Andreas Dunn

D207

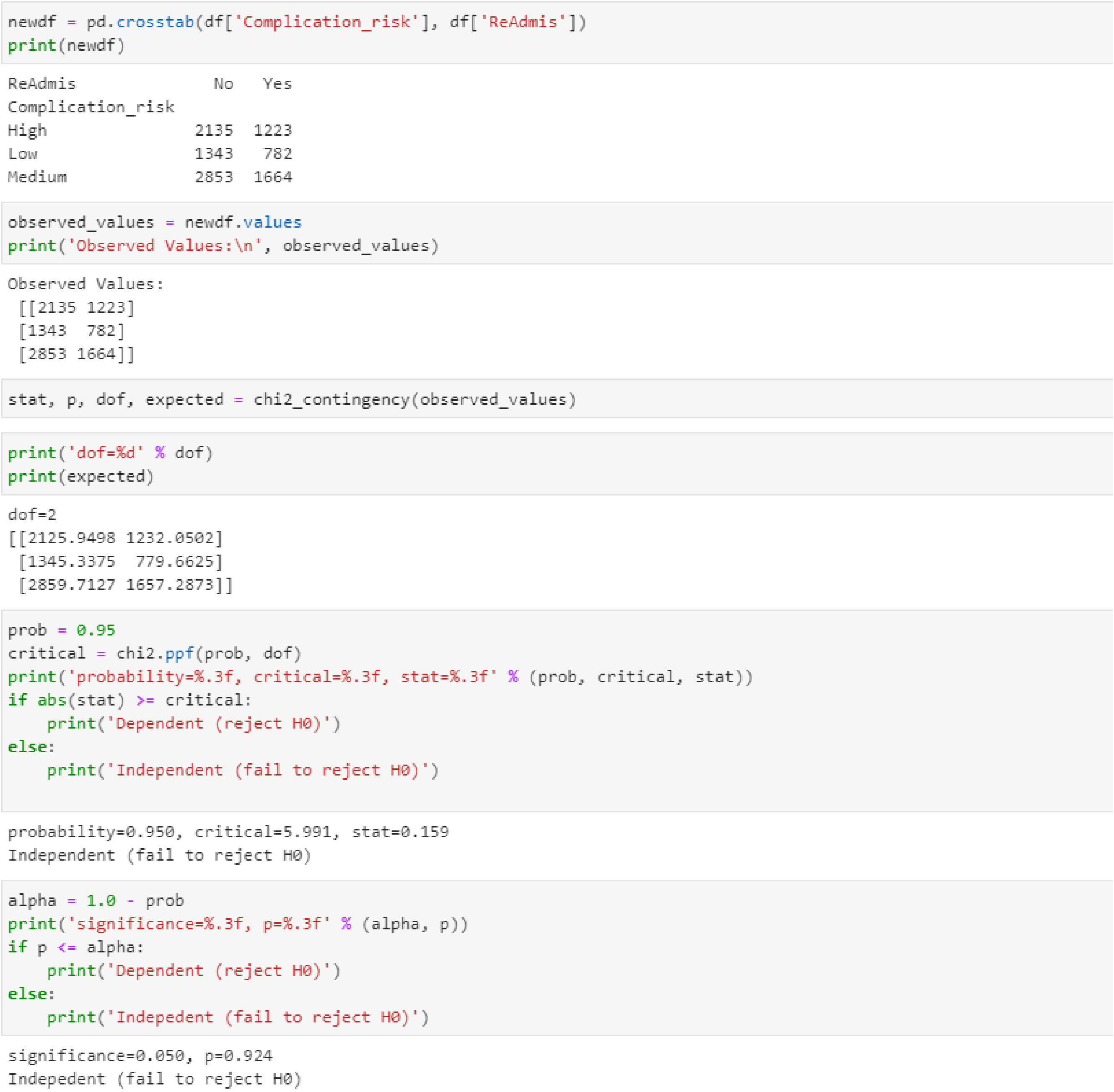
8/22/2021

A.

