Marisela Monterroza

D207: Data Exploration

Instructor: David Gagner

March 24, 2023

Table of Contents

Part I: Organizational Situation

A. Question for Analysis:

According to the medical industry,readmissions are to be taken seriously because it shows that there is an underlying issue that involves other variables. For my research question, I would like to know if the initial_admin variable correlates to readmissions.

B. Benefit from an Analysis:

The analysis can contribute to help lower the readmissions rate which will help increase patient satisfaction towards medical providers. If patient satisfaction increases then it will encourage more patients to choose the hospital for care.Medicare's Hospital Readmissions Reduction Program applies penalties to hospitals which causes a hospital to lose money when a patient readmits. A hospital with low readmission rates will in turn produce more revenue for the hospital, increasing the stakeholders return on investment.

C. Data Identification:

I will be using ReAdmis as the dependent variable and as well as a categorical variable.The independent variable will be initial_admin. For the continuous variables, it will be income and age. For the categorical variables it will include Initial_admin and Complication_risk.

Part II: Data Analysis Description

A. Code for analysis

For the data analysis, I first began by checking for any missing values by using the df.info() function. In my data analysis the columns showed the total as non-null which means there are no missing values. The rest of the data was clean since there were no missing values or duplicated values.To search for any duplicated values, I used the df.duplicated() function. I chose the chi-square test to see if two variables are related.I also used python for coding.

B. Code and Results

```
import numpy as np
```

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import pylab
import statsmodels.api as sm
import statistics
from scipy import stats
from scipy.stats import chisquare
from scipy.stats import chi2_contingency

df = pd.read_csv('medical_clean.csv')

df.info()

contingency = pd.crosstab(df['ReAdmis'], df['Initial_admin'])
contingency
```

```
contingency = pd.crosstab(df['ReAdmis'], df['Initial_admin'])
contingency
```

ReAdmis	Elective Admission	Emergency Admission	Observation Admission
No	1608	3156	1567
Yes	896	1904	869

```
contingency_pct = pd.crosstab(df['ReAdmis'], df['Initial_admin'], normalize='index')
contingency_pct
```

ReAdmis	Elective Admission	Emergency Admission	Observation Admission
No	0.253988	0.498499	0.247512
Yes	0.244208	0.518942	0.236849

c, p, dof, expected = chi2_contingency(contingency)
print('p-value = ' + str(p))

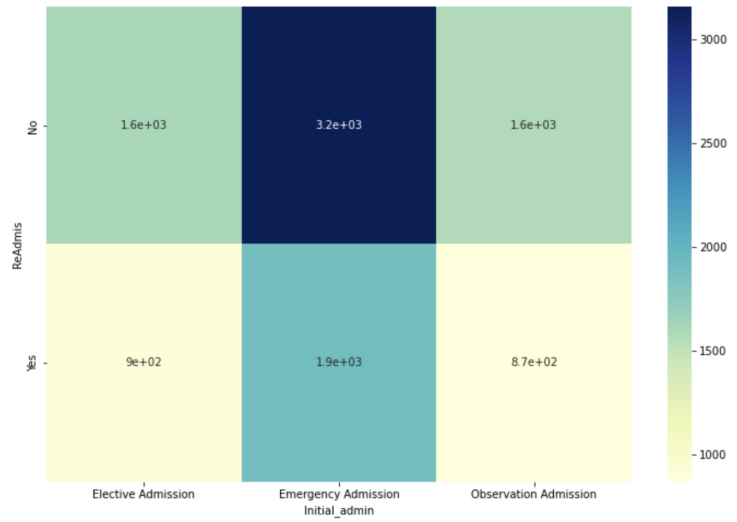
```
c, p, dof, expected = chi2_contingency(contingency)
print('p-value = ' + str(p))
```

```
p-value = 0.14298951184306222
```

plt.figure(figsize=(12,8))
sns.heatmap(contingency, annot=True, cmap="YlGnBu")

```
plt.figure(figsize=(12,8))
sns.heatmap(contingency, annot=True, cmap="YlGnBu")
```

```
<AxesSubplot:xlabel='Initial_admin', ylabel='ReAdmis'>
```



df.describe()

df[['Income','Age']].hist()

plt.savefig('medical_pyplot.jpg')

plt.tight_layout()

groupedInitial_admin=df.groupby(by='Initial_admin').size()
groupedInitial_admin

```
groupedInitial_admin=df.groupby(by='Initial_admin').size()
groupedInitial_admin
```

