­

**Performance Assessment for D207: Exploratory Data Analysis**

Christopher Fischer - 011933891

College of Information Technology, Western Governors University

April 20, 2024

Performance Assessment for D207: Exploratory Data Analysis

This document contains the tasks and outputs required for the “OEM2 TASK 1: EDA: EXPLORATORY DATA ANALYSIS” assessment. All work is original to the author unless otherwise indicated by a citation.

# Problem Statement (A1 – A3)

The company’s stakeholders, both internal and external, are disappointed with the rising cost of readmittance penalties received from the Centers for Medicare and Medicaid Services (CMMS) over the last several years. While it is obvious that significant efforts need to be made to reduce the number of readmittances our patients experience, without knowing more about which patients are more likely to be readmitted, no reasonable plan can be formulated. Management wants to know *if there are certain variables which are associated with more likely readmittance*. Having this information will allow teams to zero in on the patient populations with the highest likelihood of readmittance first.

To complete this study, the team will explore the data from the **medical\_clean.csv** file provided as input. In specific, the study will focus on **ReAdmis**, a Boolean (yes/no) variable indicating if a patient was readmitted to our facility. The following categorical values will be tested against **ReAdmis** to determine if any show a strong association. These variables are:

| Variable | Description |
| --- | --- |
| HighBlood | Patient has high blood pressure (yes/no) |
| Stroke | Patient has a history of stroke (yes/no) |
| Overweight | Patient is statistically overweight (yes/no) |
| Arthritis | Patient has a history of arthritis (yes/no) |
| Diabetes | Patient has a history of diabetes mellitus (yes/no) |
| Hyperlipidemia | Patient has a history of high cholesterol (yes/no) |
| BackPain | Patient has a history of chronic back pain (yes/no) |
| Anxiety | Patient has a history of an anxiety disorder (yes/no) |
| Allergic\_rhinitis | Patient has a history of runny nose due to allergies (yes/no) |
| Asthma | Patient has a history of asthma (yes/no) |
| Soft\_drink | Patient reports consuming three or more sodas per day (yes/no) |
| Item1 | Patient response to the survey question on timely admission (1-8) |
| Item2 | Patient response to the survey question on timely treatment (1-8) |
| Item3 | Patient response to the survey question on timely visits (1-8) |
| Item4 | Patient response to the survey question on reliability (1-8) |
| Item5 | Patient response to the survey question on options (1-8) |
| Item6 | Patient response to the survey question on hours of treatment (1-8) |
| Item7 | Patient response to the survey question on staff courtesy (1-8) |
| Item8 | Patient response to the survey question on active listening (1-8) |
| State | State of residence reports on patient’s billing records |
| Gender | Patient’s self-reported gender identity (male, female, non-binary) |
| Marital | Patient’s self-reported marital status (Married, Divorced, Separated, Widowed, Never Married) |

Table - List of Study Independent Variables

# Data Analysis (B1 – B3)

To assess whether certain variable have a greater association with being readmitted, a **ꭓ2 Test of Independence** will be performed using each of the independent variables shown in table 1 above against **ReAdmis** (the dependent variable). The **ꭓ2** test is well-suited for analyzing two categorical variables (Sewell, 2021). As a non-parametric test, **ꭓ2** does not require the variables to conform to a normal distribution (Turney, 2023). The **ꭓ2** test requires a crosstab of the study variables (Sewell, 2021). One important limitation of the test is that each cell in the table must contain a value >= 5 (Learn Statistics Easily, 2023). During the study, variables for which one or more cells of the contingency table had a value under 5 were excluded from further consideration.

The following graphic shows the contingency table for each variable excluded from the study due to failing to meet the “minimum 5” rule.

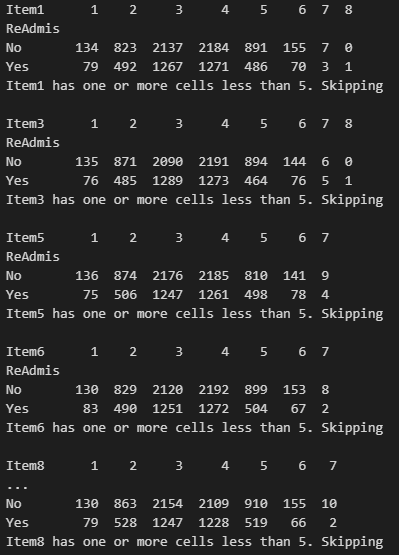


Figure - Variables Excluded from X2 Test

# Univariate Statistics (C)

## Continuous Variables

Two continuous variables were explored to determine the distribution of values.