In this paper I will attempt to answer the following: is there a statistically significant difference between the low, medium, and high levels of complication risk on readmission rates? Put another way, does the proportion of the ReAdmis variable differ across the three levels of the Complication_risk variable? In this analysis I will break down the Complication_risk column and discover any differences between those who answer low, medium, or high in trying to better understand how the data could impact ReAdmis(readmission rates). Internal stakeholders benefit from this paper as they can further their understanding of how complication risks can impact readmission rates and perhaps save the hospital money by seeking new policies. External stakeholders can benefit by knowing if they have an increased risk of being readmitted to the hospital. Firstly I will deal with performing the chi-squared test on our two variables: Complication_risk and ReAdmis. Later in the paper I will visually show the distribution of some other variables from our data set. Finally, I will conclude with a discussion about the previous two sections, some downsides to chi-square, and closing thoughts on the paper.

B.

I will be using the chi-squared analysis method to see if we can tell if there is a statistically significant difference between our independent variable Complication_risk and our dependent variable ReAdmis. I chose this technique because both our variables are categorical and the chi-squared test is the most suitable given our variables. Our null hypothesis is that there is no difference between the 3 groups of Complication_risk on readmission rates while our alternative hypothesis is that there is a difference. The code below shows how I coded the analysis using Python:



```
newdf = pd.crosstab(df['Complication_risk'], df['ReAdmis'])
print(newdf)
```

```
ReAdmis              No   Yes
Complication_risk
High               2135  1223
Low                1343   782
Medium             2853  1664
```

```
observed_values = newdf.values
print('Observed Values:\n', observed_values)
```

```
Observed Values:
 [[2135 1223]
 [1343  782]
 [2853 1664]]
```

```
stat, p, dof, expected = chi2_contingency(observed_values)
```

```
print('dof=%d' % dof)
print(expected)
```

```
dof=2
[[2125.9498 1232.0502]
 [1345.3375  779.6625]
 [2859.7127 1657.2873]]
```

```
prob = 0.95
critical = chi2.ppf(prob, dof)
print('probability=%.3f, critical=%.3f, stat=%.3f' % (prob, critical, stat))
if abs(stat) >= critical:
    print('Dependent (reject H0)')
else:
    print('Independent (fail to reject H0)')
```

```
probability=0.950, critical=5.991, stat=0.159
Independent (fail to reject H0)
```

```
alpha = 1.0 - prob
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
    print('Dependent (reject H0)')
else:
    print('Indepedent (fail to reject H0)')
```

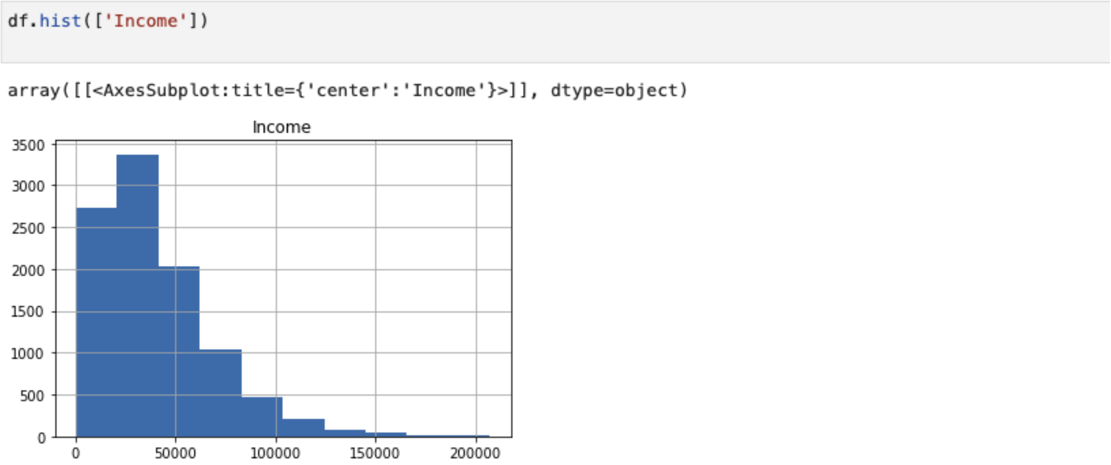
```
significance=0.050, p=0.924
Indepedent (fail to reject H0)
```

As you can see from our analysis, we compare the recorded values and then calculate expected values and then compare them to each other. I felt that chi-squared was appropriate for this analysis because both our variables are categorical and it provides a good explanation of how complication risk can impact ReAdmis.

