Western Governors University

D207 – Exploratory Data Analysis - PA

By: Krista Moik

Table of Contents

A1: Question for Analysis

Using the provided medical_clean CSV, my research question is: **Do patients with a high complication risk factor experience higher rates of readmission than patients with a low complication risk factor?** I will be using an alpha value of 0.05 for this analysis.

A2: Benefit from Analysis

As readmission within a one-month period can and does result in penalties against the provider, it is beneficial that stakeholders are aware of any variable that has a relationship with readmission rates. By knowing which variables have a relationship with readmission rates and the effect of that variable on those rates, it would be possible for stakeholders to not only be aware of patients at greater risk for readmission to plan for possible fees, but they may also be able to learn from the data to implement strategies or procedures that could reduce or eliminate this risk altogether.

A3: Data Identification

One of the variables I will look at to answer my research question is ReAdmis. ReAdmis is represented by either yes or no to indicate whether the patient was readmitted within one month of release. This variable is qualitative. An example of this variable from row 2 of the CSV is: No.

The other variable I will look at to answer my research question is Complication_risk. This variable is represented as either high, medium, or low to indicate whether the risk level of the patient for complications per the primary assessment. This variable is qualitative. An example of this variable from row 2 of the CSV is: Medium.

B1: Code

I used the Chi-square technique to test my hypothesis as this method is well suited to categorical and nominal data and I am using categorical variables. Using the code example from the provided WGU course materials, I imported chi2_contingency from the scipy.stats library and ran the following functions (Western Governors University, n.d.):

Import packages:

```
import pandas as pd
import numpy as np
from scipy.stats import chi2_contingency
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Create Contingency Table for Chi-Square:

```
table=pd.crosstab(df.ReAdmis, df.Complication_risk, margins=True)
print(table)
```

To obtain the p-value:

```
df.head()
contingency=pd.crosstab(df['ReAdmis'], df['Complication_risk'])
contingency
contingency_pct=pd.crosstab(df['ReAdmis'], df['Complication_risk'])
contingency_pct
plt.figure(figsize=(12,8))
sns.heatmap(contingency, annot=True, cmap="PuBuGn")
```

To complete the Chi-Square test of independence:

```
c, p, dof, expected=chi2_contingency(contingency)
print(p)
```

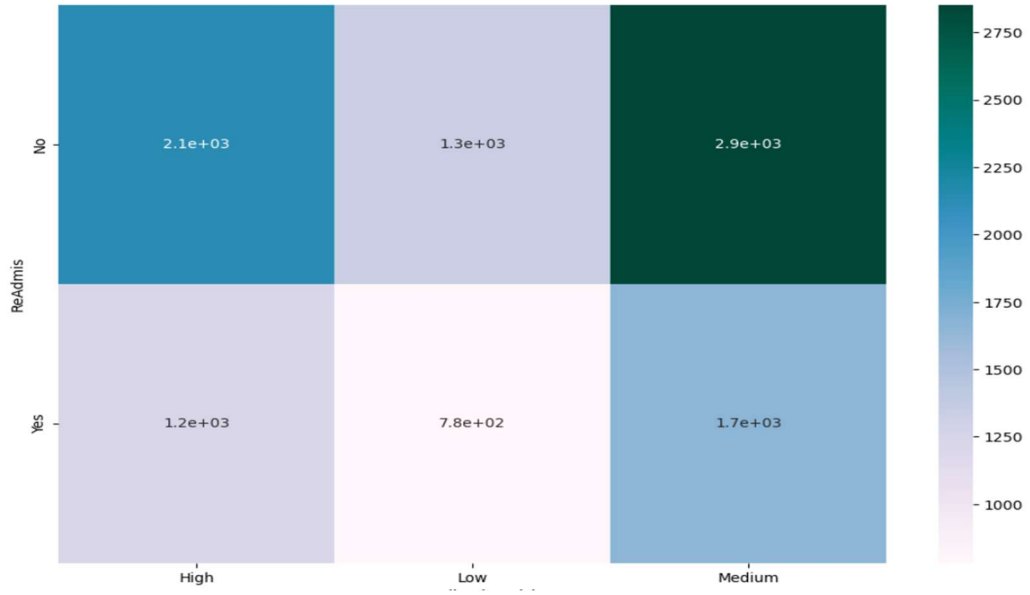
B2: Output

The contingency table counted all of the Complication_risk variables against whether or not those patients were readmitted. The table shows most patients in all 3 ratings of low, medium, and high complication risk were NOT readmitted within one month. This appears to show that there is no relationship between complication risk and readmission. The table is below:

