Environmental and Health Factors in Hospital Readmissions

Jon Bexell
Western Governor's University
Student: 000215599
Course D207: Data Analysis
Course Instructor: William Sewell

October 31, 2024

Table of Contents

A. Introduction

Frequent hospital readmissions in the U.S. can often be a sign of ineffective treatment, and as a result the government has enacted certain fines that can correspond to high readmission rates. More specifically, penalties can cost 1% of a hospital's Medicare payments through any given fiscal year. (Jordan Rau, 2023)

This paper will review the Medical dataset to identify variables that correlate to readmissions. From here, we can analyze the data to find levers within the metrics that leadership can use to curtail readmission rates and limit the risk of fines from the government. By prioritizing variables are identified that lead to higher risks of readmissions, readmission rates may be lowered by the added focus.

A1. What correlates to hospital readmissions?

Identifying which categorical variables correlate to higher readmissions will be the first step in finding metrics that actually cause readmissions. From there, we can investigate select values from those trigger variables to see which factors could be contributing to readmissions. With the sources of readmission discovered, we can then begin to create a plan that can mitigate readmissions and protect the hospital and it's stakeholders from unnecessary and burdensome fines.

A2. Stakeholder Benefits

By reducing hospital readmissions, stakeholders can ensure the hospital won't be liable for fees that reduce the overall Medicare payments received. This will improve the bottom line for the hospital and finances. Practitioners at the hospital will likely have improved metrics for their patients as a result of improved care, and patients benefit by receiving needed care at the time of hospitalization, rather than having to return for secondary or tertiary visits.

A3. Relevant Data

Within the medical data, we can focus on health factors (such as diet, blood pressure, stroke, weight, arthritis, diabetes, hyperlipidemia, back pain, anxiety, allergic rhinitis, and acid reflux) that might relate to readmissions. We can also look at environmental factors. This will include geographic area, family composition, age, education, employment, gender and income. The state of initial admission is also important, and this categorical variable includes emergency, elective or observational as potential factors. Finally, there are eight survey questions we can look into as well.

To simplify our search and create a solid starting point, we will review categorical data first (area, marital status, gender, soft drinks, initial admission, high blood pressure, stroke, complication risk, overweight, arthritis, diabetes, hyperlipidemia, back pain, anxiety, allergic rhinitis, reflex esophagitis, asthma, services, and our eight survey questions).

Categorical data with high cardinality will not be suitable candidates for analysis. This includes columns like State, City and Job. Those variables will be disregarded for the purposes of our analysis.

The data for this analysis has already been processed, but there are still some things we can do to clean the data further to streamline our review. Before we begin analyzing data, we will look for potential duplicates in unique variables. This includes CaseOrder, Customer_id, Interaction and UID. These four variables are all identifiers - unique codes that identify a specific case, customer, interaction, or UID. They should not repeat, so identifying duplicates here is a precautionary measure aimed at validating the data. Thankfully, there are no duplicates in our data.

Null values can skew data, so we will also perform a quick search to look for null values. Nulls can be imputed with data depending on their shape, or removed to avoid any skewing. Luckily, we have no nulls in our data.

We will also check for data integrity by searching for categorical data with single entries. For example, if Gender was broken down into Male (5,000), Female (4,999), and Vemale (1), this would indicate a likely typo that needs correcting. It's a relatively simple process to seek out these errors to validate the data, and it's prudent to perform if we cannot confirm it was done previously. On this dataset, there were no such typos in the categorical data.

One last point of data integrity is identifying binary fields of categorical data where data will be "Yes" or "No", and making sure they are expressed in a single format. Sometimes data in these fields can be expressed in a combination of "Yes", "1", "No", "0" if sources are mixed up without consistency and care, so it is important to unify the formatting in those fields.

B. Data Analysis

B1. Technique for Review

Our analysis begins by selecting our technique for analysis. For the purposes of this paper, we will use chi-square to determine the relationships between categorical variables (area, marital status, gender, soft drinks, initial admission, high blood pressure, stroke, complication risk, overweight, arthritis, diabetes, hyperlipidemia, back pain, anxiety, allergic rhinitis, reflex esophagitis, asthma, services, and our eight survey questions) and readmissions. While other methods like T-Test and ANOVA focus on continuous variables, Chi-Square is best used on categorical variables like the ones we are looking at today.

First we define our chi-square function:

```
# Define the Chi-Square Test
def chi_square(df, target_col, cat_col):
    # Contingency table
    contingency_table = pd.crosstab(df[cat_col], df[target_col])

    # Chi-square test
    chi2, p, dof, expected = chi2_contingency(contingency_table)

    # Show test and p-value
    return chi2, p
```

Then we run the test:

```
target_column = 'ReAdmis'
categorical_columns = [
    'Area', 'Marital', 'Gender', 'Soft_drink', 'Initial_admin',
    'HighBlood', 'Stroke', 'Complication_risk', 'Overweight',
    'Arthritis', 'Diabetes', 'Hyperlipidemia', 'BackPain', 'Anxiety',
    'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma', 'Services',
    'Item1', 'Item2', 'Item3', 'Item4', 'Item5', 'Item6', 'Item7', 'Item8'
]

# Store results in a dictionary
chi_square_results = {}

for col in categorical_columns:
    chi2, p = chi_square(medical_data_expressed, target_column, col)
    chi_square_results[col] = {'Chi2': chi2, 'p-value': p}

# Convert to DataFrame for easy viewing
chi_square_df = pd.DataFrame(chi_square_results).T
print(chi_square_df)
```

In our results, Chi2 will represent the accumulated difference between observed and expected data. A higher Chi2 number corresponds to higher differences versus expectations, so we'll be looking for results that are above the baseline. P-Value will represent the probability of observing this data if there were no correlation between that variable and ReAdmis. A lower number, especially anything below 0.05, would indicate a statistically significant relationship between the variable and ReAdmis. Finding variables with high Chi2 and low p-value results will be red flags toward identifying variables with relationships to readmissions.