C.

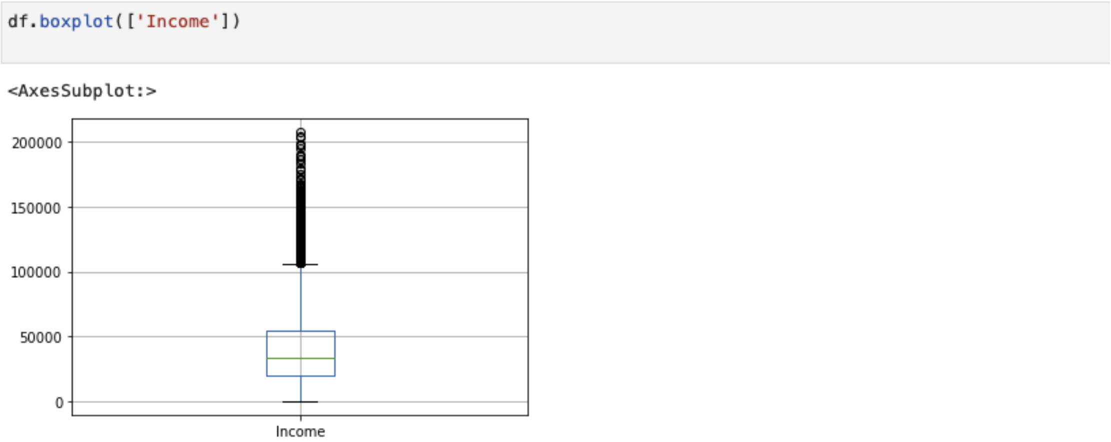
To further our understanding of this data I would like to talk about Income and Age and their possible relationship with readmission rates, as well as Services and ReAdmis and see what kind of distribution these variables have. Both Income and Age are continuous variables and Services and ReAdmis are categorical. I will first run uni-variate statistics and then conclude with bi-variate.

As you can see below, this histogram shows that income and readmission rates are negatively correlated to a high degree with a lot of outliers near the top of income levels.



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

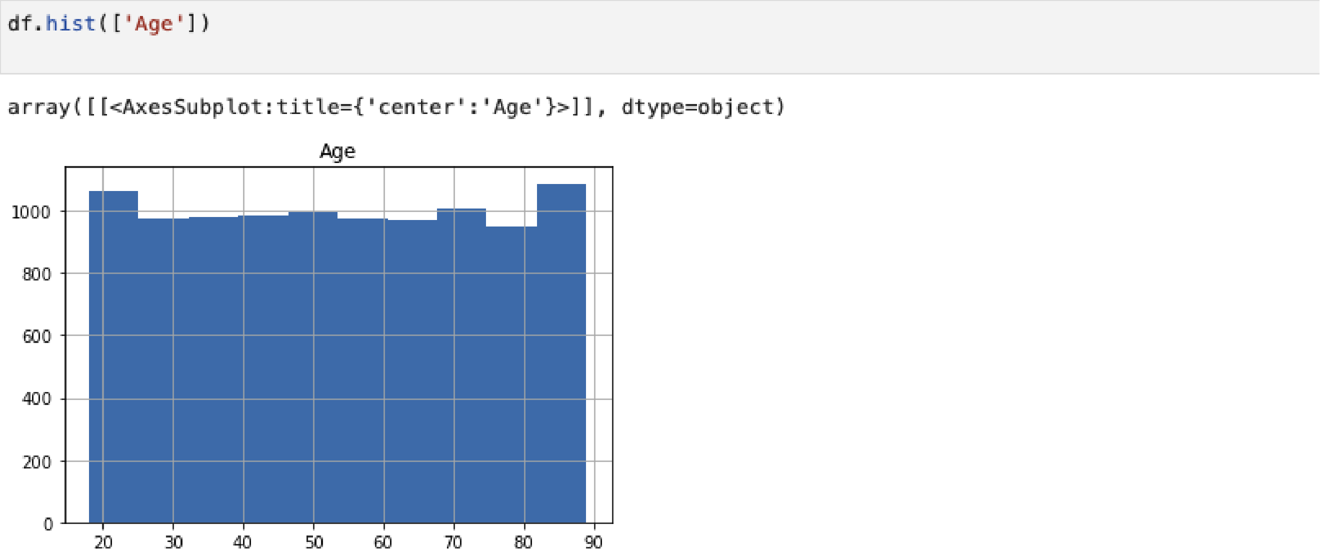
```
array([[<AxesSubplot:title={'center':'Income'}>]], dtype=object)
```