### Income

The observations in the variable **Income** exhibit the following summary statistics:

| Statistic | Value |
| --- | --- |
| Count | 10,000 |
| Mean | $40,409.50 |
| Standard Deviation | $28,251.15 |
| Minimum | $154.08 |
| 1st Quartile | $19,598.78 |
| 2nd Quartile | $33,768.42 |
| 3rd Quartile | $54296.40 |
| Maximum | $207,249.10 |

Table - Univariate Statistics for Income

A graph showing the amount of income

Description automatically generated

Figure - Histogram and Density Plot of Income

The variable **Income** shows a strong right skew with most observations being under $50,000.

### Vitamin D Levels

The observations in the variable **VitD\_levels** exhibit the following summary statistics:

| Statistic | Value |
| --- | --- |
| Count | 10,000 |
| Mean | 17.96 ng/mL |
| Standard Deviation | 2.02 ng/mL |
| Minimum | 9.81 ng/mL |
| 1st Quartile | 16.63 ng/mL |
| 2nd Quartile | 17.95 ng/mL |
| 3rd Quartile | 19.35 ng/mL |
| Maximum | 26.39 ng/mL |

Table - Univariate Statistics for VitD\_levels

A chart of vitamin d levels

Description automatically generated

Figure - Boxplot of VitD\_levels

The boxplot for **VitD\_levels** reveals a likely standard distribution of values with a comparably large number of outliers at both ends of the scale. This should be confirmed with a histogram or density plot.

## Categorical Variables

Two categorical variables were explored to determine the distribution of values.

### Marital Status

The observations of **Marital** are grouped as follows:

| Value | Count |
| --- | --- |
| Widowed | 2,045 |
| Married | 2,023 |
| Separated | 1,987 |
| Never Married | 1,984 |
| Divorced | 1,961 |

Table - Value Counts for Marital

A colorful circle with text

Description automatically generated

Figure – Donut Chart of Marital Distribution

### Intake Method

The observations of **Initial\_admin** are grouped as follows:

| Value | Count |
| --- | --- |
| Emergency | 5,060 |
| Elective | 2,504 |
| Observation | 2,436 |

Table - Value Counts for Initial\_admin

A graph with blue bars

Description automatically generated

Figure - Frequency Chart of Initial\_admin

# Bivariate Statistics (D)

## Continuous Variables

The continuous variables **Income** and **TotalCharge** were compared with *Pearson’s r* and *Spearman’s r* to test for correlation. The results indicate no correlation between these variables:

| Statistic | Result |
| --- | --- |
| Pearson’s r | -0.014 |
| Spearman’s r | -0.013 |

Figure - Correlation Tests of Income vs. TotalCharge

A graph with blue dots and a red line

Description automatically generated

Figure - Scatter Plot of Income vs TotalCharge with Regression

## Categorical Variables

The categorical variables **Gender** and **Readmis** were plotted against each other with a stacked histogram:

A graph with blue and orange squares

Description automatically generated

Figure - Readmis by Gender Histogram

# Summary of Findings (E1 – E3)

This study attempted to find one or more categorical variables with a statistical correlation to a patient being readmitted. The goal of the study was to identify patient populations upon which to focus care plan improvements to reduce the likelihood of readmission, thereby reducing the company’s total penalties.

Using a standard alpha value of 0.05, none of the studied variables demonstrated statistically significant correlation. The results of the **ꭓ2 Test of Independence** for each studied variable can be seen in the following table and graph:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **X**2 | **p-value** | **Dof** | **Accept H0**0 |
| HighBlood | 0.0424 | 0.8369 | 1 | Yes |
| Stroke | 0.0043 | 0.9475 | 1 | Yes |
| Overweight | 0.6985 | 0.4033 | 1 | Yes |
| Arthritis | 0.5545 | 0.4565 | 1 | Yes |
| Diabetes | 0.0798 | 0.7775 | 1 | Yes |
| Hyperlipidemia | 0.1671 | 0.6827 | 1 | Yes |
| BackPain | 1.7166 | 0.1901 | 1 | Yes |
| Anxiety | 0.0477 | 0.8271 | 1 | Yes |
| Allergic \_rhinitis | 0.1970 | 0.6572 | 1 | Yes |
| Asthma | 2.8575 | 0.0910 | 1 | Yes |
| Soft drink | 0.5573 | 0.4553 | 1 | Yes |
| State | 46.7579 | 0.6427 | 51 | Yes |
| Gender | 1.5858 | 0.4525 | 2 | Yes |
| Marital | 5.0852 | 0.2787 | 4 | Yes |
| Item2 | 9.1291 | 0.1664 | 6 | Yes |
| Item4 | 5.0076 | 0.5428 | 6 | Yes |
| Item7 | 6.2897 | 0.3915 | 6 | Yes |

