PART I: RESEARCH QUESTION

A1. RESEARCH QUESTION

What factors contribute significantly to the variability in additional charges incurred by the patients?

A2. GOALS

The objective of this analysis is to improve patient financial planning. Patients benefit from knowing what factors contribute to additional charges. This can help them make informed decisions about their healthcare and plan for potential expenses. It enhances transparency and reduces financial surprises for patients.

PART II: METHOD JUSTIFICATION

B1. SUMMARY OF ASSUMPTIONS

There are several assumptions for multiple linear regression. These assumptions are needed for the model to be valid and for the statistical inferences to be accurate. These are the four key assumptions to keep in mind.

Firstly, the relationship between the independent and dependent variables is assumed to be linear. The changes in the independent variables are associated with the constant change in the dependent variable. Secondly, the residuals should be independent of each other. The residual value for one observation should not be related to the residual for further observation. Thirdly, the variance of the residuals should be constant across all levels of the independent variables. Lastly, residuals should have a normal distribution. This is especially essential for smaller sample sizes to ensure the validity of statistical tests and confidence intervals.

B2. TOOL BENEFITS

For this analysis, I used R. This language has several data-cleaning features and capabilities.

Firstly, R has a vast ecosystem of packages designed explicitly for data cleaning. In this assessment, **naniar**, **dplyr,** and **plyr** packages were used. Naniar was used for finding missing data. Dplyr was used for data manipulation and analysis. Plyr was used to convert and revalue variables. All these packages were essential to accomplish a clean dataset.

Lastly, R has easy-to-use data visualization capabilities. Visualization can aid in identifying outliers, missing values, and patterns between variables.

B3. APPROPRIATE BENEFIT

Multiple linear regression is an appropriate technique to find the factors that affect additional charges incurred by patients for numerous reasons.

Firstly, my research question involves understanding the impact of multiple factors on additional charges. Multiple linear regression allows the manipulation of more than one independent variable to capture the complex relationships among various factors. Moreover, multiple linear regression can help identify which independent variables significantly contribute to the variability in additional charge. This is critical to explore the relative importance of different factors. This technique also allows for the assessment of variable importance. I can prioritize the most influential factors impacting additional charges by examining the coefficients and their significance.

PART III: DATA PREPARATION

C1. DATA CLEANING

Data cleaning is vital to data preparation. This ensures the quality, accuracy, and reliability of the data used for analysis. Firstly, find missing values. Ignoring missing data can lead to inaccurate conclusions and affect the study's trustworthiness. Then, find duplicates. Duplicates can introduce redundancy to the analysis. It can lead to overestimation of specific trends, which affects the reliability of insights. Lastly, count the outliers. Outliers can affect statistical measures. This leads to distorted results, affects the distribution of data, and leads to misleading conclusions.

To find the missing values, use the **naniar** package. **Naniar** provides valuable functions to visualize and handle missing data. Meanwhile, R makes it easier to find duplicates using **duplicated().** This function identifies duplicate rows. Lastly, outliers can be identified using the interquartile range. Values beyond the lower and upper bound are considered outliers.

The code for detection is attached.

C2. SUMMARY STATISTICS

Below is a screenshot of the summary statistics of the medical_clean dataset.

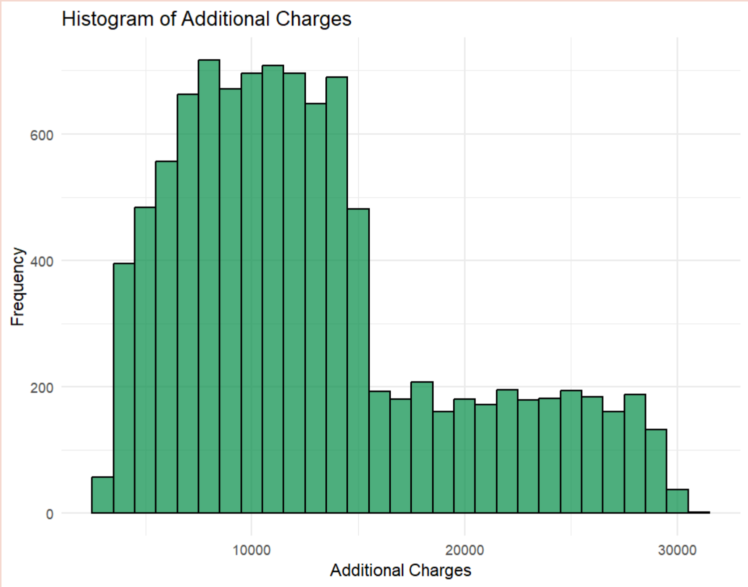
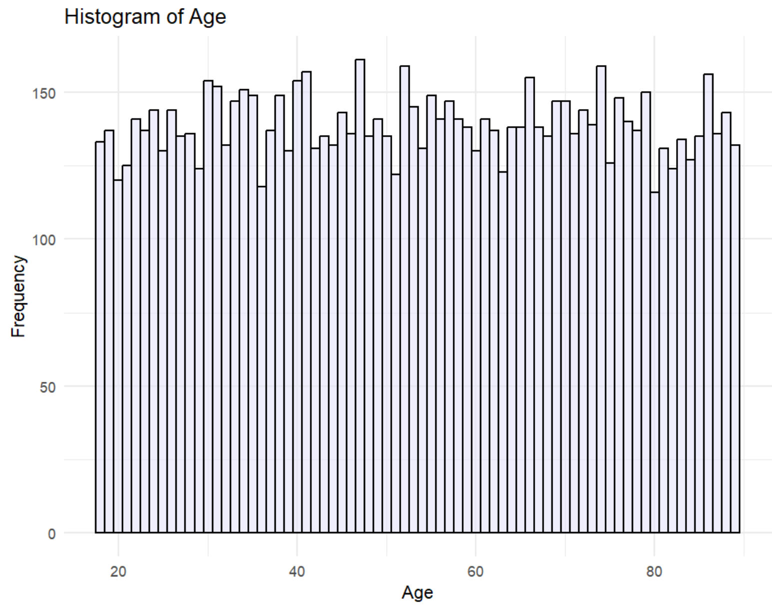
```
> summary(df)
   CaseOrder      Customer_id         Interaction            UID                City          
 Min.   :    1   Length:10000       Length:10000       Length:10000       Length:10000      
 1st Qu.: 2501   Class :character   Class :character   Class :character   Class :character  
 Median : 5000   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   : 5000                                                                              
 3rd Qu.: 7500                                                                              
 Max.   :10000                                                                              
    State              County               Zip             Lat             Lng         
 Length:10000       Length:10000       Min.   :  610   Min.   :17.97   Min.   :-174.21  
 Class :character   Class :character   1st Qu.:27592   1st Qu.:35.26   1st Qu.: -97.35  
 Mode  :character   Mode  :character   Median :50207   Median :39.42   Median : -88.40  
                                       Mean   :50159   Mean   :38.75   Mean   : -91.24  
                                       3rd Qu.:72412   3rd Qu.:42.04   3rd Qu.: -80.44  
                                       Max.   :99929   Max.   :70.56   Max.   : -65.29  
   Population          Area             TimeZone              Job               Children     
 Min.   :     0.0   Length:10000       Length:10000       Length:10000       Min.   : 0.000  
 1st Qu.:   694.8   Class :character   Class :character   Class :character   1st Qu.: 0.000  
 Median :  2769.0   Mode  :character   Mode  :character   Mode  :character   Median : 1.000  
 Mean   :  9965.2                                                            Mean   : 2.097  
 3rd Qu.: 13945.0                                                            3rd Qu.: 3.000  
 Max.   :122814.0                                                            Max.   :10.000  
      Age            Income            Marital             Gender             ReAdmis         
 Min.   :18.00   Min.   :   154.1   Length:10000       Length:10000       Length:10000      
 1st Qu.:36.00   1st Qu.: 19598.8   Class :character   Class :character   Class :character  
 Median :53.00   Median : 33768.4   Mode  :character   Mode  :character   Mode  :character  
 Mean   :53.51   Mean   : 40490.5                                                           
 3rd Qu.:71.00   3rd Qu.: 54296.4                                                           
 Max.   :89.00   Max.   :207249.1                                                           
  VitD_levels       Doc_visits    Full_meals_eaten   vitD_supp        Soft_drink       
 Min.   : 9.806   Min.   :1.000   Min.   :0.000    Min.   :0.0000   Length:10000      
 1st Qu.:16.626   1st Qu.:4.000   1st Qu.:0.000    1st Qu.:0.0000   Class :character  
 Median :17.951   Median :5.000   Median :1.000    Median :0.0000   Mode  :character  
 Mean   :17.964   Mean   :5.012   Mean   :1.001    Mean   :0.3989                     
 3rd Qu.:19.348   3rd Qu.:6.000   3rd Qu.:2.000    3rd Qu.:1.0000                     
 Max.   :26.394   Max.   :9.000   Max.   :7.000    Max.   :5.0000                     
```

