**EXPLORATORY DATA ANALYSIS**

**D207**

**Western Governors University**

**By Michael Rupert**

Table of Contents

[A1: Analysis Question 3](#_Toc104295523)

[A2: Benefits from the Analysis 3](#_Toc104295524)

[A3: Identification of Variables within Dataset 3](#_Toc104295525)

[B1: Code 4](#_Toc104295526)

[B3: Justification of Approach 6](#_Toc104295527)

[C: Univariate Statistics 8](#_Toc104295528)

[D: Bivariate 10](#_Toc104295529)

[E1: Results of Analysis 12](#_Toc104295530)

[E2: Limitations 12](#_Toc104295531)

[E3: Course of Action 13](#_Toc104295532)

[G: Sources for Third Party Code 13](#_Toc104295533)

[H: Sources 13](#_Toc104295534)

# A1: Analysis Question

We wish to answer the following question: **Based on the sample of historical data, can we develop a model that shows which patients are at risk of readmission?** From this question we can form a hypothesis that indicates if a patient has a higher probability of readmission based on certain reported medical conditions or other reported factors.

H0: There is no correlation between Readmission and other recorded variables.

H1: There is a correlation between Readmission and other recorded variables.

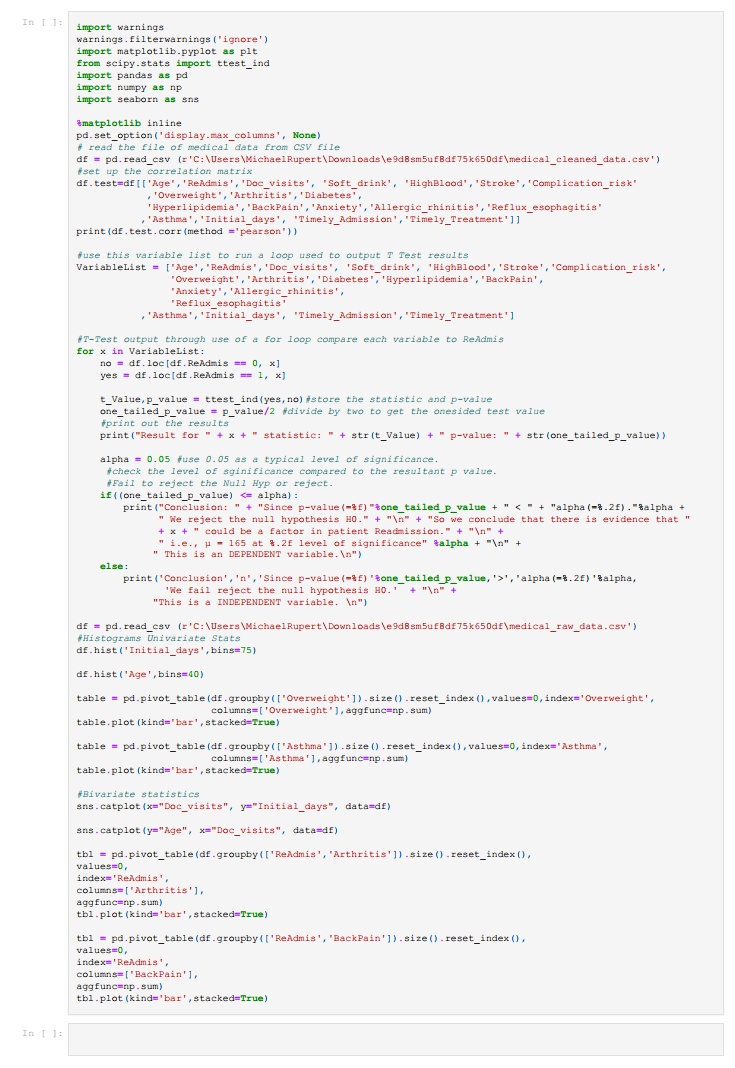
# A2: Benefits from the Analysis

By answering this question many organizational stake holders could look at the key factors of readmission and better prepare a facility, assess how risk exposed they are for readmission violations that would lead to a fine. As we explore the data, we will be able to discover what variables may influence Readmission of patients and eliminate others as less likely factors, we can assess, with a certain level of confidence, what factors institutions can focus on to reduce the rate of readmission.

