Exploratory Data Analysis – D207

Western Governor's University

Performance Assessment

Matthew Morgan

Student ID: 010471280

9/27/2022

A1. For this assessment I am using the medical data set. The purpose of this assessment is to answer a question regarding patient readmissions. I chose to answer the following question, “Is there a relationship between the initial_days column and the ReAdmis column?”

A2. This data set points out that an external organization penalizes hospitals for excessive readmissions. This analysis of the data is the first step towards reducing patient readmissions for the hospital. The question I am answering can help identify a specific variable which leads to patients having a higher chance for readmission than others. This can then be used in the future to start enacting change to reduce readmissions. Ultimately, if readmissions can be reduced that will lead to less penalty payouts as well as better overall outcomes for patients.

A3. Because my question is very specifically about the relationship between the initial_days and ReAdmis columns, I will only be using those two columns for my analysis.

B1. To analyze my question, I ran a two-sample t-test. The code (Bowne-Anderson et al., n.d.) is as follows:

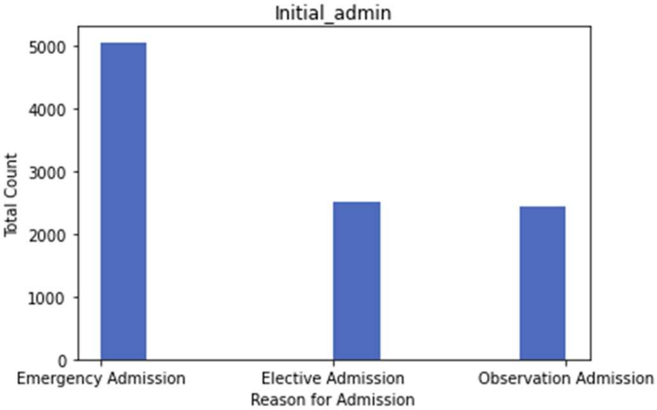
```
stats.ttest_ind(df['ReAdmis_numeric'], df['Initial_days'])
```

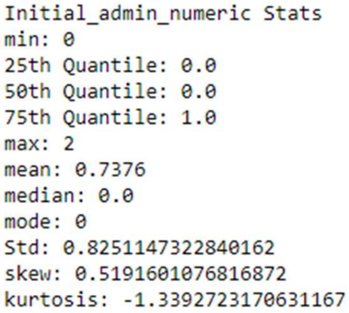
B2. The output was:

```
Ttest_indResult(statistic=-129.54592813419822, pvalue=0.0)
```

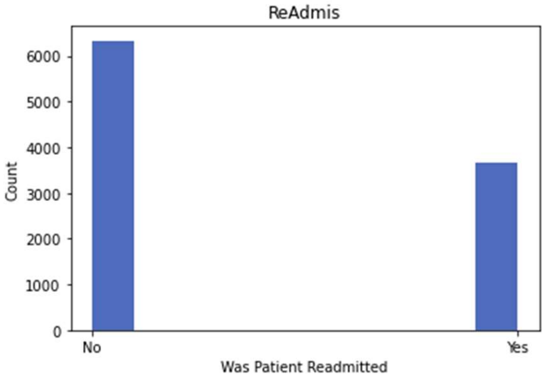
B3. I chose a t-test as I have one numeric and one categorical variable to compare. The initial_days column being numeric data and the ReAdmis being categorical.

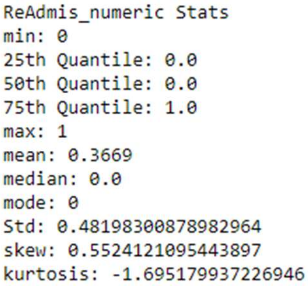
C1. Below are the distributions for four required variables used in the analysis for this report. They include two categorical (Initial_admin, and ReAdmis) variables as well as two continuous (Initial_days, Children) variables.