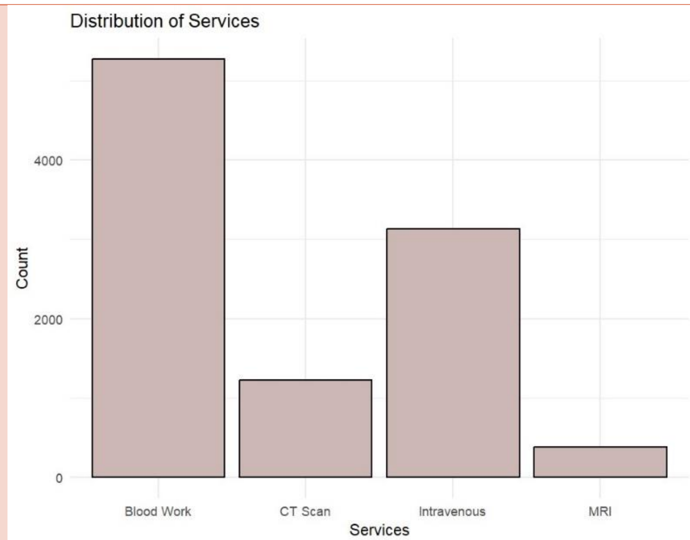
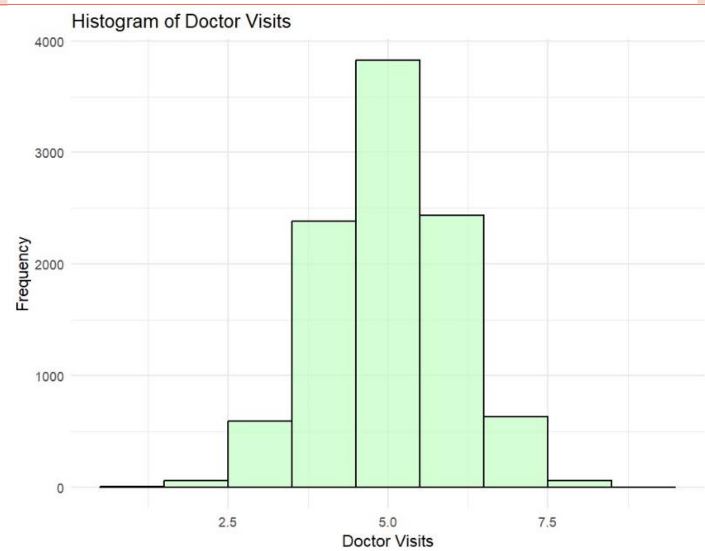
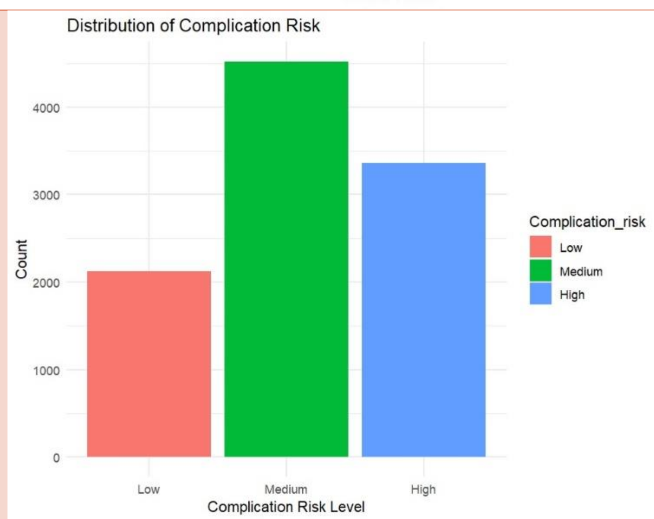
```
 Initial_admin        HighBlood            Stroke           Complication_risk   Overweight       
 Length:10000       Length:10000       Length:10000       Length:10000       Length:10000      
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                                                                                               
                                                                                               
                                                                                               
  Arthritis           Diabetes          Hyperlipidemia       BackPain           Anxiety         
 Length:10000       Length:10000       Length:10000       Length:10000       Length:10000      
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                                                                                               
                                                                                               
                                                                                               
 Allergic_rhinitis  Reflux_esophagitis    Asthma             Services          Initial_days    
 Length:10000       Length:10000       Length:10000       Length:10000       Min.   : 1.002  
 Class :character   Class :character   Class :character   Class :character   1st Qu.: 7.896  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :35.836  
                                                                             Mean   :34.455  
                                                                             3rd Qu.:61.161  
                                                                             Max.   :71.981  
  TotalCharge    Additional_charges     Item1           Item2           Item3      
 Min.   :1938   Min.   : 3126      Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3179   1st Qu.: 7986      1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000  
 Median :5214   Median :11574      Median :4.000   Median :3.000   Median :4.000  
 Mean   :5312   Mean   :12935      Mean   :3.519   Mean   :3.507   Mean   :3.511  
 3rd Qu.:7460   3rd Qu.:15626      3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :9181   Max.   :30566      Max.   :8.000   Max.   :7.000   Max.   :8.000  
     Item4           Item5           Item6           Item7           Item8      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00  
 1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.00  
 Median :4.000   Median :3.000   Median :4.000   Median :3.000   Median :3.00  
 Mean   :3.515   Mean   :3.497   Mean   :3.522   Mean   :3.494   Mean   :3.51  
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00  
 Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :7.00  
> 
```

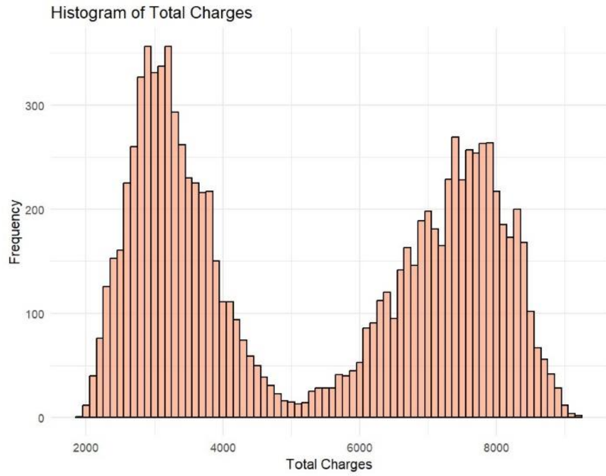
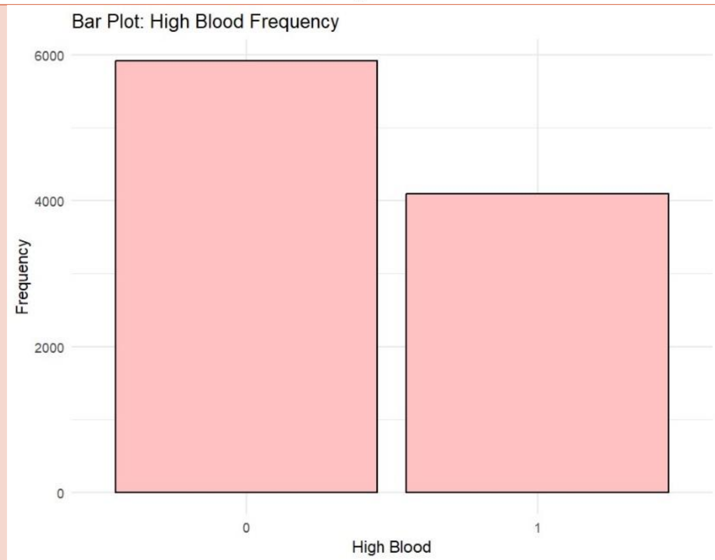
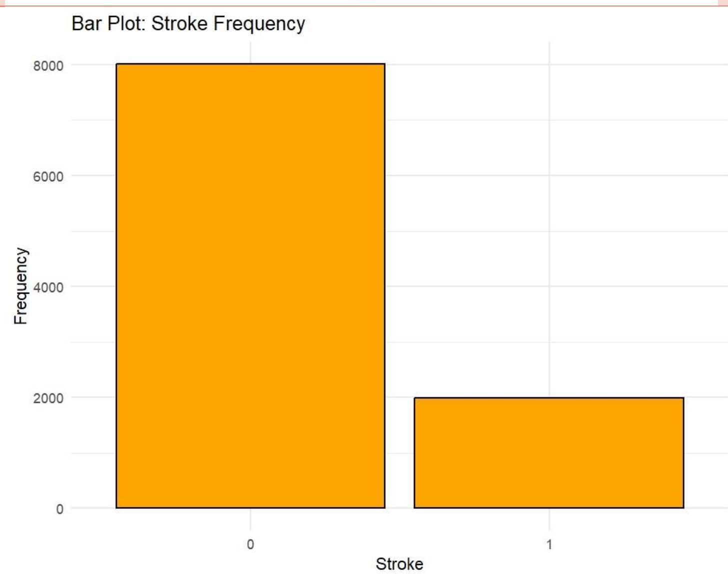
Variable	Data type	Variable type	Summary (for Categorical) *before data transformation
Dependent Variable: Additional Charges	Quantitative	Continuous	
Independent Variables:			
Age	Quantitative	Discrete	
Services	Qualitative	Categorical	Blood Work 5265, CT Scan 1225, Intravenous 3130, MRI 380
Doc visits	Quantitative	Discrete	
Complication risk	Qualitative	Categorical	High 3358, Low 2125, Medium 4517
Total charges	Quantitative	Continuous	
High blood	Qualitative	Categorical	No 5910, Yes 4090
Stroke	Qualitative	Categorical	No 8007, Yes 1993

Overweight	Qualitative	Categorical	No Yes 2906 7094
Arthritis	Qualitative	Categorical	No Yes 6426 3574
Diabetes	Qualitative	Categorical	No Yes 7262 2738
Hyperlipidemia	Qualitative	Categorical	No Yes 6628 3372
Back pain	Qualitative	Categorical	No Yes 5886 4114
Anxiety	Qualitative	Categorical	No Yes 6785 3215
Allergic rhinitis	Qualitative	Categorical	No Yes 6059 3941
Reflux esophagitis	Qualitative	Categorical	No Yes 5865 4135
Asthma	Qualitative	Categorical	No Yes 7107 2893

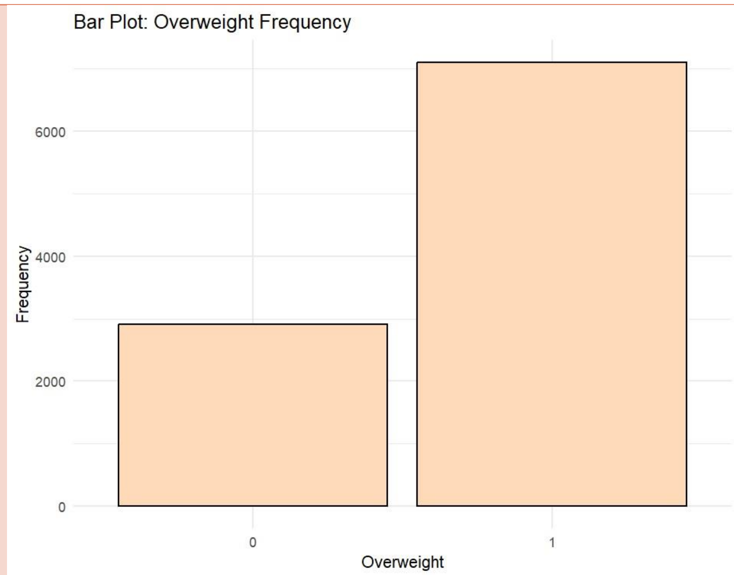
C3. VISUALIZATIONS

Variables	**Univariate Visualizations**
Dependent Variable: **Additional Charges**	
Age	

