D208 Performance Assessment

Hillary Osei (Student ID #011039266)

Western Governors University, College of Information Technology

Program Mentor: Dan Estes

January 2, 2025

Table of Contents

Part I: Research Question

Section A1) Research Question

The research question is “Which factors contribute to the length of a patient’s initial hospital stay?”

Section A2) Goals

The goal of analyzing the research question, “Which factors contribute to the length of a patient’s initial hospital stay?” is to identify and measure important predictors that influence a patient’s initial hospital stay. The data analysis explores variables concerning patient demographics, hospital administration type, and health conditions to determine which variable has the most significant impact and influences variations in stay lengths.

For example, identifying medical conditions that lead to prolonged stays helps healthcare providers to focus on addressing those specific conditions more effectively. Hospitals could develop specialized treatment plans to manage these conditions effectively. On the financial side, organizations can also reduce costs caused by prolonged hospitalizations.

Part II: Method Justification

Section B1) Summary of Assumptions

The four assumptions of a multiple linear regression model are as followed (Bobbit, 2021):

1. There is a linear relationship between the independent and dependent variables
2. There is no high correlation between independent variables, also known as multicollinearity.
3. Multivariate Normality: There is a normal distribution in the residuals (the difference between observed and predicted values)
4. A constant variance exists across every point of the linear model (Homoscedasticity).

Section B2) Tool Benefits

Python has many versatile libraries that can help with the data analysis process. Libraries such as pandas, numpy, and matplotlib allow efficient data cleaning, exploration, and visualization. For instance, pandas can be used with functions like df.duplicated() and df.isnull().sum() to quickly identify duplicates and missing values. NumPy provides tools for numerical computations that help with consistency in data processing. Matplotlib and Seaborn also make it easy to create informative visual plots and explore trends in the data.

For statistical modeling, Python's statsmodels library offers tools to build and evaluate regression models, providing detailed outputs such as coefficients, p-values, R2 values, F-statistic, and other metrics. These metrics help pinpoint the most significant predictors of hospital stay lengths. The package sklearn.preprocessing.OrdinalEncoder simplifies re-expressing ordinal variables into numerical values. To address multicollinearity, the variance_inflation_factor function from statsmodels ensures that predictors are independent, improving the stability of the regression model. By combining these libraries, Python facilitates efficient and accurate analysis, from data preparation to advanced modeling, supporting meaningful insights that align with the research question.

Section B3) Appropriate Technique

Multiple linear regression is an effective method for analyzing the research question as it examines the relationship between a single dependent variable and multiple independent variables. The dependent variable, "Initial_days," is a continuous measure of the number of days a patient stayed in the hospital during their initial visit (*Medical Data Considerations and Dictionary,* n.d.). Because this technique specifically requires a continuous dependent variable, multiple linear regression is an appropriate choice for analyzing the research question. (JMP, n.d.).

Multiple linear regression allows for the evaluation of how each independent variable influences "Initial_days," identification of which predictors significantly impact the length of hospital stays, and determination of how much of the variability in hospital stays can be explained by these predictors. Additionally, the technique provides insights into the strength and direction of relationships between the predictors and "Initial_days" (Soetewey, 2021). For example, it identifies whether certain factors are associated with longer or shorter hospital stays. Multiple linear regression also assesses the goodness of fit of the model, ensuring that the included variables adequately explain the variability in hospital stay lengths.

Part III: Data Preparation

Section C1) Data Cleaning

Data cleaning plays a vital role in preparing a dataset for analysis, ensuring the data is accurate, consistent, and ready for generating insights. The process involves key tasks such as identifying and removing duplicate entries, addressing missing values, standardizing numerical data, and transforming categorical variables into formats suitable for statistical modeling. Duplicate records are removed to prevent distorted results, while missing values are handled to avoid

biases in the analysis. Numerical data is standardized, such as rounding values to a consistent number of decimal places, to align with real-world conventions and improve interpretability.

Categorical variables are transformed into numeric formats to work effectively with statistical models. For example, binary variables with “Yes” or “No” values are converted into numerical representations like 1 and 0. Variables with multiple categories are encoded using one-hot or ordinal encoding, depending on whether the categories have a natural order. Column names are simplified for better readability, and boolean values are converted into numeric values to make the dataset easier to work with and more practical for analysis.

Through these steps, the dataset becomes clean, reliable, and ready for analysis. This process minimizes errors and allows statistical models to uncover meaningful patterns and relationships, ultimately leading to more accurate and actionable insights.

Section C2) Summary Statistics

For the analysis, the dependent variable selected is “Initial_days,” and the independent variables selected are “Age,” “Income,” “Marital,” “Gender,” “Area,” “TotalCharge,” “VitD_levels,” “HighBlood,” “Stroke,” “Complication_risk,” “Overweight,” “Arthritis,” “Diabetes,” “Hyperlipidemia,” “BackPain,” “Anxiety,” “Allergic_rhinitis,” “Reflux_esophagitis,” “Asthma,” “Initial_admin,” “Doc_visits,” and “Services.”

To get the summary statistics for the dependent and continuous independent variables, the describe() function was used.

For the “Age” variable,” the mean is 23.511700 years old. The standard deviation is at 20.638538. The “Age” variable ranges from 18 to 89 years old. The 25th percentile is at 36 years old, the 50th percentile/median is at 53 years old, and the 75th percentile is at 71 years old.

For the “Income” variable, the mean is $40490.495160. The standard deviation is at 28,521.153293. The “Income” variable ranges from $154.080000 to $207,249.100000. The 25th percentile is at $19,598.775000, the 50th percentile/median is at $33,768.420000, and the 75th percentile is at $54,296.402500.

For the “Docs_visits” variable, the average is 5.012200. The standard deviation is 1.045734. The “Docs_visits” variable ranges from 1 to 9. The 25th percentile is at 4, the 50th percentile/median is at 5, and the 75th percentile is at 6.

For the “TotalCharge” variable, the average is $5312.172769. The standard deviation is at 2180.393838. The “TotalCharge” variable ranges from $1,938.312067 to $9,180.728000. The

25th percentile is at $3,179.374015, the 50th percentile/median is at $5,213.952000, and the 75th percentile is at $7,459.699750

For the “VitD_levels” variable, the average is at 17.964262. The standard deviation is at 2.017231. The “VitD_levels” variable ranges from 9.806483 to 26.394449. The 25th percentile is at 16.626439, the 50th percentile/median is at 17.951122, and the 75th percentile is at 19.347963.

For the “Initial_days” variable, the average is 34.455299. The standard deviation is 26.309341. The range for the “Initial_days” variable is from 1.001981 to 71.981490. The 25th percentile is at 7.896215, the 50th percentile/median is at 35.836244, and the 75th percentile is at 61.161020.

	Age	Income	Children	Doc_visits	TotalCharge	VitD_levels	Initial_days
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	53.511700	40490.495160	2.097200	5.012200	5312.172769	17.964262	34.455299
std	20.638538	28521.153293	2.163659	1.045734	2180.393838	2.017231	26.309341
min	18.000000	154.080000	0.000000	1.000000	1938.312067	9.806483	1.001981
25%	36.000000	19598.775000	0.000000	4.000000	3179.374015	16.626439	7.896215
50%	53.000000	33768.420000	1.000000	5.000000	5213.952000	17.951122	35.836244
75%	71.000000	54296.402500	3.000000	6.000000	7459.699750	19.347963	61.161020
max	89.000000	207249.100000	10.000000	9.000000	9180.728000	26.394449	71.981490

To get the summary statistics for the categorical independent variables, the values_count() function was used.

For the “Marital” variable, the number of patients who are widowed is 2,045. The number of patients who are married is 2,023. The number of patients who are separated is 1,987. The number of patients who are divorced is 1,961

```
Marital
Widowed          2045
Married          2023
Separated        1987
Never Married    1984
Divorced         1961
Name: count, dtype: int64
```

For the “Gender” variable, the number of female patients is 5,018. The number of male patients is 4,768. The number of nonbinary patients is 214.

```
Gender
Female       5018
Male         4768
Nonbinary     214
Name: count, dtype: int64
```

For the “Area” variable, the number of patients who live in a rural area is 3,369. The number of patients who live in a suburban area is 3,328. The number of patients who live in an urban area is 3,303.

```
Area
Rural       3369
Suburban    3328
Urban       3303
Name: count, dtype: int64
```

For the “Initial_admin” variable, the number of patients who were initially admitted to the hospital for emergency admission was 5,060. The number of patients who were initially admitted to the hospital for elective admission is 2,504. The number of patients who were initially admitted for observation admission is 2,436.

```
Initial_admin
Emergency Admission      5060
Elective Admission       2504
Observation Admission    2436
Name: count, dtype: int64
```

For the “Services” variable, the number of patients who received primarily blood work while hospitalized is 5,265. The number of patients who received primarily IV is 3,130. The number of patients who received primarily CT scans is 1,225. The number of patients who received primarily MRI scans is 380.

```
Services
Blood Work     5265
Intravenous    3130
CT Scan        1225
MRI             380
Name: count, dtype: int64
```

For the “HighBlood” variable, the number of patients who do not have high blood pressure is 5,910 and the number of patients who do have high blood pressure is 4,090.

```
HighBlood
No     5910
Yes    4090
Name: count, dtype: int64
```

