Eric Yarger

Chi-Square Analysis and Findings

A1 Question for Analysis

In the medical industry readmission rates are a major cause of concern. The question this paper will address is: what categorical patient medical condition variables from the dataset show a strong correlation with readmission? Chi-Square analysis will be conducted for each chosen variable paired in conjunction with ReAdmis, the variable for readmission.

A2 Benefit From Analysis

The benefits of reducing hospital readmission rates are rewarding for hospital leaders, administrators, staff, and patients. A meta analysis of 30 studies published in the Journal of the American Medical Association looking to predict the risk of hospital readmission concluded that a median of 27% of hospital readmissions are preventable (Kansagara, et. al., 2011). These preventable readmissions provide our organization, and hospital organizations nationwide, with a rewarding and profitable challenge to solve.

This data analysis, conducted to find correlation between Yes/No responses to medical conditions in the provided dataset correlated with readmission rates, is highly beneficial to the stakeholders of our organization. From a financial perspective the organization can use the results of this analysis to lower readmission rates. This lowers the chance of accruing penalties for excessive readmissions. This also has the benefit of displaying a strong positive image of leadership to other corporations in the healthcare community. This ups the chances that top leaders, physicians, and staff will want to work in our organization as compared to other hospitals.

From an organizational perspective, the hospital's medical staff will have a useful indicator that a patient potentially needs more attention when certain medical conditions are answered with a 'Yes' to prevent readmission. This is a positive for patients who won't have to come back. It's also a positive to the doctors and nurses who will now have more time to treat new patients. Reducing readmission rates is also a good way of improving the hospital's overall customer satisfaction metrics, which will contribute to the organization's value and reputation for providing outstanding medical treatment.

A3 Data Identification

The specific variables analyzed to answer this paper's central question are organized below. These variables, as per the PDF included with the data download, are categorized as patient medical conditions. These variables were chosen for their relevance at providing a quick, simple, and effective means of gathering specific information as to why the patient was potentially at the hospital initially. The variables are all categorical with the choice of answering Yes or No. During analysis, the variable 'ReAdmis' will be paired with the other variables on the list to determine their dependency and readmission predictive capacity. The provided dataset had no missing values for the variables. The table below lists the variables used in this analysis.

Variable Name	Categorical or Continuous	Example of cell entry
ReAdmis	Categorical	Yes
HighBlood	Categorical	No
Stroke	Categorical	No
Asthma	Categorical	Yes
BackPain	Categorical	Yes
Overweight	Categorical	No
Arthritis	Categorical	Yes
Diabetes	Categorical	No
Hyperlipidemia	Categorical	Yes
Anxiety	Categorical	No
Allergic_rhinitis	Categorical	Yes
Reflux_esophagitis	Categorical	No

B1 Code

Attached to this submission is a PDF of my code for this section titled 'D207AandB'. I use a Laptop running Windows 10. The IDE used was JupyterLab, using Python3 as the language. Libraries and tools used for analysis are shown in the first cell of the JupyterLab notebook. The code is warning and error-free after

being run, and analyzes the chosen variables from the dataset using Chi-Square Test of Independence analysis.

B2 Output

Attached to this submission is a PDF of my JupyterLab notebook with all code used titled 'D207AandB'. Coded inputs with generated outputs are labeled and organized in order of execution.

B3 Justification

Chi-Square Test of Independence was chosen for use in this analysis. Justification for choosing this method over T-Test or ANOVA is because the variables to analyze were all categorical. Analyzing categorical by categorical variables is what Chi-Square analysis excels in.

In contrast, T-Tests require one independent categorical variable with two groups, one continuous variable, and the dataset cases to be independent of each other. ANOVA, used for multivariate (3+) analysis, also does not fill the requirement of being used for purely categorical variables. ANOVA also wasn't suitable because this analysis was performed on 2 variables at a time. Neither of these analytic techniques would have been suitable for the variables chosen to answer the PA's question.

Chi-Square Test of Independence results determine if there's an association/relation between two categorical variables. Each test variable was crosstabbed into a 2X2 matrix with 'ReAdmis' and the variable tested. The test statistic and p-value were calculated and analyzed to statistically determine if the variable showed dependence with variable 'ReAdmis'.

Assumptions of the Chi-Square test include (McHugh, 2013):

- The data in cells should be frequencies or counts of cases, not percentages or other transformations of the data.
- The categories of the variable are mutually exclusive.
- There are 2 variables, and both are measured as categories.

The variables chosen for analysis using Chi-Square are all applicable under these assumptions.

