Predicting Initial Hospital Stay Length (Multiple Linear Regression)

Ashley Traore

Research question:

What factors influence the patient's initial days spent in the hospital most significantly?

Objectives and Goals for Analysis:

My analysis aims to gain greater insight into determining what patient factors correlate to the initial number of days the patient stays in the hospital. The company can use this to predict the length of the hospital stay based on their symptoms upon admission and make better business decisions.

Assumptions:

Four key assumptions that a multiple linear regression model operates under are: Firstly, it assumes a linear relationship between the dependent variable and each independent variable. Secondly, it assumes that the independent variables are not highly correlated, a condition known as multicollinearity, which can create challenges in determining the specific variable contributing to the variance in the dependent variable. Thirdly, it assumes homoscedasticity, meaning that the error in the residuals is consistent across all points of the linear model. Lastly, it assumes that each observation is independent of the others. (Taylor, 2024)

Programming Language and Benefits:

Python was chosen as the programming language for this project for two key reasons. Firstly, Python has simple and consistent syntax, which facilitates easier coding and debugging. Secondly, Python has a large and active community, providing an abundance of resources and tutorials for support and knowledge sharing.

Several modeling and data analysis libraries were utilized for this project:

- **Pandas** - The main library for data handling, it facilitates easy manipulation and cleaning of datasets with its powerful DataFrame structure. It was used to handle and prepare the data for analysis.

- **Scikit-learn** - A machine learning library. It provides efficient tools for data analysis and modeling. Specifically,the LabelEncoder, LinearRegression, and train_test_split classes were used. LabelEncoder was used to encode the categorical variables as numeric, LinearRegression for implementing the linear regression model, and train_test_split for splitting the data into training and testing sets.
- **Matplotlib/Seaborn** - Both libraries were used for visualizing the complex dataset, aiding in understanding the data through various charts and plots.
- **Numpy** - Essential for scientific computing, it efficiently supports large mult-dimensional arrays and matrices, along with a suite of mathematical functions to operate on them. It was used for numerical operations.
- **Statsmodels** - Provides classes and functions for the estimation of statistical models, and for conducting statistical tests. Specifically, the formula.api and api modules were used to create the OLS models. Statsmodels.stats.outliers_influence was utilized to check for multicollinearity.

Justification of using Linear Regression:

Multiple Linear Regression is the appropriate technique to answer the research question because as it quantifies the relationship between the dependent variable and multiple independent variables. This technique helps understand how changes in various factors influence the number of days a patient is hospitalized. It allows for the identification of significant factors, which can guide improvements in patient care. Since the dependent variable, initial_days, is continuous, Linear Regression is the most suitable for answering this research question.

Data Cleaning:

To clean the medical dataset, I identified and addressed nulls, outliers, and duplicates. There were no duplicates or nulls identified in the dataset, so no cleaning was needed for those aspects. However, outliers were present, which can significantly impact linear regression models in various ways, some of which include distorted results, inflated errors, and influence on slope and intercept. To mitigate these issues, I applied z-score capping with a threshold set at 3 standard deviations from the mean. This method is straightforward and easy to implement, efficiently capping extreme

values without removing them from the dataset, thus retaining as much data as possible.

Data Exploration (EDA):

	Income
count	10000.000000
mean	40490.495160
std	28521.153293
min	154.080000
25%	19598.775000
50%	33768.420000
75%	54296.402500
max	207249.100000

Shown in the screenshot above are the summary statistics of the continuous variable used in the analysis. Listed below are each of the calculated values and a corresponding description.

- **Count**: Total number of non-null values
- **Mean**: Average value
- **STD:** Standard deviations from the mean. This indicates how spread out the data is compared to the mean.
- **Min**: Minimum value
- **25%**: First quartile
- **50%**: Second quartile
- **75%**: Third quartile
- **Max**: Maximum value.

Summary statistics apply exclusively to quantitative variables, rendering the mean, median, or mode unsuitable for qualitative variables. Instead, for each categorical variable, I provide below the percentage distributions of every distinct value within each column.

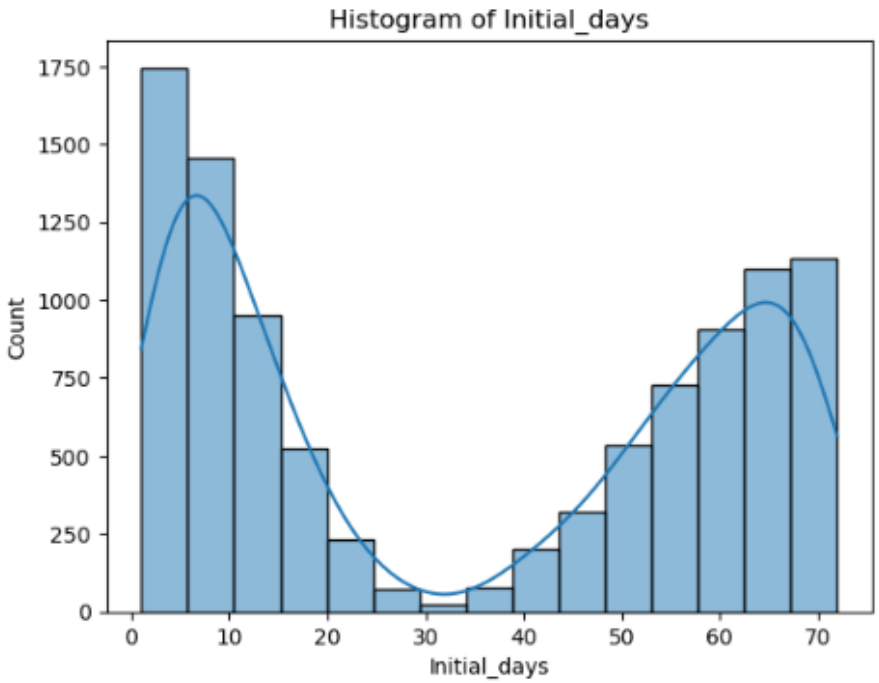
- Initial_admin:
    - Emergency Admission - 50.6%
    - Elective Admission - 25.04%
    - Observation Admission - 24.36%

- Stroke:
    - No - 80.07%
    - Yes - 19.93%
- Complication_risk:
    - Medium - 45.17%
    - High - 33.58%
    - Low - 21.25%
- Arthritis:
    - No - 64.26%
    - Yes - 35.74%
- Diabetes:
    - No - 72.62%
    - Yes - 27.38%
- Hyperlipidemia:
    - No - 66.28%
    - Yes - 33.72%
- Asthma:
    - No - 71.07%
    - Yes - 28.93%
- Anxiety:
    - No - 67.85%
    - Yes - 32.15%