B2. Output

After running our code, we find the following results:

	Chi2	*p-value*
Area	*0.713313*	*0.700013*
Marital	*5.085194*	*0.278667*
Gender	*1.585771*	*0.452537*
Soft_drink	*0.557316*	*0.455344*
Initial_admin	*3.889968*	*0.142990*
HighBlood	*0.042397*	*0.836866*
Stroke	*0.004340*	*0.947477*
Complication_risk	*0.159022*	*0.923568*
Overweight	*0.698480*	*0.403295*
Arthritis	*0.554512*	*0.456480*
Diabetes	*0.079833*	*0.777524*
Hyperlipidemia	*0.167074*	*0.682726*
BackPain	*1.716615*	*0.190129*
Anxiety	*0.047705*	*0.827107*
Allergic_rhinitis	*0.196991*	*0.657161*
Reflux_esophagitis	*0.271563*	*0.602285*
Asthma	*2.857452*	*0.090951*
Services	*8.892645*	*0.030753*
Item1	*6.826957*	*0.447117*
Item2	*9.129144*	*0.166444*
Item3	*9.876953*	*0.195654*
Item4	*5.007596*	*0.542839*
Item5	*1.615424*	*0.951462*
Item6	*6.090464*	*0.413133*
Item7	*6.289736*	*0.391528*
Item8	*7.680676*	*0.262444*

We have several variables with high Chi2 values like Marital, Initial Admin, and the our survey results, but those all have higher p-values so they're unlikely to be significant. The most important variable is Services, with a Chi2 value of 8.892645 and a p-value of 0.030753. This is the only p-value that is statistically significant and it invites us to explore the relationship to hospital readmissions in more depth. Asthma was our closes "second place" variables with a Chi2 of 2.857452 and a p-value of 0.090951 (not below 0.05 but worth investigating further).

B3. Explanation

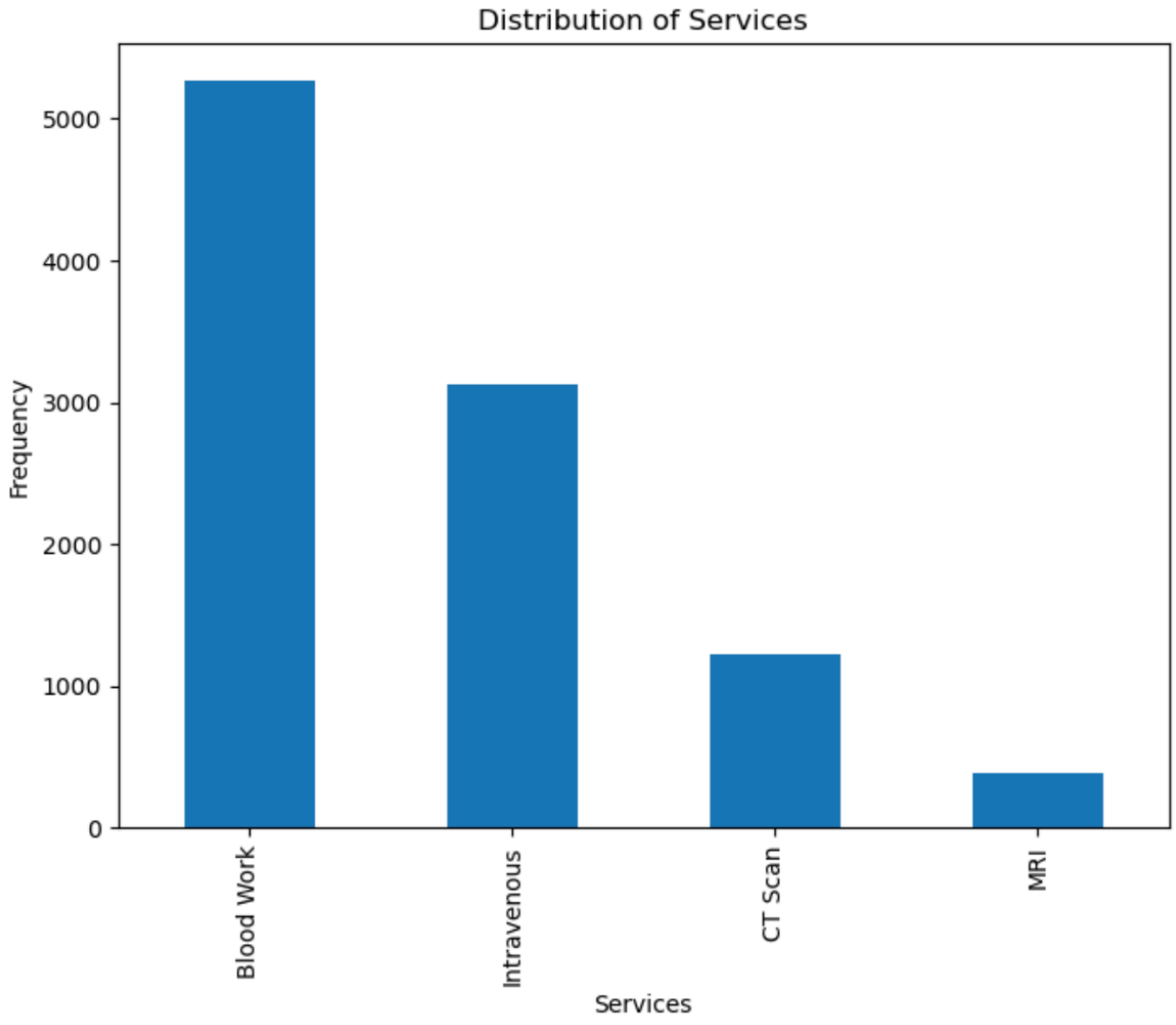
We opted to leverage chi-square for our test because my goal was to investigate categorical data for potential leads first. Across the three analysis tools