```
Initial_admin
Elective Admission       2504
Emergency Admission      5060
Observation Admission    2436
dtype: int64
```

%matplotlib inline
groupedInitial_admin.plot.bar()

groupedComplication_risk=df.groupby(by='Complication_risk').size()
groupedComplication_risk

```
groupedComplication_risk=df.groupby(by='Complication_risk').size()
groupedComplication_risk
```

```
Complication_risk
High      3358
Low       2125
Medium    4517
dtype: int64
```

groupedComplication_risk.plot.bar()

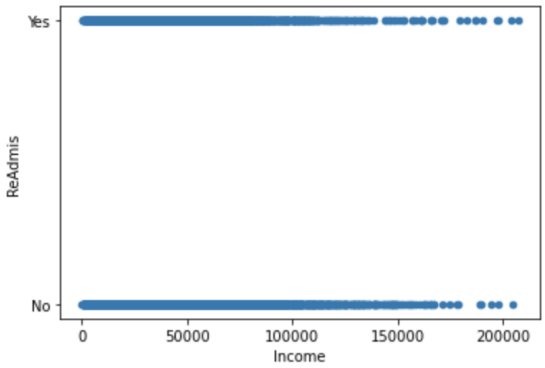
df.boxplot(['Age'])

df.boxplot(['Income'])

df.plot.scatter(x='Income',y='ReAdmis')

```
df.plot.scatter(x='Income',y='ReAdmis')
```

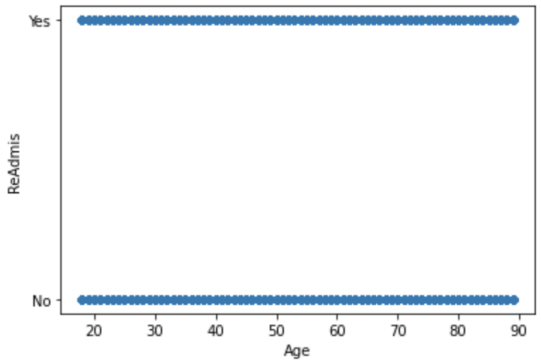
```
<AxesSubplot:xlabel='Income', ylabel='ReAdmis'>
```



df.plot.scatter(x='Age',y='ReAdmis')

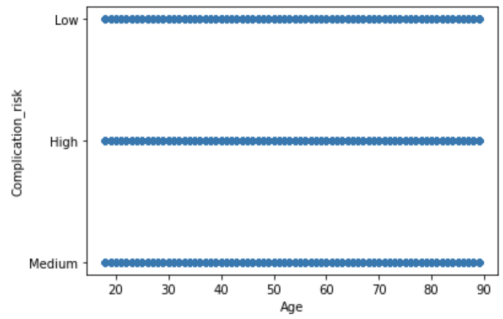
```
df.plot.scatter(x='Age',y='ReAdmis')
```

```
<AxesSubplot:xlabel='Age', ylabel='ReAdmis'>
```



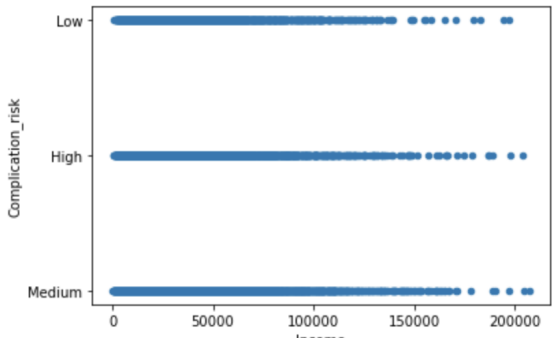
df.plot.scatter(x='Age',y='Complication_risk')

```
: df.plot.scatter(x='Age',y='Complication_risk')
: <AxesSubplot:xlabel='Age', ylabel='Complication_risk'>
```



df.plot.scatter(x='Income',y='Complication_risk')

