A1: Research Question

The question that I will be answering is if a patient’s complication risk influences whether they are re-admitted into the hospital within one month of their last visit.

A2: Benefit of Analysis

A complication in medicine is an unfavorable result, whether it be from disease, surgery, or treatment.[3] Stakeholders would benefit from this analysis by knowing, with a measure of confidence, if patients being given a low, medium, or high complication risk will affect if they are readmitted. This knowledge could allow them to make changes that could benefit both the patients and doctors.

A3: Data Identification

The most relevant data to answering our question are ‘ReAdmis’, a categorical variable of whether the patient has been readmitted to the hospital within a month with two values (yes, no) and ‘Complication\_risk’, another categorical variable of each patient’s complication risk, as assessed by the physician, has three values (low, medium, high).

Since my two main variables are both categorical, I will be using the chi-square test.

B1: Code

*#Import all packages*

*import pandas as pd*

*import numpy as np*

*import seaborn as sns*

*import matplotlib.pyplot as plt*

*%matplotlib inline*

*from scipy import stats*

*from scipy.stats import chi2\_contingency*

*#Load my data set into pandas*

*df = pd.read\_csv('medical\_clean.csv')*

*#Create a second one with only the two columns to look at the descriptions*

*df2 = df[['ReAdmis', 'Complication\_risk']]*

*df2.describe()*

*#Create a contingency table for the ReAdmis and Complication\_risk variables*

*contingency = pd.crosstab(df['ReAdmis'], df['Complication\_risk'])*

*contingency*

*#Second contingency table to see the percentages of each value compared to the total*

*contingency\_pct = pd.crosstab(df['ReAdmis'], df['Complication\_risk'], normalize = 'index')*

*contingency\_pct*

*#Run Chi-Square test for the p-value then test it against the alpha value of 0.05*

*c, p, dof, expected = chi2\_contingency(contingency)*

*alpha = 0.05*

*print("p-value is: " + str(p))*

*if p <= alpha:*

*print('Dependent: Reject H0.')*

*else:*

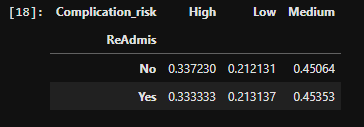
*print('Independent: H0 holds true.')*

B2: Output

*c, p, dof, expected = chi2\_contingency(contingency)*

p-value: 0.923567890607327

A screenshot of a computer screen

Description automatically generated with medium confidence

B3: Justification

For this analysis, the two variables I used were ‘ReAdmis’ and ‘Complication\_risk’. Both variables are categorical, ‘ReAdmis’ is either a yes or no question and ‘Complication\_risk’ has three possible values, which are high, medium or low. Given that both variables are categorical I used chi-square testing for analysis.

C: Univariate Statistics

Continuous variables:

1. Age
2. TotalCharge

Categorical variables:

1. Item1 – Timely\_Admission
2. Item2 – Timely\_Treatment

*#Rename the 'item' columns I'm using for better visualization*

*df.rename(columns = {'Item1': 'Timely\_Admission',*

*'Item2': 'Timely\_Treatment',*

*'Item6': 'Hours\_of\_Treatment',*

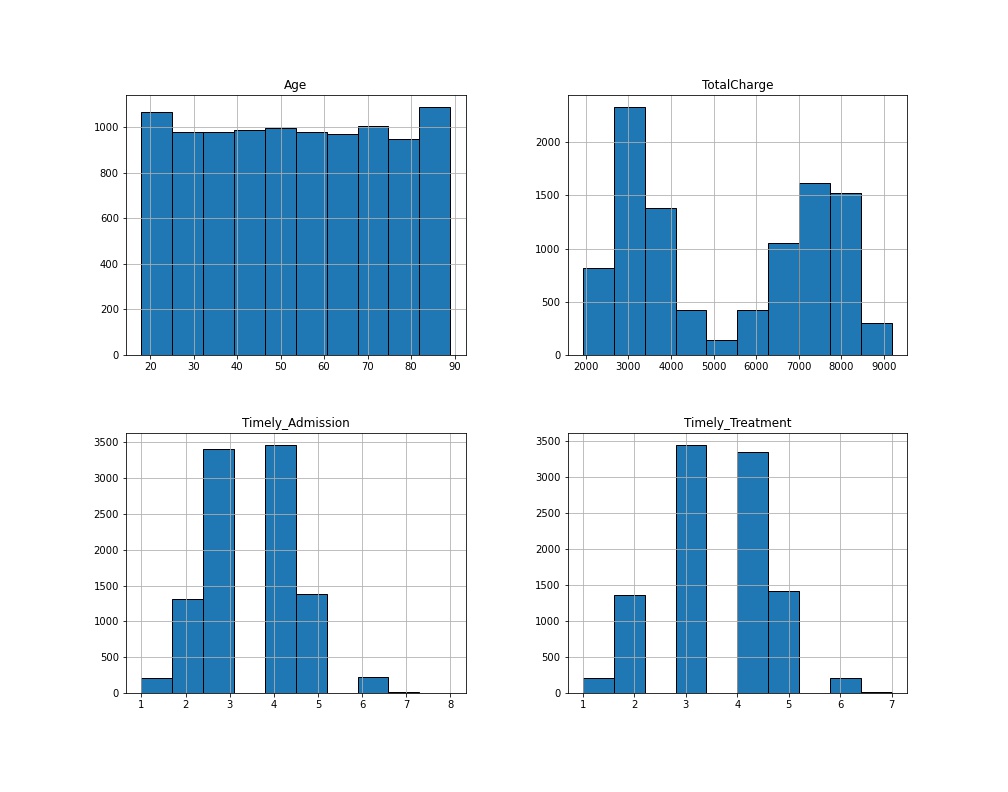
*'Item8': 'Evidence\_of\_Listening'},*

*inplace = True)*

*#Create histograms for the continuous and categorical variables*

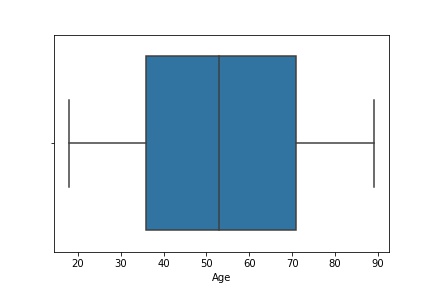
*df[['Age', 'TotalCharge', 'Timely\_Admission', 'Timely\_Treatment']].hist(ec = "black", figsize = (14, 11))*

*#plt.savefig('Univariate Histogram.jpg')*



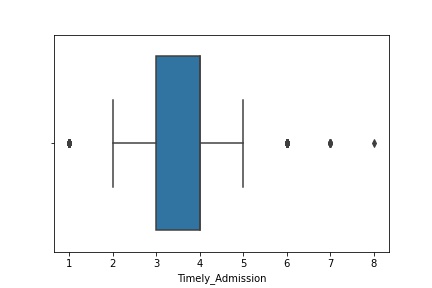
*#Create boxplots for the continuous and categorical variables [2]*

*sns.boxplot(x = df['Age'])*

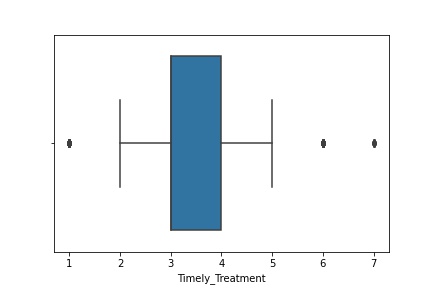
*sns.boxplot(x = df['TotalCharge'])* Chart, box and whisker chart

Description automatically generated

*sns.boxplot(x = df['Timely\_Admission'])*

**

*sns.boxplot(x = df['Timely\_Treatment'])*

**

D: Bivariate Statistics

Continuous variables:

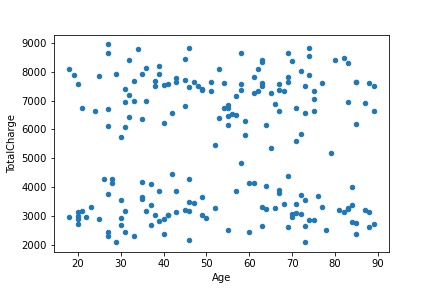
1. Age
2. TotalCharge

Categorical variables:

1. Item6 – Hours\_of\_Treatment
2. Item8 – Evidence\_of\_Listening

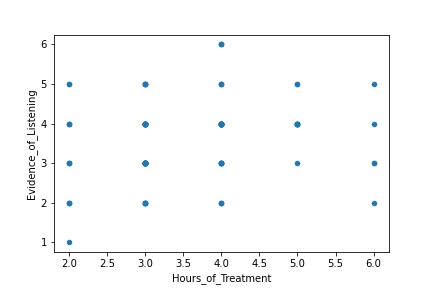
*#Create a scatter plot for the continuous variables [1]*

*df[df['Age'] < 100].sample(200).plot.scatter(x = 'Age', y = 'TotalCharge')*



*#Create a scatter plot for the categorical variables*

*df[df['Hours\_of\_Treatment'] < 10].sample(100).plot.scatter(x = 'Hours\_of\_Treatment', y = 'Evidence\_of\_Listening')*



E1: Result of Analysis

For this test, our hypothesis was that the patient’s complication risk did influence whether they were readmitted, the null hypothesis being that the complication risk did not influence readmittance.

Looking at the two contingency tables above, we see that the percentage of patients who were readmitted with low, medium, and high complication risks are nearly identical to the percentage of patients who were not readmitted with low, medium, and high complication risks. This, coupled with our p-value of 0.9235 from our chi-square test leads us to be able to say that the complication risk has no influence on whether a patient is readmitted.

E2: Limits of Analysis

With our p-value so high we may need to investigate further, to be able to determine with certainty that complication risk has no impact on readmittance; this is too important to be uncertain. For the analysis I used the entire sample size of 10,000 entries for the two variables, so it may be prudent to use a bigger sample size.

E3: Course of Action

Based upon the chi-square test and high p-value I believe the best course of action would be further analysis.

F: Video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=86ef6ed3-44a4-4c32-acc5-ae2c00240ba1

G: Sources for Third-Party Code

[1] “Bivariate Plotting with Pandas.” *Kaggle*, Kaggle, 19 Sept. 2018, https://www.kaggle.com/residentmario/bivariate-plotting-with-pandas

[2] “Seaborn.boxplot.” *Seaborn.boxplot - Seaborn 0.11.2 Documentation*, https://seaborn.pydata.org/generated/seaborn.boxplot.html

H: Sources

[3] “Complication (Medicine).” *Wikipedia*, Wikimedia Foundation, 16 Jan. 2022, <https://en.wikipedia.org/wiki/Complication_(medicine)>

[4] Frost, Jim, et al. “Can High P-Values Be Meaningful?” *Statistics By Jim*, 26 Feb. 2019, https://statisticsbyjim.com/hypothesis-testing/high-p-values/