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. Does customer service, as measured by patient response data, impact the length of initial stays, as measured by Initial\_days?
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 have established a correlation between readmission and the longer initial stay lengths. Now, I will look deeper at how patient satisfaction affects that stay length. I 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. I will also be using matplotlib and seaborn for data visualization.
   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 and Manipulations
   1. A clean, workable data set is necessary in order to run an analysis that is not skewed. After importing my packages and data, I began by renaming the vague columns in order to better see what I was working with.

Graphical user interface, text

Description automatically generated

* 1. After renaming my columns, I dropped all obviously irrelevant data, such as the demographic and location information, and rearranged the order so that my primary variable was first. Though I am primarily interested in the patient response variables, I maintained predictor variables from previous analysis to ensure my view was not skewed. I also needed to keep categorical variables, as my research question focuses entirely on continuous responses.

Text

Description automatically generated

* 1. Lastly, I checked to make sure there were no null or missing entries in my data set.

Text

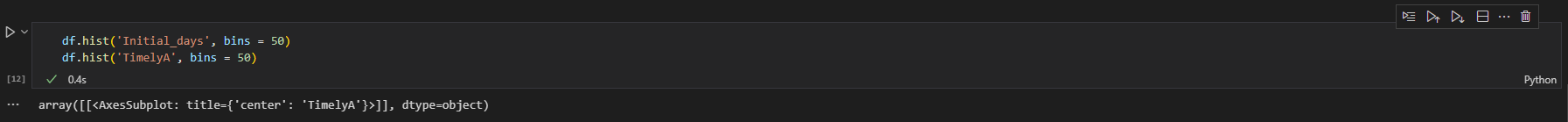
Description automatically generated with low confidence

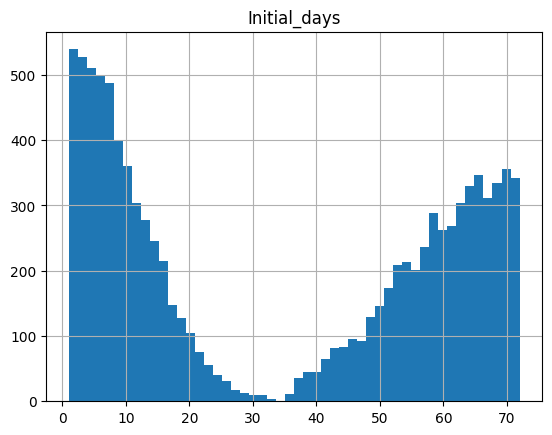
1. Summary Statistics:
   1. Next I used .describe() and .corr() to look at some of the summary statistics for each column. In particular, I paid attention to the correlation matrix, as this would help me decide which of the remaining variables would be the most useful for further analysis. At this point, I felt confident in using Doc\_visits and Age as predictor variables—particularly because they had also been identified as having some correlation to ReAdmis in my D207 assessment and because they were normally distributed in both assessments. On the other hand, while looking at the summary for the response variables, I had immediate concerns about their correlation, as TimelyA and TimelyV are largely not distributed normally compared to Initial\_days.
      1. Target variable: Initial\_days
      2. Predictor variables: Age, Doc\_visits
      3. Tested variables: TimelyV, Reliability, Options, HoT, Courteous, ActiveLis

Graphical user interface

Description automatically generated

1. Data Preparation:
   1. Steps and code have been provided via screenshots in C1 and C2. No categorical data was necessary to encode during these steps.
2. Univariate and bivariate visualizations:
   1. Histograms were generated the continuous variables Initial\_days and TimelyA to check my observation as to distribution. As was noted in the summary previously, distributions are off for the patient response variables and the visualization supports this skew. I also double checked to make sure there were no outliers in Initial\_days with a boxplot, since the histogram was also so unevenly distributed.
   2. Lastly, I compared the distribution for ReAdmis and Asthma once more to decide whether or not to continue to look at the variables.
   3. Both Categorical and Continuous variables were considered in my assessments, beginning with **the Univariate visualizations**:

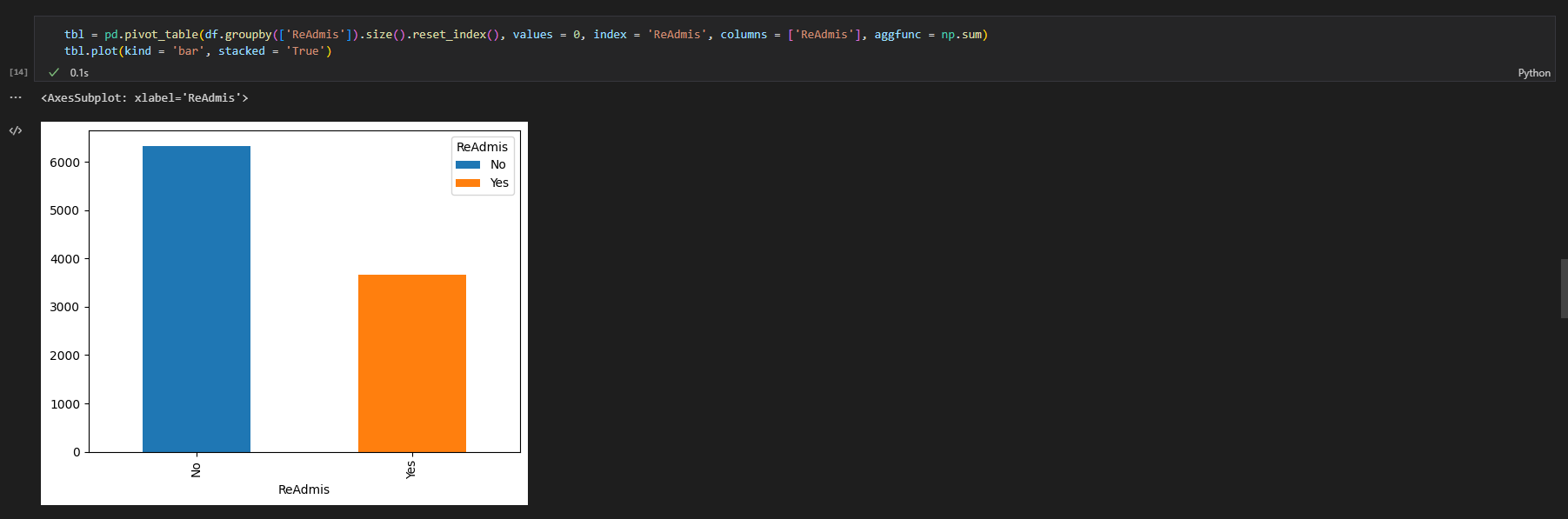


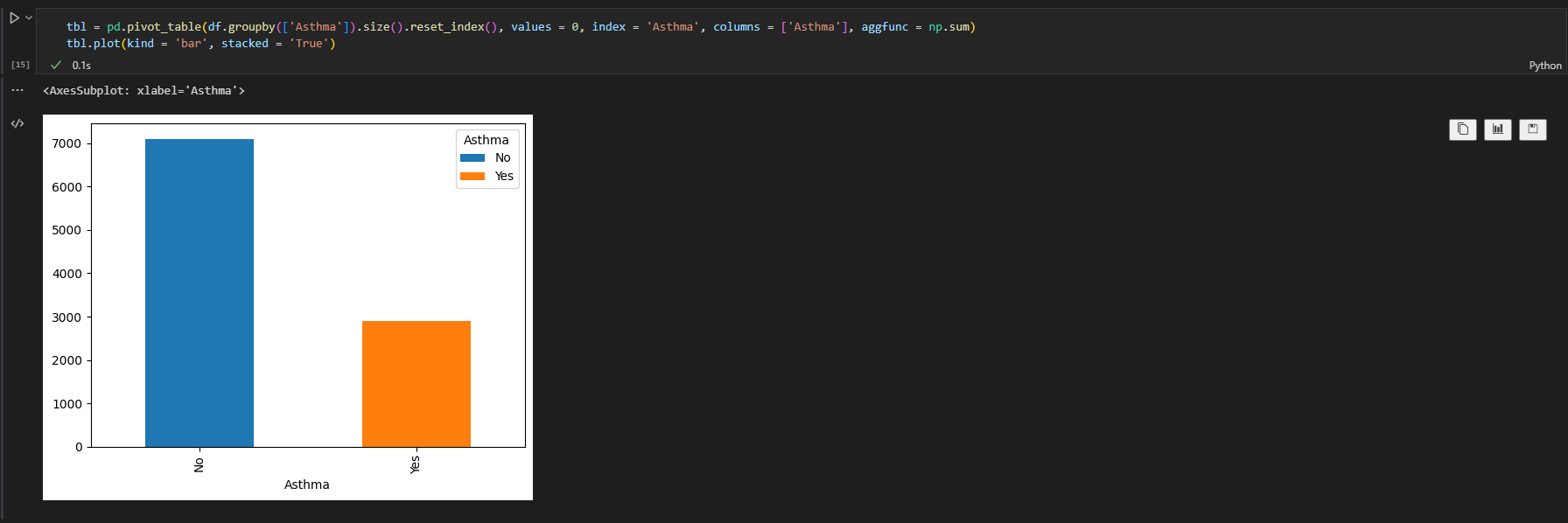
Chart

Description automatically generated