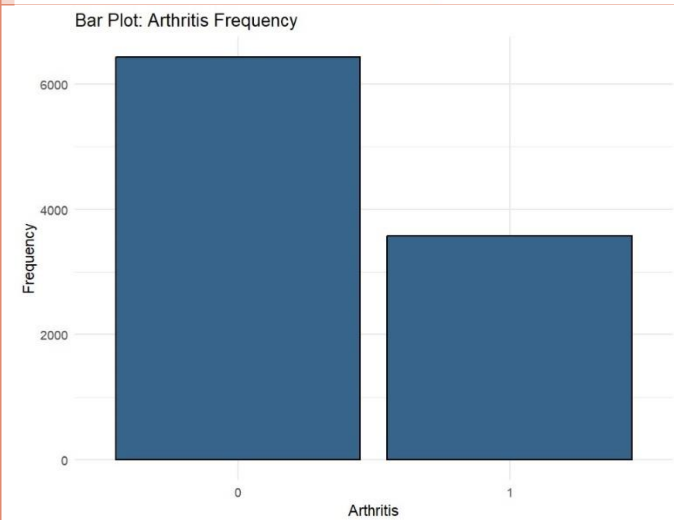
Services	
Doc Visits	
Complication Risk	

<table>
<tr><td>Total Charges</td><td>
</td></tr>
<tr><td>High Blood</td><td>
</td></tr>
<tr><td>Stroke</td><td>
</td></tr>
</table>

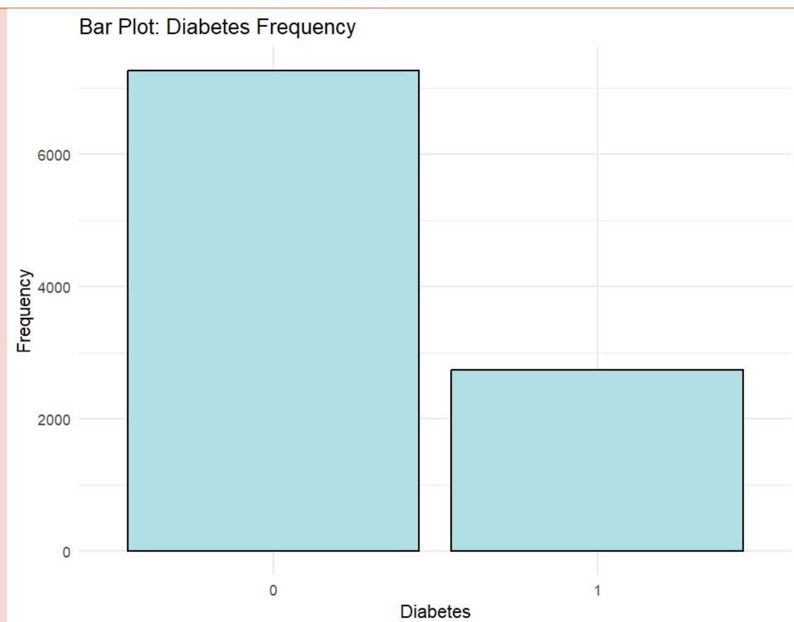
Overweight

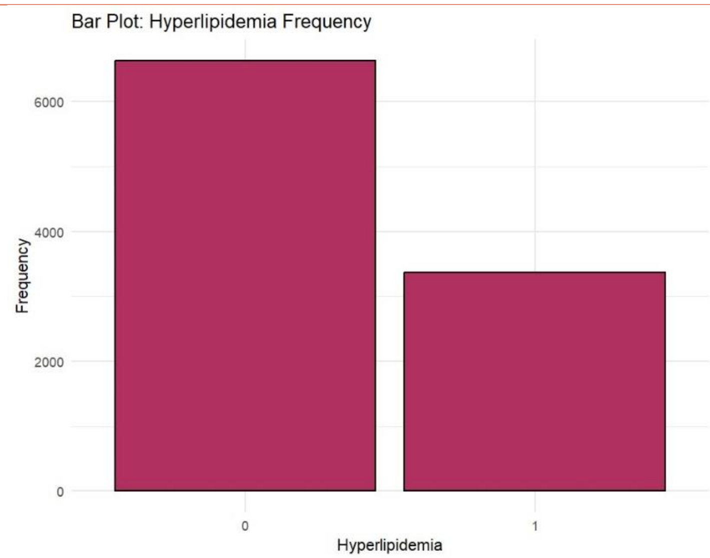
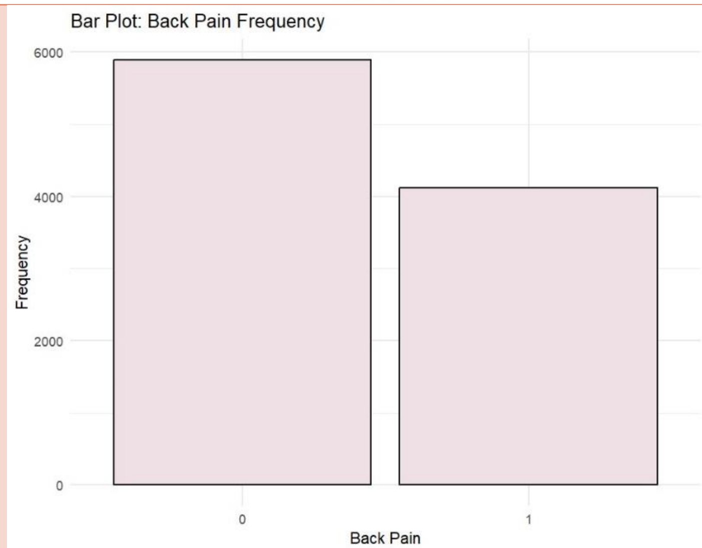
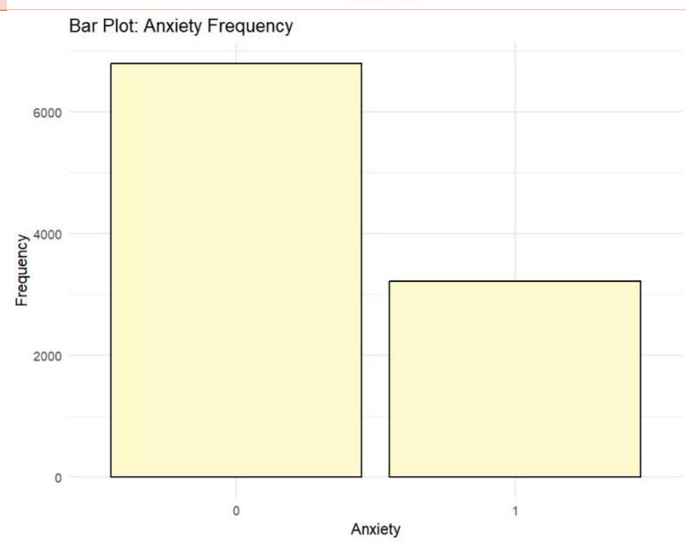


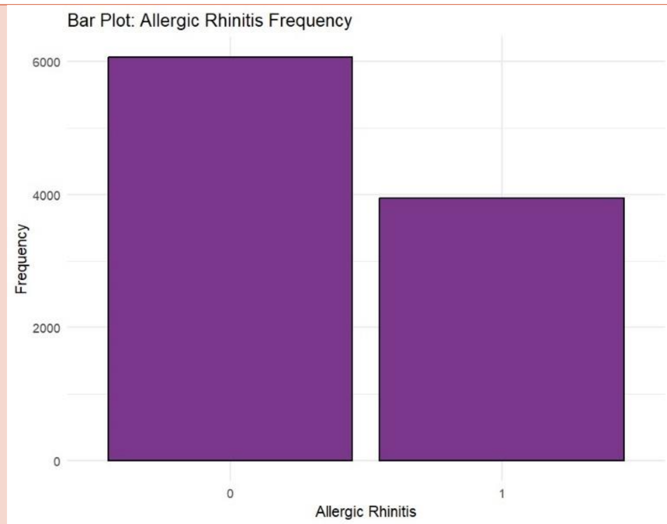
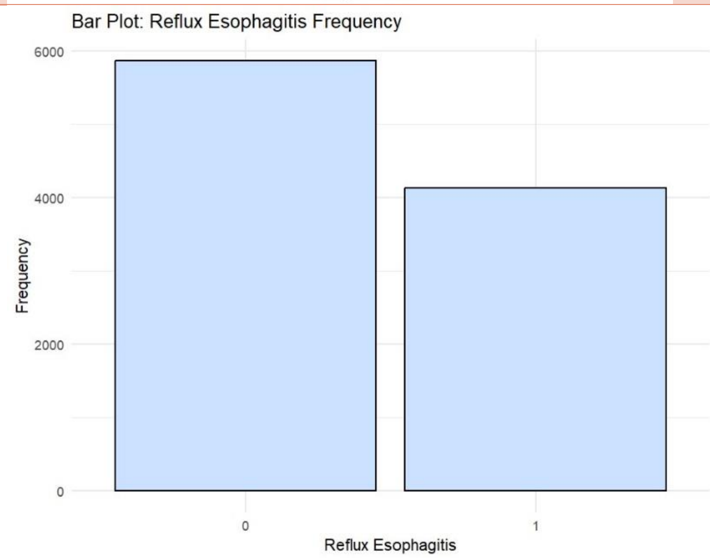
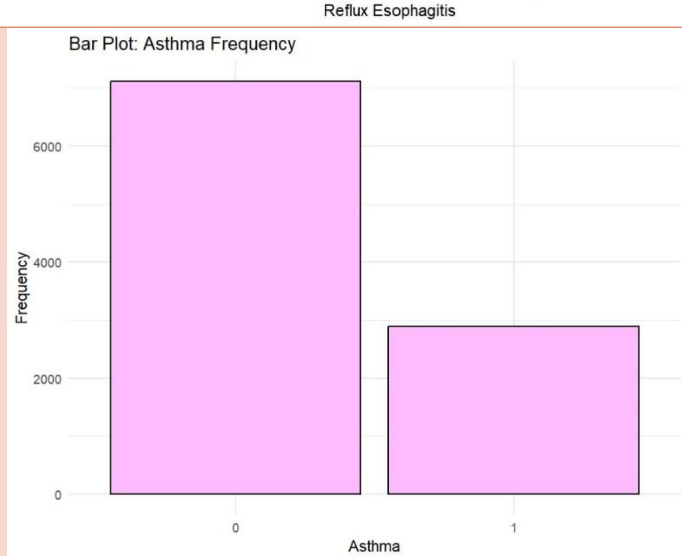
Arthritis

