Multiple Regression for Predictive Modeling – D208

Gooden, Nina S. Gooden [ID #: 009823504]

Dr. Eric Straw

Multiple Regression for Predictive Modeling – D208

This analysis will explore the medical data of a potentially real-world organization. In creating this analysis, I will present an evaluation of the provided Medical Data and Dictionary Files to provide data-driven interpretations that will benefit the investigation into a chain of hospital’s readmission problem.

## **A. Describe the Purpose of this Data Analysis:**

1. Research Question
   1. Is there a correlation between Initial\_days and customer (patient) responses?
2. Data Analysis Goals and Objective
   1. The goal for this evaluation is to further explore variables that were identified as having correlation with readmission. We will be exploring independent variables to determine fit for use with linear regression.

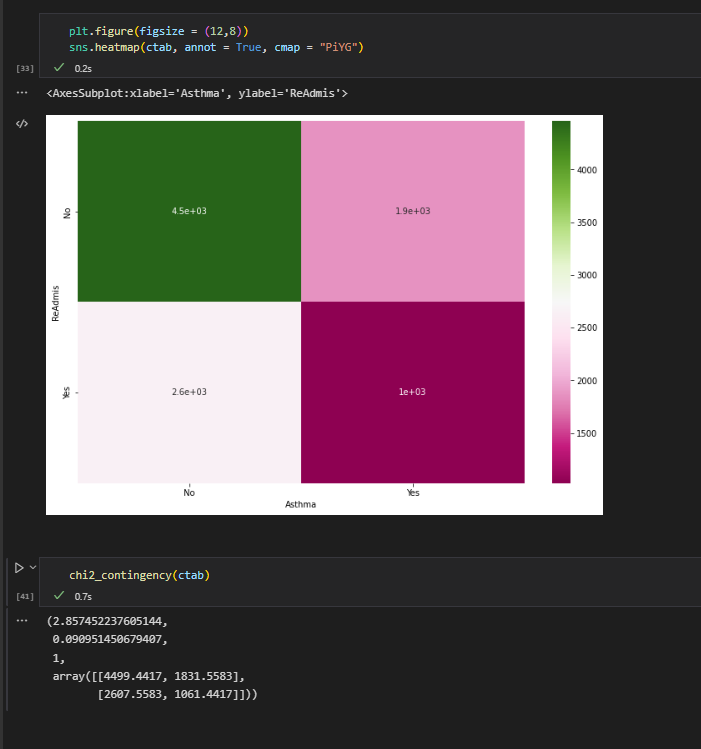
## **B. Justification for Method**

1. Multiple Regression Model Assumptions:
   1. Linear relationship: a relationship exists between the independent and dependent variables that is linear.
      1. Proven with scatterplots.
   2. Multicollinearity: which means independent variables are none of the independent variables are highly correlated with one another.
      1. Proven with a correlation matrix or VIF.
   3. Independence of observations.
      1. Tested with a Durbin-Watson test.
   4. Multivariate normality: that means the residuals—after the model is created—are normally distributed.
      1. Proven with histogram or QQ plot.
   5. Homoscedasticity: which means there is no obvious patten in the distribution of the residual vs. predicted data.
2. Benefits of Tools:
   1. I’ve opted to use Python for my analysis as this is the language, I am most comfortable with. Python offers packages and libraries that make visualizations and analysis easy and straightforward. I will be using pandas, SciPy, and statsmodels in this evaluation, all of which are uniquely designed for data analysis.
   2. Additionally, Python has functions that are uniquely designed for regression, which I will be utilizing in my assessment.
3. Why Multiple Regression is Appropriate:
   1. Since Initial\_days was previously identified as having an impact on readmission, there is merit to further exploring the variable. Due to the fact that the information I will be exploring is largely numerical or convertible, evaluation is a good fit for regression techniques. This analysis allows for exploring the correlation between Initial\_days and the patient responses. The information gleaned from this analysis will allow the hospital chain to reduce readmission likelihood, based on correlation between patient care satisfaction and how long the patient stayed during initial visit.