# 

# A3: Identification of Variables within Dataset

Several variables are represented that can be examined to help us in our exploration of this dataset. The following variables Age, ReAdmis, Doc\_Visits, Soft\_drink, HighBlood, Stroke, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, Services , and Initial\_days, Timely\_addmission, Timely\_treatment. The data with in the dataset as been cleaned in previous assignments, missing variables have been replaced, outliers have either been removed or normalized.

B1: Code

B2: Code Output

There will be many visualizations produced by the code and will be displayed in the visualizations section.

A picture containing calendar

Description automatically generated

*Figure 1*

The above figure (Figure 1.) shows a correlation matrix output for the variables. We can see from this comparison that there is a strong correlation between Readmission and Initial\_days of the preceding hospital visit and a weak correlation between Readmission and stroke and Readmission and Doctor\_visits. The other variables range between the ones mentioned above.

As we can see in the chart below once a T-test is performed using ttest\_ind from the scipy.stats lib the statistic and pvalue outputs help us to determining the dependency of the variable. A score for the statistic being higher than the pvalue indicates a dependent variable if the reverse is true, it indicates an independent variable. Also, we can determine if a variable is dependant or independent by checking it against the alpha of 0.05 (5%) (HAYES, 2022). The lower pvlaue that is lower than the statistic and closer to zero indicates a very good probability the variable leads to readmission. Arthritis, BackPain, Initial\_days are all variables that could indicate Readmission risk with initial\_days having a pvalue of 0.0 which could indicate a very strong probability of readmission.

|  |  |  |
| --- | --- | --- |
| **Variables of Comparison** | **Dependancy** | **Observed values** |
| **ReAdmis, Age** | Independent | statistic=0.6318910887022298 pvalue=0. 26373624804223583 |
| **ReAdmis, Doc\_visits** | Independent | statistic=0.024720337146443908 pvalue=0. 4901392633656937 |
| **ReAdmis, Soft\_drinks** | Independent | statistic=0.3155456360272841  pvalue=0. 37617699680788796 |
| **ReAdmis, HighBlood** | Independent | statistic=0.23792794718833934  pvalue=0. 4059709276082407 |
| **ReAdmis, Stroke** | Independent | statistic=0.09839466538009183  pvalue=0. 4608104416125448 |
| **ReAdmis, Compilation\_risk** | Independent | statistic=-0.3034171623425427  pvalue=0. 3807891410298775 |
| **ReAdmis, Overweight** | Independent | statistic=-0.9967977184641879 pvalue=0. 15944341601952894 |
| **ReAdmis, Arthritis** | Independent | statistic=0.7487285970611021 pvalue=0. 22701920943044934 |
| **ReAdmis, Diabetes** | Independent | statistic=-0.29773426761366545 pvalue=0. 3829560869772419 |
| **ReAdmis, Hyperlipidemia** | Independent | statistic=0.4122901122227031 pvalue=0. 34006781779248507 |
| **ReAdmis, BackPain** | Independent | statistic=1.3423444567266545 pvalue=0. 08975740836796603 |
| **ReAdmis, Anxiety** | Independent | statistic=0.12286041147869296 pvalue=0. 45111005241412266 |
| **ReAdmis, Allergic\_rhinitis** | Independent | statistic=-0.4813681169156076 pvalue=0. 3151327142702056 |
| **ReAdmis, Reflux\_espohigitis** | Independent | statistic=0.5532068942326542 pvalue=0. 2900670586345776 |
| **ReAdmis, Asthma** | Dependent | statistic=-1.7050714554663726  pvalue=0. 04410608348132783 |
| **ReAdmis, Initial\_days** | Dependent | statistic=134.36783397363934 pvalue=0.0 |
| **ReAdmis, Timely\_admission** | Dependent | statistic=-1.6847491760925344 pvalue=0. 04410608348132783 |
| **ReAdmis, Timely\_Treatment** | Independent | statistic=-0.23582263059303873 pvalue=0. 4067875687361996 |

Actual Output

Result for Age statistic: 0.6318910887022298 p-value: 0.26373624804223583

Conclusion n Since p-value(=0.263736) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Doc\_visits statistic: 0.024720337146443908 p-value: 0.4901392633656937