```
#Create Contingency Table for Chi-Square
table=pd.crosstab(df.ReAdmis, df.Complication_risk, margins=True)
print(table)
```

Complication_risk	High	Low	Medium	All
ReAdmis				
No	2135	1343	2853	6331
Yes	1223	782	1664	3669
All	3358	2125	4517	10000

I then used a heatmap to get a different view of the data:



The heat map provides another view of what the contingency table is showing – a larger number of patients at medium risk were NOT readmitted within one month. A slightly smaller group of high complication risk patients were also NOT readmitted within one month. This does not appear to support a hypothesis that those considered at high complication risk were more likely to be readmitted within one month.

Once the contingency table and heat map were created, I used Python to perform statistics to determine the P-value. The P-value was determined to be 0.923567890607327. This value is far greater than the established alpha value of .05. This supports the above visualizations of the variables and means we must accept the null hypothesis that there is NO relationship between Complication_risk and ReAdmis.

```
#Chi-Square test of independence
c, p, dof, expected=chi2_contingency(contingency)
```

```
#Print P-value
print(p)
```

```
0.923567890607327
```

B3: Justification

Since both ReAdmis and Complication_risk are qualitative variables, I chose to use Chi-square as it is useful in comparing qualitative variables and "can provide information not only on the significance of any observed differences, but also provides detailed information on exactly which categories account for any differences found" (McHugh M. L., 2013). In addition to using pandas and numpy, I also used matplotlib.pyplot as plt and seaborn as sns as I knew I would be using their visualization tools. I used chi2_contingency from scipy.stats once I knew I would be using the Chi-square method as it has the statistical capabilities needed (Western Governors University, n.d.).

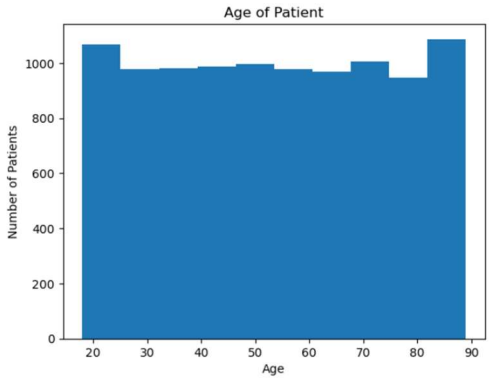
C: Univariate Statistics

I chose to view columns Age and Initial_days for my continuous variables and columns Marital and Initial_admin for my categorical variables.

C1: Visual of Findings

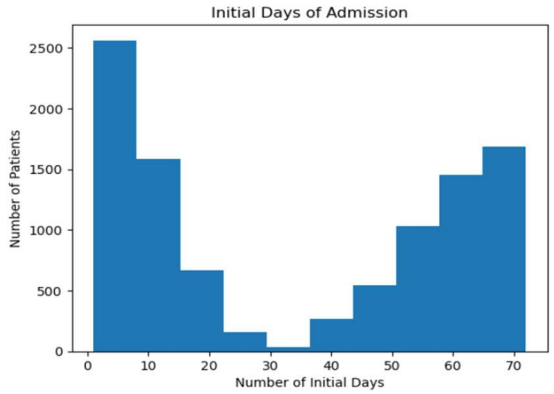
Here is the histogram for the column Age. The histogram shows what appears to be a uniform distribution.

```
#Histogram of Age column - continuous
plt.hist(df['Age'])
plt.title("Age of Patient")
plt.xlabel("Age")
plt.ylabel("Number of Patients")
plt.show()
```



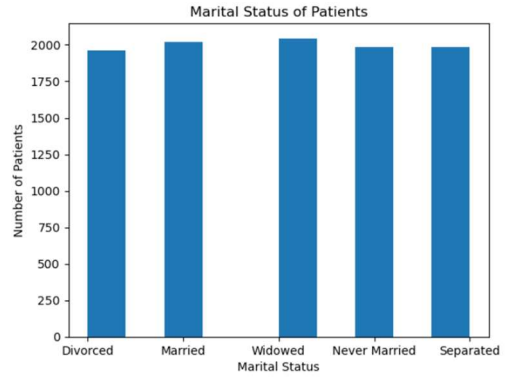
Below is the histogram for the column Initial_days. This histogram shows a non-symmetrical bimodal skew.