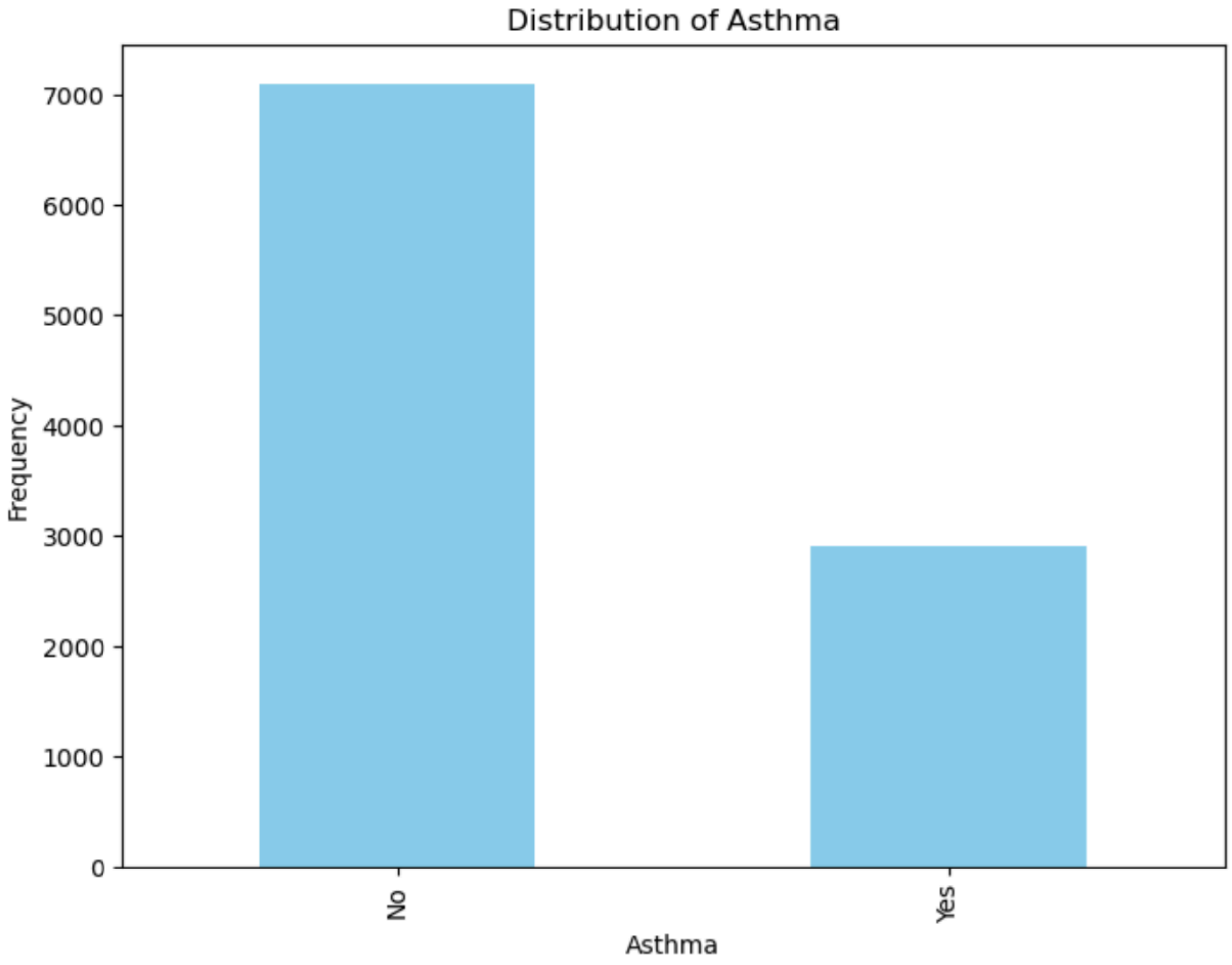
we had at our disposal, chi-square is the only one tailored towards categorical analysis. Using this produced one key variable that we can analyze further to find out how it relates to readmissions. This should give us the insight we need to determine how to use that data as a lever to reduce readmissions in the future.

C. Univariate Distribution

C1. Variable Distribution

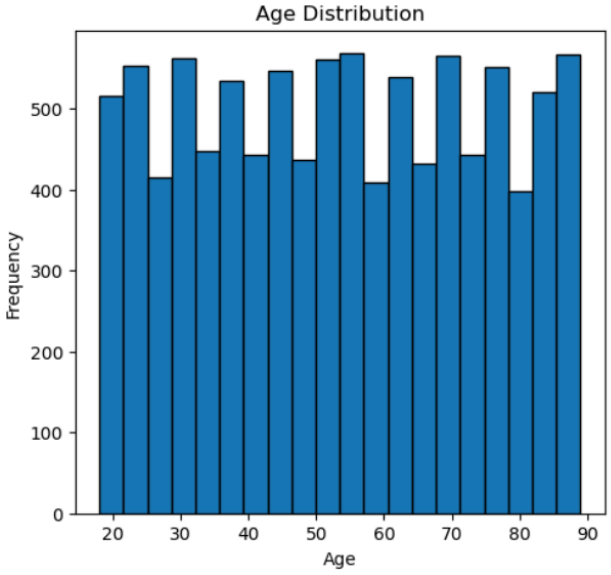
Using univariate statistics, we should analyze specific variables in more depth. Let's take a look at our two interesting categorical variables (Services and Asthma). Afterwards, we'll review two continuous variables as well. By doing this we can better see the distributions and spreads for these metrics.