Conclusion n Since p-value(=0.490139) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Soft\_drink statistic: 0.3155456360272841 p-value: 0.37617699680788796

Conclusion n Since p-value(=0.376177) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for HighBlood statistic: 0.23792794718833934 p-value: 0.4059709276082407

Conclusion n Since p-value(=0.405971) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Stroke statistic: 0.09839466538009183 p-value: 0.4608104416125448

Conclusion n Since p-value(=0.460810) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Complication\_risk statistic: -0.3034171623425427 p-value: 0.3807891410298775

Conclusion n Since p-value(=0.380789) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Overweight statistic: -0.9967977184641879 p-value: 0.15944341601952894

Conclusion n Since p-value(=0.159443) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Arthritis statistic: 0.7487285970611021 p-value: 0.22701920943044934

Conclusion n Since p-value(=0.227019) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Diabetes statistic: -0.29773426761366545 p-value: 0.3829560869772419

Conclusion n Since p-value(=0.382956) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Hyperlipidemia statistic: 0.4122901122227031 p-value: 0.34006781779248507

Conclusion n Since p-value(=0.340068) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for BackPain statistic: 1.3423444567266545 p-value: 0.08975740836796603

Conclusion n Since p-value(=0.089757) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Anxiety statistic: 0.12286041147869296 p-value: 0.45111005241412266

Conclusion n Since p-value(=0.451110) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Allergic\_rhinitis statistic: -0.4813681169156076 p-value: 0.3151327142702056

Conclusion n Since p-value(=0.315133) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Reflux\_esophagitis statistic: 0.5532068942326542 p-value: 0.2900670586345776

Conclusion n Since p-value(=0.290067) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

Result for Asthma statistic: -1.7050714554663726 p-value: 0.04410608348132783

Conclusion: Since p-value(=0.044106) < alpha(=0.05). We reject the null hypothesis H0.

So we conclude that there is evidence that Asthma could be a factor in patient Readmission.

i.e., μ = 165 at 0.05 level of significance

This is an DEPENDENT variable.

Result for Initial\_days statistic: 134.36783397363934 p-value: 0.0

Conclusion: Since p-value(=0.000000) < alpha(=0.05). We reject the null hypothesis H0.

So we conclude that there is evidence that Initial\_days could be a factor in patient Readmission.

i.e., μ = 165 at 0.05 level of significance

This is an DEPENDENT variable.

Result for Timely\_Admission statistic: -1.6847491760925344 p-value: 0.046034094009676355

Conclusion: Since p-value(=0.046034) < alpha(=0.05). We reject the null hypothesis H0.

So we conclude that there is evidence that Timely\_Admission could be a factor in patient Readmission.

i.e., μ = 165 at 0.05 level of significance

This is an DEPENDENT variable.

Result for Timely\_Treatment statistic: -0.23582263059303873 p-value: 0.4067875687361996

Conclusion n Since p-value(=0.406788) > alpha(=0.05) We fail reject the null hypothesis H0.

This is a INDEPENDENT variable.

We can see from the above output which p-values are above or below the 0.05 level of significance. This indicates if a variable is dependent or independent and if we reject or fail to reject the Null Hypothesis, H0. This can help use determine which variables could likely affect Readmission. In the above output the most likely variables that could affect Readmission are Asthma, Initial\_days, Timely\_Admission.

# B3: Justification of Approach

After exploring the three possible ways to test this data, ANOVA, T-test and Chi-square I have decided on the T-test since we are doing a pairwise comparison of Readmission to many other possible variables. A T-test is used to compare the means of two groups, (HAYES, 2022). Using T-test will help discover if certain variables influence Readmission and which of these has a higher probability.

# 

# C: Univariate Statistics

From the data two continuous variables can be identified, Age and Initial\_days. When a univariate analysis is performed on these variables, we can see what the distribution looks like. Since these variables can have any values within the range of finite values this makes them continuous. Figure 2 and 3 below. Univariate Statistics refer to statistical analysis that include a single dependent variable and could include one or more independent variables. (Allen, 2018)

Chart, histogram

Description automatically generated

*Figure 2*

*Chart, histogram

Description automatically generated*

*Figure 3*

Categorical variables, Overweight and Asthma, since they are either the category of yes or no are easy to render the distributions through visualization. This can be seen in Figure 4 and 5.

