**D208 Performance Assessment II**

Renzo Espinoza

College of Information Technology, Western Governors University

D208: Predictive Modeling

Dr. Keiona Middleton

September 14, 2022

**A1: Research Question**

Using logistic regression, the research question that I will answer is: which independent variables are related to the likelihood of a patient being readmitted within a month of release. According to the data dictionary provided to me, for this scenario I am asked to investigate whether there is a readmission problem in a popular medical hospital chain with patients all throughout the United States. For this reason, my research question is focused on the response variable found in the dataset called ‘Re\_admis’.

**A2: Objectives and Goals**

The goal of my analysis is to identify which variables in the data can accurately predict whether a patient will be readmitted within a month of release. The dataset contains many features of patients that I could use as potential predictor variables.

**B1: Summary of Assumptions**

Logistic regression models come with many assumptions. According to Statistics Solutions (“Assumptions of Logistic Regression”, 2021), one of these assumptions is that for binary logistic regression, then the dependent variable must only have two possible outcomes. The dependent variable ‘Re\_admis’ fits that assumption. Another assumption is that the observations of the dataset are independent of one another. The independent variables should also not be highly correlated with each other. Also, the sample size of the dataset should be sufficiently large. Finally, there should be a linear relationship between the predictor variables and the log odds, or logit of the dependent variable.

**B2: Tool Benefits**

To perform my analysis, I will use the Python programming language alongside the many data analysis tools in the Python ecosystem. For storing and manipulating the raw data provided to me, I used the NumPy and Pandas libraries. I also utilized Matplotlib and Seaborn to perform visualizations of the data. I used the Statsmodels library to build the logistic regression model. All of these tools, conveniently in the same programming environment, benefitted me greatly throughout all stages of the analysis.

**B3: Appropriate Technique**

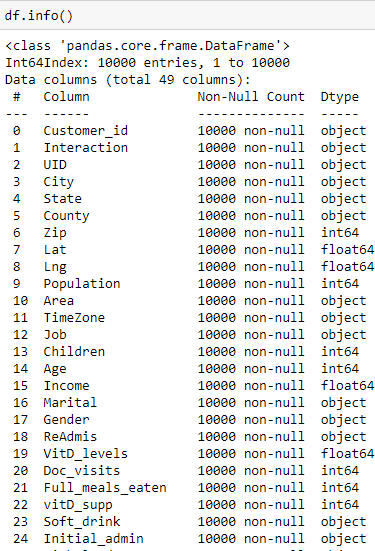
I believe logistic regression is an appropriate technique to analyze which variables in the dataset have an influence on the likelihood of a patient being readmitted. According to Çetinkaya-Rundel & Hardin (2021), logistic regression is used to model the probability of an event, where the outcome is a binary categorical variable. Therefore, I believe this technique is most appropriate for analyzing the relationship between the dichotomous dependent variable ‘Re\_admis’ and the many potential independent variables found in the dataset.

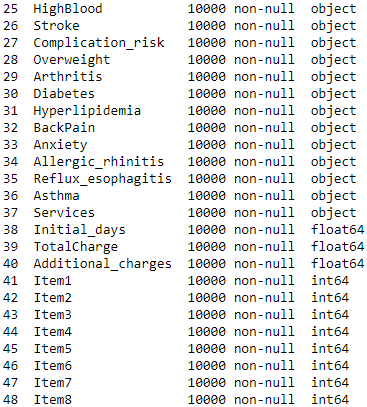
**C1: Data Goals**

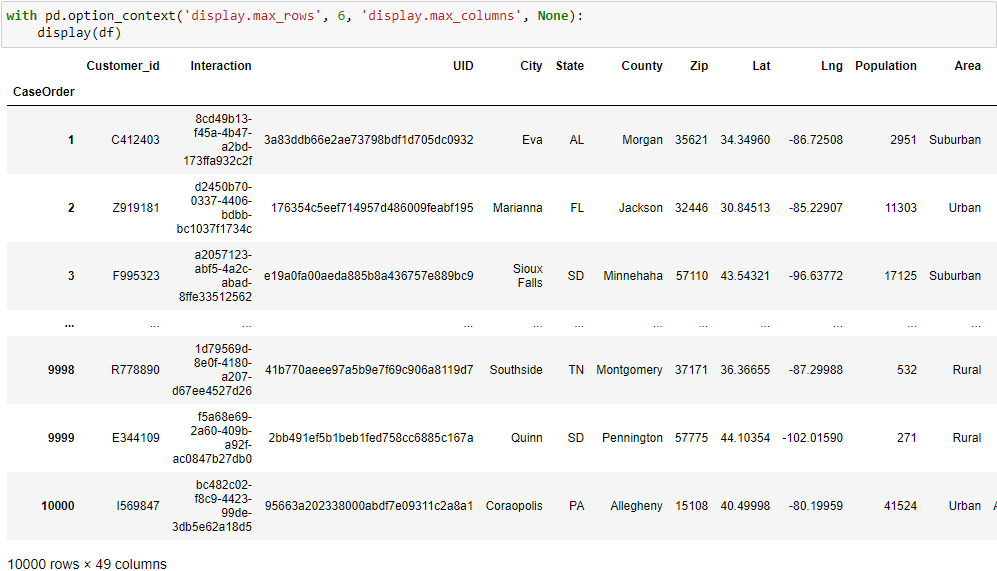
My data preparation goal is to refine the data so the analysis process goes smoothly for both myself and the other analysts on my team in this scenario. To achieve this I will find and handle outliers, convert variables to the proper datatypes, rename columns that do not fit the general naming conventions, reduce the cardinality of categorical variables where needed, and perform one-hot encoding. After reading a passage from Çetinkaya-Rundel & Hardin (2021) where they warned that ignoring exceptional cases can lead to a model performing poorly, I decided that I will only remove outliers that are an obvious input error instead of a natural observation.

**C2: Summary Statistics**

To start off my data preparation process, I used the ‘Dataframe.info’ method from the Pandas library to get a bird’s-eye view of the data that I imported into a Pandas Dataframe. I also used the ‘Ipython.display’ method to take a quick peek at the dataset without displaying all 10000 entries.

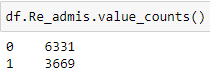


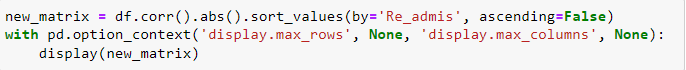


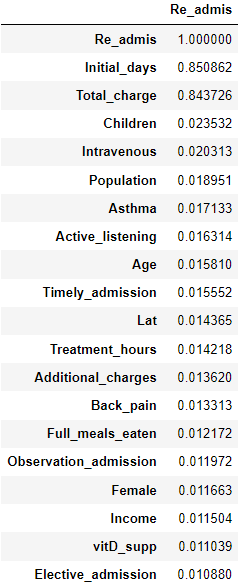


From the output displayed above, I was able to quickly identify the categorical variables that need to be converted into numerical values through basic conversion and one-hot encoding. For specific variables, I used the ‘Dataframe.describe’ and ‘Dataframe.value\_counts’ methods to generate summary statistics of each variable.

Below is the summary of the target variable, ‘Re\_admis’. From this summary you can see that more than a third of patients are readmitted within a month of release, suggesting that this chain of hospitals does have a readmission problem.



