D207 Exploratory Data Analysis

Sam Spencer

Western Governors University

**A1: Research Question**

Can we correlate how long a patient stays in the hospital (initial\_days) to the likelihood of said patient being readmitted?

**A2: Benefit from Analysis**

By analyzing these data for patterns and trends, the organization and its stakeholders will be able to predict the likelihood of a patient being readmitted with greater accuracy. Furthermore, having more insight and intelligence into the potential outcomes would allow the company to better estimate both the cost and resources required by their facilities, as well as the potential fines if they fail to meet CMS's guidelines.

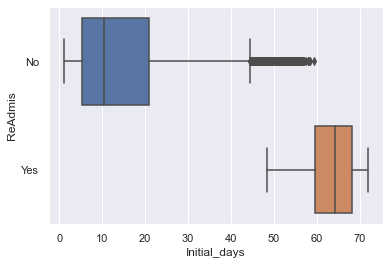
**A3: Data Identification**

The "Initial days" and "ReAdmis" variables from the medical clean.csv dataset are pertinent information needed to investigate the aforementioned query. These two should be sufficient to determine whether or not a patient's initial hospital stay is a reliable indicator of their likelihood of readmission.

**B1. &B2. Code & Output**

| #import the required libraries and packages import numpy as np import pandas as pd from scipy import stats import matplotlib.pyplot as plt import seaborn as sns  #Reading the variables that are required.  df\_b = pd.read\_csv('/Users/samtravel/Desktop/3kfid8emf9rkc9ek30sf (1)/medical\_clean.csv',   usecols=['Initial\_days',   'ReAdmis']) |
| --- |

| # visualization of relationship between initial days hospitalized and readmission yes/no sns.boxplot(y='ReAdmis',   x='Initial\_days',   data=df\_b) |
| --- |



**T-Test code and output**

| # create groups readmitted yes/no with length of initial hospitalization  readmit\_yes = df\_b[df\_b['ReAdmis'] == 'Yes'].Initial\_days readmit\_no = df\_b[df\_b['ReAdmis'] == 'No'].Initial\_days  # set t-test and print raw result ttest\_result = stats.ttest\_ind(readmit\_yes,   readmit\_no) print(ttest\_result)  # if t-test p-value is less than alpha (.05), reject null hypothesis, else do not reject alpha = .05 if (ttest\_result[1] < alpha):  print(f'T-Test p-value returned as: {str(ttest\_result[1])} with an alpha of {str(alpha)}. \nP-value falls within .05 alpha: \nReject the null hypothesis.') else:  print(f'T-Test p-value returned as: {str(ttest\_result[1])} with an alpha of {str(alpha)}. \nP-value does not fall within .05 alpha: \nAccept the null hypothesis.') |
| --- |

**T-Test output:** Ttest\_indResult(statistic=161.9324105776643, pvalue=0.0)

T-Test p-value returned as: 0.0 with an alpha of 0.05.

P-value falls within .05 alpha:

Reject the null hypothesis.

**B3: Justify**

Because of the alternative hypothesis, there is a direct correlation between the initial length of a patient's hospitalization and the said patient's probability of being readmitted, requiring the analysis of one continuous variable, which is initial\_days, and one binary, categorical variable (ReAdmis).

### C: Univariate Statistics

We will examine the distributions of observed patient levels as well as the initial length of time (in days) patients were hospitalized for (Initial days) for our visualization of two continuous variables.

We will examine the binary distributions of whether or not (yes/no) a patient was readmitted after initial hospitalization (ReAdmis) as well as whether or not (yes/no) a patient was identified as having high blood pressure for our visualization of two categorical variables (HighBlood).

#### **Identifying continuous variable distributions**

As shown by the two peaks, the distribution of the initial length of hospitalization is bimodal.

Vitamin D levels have a normal distribution.

#### **Categorical variable distributions**

Because of its binary nature, patient readmission and high blood pressure are composed of a Bernoulli distribution.

### C1: Visual of Findings

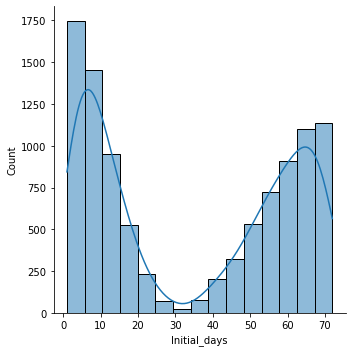
Visualizations are attached. See Panapto for demonstrations of the code

| #Univarete stat table df\_c = pd.read\_csv('/Users/samtravel/Desktop/3kfid8emf9rkc9ek30sf (1)/medical\_clean.csv',  usecols=['VitD\_levels',   'Initial\_days',  'ReAdmis',  'HighBlood']) |
| --- |

#### 

#### Continuous Variables

| #distribution of Initial\_days, using seaborn sns.displot(df\_c.Initial\_days,   kde=True) |
| --- |



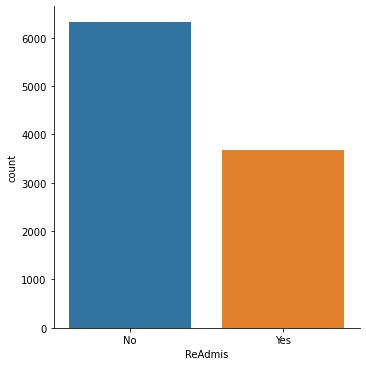
| # plotting distribution of VitD\_levels # red vertical line indicates the accepted minimum Vit D threshold of 20ng/mL sns.displot(df\_c.VitD\_levels,   kde=True) plt.axvline(20,   color='r') |
| --- |

According to the National Institutes for health “Levels of 50 nmol/L (20 ng/mL) or above are adequate for most people for bone and overall health.”