Bar chart

Description automatically generated

*Figure 4*

Logo, bar chart

Description automatically generated

*Figure 5*

# D: Bivariate

Bivariate data is when you are studying two variables. (Glen, 2015) If we take the continuous variables mentioned above beginning with Initial\_days and Age comparing them with Doc\_visits this will give a bivariate comparison of continuous variables. Figure 5 and 6 shows the seaborn categorical plot of the bivariate visualization.

Chart

Description automatically generated with low confidence

*Figure 5*

A picture containing text, screenshot, vector graphics

Description automatically generated

*Figure 6*

For the bivariate analysis of the categorical variables identified are BackPain and Arthritis which can be compared to ReAdmis. Below we can see their visualizations in Figure 7 and 8.

Chart, bar chart

Description automatically generated

*Figure 7*

Chart, bar chart

Description automatically generated

*Figure 8*

# E1: Results of Analysis

The results of this analysis allow us to reject the null hypothesis for certain variables, claiming that the variables Arthritis, BackPain, Initial\_days can increase the probability of readmission. Patients could have a higher risk of being if suffering from arthritis and/or back pain. The initial number of days spent at the preceding visit to the Readmit could also be a factor that leads to the readmission.

# E2: Limitations

We have no direct evidence from the hypothesis that these readmits are caused by the variables, we are limited to make the claim and only state there is a higher probability with patients in the categories. The data and analysis show that there are lesser correlations to other variables. There could also be unrecorded variables that could play a factor or bare out more information if collected and examined. Some assumptions when doing a T-Test are considered.

* The scale of measurement. The assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale, such as the scores for an IQ test. (HAYES, 2022)
* random sampling. The data is collected from a representative, randomly selected portion of the total population. (HAYES, 2022)
* data is normally distributed. (Bevens, 2020)
* the two population have the same variance. This can be adjusted though. (Bevens, 2020)

# E3: Course of Action

We can indeed build a model that potentially helps to reduce readmissions. Possible benefits by working to change the number of days a patient has spent at the facility on their initial visit could help facilities lower their risk of fine due to violation. Since medical staff tend to be focused on healing and healing takes time it might cause them to miss this opportunity to reduce readmits and discharge a well patient sooner. Patients suffer from asthma or arthritis might require specialized types of care. Additionally, it is recommended that more information is gathered, cause of initial admission, what was the metric used to determine release, where they considered fully healed, was there follow ups recommended, outpatient appointments. All these could point to unknown factors not represented in the data.

# G: Sources for Third Party Code

Using TTest\_ind

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html>

Using seaborn for Continuous data representations.

<https://seaborn.pydata.org/tutorial/distributions.html>

Accessing pandas dataframe columns, rows, and cells. Retrieved March 18, 2020, from <https://pythonhow.com/accessing-dataframe-columns-rows-and-cells/>

Bezden, V. (2018, October 5). How to set seaborn plot size in Jupyter Notebook. Retrieved March 18, 2020, from [https://medium.com/@vladbezden/how-to-set-seaborn-plot-size-injupyter-notebook-63ffb141543](https://medium.com/@vladbezden/how-to-set-seaborn-plot-size-in%20jupyter-notebook-63ffb141543)

# H: Sources

The use of T-Test why it is logical to use

Geeks for Geeks using Correlation in Panads

<https://www.geeksforgeeks.org/python-pandas-dataframe-corr/>

Allen, M. (2018, 12 19). *Univariate Statistics*. doi: https://dx.doi.org/10.4135/9781483381411.n646

Bevens, R. (2020, January 31). *An Introduction to T-Tests | Definitions, Formula and Examples*. Retrieved from Scribbr: https://www.scribbr.com/statistics/t-test/#:~:text=A%20t%2Dtest%20is%20a,are%20different%20from%20one%20another.

Glen, S. (2015, 7 9). *Elementary Statistics for the rest of us!* Retrieved from StatisticsHowTo: https://www.statisticshowto.com/bivariate-analysis/

HAYES, A. (2022, March 12). *T-Test*. Retrieved from Investopedia: https://www.investopedia.com/terms/t/t-test.asp