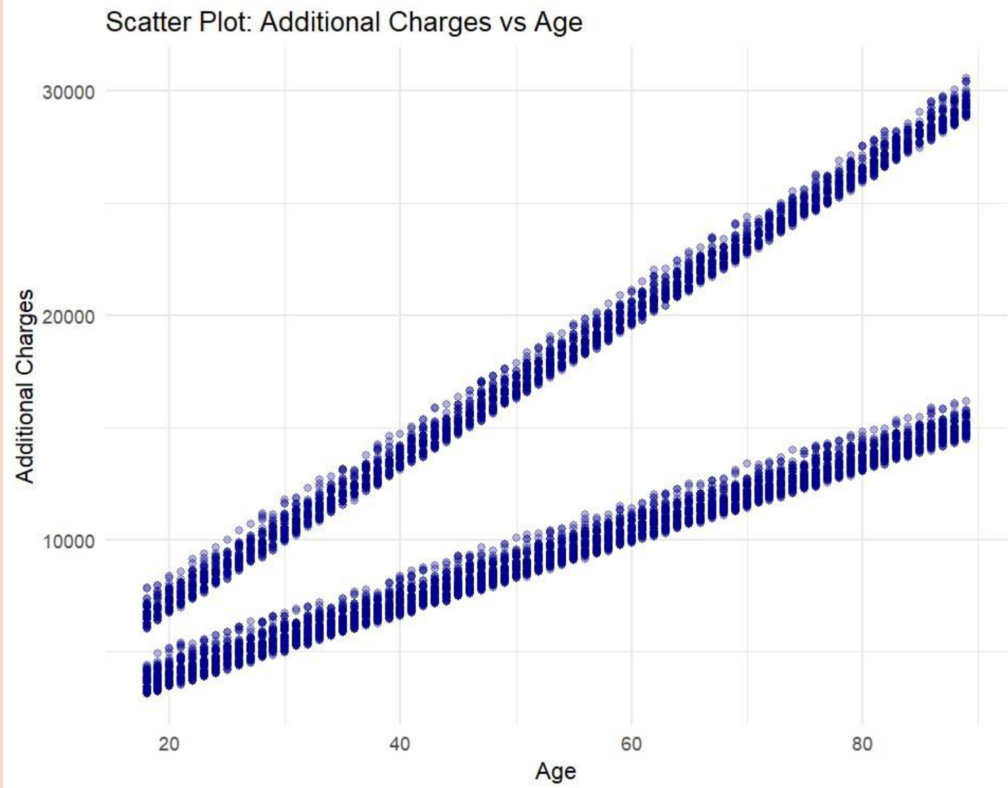
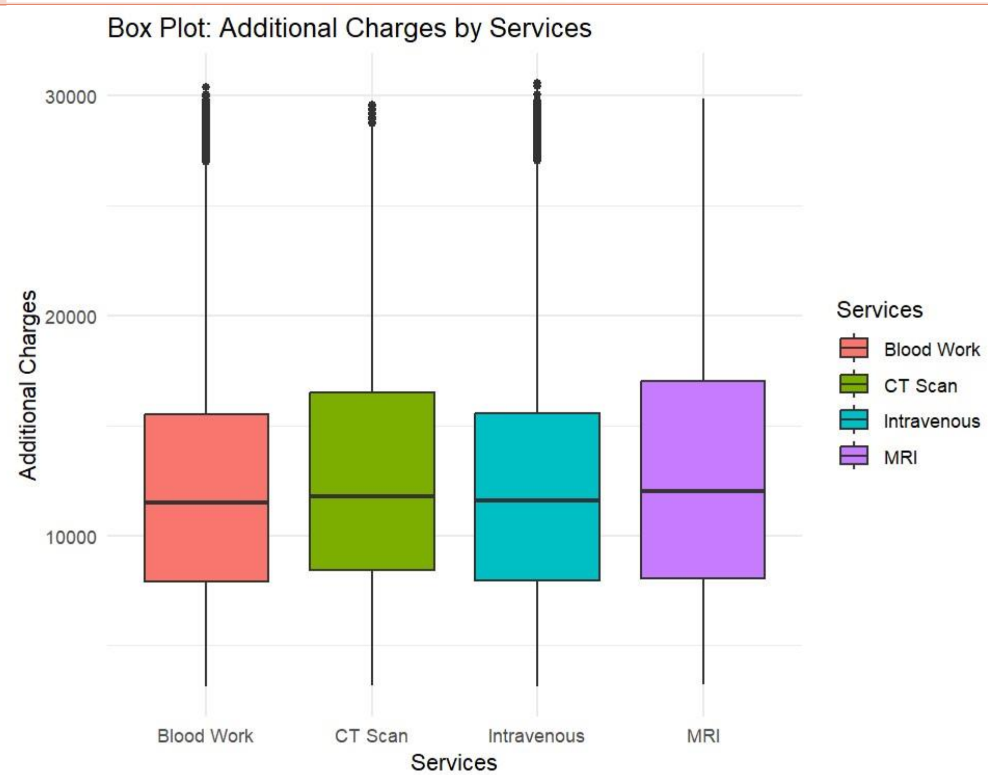


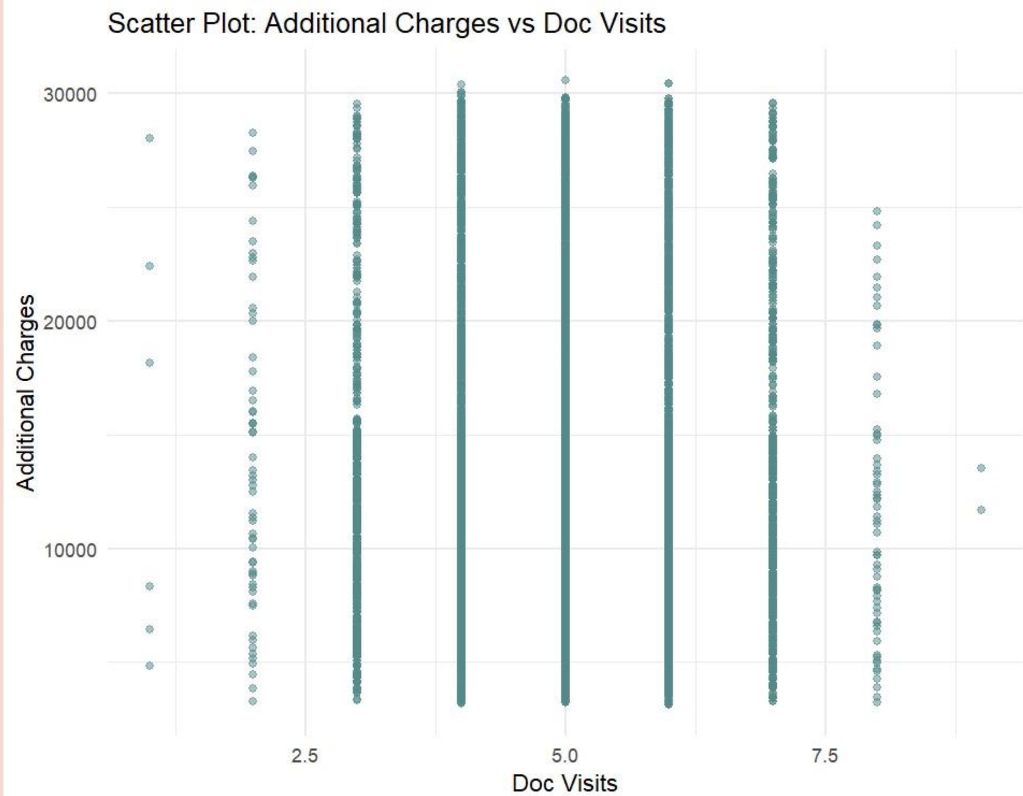
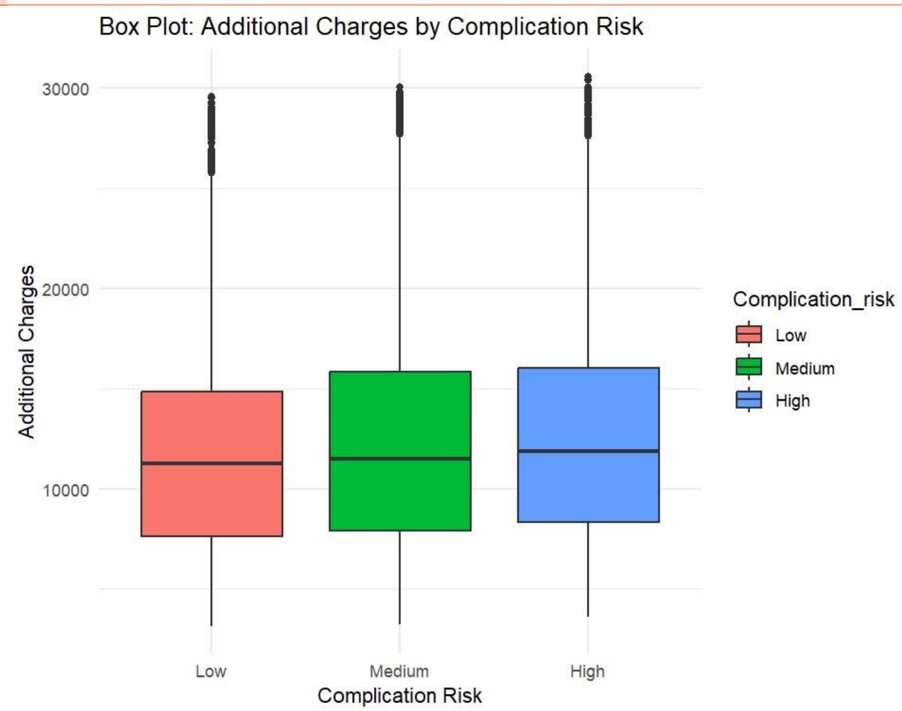
Diabetes