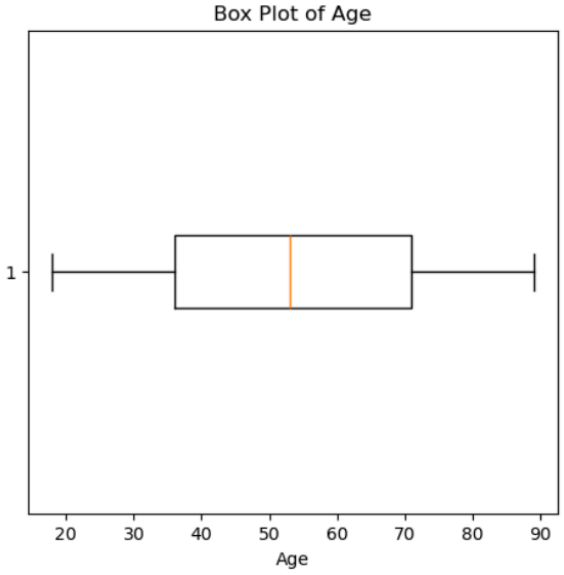


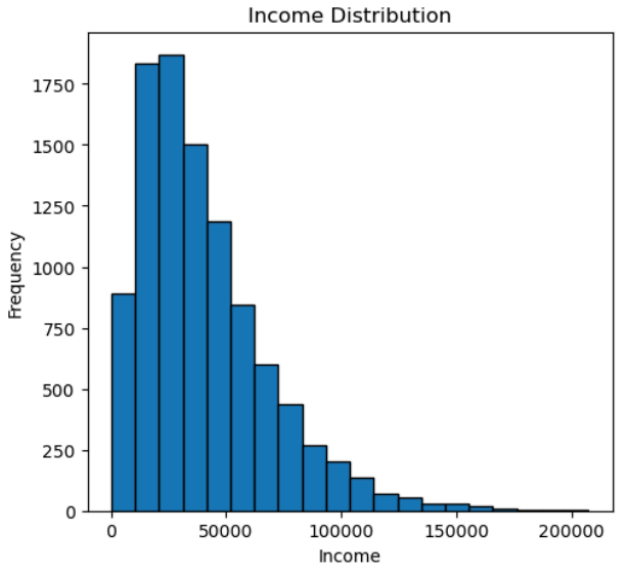


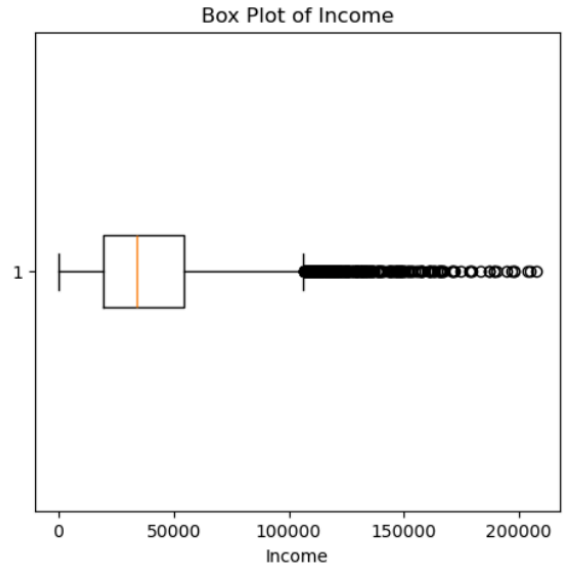
This was performed in Python using the following code (Larose & Larose, 2019).

```
# Bar plot: Services
plt.figure(figsize=(8, 6))
services_counts.plot(kind='bar')
plt.title("Distribution of Services")
plt.xlabel("Services")
plt.ylabel("Frequency")
plt.show()

# Bar plot: Asthma
plt.figure(figsize=(8, 6))
asthma_counts.plot(kind='bar', color='skyblue')
plt.title("Distribution of Asthma")
plt.xlabel("Asthma")
plt.ylabel("Frequency")
```










```
plt.show()
```

This was performed in Python using the following code.

```
age_summary = medical_data_expressed['Age'].describe()
income_summary = medical_data_expressed['Income'].describe()

# Hist and Box: Age
plt.figure(figsize=(12, 5))

# Hist
plt.subplot(1, 2, 1)
plt.hist(medical_data['Age'], bins=20, edgecolor='black')
```