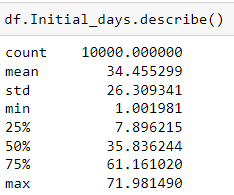
Instead of performing the tedious task of visualizing the relationship between every potential predictor variable and target variable, I used the ‘Dataframe.corr’ method to create a correlation matrix, allowing me to identify which independent variables have the strongest linear relationship with the target variable. Note that I performed this after cleaning the data and that I’m displaying the absolute value of the resulting correlation coefficients.



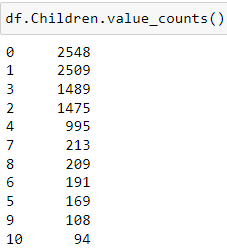
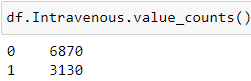
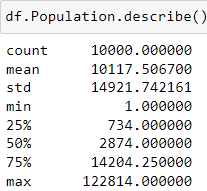
The correlation matrix was much larger than the screenshot I displayed above, I chose the correlation coefficient of 0.01 as an arbitrary cutoff point for what predictors I will consider for the model. . Every variable displayed above, excluding ‘Lat’ and ‘Re\_admis’ will be my chosen predictor variables. ‘Initial\_days’ and ‘Total\_charge’ are the most promising variables with correlation coefficients around 85%, but after those two there is a considerable drop-off with ‘Children’ being the next highest at 2%.

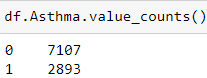
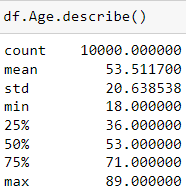
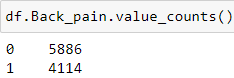
To detect multicollinearity amongst these predictors, I used the ‘variance\_inflation\_factor’ method from the Statsmodels library to calculate the VIF of the independent variables. Both ‘Initial\_days’ and ‘Total\_charge’ have a very high VIF, so I removed the higher of the two, which resulted in the VIF of ‘Initial\_days’ being reduced to 2.6. From there I went one-by-one removing the variable with the highest VIF until there none over 5. This resulted in my pool of predictors being reduced from 18 to 13 variables.

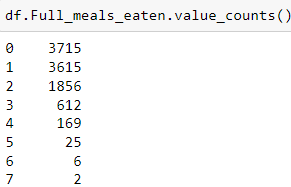
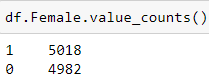
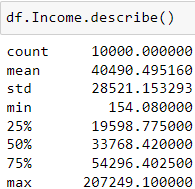
Now that I narrowed down my pool of predictors, I can give a summary of each variable. To start off, a summary of ‘Initial\_days’ is displayed below. The average amount of days that a patient stays for during their initial visit is around 34 days, and the highest amount of days recorded is 71 days. The relatively high standard deviation suggests that the values are fairly spread out.



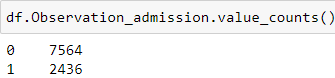
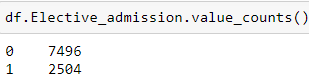
Below are the summaries for the next three predictors, ‘Children’, ‘Intravenous’, and ‘Population’. According to the summary for ‘Children’, a majority of patients fall in the range of having 0 to 4 kids. For ‘Intravenous’, around 31% of patients received some form of intravenous primary service. The summary of ‘Population’ tells us that average amount of people within a mile radius of a patient is around 10,000. With 14,000 people being the 75th percentile and the max value being 123,000, this suggests that a wide majority of patients live in lower density areas.

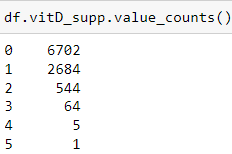


 Below are the summaries for the next set of predictors, ‘Asthma’, ‘Age’, and ‘Back\_pain’. According to the summary for ‘Asthma’, around 29% of patients reported having asthma. The summary of ‘Age’ tells us that average age of a patient is 53 years old. For ‘Back\_pain’, a surprisingly high amount of patients reported having chronic back pain, at a rate of 41%.

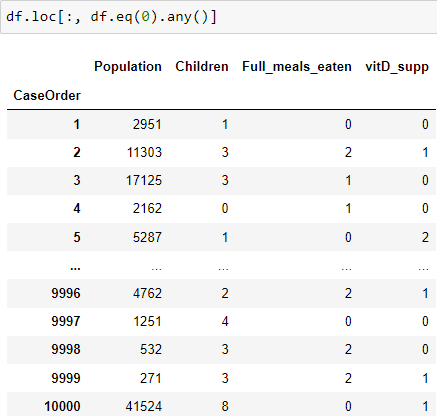
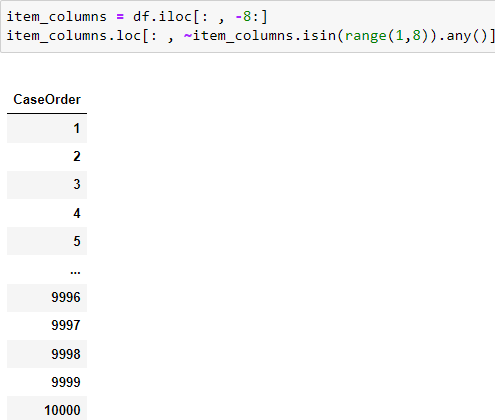
 Below are the summaries for the three predictors, ‘Full\_meals\_eaten’, ‘Female’, and ‘Income’. For ‘Full\_meals\_eaten’, a majority of patients ate 0 to 2 full meals throughout their hospitalization. According to the summary for ‘Female’, roughly half of all patients are women. The summary of ‘Income’ tells us that average annual income of patients is around $40,500.

Below are the summaries for ‘Observation\_admission’, ‘Elective\_admission’, and ‘vitD\_supp’. According to the summary for ‘vitD\_supp’, 67% of patients were administered no vitamin D supplements throughout their stay. The other two summaries show that around 24% and 25% of patients were initially admitted for observation or electively. From this we can gather that around 50% of patients were admitted under an emergency.

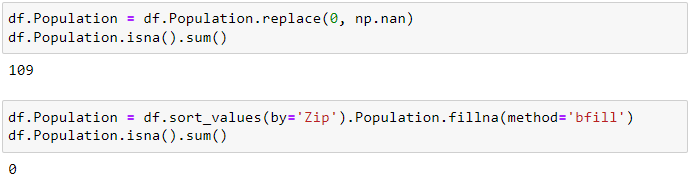




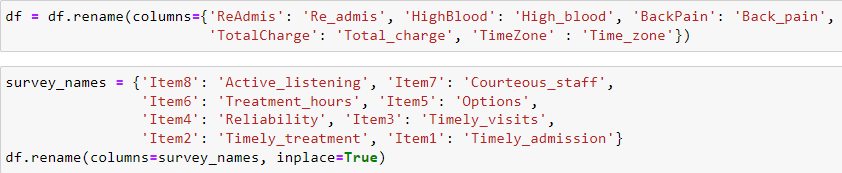
**C3: Steps to Prepare the Data**

 For finding input errors, I used the ‘Dataframe.loc’ functionality to query for 0 values appearing in the dataset and values outside of the range of the survey columns. Thankfully, no outliers appeared in the survey columns. For the columns containing zero-values, ‘Population’ is the only column where a 0 would signify an input error.

To perform imputation on the missing ‘Population’ entries, I sorted the Dataframe by the value of the ‘Zip’ column, then I replaced the missing values using the ‘Dataframe.fillna’ method and set the imputation method to backwards-fill.