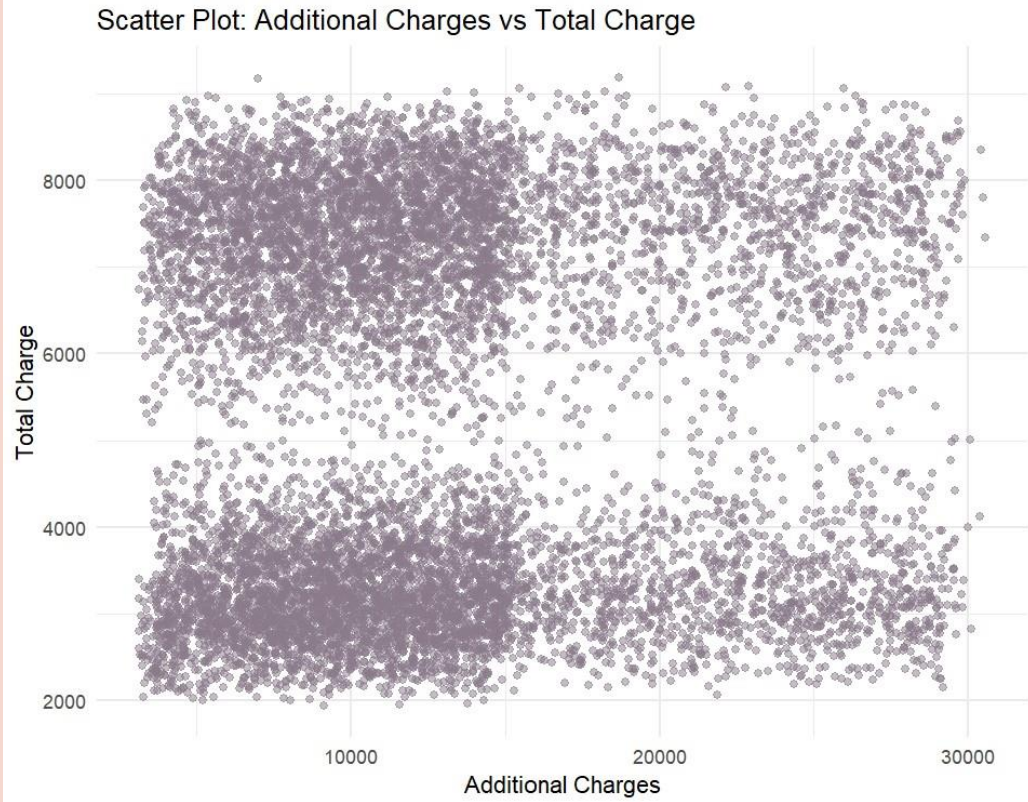
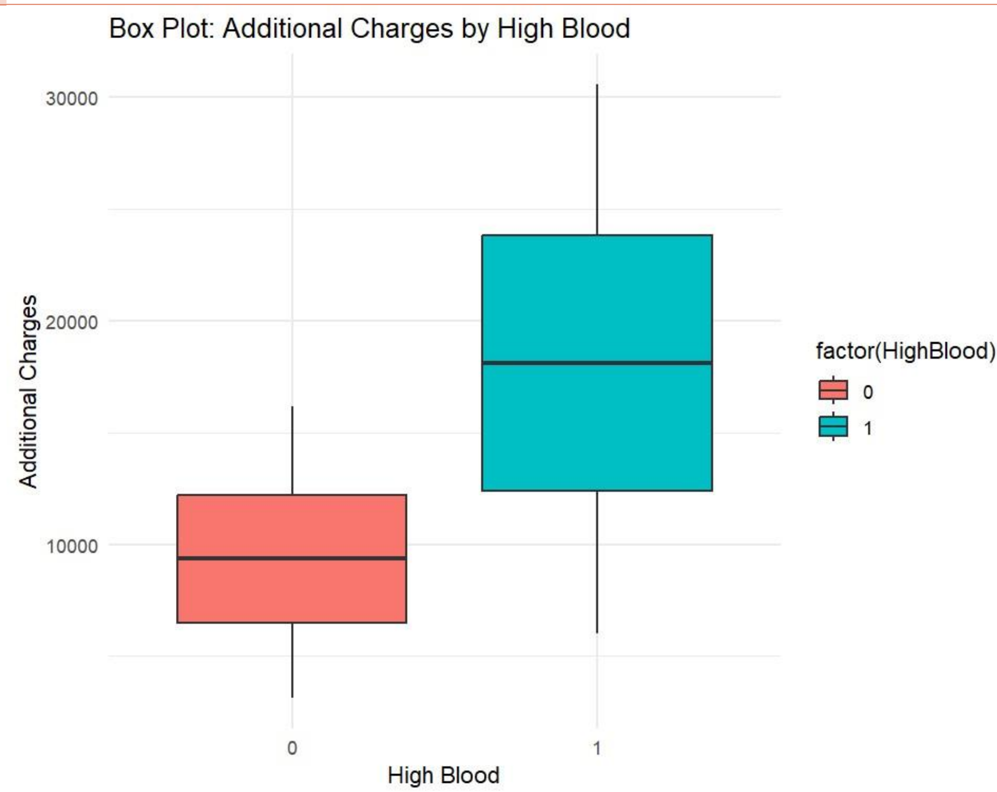


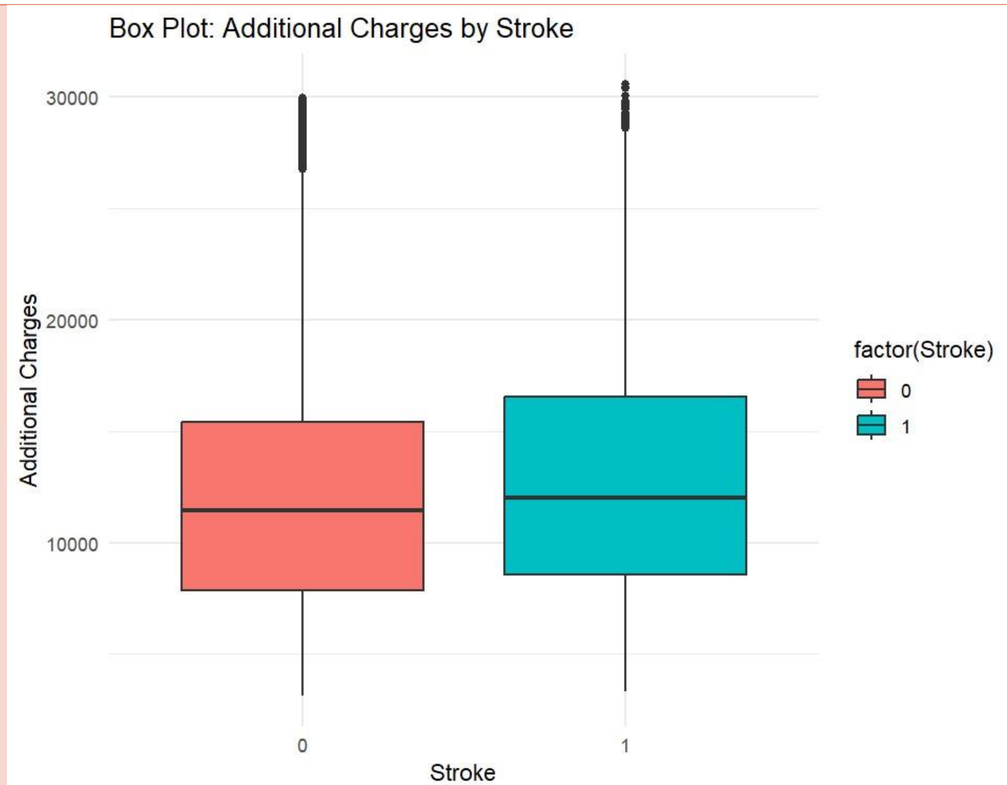
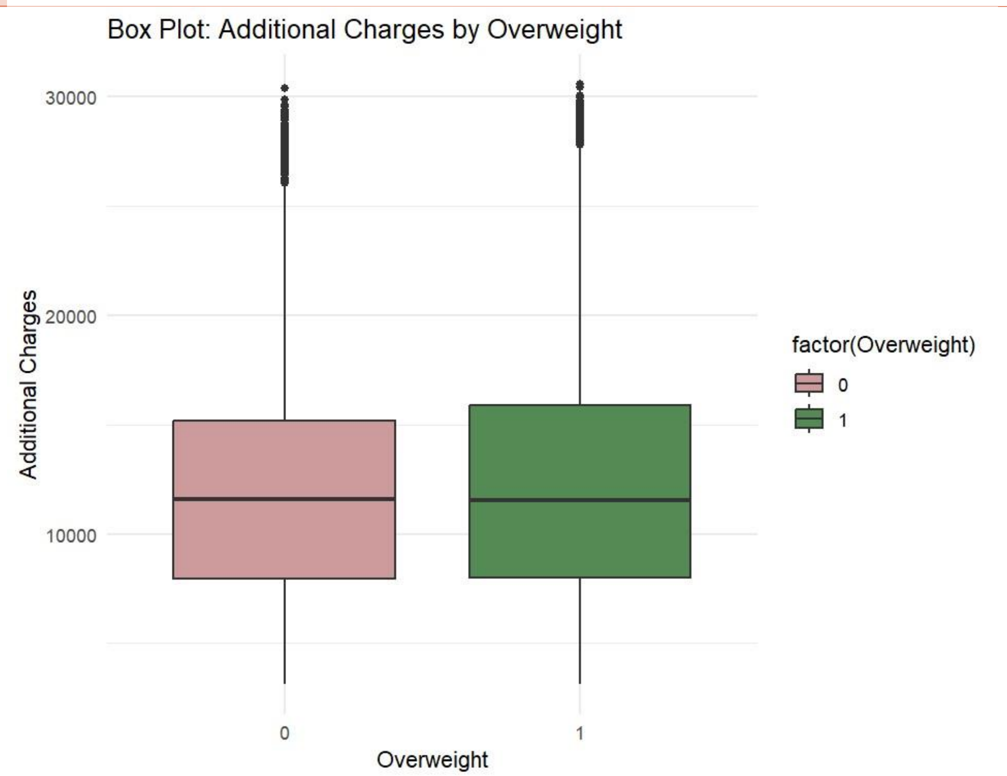
Hyperlipidemia	
Backpain	
Anxiety	