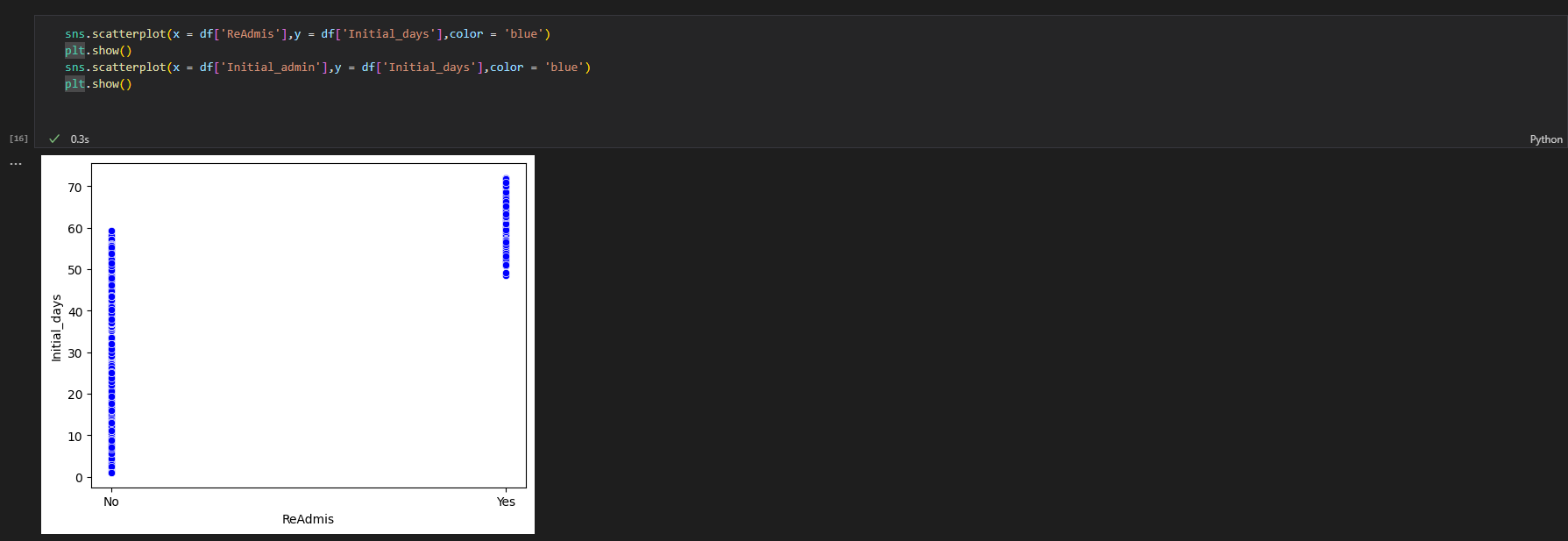
Text

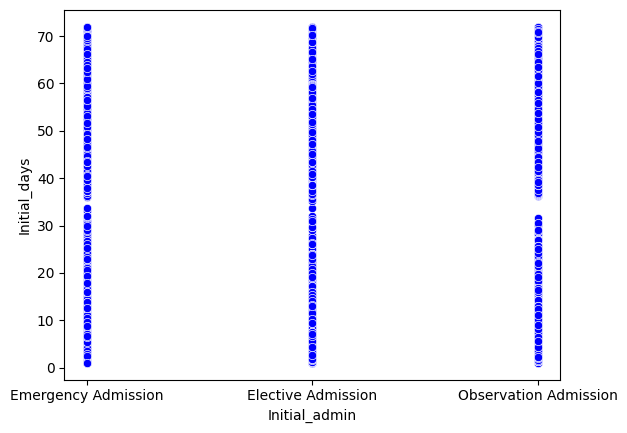
Description automatically generated with medium confidence