For the “Stroke” variable, the number of patients who have not had a stroke is 8,007 and the number of patients who have had a stroke is 1,993.

```
Stroke
No     8007
Yes    1993
Name: count, dtype: int64
```

For the “Complication_risk” variable, the number of patients who were assessed to be at a medium-level complication risk is 4,517. The number of patients who were assessed to be at a high-level complication risk is 3,358. The number of patients who were assessed to be at a low-level complication risk is 2,125.

```
Complication_risk
Medium    4517
High      3358
Low       2125
Name: count, dtype: int64
```

For the “Overweight” variable, the number of patients who are considered overweight based on age, gender, and height are 7,094 and the number of patients who are not considered overweight are 2,906.

```
Overweight
Yes    7094
No     2906
Name: count, dtype: int64
```

For the “Arthritis” variable, the number of patients who do not have arthritis is 6,426 and the number of patients who do have arthritis is 3,574.

```
Arthritis
No     6426
Yes    3574
Name: count, dtype: int64
```

For the “Diabetes” variable, the number of patients who are not diabetic are 7,262 and the number of patients who are diabetic are 2,738.

```
Diabetes
No     7262
Yes    2738
Name: count, dtype: int64
```

For the “Hyperlipidemia” variable, the number of patients who do not have hyperlipidemia are 6,628 and the number of patients who do have hyperlipidemia are 3,372.

```
Hyperlipidemia
No     6628
Yes    3372
Name: count, dtype: int64
```

For the “BackPain” variable, the number of patients who do not have chronic back pain are 5,886 and the number of patients who do have chronic back pain are 4,114.

```
BackPain
No     5886
Yes    4114
Name: count, dtype: int64
```

For the “Anxiety” variable, the number of patients who do not have an anxiety disorder are 6,785 and the number of patients who do have an anxiety disorder are 3,215.

```
Anxiety
No     6785
Yes    3215
Name: count, dtype: int64
```

For the “Allergic_rhinitis” variable, the number of patients who do not have allergic rhinitis are 6,059 and the number of patients who do have allergic rhinitis is 3,941.

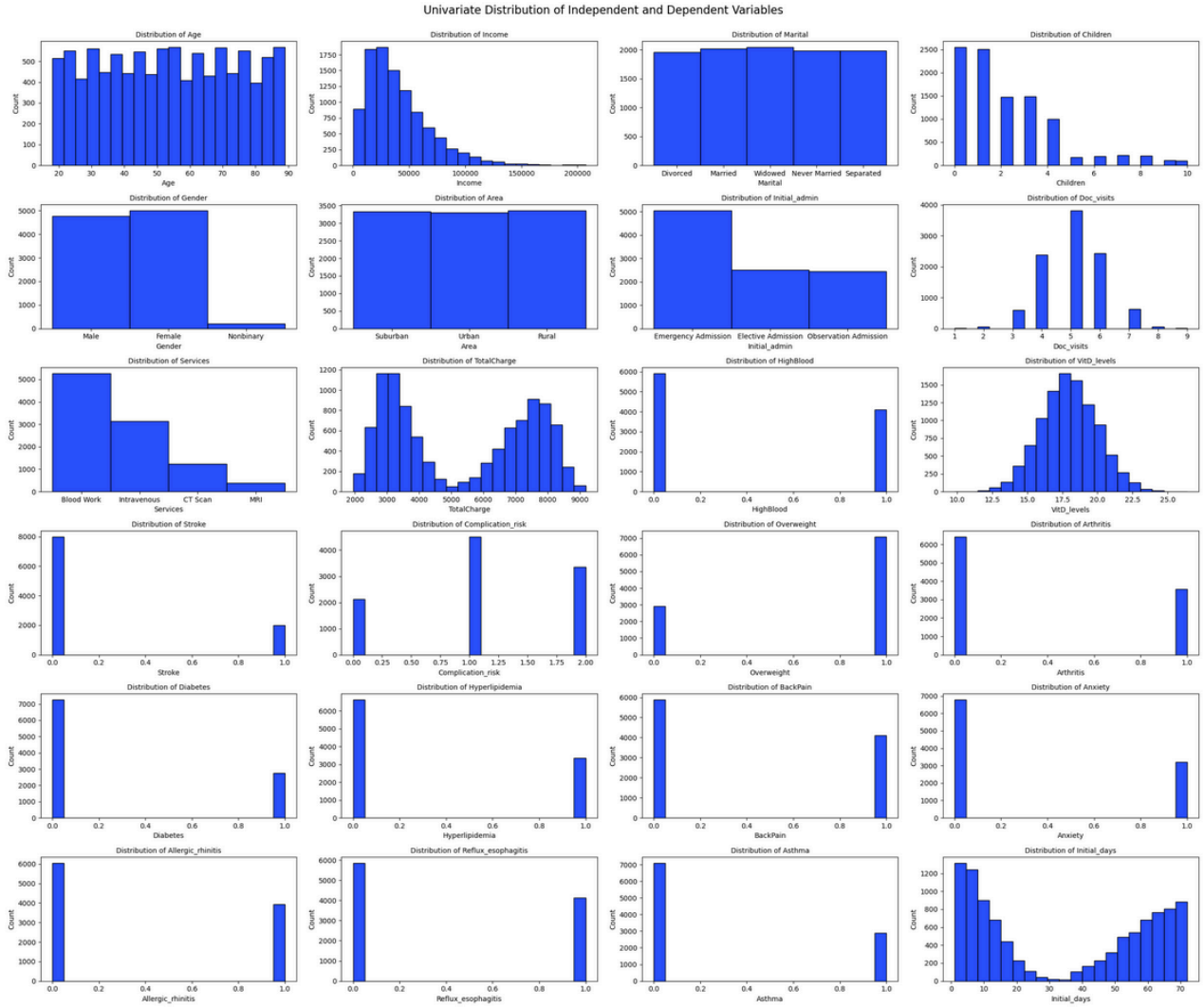
```
Allergic_rhinitis
No     6059
Yes    3941
Name: count, dtype: int64
```

For the “Reflux_esophagitis” variable, the number of patients who do not have reflux esophagitis is 5,865 and the number of patients who do have reflux esophagitis is 4,135.

```
Reflux_esophagitis
No     5865
Yes    4135
Name: count, dtype: int64
```

Section C3) Visualizations

Univariate:



The histograms provide insights into the distribution of each variable in the dataset. For “Age,” the distribution is somewhat uniform because there are several fluctuations. “Income” is right-skewed, with most patients earning less than $100,000 and a smaller number earning higher incomes. The “Marital” variable is relatively balanced for categories like “Married” and “Never Married,” while slightly fewer patients fall into the “Widowed” and “Separated” categories. For “Children,” the majority of patients have fewer than two children, with the frequency decreasing sharply for larger numbers of children.

The “Gender” variable shows that there are more female patients, with a smaller representation of nonbinary and male patients. “Area” shows there are slightly more patients living in rural areas, compared to suburban and urban areas. For “Initial_admin,” there are significantly more

emergency admissions while patients admitted for observation and elective reasons occur less. “Doc_visits” shows a roughly symmetric distribution, peaking at around 5 visits, suggesting this is the most common range for patients.