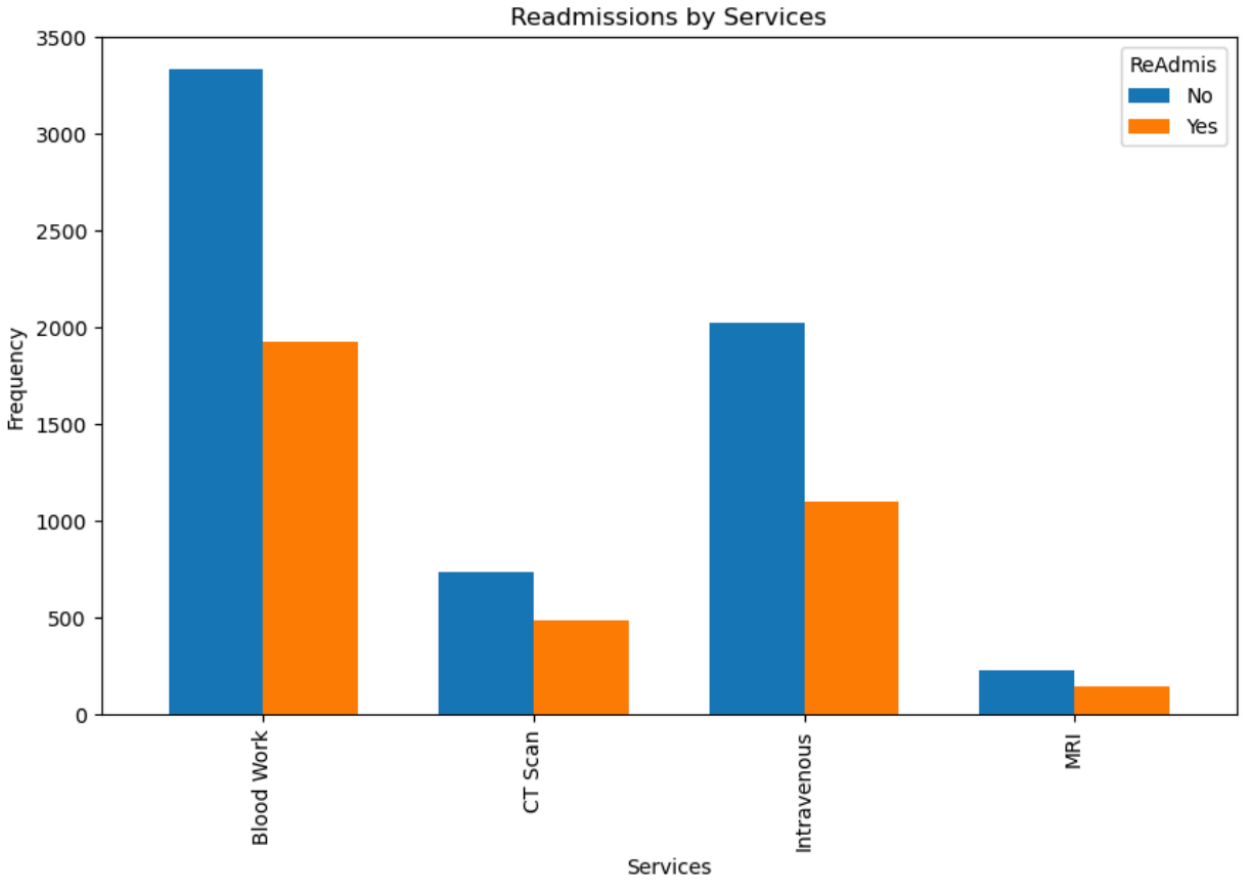
```
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Frequency")

# Box
plt.subplot(1, 2, 2)
plt.boxplot(medical_data['Age'], vert=False)
plt.title("Box Plot of Age")
plt.xlabel("Age")

plt.show()

# Hist and Box: Income
plt.figure(figsize=(12, 5))

# Hist
plt.subplot(1, 2, 1)
plt.hist(medical_data['Income'], bins=20, edgecolor='black')
plt.title("Income Distribution")
plt.xlabel("Income")
plt.ylabel("Frequency")

# Box
plt.subplot(1, 2, 2)
plt.boxplot(medical_data['Income'], vert=False)
plt.title("Box Plot of Income")
plt.xlabel("Income")

plt.show()
```

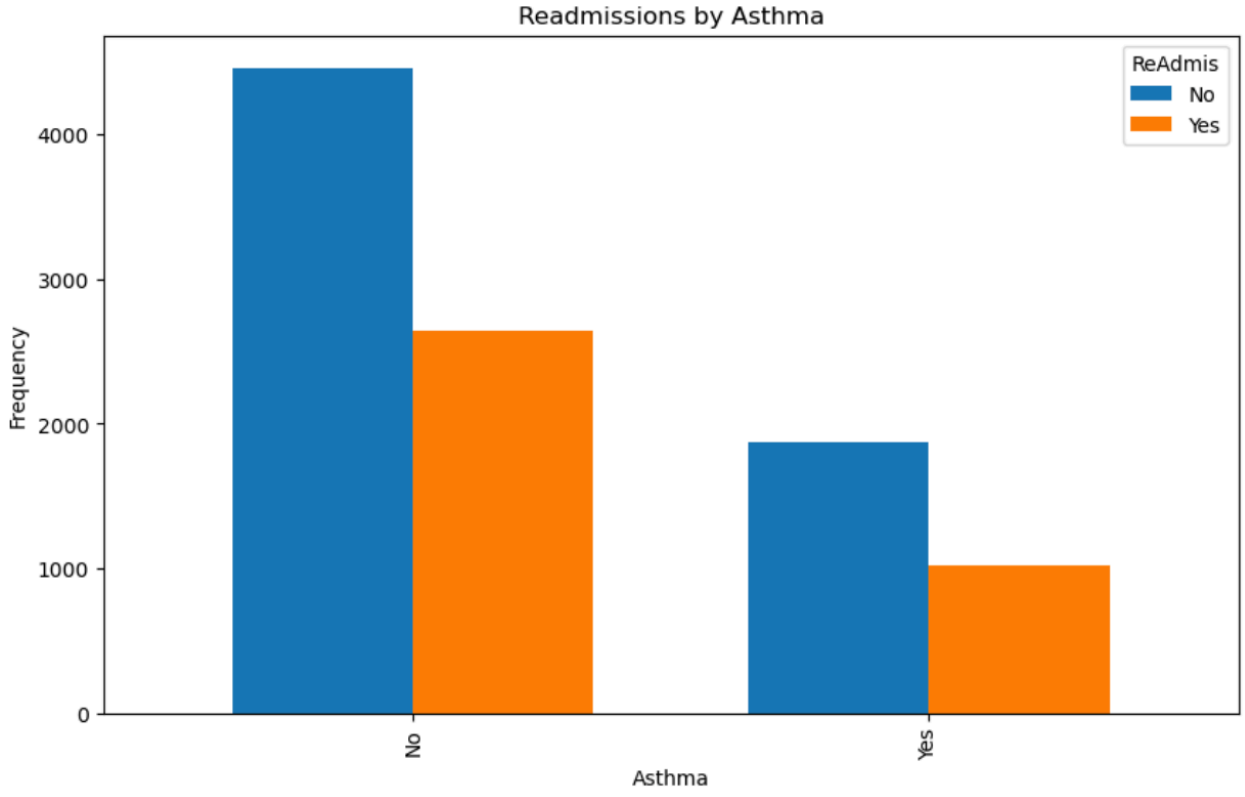
It is worth noting that there are outliers in the Income category. All of the relevant outliers within Income exist at the high end of the spectrum (the skew). If these outliers could jeopardize our research question, it would be worth reviewing and perhaps even removing the outliers altogether. Our primary question is readmissions, however; so we will leave this data in so can see a full perspective from our data without any adjustments.