## **C. Data Preperation Process**

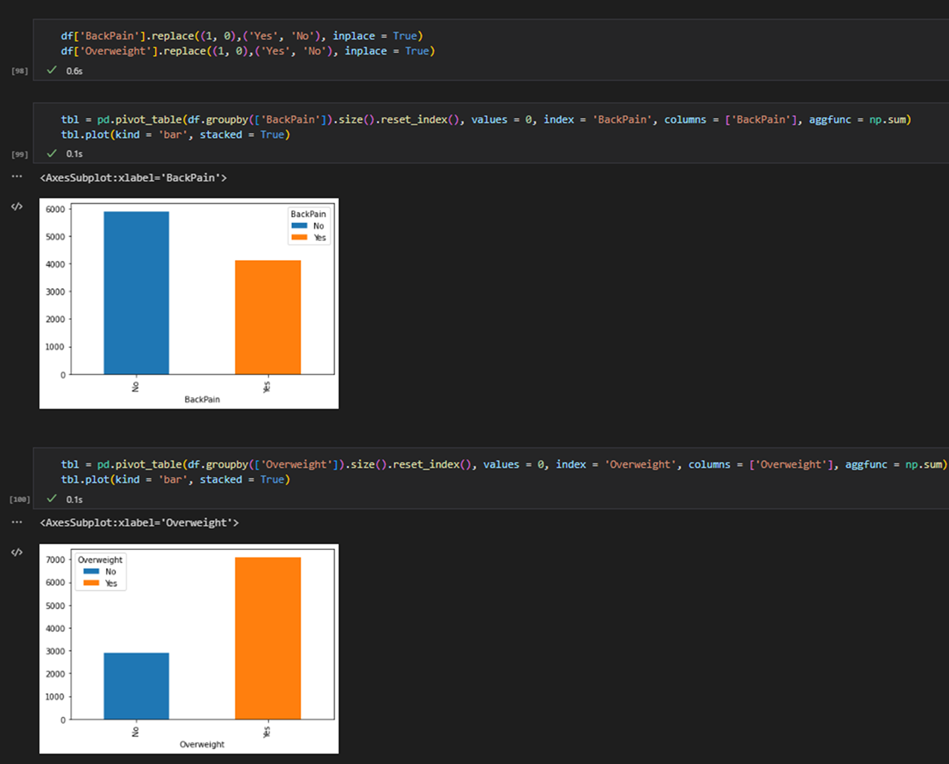
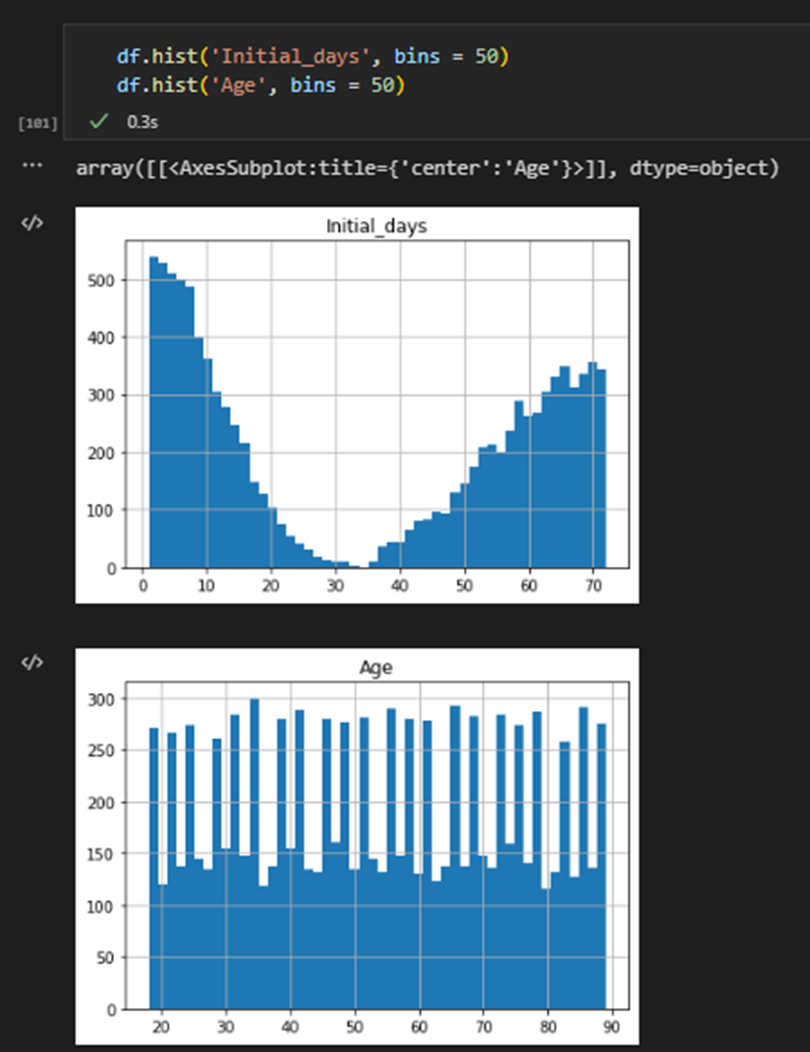
1. Data Preparation Goals
   1. A clean, workable data set is necessary in order to run an analysis that is not skewed. Columns must be checked for missing and null values, which will need to be removed or replaced with statistically valuable measures.
   2. Additionally, patient response columns are necessary, in addition to the dependent variable of Initial\_days. Data should be stripped to only include columns necessary for the valuation.
2. Summary Statistics:
   1. Next, I normalized the data into a second contingency table. A screenshot of a computer