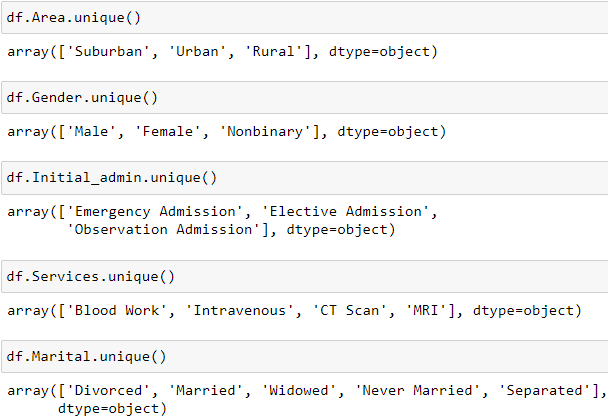


 I identified 13 columns in the dataset that have ‘Yes’ or ‘No’ as the two possible outcomes. I converted these columns from String to integer, allowing me to use these variables in the logistic regression model.

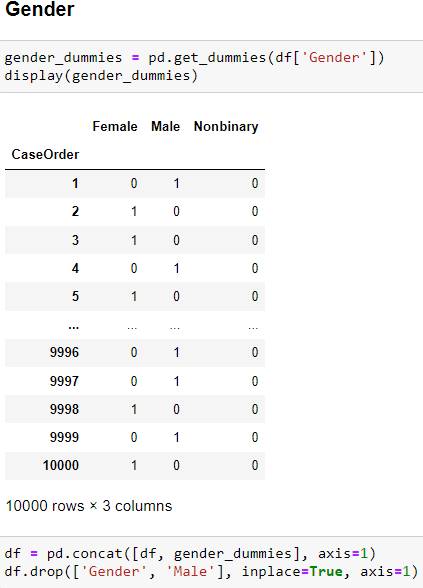
 For clarity, I renamed the survey columns to explain what company qualities the survey pertains to. I also renamed columns that did not fit the naming conventions of separating words with an underscore.

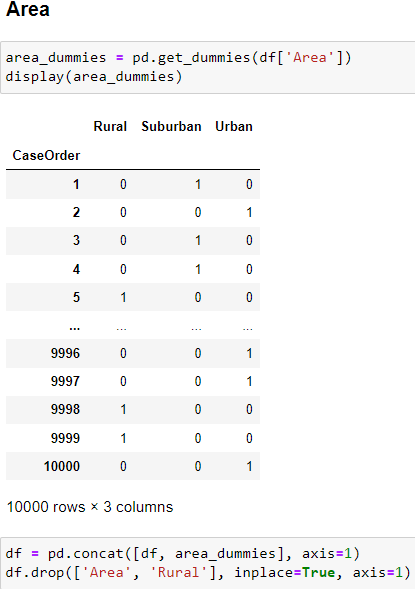
 I also recoded ‘Complication\_risk’ to an integer to give ordinality to the variable.

 I reduced the cardinality of the 8 survey variables by performing ceiling division on each column. This effectively reduced the cardinality of the predictors from 8 to 4.

 There were certain categorical variables that I felt were suitable for one-hot encoding. I used the ‘Dataframe.unique’ method to get an array of every unique value in those columns.

All of the variables are suitable for one-hot encoding, except for ‘Marital’ because I believe the cardinality is too high. To create the dummy variables for each predictor, I used the ‘Pandas.get\_dummies’ method. I then added the new columns to the dataset and then dropped the original predictor along with one of the dummy variables to serve as the reference group.