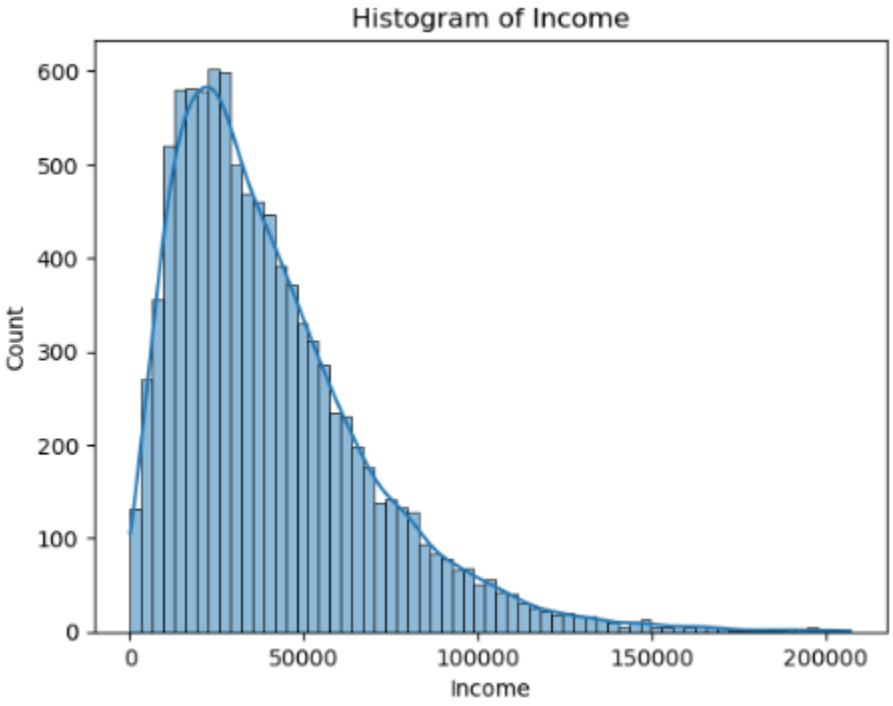
Visualizations:

Univariate visuals:

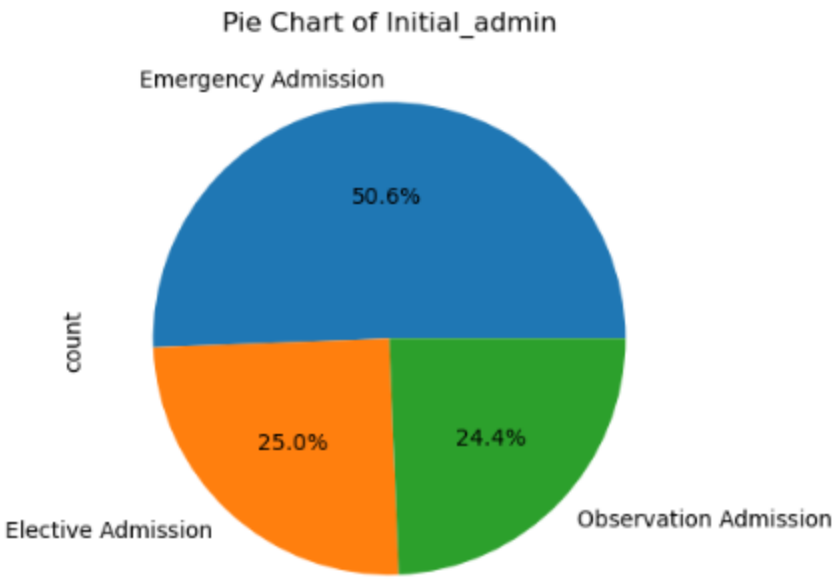
Initial Days:



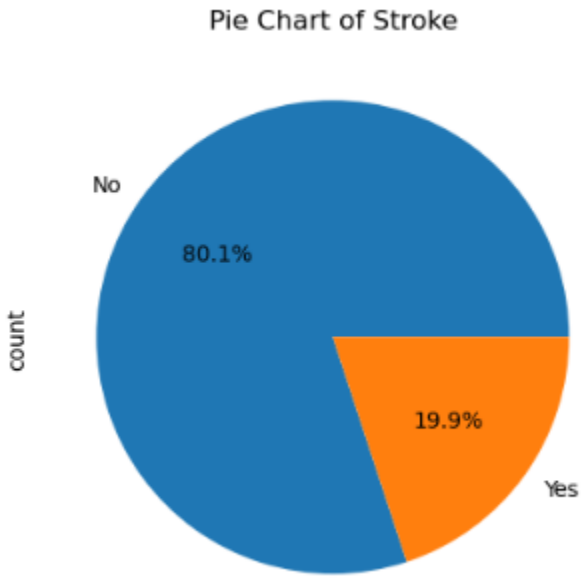
Income:



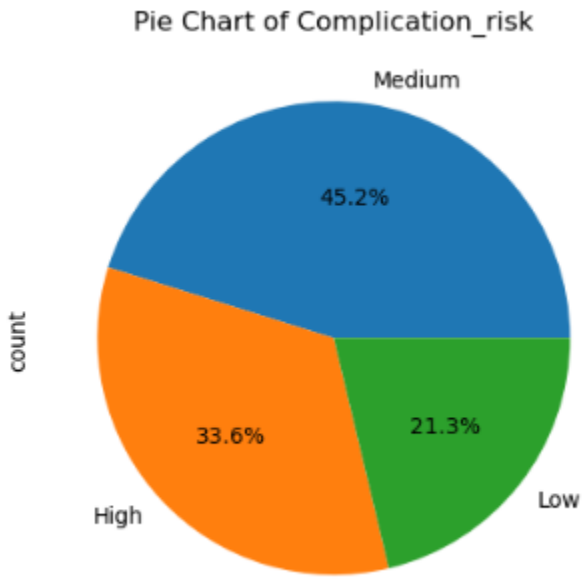
Initial_Admin:


Pie Chart of Initial_admin
Emergency Admission
50.6%
count
25.0%
24.4%
Elective Admission
Observation Admission

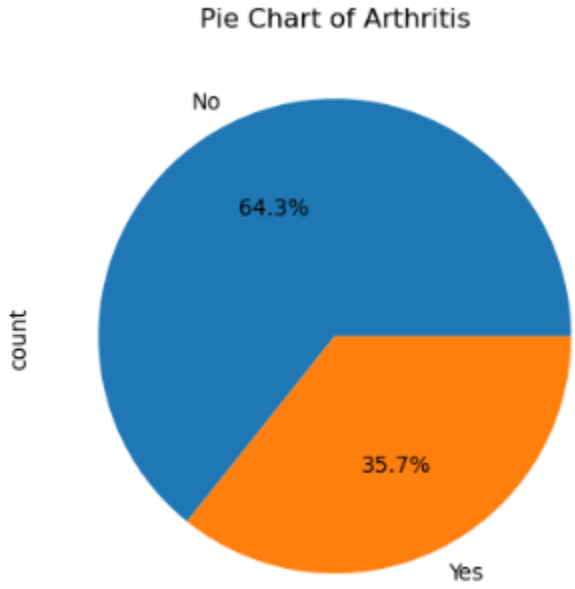
Stroke:


Pie Chart of Stroke
No
80.1%
count
19.9%
Yes

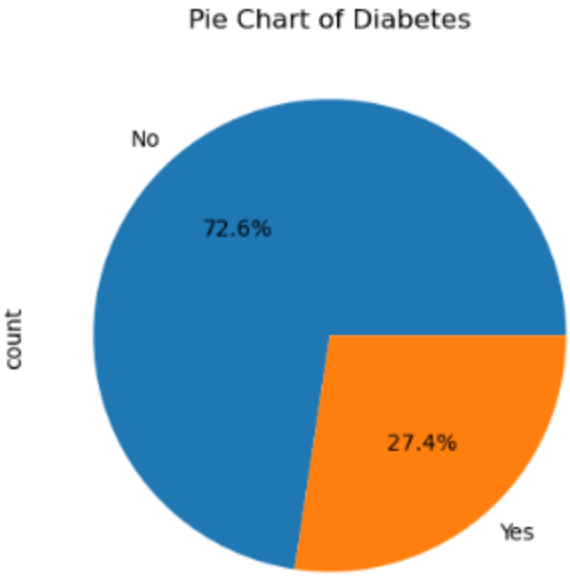
Complication Risk:



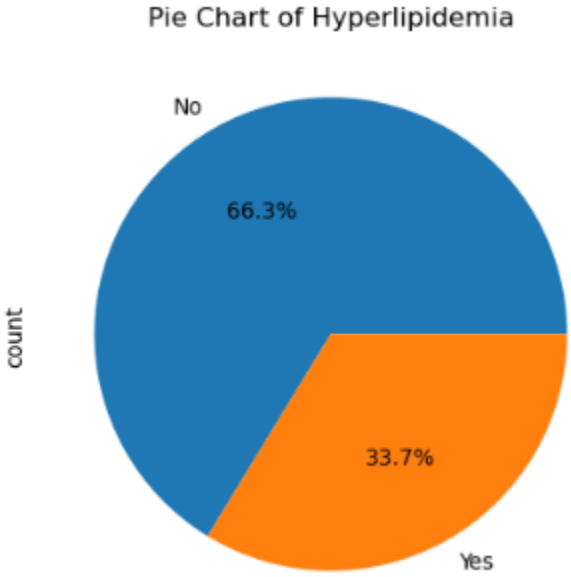
Arthritis:



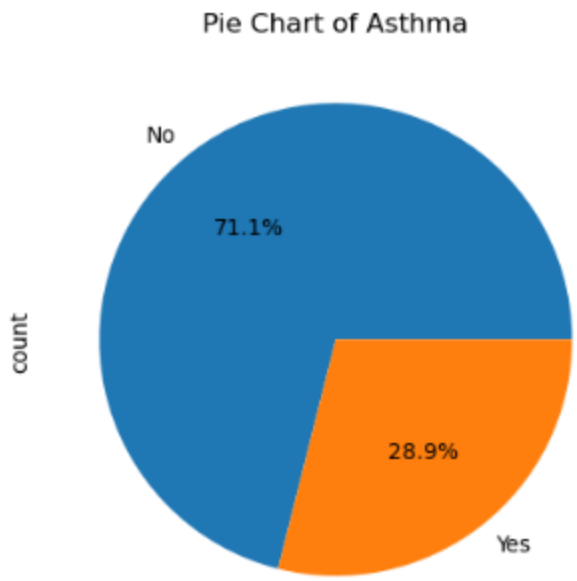
Diabetes:



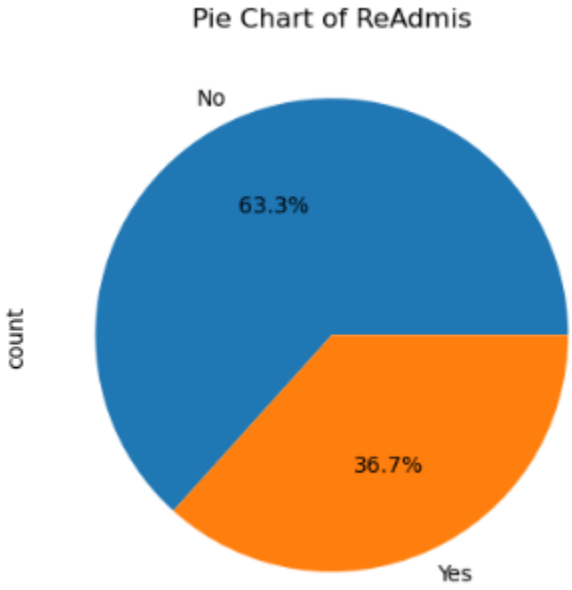
Hyperlipidemia:



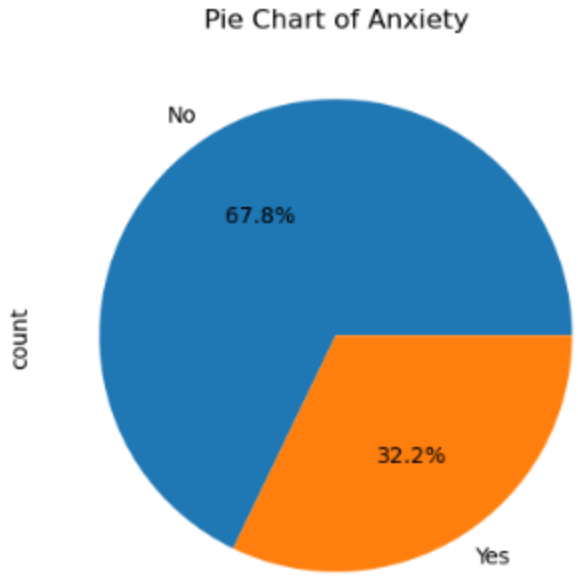
Asthma:



ReAdmis:

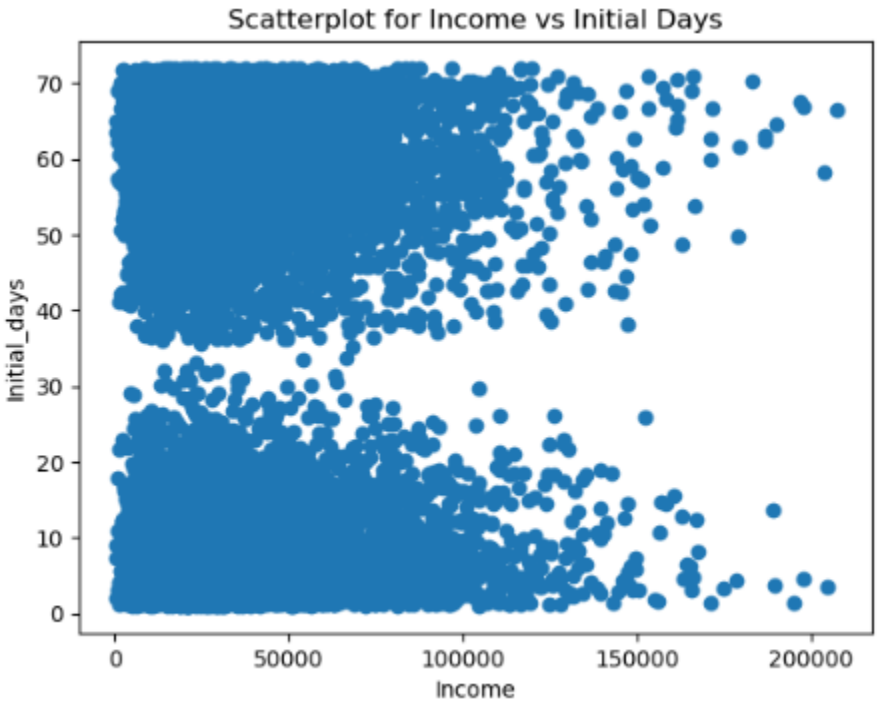


Anxiety:

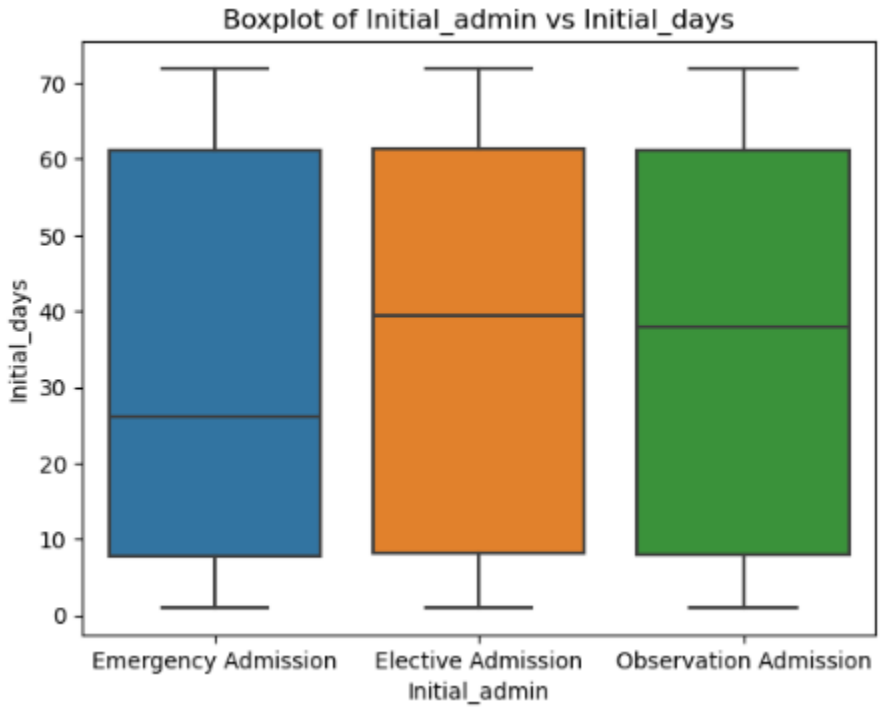


Bivariate Visuals:

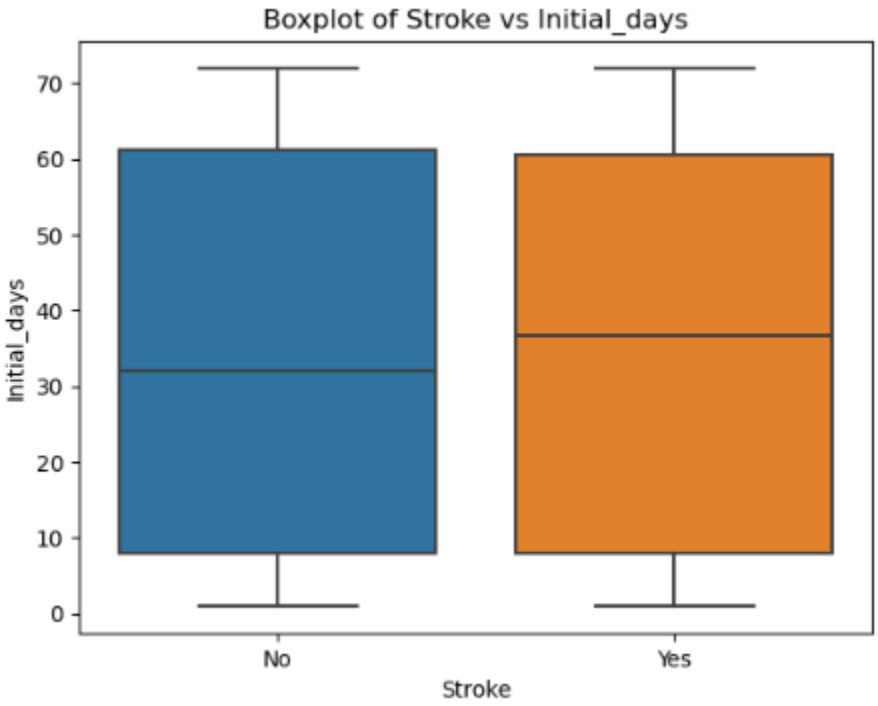
Income:



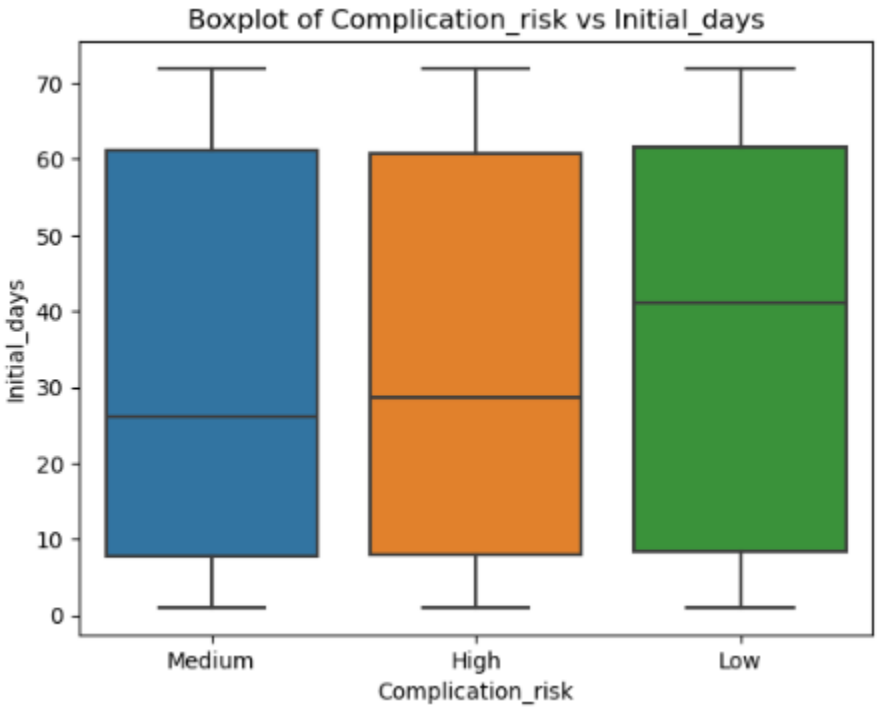
Initial Admin:



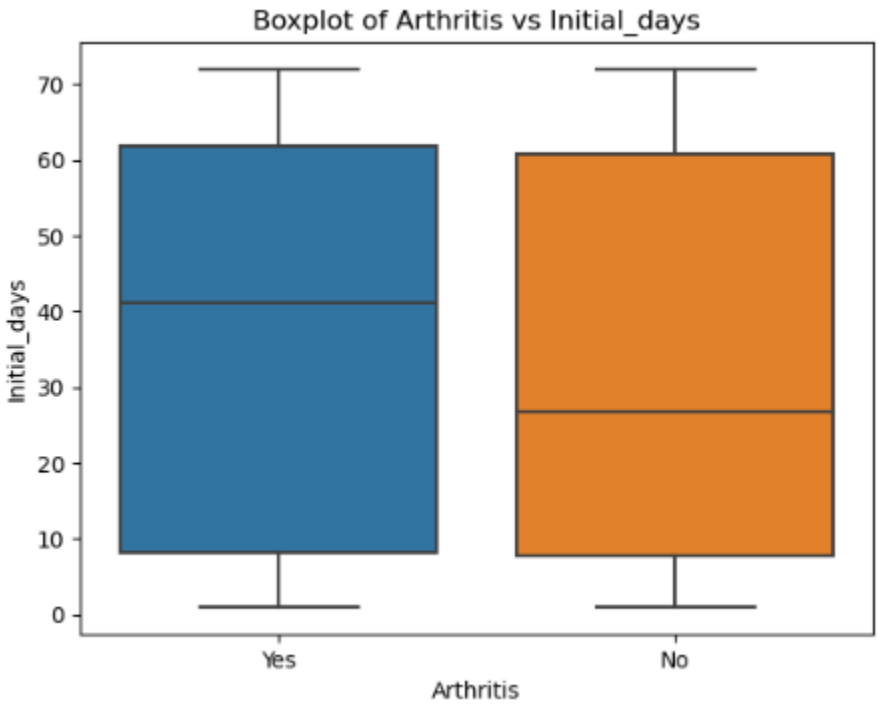
Stroke:



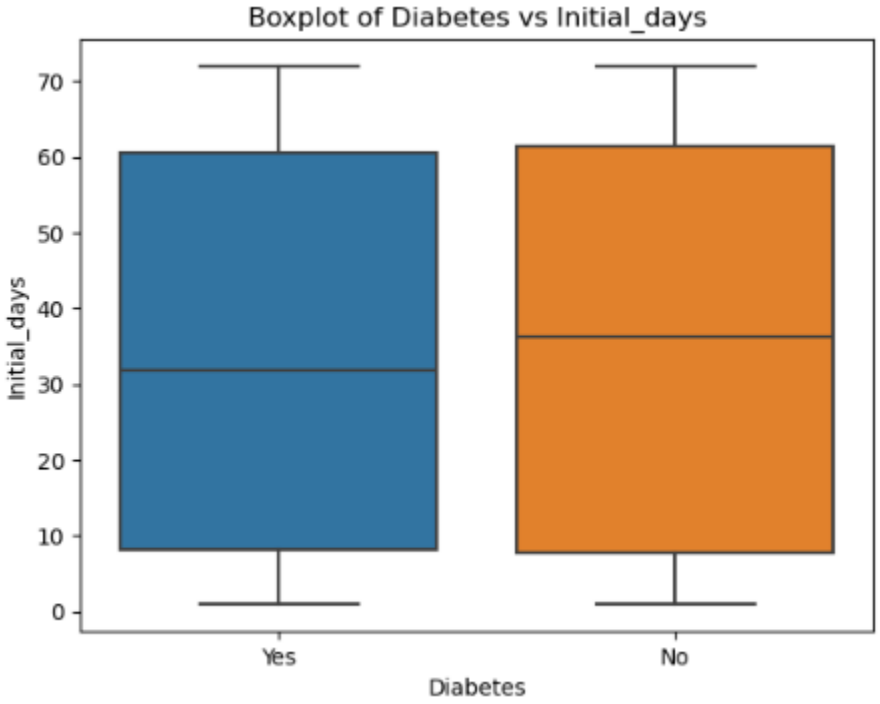
Complication Risk:



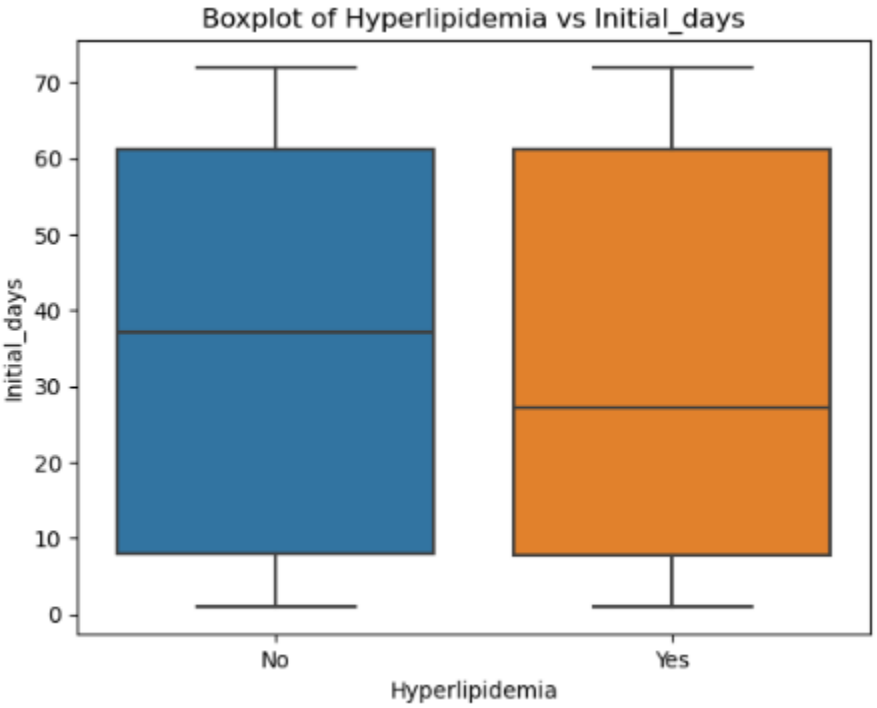
Arthritis:



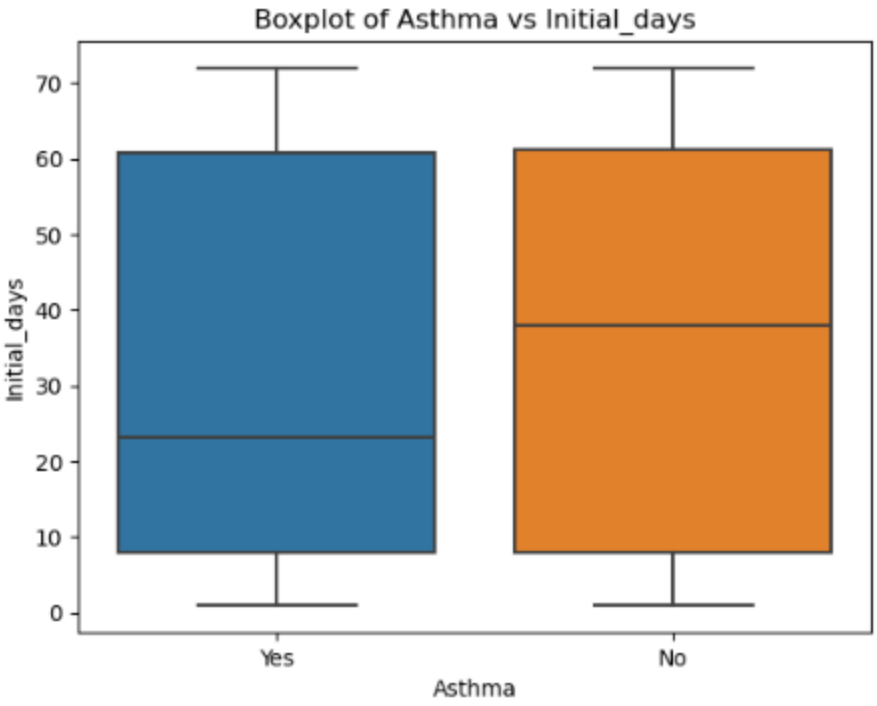
Diabetes:



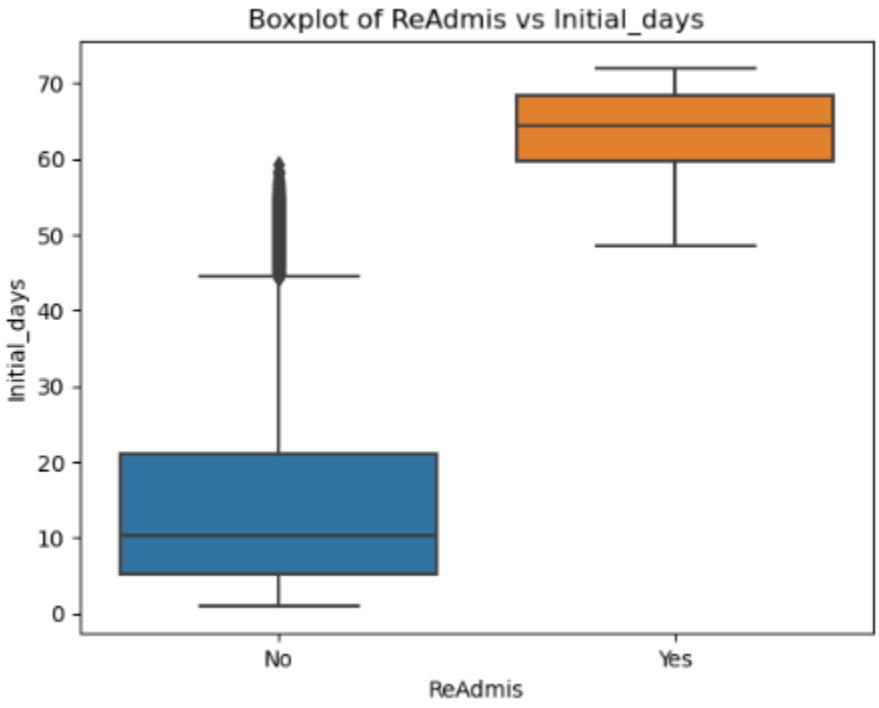
Hyperlipidemia:



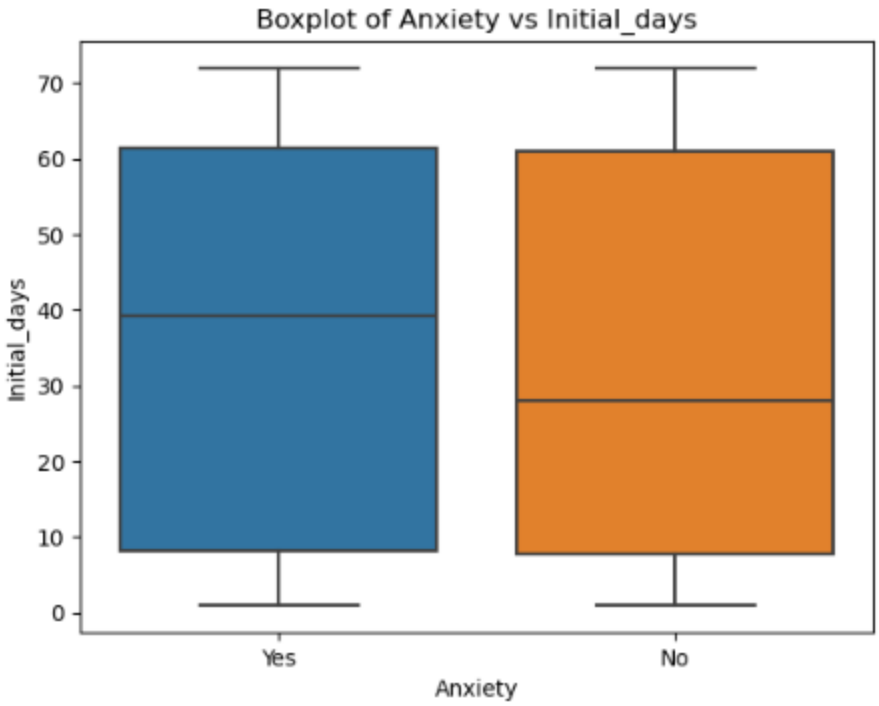
Asthma:



ReAdmis:



Anxiety:



Data Transformation (Data Wrangling):

Linear regression models require all variables to be numeric. Therefore, I converted categorical variables into numerical values. For yes/no variables, I employed a loop and the replace() function to transform them into 1/0. For other categorical variables, I utilized the LabelEncoder() function to assign numeric values to each category.