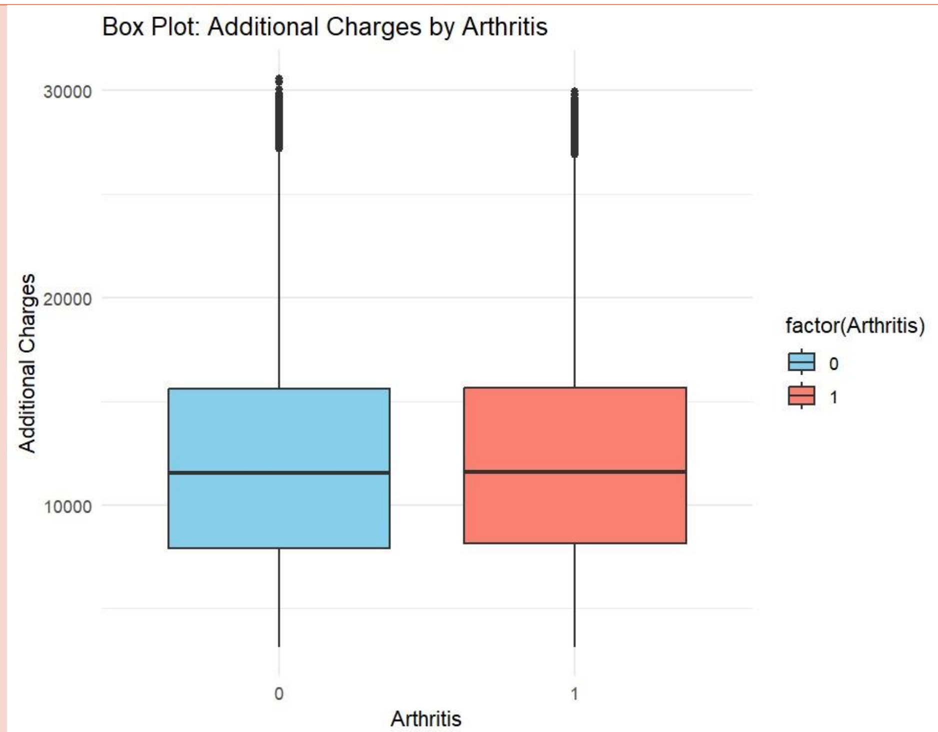
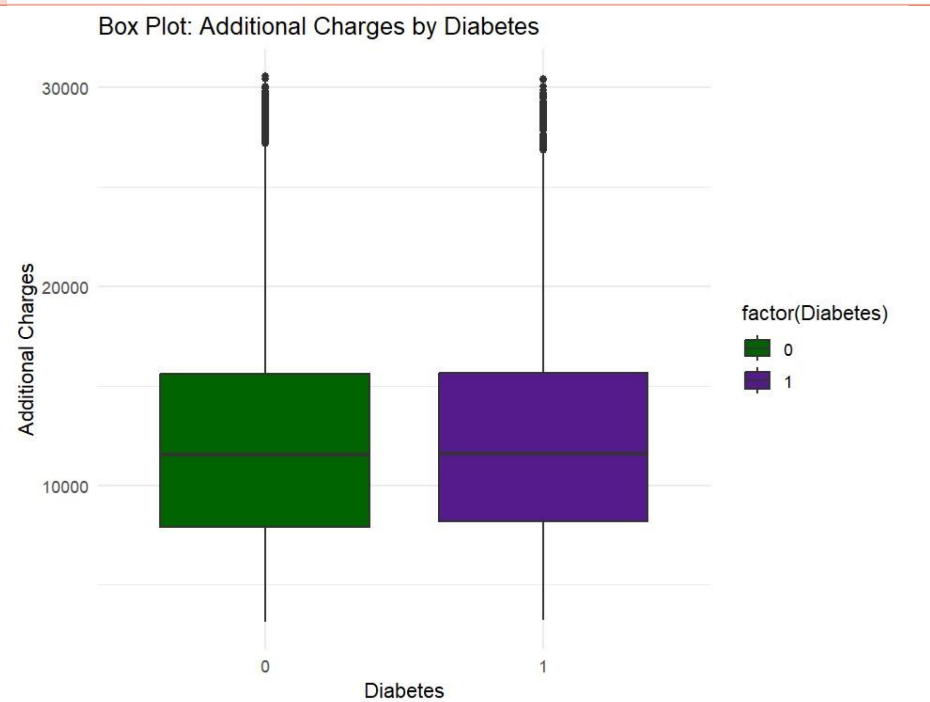
Allergic rhinitis	
Reflux esophagitis	
Asthma	

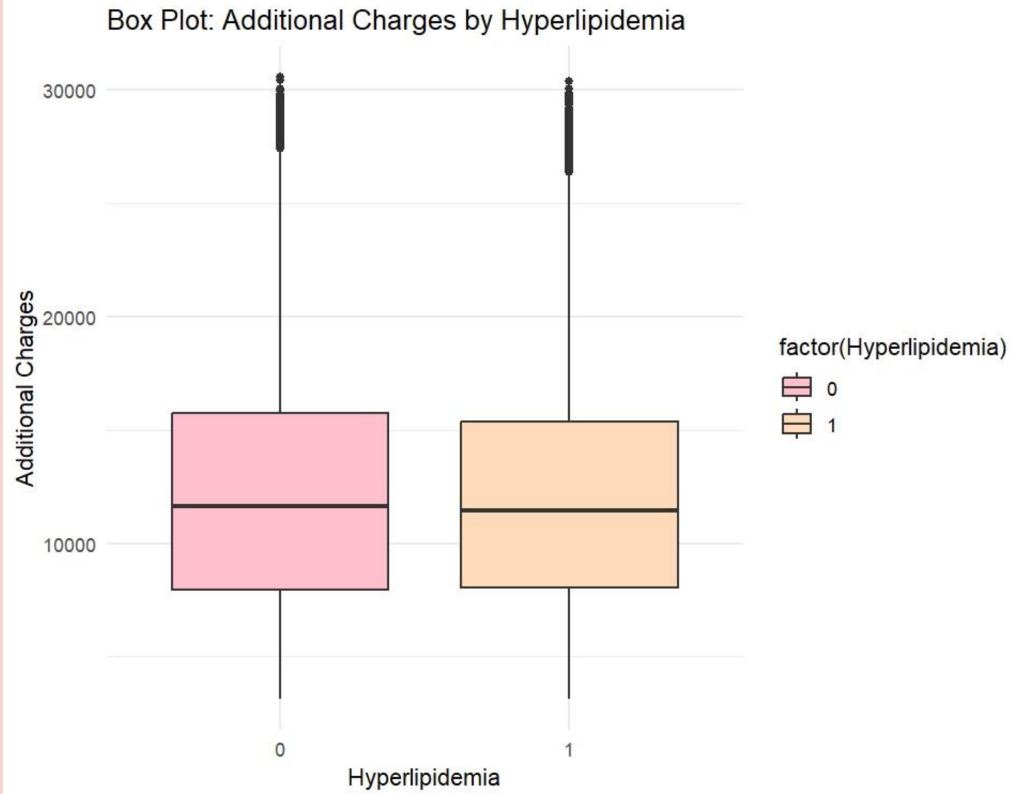
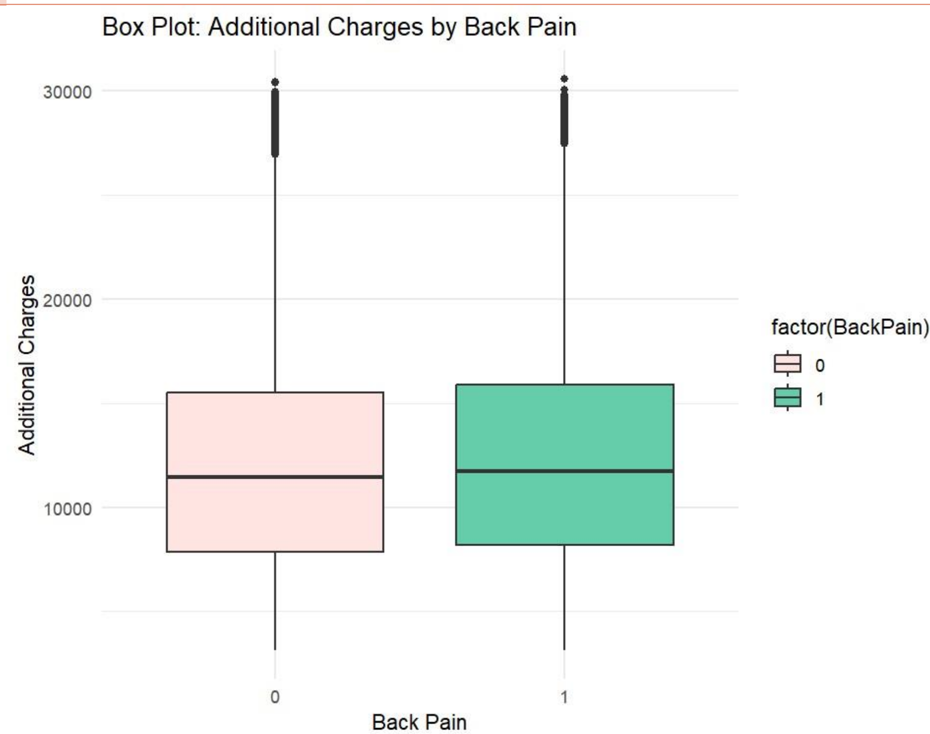
Variables	Bivariate Visualizations
Age	
Services	