```
#Histogram of Initial_days column - continuous
plt.hist(df['Initial_days'])
plt.title("Initial Days of Admission")
plt.xlabel("Number of Initial Days")
plt.ylabel("Number of Patients")
plt.show()
```



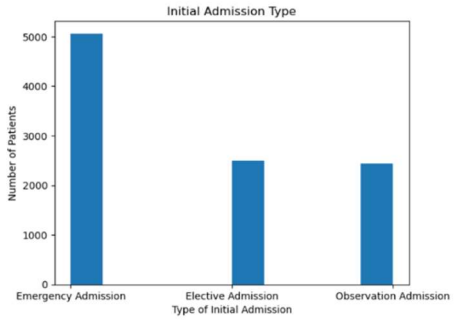
Below is the histogram for the categorical column Marital. It is interesting that the number of patients for each marital status is relatively similar and could be considered uniform distribution.

```
#Histogram of Marital - categorical
plt.hist(df['Marital'])
plt.title("Marital Status of Patients")
plt.xlabel("Marital Status")
plt.ylabel("Number of Patients")
plt.show()
```



Below is the histogram for column Initial_admin. It shows that most patients are admitted as an emergency, but a similar number of patients are admitted as elective and observation admissions.

```
#Histogram of Initial_admin - categorical
plt.hist(df['Initial_admin'])
plt.title("Initial Admission Type")
plt.xlabel("Type of Initial Admission")
plt.ylabel("Number of Patients")
plt.show()
```



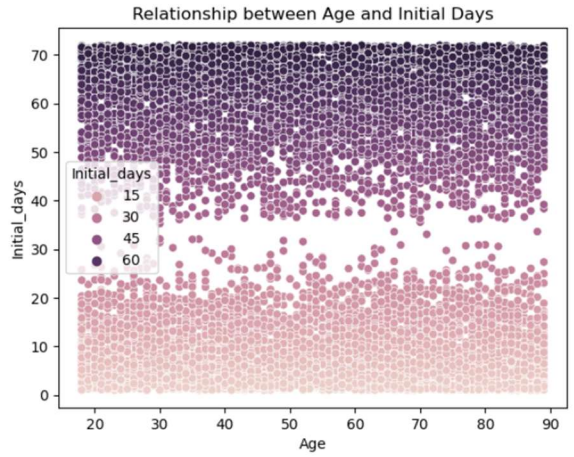
D: Bivariate Statistics

I will use the same variables from section C - columns Age and Initial_days for my continuous variables, and columns Marital and Initial_admin for my categorical variables.

D1: Visual of Findings

First, I made a scatterplot comparing Age with Initial_days. It is interesting that the plot shows a high density among all age groups for the first 20 days and then it becomes sparser from days 20 to about 40, before growing denser again up to the maximum of 70 days.

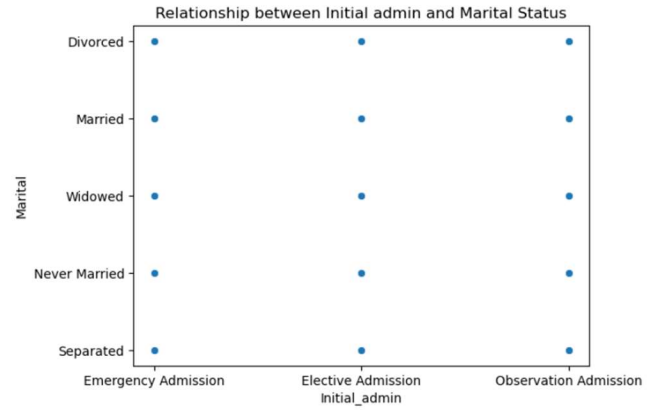
```
#Scatterplot of Age and Initial_days - continuous
sns.scatterplot(x='Age', y='Initial_days', hue='Initial_days', data=df).set(title='Relationship between Age and Initial Days')
plt.show()
```



Below is a scatterplot comparing the relationship between Initial_admin and Marital. This does not provide much insight other than all marital statuses are represented in all the types of initial admission.

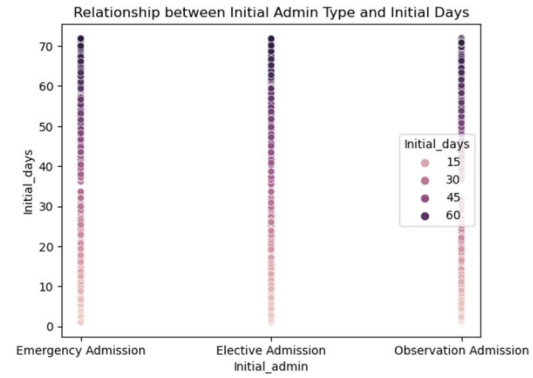
```
#Scatter plot of Initial_admin and Marital - categorical
sns.scatterplot(x='Initial_admin', y='Marital', data=df).set(title='Relationship between Initial admin and Marital Status')
```