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

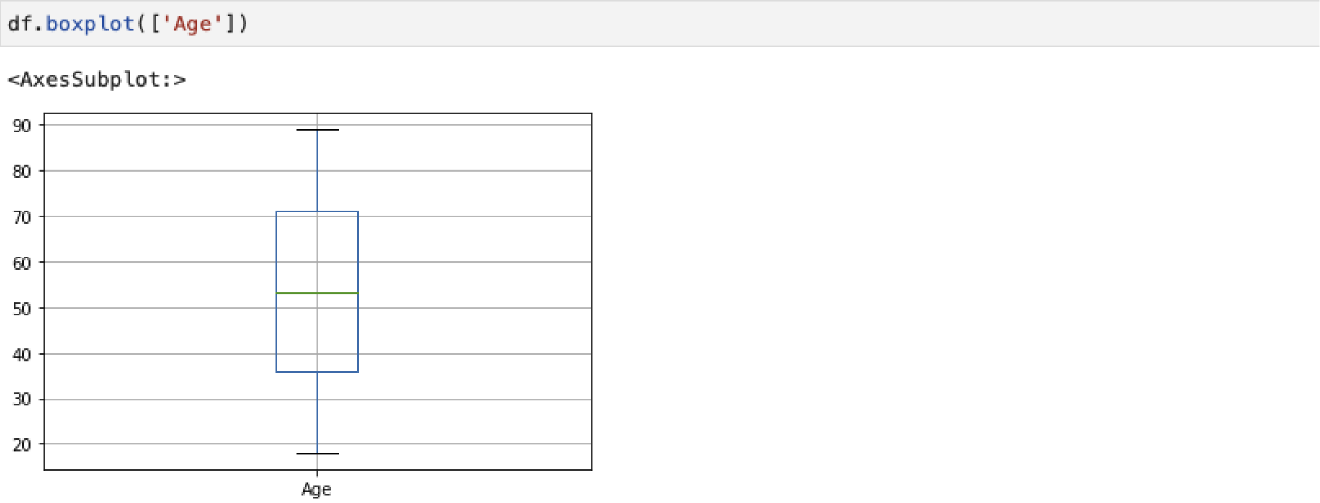
```
<AxesSubplot:>
```

Age, on the other hand, age has very little correlated data as we can tell from the following plots.