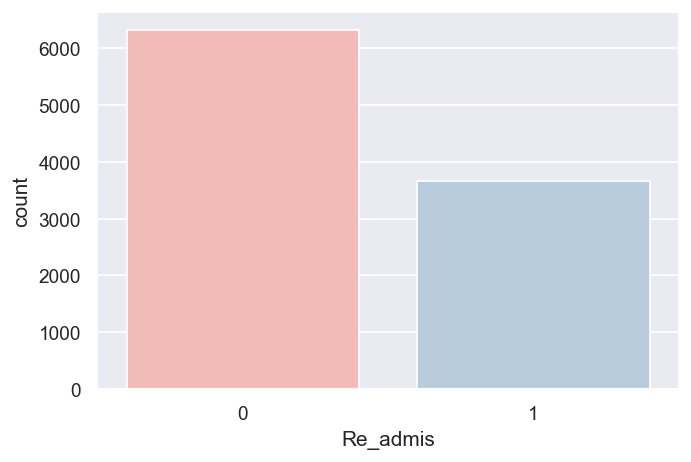


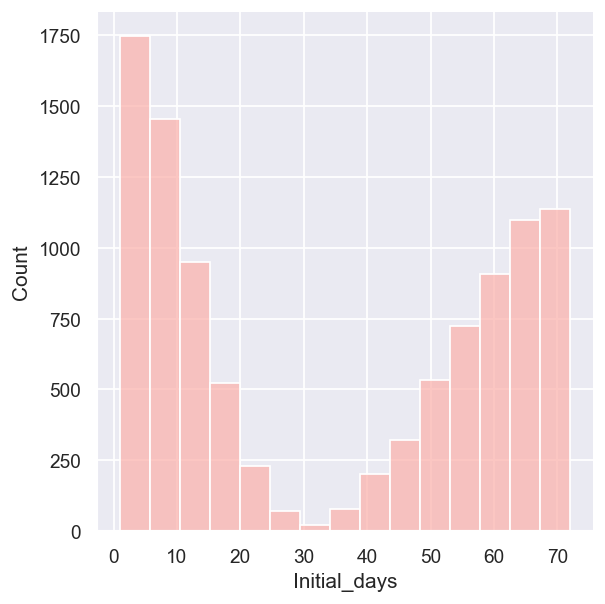


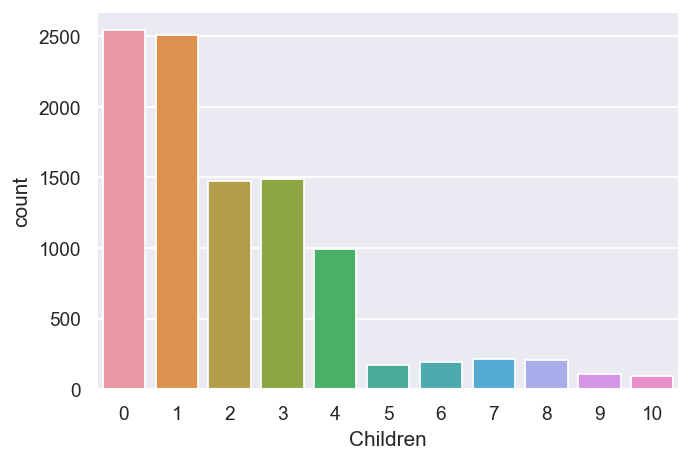


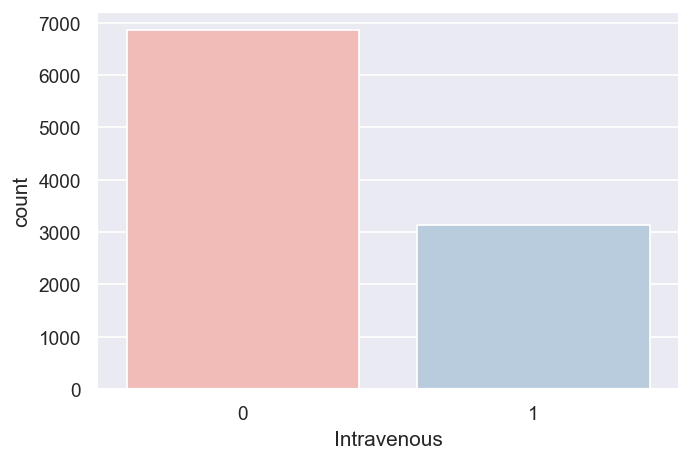
**C4: Visualizations**

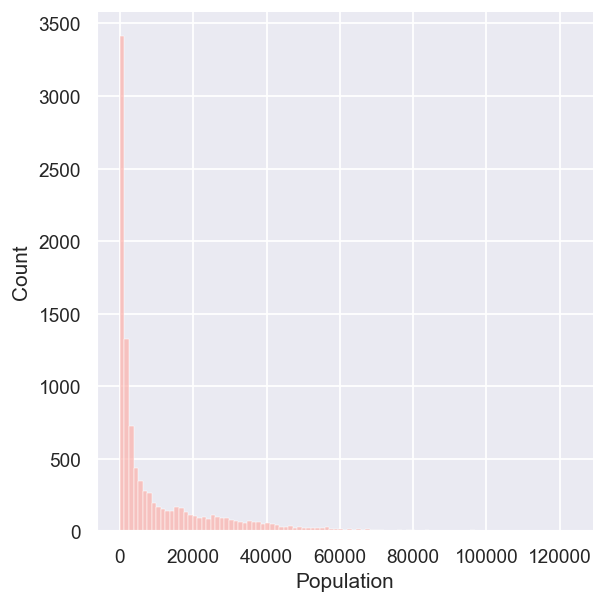
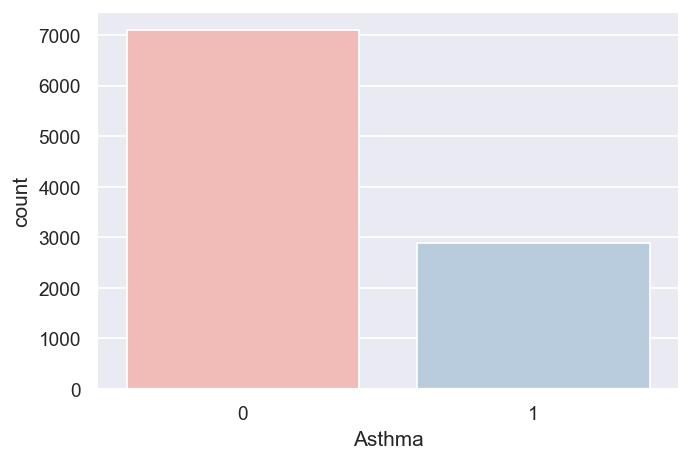
Univariate graphs:

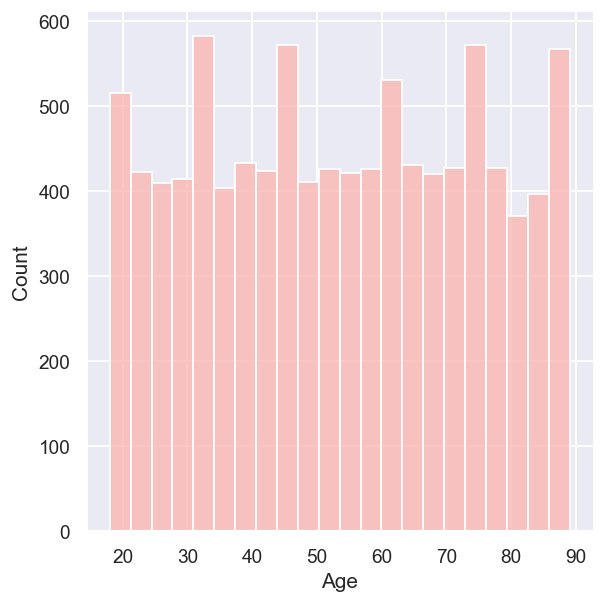
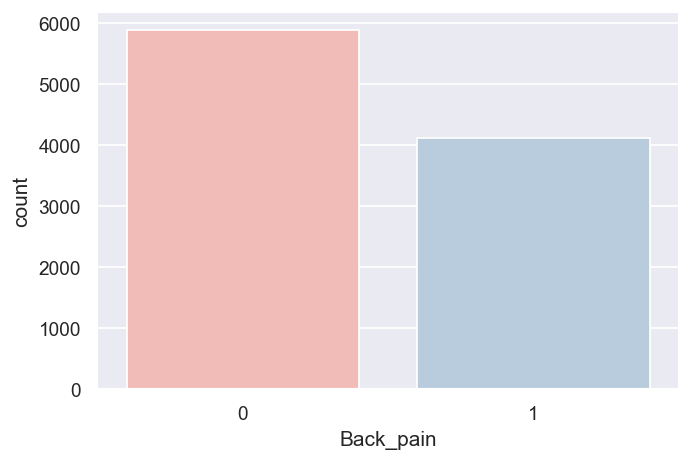


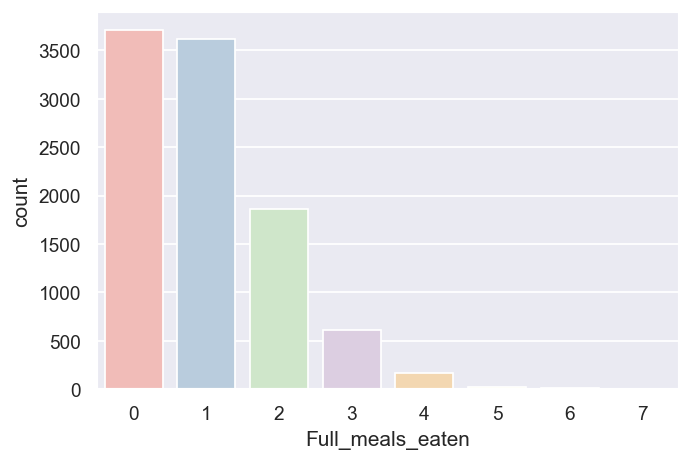
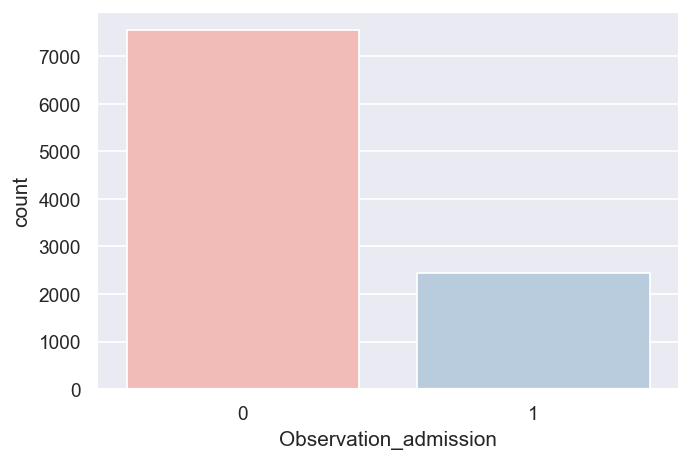