```
<AxesSubplot:xlabel='Income', ylabel='Complication_risk'>
```



sns.catplot(x="Doc_visits", y="Age",data=df)

sns.catplot(x="Doc_visits", y="Income",data=df)

pd.crosstab(df['ReAdmis'], df['Initial_admin']).plot(kind='bar', stacked=True)

pd.crosstab(df['ReAdmis'], df['Complication_risk']).plot(kind='bar', stacked=True)

C. **Justification**

The ReAdmis variable shows whether the patient was readmitted within a month of release or not. ReAdmis is a dependent categorical variable, therefore I chose to use the Chi-square testing. Chi-square testing is a test for independence between categorical variables

(GeeksforGeeks, 2020).The other variable I will be looking into will be Initial Admin which is also a categorical variable.The analysis shows if the two variables are dependent of each other.

Part III Distribution of Variables-Univariate

A:**Univariate Statistics**

For the univariate statistics, I will be using visual representations such as histograms and boxplots. For a univariate data visualization, the histogram displays the distribution of the data.Boxplots are also very useful to see the distribution of the data as well as to detect the outliers.Univariate visualization is when data is visualized in one-dimension(G, 2022).

Distribution of two continuous variables:

1.Age
2.Income

Distribution of two Categorical variables:

1.Initial Admin
2.Complication risk

The histogram for the income variable shows a skewed distribution.The histogram for the age variable shows a uniform distribution. The bar graph for the Initial Admin variable shows that most were initially admitted as an emergency. For the bar graph of the complication risk variable, it shows that most had a medium complication risk.

For the boxplots, the income shows many outliers and the age boxplot has no outliers.

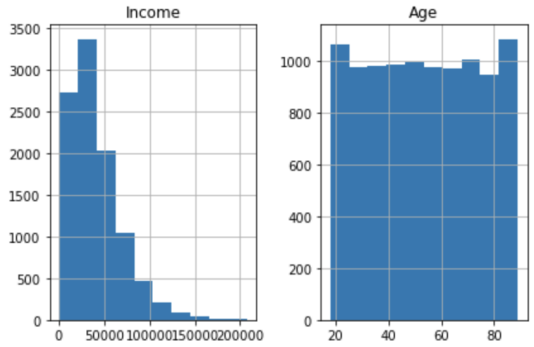
B: **Visual Findings**

```
df.describe()
```

	CaseOrder	Zip	Lat	Lng	Population	Children	Age	Income	VitD_levels	Doc_visits	...
count	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	...
mean	5000.50000	50159.323900	38.751099	-91.243080	9965.253800	2.097200	53.511700	40490.495160	17.964262	5.012200	...
std	2886.89568	27469.588208	5.403085	15.205998	14824.758614	2.163659	20.638538	28521.153293	2.017231	1.045734	...
min	1.00000	610.000000	17.967190	-174.209700	0.000000	0.000000	18.000000	154.080000	9.806483	1.000000	...
25%	2500.75000	27592.000000	35.255120	-97.352982	694.750000	0.000000	36.000000	19598.775000	16.626439	4.000000	...
50%	5000.50000	50207.000000	39.419355	-88.397230	2769.000000	1.000000	53.000000	33768.420000	17.951122	5.000000	...
75%	7500.25000	72411.750000	42.044175	-80.438050	13945.000000	3.000000	71.000000	54296.402500	19.347963	6.000000	...
max	10000.00000	99929.000000	70.560990	-65.290170	122814.000000	10.000000	89.000000	207249.100000	26.394449	9.000000	...

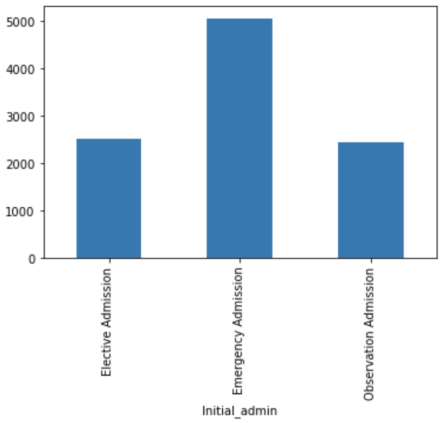
8 rows × 24 columns

```
df[['Income','Age']].hist()
plt.savefig('medical_pyplot.jpg')
plt.tight_layout()
```



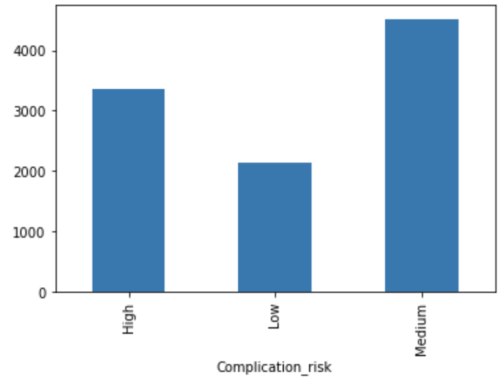

```
groupedInitial_admin.plot.bar()
```

```
<AxesSubplot:xlabel='Initial_admin'>
```



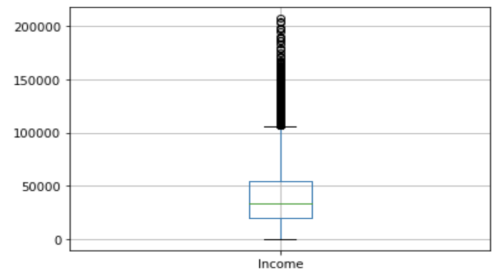
```
groupedComplication_risk.plot.bar()
```