      Description automatically generated with medium confidence
3. Output and Results
   1. In order to better understand what the contingency table was telling me, I visualized the output with a heatmap.
   2. I used the chi-square test of independence, *chi2\_contigency* to call the test statistic and p-value. The p-value is .09 and much lower than the statistic, which means I can reject the null hypothesis—“Readmis” and “Asthma” are correlated.

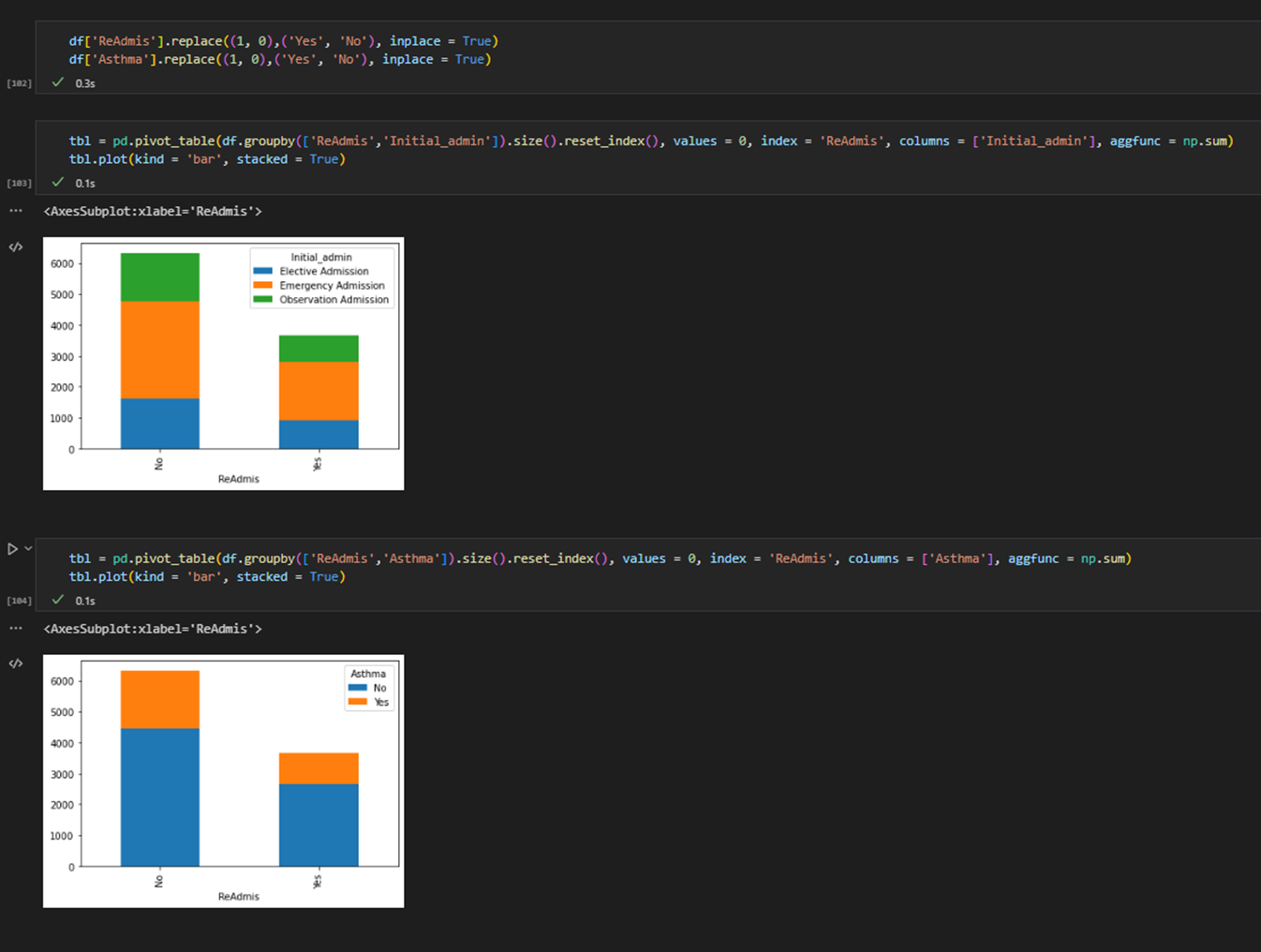
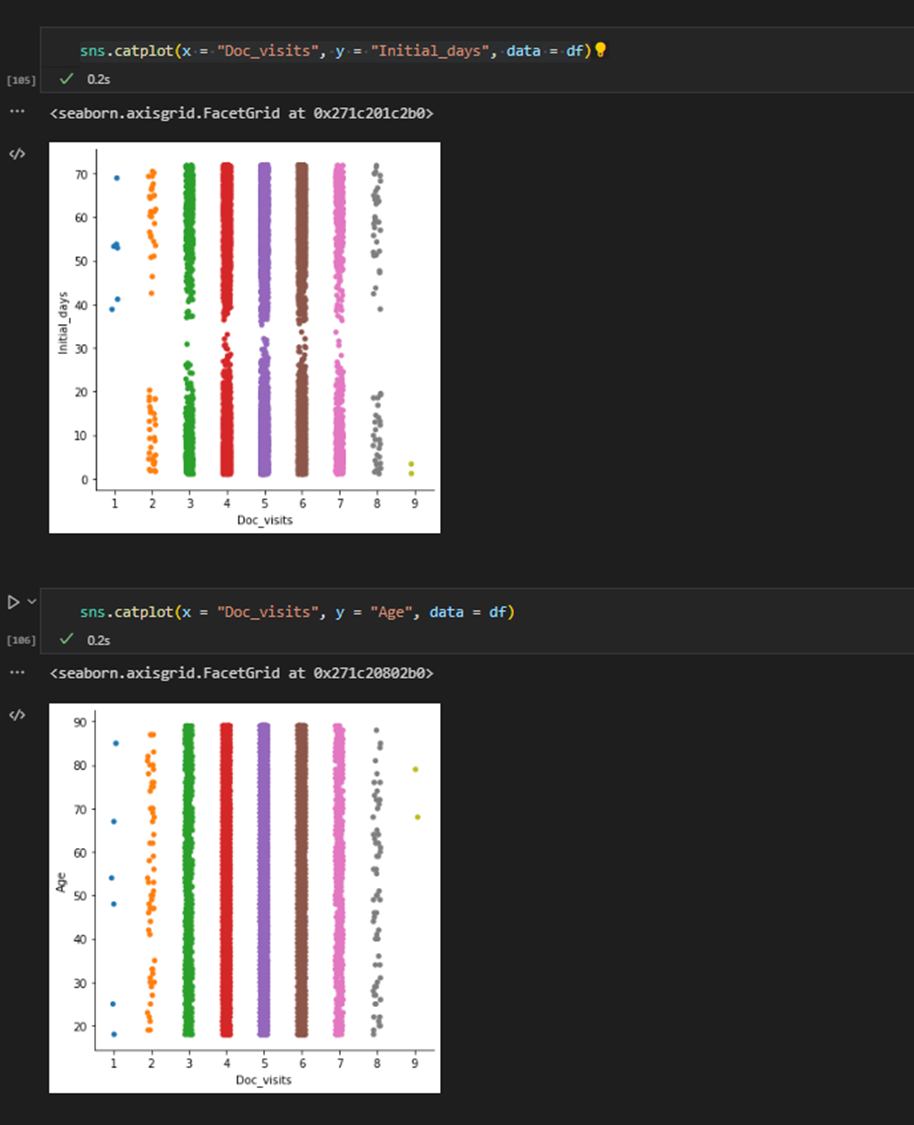