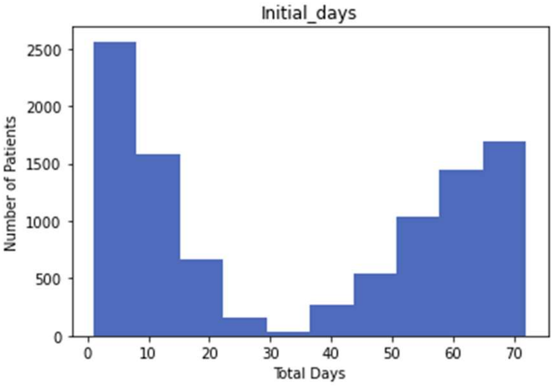


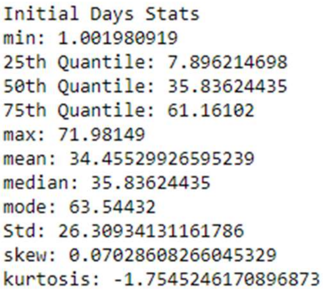
```
Initial_admin_numeric Stats
min: 0
25th Quantile: 0.0
50th Quantile: 0.0
75th Quantile: 1.0
max: 2
mean: 0.7376
median: 0.0
mode: 0
Std: 0.8251147322840162
skew: 0.5191601076816872
kurtosis: -1.3392723170631167
```



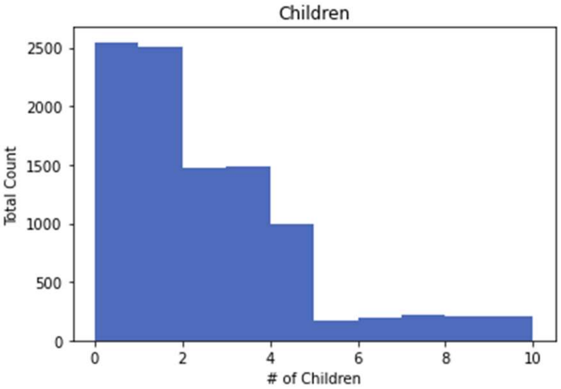


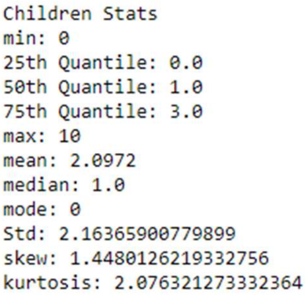
```
ReAdmis_numeric Stats
min: 0
25th Quantile: 0.0
50th Quantile: 0.0
75th Quantile: 1.0
max: 1
mean: 0.3669
median: 0.0
mode: 0
Std: 0.48198300878982964
skew: 0.5524121095443897
kurtosis: -1.695179937226946
```





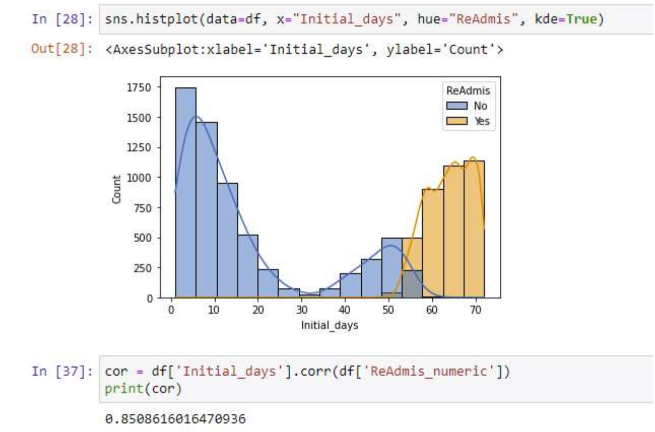
```
Initial Days Stats
min: 1.001980919
25th Quantile: 7.896214698
50th Quantile: 35.83624435
75th Quantile: 61.16102
max: 71.98149
mean: 34.45529926595239
median: 35.83624435
mode: 63.54432
Std: 26.30934131161786
skew: 0.07028608266045329
kurtosis: -1.7545246170896873
```





```
Children Stats
min: 0
25th Quantile: 0.0
50th Quantile: 1.0
75th Quantile: 3.0
max: 10
mean: 2.0972
median: 1.0
mode: 0
Std: 2.16365900779899
skew: 1.4480126219332756
kurtosis: 2.076321273332364
```

D1. Included below is the python code (Bowne-Anderson et al., n.d.) used to produce 3 bivariate distributions comparing Initial_days, Children, and Initial_admin to ReAdmis.

