Exploratory Data Analysis

D207 WGU

Tresa (Tessie) Austin

A.  Research Question Information

1. After looking generically at what conditions influence hospital readmission, taking that a step further now and addressing the question “Do patients with high blood pressure have higher readmission rates than patients without high blood pressure?”

The goal here is to evaluate the below hypothesis:

𝐻0​:*HBPReadmission=NonHBPReadmission* againstthe alternative hypothesis

𝐻1: *HBPReadmission ≠ NonHBPReadmission* with an alpha value of 0.05 or 95% certainty.

1. According to data, CMS penalizes hospitals for readmission rates. Our stakeholders encompass hospital executives, staff, patients, and the wider community. By investigating whether high blood pressure, typically treatable according to the American Medical Association (AMA), correlates with increased readmission rates, healthcare professionals can better ensure patients receive and adhere to necessary medications ("Patients Can Take These Steps to Lower Their High Blood Pressure," American Medical Association). This in turns allows hospital resources to address other illnesses that are not as treatable, may lower overall healthcare costs, and contribute to the overall community populations health.
2. From the medical\_clean.csv the following data points are relevant to addressing the research question “Do patient with high blood pressure have higher readmission rates than patients without high blood pressure?”

ReAdmis: Whether the patient was readmitted within a month of release or not (Yes, No)

HighBlood: Whether the patient has high blood pressure (Yes, No)

Based on work done with this dataset previously, there are multiple columns with data types that need to be addressed. Addressing data types that would more readily align with the needed testing done here encompasses changing some items to Booleans or categorical variables with numeric scales. To achieve this, initial coding for conversion of Boolean and category columns is set.

First, columns ReAdmis, Soft\_drink, HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain,Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, and Asthma will be changed to Booleans.

Secondly, columns Marital, Gender, Initial\_admin, Complication\_risk, and Services will be changed to categories and set with a numerical scale based on the number of options for each.

*# Function to convert columns to bool*

def convert\_to\_boolean(df, columns):

for col in columns:

try:

df[col] = df[col].str.startswith('Yes')

except AttributeError:

print(f"{col}'not a string column. Next!")

df[col] = df[col].astype(bool)

return df

*# Function to convert category and assign numeric*

def convert\_to\_category(df, column, categories, levels):

df[column] = df[column].astype("category")

df[column] = df[column].cat.set\_categories(categories).cat.codes

return df

*# Convert columns to bool*

boolean\_columns = [

'ReAdmis', 'Soft\_drink', 'HighBlood', 'Stroke', 'Overweight', 'Arthritis',

'Diabetes', 'Hyperlipidemia','BackPain','Anxiety', 'Allergic\_rhinitis',

'Reflux\_esophagitis', 'Asthma'

]

df = convert\_to\_boolean(df, boolean\_columns)

*# Convert category and assign numeric*

df = convert\_to\_category(df, "Marital", ['Divorced', 'Married', 'Never Married', 'Separated', 'Widowed'], range(5))

df = convert\_to\_category(df, "Gender", ['Female', 'Male', 'Nonbinary'], range(3))

df = convert\_to\_category(df, "Initial\_admin", ['Elective Admission', 'Emergency Admission', 'Observation Admission'], range(3))

df = convert\_to\_category(df, "Complication\_risk", ['High', 'Medium', 'Low'], range(3))

df = convert\_to\_category(df, "Services",['Blood Work', 'CT Scan', 'Intravenous', 'MRI'], range(4))

*# Display dtypes after changes*

df.info()

B.  Data analysis

1. For the chosen hypothesis using readmission and high blood pressure, the chi-square test is the best option. According to McHugh(2013) “the chi-square statistic is robust with respect to the distribution of the data.” McHugh(2013) also goes on to say that the richness of the details chi-square functions present provide the researcher betting understanding of results and a way to derive more information.

The chi-square analysis is a test of independence of our two variables. This helps us determine if there is a strong link between the two variables which can help our stakeholders make better medical and personal decisions.

To see the relationship between the variables, a pivot table created using Python will be done.

*# Create a pivot table*

pivot\_table = pd.pivot\_table(df, index='HighBlood', columns='ReAdmis', aggfunc='size', fill\_value=0)

print(pivot\_table)

After dissecting the pivot table to see how the two variables relate, the Chi-Square statistic is done using the following code and results printed listing the chi-square value and the p-value allowing us to determine information about the hypothesis.