1. Justification
   1. “Readmis” and “Asthma” are both categorical variables and I wanted to test whether or not they were related to one another. As such, the chi-square test was the most appropriate option.

## **C. Univariate Statistics**

1. Categorical Variables
   1. BackPain, Overweight
   2. BackPain showed the potential for readmission correlation. Being overweight is a hot-button issue in the medical community. I wanted to check this column to see if there was merit for that concern.
   3. 
2. Continuous Variables
   1. Initial\_days and Age
   2. Initial\_days, and Age showed the potential for readmission correlation.
   3. 

## **D. Bivariate Statistics**

1. Categorical Variables
   1. Initial\_admin, Asthma
   2. Initial\_admin and Asthma showed the potential for readmission correlation. In order to get an accurate view of these variables, I needed to look at them broken down further.
   3. 
2. Continuous Variables
   1. Initial\_days, Age
   2. Initial\_days and Age showed the potential for readmission correlation. I used Doc\_visits as a comp because the spread for the secondary variable was easy to categorize visually.
   3. 
   4. Chart, histogram

      Description automatically generated

## **E. Implication Summary**

1. **Results:** Conducting the chi-square test brought the following conclusions:
   1. Per the chi-square test, a relationship exists between Asthma and Readmis, as the p-value of .09 is less than the significance level of 2.85.
2. **Limitations:** There are a number of limitations presented by the data as it is and the test I used:
   1. We would not be able to use the chi-square test on non-categorical variables, should the scope of our test expand.
3. **Recommendations:** Because Asthma was found to contribute to Readmis, this medical center should take mitigation action for patients with this illness.

## **F. Panopto**

Multiple Regression Modeling – NBM2 | D208

## **G. Reference web sources**

Seaborn.histplot#. seaborn.histplot - seaborn 0.12.0 documentation. (n.d.). Retrieved September 30, 2022, from https://seaborn.pydata.org/generated/seaborn.histplot.html

https://stackoverflow.com/questions/72009138/pandas-dataframe-label-columns-encoding

pandas.DataFrame.sort\_values — pandas 1.5.0 documentation. (n.d.). Retrieved September 30, 2022, from https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sort\_values.html

GeeksforGeeks. (2022, September 28). Python program to convert a list to string. Retrieved September 30, 2022, from https://www.geeksforgeeks.org/python-program-to-convert-a-list-to-string/?ref=leftbar-rightbar

scipy.stats.chi2\_contingency — SciPy v0.14.0 Reference Guide. het.as.utexas.edu/HET/Software/Scipy/generated/scipy.stats.chi2\_contingency.html. Accessed 09 Oct. 2022.

“Chi-squared Test With Scipy: What’s the Difference Between Chi2\_Contingency and Chisquare?” Cross Validated, 5 Aug. 2014, stats.stackexchange.com/questions/110718/chi-squared-test-with-scipy-whats-the-difference-between-chi2-contingency-and.

scipy.stats.chi2\_contingency — SciPy v1.9.2 Manual. docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2\_contingency.html. Accessed 11 Oct. 2022.

## **H. Acknowledge sources**

LaPointe, J. (2019) 3 strategies to reduce hospital readmission rates, costs, Rev Cycle Intelligence. Agency for Healthcare Research and Quality (AHRQ). Available at: https://revcycleintelligence.com/news/3-strategies-to-reduce-hospital-readmission-rates-costs (Accessed: October 12, 2022).