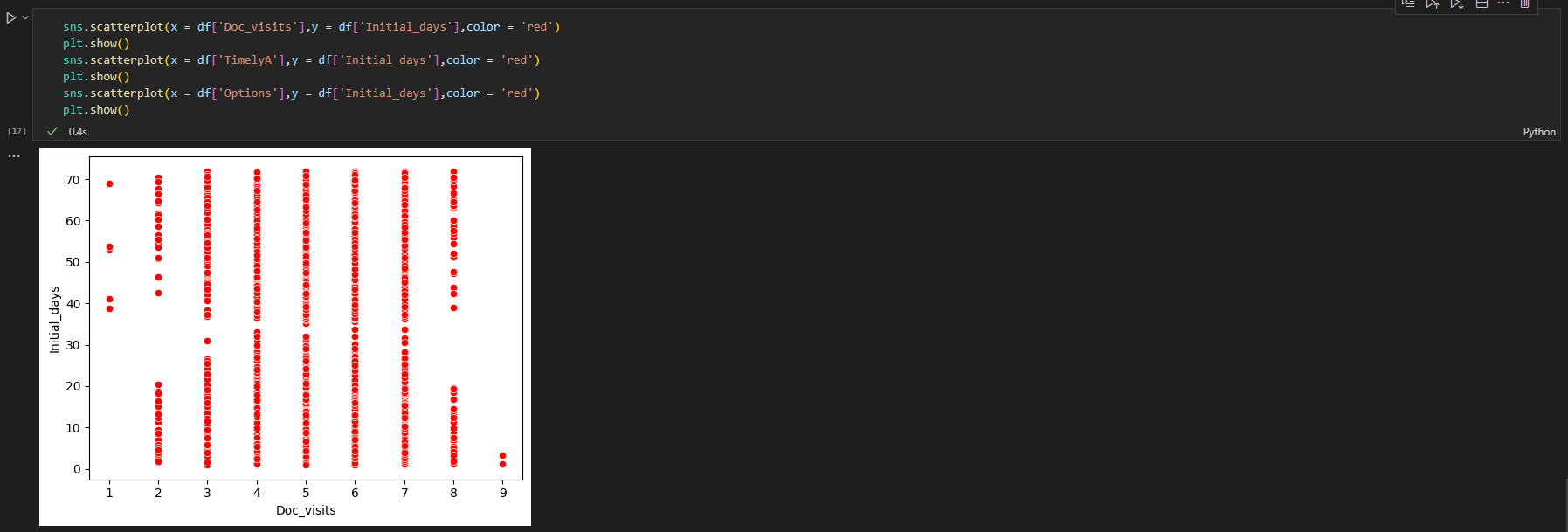


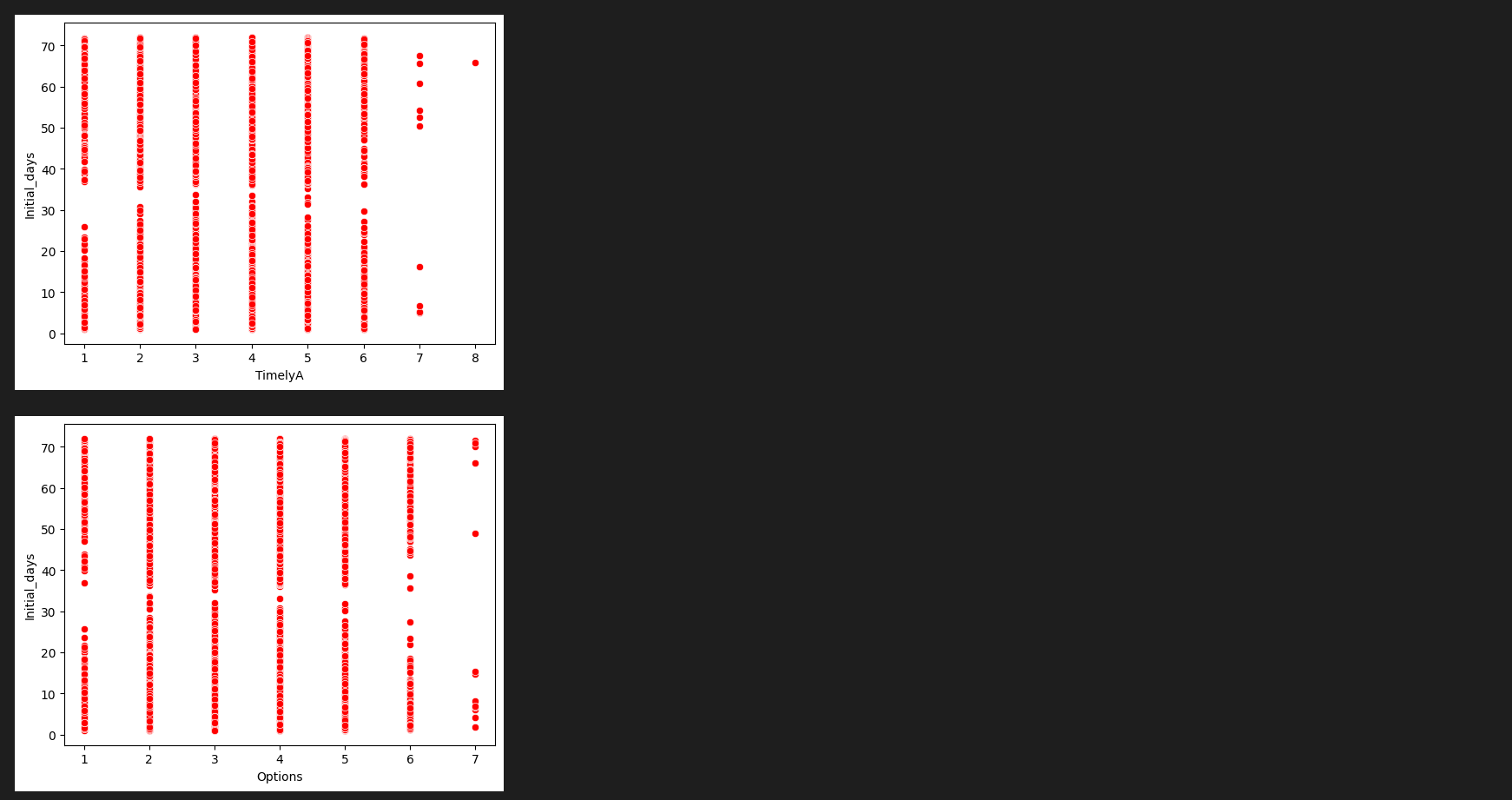


* 1. When moving on to my bivariate visualizations, I was looking for anything that might narrow my focus, as I was concerned that my hypothesis that there *was* a correlation between patient responses and readmissions was incorrect. That said, Initial\_admin looked promising. I continued to observe using the null hypothesis.