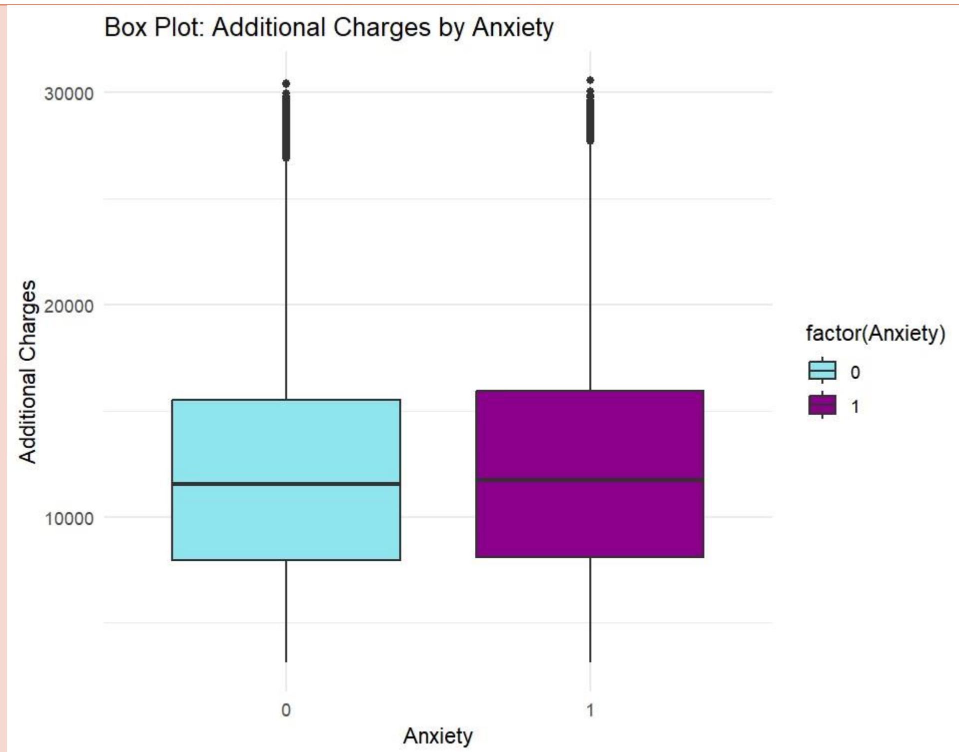
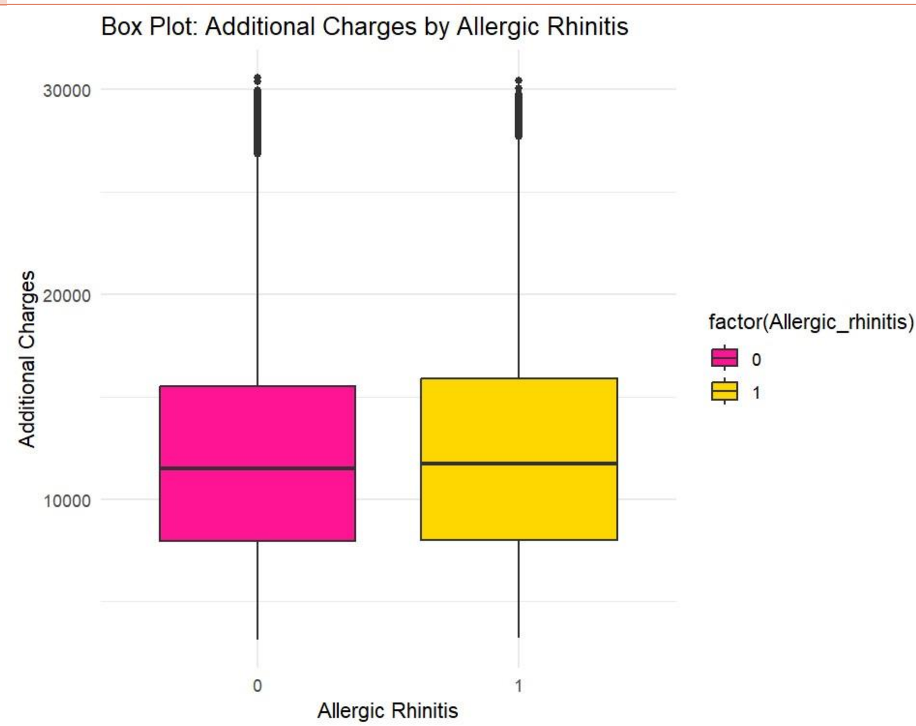
Doc Visits	
Complication Risk	

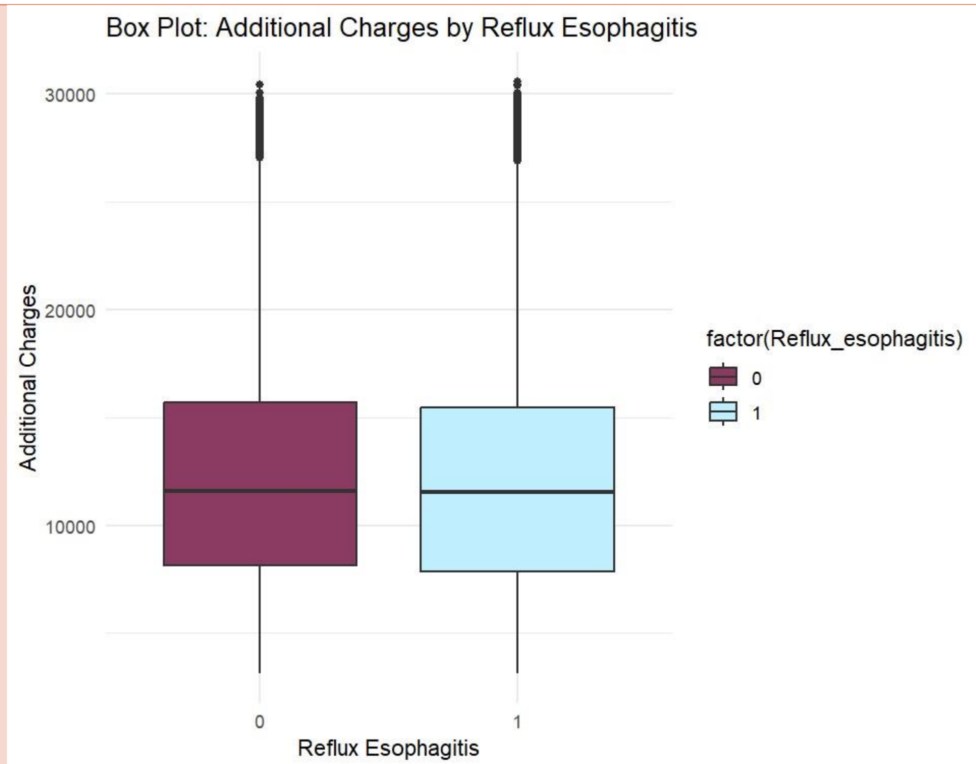
Total Charges	
High Blood	

Stroke	
Overweight	

Arthritis	
Diabetes	

Hyperlipidemia	
Backpain	

Anxiety	
Allergic rhinitis	

Reflux esophagitis	
Asthma	

C4. DATA TRANSFORMATION

The data wrangling activities performed on the dataset are converting categorical variables to numeric and factor-level adjustments.

Several categorical variables, such as high blood, stroke, overweight, arthritis, diabetes, hyperlipidemia, back pain, anxiety, allergic rhinitis, reflux esophagitis, and asthma, were converted into numeric factors. This conversion was achieved using the plyr package with the revalue function. The conversion turned Yes to 1, No to 0, and NA to NA. The variable complication risk was adjusted to Low, Medium, and High levels. This was done with factor().

Performing the conversion and adjustment is essential to answer my research question. Firstly, this ensures consistency in data types and makes it easier to work with the dataset. Numeric representation of categorical variables helps maintain uniformity across the data set. Furthermore, adjusting factor levels is essential for ensuring that categorical variables are treated correctly in statistical models. Overall, data wrangling activities contribute to data quality. Uniform and well-structured data are necessary for obtaining reliable and meaningful insights from analysis.

Code for wrangling is attached

C5. PREPARED DATA SET

.csv attached

PART IV: MODEL COMPARISON & ANALYSIS

D1. INITIAL MODEL

```
Call:
lm(formula = Additional_charges ~ Age + Services + Doc_visits +
    Complication_risk + TotalCharge + HighBlood + Stroke + Overweight +
    Arthritis + Diabetes + Hyperlipidemia + BackPain + Anxiety +
    Allergic_rhinitis + Reflux_esophagitis + Asthma, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3912.6 -1308.5    17.4  1328.1  3925.9

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -2.963e+03  1.138e+02 -26.045  < 2e-16 ***
Age                     2.257e+02  7.978e-01 282.848  < 2e-16 ***
ServicesCT Scan        -2.497e+01  5.222e+01  -0.478  0.63250
ServicesIntravenous     2.779e+01  3.715e+01   0.748  0.45437
ServicesMRI             1.230e+02  8.742e+01   1.406  0.15961
Doc_visits             -1.801e+01  1.575e+01  -1.143  0.25286
Complication_riskMedium 1.201e+02  4.332e+01   2.773  0.00556 **
Complication_riskHigh   5.077e+02  4.570e+01  11.109  < 2e-16 ***
TotalCharge             6.267e-03  7.596e-03   0.825  0.40935
HighBlood               8.631e+03  3.350e+01 257.642  < 2e-16 ***
Stroke                  3.549e+02  4.121e+01   8.612  < 2e-16 ***
Overweight              2.976e+01  3.626e+01   0.820  0.41195
Arthritis              -5.704e+01  3.438e+01  -1.659  0.09711 .
Diabetes                5.264e+01  3.692e+01   1.426  0.15397
Hyperlipidemia          1.943e+01  3.483e+01   0.558  0.57697
BackPain               -3.506e+01  3.349e+01  -1.047  0.29519
Anxiety                 3.239e+01  3.526e+01   0.919  0.35827
Allergic_rhinitis      -5.137e-01  3.368e+01  -0.015  0.98783
Reflux_esophagitis      2.431e+01  3.344e+01   0.727  0.46736
Asthma                  5.360e+01  3.631e+01   1.476  0.13995
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1645 on 9980 degrees of freedom
Multiple R-squared:  0.9369,	Adjusted R-squared:  0.9368
F-statistic:  7801 on 19 and 9980 DF,  p-value: < 2.2e-16
```

D2. JUSTIFICATION OF MODEL REDUCTION

For this model, I used the backward stepwise elimination. Firstly, construct the initial regression model with all the predictor variables in your dataset. Then, the **step()** function does all the work. It utilizes multiple iterations that evaluate the contribution of each predictor variable. Every predictor variable that has the highest p-value and thus contributes the least to the model is removed. The function stops when the iteration with the lowest AIC is obtained.