*# Perform chi-square test*

chi\_sq, p\_value, \_, \_ = chi2\_contingency(pivot\_table)

*# Print the results*

print("Chi-square statistic:", chi\_sq)

print("P-value:", p\_value)

print(f"The p-value, or probability under the null hypothesis of obtaining a result as extreme as the one observed, is {p\_value:.3f}.")

1. Output and results analysis performed from above Python code

The results of the pivot table are as follows:

ReAdmis False True

HighBlood

False 3747 2163

True 2584 1506

* 3747 individuals do not have high blood pressure (HighBlood=False) and were not readmitted within 30 days(ReAdmis=False).
* 2163 individuals do not have high blood pressure (HighBlood=False) but were readmitted within 30 days (ReAdmis=True).
* 2584 individuals have high blood pressure (HighBlood=True) and were not readmitted within 30 days(ReAdmis=False).
* (ReAdmis=False).
* 1506 individuals have high blood pressure (HighBlood=True) and were readmitted within 30 days (ReAdmis=True).

The result of the chi-square test are as follows:

Chi-square statistic: 0.04239657973011679

P-value: 0.8368656684578771

The p-value, or probability under the null hypothesis of obtaining a result as extreme as the one observed, is 0.837.

Using this statistical information we can determine that there is no association between “HighBlood” and “ReAdmis” variables. Our stakeholders can use the knowledge that even though high blood pressure is a significant illness that many Americans have, it does not play a role in if a patient is readmitted to the hospital within 30 days. ​

1. Justification of using chi-square analysis

Based on fact that there were two variables being analyzed for this study, the most relevant technique to use is a chi-square analysis. In this case, the variables are categorical and the need to determine if the output variable is dependent or independent of the input variable aligns with the selected option of using a chi-square analysis( Brownlee, 2019). This requires the use of chi2\_contingency imported via scipy.stats. Using this imported function, we can compute the p-value as well as the chi-square statistic. To determine significance an alpha value of 0.05 was used and to determine that there is no significant association between high blood pressure and readmission within 30 days. This allows our stakeholders to use resources wisely at the hospital level to address patients needs. This also allows our community stakeholders to know that even if you have a disease such as high blood pressure it does not always lead to repeated hospital stays. There is more research that would need to be done on why patients are readmitted to assist our stakeholders with decisions.

C.  Identify the distribution of **two** continuous variables and **two** categorical variables using univariate statistics and visually represent.

Two categorical variables:

* Initial\_admin: The means by which the patient was admitted into the hospital initially (emergency admission, elective admission, observation)
* Services: Primary service the patient received while hospitalized (blood work, intravenous, CT scan, MRI) (Note: the patient may have received more services, but only the primary service is reported)

Two continuous variables:

* Initial\_days: The number of days the patient stayed in the hospital during the initial visit
* TotalCharge: The amount charged to the patient daily. This value reflects an average per patient based on the total charge divided by the number of days hospitalized. This amount reflects the typical charges billed to patients, not including specialized treatments.

*# Univariate categorical variable 1*

*# Pie chart*

*plt.plot(1, 2, 1)*

*plt.title("Patient Initial Admission")*

*admin\_percent = df["Initial\_admin"].value\_counts()*

*admin\_labels = [ "Emergency Admission","Elective Admission", "Observation"]*

*plt.pie(admin\_percent, labels=admin\_labels, autopct='%1.1f%%', startangle=90, counterclock=False)C*

*plt.axis('square')*

*plt.show()*

A pie chart with text on it

Description automatically generated

*# Mapping to numeric value*

value\_descriptions = {

0: "Elective Admission",

1: "Emergency Admission",

2: "Observation"

}

*# Percent of each Admit type*

value\_counts = df['Initial\_admin'].value\_counts(normalize=True) \* 100

*# Get description from dictionary, default to "Unknown" if not found*

for value, count in value\_counts.items():

description = value\_descriptions.get(value, "Unknown")

print(f"{description}: {count:.2f}%")

Emergency Admission: 50.60%

Elective Admission: 25.04%

Observation: 24.36%

*# Univariate categorical variable 2*

*# Frequency of each service*

service\_percent = df["Services"].value\_counts()