  2. Both Categorical (blue) and Continuous variables (red) were considered in my assessments, concluding with **the Bivariate visualizations**:





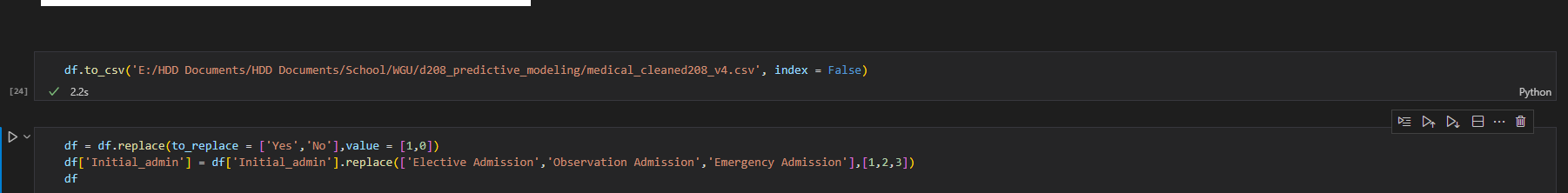




1. Provide prepared data set:

Text

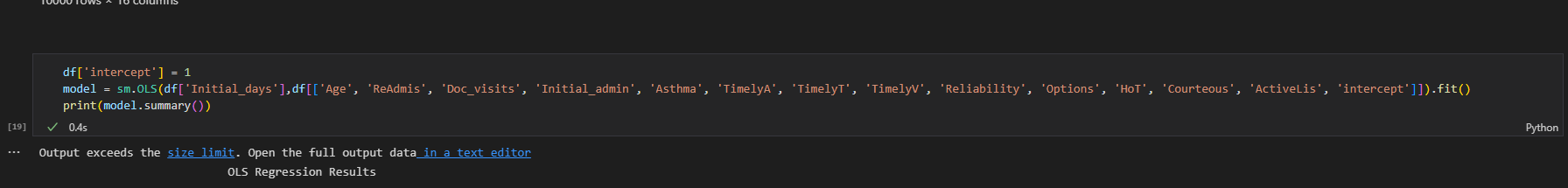
Description automatically generated



Data provided.