C Univariate Statistics and C1 Visual of Findings

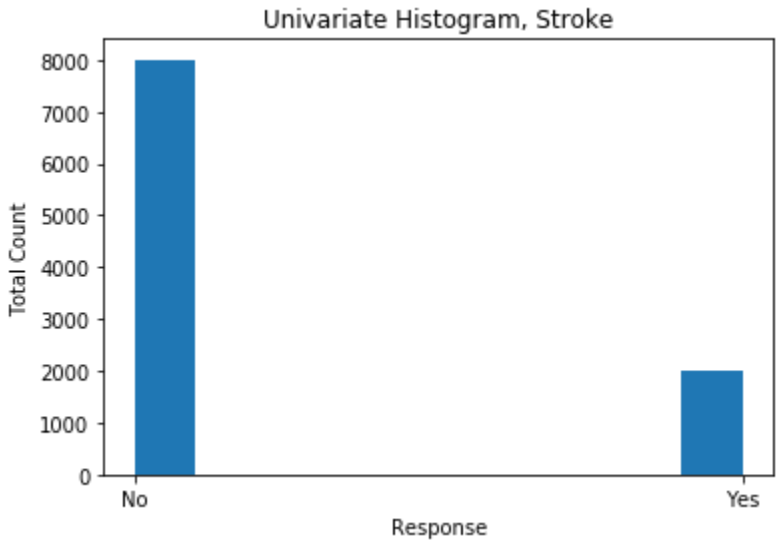
Two categorical variables, Stroke and Initial_admin, and two continuous variables, Income and TotalCharge, were identified and analyzed using univariate analysis. Included in the attached PDF titled 'D207CandD' to this submission is the JupyterLab notebook with all libraries and code used for this analysis. Below is the overview of this analysis for each variable.

Univariate categorical Variable #1: Stroke

Categorical variable Stroke was analyzed by creating a Frequency Table and plotting a Histogram.

Frequency Table: **Variable: Stroke, datatype: int64**

Response	Count
No	8007
Yes	1993



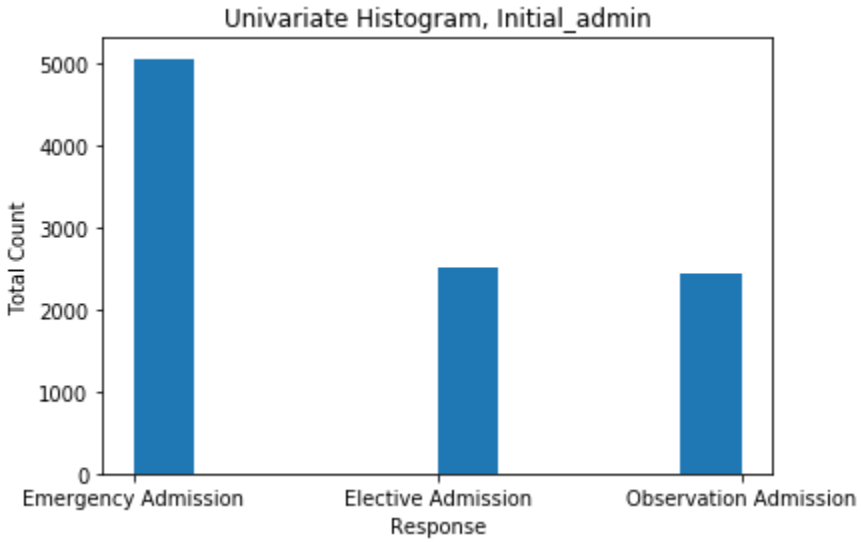
Univariate categorical variable #2: Initial_admin

Categorical variable Initial_admin was analyzed by creating a Frequency Table and plotting a Histogram.

