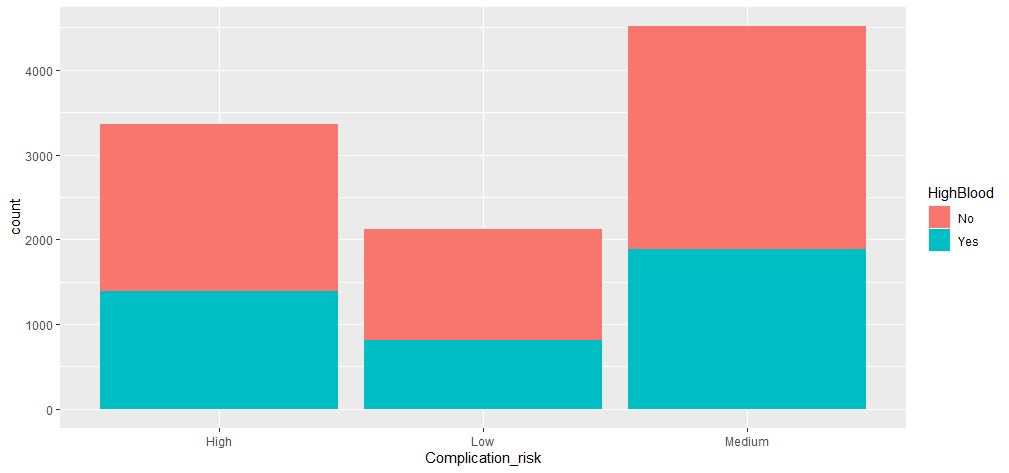
ggplot(med\_clean, aes(BackPain, fill = Overweight)) +

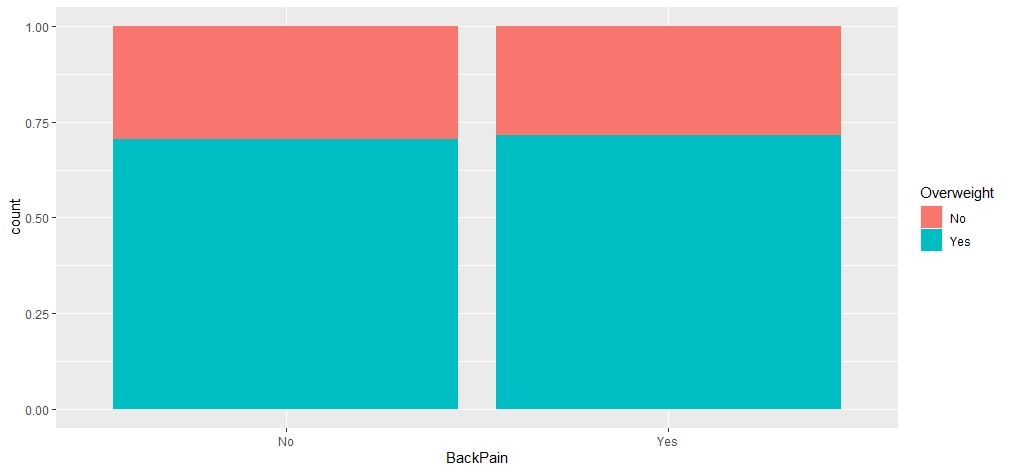
geom\_bar(position = ‘fill’)

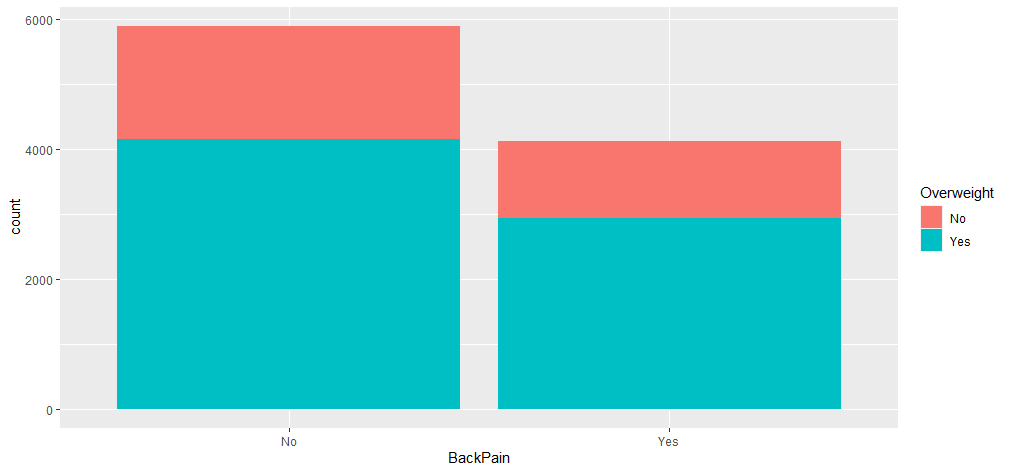
ggplot(med\_clean, aes(BackPain, fill = Overweight)) +

geom\_bar()









# Summarize the implications of your data analysis by doing the following:

## Discuss the results of the hypothesis test.

As can be seen above in the output of my Chi-squared test in B1 and B2, the p-value of my test was greater than 0.9.

For a Chi-square test such as this, my null hypothesis is that the two variables are independent, or that one does not affect the other. This can be stated in another way, such that a patient’s complication risk is independent of his or her readmission status.

The alternate hypothesis would stat that the variables complication risk and readmission status are not independent from one another.

The standard alpha is set to 0.05, therefore, to reject the null hypothesis, we would require a p-value less than or equal to 0.05.

In this case, we would not reject the null hypothesis, as our data does not provide compelling evidence that complication risk and readmission status affect one another.

## Discuss the limitations of your data analysis.

One limitation of this analysis is that if we were to have seen a relationship between Complication\_risk and ReAdmis, we would not be able to determine the strength of the relationship from the Chi-squared test alone. Furthermore, without further testing we could not determine which levels of Complication\_risk would be significantly related to ReAdmission.

Another limitation from my analysis is that I began with a very narrow scope. My initial question only addressed two variables, while our dataset contains a total of approximately 20 variables that may affect a patient’s readmission status or even his or her complication risk.

## Recommend a course of action based on your results.

I believe that the next best course of action would be to pursue exploring other variables that were noted to have increased weighting with the PCA in D206. HighBlood (pressure) was one such variable. I think it would save some time to focus on these variables first, rather than blindly exploring the relationship between ReAdmis and all potential factors, for example, ReAdmis and Overweight, ReAdmis and BackPain, ReAdmis and Arhtritis, etc. Beyond that, addressing variables that are known to have increased risk of readmission, such as those that may put a patient at greater risk for infection, would be beneficial to explore. We would still like to keep healthcare costs low for the patients and the company, and determining what factors patients present with that may be related to an increased rate of readmission is still our goal.

# Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the tool(s) used.

Please see the attached Panopto video.

# Reference the web sources used to acquire segments of third-party code to support the analysis.

No web sources were utilized to assist with compiling the code within this performance assessment. DataCamp videos were completed over the course of the previous 4 weeks, which provided the foundation for completion of this task.

# Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

No sources were utilized as a reference, quoted, paraphrased, or summarized within this performance assessment.

# Demonstrate professional communication in the content and presentation of your submission.