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

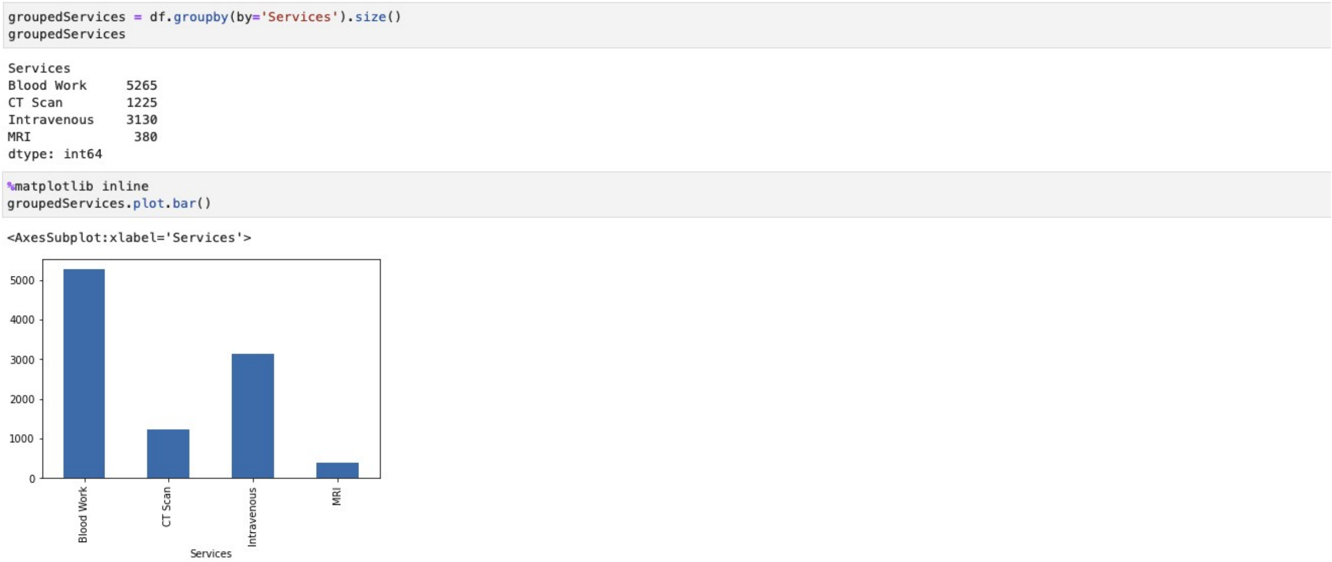
```
array([[<AxesSubplot:title={'center':'Age'}>]], dtype=object)
```



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

```
<AxesSubplot:>
```

Moving onto Services we see that Blood Work and Intravenous services are performed the most with the other two trailing far behind. Finally, we see that ReAdmis has many more in the ‘no' column over the ‘yes’ column.

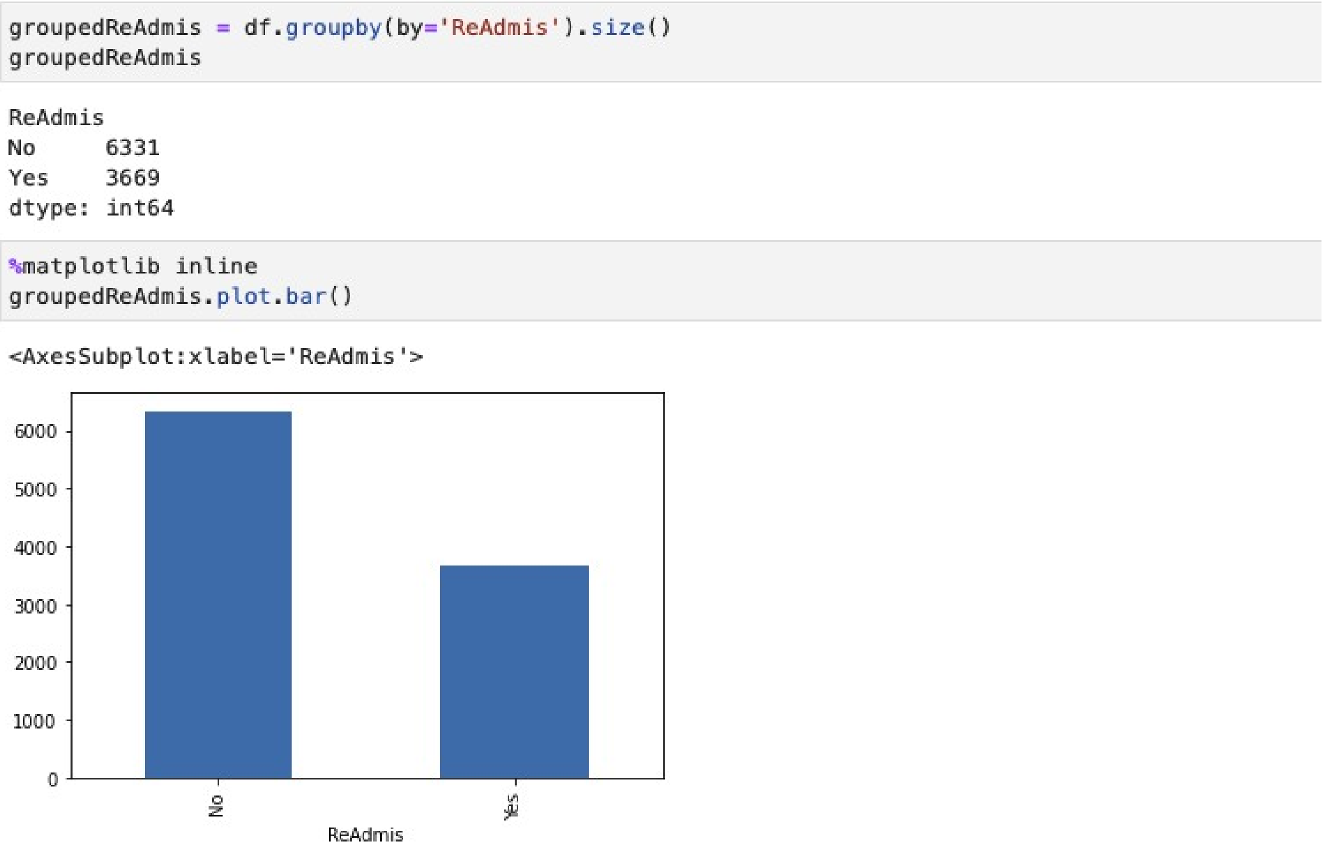


```
groupedServices = df.groupby(by='Services').size()
groupedServices
```

```
Services
Blood Work     5265
CT Scan        1225
Intravenous    3130
MRI             380
dtype: int64
```

```
%matplotlib inline
groupedServices.plot.bar()
```