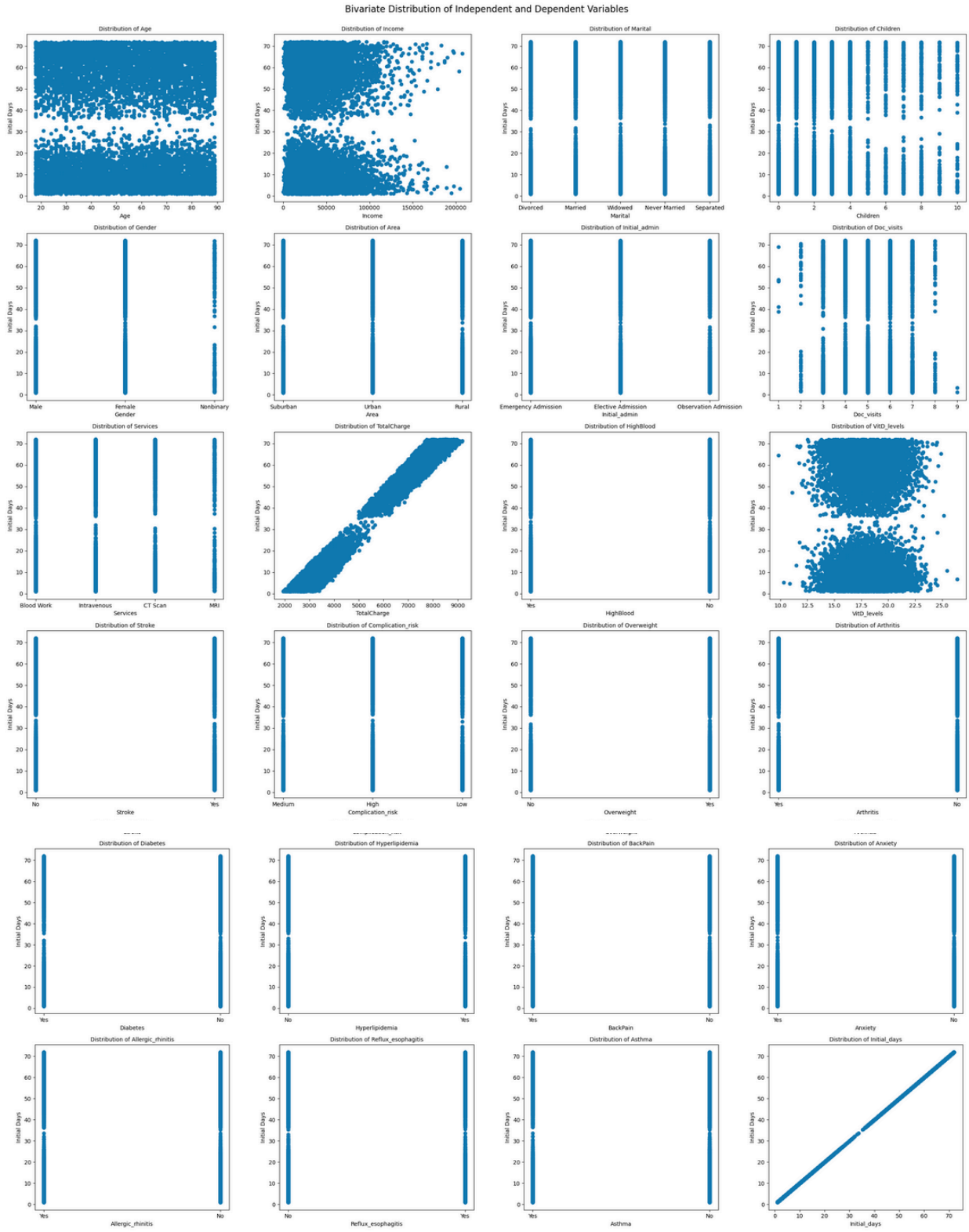
In the “Services” category, Blood Work is the most frequently used service, followed by Intravenous treatments and CT Scans, while MRI scans being the least used service. “TotalCharge” has a slight bimodal distribution, with most patients incurring charges around $3,000. “HighBlood” indicates that a significant portion of patients do not have high blood pressure. “VitD_levels” exhibit a bell-shaped, normal distribution, with values typically peaking around 17.5 units.

For “Stroke,” most patients have not had a stroke. “Complication_risk” shows that there are more patients with medium-level complication risks than low-level and high-level. “Overweight” has a clear majority of patients who are classified as overweight. Similarly, “Arthritis” shows that a significant number of patients do not have the condition.

The “Diabetes” variable reveals that the number of patients with diabetes is smaller compared to those without it. “Hyperlipidemia” follows a similar trend, with a smaller group of patients having the condition compared to those who are not. “BackPain” also has a lower frequency but still appears in a substantial portion of patients. For “Anxiety,” the distribution shows more patients do not experience this condition.

For “Allergic_rhinitis” and “Reflux_esophagitis,” most patients do not have the condition, similarly with “Asthma.”

Bivariate:


Bivariate Distribution of Independent and Dependent Variables
Distribution of Age
Distribution of Income
Distribution of Marital
Distribution of Children
Distribution of Gender
Distribution of Area
Distribution of Initial_admin
Distribution of Doc_visits
Distribution of Services
Distribution of TotalCharge
Distribution of HighBlood
Distribution of VitD_levels
Distribution of Stroke
Distribution of Complication_risk
Distribution of Overweight
Distribution of Arthritis
Distribution of Diabetes
Distribution of Hyperlipidemia
Distribution of BackPain
Distribution of Anxiety
Distribution of Allergic_rhinitis
Distribution of Reflux_esophagitis
Distribution of Asthma
Distribution of Initial_days
Initial Days
Age
Income
Marital
Children
Gender
Area
Initial_admin
Doc_visits
Services
TotalCharge
HighBlood
VitD_levels
Stroke
Complication_risk
Overweight
Arthritis
Diabetes
Hyperlipidemia
BackPain
Anxiety
Allergic_rhinitis
Reflux_esophagitis
Asthma
Initial_days

The scatterplots show the relationships between the dependent variable, “Initial_days,” and each independent variable. No clear relationship is evident for “Age,” as the points are evenly scattered across all ages. Similarly, “Income” shows a pattern where higher incomes are associated with shorter hospital stays, as the data points thin out at higher income levels. “Marital” has no clear trend for the various marital statuses. “Children” also shows no clear trend.

For “Gender,” the distributions of hospital stays are similar for the male, female, and nonbinary categories. “Area” also shows no distinct differences between rural, suburban, and urban areas. In “Initial_admin,” there is no clear relationship. “Doc_visits” does not show a strong relationship with “Initial_admin.”

The “Services” variable does not correlate with “Initial_admin.” On the other hand, “TotalCharge” displays a strong positive linear relationship, with higher charges corresponding to more extended hospital stays. For “HighBlood,” no noticeable differences are observed in hospital stays between patients with or without high blood pressure. “VitD_levels” exhibits a bimodal distribution. The upper cluster represents patients with more extended hospital stays, while the lower cluster corresponds to shorter stays.

“Stroke” shows no significant differences in hospital stays between patients who have and have not had a stroke. For “Complication_risk,” there is no clear relationship. The “Overweight” variable does not exhibit any apparent relationship with hospital stays, and the same is true for “Arthritis,” as both groups show similar distributions. “Diabetes” and “Hyperlipidemia” also show no clear trends, with patients having similar hospital stay lengths regardless of these conditions.

For “BackPain” and “Anxiety,” no strong relationships with hospital stays are evident. Similarly, “Allergic_rhinitis,” “Reflux_esophagitis,” and “Asthma” do not appear to correlate with “Initial_admin”

Section C4) Data Transformation

Data transformation is the process of converting raw data into a standardized format to improve data quality and ensure it is ready for analysis (IBM, 2024). It is a critical aspect of the data wrangling process, alongside data cleaning (Lu, 2024), and is performed before the initial regression modeling to ensure the data meets the assumptions of the regression model and reduces the chance of skewed, inaccurate data.

The data transformation process begins by using the df.info() function to obtain a summary of the DataFrame's data types. This step is necessary to identify the types of variables (e.g., numerical, categorical, or object) and determine the appropriate transformations needed for each. In this case, the dataset contained 7 columns with the float64 data type, 16 columns with the int64 data type, and 27 columns with the object data type.

The first step in data cleaning process involves checking for duplicates using the df.duplicated() function. This function returns a boolean value (True or False) for each row, indicating whether it is a duplicate of another row. In this dataset, all rows returned False, meaning there were no duplicate records. Next, missing values were checked using the df.isnull().sum() function, which provides a count of missing values for each column. Here, all columns returned 0, indicating no missing values.

Then, the columns "TotalCharge" and "VitD_levels" were rounded to two decimal places using the round() function. This step is important to improve the readability of numerical values:

```
# Round "TotalCharge" and VitD_levels
df['TotalCharge'] = df['TotalCharge'].round(2)
df['VitD_levels'] = df['VitD_levels'].round(2)
```

Re-expressing categorical variables is an essential step because regression modeling requires numerical values, and doing so ensures that data is usable, accurate, and easily interpreted. Categorical variables with binary (yes/no) values, such as "HighBlood," "Stroke," "Overweight," "Arthritis," "Diabetes," "Hyperlipidemia," "BackPain," "Anxiety," "Allergic_rhinitis," "Reflux_esophagitis," "Asthma," "ReAdmis," and "Soft_drink," were re-expressed using a dictionary. The dictionary (yes_no_dict) assigned the value 0 to "No" and 1 to "Yes." This transformation was applied using the .replace() method:

```
# Re-express categorical variables for "HighBlood", "Stroke", "Overweight", "Arthritis",
"Diabetes",
# "Hyperlipidemia", "BackPain", "Anxiety", "Allergic_rhinitis", "Reflux_esophagitis",
"Asthma", "ReAdmis," "Soft_drink"
yes_no_dict = {'No': 0, 'Yes': 1}
df['HighBlood'] = df['HighBlood'].replace(yes_no_dict).infer_objects(copy=False)
df['Stroke'] = df['Stroke'].replace(yes_no_dict).infer_objects(copy=False)
df['Overweight'] = df['Overweight'].replace(yes_no_dict).infer_objects(copy=False)
df['Arthritis'] = df['Arthritis'].replace(yes_no_dict).infer_objects(copy=False)
df['Diabetes'] = df['Diabetes'].replace(yes_no_dict).infer_objects(copy=False)
df['Hyperlipidemia'] = df['Hyperlipidemia'].replace(yes_no_dict).infer_objects(copy=False)
df['BackPain'] = df['BackPain'].replace(yes_no_dict).infer_objects(copy=False)
df['Anxiety'] = df['Anxiety'].replace(yes_no_dict).infer_objects(copy=False)
df['Allergic_rhinitis'] = df['Allergic_rhinitis'].replace(yes_no_dict).infer_objects(copy=False)
```

```
df['Reflux_esophagitis'] =
df['Reflux_esophagitis'].replace(yes_no_dict).infer_objects(copy=False)
df['Asthma'] = df['Asthma'].replace(yes_no_dict).infer_objects(copy=False)
# Re-express other categorical variables
df[”ReAdmis”] = df[”ReAdmis”].replace(yes_no_dict).infer_objects(copy=False)
df[”Soft_drink”] = df[”Soft_drink”].replace(yes_no_dict).infer_objects(copy=False)
```

Dummy variables were created for categorical variables such as “Gender,” “Area,” “Marital,” “Services,” and “Initial_admin” using pd.get_dummies(). This method is a simple way to one-hot encode categorical variables while avoiding multicollinearity by dropping the first category for each variable:

```
# Dummy variables for “Gender,” “Area,” “Initial_admin,” “Services,” “Marital”
df = df.join(pd.get_dummies(df['Gender'], prefix='Gender', drop_first=True))
df = df.join(pd.get_dummies(df['Area'], prefix='Area', drop_first=True))
df = df.join(pd.get_dummies(df['Initial_admin'], prefix='Initial_admin', drop_first=True))
df = df.join(pd.get_dummies(df['Services'], prefix='Services', drop_first=True))
df = df.join(pd.get_dummies(df['Marital'], prefix='Marital', drop_first=True))
```

In the above code, the first category “Female” was dropped for “Gender.”For “Area,” “Rural” was dropped. For “Initial_admin,” “Elective Admission” was dropped. For “Services,” “Blood Work” was dropped. For “Marital,” “Divorced” was dropped.