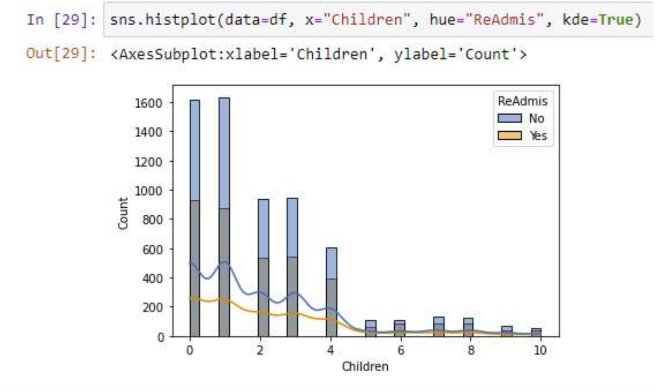


```
In [28]: sns.histplot(data=df, x="Initial_days", hue="ReAdmis", kde=True)

Out[28]: <AxesSubplot:xlabel='Initial_days', ylabel='Count'>
```

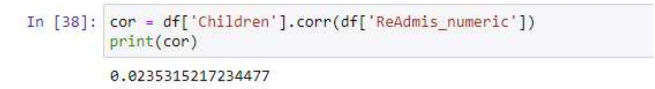
```
In [37]: cor = df['Initial_days'].corr(df['ReAdmis_numeric'])
         print(cor)

         0.8508616016470936
```



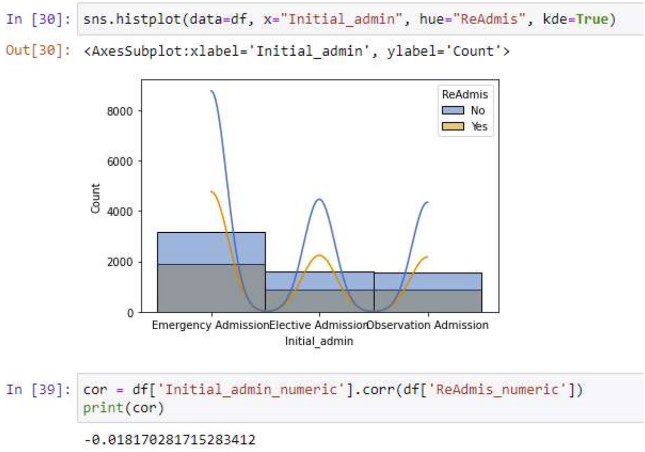
```
In [29]: sns.histplot(data=df, x="Children", hue="ReAdmis", kde=True)

Out[29]: <AxesSubplot:xlabel='Children', ylabel='Count'>
```



```
In [38]: cor = df['Children'].corr(df['ReAdmis_numeric'])
         print(cor)

         0.0235315217234477
```



```
In [30]: sns.histplot(data=df, x="Initial_admin", hue="ReAdmis", kde=True)

Out[30]: <AxesSubplot:xlabel='Initial_admin', ylabel='Count'>
```

```
In [39]: cor = df['Initial_admin_numeric'].corr(df['ReAdmis_numeric'])
         print(cor)

         -0.018170281715283412
```

E1. Based on the results of the t-test, the p-value between Initial_days and ReAdmis is 0. This shows that we can reject the null hypothesis. Based on the question being asked, “Is there a relationship between the Initial_days column and the ReAdmis column?” the alternative and null hypothesis would be as follows:

Ha: There is a relationship between the Initial_days and the ReAdmis column.
Ho: There is no relationship between Initial_days and the ReAdmis column.

Again, because the p-value of our t-test with Initial_days vs ReAdmis is < 0.05 we have sufficient evidence to reject our null hypothesis.

E2. I only tested for a correlation between readmissions and one other variable. In the exploratory data analysis phase I imagine we’d want to try and find all the variables that have a relationship with readmissions before proceeding.

E3. Because Initial_days has a strong relationship with patient readmissions the next course of action would be to find other variables that are related to readmissions as well. One possible avenue is to also look at variables or trends for patients who have long initial hospital stays. If we can modify something about patients who have long hospital stays then maybe we can lower patient readmissions.

F. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e039cfe9-f5d6-428f-a3c8-af1e01650db5

G. Bowne-Anderson, H., Scouwenaars, F., Matsui, Maggie., Ramirez M., Alexander A.., (n.d.) *D207 – Exploratory Data Analysis* [OC]. Datacamp. https://app.datacamp.com/learn/custom-tracks/custom-d207-exploratory-data-analysis

H. No sources quoted, paraphrased or summarized.