The backward stepwise elimination method is vital to my research question since it enables me to select the variables that are statistically significant. These variables are considered the most important contributors to the increase in additional charges.

D3. REDUCED LINEAR REGRESSION MODEL

```
> summary(reduced_model)

Call:
lm(formula = Additional_charges ~ Age + Complication_risk + HighBlood +
    Stroke + Arthritis + Diabetes + Asthma, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3911.5 -1313.2    14.7  1323.2  3967.2 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -2976.5087    60.4656 -49.226  < 2e-16 ***
Age                       225.6459     0.7971 283.098  < 2e-16 ***
Complication_riskMedium   120.7432    43.2788   2.790  0.00528 ** 
Complication_riskHigh     508.9442    45.6075  11.159  < 2e-16 ***
HighBlood                8631.2481    33.4679 257.896  < 2e-16 ***
Stroke                    352.7240    41.1815   8.565  < 2e-16 ***
Arthritis                 -54.9452    34.3354  -1.600  0.10957    
Diabetes                   53.4839    36.8914   1.450  0.14716    
Asthma                     53.2815    36.2801   1.469  0.14197    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1645 on 9991 degrees of freedom
Multiple R-squared:  0.9369,	Adjusted R-squared:  0.9368 
F-statistic: 1.853e+04 on 8 and 9991 DF,  p-value: < 2.2e-16
```

E1. MODEL COMPARISON

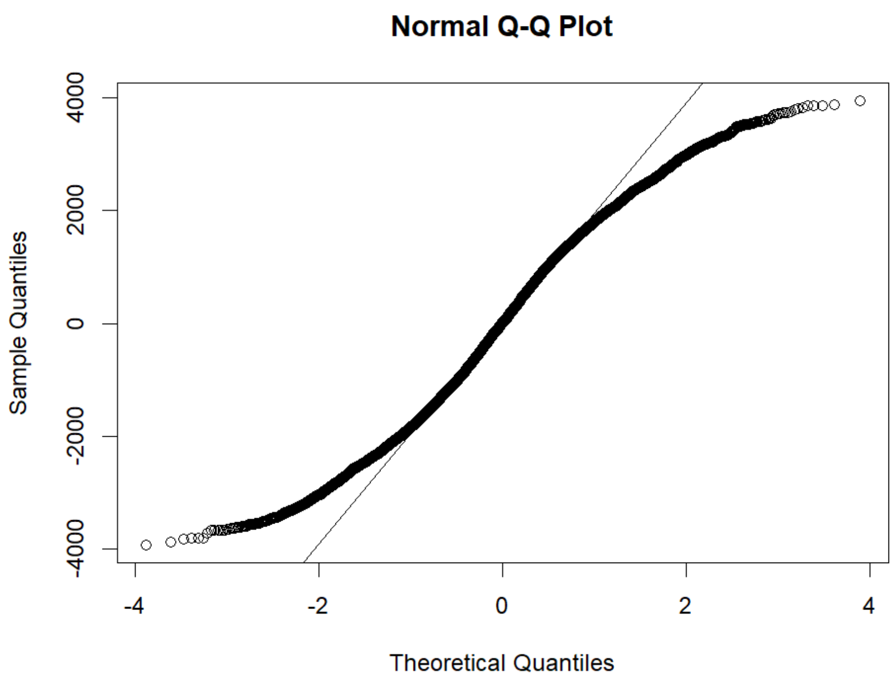
Metrics	Initial Model	Reduced Model
Residual Standard Error (RSE)	1645 on 9980 degrees of freedom	1645 on 9991 degrees of freedom
R-squared	0.9369	0.9369
Adjusted R-squared	0.9368	0.9368
F-statistic	7801 on 19 and 9980 degrees of freedom	1.853e+04 on 8 and 9991 degrees of freedom
p-value	p-value < 2.2e-16	p-value < 2.2e-16

The residual standard error is similar in both models. This indicates a similar goodness of fit regarding the spread of residuals. R-squared values are identical in both models. This demonstrates that models have similar variability in the response variable. The p-values are both extremely low, which indicates that both models have strong evidence against the null hypothesis. The F-statistic is much higher in the final model. This change suggests that the reduced model is a more statistically significant improvement over the initial model.

The F-statistic is vital to choosing my model since this metric measures the overall model significance. In this analysis, the F-statistic increased substantially in the reduced model. This indicates that the reduced model should be chosen over the initial model.

E2. OUTPUT & CALCULATIONS

RESIDUAL PLOT OF THE REDUCED MODEL



RESIDUAL STANDARD ERROR FOR THE REDUCED MODEL

```
Residual standard error: 1645 on 9991 degrees of freedom
Multiple R-squared:  0.9369,	Adjusted R-squared:  0.9368
F-statistic: 1.853e+04 on 8 and 9991 DF,  p-value: < 2.2e-16
```

E3. CODE

See .R code attached

PART V: DATA SUMMARY & IMPLICATIONS

F1. RESULTS

REGRESSION EQUATION

$$
\begin{aligned}
Additional\_charges \\
&= \beta_0 + \beta_1 \times Age + \beta_2 \times Complication\_riskMedium \\
&+ \beta_3 \times Complication\_riskHigh + \beta_4 \times HighBlood + \beta_5 \times Stroke + \epsilon
\end{aligned}
$$

INTERPRETATION

$$
\begin{aligned}
Additional\_charges \\
&= -2966.5087 + 225.6459 \times Age + 120.7432 \times Complication\_riskMedium \\
&+ 508.9442 \times Complication\_riskHigh + 8631.2481 \times HighBlood \\
&+ 352.7240 \times Stroke + \epsilon
\end{aligned}
$$

β_0 with the value of -2966.5087 is the intercept. This represents the estimated additional charges when all predictor variables are zero. β_1 with the value of 225.6459 means that for every one unit increase in age, the additional charges increase by 225.6459, holding other variables constant. If the complication risk is medium, additional charges will increase by β_2, which is 120.7432, holding other variables constant. If the complication risk is high, additional charges will increase by β_3, which is 508.9442, holding other variables constant. If a patient has high blood, additional charges will increase by β_4 which is 8631.2481, holding other variables constant. If a patient has experienced a stroke, additional charges will increase by β_5, which is 354.7240, holding other variables constant.

STATISTICAL SIGNIFICANCE

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -2976.5087    60.4656 -49.226  < 2e-16 ***
Age                     225.6459     0.7971 283.098  < 2e-16 ***
Complication_riskMedium 120.7432    43.2788   2.790  0.00528 **
Complication_riskHigh   508.9442    45.6075  11.159  < 2e-16 ***
HighBlood              8631.2481    33.4679 257.896  < 2e-16 ***
Stroke                  352.7240    41.1815   8.565  < 2e-16 ***
Arthritis               -54.9452    34.3354  -1.600  0.10957
Diabetes                 53.4839    36.8914   1.450  0.14716
Asthma                   53.2815    36.2801   1.469  0.14197
---

Residual standard error: 1645 on 9980 degrees of freedom
Multiple R-squared:  0.9369,    Adjusted R-squared:  0.9368
F-statistic:  7801 on 19 and 9980 DF,  p-value: < 2.2e-16
```

The p-values of the reduced model and the F-statistic are essential to assessing the statistical significance. The p-values should be very small to indicate that the variable is statistically significant in predicting additional charges. All the corresponding variables chosen are below the significance level (0.05). Also, the overall reduced model's F-statistic has a very low p-value. This suggests that at least one predictor variable contributes significantly to the variability.

Since both metrics are statistically significant, this implies that the observed relationships are unlikely to have occurred by chance alone. The results are more likely due to the relationships between variables. This strengthens the validity and reliability of the model.

PRACTICAL SIGNIFICANCE

It is crucial to find the practical significance of the model to focus on the relevance of the results. The model suggests that age, complication risk, high blood, and stroke are significant factors that influence additional charges. Older patients should expect higher charges in their hospital bills. Moreover, patients with medium and high complication risk levels should expect higher additional charges. Patients with the medical condition of high blood and stroke should be wary of the additional financial impact on their hospital bills.

The model's practical significance is its ability to identify specific patient characteristics directly impacting healthcare costs. Practical significance is found in understanding these variables' effects on patients and healthcare professionals. It can help with resource allocation, budgeting, and patient care planning.

LIMITATIONS OF THE ANALYSIS

It is vital to be aware of the limitations of the analysis. The transformation of categorical variables in regression analysis is one of the potential limitations. Firstly, the dataset may lose some information inherent in the categories. In this analysis, this transformation is evident for variables like high blood and stroke, and potential complication risk. Moreover, recoding categorical variables into numerical values might lead to a loss of meaningful information. Assigning 1 to Yes and 0 to No may not capture the nuances of each category. This can lead to a risk of misinterpretation. Users of the model may interpret the binary if not adequately communicated.

Furthermore, the linear regression model assumes a linear relationship between predictors and the response variable. If the true relationship is non-linear, the model may not capture the complexity of the data. It is also important to note causation and correlation. The analysis identifies an association between predictor values and additional charges but does not establish causation. Lastly, the final reduced model simplifies the true underlying relationships. While this simplification aids interpretation, it may overlook nuanced interactions present in the data.

F2. RECOMMENDATIONS

Based on the regression analysis where I identified the significant predictors for additional charges on the hospital bill, I have several recommendations for the organization.

It is crucial to integrate regression findings into clinical decision support systems. Ensure that healthcare professionals are aware of the significant predictors. Training can support a more informed approach to patient care. Also, consider allocating resources based on identified risk factors. The organization should prioritize resources and preventive measures for patients with a patient with higher complication risk and those with a history of high blood and stroke.

It is widely known that prevention is better than cure. The organization should implement strategies to engage patients. Actively involve patients in their care plans, especially those with identified risk factors. Patient engagement can contribute to better health outcomes. Also, there is a need to develop patient education programs. Teach patients about the importance of preventive measures. It should be highly emphasized that this initiative is all for the sake of patients. This not only reduces their hospital bills but improves their overall quality of life.

These recommendations gained from the analysis are to enhance the effectiveness of healthcare and improve patient outcomes. It is vital that the organization implement these recommendations with a focus on continuous improvement and further development in alignment with its mission and values.