Spaces in column names resulting from pd.get_dummies() were replaced with underscores for readability using:

```
# Replace spaces with '_'
df.columns = df.columns.str.replace(' ', '_')
```

This process resulted in the creation of the variables “Gender_Male,” “Gender_Nonbinary,” “Area_Suburban,” “Area_Urban,” “Initial_admin_Emergency_Admission,” “Initial_admin_Observation_Admission,” “Services_CT_Scan,” “Services_MRI,” “Services_Intravenous,” “Marital_Married,” “Marital_Never_Married,” “Marital_Separated,” amd “Marital_Widowed.”

Furthermore, the dummy variables with boolean values (True or False) were converted to binary values (1 or 0) using a dictionary (true_false_dict), where the value “True” was set to 1 and the value “False” was set to 0:

```
# Convert dummy variables into 0/1
true_false_dict = {True: 1, False: 0}
df['Gender_Male'] = df['Gender_Male'].replace(true_false_dict).infer_objects(copy=False)
```

```
df['Gender_Nonbinary'] =
df['Gender_Nonbinary'].replace(true_false_dict).infer_objects(copy=False)
df['Area_Suburban'] = df['Area_Suburban'].replace(true_false_dict).infer_objects(copy=False)
df['Area_Urban'] = df['Area_Urban'].replace(true_false_dict).infer_objects(copy=False)
df['Initial_admin_Emergency_Admission'] =
df['Initial_admin_Emergency_Admission'].replace(true_false_dict).infer_objects(copy=False)
df['Initial_admin_Observation_Admission'] =
df['Initial_admin_Observation_Admission'].replace(true_false_dict).infer_objects(copy=False)
df['Services_CT_Scan'] =
df['Services_CT_Scan'].replace(true_false_dict).infer_objects(copy=False)
df['Services_Intravenous'] =
df['Services_Intravenous'].replace(true_false_dict).infer_objects(copy=False)
df['Services_MRI'] = df['Services_MRI'].replace(true_false_dict).infer_objects(copy=False)
df['Marital_Married'] = df['Marital_Married'].replace(true_false_dict).infer_objects(copy=False)
df['Marital_Never_Married'] =
df['Marital_Never_Married'].replace(true_false_dict).infer_objects(copy=False)
df['Marital_Separated'] =
df['Marital_Separated'].replace(true_false_dict).infer_objects(copy=False)
df['Marital_Widowed'] =
df['Marital_Widowed'].replace(true_false_dict).infer_objects(copy=False)
```

The “Complication_risk” variable was re-expressed using ordinal encoding because its values—“Low,” “Medium,” and “High”—follow a logical order. Using OrdinalEncoder from sklearn.preprocessing, the values were transformed to integers, where “Low” = 0, “Medium” = 1, and “High” = 2:

```
# Ordinal encoding for “Complication_risk”
# Code adapted from GeeksforGeeks:
# GeeksforGeeks. (n.d.). How to perform ordinal encoding using sklearn. Retrieved December
10, 2024, from https://www.geeksforgeeks.org/how-to-perform-ordinal-encoding-using-sklearn/

# Initialize and fit encoder
encoder = OrdinalEncoder(categories=[['Low', 'Medium', 'High']])
df['Complication_risk'] = encoder.fit_transform(df[['Complication_risk']])

# Convert values to integer
df['Complication_risk'] = df['Complication_risk'].astype(int)

# Print result
print(df[['Complication_risk']])
```

Section C5) Prepared Data Set

Cleaned and prepared data set attached and saved as “updated_medical.csv”

Part IV: Model Comparison and Analysis

Section D1) Initial Model

Multiple linear regression is used to model the relationship between a dependent variable and multiple independent variables (JMP, n.d.). Below is the code for the initial model:

```
# Start initial multiple linear regression model

# Set up your independent and dependent variables
X = df[[”Age”, “Income”, “Marital_Married”, “Marital_Never_Married”, “Marital_Separated”,
“Marital_Widowed”, “Gender_Male”, “Gender_Nonbinary”,
        “Area_Suburban”, “Area_Urban”, “TotalCharge”, “Children”,
“Initial_admin_Emergency_Admission”, “Initial_admin_Observation_Admission”,
        “Doc_visits”, “Services_CT_Scan”, “Services_Intravenous”, “Services_MRI”,
“HighBlood”, “Stroke”, “Complication_risk”, “Overweight”, “Arthritis”, “Diabetes”,
        “Hyperlipidemia”, “BackPain”, “Anxiety”, “Allergic_rhinitis”, “Reflux_esophagitis”,
“VitD_levels”, “Asthma”]]
y = df['Initial_days']

from statsmodels.stats.outliers_influence import variance_inflation_factor
import statsmodels.api as sm

# Calculate VIFs for each independent variable
vif_data = pd.DataFrame({
        'Variable': X.columns,
        'VIF': [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
})

# Display the VIF DataFrame
print(vif_data)

# Add a constant term to the independent variable matrix
```

```
X = sm.add_constant(X)

# Fit an OLS regression model on the data
initial_model = sm.OLS(y, X).fit()

# Print the model summary
print(initial_model.summary())
```

	Variable	VIF
0	Age	7.414393
1	Income	2.972303
2	Marital_Married	2.002355
3	Marital_Never_Married	1.975333
4	Marital_Separated	1.975439
5	Marital_Widowed	2.008886
6	Gender_Male	1.934757
7	Gender_Nonbinary	1.045707
8	Area_Suburban	1.977834
9	Area_Urban	1.977721
10	TotalCharge	6.871342
11	Children	1.936926
12	Initial_admin_Emergency_Admission	3.018355
13	Initial_admin_Observation_Admission	1.953101
14	Doc_visits	20.240376
15	Services_CT_Scan	1.233710
16	Services_Intravenous	1.589342
17	Services_MRI	1.072562
18	HighBlood	1.693950
19	Stroke	1.249203
20	Complication_risk	3.344253
21	Overweight	3.396865
22	Arthritis	1.558945
23	Diabetes	1.376959
24	Hyperlipidemia	1.506040
25	BackPain	1.703525
26	Anxiety	1.475681
27	Allergic_rhinitis	1.646759
28	Reflux_esophagitis	1.699297
29	VitD_levels	34.546984
30	Asthma	1.406712

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Initial_days   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                 1.451e+05
Date:                Fri, 27 Dec 2024   Prob (F-statistic):               0.00
Time:                        17:53:19   Log-Likelihood:                -16315.
No. Observations:               10000   AIC:                         3.269e+04
Df Residuals:                    9968   BIC:                         3.293e+04
Df Model:                          31
Covariance Type:            nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                  -21.1762      0.145   -146.151      0.000     -21.460     -20.892
Age                                     -0.0003      0.001     -0.471      0.638      -0.001       0.001
Income                              -3.549e-07   4.35e-07     -0.816      0.415   -1.21e-06    4.98e-07
Marital_Married                         -0.0280      0.039     -0.711      0.477      -0.105       0.049
Marital_Never_Married                   -0.0107      0.040     -0.271      0.786      -0.088       0.067
Marital_Separated                       -0.0211      0.039     -0.534      0.593      -0.098       0.056
Marital_Widowed                         -0.0087      0.039     -0.223      0.824      -0.086       0.068
Gender_Male                              0.0227      0.025      0.905      0.366      -0.026       0.072
Gender_Nonbinary                         0.0015      0.087      0.017      0.986      -0.168       0.171
Area_Suburban                            0.0381      0.030      1.256      0.209      -0.021       0.098
Area_Urban                              -0.0083      0.030     -0.273      0.785      -0.068       0.051
TotalCharge                              0.0122   5.75e-06   2117.085      0.000       0.012       0.012
Children                                -0.0025      0.006     -0.438      0.661      -0.014       0.009
Initial_admin_Emergency_Admission       -6.2332      0.030   -204.828      0.000      -6.293      -6.174
Initial_admin_Observation_Admission      0.0639      0.035      1.809      0.071      -0.005       0.133
Doc_visits                              -0.0083      0.012     -0.700      0.484      -0.032       0.015
Services_CT_Scan                         0.0237      0.039      0.602      0.547      -0.053       0.101
Services_Intravenous                    -0.0091      0.028     -0.326      0.744      -0.064       0.046
Services_MRI                             0.0726      0.066      1.102      0.271      -0.057       0.202
HighBlood                               -1.3239      0.025    -52.445      0.000      -1.373      -1.274
Stroke                                   0.0037      0.031      0.121      0.904      -0.057       0.065
Complication_risk                       -2.7801      0.017   -163.304      0.000      -2.813      -2.747
Overweight                               0.0473      0.027      1.730      0.084      -0.006       0.101
Arthritis                               -0.8331      0.026    -32.154      0.000      -0.884      -0.782
Diabetes                                -0.9229      0.028    -33.145      0.000      -0.977      -0.868
Hyperlipidemia                          -1.1148      0.026    -42.462      0.000      -1.166      -1.063
BackPain                                -1.0645      0.025    -42.144      0.000      -1.114      -1.015
Anxiety                                 -1.0325      0.027    -38.860      0.000      -1.085      -0.980
Allergic_rhinitis                       -0.7823      0.025    -30.814      0.000      -0.832      -0.733
Reflux_esophagitis                      -0.7371      0.025    -29.241      0.000      -0.786      -0.688
VitD_levels                              0.0068      0.006      1.100      0.271      -0.005       0.019
Asthma                                   0.0129      0.027      0.473      0.636      -0.041       0.067
==============================================================================
Omnibus:                    37550.478   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1471.585
Skew:                           0.121   Prob(JB):                         0.00
Kurtosis:                       1.136   Cond. No.                     5.86e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.86e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Section D2) Justification of Model Reduction