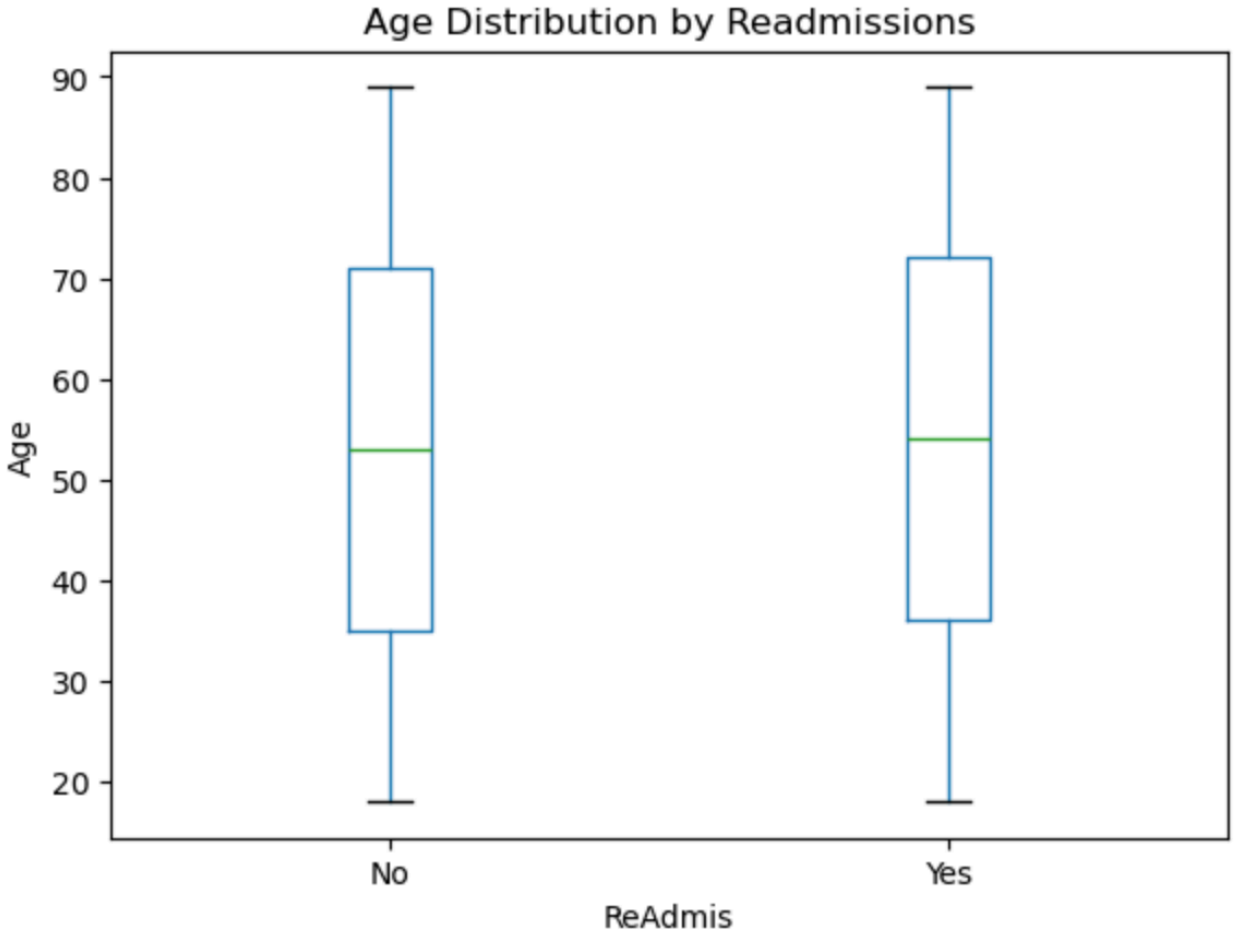
D. Bivariate Distribution

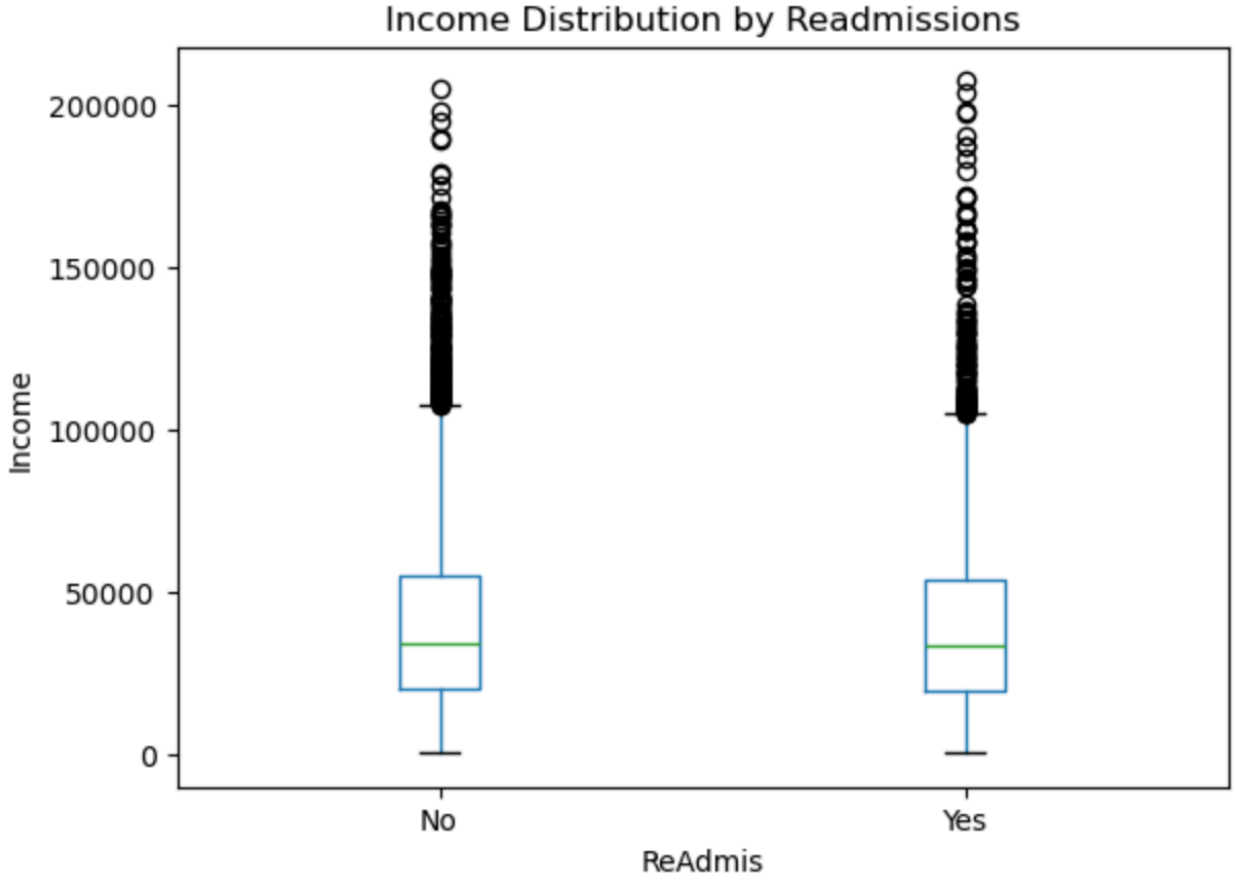
D1. Bivariate Distribution

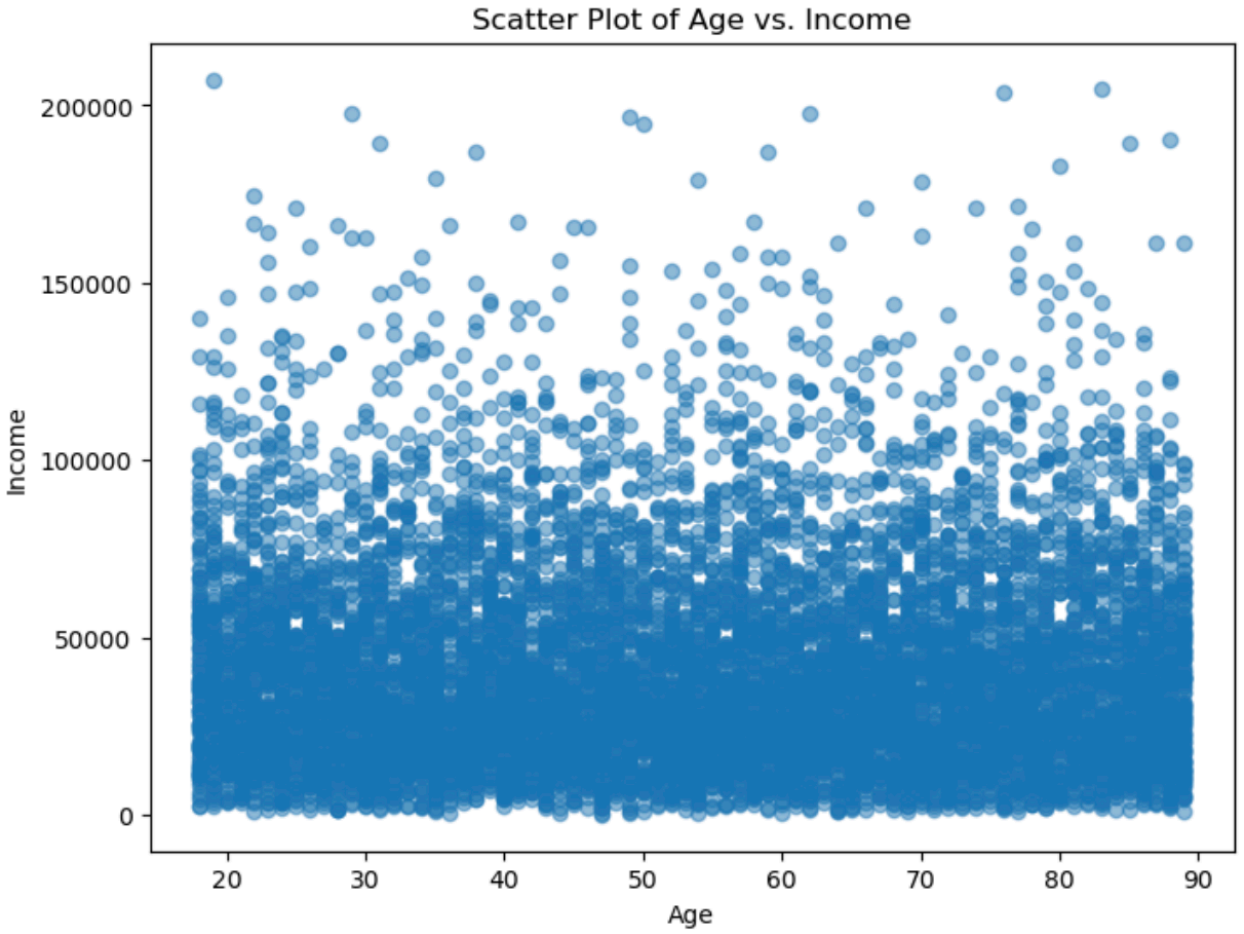
We've looked at univariate distribution with our four variables, so let's now turn to a bivariate analysis - reviewing each variable against hospital readmissions. For