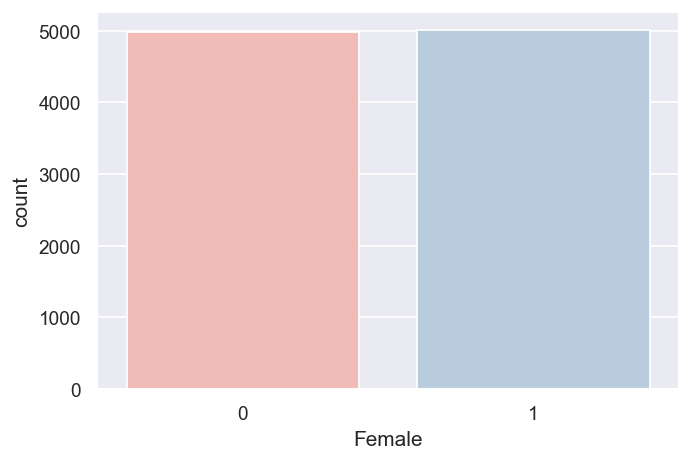
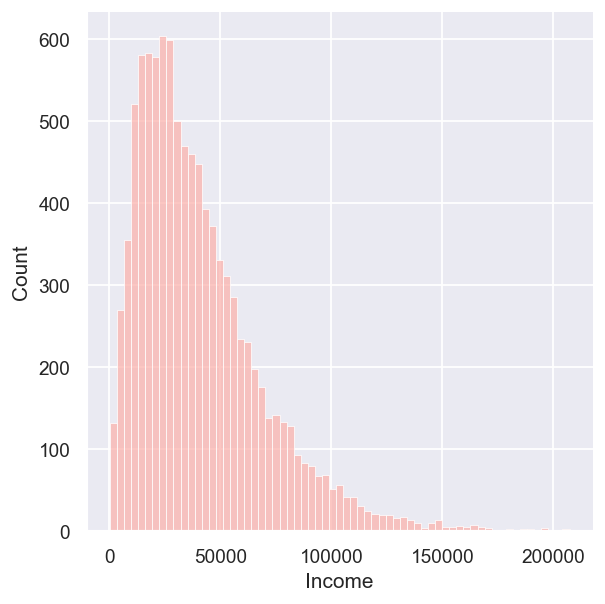