Features in the initial model are reduced using backward stepwise elimination. This method starts with all initial features and iteratively removes the least significant feature until no further improvement is needed (Hayes, 2022). In this process, features were removed based on two criteria: the Variance Inflation Factor (VIF) and the p-value. The VIF threshold is set at 10, and the p-value threshold is set at 0.05. The process iterates until no features exceed a VIF value of 10 or a p-value greater than 0.05. At this point, all remaining variables are considered statistically significant and independent.

The VIF is a tool for detecting multicollinearity in regression analysis by measuring how much the variance of a regression coefficient is inflated due to correlations among predictors (Singh, 2024). Multicollinearity occurs when independent variables are highly correlated, leading to

unstable coefficient estimates and inflated standard errors. (Frost, n.d.) A VIF value exceeding 10 is considered an indicator of severe multicollinearity. When this occurs, removing the variable is necessary to stabilize the model (Singh, 2024). Eliminating variables with high VIF values ensures that the remaining predictors independently contribute to explaining the variation in the lengths of initial hospital stays.

In addition to addressing multicollinearity, backward stepwise elimination using p-values ensures that only predictors with statistically significant contributions remain in the model. Backward stepwise elimination removes irrelevant variables that could warp the relationships between the predictors and the outcome but also ensures that the final model is more straightforward, easier to interpret, and directly tied to the research question.

Section D3) Reduced Linear Regression Model

Removal of "VitD_levels" with VIF, value of 34.546984 and p-value of 0.271:

	Variable	VIF
0	Age	6.902773
1	Income	2.919162
2	Marital_Married	1.955396
3	Marital_Never_Married	1.922662
4	Marital_Separated	1.922430
5	Marital_Widowed	1.959924
6	Gender_Male	1.914202
7	Gender_Nonbinary	1.044617
8	Area_Suburban	1.952489
9	Area_Urban	1.941257
10	TotalCharge	6.585008
11	Children	1.918612
12	Initial_admin_Emergency_Admission	2.941608
13	Initial_admin_Observation_Admission	1.915410
14	Doc_visits	14.805456
15	Services_CT_Scan	1.230288
16	Services_Intravenous	1.577437
17	Services_MRI	1.071302
18	HighBlood	1.686852
19	Stroke	1.244853
20	Complication_risk	3.279340
21	Overweight	3.308734
22	Arthritis	1.553247
23	Diabetes	1.374949
24	Hyperlipidemia	1.499017
25	BackPain	1.697913
26	Anxiety	1.469840
27	Allergic_rhinitis	1.637997
28	Reflux_esophagitis	1.689121
29	Asthma	1.399203

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	1.500e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:20	Log-Likelihood:	-16316.
No. Observations:	10000	AIC:	3.269e+04
Df Residuals:	9969	BIC:	3.292e+04
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.0556	0.095	-222.035	0.000	-21.242	-20.870
Age	-0.0003	0.001	-0.460	0.646	-0.001	0.001
Income	-3.61e-07	4.35e-07	-0.830	0.407	-1.21e-06	4.92e-07
Marital_Married	-0.0281	0.039	-0.714	0.475	-0.105	0.049
Marital_Never_Married	-0.0110	0.040	-0.278	0.781	-0.088	0.066
Marital_Separated	-0.0213	0.039	-0.538	0.590	-0.099	0.056
Marital_Widowed	-0.0090	0.039	-0.230	0.818	-0.086	0.068
Gender_Male	0.0224	0.025	0.892	0.372	-0.027	0.072
Gender_Nonbinary	0.0020	0.087	0.024	0.981	-0.168	0.172
Area_Suburban	0.0384	0.030	1.266	0.205	-0.021	0.098
Area_Urban	-0.0072	0.030	-0.237	0.812	-0.067	0.052
TotalCharge	0.0122	5.75e-06	2117.076	0.000	0.012	0.012
Children	-0.0025	0.006	-0.428	0.669	-0.014	0.009
Initial_admin_Emergency_Admission	-6.2325	0.030	-204.855	0.000	-6.292	-6.173
Initial_admin_Observation_Admission	0.0640	0.035	1.811	0.070	-0.005	0.133
Doc_visits	-0.0082	0.012	-0.689	0.491	-0.031	0.015
Services_CT_Scan	0.0240	0.039	0.609	0.542	-0.053	0.101
Services_Intravenous	-0.0092	0.028	-0.329	0.743	-0.064	0.046
Services_MRI	0.0717	0.066	1.089	0.276	-0.057	0.201
HighBlood	-1.3238	0.025	-52.440	0.000	-1.373	-1.274
Stroke	0.0040	0.031	0.128	0.898	-0.057	0.065
Complication_risk	-2.7800	0.017	-163.300	0.000	-2.813	-2.747
Overweight	0.0476	0.027	1.741	0.082	-0.006	0.101
Arthritis	-0.8331	0.026	-32.154	0.000	-0.884	-0.782
Diabetes	-0.9237	0.028	-33.183	0.000	-0.978	-0.869
Hyperlipidemia	-1.1151	0.026	-42.477	0.000	-1.167	-1.064
BackPain	-1.0648	0.025	-42.157	0.000	-1.114	-1.015
Anxiety	-1.0323	0.027	-38.853	0.000	-1.084	-0.980
Allergic_rhinitis	-0.7825	0.025	-30.820	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7375	0.025	-29.259	0.000	-0.787	-0.688
Asthma	0.0131	0.027	0.479	0.632	-0.041	0.067

Omnibus:	37538.682	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1472.276
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.136	Cond. No.	3.97e+05

Removal of "Doc_visits" with VIF, value of 14.805456 and p-value of 0.491:

	Variable	VIF
0	Age	6.335914
1	Income	2.836570
2	Marital_Married	1.910406
3	Marital_Never_Married	1.874517
4	Marital_Separated	1.870596
5	Marital_Widowed	1.910568
6	Gender_Male	1.891076
7	Gender_Nonbinary	1.043836
8	Area_Suburban	1.920072
9	Area_Urban	1.913063
10	TotalCharge	6.286051
11	Children	1.902539
12	Initial_admin_Emergency_Admission	2.852128
13	Initial_admin_Observation_Admission	1.857556
14	Services_CT_Scan	1.225114
15	Services_Intravenous	1.566901
16	Services_MRI	1.070506
17	HighBlood	1.677375
18	Stroke	1.241343
19	Complication_risk	3.193883
20	Overweight	3.201971
21	Arthritis	1.547163
22	Diabetes	1.367176
23	Hyperlipidemia	1.496015
24	BackPain	1.687804
25	Anxiety	1.464680
26	Allergic_rhinitis	1.626369
27	Reflux_esophagitis	1.677208
28	Asthma	1.395864