## **D. MODEL COMP AND ANALYSIS**

1. Comp initial and reduced multiple regression model:
   1. I used OLS to run an initial regression on all of the remaining variables, which required me to convert the Initial\_admin variable first.



A picture containing text, plaque

Description automatically generated

* 1. I kept the remaining observations and variables and received an R-squared value of 0.725. So, 73% of the variation is explained by this model. This is below the ideal, suggesting there the independent variables are not explaining much in way of the variation of the dependent variable. Looking at the P-values doesn’t give me a lot of hope that my hypothesis will be sound. I suspect the variables to be removed will be the ones pertaining to my hypothesis. I press on to create a heat map in order to visualize whether or not that’s correct.
  2. The initial regression model with 13 variables: Y = 20.6717 + 0.0034 [Age] + 46.4721 [ReAdmis] - 0.1724 [Doc\_visits] - 0.8528 [Initial\_admin] + 0.0397 [Asthma] - 0.4045 [TimelyA] - 0.1149 [TimelyT] + 0.1124 [TimelyV] - 0.3905 [Reliability] - 0.0209 [Options] + 0.1911 [HoT] + 0.2302 [Courteous] + 0.2070 [ActiveLis].

Graphical user interface

Description automatically generated

Chart

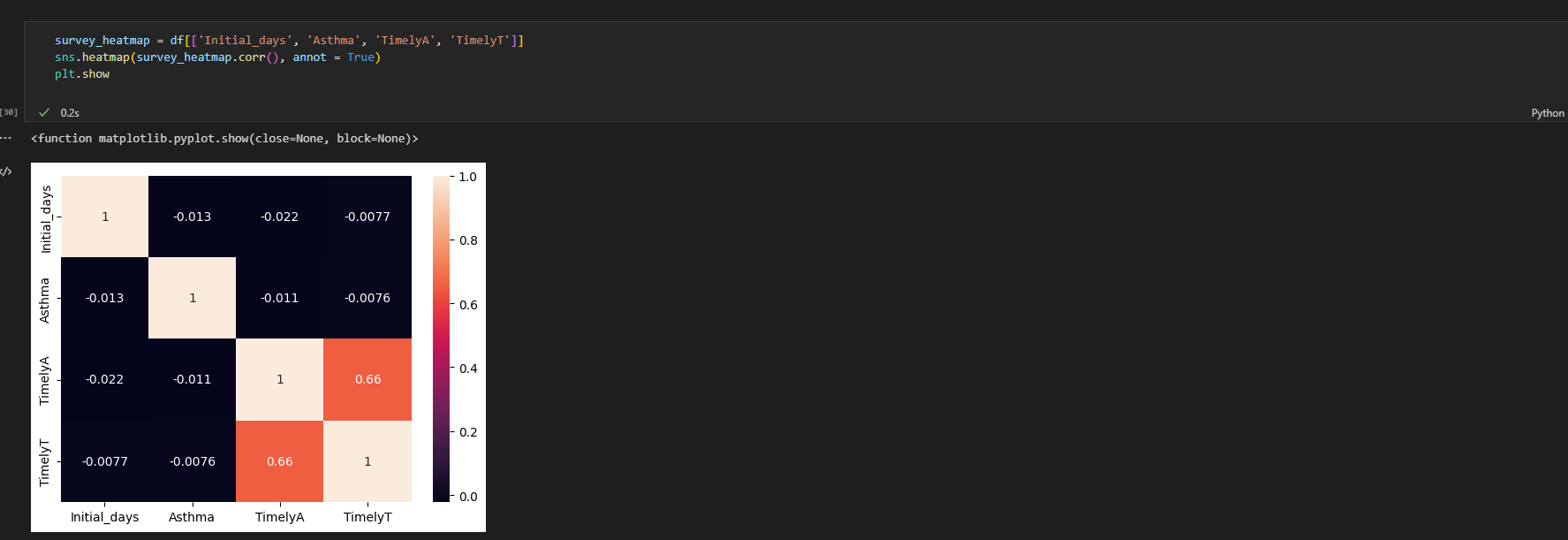
Description automatically generated

* 1. Unfortunately, the heatmap only asserts what I was already seeing. There is unlikely to be a strong correlation between Initial\_days and any of the patient response variables. I took a closer look at only the variables I’d identified as patient response variables to be sure.
  2. Before generating a reduced regression model, I needed to look at only the variables that pertained to my research question.