Frequency Table: **Variable: Initial_admin, datatype: int64**

Response	Count
Emergency Admission	5060
Elective Admission	2504

Observation Admission	2436

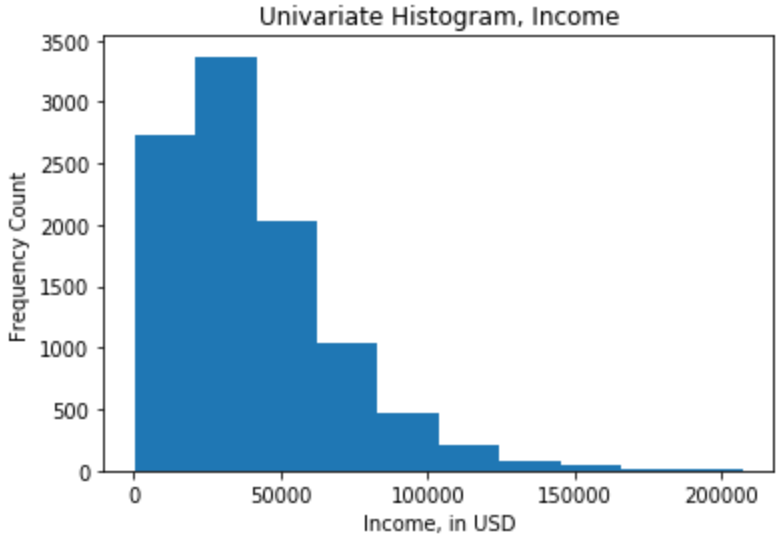


Univariate continuous variable #1: Income

Continuous variable Income was analyzed by calculating the Summary Statistics, which measures the center and spread of values in the variable. Visualization was achieved by plotting a Histogram.

Summary Statistics: **Variable Income, datatype: float64**

Summary statistic	**Value**
Mean	40490.50
Median	33768.42
Standard Deviation	28521.15

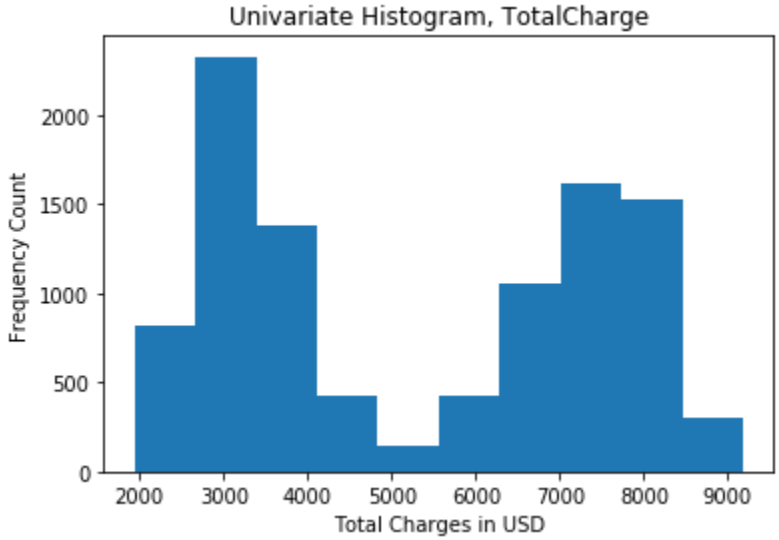


Univariate continuous variable #2: TotalCharge

Continuous variable TotalCharge was analyzed by calculating the Summary Statistics, which measures the center and spread of values in the variable. Visualization was achieved by plotting a Histogram.

Summary Statistics: **Variable TotalCharge, datatype: float64**

Summary statistic	Value
Mean	5312.17
Median	5213.95
Standard Deviation	2180.39

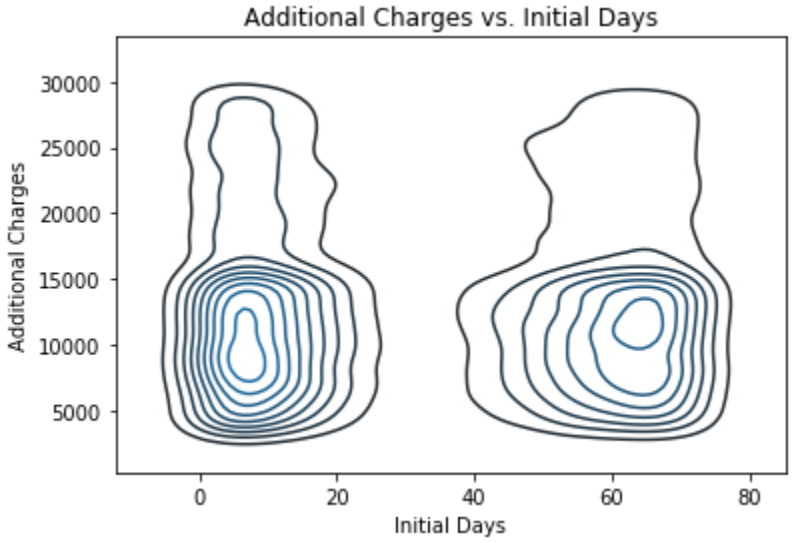


D Bivariate Statistics and D1 Visual of Findings

Two categorical variables, ReAdmis and Arthritis, and two continuous variables, Additional_charges and Initial_days, were analyzed and visualized using bivariate statistics. Included in the attached PDF titled 'D207CandD' to this submission is the JupyterLab notebook with all libraries and code used for this analysis. Below is the overview of this analysis for each variable.