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	1.552e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:20	Log-Likelihood:	-16316.
No. Observations:	10000	AIC:	3.269e+04
Df Residuals:	9970	BIC:	3.291e+04
Df Model:	29		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.0962	0.074	-283.580	0.000	-21.242	-20.950
Age	-0.0003	0.001	-0.465	0.642	-0.001	0.001
Income	-3.652e-07	4.35e-07	-0.840	0.401	-1.22e-06	4.87e-07
Marital_Married	-0.0278	0.039	-0.708	0.479	-0.105	0.049
Marital_Never_Married	-0.0106	0.040	-0.268	0.789	-0.088	0.067
Marital_Separated	-0.0210	0.039	-0.532	0.595	-0.098	0.056
Marital_Widowed	-0.0088	0.039	-0.224	0.823	-0.086	0.068
Gender_Male	0.0225	0.025	0.895	0.371	-0.027	0.072
Gender_Nonbinary	0.0022	0.087	0.025	0.980	-0.168	0.172
Area_Suburban	0.0381	0.030	1.258	0.208	-0.021	0.098
Area_Urban	-0.0072	0.030	-0.237	0.812	-0.067	0.052
TotalCharge	0.0122	5.75e-06	2117.173	0.000	0.012	0.012
Children	-0.0024	0.006	-0.426	0.670	-0.014	0.009
Initial_admin_Emergency_Admission	-6.2328	0.030	-204.901	0.000	-6.292	-6.173
Initial_admin_Observation_Admission	0.0634	0.035	1.796	0.073	-0.006	0.133
Services_CT_Scan	0.0237	0.039	0.602	0.547	-0.053	0.101
Services_Intravenous	-0.0090	0.028	-0.323	0.747	-0.064	0.046
Services_MRI	0.0723	0.066	1.098	0.272	-0.057	0.201
HighBlood	-1.3240	0.025	-52.449	0.000	-1.373	-1.274
Stroke	0.0040	0.031	0.130	0.896	-0.057	0.065
Complication_risk	-2.7802	0.017	-163.327	0.000	-2.814	-2.747
Overweight	0.0473	0.027	1.733	0.083	-0.006	0.101
Arthritis	-0.8331	0.026	-32.154	0.000	-0.884	-0.782
Diabetes	-0.9239	0.028	-33.197	0.000	-0.978	-0.869
Hyperlipidemia	-1.1146	0.026	-42.475	0.000	-1.166	-1.063
BackPain	-1.0649	0.025	-42.165	0.000	-1.114	-1.015
Anxiety	-1.0323	0.027	-38.853	0.000	-1.084	-0.980
Allergic_rhinitis	-0.7825	0.025	-30.823	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7374	0.025	-29.257	0.000	-0.787	-0.688
Asthma	0.0134	0.027	0.492	0.623	-0.040	0.067

Omnibus:	37533.061	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1472.579
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.136	Cond. No.	3.51e+05

Removal of "Gender_Nonbinary", with p-value of 0.980:

	Variable	VIF
0	Age	6.332795
1	Income	2.835991
2	Marital_Married	1.910048
3	Marital_Never_Married	1.874229
4	Marital_Separated	1.870585
5	Marital_Widowed	1.910527
6	Gender_Male	1.856301
7	Area_Suburban	1.919731
8	Area_Urban	1.912264
9	TotalCharge	6.282818
10	Children	1.902502
11	Initial_admin_Emergency_Admission	2.852030
12	Initial_admin_Observation_Admission	1.857519
13	Services_CT_Scan	1.225099
14	Services_Intravenous	1.566797
15	Services_MRI	1.069953
16	HighBlood	1.676680
17	Stroke	1.241274
18	Complication_risk	3.192365
19	Overweight	3.201000
20	Arthritis	1.546546
21	Diabetes	1.367166
22	Hyperlipidemia	1.495495
23	BackPain	1.686834
24	Anxiety	1.464603
25	Allergic_rhinitis	1.626369
26	Reflux_esophagitis	1.677029
27	Asthma	1.395768

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	1.607e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:21	Log-Likelihood:	-16316.
No. Observations:	10000	AIC:	3.269e+04
Df Residuals:	9971	BIC:	3.290e+04
Df Model:	28		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.0961	0.074	-283.900	0.000	-21.242	-20.950
Age	-0.0003	0.001	-0.465	0.642	-0.001	0.001
Income	-3.652e-07	4.35e-07	-0.840	0.401	-1.22e-06	4.87e-07
Marital_Married	-0.0279	0.039	-0.709	0.479	-0.105	0.049
Marital_Never_Married	-0.0106	0.040	-0.269	0.788	-0.088	0.067
Marital_Separated	-0.0210	0.039	-0.533	0.594	-0.098	0.056
Marital_Widowed	-0.0088	0.039	-0.224	0.823	-0.086	0.068
Gender_Male	0.0224	0.025	0.901	0.368	-0.026	0.071
Area_Suburban	0.0381	0.030	1.258	0.208	-0.021	0.098
Area_Urban	-0.0072	0.030	-0.237	0.813	-0.067	0.052
TotalCharge	0.0122	5.75e-06	2117.312	0.000	0.012	0.012
Children	-0.0024	0.006	-0.426	0.670	-0.014	0.009
Initial_admin_Emergency_Admission	-6.2328	0.030	-204.915	0.000	-6.292	-6.173
Initial_admin_Observation_Admission	0.0634	0.035	1.796	0.073	-0.006	0.133
Services_CT_Scan	0.0237	0.039	0.602	0.547	-0.053	0.101
Services_Intravenous	-0.0091	0.028	-0.323	0.746	-0.064	0.046
Services_MRI	0.0723	0.066	1.098	0.272	-0.057	0.201
HighBlood	-1.3240	0.025	-52.457	0.000	-1.373	-1.274
Stroke	0.0040	0.031	0.130	0.896	-0.057	0.065
Complication_risk	-2.7802	0.017	-163.343	0.000	-2.814	-2.747
Overweight	0.0473	0.027	1.733	0.083	-0.006	0.101
Arthritis	-0.8330	0.026	-32.159	0.000	-0.884	-0.782
Diabetes	-0.9239	0.028	-33.199	0.000	-0.978	-0.869
Hyperlipidemia	-1.1146	0.026	-42.481	0.000	-1.166	-1.063
BackPain	-1.0649	0.025	-42.174	0.000	-1.114	-1.015
Anxiety	-1.0323	0.027	-38.855	0.000	-1.084	-0.980
Allergic_rhinitis	-0.7825	0.025	-30.825	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7374	0.025	-29.263	0.000	-0.787	-0.688
Asthma	0.0135	0.027	0.492	0.623	-0.040	0.067

Omnibus:	37533.033	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1472.581
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.136	Cond. No.	3.31e+05

Removal of "Marital_Widowed" with p-value of 0.823:

	Variable	VIF
0	Age	6.224059
1	Income	2.827437
2	Marital_Married	1.480628
3	Marital_Never_Married	1.459252
4	Marital_Separated	1.457978
5	Gender_Male	1.851777
6	Area_Suburban	1.916518
7	Area_Urban	1.906507
8	TotalCharge	6.191979
9	Children	1.901121
10	Initial_admin_Emergency_Admission	2.834054
11	Initial_admin_Observation_Admission	1.846770
12	Services_CT_Scan	1.224790
13	Services_Intravenous	1.561726
14	Services_MRI	1.069258
15	HighBlood	1.676393
16	Stroke	1.241158
17	Complication_risk	3.176529
18	Overweight	3.174517
19	Arthritis	1.544114
20	Diabetes	1.366839
21	Hyperlipidemia	1.492077
22	BackPain	1.686254
23	Anxiety	1.463870
24	Allergic_rhinitis	1.625681
25	Reflux_esophagitis	1.674129
26	Asthma	1.393314

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	1.667e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:22	Log-Likelihood:	-16316.
No. Observations:	10000	AIC:	3.269e+04
Df Residuals:	9972	BIC:	3.289e+04
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1004	0.072	-294.150	0.000	-21.241	-20.960
Age	-0.0003	0.001	-0.465	0.642	-0.001	0.001
Income	-3.639e-07	4.35e-07	-0.837	0.403	-1.22e-06	4.89e-07
Marital_Married	-0.0234	0.034	-0.691	0.490	-0.090	0.043
Marital_Never_Married	-0.0061	0.034	-0.180	0.857	-0.073	0.061
Marital_Separated	-0.0165	0.034	-0.486	0.627	-0.083	0.050
Gender_Male	0.0224	0.025	0.901	0.368	-0.026	0.071
Area_Suburban	0.0382	0.030	1.261	0.207	-0.021	0.098
Area_Urban	-0.0072	0.030	-0.237	0.812	-0.067	0.052
TotalCharge	0.0122	5.75e-06	2117.866	0.000	0.012	0.012
Children	-0.0024	0.006	-0.423	0.673	-0.014	0.009
Initial_admin_Emergency_Admission	-6.2329	0.030	-204.932	0.000	-6.292	-6.173
Initial_admin_Observation_Admission	0.0634	0.035	1.795	0.073	-0.006	0.133
Services_CT_Scan	0.0237	0.039	0.604	0.546	-0.053	0.101
Services_Intravenous	-0.0092	0.028	-0.327	0.743	-0.064	0.046
Services_MRI	0.0722	0.066	1.096	0.273	-0.057	0.201
HighBlood	-1.3239	0.025	-52.463	0.000	-1.373	-1.274
Stroke	0.0041	0.031	0.133	0.894	-0.057	0.065
Complication_risk	-2.7802	0.017	-163.352	0.000	-2.814	-2.747
Overweight	0.0473	0.027	1.730	0.084	-0.006	0.101
Arthritis	-0.8331	0.026	-32.165	0.000	-0.884	-0.782
Diabetes	-0.9239	0.028	-33.200	0.000	-0.978	-0.869
Hyperlipidemia	-1.1147	0.026	-42.494	0.000	-1.166	-1.063
BackPain	-1.0649	0.025	-42.176	0.000	-1.114	-1.015
Anxiety	-1.0323	0.027	-38.856	0.000	-1.084	-0.980
Allergic_rhinitis	-0.7824	0.025	-30.827	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7374	0.025	-29.265	0.000	-0.787	-0.688
Asthma	0.0134	0.027	0.489	0.625	-0.040	0.067