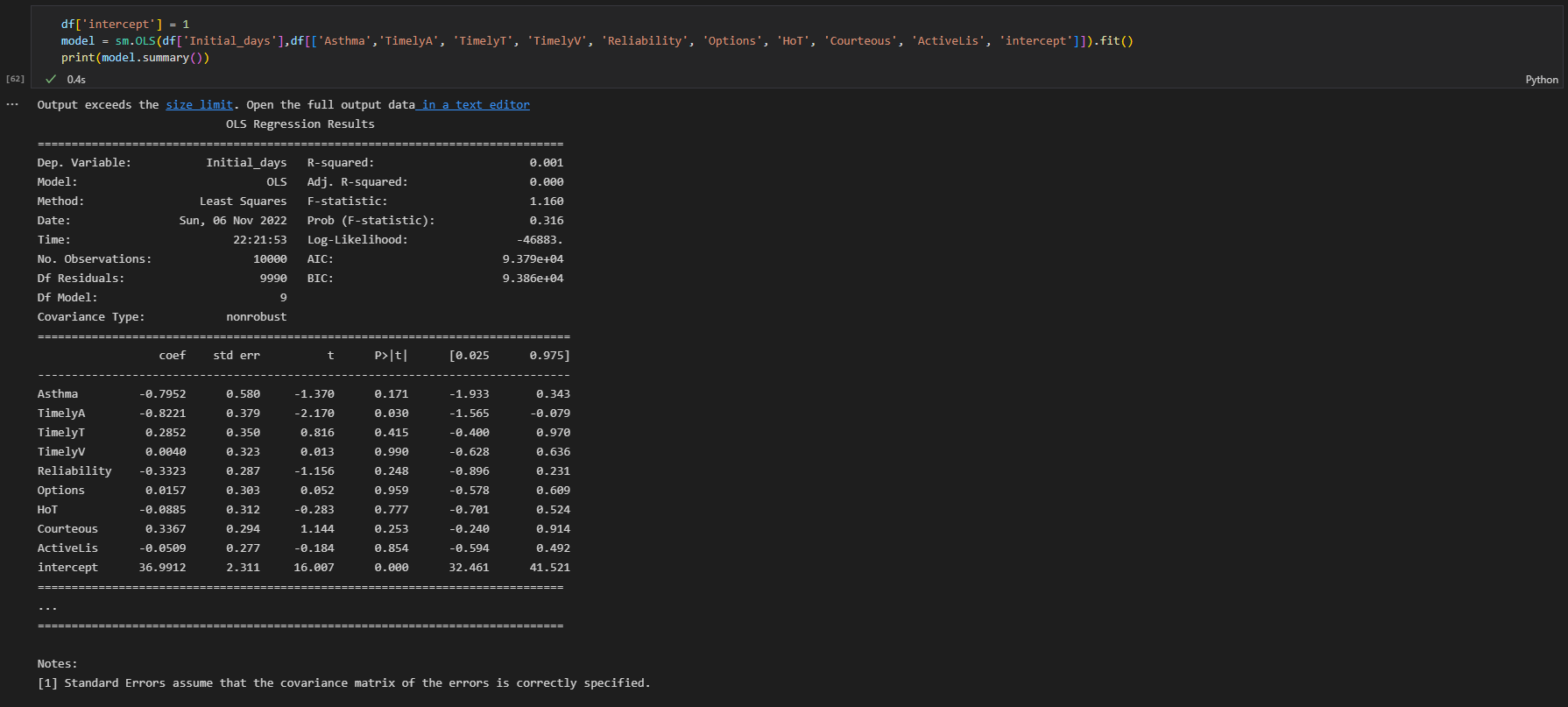
Graphical user interface, application, Teams

Description automatically generated

* 1. The narrowed vision of the heatmap is clearer and brings me back to the assertion that the patient responses have little to do with how long an initial visit lasts.



* 1. Despite this finding, I wanted to put together one more matrix to compare my assumptions from D207 to the newfound skills and abilities I’ve gained in D208. I’d previously ascertained that Asthma has a correlation with readmissions. I added it back to my reduced regression model.
  2. The updated model has an insignificant R-squared. None of the listed variables are strong predictors of the variance. As such, there is no linear relationship between patient responses and the length of their stay.
  3. The reduced regression model with nine variables: Y = 36.9912 – 0.7952 [Asthma] - 0.8221 [TimelyA] + 0.2852 [TimelyT] + 0.0040 [TimelyV] - 0.3323 [Reliability] + 0.0157 [Options] - 0.0885 [HoT] + 0.3367 [Courteous] - 0.0509 [ActiveLis].