*# Percentage of each service*

sample\_size = len(df)

service\_percentages = (service\_percent / sample\_size) \* 100

*# Custom labels*

custom\_labels = {0: "Blood Work", 1: "Intravenous", 2: "CT Scan", 3: "MRI"}

*# Bar chart*

plt.figure(figsize=[10, 6])

plt.bar(service\_percentages.index, service\_percentages, color='skyblue')

plt.title("Primary Service Received While Hospitalized")

plt.xlabel("Service")

plt.ylabel("Percentage")

plt.xticks(rotation=45, ticks=service\_percentages.index, labels=[custom\_labels.get(i, '') for i in service\_percentages.index])

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()

*# Mapping to numeric values*

value\_descriptions = {

0: "Blood Work",

1: "Intravenous",

2: "CT Scan",

3: "MRI"

}

*# Percent of each service*

value\_counts = df['Services'].value\_counts(normalize=True) \* 100

*# Get description from dictionary, default to "Unknown" if not found*

for value, count in value\_counts.items():

description = value\_descriptions.get(value, "Unknown")

print(f"{description}: {count:.2f}%")

Blood Work: 52.65%

CT Scan: 31.30%

Intravenous: 12.25%

MRI: 3.80%

A graph of blue rectangular bars

Description automatically generated with medium confidence

*# Violin of Initial Days*

plt.figure(figsize=[8, 6])

sns.violinplot(data=df, x="Initial\_days", color='skyblue')

plt.title("Distribution of Initial Days Patient is in Hospital")

plt.xlabel("Initial Days in Hospital")

plt.grid(axis='x', linestyle='--', alpha=0.7)

plt.show()

*# Measures of Central Tendency*

df.Initial\_days.describe()

count 10000.000000

mean 34.455299

std 26.309341

min 1.001981

25% 7.896215

50% 35.836244

75% 61.161020

max 71.981490

Name: Initial\_days, dtype: float64

A diagram of a patient

Description automatically generated

*# Histogram of Total Charges*

plt.figure(figsize=[8, 6])

plt.hist(df["TotalCharge"], bins=30, color='skyblue', edgecolor='black')

plt.title("Distribution of Total Charges")

plt.xlabel("Total Charges")

plt.ylabel("Number of Patients")

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()

*# Measures of Central Tendency*

df.TotalCharge.describe()

count 10000.000000

mean 5312.172769

std 2180.393838

min 1938.312067

25% 3179.374015

50% 5213.952000

75% 7459.699750

max 9180.728000

Name: TotalCharge, dtype: float64

A graph of a number of charges

Description automatically generated with medium confidence

Both categorical variables, Initial\_admin and Services were in line with the expected percentages. Most patients enter the hospital after an emergency. Our results showed 50.6% of our sample were admitted via this scenario. Also, one of the main tool’s hospitals use for diagnosis is blood work and 52.65% of patients had that service performed.

Both continuous variables were a bit puzzling. Initial days in the hospital and Total Charges, while that variable addresses the average per patient divided by the number of days that patient was hospitalized, are quite skewed on both ends. Of course the less days you are in the hospital the less charges you acquire, however, there appears to be two separate peaks for these items.

D.  Identify the distribution of **two** continuous variables and **two** categorical variables using bivariate statistics and visually represent

Set 1:

Categorical

Complication\_risk: Level of complication risk for the patient as assessed by a primary patient assessment (high, medium, low)

Continuous

Additional\_charges: The average amount charged to the patient for miscellaneous procedures, treatments, medicines, anesthesiology, etc.

Set 2:

Categorical:

Gender: Customer self-identification as male, female, or nonbinary

Continuous:

TotalCharge: The amount charged to the patient daily. This value reflects an average per patient based on the total charge divided by the number of days hospitalized. This amount reflects the typical charges billed to patients, not including specialized treatments.