Omnibus:	37532.601	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1472.602
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.136	Cond. No.	3.07e+05

Removal of "Stroke" with p-value of 0.894:

	Variable	VIF
0	Age	6.203040
1	Income	2.825531
2	Marital_Married	1.480501
3	Marital_Never_Married	1.459050
4	Marital_Separated	1.457389
5	Gender_Male	1.851492
6	Area_Suburban	1.916156
7	Area_Urban	1.905794
8	TotalCharge	6.185736
9	Children	1.900412
10	Initial_admin_Emergency_Admission	2.833223
11	Initial_admin_Observation_Admission	1.845699
12	Services_CT_Scan	1.224471
13	Services_Intravenous	1.561696
14	Services_MRI	1.069258
15	HighBlood	1.675871
16	Complication_risk	3.174708
17	Overweight	3.172501
18	Arthritis	1.543997
19	Diabetes	1.366516
20	Hyperlipidemia	1.492043
21	BackPain	1.685947
22	Anxiety	1.463834
23	Allergic_rhinitis	1.625603
24	Reflux_esophagitis	1.673848
25	Asthma	1.393132

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	1.731e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:22	Log-Likelihood:	-16316.
No. Observations:	10000	AIC:	3.269e+04
Df Residuals:	9973	BIC:	3.288e+04
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.0996	0.071	-295.332	0.000	-21.240	-20.960
Age	-0.0003	0.001	-0.464	0.643	-0.001	0.001
Income	-3.637e-07	4.35e-07	-0.836	0.403	-1.22e-06	4.89e-07
Marital_Married	-0.0234	0.034	-0.691	0.490	-0.090	0.043
Marital_Never_Married	-0.0061	0.034	-0.180	0.857	-0.073	0.061
Marital_Separated	-0.0165	0.034	-0.485	0.627	-0.083	0.050
Gender_Male	0.0224	0.025	0.900	0.368	-0.026	0.071
Area_Suburban	0.0382	0.030	1.260	0.208	-0.021	0.098
Area_Urban	-0.0072	0.030	-0.237	0.812	-0.067	0.052
TotalCharge	0.0122	5.75e-06	2117.976	0.000	0.012	0.012
Children	-0.0024	0.006	-0.422	0.673	-0.014	0.009
Initial_admin_Emergency_Admission	-6.2329	0.030	-204.949	0.000	-6.293	-6.173
Initial_admin_Observation_Admission	0.0634	0.035	1.795	0.073	-0.006	0.133
Services_CT_Scan	0.0238	0.039	0.605	0.545	-0.053	0.101
Services_Intravenous	-0.0092	0.028	-0.330	0.742	-0.064	0.046
Services_MRI	0.0721	0.066	1.095	0.273	-0.057	0.201
HighBlood	-1.3238	0.025	-52.466	0.000	-1.373	-1.274
Complication_risk	-2.7802	0.017	-163.360	0.000	-2.814	-2.747
Overweight	0.0473	0.027	1.730	0.084	-0.006	0.101
Arthritis	-0.8332	0.026	-32.175	0.000	-0.884	-0.782
Diabetes	-0.9238	0.028	-33.202	0.000	-0.978	-0.869
Hyperlipidemia	-1.1148	0.026	-42.503	0.000	-1.166	-1.063
BackPain	-1.0648	0.025	-42.178	0.000	-1.114	-1.015
Anxiety	-1.0323	0.027	-38.864	0.000	-1.084	-0.980
Allergic_rhinitis	-0.7825	0.025	-30.829	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7374	0.025	-29.266	0.000	-0.787	-0.688
Asthma	0.0134	0.027	0.489	0.625	-0.040	0.067

Omnibus:	37532.466	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1472.614
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.136	Cond. No.	3.05e+05

Removal of "Marital_Never_Married" with p-value of 0.857:

	Variable	VIF
0	Age	6.177562
1	Income	2.818036
2	Marital_Married	1.331768
3	Marital_Separated	1.314501
4	Gender_Male	1.848809
5	Area_Suburban	1.915102
6	Area_Urban	1.904308
7	TotalCharge	6.148822
8	Children	1.900236
9	Initial_admin_Emergency_Admission	2.831376
10	Initial_admin_Observation_Admission	1.843980
11	Services_CT_Scan	1.224380
12	Services_Intravenous	1.561182
13	Services_MRI	1.069206
14	HighBlood	1.675676
15	Complication_risk	3.165903
16	Overweight	3.160891
17	Arthritis	1.541676
18	Diabetes	1.365934
19	Hyperlipidemia	1.491432
20	BackPain	1.685276
21	Anxiety	1.463765
22	Allergic_rhinitis	1.625012
23	Reflux_esophagitis	1.672622
24	Asthma	1.392991

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	1.801e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:23	Log-Likelihood:	−16316.
No. Observations:	10000	AIC:	3.268e+04
Df Residuals:	9974	BIC:	3.287e+04
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	−21.1016	0.070	−299.344	0.000	−21.240	−20.963
Age	−0.0003	0.001	−0.461	0.645	−0.001	0.001
Income	−3.643e−07	4.35e−07	−0.838	0.402	−1.22e−06	4.88e−07
Marital_Married	−0.0214	0.032	−0.669	0.503	−0.084	0.041
Marital_Separated	−0.0145	0.032	−0.451	0.652	−0.077	0.048
Gender_Male	0.0223	0.025	0.899	0.369	−0.026	0.071
Area_Suburban	0.0382	0.030	1.262	0.207	−0.021	0.098
Area_Urban	−0.0072	0.030	−0.236	0.813	−0.067	0.052
TotalCharge	0.0122	5.75e−06	2118.257	0.000	0.012	0.012
Children	−0.0024	0.006	−0.419	0.675	−0.014	0.009
Initial_admin_Emergency_Admission	−6.2328	0.030	−204.992	0.000	−6.292	−6.173
Initial_admin_Observation_Admission	0.0635	0.035	1.797	0.072	−0.006	0.133
Services_CT_Scan	0.0238	0.039	0.606	0.545	−0.053	0.101
Services_Intravenous	−0.0092	0.028	−0.329	0.742	−0.064	0.046
Services_MRI	0.0721	0.066	1.096	0.273	−0.057	0.201
HighBlood	−1.3238	0.025	−52.469	0.000	−1.373	−1.274
Complication_risk	−2.7802	0.017	−163.377	0.000	−2.814	−2.747
Overweight	0.0472	0.027	1.728	0.084	−0.006	0.101
Arthritis	−0.8333	0.026	−32.187	0.000	−0.884	−0.783
Diabetes	−0.9239	0.028	−33.204	0.000	−0.978	−0.869
Hyperlipidemia	−1.1148	0.026	−42.505	0.000	−1.166	−1.063
BackPain	−1.0648	0.025	−42.180	0.000	−1.114	−1.015
Anxiety	−1.0323	0.027	−38.865	0.000	−1.084	−0.980
Allergic_rhinitis	−0.7824	0.025	−30.831	0.000	−0.832	−0.733
Reflux_esophagitis	−0.7374	0.025	−29.268	0.000	−0.787	−0.688
Asthma	0.0134	0.027	0.491	0.624	−0.040	0.067

Omnibus:	37532.138	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1472.632
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.136	Cond. No.	3.00e+05

Removal of "Area_Urban" with p-value of 0.813:

	Variable	VIF
0	Age	6.120928
1	Income	2.805702
2	Marital_Married	1.331012
3	Marital_Separated	1.313932
4	Gender_Male	1.846838
5	Area_Suburban	1.482879
6	TotalCharge	6.092709
7	Children	1.898801
8	Initial_admin_Emergency_Admission	2.816292
9	Initial_admin_Observation_Admission	1.834342
10	Services_CT_Scan	1.224130
11	Services_Intravenous	1.559208
12	Services_MRI	1.069185
13	HighBlood	1.673317
14	Complication_risk	3.157718
15	Overweight	3.150544
16	Arthritis	1.541140
17	Diabetes	1.365933
18	Hyperlipidemia	1.489778
19	BackPain	1.684812
20	Anxiety	1.460212
21	Allergic_rhinitis	1.621765
22	Reflux_esophagitis	1.670235
23	Asthma	1.392472