Refer to the code below and the attached “PA Task 1 - D208 Script” document for more details.

```
# ## Data Wrangling

# In[42]:

# Re-express all yes/no categorical variables to be expressed as 1/0.

# Create a variable for categorical columns that need to be replaced.

yes_no_Columns = ['ReAdmis', 'Soft_drink', 'HighBlood', 'Stroke', 'Arthritis', 'Diabetes',
'Hyperlipidemia', 'BackPain',
                  'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma', 'Overweight', 'Anxiety']
```

```
# Loop through each column and use replace to change yes/no values to 1/0.
for i in yes_no_Columns:
    medicalClean[i].replace(to_replace =['Yes','No'], value = [1,0], inplace = True)


# Re-express initial admin reasons and complication risk level to be expressed numerically.
label_encoder = LabelEncoder()
encodedLabelsInitialAdmin = label_encoder.fit_transform(medicalClean['Initial_admin'])
medicalClean['Initial_admin'] = encodedLabelsInitialAdmin
encodedLablesCompRisk = label_encoder.fit_transform(medicalClean['Complication_risk'])
medicalClean['Complication_risk'] = encodedLablesCompRisk
```

Initial Model:

Before performing multiple linear regression, it is important to check for multicollinearity. Multicollinearity occurs when independent variables, also known as predictor variables, in a regression model, are highly correlated, which can distort the results of the model and make it difficult to determine the exact variable responsible for the change in the dependent variable. To detect multicollinearity, I utilized the Variance Inflation Factor (VIF) method. VIF measures the extent of variance inflation due to multicollinearity among the predictor variables in the model. (The Investopedia Team) A VIF values greater than 5 indicates multicollinearity; however, since all VIF scores for the independent variables used in the model are below 5, there is no multicollinearity issue.

	Variable	VIF
0	Income	2.467622
1	Initial_admin	2.426295
2	Stroke	1.215017
3	Complication_risk	2.222618
4	Arthritis	1.474117
5	Diabetes	1.325700
6	Hyperlipidemia	1.433073
7	Asthma	1.352791
8	ReAdmis	1.482308
9	Anxiety	1.405114

Having verified the absence of multicollinearity, the next step is to develop the initial model. This model will encompass all the variables previously mentioned. Below are the Ordinary Least Squares (OLS) regression results, the adjusted coefficient of determination, and the residual standard error from the initial model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.724
Model:                            OLS   Adj. R-squared:                  0.724
Method:                 Least Squares   F-statistic:                     2625.
Date:                Fri, 24 Jan 2025   Prob (F-statistic):               0.00
Time:                        11:53:29   Log-Likelihood:                -40445.
No. Observations:               10000   AIC:                         8.091e+04
Df Residuals:                    9989   BIC:                         8.099e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     17.5722      0.432     40.662      0.000      16.725      18.419
Income     -2.009e-06   5.07e-06     -0.396      0.692    -1.2e-05    7.94e-06
Initial_admin  -0.1346      0.197     -0.684      0.494      -0.520       0.251
Stroke         -0.1694      0.346     -0.489      0.625      -0.848       0.509
Complication_risk  -0.1836      0.157     -1.168      0.243      -0.492       0.125
Arthritis       0.6750      0.289      2.339      0.019       0.109       1.241
Diabetes        0.0093      0.310      0.030      0.976      -0.598       0.617
Hyperlipidemia  -0.4103      0.292     -1.403      0.161      -0.984       0.163
Asthma          0.0610      0.305      0.200      0.841      -0.537       0.659
ReAdmis        46.4409      0.287    161.906      0.000      45.879      47.003
Anxiety         0.5408      0.296      1.827      0.068      -0.039       1.121
==============================================================================
Omnibus:                     1985.615   Durbin-Watson:                   1.270
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3394.364
Skew:                           1.334   Prob(JB):                         0.00
Kurtosis:                       4.013   Cond. No.                     1.70e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.7e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Adjusted Coefficient of Determination: 0.7240561077299028

Residual Standard Error: 13.820392207894772
```

Model Reduction Method:

The feature selection procedure used to reduce the initial model is the Backward Stepwise Elimination procedure. This is a variable elimination approach that begins with a full model and gradually removes variables based on the p-value. For each step, an alpha-to-remove value is used to determine which variables to eliminate. An alpha-to-remove value of 0.1 was used for this project, allowing for a more lenient threshold during the elimination process. This method helps ensure that only the most

significant predictors remain in the final model, therefore enhancing its accuracy and interpretability. Additionally, Backward Stepwise Elimination simplifies the model, improves predictive performance by focusing on the most impactful variables, and provides an automated, unbiased approach to variable selection.

Reduced Model:

```
Removing Diabetes, p-value: 0.975968304703658
Removing Asthma, p-value: 0.8410422055359302
Removing Income, p-value: 0.6929823664612735
Removing Stroke, p-value: 0.6243173905058623
Removing Initial_admin, p-value: 0.4906199475512958
Removing Complication_risk, p-value: 0.23712772316789557
Removing Hyperlipidemia, p-value: 0.1575920998845058
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.724
Model:                            OLS   Adj. R-squared:                  0.724
Method:                 Least Squares   F-statistic:                     8750.
Date:                Fri, 24 Jan 2025   Prob (F-statistic):               0.00
Time:                        12:48:40   Log-Likelihood:                -40447.
No. Observations:               10000   AIC:                         8.090e+04
Df Residuals:                    9996   BIC:                         8.093e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         17.0004      0.222     76.499      0.000      16.565      17.436
Arthritis      0.6736      0.288      2.336      0.020       0.108       1.239
ReAdmis       46.4384      0.287    161.961      0.000      45.876      47.000
Anxiety        0.5471      0.296      1.849      0.064      -0.033       1.127
==============================================================================
Omnibus:                     1988.574   Durbin-Watson:                   1.269
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3402.208
Skew:                           1.335   Prob(JB):                         0.00
Kurtosis:                       4.016   Cond. No.                         2.99
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Adjusted Coefficient of Determination: 0.7241304840258929