*# Additional Charges info*

print("Summary statistics of 'Additional\_charges':")

print(df['Additional\_charges'].describe())

*# Complication Risk info*

print("Unique values of 'Complication\_risk' and their counts:")

print(df['Complication\_risk'].value\_counts())

Summary statistics of 'Additional\_charges':

count 10000.000000

mean 12934.528587

std 6542.601544

min 3125.703000

25% 7986.487755

50% 11573.977735

75% 15626.490000

max 30566.070000

Name: Additional\_charges, dtype: float64

Unique values of 'Complication\_risk' and their counts:

Complication\_risk

1 4517

0 3358

2 2125

Name: count, dtype: int64

*# Mapping for Complication Risk categories*

risk\_mapping = {0: 'Low', 1: 'Medium', 2: 'High'}

df['Complication\_risk'] = df['Complication\_risk'].map(risk\_mapping)

*# Create a violin plot of Complication Risk and Additional Charges*

plt.figure(figsize=(10, 6))

sns.violinplot(data=df, x='Complication\_risk', y='Additional\_charges', order=['Low', 'Medium', 'High'], palette='viridis')

plt.xlabel('Complication Risk')

plt.ylabel('Additional Charges')

plt.title('Complication Risk vs Additional Charges')

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()

A diagram of different colored shapes

Description automatically generated

*# Total Charges info*

print("Summary statistics of 'Total Charges':")

print(df['TotalCharge'].describe())

*# Gender info*

print("Unique values of 'Gender' and their counts:")

print(df['Gender'].value\_counts())

Summary statistics of 'Total Charges':

count 10000.000000

mean 5312.172769

std 2180.393838

min 1938.312067

25% 3179.374015

50% 5213.952000

75% 7459.699750

max 9180.728000

Name: TotalCharge, dtype: float64

Unique values of 'Gender' and their counts:

Gender

0 5018

1 4768

2 214

Name: count, dtype: int64

*# Create a box plot for Total Charges and Gender*

plt.figure(figsize=(8, 6))

sns.boxplot(data=df, x='Gender', y='TotalCharge', palette='muted')

plt.xlabel('Gender')

plt.ylabel('Total Charges')

plt.title('Total Charges by Gender')

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.xticks([0, 1, 2], ['Female', 'Male', 'Nonbinary'])

plt.show()

A chart of different colored rectangular shapes

Description automatically generated with medium confidence

E.  Summarize the implications of your data analysis

1.   Our analysis was performed with an alpha of 0.05 or 95% certainty in testing the null hypothesis (𝐻0​) against the alternative hypothesis (𝐻1). The null hypothesis states that there is no difference between the readmission rate of patients with high blood pressure and those without. Where the alternate hypothesis poses that there is a difference between readmission rates of patients with high blood pressure and those without. We are testing that there the proportion of patients readmitted to the hospital with high blood pressure is the same readmission rate at those without.

The following statistical values were determined using our data set.

Chi-square statistic: 0.04239657973011679

P-value: 0.8368656684578771

The p value found when performing our test was 0.836 which is higher than the alpha or significance level of 0.05 then there is insufficient evidence to reject the null hypothesis and it can be inferred that there is no statistically significant difference in the readmission rates of patients with high blood pressure and those without high blood pressure.

2.  One limitation of the data was that there was no information around readmission other than if a patient was readmitted within 30 days. There was also no information other than yes or no around a patient’s blood pressure. This data set only included 10,000 records, since an illness such as high blood pressure can increase based on socioeconomic factors this is a small sample size. Readmission rates were only within the first 30 days of dismissal from the hospital, a longer time range could yield different results as some patients may be more diligent within the first month at aftercare such as staying on medications which significantly decreases blood pressure.

3.  Recommendations would be to look at larger samples, longer readmission ranges, and add in other factors around the research such as if a patient were already on medication for blood pressure before admittance to the hospital or not. There are too many factors that could contribute one way or another to this to provide many recommendations since we found no significance in the data analyzed.

F.  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.

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=fbecf034-7ebb-4d30-a14d-b16e0141011c>

G.  Code Sources

McHugh, M. L. (2013). The chi-square test of independence. Biochemia medica, 24(2), 143-149. <https://doi.org/10.11613/BM.2014.018>

H.  Sources

American Medical Association. (2024) Patients can take these steps to lower their high blood pressure. AMA. Retrieved from <https://www.ama-assn.org/delivering-care/hypertension/patients-can-take-these-steps-lower-their-high-blood-pressure#:~:text=While%20there%20is%20no%20cure,as%20prescribed%20by%20their%20physicians>.

Brownlee, J. (2019). Chi-Squared Test for Machine Learning. *Machine Learning Mastery*. Retrieved from <https://machinelearningmastery.com/chi-squared-test-for-machine-learning/>