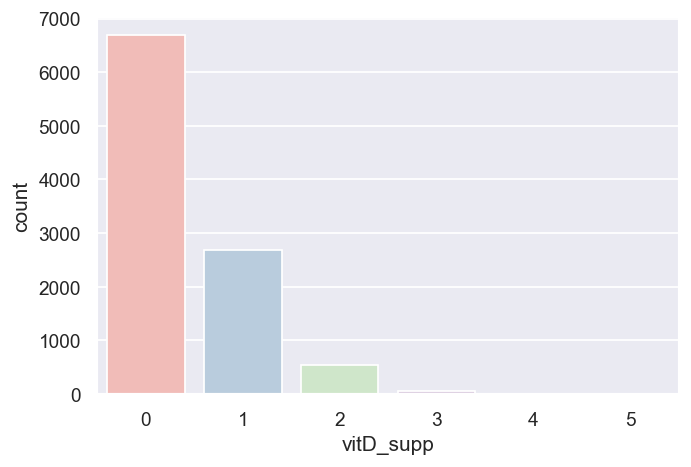


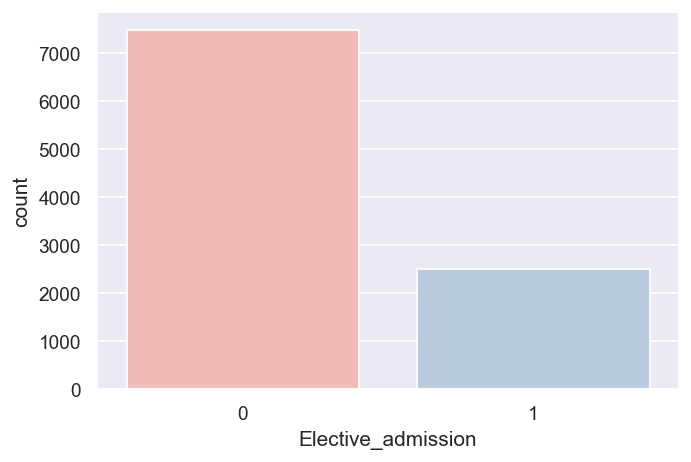


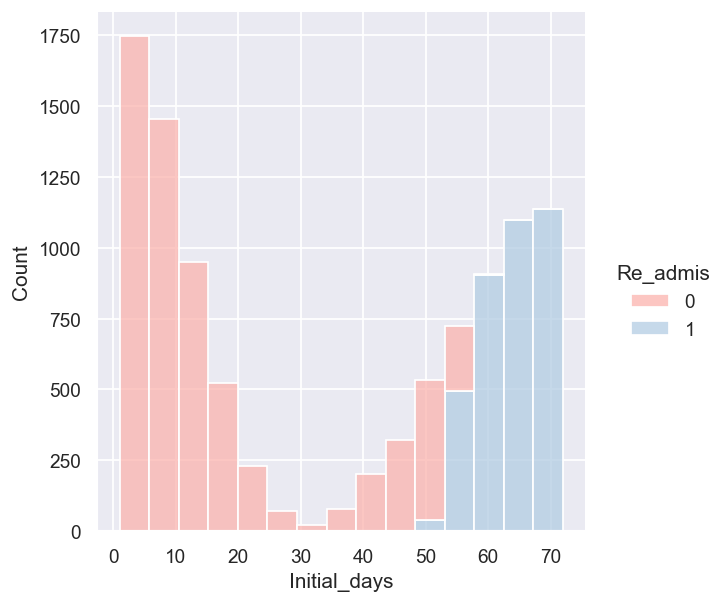


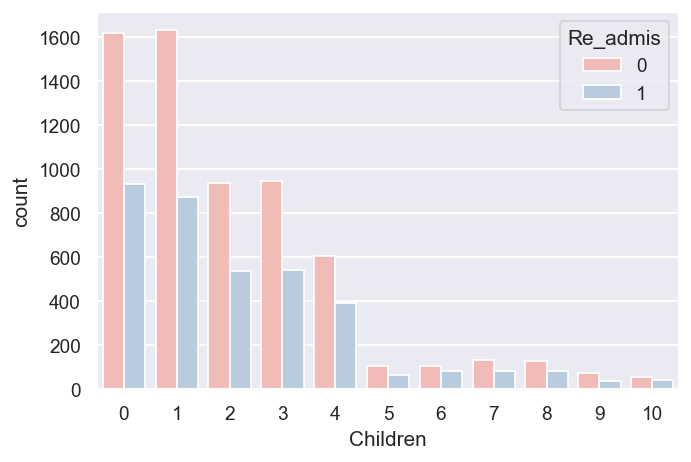


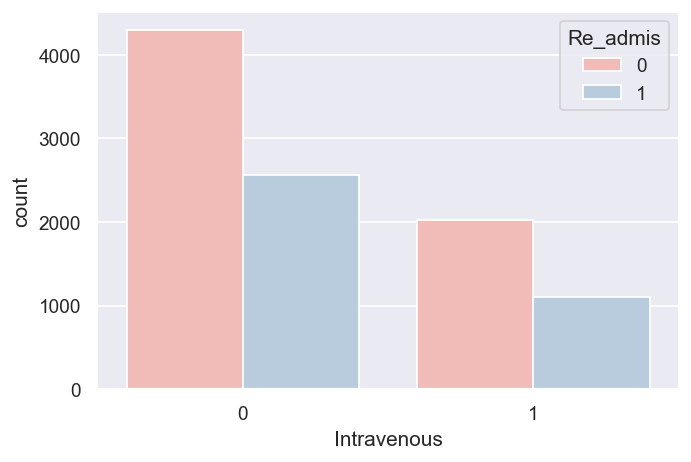
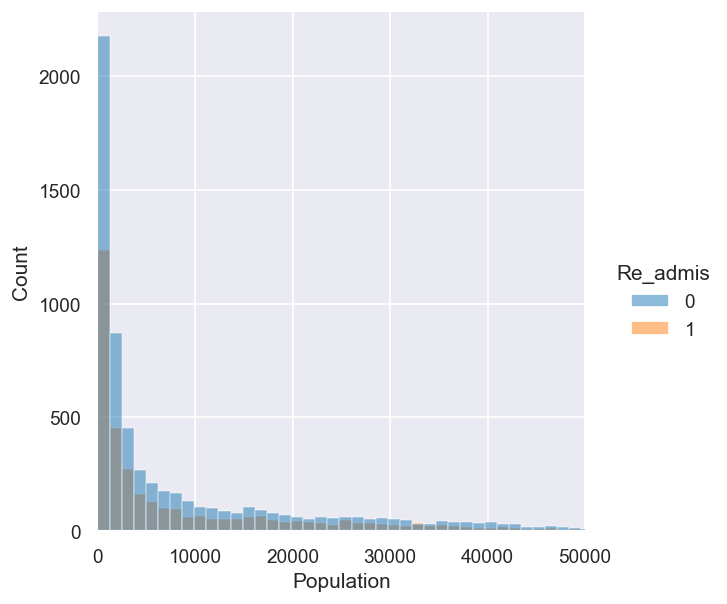


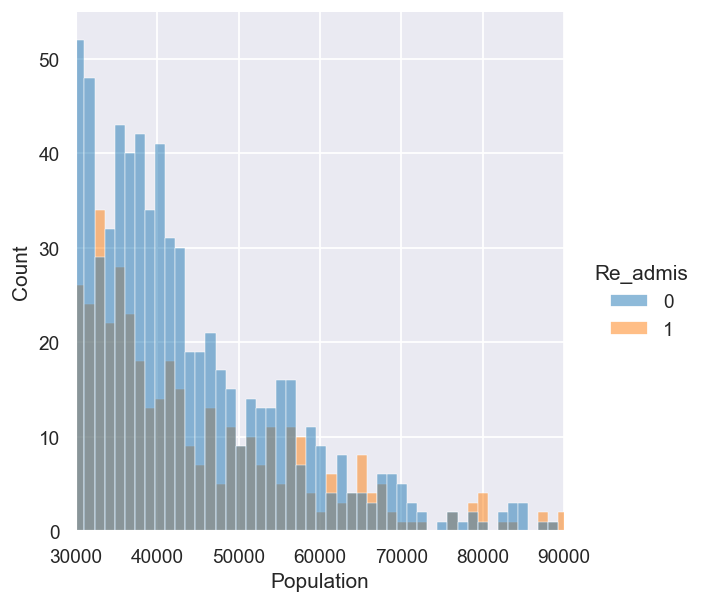
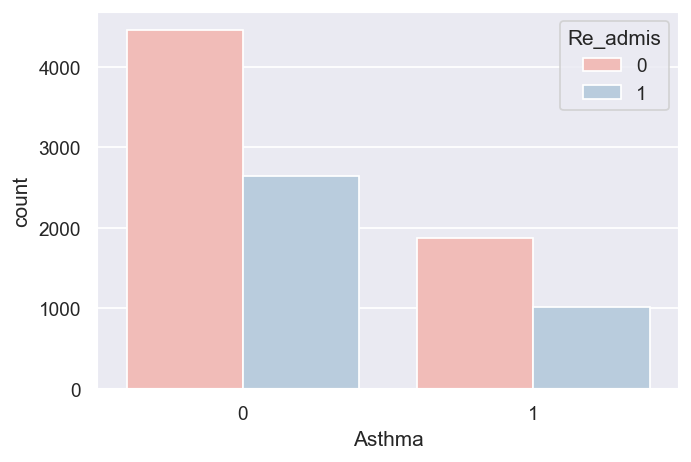


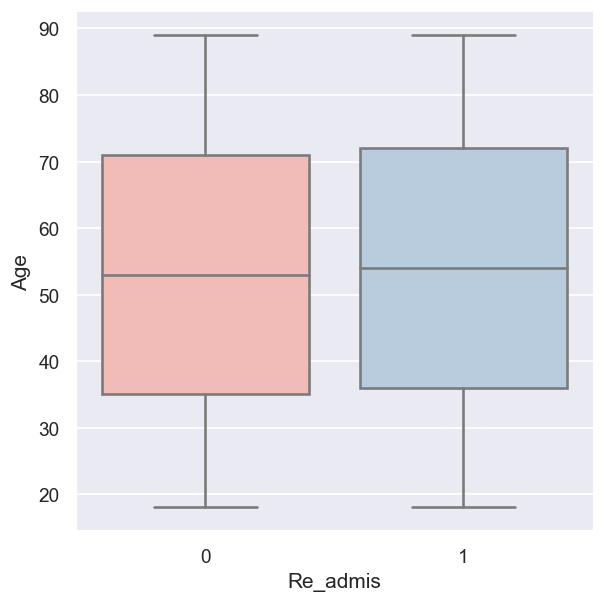
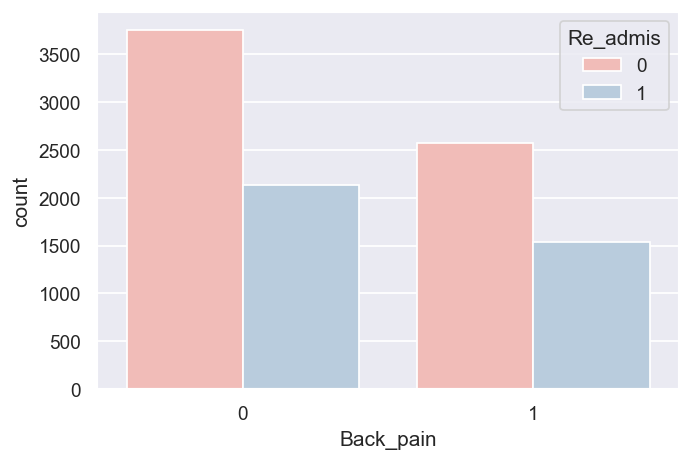