Table - X2 Results for Studied Variables vs. Readmis

A graph with blue dots

Description automatically generated

Figure - P-values vs. Alpha

One limitation of the study performed is that several customer satisfaction survey questions were excluded because of having insufficient (< 5) frequency counts in one or more crosstab cells. Upon further evaluation, a better way to analyze an association between the survey responses and readmittance would be to compute a mean survey score for each patient. Then, use a t-test to compare the mean survey score against **ReAdmis**. Another possibility would be to bucket the survey responses into fewer values such as Poor, Average, Excellent and repeat the ꭓ2 tests.

Unfortunately, this study does not reveal any clear course of action to reduce the amount of readmission penalties paid. The team recommends gathering additional clinical data for further study to answer the problem statement expressed at the outset of this report.

# Recorded Code Review (F)

A recording of the code review presentation was uploaded with this submission. For quick reference, that video may be found here: [Panopto Recording](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3c057a27-f453-4362-9c0a-b163013dbc37)

# Third-Party Code References (G)

Bobbitt, Z. (2020, 08 13). *How to Create a Scatterplot with a Regression Line in Python*. Retrieved 04 29, 2024, from Statology: https://www.statology.org/scatterplot-with-regression-line-python/

Collado, S. D. (2022, 03 17). *Show both value and percentage on a pie chart*. Retrieved 04 29, 2024, from Stack Overflow: https://stackoverflow.com/questions/59644751/show-both-value-and-percentage-on-a-pie-chart

DataScienceTribe Staff. (2023, 10 07). *How to Indicate if Any of Multiple Columns Have Values Greater Than 0 in Pandas*. Retrieved 04 27, 2024, from DataScienceTribe: https://www.datasciencebyexample.com/2023/10/07/fast-way-to-create-column-to-check-postive-value-in-multiple-columns-in-pandas/

Holtz, Y. (2021, 03 05). *Custom donut plot using matplotlib*. Retrieved 04 29, 2024, from Python Graph Gallery: https://python-graph-gallery.com/161-custom-matplotlib-donut-plot/

Landup, D. (2023, 09 20). *How to Set Axis Range (xlim, ylim) in Matplotlib*. Retrieved 04 27, 2024, from StaclAbuse: https://stackabuse.com/how-to-set-axis-range-xlim-ylim-in-matplotlib/

Little, S. (2016, 06 08). *Rotate axis tick labels*. Retrieved 04 27, 2024, from Stack Overflow: https://stackoverflow.com/questions/10998621/rotate-axis-tick-labels

Martin, M. (2021, 02 21). *Plot Two Categorical Variables*. Retrieved 04 29, 2024, from Stack Exchange: https://datascience.stackexchange.com/questions/89692/plot-two-categorical-variables

Morwenn. (2023, 02 25). *Add a label to y-axis to show the value of y for a horizontal line in matplotlib*. Retrieved 04 27, 2024, from Stack Overflow: https://stackoverflow.com/questions/42877747/add-a-label-to-y-axis-to-show-the-value-of-y-for-a-horizontal-line-in-matplotlib

Ray, S. (2020). *Show all colums of a pandas-dataframe in ".describe()"*. Retrieved 04 08, 2024, from Stack Overflow: https://stackoverflow.com/questions/64455605/show-all-colums-of-a-pandas-dataframe-in-describe

# Referenced Works (H)

# Learn Statistics Easily. (2023, 06 15). Mastering the Chi-Square Test: A Comprehensive Guide. Retrieved 05 01, 2024, from Learn Statistics Easily: https://statisticseasily.com/chi-square-test/

Sewell, D. W. (2021, 02 13). *D207 Exploratory Data Analysis - Chi Square.* Retrieved from WGU Sharepoint: https://westerngovernorsuniversity-my.sharepoint.com/:p:/g/personal/william\_sewell\_wgu\_edu/Eei\_h7iUH6pKnTcyvdymHb0BBxrkKtOmu1vKlU5Xb-GpEQ?e=55SJzh&previoussessionid=c8b9d525-85ff-eeb7-e3b2-3e52edfae6c8

Turney, S. (2023, 06 22). *Chi-Square (Χ²) Tests | Types, Formula & Examples*. Retrieved 05 01, 2024, from Scribbr: https://www.scribbr.com/statistics/chi-square-tests/