Readmissions by Services
ReAdmis
No
Yes
Frequency
3500
3000
2500
2000
1500
1000
500
0
Blood Work
CT Scan
Intravenous
MRI
Services


Readmissions by Asthma
ReAdmis
No
Yes
Frequency
4000
3000
2000
1000
0
No
Yes
Asthma


Age Distribution by Readmissions
90
80
70
60
50
40
30
20
Age
No
Yes
ReAdmis


Income Distribution by Readmissions
200000
150000
100000
50000
0
Income
No
Yes
ReAdmis



categorical data, we'll review by graphing using a clustered bar chart to see values directly corresponding to "Yes" and "No" on ReAdmis. For continuous data, we'll look at box plots.

Although not corresponding to readmissions, I also performed a bivariate analysis on Age and Income. This was to demonstrate the bivariate technique on two continuous variables, and is not likely to result in any eureka moments.

E. Conclusion

E1. Results

Reviewing our analysis above, we can draw a few conclusions about our data. Overall, with our categorical data we identified Services as having the only statistically significant p-values correlating them to readmissions. This certainly warranted further review. While some survey questions had high Chi2 values, they all had relatively high p-values as well. So when opting to identify a secondary categorical variable to investigate, we opted for Asthma, which had a relatively low p-value compared to others at 0.09 (which is still higher than our 0.05 threshold for statistical significance).

When we look at univariate distribution of our categorical variables, we can see that Services has a right skew (also referred to as a positive skew) from Blood Work (peak) to Intravenous, to CT Scan and finally MRI. Asthma had a "No" skew in overall responses compared to "Yes" for this metric, indicating not as many patients had asthma.

Continuing our univariate analysis to continuous variables, we opted to review Age and Income. Age had a uniform distribution with no strong skew or clustering. The median was well centered and there were no outliers within the data. Income was not uniformly distributed - it was right-skewed with a concentration in the lower end of the spectrum. The median represented that by being centered towards the left of the box, with the box on the lower side of the full spectrum. A wide range of outliers exist within income - all positioned on the high end of the income spectrum.

When we look at a bivariate analysis, things begin to get more interesting because we can see relationships (or a lack thereof) to our readmissions statistic. For example, when comparing Services to Readmissions, we can see that Blood Work and Intravenous had high volumes of readmissions. It may be worth

investigating further, but we also see that CT Scans and MRI appeared to have a higher proportion of readmissions (when compared to it's value count), versus Blood Work and Intravenous. That could indicate that although blood work and IVs are more common, the chance of a readmission following a CT Scan or MRI might be higher.

For Asthma, there did not appear to be a proportional skew similar to CT Scan or MRI, but there does appear to be a slight bump in the proportion of asthma sufferers having readmission compared to otherwise. While we can say that more people without asthma needed readmissions, but that appears to just be the result of fewer people having asthma in general. Correlation does not equal causation in this case.

Transitioning now to our bivariate analysis of continuous variables (age and income) compared to the categorical Readmissions, we can see that median age sits very centered in the box, showing that no age groups skew the likelihood of readmissions for this metric. Readmissions to not appear to concentrate on any specific quartiles, and there are no outliers in the group. Age does not appear to play a role in readmissions.

In our income box plot, we can see that the median centered around the lower end of the full spectrum (most incomes were in the lower range), with outliers on the high end of income. That long right tail of outliers on the higher side of the income spectrum is not a source of distinction between responses on readmissions though. Here, we can say that income does not appear to play a role in readmissions.

Our last graph was a scatter plot of Age and Income - two continuous variables. This was to demonstrate the bivariate technique with continuous-only values, and it did not result in any significant realizations. The income is

distributed evenly across age, with outliers on the high end of the income spectrum not clustering on any particular group.

E2. Limitations

This review was limited to only a few select variables within the data. A broader exploratory analysis could be performed by reviewing other continuous variables in more depth with ANOVA or T-Tests. After identifying other variables worth exploring, a univariate and bivariate analysis of that data may produce additional levers to lower readmissions overall. Alternatively, the continuous data could be broken down into groups based on the distribution of their values, and then a chi-square test could be performed on that data as well.

E3. Next Steps

With the information we did gain, however; administration does have some options to focus on improving readmission rates. Blood work and IVs appear to have higher volumes of readmissions, so it might be useful to provide extra care up front to patients receiving these services. Blood work might produce more readmissions due to abnormal results, so processing blood work on-site immediately could reduce readmissions significantly for this metric.

We also noted that the proportion of readmissions for CT Scans and MRIs is greater than other services. It would be useful to analyze this in more detail by looking at the exact proportions to confirm our suspicions. If that confirmation is statistically significant, readmissions might be reduced by having additional resources devoted to immediate analysis of scan results. Like blood work, the idea here is that readmissions may be occurring due to scan reviews that take place after the patient has been discharged. Offering immediate analysis might allow for immediate treatments of unexpected results. Patients might also then be referred

to specialties that directly target whatever sources caused the abnormal results, improving overall patient care in the same process.

Lastly, as mentioned above, further analysis of continuous data should also be performed to identify other vectors of readmissions management. If no other statistically significant variables are found, management will need to look at other, more simplified ways of exploiting the data (like reducing the proportion of readmissions within specific areas, or targeting high-volume metrics with smaller process changes to incrementally improve performance).

References

Rau, J. (2022, October 31). Look Up Your Hospital: Is It Being Penalized By Medicare? KFF Health News. https://kffhealthnews.org/news/hospital-penalties/

Larose, C. & Larose, D. (2019). *Data Science Using Python and R*. Wiley.