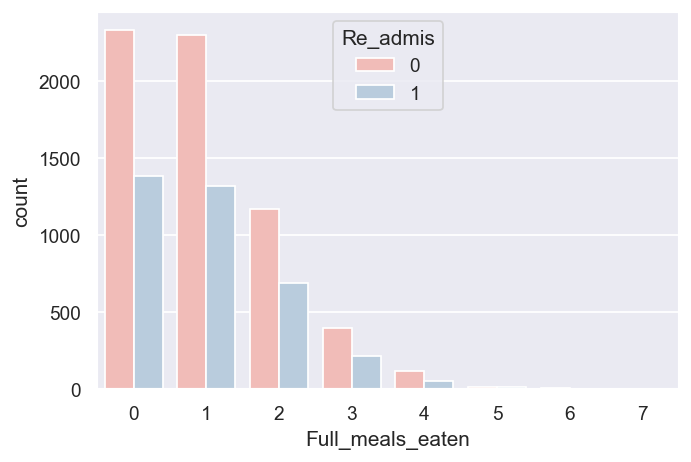
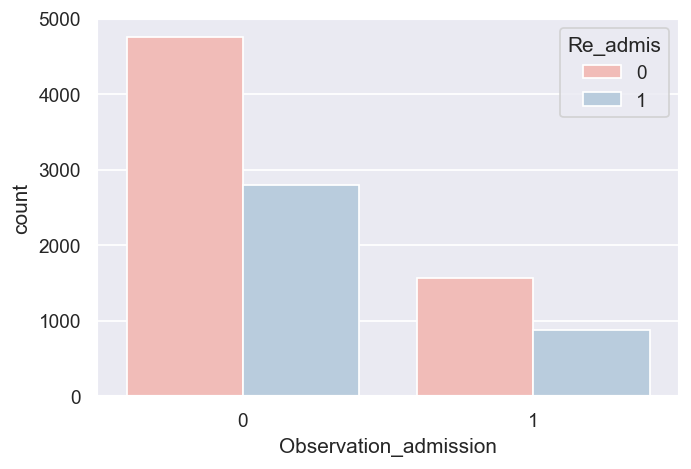
Bivariate graphs:

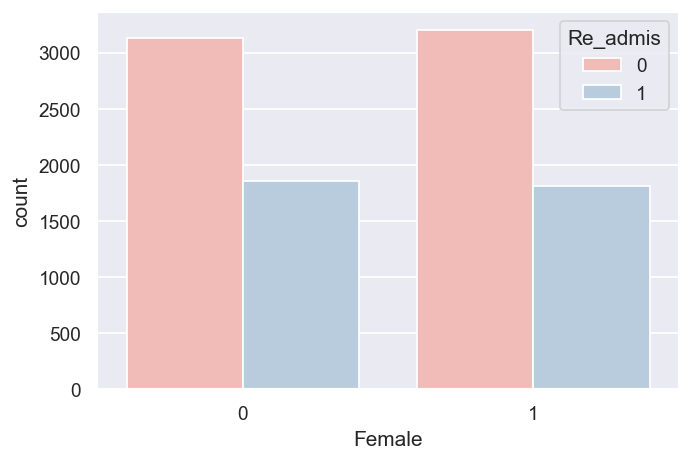
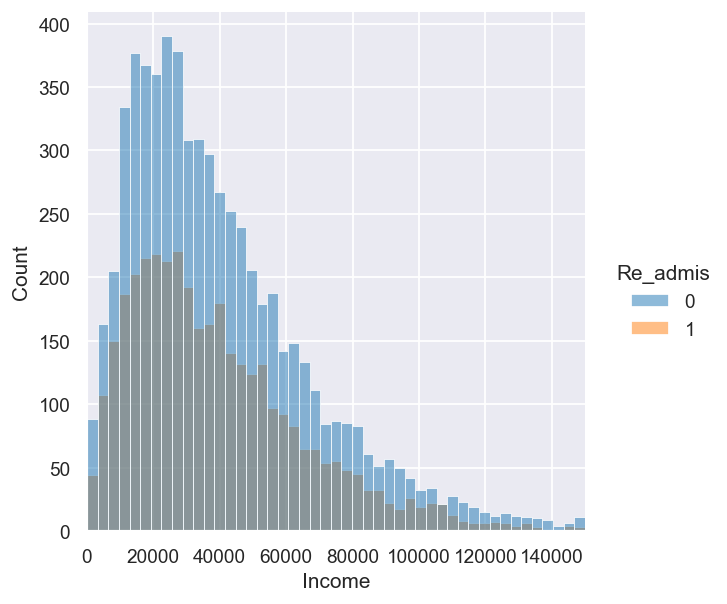


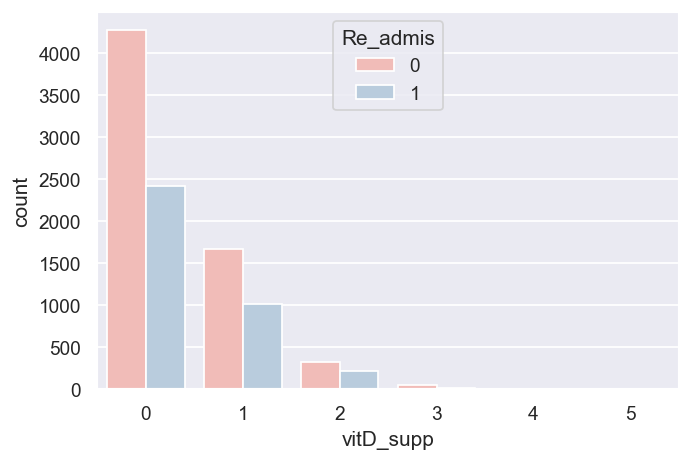


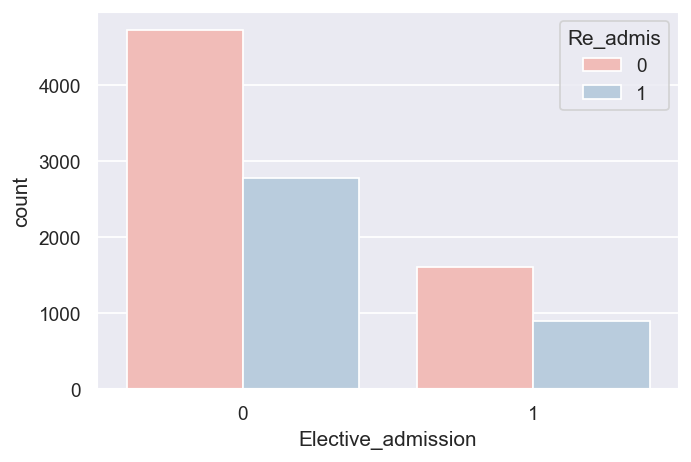






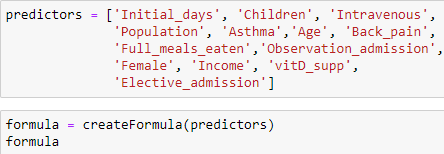
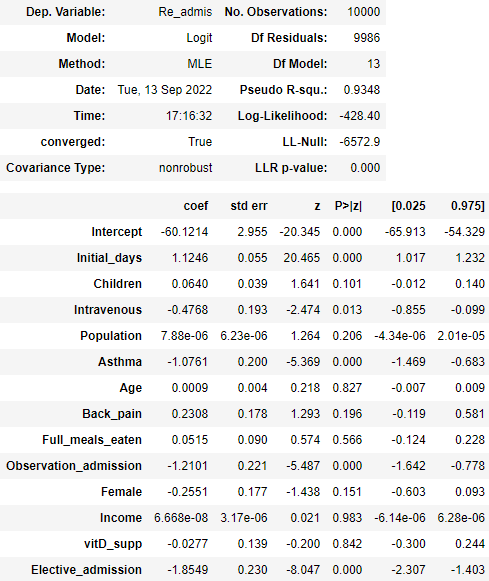
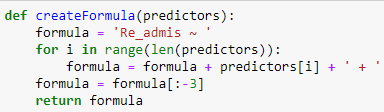






**D1: Initial Model**

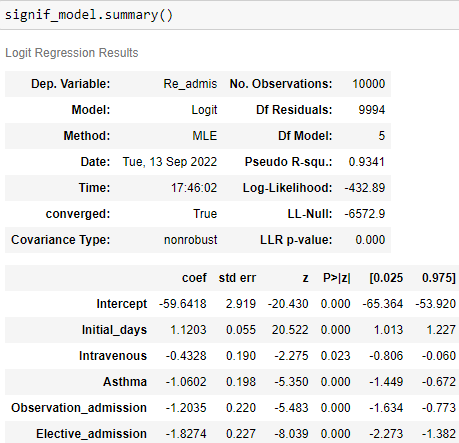
Using all of the predictor variables I identified in Part C, I constructed the initial logistic regression model using the Statsmodels package. As you can see from the results, the model looks promising with a pseudo R-squared value of 0.9348 and the LLR p-value suggesting that the model is useful. The model contains several variables that are statistically insignificant and require removal.