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



```
groupedReAdmis = df.groupby(by='ReAdmis').size()
groupedReAdmis
```

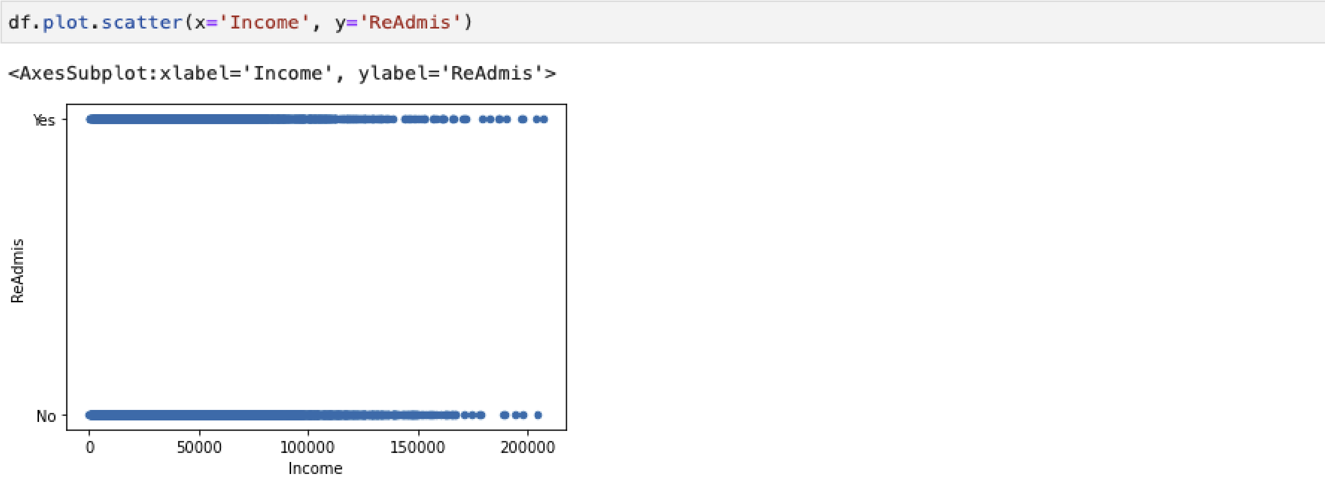
```
ReAdmis
No     6331
Yes    3669
dtype: int64
```

```
%matplotlib inline
groupedReAdmis.plot.bar()
```