```
[Text(0.5, 1.0, 'Relationship between Initial admin and Marital Status')]
```



Below is the scatterplot comparing Initial_admin with Initial_days. It is interesting that the sparsity between 30 and 40 days in the Initial_days columns is present in the emergency and observation admin types.

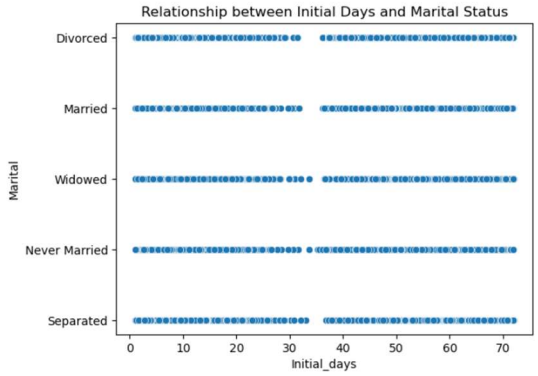
```
#Scatterplot of Initial_admin and Initial_days - continuous
sns.scatterplot(x='Initial_admin', y='Initial_days', hue='Initial_days', data=df).set(title='Relationship between Initial Admin
plt.show()
```



Below is the scatterplot comparing Initial_days with Marital. It is interesting that this plot also shows a similar sparsity between 30 and 40 days across all marital statuses.

```
#Scatter plot of Initial_admin and Marital - categorical
sns.scatterplot(x='Initial_days', y='Marital', data=df).set(title='Relationship between Initial Days and Marital Status')
```

```
[Text(0.5, 1.0, 'Relationship between Initial Days and Marital Status')]
```



E1: Results of Analysis

My findings determined that there is no correlation that more patients labeled as high for Complication_risk were readmitted within one month than the other risk levels. This was supported visually by the contingency table and the heatmap, as well as statistically using the Chi-square method. Since an alpha value of 0.05 was used, my analysis would have needed a p-value that was less than 0.05 to confirm the probability of my hypothesis being true. Unfortunately, the analysis found a p-value of approximately 0.92, indicating that the Complication_risk variable of the patient had no statistically significant relationship with higher rates of readmission per the ReAdmis variable.

E2: Limitations of Analysis

Some of the limitations of my analysis stem from the data itself. It would be helpful to know the reason for the initial admission as well as how and what determines the Complication_risk scoring of low, medium, or high. Statistically speaking, while the Chi-square test is one of the more useful tests for analyzing categorical data, it has its limitations when it comes to sample

size requirements, interpretation difficulties, and that sometimes low correlations are determined even when there are significant results (McHugh M. L., 2013).

E3: Recommended Course of Action

As my analysis did not show any significant correlation, there are no recommendations to be made regarding the Complication_risk, other than it does not seem likely to affect ReAdmis and therefore may be unnecessary to keep collecting data on. However, this lack of correlation could be due to errors in the data. For instance, the hospital may need to re-evaluate how patients are classified as low, medium, or high risk for complications as errors in this process may be affecting the data. Although I do not have experience in the health field, I would recommend that the hospital reevaluate and recalibrate their determination of complication risk levels, and then review the data after to see if there are any changes or relationships found in terms of readmission. Additionally, I found the sparsity seen between 30-40 days in the Initial_days column to be intriguing and would recommend further analysis on what causes that sparsity and if it could be applied somehow to the ReAdmis rates.

F: Video

This link to my Panopto video recording is:
https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c98e2e5b-2cea-403c-8827-b0fd01098948

G: Sources for Third-Party Code

Western Governors University. (n.d.) *D207 Exploratory Data Analysis Webinar Episode 5* [Video]. https://wgu.webex.com/recordingservice/sites/wgu/recording/57b9a2e55ba9103ab7f70050568fa8a9/playback

H: Sources

McHugh M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2), 143–149. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/#:~:text=The%20assumptions%20of%20the%20Chi,the%20variables%20are%20mutually%20exclusive

Western Governors University. (n.d.) *D207 Exploratory Data Analysis Webinar Episode 5* [Video]. https://wgu.webex.com/recordingservice/sites/wgu/recording/57b9a2e55ba9103ab7f70050568fa8a9/playback

I: Professional Communication
Demonstrate professional communication in the content and presentation of your submission.