## **E. Data Set Analysis**

1. The technique I used for variable selection is explained in my step-by-step explanations. While searching for answers to my research question, I quickly hit a wall when it became obvious through the correlation matrix that I wasn’t going to find connection between patient reviews and readmission likelihood. I continued to explore this avenue against the Initial\_days variable. The residual plot for the model is ultimately the same as the original, since there was no discernable connection between the two tested variables.
2. **The most statistically significant variable was TimelyA, based on a negative coef less than – 0.8 and a p-value of 0.030.**

Text

Description automatically generated

1. Output and calculations are included in the code section of this communication.
2. Code is included in the code section of this communication.

## **F. DATA/Implications Summary**

1. Results:
   1. The final multiple linear regression equation is Ŷ = 36.9912 – 0.7952 [Asthma] –0.8221 [TimelyA] + 0.2852 [TimelyT] + 0.0040 [TimelyV] – 0.3323 [Reliability] + 0.0157 [Options] – 0.0885 [HoT] + 0.3367 [Courteous] – 0.0509 [ActiveLis].
   2. The coefficients suggest for every unit of:
      1. Asthma, Initial\_days decreases by 0.7952 units.
      2. TimelyA, Initial\_days decreases by 0.8221 units.
      3. TimelyT, Initial\_days increases by 0.2852 units.
      4. TimelyV, Initial\_days increases by 0.0040 units.
      5. Reliability, Initial\_days decreases by 0.3323 units.
      6. Options, Initial\_days increases by 0.0157 units.
      7. HoT, Initial\_days decreases by 0.0885 units.
      8. Courteous, Initial\_days increases by 0.3367 units.
      9. ActiveLis, Initial\_days decreases by 0.0509 units.
   3. The statistical and practical significance of this model is primarily exclusionary. The efforts made to identify potential customer service impacts on readmission resulted in an understanding that there is not a strong correlation between patient responses and Initial\_days.
2. Limitations:
   1. This analysis only looked at patient response numbers in correlation with Initial\_days. As such, the information is limited to which patients actually responded to the survey and may not be a clear picture of overall impact. In addition, while the testing parameters for this assignment recommended using categorical and continuous data, it is possible that adding the categorical variables impacted the outcome of the initial investigation—before the focuses was narrowed in the reduced model.
3. Recommendations:
   1. Because the patient responses did not have a significant impact on Initial\_days, this medical center should not spend resources to address potential customer service concerns in order to offset readmission numbers.
   2. In order to maintain the health of this analysis, the medical center should launch campaigns to gather as much patient response data as possible. I would also recommend keeping track of the response vs. ignored KPI in order to segment response efforts.
   3. This analysis should be run quarterly in order to ensure the value of this assessment.

## **G. Panopto**

[MULTIPLE REGRESSION FOR PREDICTIVE MODELING | TASK 1](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=91335005-ed77-4219-8e72-af460063c560)

## **H. Reference web sources**

GeeksforGeeks. (2022, June 8). Working with Missing Data in Pandas. https://www.geeksforgeeks.org/working-with-missing-data-in-pandas/

Abba, I. V. (2022, May 25). TypeError: can't multiply sequence by non-int of type 'float' [SOLVED]. freeCodeCamp.org. https://www.freecodecamp.org/news/typeerror-cant-multiply-sequence-by-non-int-of-type-float-solved/

N, Z. (2021, June 1). How to Create a Residual Plot in Python. Statology. https://www.statology.org/residual-plot-python/

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

How to plot multiple scatter plots in the same window with matplotlib? (2018, April 11). Stack Overflow. https://stackoverflow.com/questions/49783682/how-to-plot-multiple-scatter-plots-in-the-same-window-with-matplotlib

How to Create a Correlation Matrix using Pandas. (n.d.). Retrieved November 6, 2022, from https://datatofish.com/correlation-matrix-pandas/

Libretexts. (2022, July 18). 4.5: Building Your Own Model Equation. Mathematics LibreTexts. https://math.libretexts.org/Bookshelves/Scientific\_Computing\_Simulations\_and\_Modeling/Book:\_Introduction\_to\_the\_Modeling\_and\_Analysis\_of\_Complex\_Systems\_(Sayama)/04:\_DiscreteTime\_Models\_I\_\_Modeling/405:\_Building\_Your\_Own\_Model\_Equation

## **G. Acknowledge sources**

None.