**D2: Justification of Model Reduction**

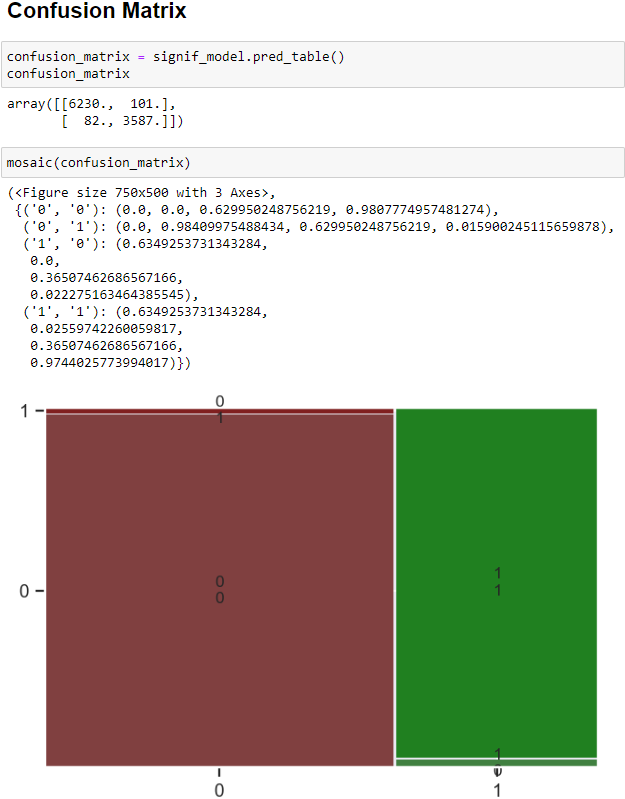
The initial model contains several variables that have p-values higher than the cutoff for significance, 0.05. To reduce the model, I removed the predictor with the highest p-value until only statistically significant predictors remained. To further reduce this model, I decided to perform backward selection, where I start off with the full model and one-by-one I remove a single variable and compare the resulting models, keeping the model with the greatest improvement in AIC. I would repeat this process until there are no improvements made from removing a variable. The reason why I chose AIC as the model evaluation metric for my backward selection process is because according to Çetinkaya-Rundel & Hardin (2021), AIC is a statistic commonly used for model comparison.

**D3: Reduced Logistic Regression Model**



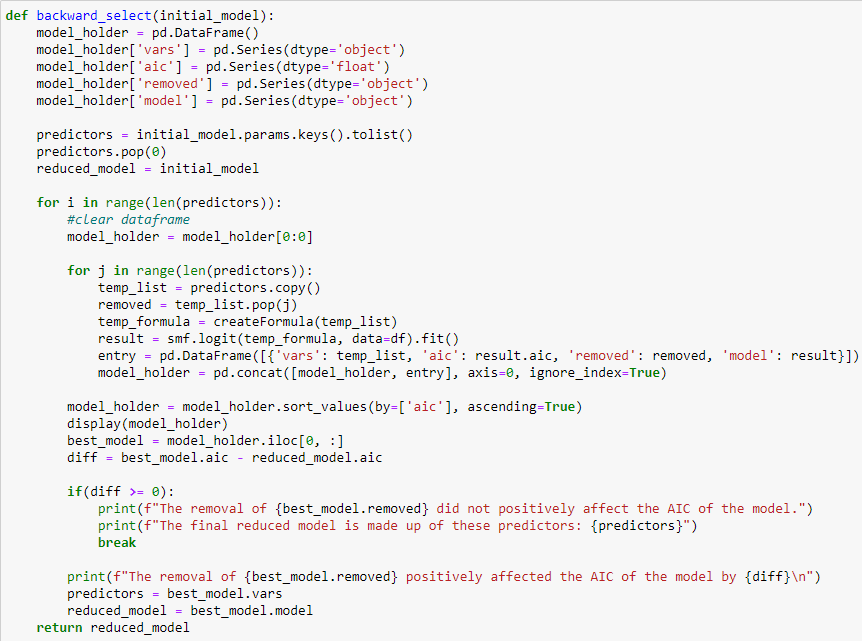
**E1: Model Comparison**

My model reduction process resulted in a reduced model containing the five predictors ‘Initial\_days’, ‘Intravenous’, ‘Asthma’, ‘Observation\_admission’, and ‘Elective\_admission’. This caused an improvement of the model from an initial AIC of 884.81 to 877.78. The reduction of predictors from 13 variables to 5 resulted in a very minor decrease in the pseudo R-squared value of 0.0007. Most importantly, the model now contains only statistically significant predictors and the LLR p-value of the reduced model suggests that the new model is useful.

**E2: Output and Calculations**

**E3: Code**

Here is the code I used to perform backward selection to further reduce my model.



**F1: Results**

The regression equation for my reduced model is:

According to the model, for each day the patient stays for in their initial visit, the odds that they are later readmitted increases by about 206%! If the primary service that the patient received was intravenous, then the chance of the patient being readmitted is reduced by about 35% compared to patients who received other primary services. If the patient has asthma then the odds that they are later readmitted to the hospital is reduced by about 65%. If the patient was admitted for observation, the odds that they are readmitted later on is 70% less than those who were admitted under emergency and if the patient's initial admission was elective, the odds of readmission is 84% less than emergency admission patients. When all of these variables are 0, the probability that the patient will be readmitted within a month is very close 0.

The LLR p-value of the model and the p-value of the individual predictors suggest that model is statistically meaningful. Though practically, I would say that it was already fairly obvious that those who were admitted to the hospital electively or for observation would face a lower risk of readmission than those were admitted under emergency. The model could be used to predict whether a patient is going to be readmitted within a month, but it is limited because of the fact that other than ‘Initial\_days’, none of the other predictors have a linear relationship with the target variable.

**F2: Recommendations**

Based on my results, my recommended course of action would be to closely monitor patients that are staying in the hospital for an extended period of time. The reason for their admission was likely serious and there might be complications during their treatment that are leading to them being readmitted at a higher rate than those who stayed for only a short time.

**G: Panopto Demonstration**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0988ac6a-7785-4d1e-bf74-af120159963d>

**H & I: Sources**

*Assumptions of logistic regression*. Statistics Solutions. (2021, August 11). Retrieved September 14, 2022, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

Çetinkaya-Rundel Mine, & Hardin, J. (2021, June 12). *Introduction to modern statistics*. OpenIntro.