#### 

#### Categorical Variables

| sns.catplot(x='ReAdmis',   kind='count',   data=df\_c) |
| --- |



| sns.catplot(x='HighBlood',   kind='count',   data=df\_c,   order=['No',   'Yes']) |
| --- |

| Vitamin D and initial days correlation  sns.scatterplot(x='Initial\_days',   y='VitD\_levels',   data=df\_c) |
| --- |

| # Y/N for high blood pressure patients readmitted. sns.countplot(x=df\_c[df\_c.ReAdmis == 'Yes'].HighBlood) plt.xlabel('Readmitted Patients w/ High Blood Pressure') |
| --- |

### D: Bivariate Statistics

For our visualization of the two continuous variables, we will again visualize observed patient vitamin D levels (VitD levels) as well as the initial length of time (in days) patients were hospitalized for (Initial days). This time, we'll see if there's a clear connection between the two.

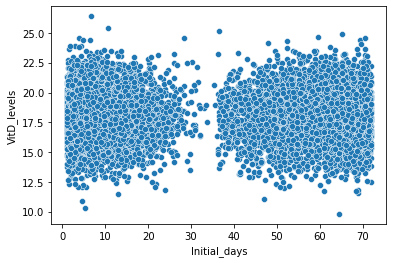
We will look at patients who were readmitted after their initial hospitalization (ReAdmis) and whether or not a patient was identified as having high blood pressure (yes/no) for our visualization of two categorical variables (HighBlood).

### 

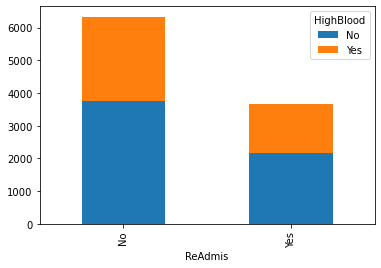
### D1: Visualization of Findings

| # scatter plot for correlation of VitD\_levels and Initial\_days sns.scatterplot(x='Initial\_days',   y='VitD\_levels',   data=df\_c) |
| --- |

Scatter plot output for correlation of Vitamin D levels and Initial Days.



| tbl = pd.pivot\_table(df.groupby(['ReAdmis','HighBlood']).size().reset\_index(), values=0, index='ReAdmis', columns=['HighBlood'], aggfunc=np.sum) tbl.plot(kind='bar',stacked=True) |
| --- |



### E1: Results of Analysis

My initial research question was to see if we could find a statistically significant link between a patient's initial length of hospitalization and the likelihood of being readmitted. The null hypothesis stated that there is no statistically significant difference in the distribution of initial length of hospitalization between patients who were readmitted and patients who were not readmitted. The alternative hypothesis was that there is a statistically significant difference. We used a t-test with a 5% alpha to examine this relationship. This test demonstrated that there was sufficient difference between two groups of the sample to reject the null hypothesis. A simple visualization showing comparative boxplots of patients who were readmitted or not was also added for ease of comprehension. The difference between the two groups is quite obvious in this visualization.

### E2: Limitations

Although a t-test is a great starting point for exploratory data analysis, the outcome can only be used to determine whether the hypothesis is worthwhile for further investigation. Furthermore, confounding variables' impacts were not taken into account in this study's limited analysis. We are now unable to identify the unknown that are both related to the metrics observed and impactful on the outcome.

### E3: Recommended Course of Action

I would suggest collecting more data on a broader set of qualities, accompanied by more analysis. Unfortunately, there were no strong indicators in the values I tested, which is a common result. This is frequently caused by not obtaining or testing enough data points that drive the conclusions.

There is a nearly infinite number of attributes that describe people and hospital visits, and many of these would interact to influence the likelihood of readmission. It could be individuals assigned to a case (nurses and doctors), the time of day or year of initial admission, or a variety of other factors.

Instead of randomly selecting some relationships to investigate, as previously stated, I would begin with an overall comparison tool, such as a PairGrid. It would provide a quick overview of the variables' distributions and whether they appear to have any tight groupings or trends in their plots.

### F: Third Party Code

*Dealing with rows and columns in pandas DataFrame*. GeeksforGeeks. (2021, October 13). Retrieved September 2, 2022, from https://www.geeksforgeeks.org/dealing-with-rows-and-columns-in-pandas-dataframe/

*Visualizing distributions of data¶*. Visualizing distributions of data - seaborn 0.11.2 documentation. (n.d.). Retrieved September 2, 2022, from https://seaborn.pydata.org/tutorial/distributions.html

Sources

U.S. Department of Health and Human Services. (n.d.). *Office of dietary supplements - vitamin D*. NIH Office of Dietary Supplements. Retrieved August 7, 2022, from https://ods.od.nih.gov/factsheets/VitaminD-Consumer/#:~:text=One%20nmol%2FL%20is%20the,and%20might%20cause%20health%20problems.