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

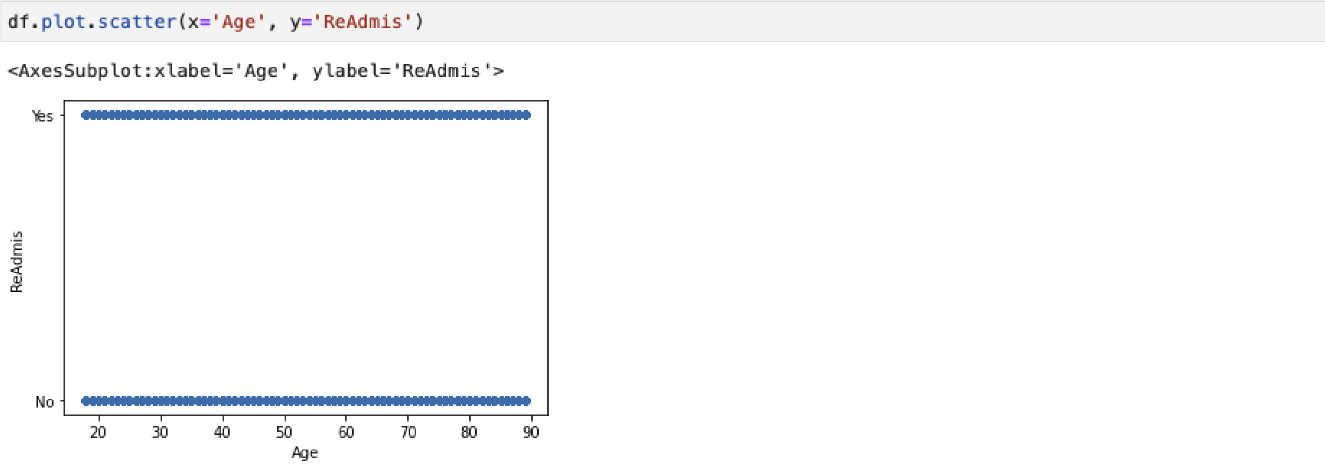
D.

Moving onto bi-variate statistics I will be using Income and Age as my two continuous variables and ReAdmis and Services as my two categorical variables.



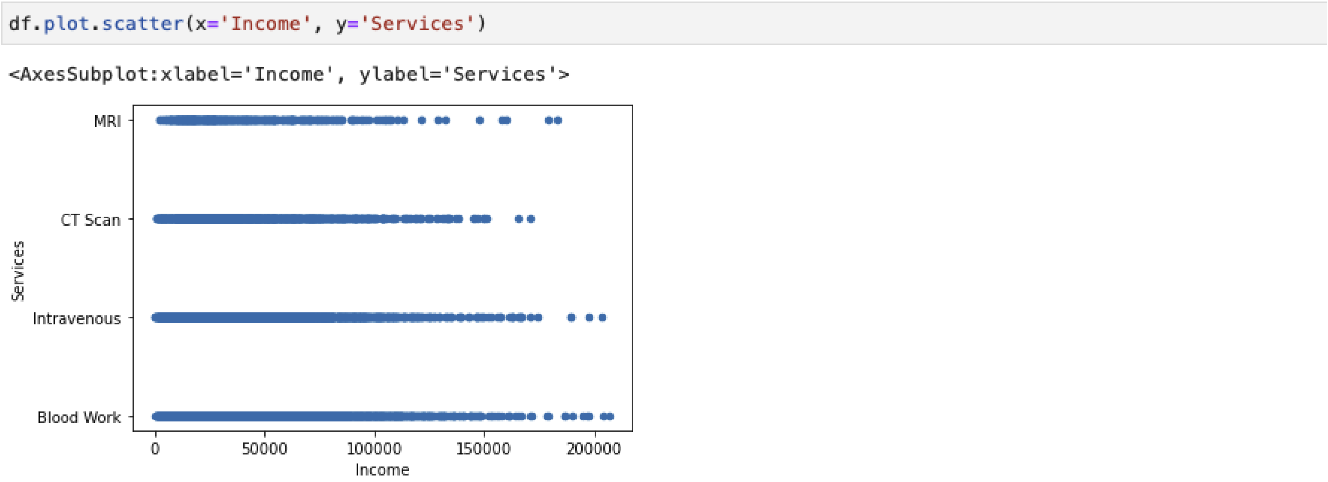
```
df.plot.scatter(x='Income', y='ReAdmis')
```