```
<AxesSubplot:xlabel='Complication_risk'>
```



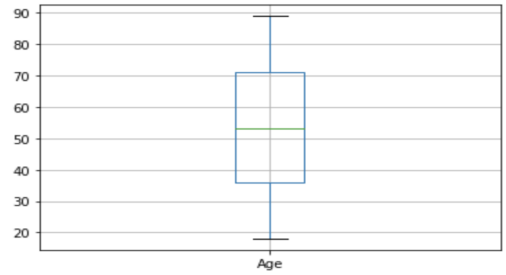

```
df.boxplot(['Income'])
```

```
<AxesSubplot:>
```




```
df.boxplot(['Age'])
```

```
<AxesSubplot:>
```



Part IV Distribution of variables-Bivariate

A. **Bivariate Statistics**

I will be using visual representations for the bivariate statistics.I will be using scatter plots, stacked bar graphs and boxplots.. Bivariate statistics analyzes two variables to understand the relationship between two variables(Z, 2021).

Comparinging Continuous variable with categorical variable

Continuous variables
1. Age and Doc visits
2. Income and Doc visits

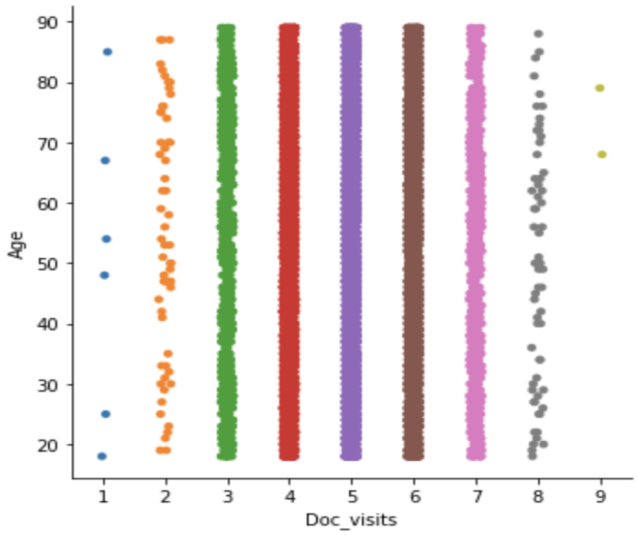
Categorical variables
1. Initial_admin and ReAdmis
2. Complication_risk and ReAdmis

The scatterplots do not show any correlation other than that most patients received about 4-7 doctor visits. I also included a boxplot which showed the same. The stacked bar chart is to show if there is a correlation between two variables with two or more categories. According to the bar chart, fewer patients were readmitted. Out of all the patients who were readmitted, the category in the yes ReAdmis group was the emergency admission. But we cannot say emergency admission is the cause for readmissions. Because the number of emergency admissions is still greater in the No ReAdmis group. We can say that those who have been admitted due to emergency reasons may be more likely to be readmitted, therefore knowing this can help lower the readmission rate by organizing a plan with the emergency department. The stacked bar graph that includes complication risk and remission shows a similar case except the median risk category seems to be more prone to being readmitted but there isn't a vast difference between the categories unlike the previous stacked bar graph.

B.**Visual Findings**

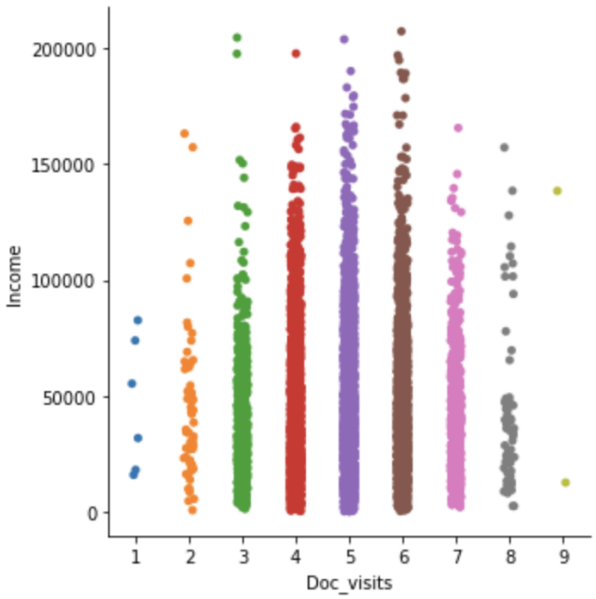
```
sns.catplot(x="Doc_visits", y="Age
```