Continuous variables #1 and #2

For continuous variables Additional_charges and Initial_days, Bivariate statistical analysis was conducted using Regression. The variables were visualized using a Kernal Density Estimate (KDE) graph. Included in the attached PDF titled 'D207CandD' to this submission is the JupyterLab notebook with all code and libraries used to perform this analysis. Scipy.stats linregress function was used to calculate R-Squared. The variables were found to not be dependent, with a P value of .6593 and R-Square of 1.9438 e -05. An R-Square value so low reveals that almost none of variability observed in the target variable is explained by the regression model.



The KDE graph above shows the density of Additional Charges measured in USD on the Y-Axis and the number of Initial Days on the X-axis. The center of each Kernel is roughly at the 10000 dollar level. If there was a strong correlation between these two variables the chart above would have the Kernel on the right showing a much higher average Additional Charge.

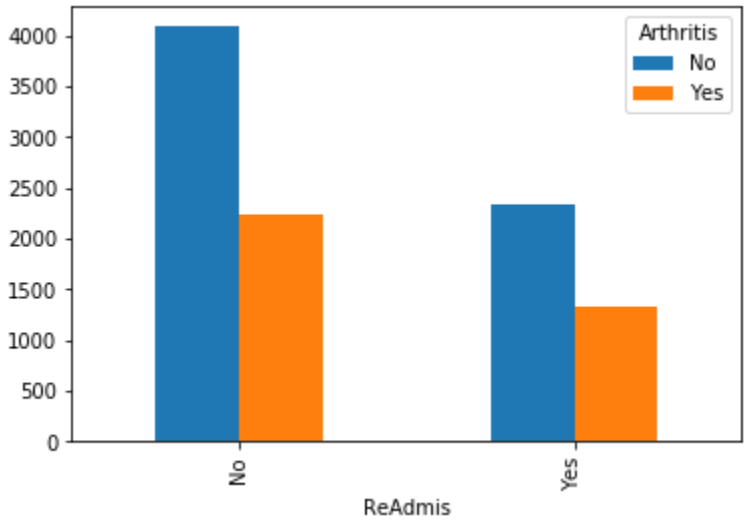
As it is, the number of Initial_days doesn't appear to be relevant at all to Additional Charges. This is what we would expect given such a low R-squared value returned from our Regression Analysis. The visual evidence provides visual validation of the statistical analysis' outcome.

Categorical variables #1 and #2

First, for categorical variables ReAdmis and Arthritis a crosstab chart was executed. To detail the result, there were four potential independent response categories. There are:

- Patient was not Readmitted and answered No on Arthritis variable
- Patient was Readmitted and answered No on Arthritis variable
- Patient was not Readmitted and answered Yes on Arthritis variable
- Patient was Readmitted and answered Yes on Arthritis variable

Arthritis	**No**	**Yes**
ReAdmis		
No	4086	2245
Yes	2340	1329



The histogram above visualizes the Bivariate analysis of ReAdmis and Arthritis. The X-axis is the Yes/No answer to ReAdmis, the Y-axis is the total count of responses. Arthritis is represented by the bars: Orange bars for a Yes response, Blue bars for a No response.

Chi-square Test of Independence was performed on these two variables to measure for dependency. The P-value was calculated to be .4565, which determines that we accept the Null Hypothesis at a required

level of confidence of 90%. Please refer to attached PDF 'D207CandD' for complete code input and output for this analysis.

E1 Results of Analysis

The results of performing Chi-Square Test of Independence on the chosen variables is outlined below. The P-Value and Test Statistic have been rounded to four places after the decimal.

Variable	Action (LOC 90%)	P-Value	Test Statistic	Degree of Freedom
HighBlood	Fail to reject null hypothesis	.8369	.0424	1
Stroke	Fail to reject null hypothesis	.9475	.0043	1
Asthma	Reject null hypothesis	.0910	2.8575	1
BackPain	Fail to reject null hypothesis	.1901	1.7166	1
Overweight	Fail to reject null hypothesis	.4033	.6985	1
Arthritis	Fail to reject null hypothesis	.4565	.5545	1
Diabetes	Fail to reject null hypothesis	.7775	.0798	1
Hyperlipidemia	Fail to reject null hypothesis	.6827	.1671	1
Anxiety	Fail to reject null hypothesis	.8271	.0477	1
Allergic_rhinitis	Fail to reject null hypothesis	.6572	.1970	1
Reflux_esophagitis	Fail to reject null hypothesis	.6022	.2716	1

The variables were analyzed to determine their level of dependence and correlation to the variable 'ReAdmis'. A chi-square test of independence was performed for each variable grouping, resulting in the p-value and test statistic noted above. The level of confidence (LOC) set for accepting or rejecting the null hypothesis was 90% for each variable analysis.