Residual Standard Error: 13.818529548702335
```

Model Comparison:

The model evaluation metrics utilized to compare the initial model to the reduced model are Residual Standard Error (RSE) and Adjusted Coefficient of Determination (Adjusted R-squared). RSE is used to measure the error rate of a linear regression model. RSE quantifies the standard deviation of the residuals (errors), which are the differences between the observed values and those predicted by the model. In other

words, this value represents how much the predicted values deviate from the actual values. The lower the value, the better the model performs, which indicates less of an error rate. Adjusted R-squared is another version of R-squared that accounts for the number of predictor variables in the model. This is a value between 0 and 1, the higher the value the better the model performs. While you can't use Adjusted R-squared to explain variation you can use it to say if a model performs better than another. Below are the Adjusted R-squared and RSE values for the initial and reduced models. Based on the Adjusted R-squared values we can say that the reduced model performs slightly better than the initial model, as the Adjust R-squared value for the reduced model is slightly higher than the initial model's. The RSE for the reduced model is also slightly lower than the RSE for the initial model. This indicates that the reduced model performs better. Overall, the reduced model performs similarly to the initial model. This means that the simplified model captures the essential information without unnecessary complexity.

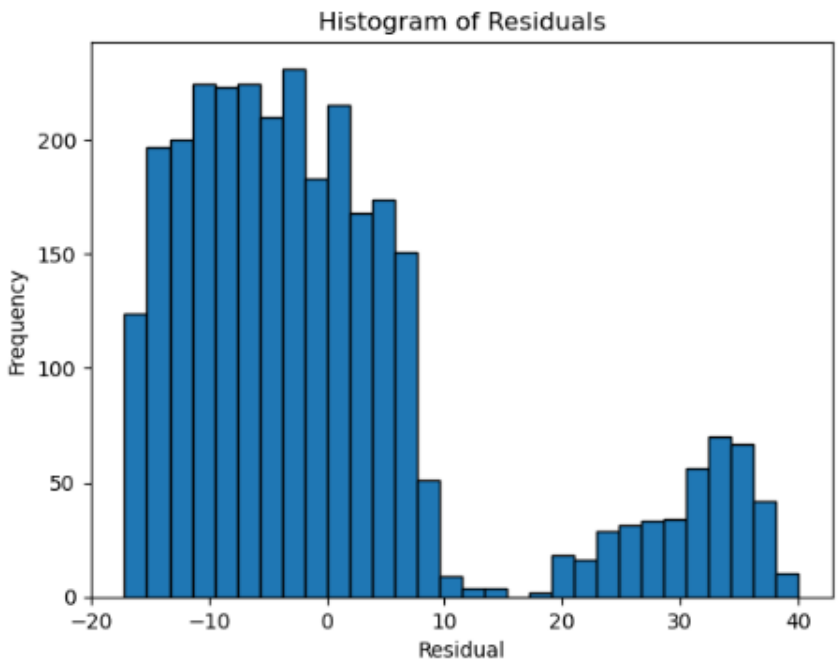
```
Adjusted Coefficient of Determination:
        Initial Model:0.7240561077299028
        Reduced Model: 0.7241304840258929

         Difference: 7.437629599005291e-05

Residual Standard Error:
        Initial Model: 13.820392207894772
        Reduced Model: 13.818529548702335

         Difference: 0.0018626591924366664
```

Residual Plot and RSE for Reduced Model:




```
Residual Standard Error: 13.818529548702335
```

Regression Equation and Interpretation of Coefficients:

Y = 17 + 0.67 (Arthritis) + 46.43 (ReAdmis) + 0.54 (Anxiety)

There are three coefficients in the reduced model. The bullet points below provide an interpretation for each coefficient.

- **Arthritis Coefficient** = 0.67 - This means that for each unit increase in the arthritis predator variable, the initial days spent in the hospital increase by 0.67 days, holding all other variables constant.
- **ReAdmis Coefficient** = 46.43 - This means that being readmitted to the hospital is associated with an increase of approximately 46.43 days spent in the hospital, compared to patients who were not readmitted when holding all other variables constant.
- **Anxiety Coefficient** = 0.54 - This means that having anxiety is associated with an increase of approximately 0.54 days spent in the hospital, compared to patients who do not have anxiety, when holding all other variables constant.

Statistical Significance:

Yes, the model is statistically significant. The model has a high F-statistic value and low p-values for each independent variable. This means that they have a significant effect on the dependent variable, and the model reliably explains the variation in the data.

Practical Significance:

Yes, the model is practically significant. The coefficient of 46.43 associated with readmission means that readmitted patients stay significantly longer in the hospital, this leads to higher operational costs and more resources being used. Additionally, patients with anxiety and arthritis require longer hospital stays. By addressing these factors, the hospital can implement targeted strategies to manage arthritis and anxiety to reduce readmission rates. This will improve patient satisfaction, enhance operational efficiency, and reduce expenses.

Disadvantages of Methods Used:

While linear regression is straightforward and easy to understand, it assumes a linear relationship between variables which often doesn't hold true, leading to inaccurate predictions. Backward stepwise elimination reduces the number of independent variables by iteratively removing the least significant ones, but it can lead to the exclusion of variables that may be important in combinations, which can result in an overly simplistic model. It is also prone to overfitting if not carefully managed, meaning it memorizes the data set instead of learning patterns. Z-score capping of outliers is a technique to mitigate the impact of extreme values. However, this can lead to the loss of valuable information by limiting the range of the data. Capping outliers can skew the results of an analysis, which can give an inaccurate representation of the data distribution.

Recommendations:

Based on the linear regression analysis, here are a couple of recommendations for the hospital.

1. Develop specialized care plans for patients with arthritis or anxiety to manage their condition more effectively and reduce their hospital stay.

2. Allocate resources more effectively to patients who are more likely to have long hospital stays, these being patients who have been readmitted or have anxiety or arthritis.

Sources:

Taylor, S. (2024, July 10). *Multiple linear regression*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/

The Investopedia Team. (n.d.). *Variance inflation factor (VIF)*. Investopedia. https://www.investopedia.com/terms/v/variance-inflation-factor.asp

Ajitesh KumarI have been recently working in the area of Data analytics including Data Science and Machine Learning / Deep Learning. I am also passionate about different technologies including programming languages such as Java/JEE. (2023, December 11). *F-Test & F-statistics in linear regression: Formula, examples*. Analytics Yogi. https://vitalflux.com/interpreting-f-statistics-in-linear-regression-formula-examples/