```
<seaborn.axisgrid.FacetGrid at 0x7fda8e43b790>
```



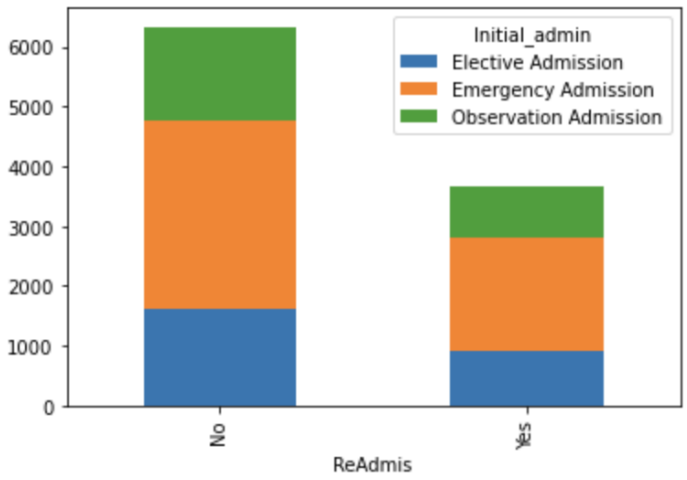

```
sns.catplot(x="Doc_visits", y="Income",data=df)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fdab9e79df0>
```



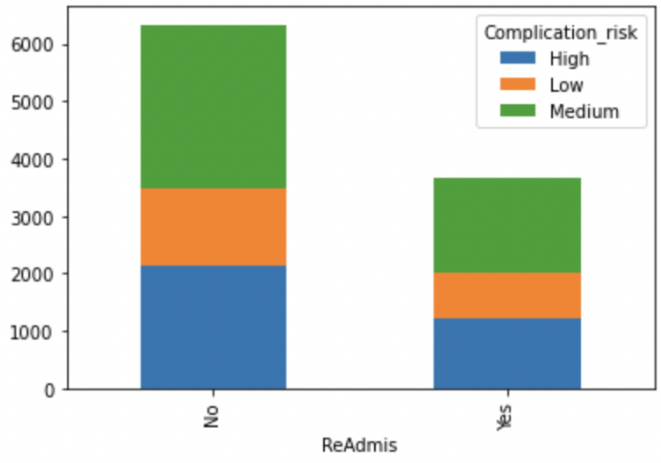
```
pd.crosstab(df['ReAdmis'], df['Initial_admin']).plot(kind='bar', stacked=True)
```

```
<AxesSubplot:xlabel='ReAdmis'>
```




```
]: pd.crosstab(df['ReAdmis'], df['Complication_risk']).plot(kind='bar', stacked=True)
```

```
]: <AxesSubplot:xlabel='ReAdmis'>
```



Part V Implications

A. Results of Analysis

For the chi-square significance test, I have two hypothesis which is ReAdmis and Initial_admin. I received a p-value of 0.14298951184306222. A p-value with a significance level of alpha that is less than 0.05 is considered to be statistically significant.When a value is significant then the null hypothesis is rejected. We cannot reject the null hypothesis because the p-value is greater than 0.05.

B. Limitations of Analysis

A limitation of this analysis is the insufficient information on why each patient was readmitted and how soon the patient was readmitted. If I had access to the information, I would be able to look at other variables to prove the causation and to help find correlations to help lower the readmission rate. The p-value that is higher than .05 shows that data gathered was not enough to prove if the values I looked into would help prevent readmissions in the future. I would need to look into other variables such as medical illnesses.

C. Recommended course of action

The chi-square significance test showed that there was no correlation between readmission and initial admission. Additional analysis is recommended to find a correlation between readmission and medical health issues such as high blood pressure, arthritis, back pain, etc.

Part VI Panopto Video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=004018db-c7d9-407a-9185-afe8016ab807

Part VII Sources for third-party code

G.(2022, December 2). *12 Univariate Data Visualizations With Illustrations in Python.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/07/univariate-analysis-visualization-with-illustrations-in-python/

GeeksforGeeks. (2020, June 23). *Python Pearson s Chi Square Test.* https://www.geeksforgeeks.org/python-pearsons-chi-square-test/

Z.(2021, November 22). *How to Perform Bivariate Analysis in Python (With Examples)*. Stratology. https://www.statology.org/bivariate-analysis-in-python/

PART VIII Sources

Western Governors University. *(n.d.).MSDA, Exploratory Data Analysis - D207.* https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=bccf2cb6-39e2-4a53-8744-ad1900e9aa91