At this LOC, the results of this analysis show that only one variable, 'Asthma''s p-value correlated strongly enough with variable 'ReAdmis' to reject the null hypothesis. All other variables tested for correlation with variable 'ReAdmis' failed to reject the null hypothesis.

The test statistic is a measurement of how the observed data compares to the expected frequencies for the data. The degrees of freedom for this test are calculated as: df = (number of rows - 1)*(number of columns -1) (Turney, 2022). For the 2X2 matrix for each analysis, this calculates the degrees of freedom to be 1.

E2 Limitations of Analysis

One of the main limitations of this analysis stems from the analysis method chosen. Chi-square tests require that both variables are categorical. This means that any continuous variables in the dataset, such as VitD_levels, TotalCharge, or Additional_charges are not suitable variables for analysis. Relying only on one analytical method leaves to chance that other variables, that can potentially have great value in reducing readmission, are left unanalyzed.

Another limitation lies in the level of confidence chosen to accept or reject the null for each variable. If we set the level too high, for instance at 99.999 level of confidence, then only the most predictive dependant variables will pass scrutiny. This can lead to financially and medically valuable insights getting screened out. If we set the level too low, for instance at 80%, then potentially unrelated variables can pass as related and predictive. This can lead to the organizations' leaders focusing on the wrong variables to keep track of - potentially wasting time and money, and possibly causing negative outcomes to the organizations patients, employees, and stakeholders.

E3 Recommend Course of Action

The question posed for analysis was, "What categorical patient medical condition variables from the dataset show a strong correlation with readmission?". After analysis the answer is that out of all the categorical patient medical condition variables tested in this analysis, only variable 'Asthma' showed a strong correlation with readmission rate with a level of confidence of 90.9%. All other tested variables did not have a strong enough dependency with 'ReAdmin' to prove to be statistically significant in their predictive capacity at a LOC of 90%.

Specific actions recommended that can be taken by our organization in response to these findings are:

- Individuals that answer the question that they have Asthma could have a specific treatment plan in place for at-home treatment before discharge from the hospital to lower the chance of being readmitted.
- The general treatment of patients with Asthma at the hospital can be reevaluated to look for better operating procedures for in-hospital treatment.
- Medications that can possibly alleviate future Asthma attacks more effectively can be prescribed after patients are discharged.

- The patient questionnaire can be expanded to include more categorical medical variables. These Yes/No medical questions are fast and easy for patients to respond to, and have the potential to incrementally inform our organization how to treat individuals with specific ailments more effectively. These answers are easy to collect, low cost, and can be added to or deleted as we find more meaningful and predictive variables.

F Video

Please see attached Panopto video for information regarding the programming environment, tools used, and code executed.

G Sources For Third-Party Code

All third-party code sources used are referenced in the two attached PDFs, 'D207AandB' and 'D207CandD', that show all of my code inputs, outputs, and comments.

pandas.crosstab. (n.d). pandas. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.crosstab.html

Sewell, William. *D207 Exploratory Data Analysis - Episode 5 featuring Chi-Square [slide 61].* College of IT. WGU. https://westerngovernorsuniversity-my.sharepoint.com/:p:/g/personal/william_sewell_wgu_edu/EbIWOqJp0oJFrByVqCO5wfgB7LHdWLxGgzYyHXd2nuIV2Q?e=PrgdzD.

seaborn.kdeplot. N.d. seaborn. https://seaborn.pydata.org/generated/seaborn.kdeplot.html.

H Sources

All sources used for this submission are cited in-text in this document.

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). *Risk prediction models for hospital readmission: a systematic review*. *JAMA*, *306*(15), 1688–1698. https://doi.org/10.1001/jama.2011.1515

McHugh M. L. (2013). *The chi-square test of independence*. *Biochemia medica*, *23*(2), 143–149. https://doi.org/10.11613/bm.2013.018

Turney, Shawn. (May 30, 2022). *Chi-Square Test of Independence | Formula, Guide & Examples.* Scribbr. https://www.scribbr.com/statistics/chi-square-test-of-independence/