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



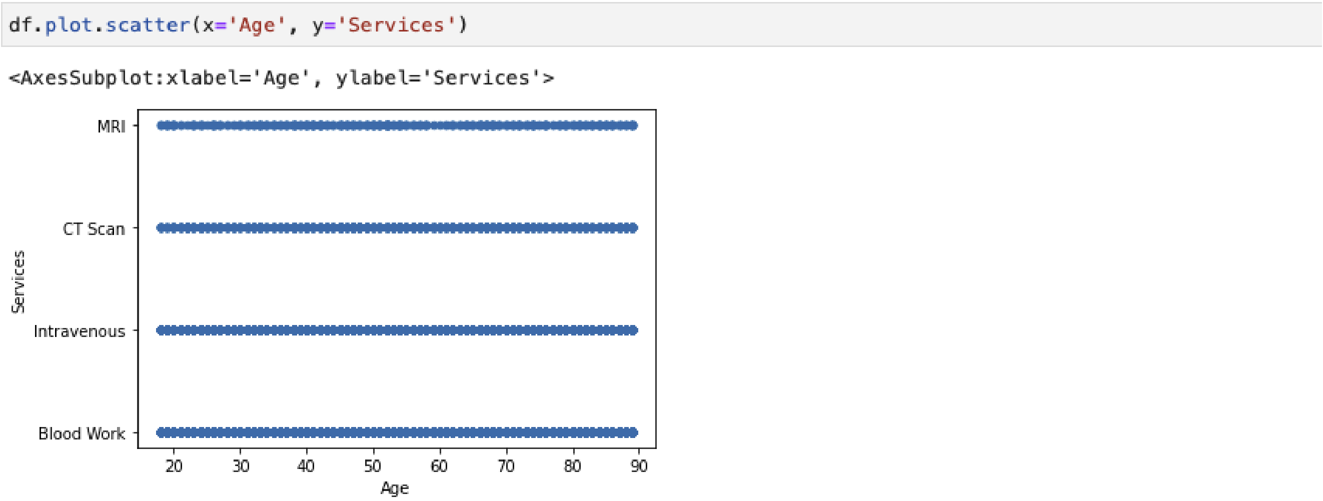
```
df.plot.scatter(x='Age', y='ReAdmis')
```

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



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

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



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

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

As is shown by the graphs above, there isn't a huge difference between levels of age when compared with services or ReAdmis but with Income we see many outliers near the top end. This suggests that those with higher incomes tend to use different services and are readmitted less often to the hospital.

E.

Our chi-squared test showed us many things, but the most important statistic to look at is our P-value which will inform us on if our data are meaningful in answering our research question or not. Our P-value was 0.924 which places it much higher than our 0.05 significance level. Therefore we have failed to reject the null hypothesis that our different groups of Complication_risk are significant in determining higher readmission rates: the difference between the groups is not statistically significant.

The limitations of using chi-squared are that it assumes that our sample is not biased which may not be the case since the hospital is in charge of collecting information on people who have at least *some* value of complication risk. The hospital did not, for instance, give anyone a zero for complication risk. Also the hospital chose to categorize complication risk into 3 choices whereas in reality maybe having a scale of 1-10 would be more accurate in determining complication risk. There is also human error to consider when dealing with assessments since there could have been different physicians that were calculating complication risk on their patients. Another limitation is that chi-squared is sensitive to sample size and since our sample size was so big (10,000 rows) we may have some biased towards finding statistical significance on samples that aren't substantive in answering our research question.

In conclusion, while there are limitations to using chi-squared and while our hypothesis test did not reveal anything significant, our analysis did show that there are some insights we can gain from the dataset such as showing a strong relationship between certain variables (i.e. Income and ReAdmis) and a weak relationship between other variables (Age and ReAdmis). My recommendation for the hospital would be to not use Complication_risk has a reliable source of determining readmission rates and instead look to use and analyze further the Income variable.

G.

All information and code was taken from course lectures or documentation found on the following:

https://pandas.pydata.org/

https://www.statsmodels.org/stable/index.html