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	1.876e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:23	Log-Likelihood:	-16316.
No. Observations:	10000	AIC:	3.268e+04
Df Residuals:	9975	BIC:	3.286e+04
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1050	0.069	-305.769	0.000	-21.240	-20.970
Age	-0.0003	0.001	-0.459	0.646	-0.001	0.001
Income	-3.653e-07	4.35e-07	-0.840	0.401	-1.22e-06	4.87e-07
Marital_Married	-0.0214	0.032	-0.670	0.503	-0.084	0.041
Marital_Separated	-0.0144	0.032	-0.450	0.653	-0.077	0.048
Gender_Male	0.0224	0.025	0.901	0.368	-0.026	0.071
Area_Suburban	0.0418	0.026	1.588	0.112	-0.010	0.093
TotalCharge	0.0122	5.74e-06	2118.533	0.000	0.012	0.012
Children	-0.0024	0.006	-0.418	0.676	-0.014	0.009
Initial_admin_Emergency_Admission	-6.2329	0.030	-205.028	0.000	-6.292	-6.173
Initial_admin_Observation_Admission	0.0633	0.035	1.794	0.073	-0.006	0.133
Services_CT_Scan	0.0239	0.039	0.607	0.544	-0.053	0.101
Services_Intravenous	-0.0092	0.028	-0.330	0.742	-0.064	0.046
Services_MRI	0.0723	0.066	1.098	0.272	-0.057	0.201
HighBlood	-1.3239	0.025	-52.479	0.000	-1.373	-1.274
Complication_risk	-2.7802	0.017	-163.385	0.000	-2.814	-2.747
Overweight	0.0472	0.027	1.730	0.084	-0.006	0.101
Arthritis	-0.8332	0.026	-32.188	0.000	-0.884	-0.783
Diabetes	-0.9237	0.028	-33.209	0.000	-0.978	-0.869
Hyperlipidemia	-1.1149	0.026	-42.511	0.000	-1.166	-1.063
BackPain	-1.0648	0.025	-42.181	0.000	-1.114	-1.015
Anxiety	-1.0324	0.027	-38.889	0.000	-1.084	-0.980
Allergic_rhinitis	-0.7825	0.025	-30.840	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7374	0.025	-29.272	0.000	-0.787	-0.688
Asthma	0.0134	0.027	0.491	0.623	-0.040	0.067

Omnibus:	37531.762	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1472.650
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.136	Cond. No.	2.94e+05

Removal of "Services_Intravenous" with p-value of 0.742:

	Variable	VIF
0	Age	6.071288
1	Income	2.799136
2	Marital_Married	1.330984
3	Marital_Separated	1.313923
4	Gender_Male	1.846068
5	Area_Suburban	1.481618
6	TotalCharge	6.077563
7	Children	1.896426
8	Initial_admin_Emergency_Admission	2.809657
9	Initial_admin_Observation_Admission	1.829986
10	Services_CT_Scan	1.144948
11	Services_MRI	1.044896
12	HighBlood	1.673020
13	Complication_risk	3.151143
14	Overweight	3.141695
15	Arthritis	1.540704
16	Diabetes	1.365043
17	Hyperlipidemia	1.489467
18	BackPain	1.684647
19	Anxiety	1.459473
20	Allergic_rhinitis	1.621081
21	Reflux_esophagitis	1.670163
22	Asthma	1.392410

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	1.957e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:23	Log-Likelihood:	-16316.
No. Observations:	10000	AIC:	3.268e+04
Df Residuals:	9976	BIC:	3.285e+04
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1085	0.068	-309.554	0.000	-21.242	-20.975
Age	-0.0003	0.001	-0.462	0.644	-0.001	0.001
Income	-3.66e-07	4.35e-07	-0.842	0.400	-1.22e-06	4.86e-07
Marital_Married	-0.0213	0.032	-0.667	0.505	-0.084	0.041
Marital_Separated	-0.0142	0.032	-0.443	0.658	-0.077	0.049
Gender_Male	0.0224	0.025	0.904	0.366	-0.026	0.071
Area_Suburban	0.0417	0.026	1.585	0.113	-0.010	0.093
TotalCharge	0.0122	5.74e-06	2118.794	0.000	0.012	0.012
Children	-0.0024	0.006	-0.422	0.673	-0.014	0.009
Initial_admin_Emergency_Admission	-6.2329	0.030	-205.039	0.000	-6.293	-6.173
Initial_admin_Observation_Admission	0.0633	0.035	1.793	0.073	-0.006	0.132
Services_CT_Scan	0.0273	0.038	0.720	0.472	-0.047	0.102
Services_MRI	0.0757	0.065	1.165	0.244	-0.052	0.203
HighBlood	-1.3238	0.025	-52.481	0.000	-1.373	-1.274
Complication_risk	-2.7802	0.017	-163.394	0.000	-2.814	-2.747
Overweight	0.0472	0.027	1.728	0.084	-0.006	0.101
Arthritis	-0.8332	0.026	-32.189	0.000	-0.884	-0.782
Diabetes	-0.9238	0.028	-33.214	0.000	-0.978	-0.869
Hyperlipidemia	-1.1148	0.026	-42.512	0.000	-1.166	-1.063
BackPain	-1.0647	0.025	-42.182	0.000	-1.114	-1.015
Anxiety	-1.0325	0.027	-38.893	0.000	-1.085	-0.980
Allergic_rhinitis	-0.7825	0.025	-30.841	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7373	0.025	-29.272	0.000	-0.787	-0.688
Asthma	0.0135	0.027	0.495	0.621	-0.040	0.067

Omnibus:	37530.769	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1472.704
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.136	Cond. No.	2.89e+05

Removal of "Children" with p-value of 0.673:

	Variable	VIF
0	Age	6.016704
1	Income	2.791062
2	Marital_Married	1.329514
3	Marital_Separated	1.313739
4	Gender_Male	1.844747
5	Area_Suburban	1.479376
6	TotalCharge	6.027632
7	Initial_admin_Emergency_Admission	2.802113
8	Initial_admin_Observation_Admission	1.826402
9	Services_CT_Scan	1.144844
10	Services_MRI	1.044896
11	HighBlood	1.672314
12	Complication_risk	3.144444
13	Overweight	3.137487
14	Arthritis	1.539733
15	Diabetes	1.362841
16	Hyperlipidemia	1.489092
17	BackPain	1.684490
18	Anxiety	1.458540
19	Allergic_rhinitis	1.621072
20	Reflux_esophagitis	1.668642
21	Asthma	1.391791

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	2.047e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:24	Log-Likelihood:	-16316.
No. Observations:	10000	AIC:	3.268e+04
Df Residuals:	9977	BIC:	3.284e+04
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1131	0.067	-313.604	0.000	-21.245	-20.981
Age	-0.0003	0.001	-0.466	0.641	-0.001	0.001
Income	-3.675e-07	4.35e-07	-0.845	0.398	-1.22e-06	4.85e-07
Marital_Married	-0.0215	0.032	-0.674	0.500	-0.084	0.041
Marital_Separated	-0.0141	0.032	-0.440	0.660	-0.077	0.049
Gender_Male	0.0225	0.025	0.906	0.365	-0.026	0.071
Area_Suburban	0.0415	0.026	1.577	0.115	-0.010	0.093
TotalCharge	0.0122	5.74e-06	2119.425	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2330	0.030	-205.052	0.000	-6.293	-6.173
Initial_admin_Observation_Admission	0.0633	0.035	1.793	0.073	-0.006	0.132
Services_CT_Scan	0.0273	0.038	0.720	0.472	-0.047	0.102
Services_MRI	0.0759	0.065	1.168	0.243	-0.052	0.203
HighBlood	-1.3239	0.025	-52.483	0.000	-1.373	-1.274
Complication_risk	-2.7802	0.017	-163.402	0.000	-2.814	-2.747
Overweight	0.0473	0.027	1.734	0.083	-0.006	0.101
Arthritis	-0.8333	0.026	-32.193	0.000	-0.884	-0.783
Diabetes	-0.9241	0.028	-33.234	0.000	-0.979	-0.870
Hyperlipidemia	-1.1148	0.026	-42.512	0.000	-1.166	-1.063
BackPain	-1.0646	0.025	-42.182	0.000	-1.114	-1.015
Anxiety	-1.0326	0.027	-38.899	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7823	0.025	-30.840	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7374	0.025	-29.276	0.000	-0.787	-0.688
Asthma	0.0135	0.027	0.494	0.621	-0.040	0.067

Omnibus:	37529.421	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1472.792
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.136	Cond. No.	2.86e+05

Removal of "Marital_Separated" with p-value of 0.660:

	Variable	VIF
0	Age	5.989723
1	Income	2.788982
2	Marital_Married	1.251900
3	Gender_Male	1.844172
4	Area_Suburban	1.479206
5	TotalCharge	6.005145
6	Initial_admin_Emergency_Admission	2.798666
7	Initial_admin_Observation_Admission	1.824116
8	Services_CT_Scan	1.144723
9	Services_MRI	1.044887
10	HighBlood	1.671885
11	Complication_risk	3.143617
12	Overweight	3.134253
13	Arthritis	1.538620
14	Diabetes	1.361983
15	Hyperlipidemia	1.489081
16	BackPain	1.683650
17	Anxiety	1.457652
18	Allergic_rhinitis	1.620713
19	Reflux_esophagitis	1.668144
20	Asthma	1.391646

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	2.144e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:24	Log-Likelihood:	-16317.
No. Observations:	10000	AIC:	3.268e+04
Df Residuals:	9978	BIC:	3.284e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1167	0.067	-315.994	0.000	-21.248	-20.986
Age	-0.0003	0.001	-0.466	0.641	-0.001	0.001
Income	-3.66e-07	4.35e-07	-0.842	0.400	-1.22e-06	4.86e-07
Marital_Married	-0.0180	0.031	-0.582	0.560	-0.079	0.043
Gender_Male	0.0226	0.025	0.909	0.364	-0.026	0.071
Area_Suburban	0.0415	0.026	1.579	0.114	-0.010	0.093
TotalCharge	0.0122	5.74e-06	2119.594	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2330	0.030	-205.060	0.000	-6.293	-6.173
Initial_admin_Observation_Admission	0.0633	0.035	1.793	0.073	-0.006	0.132
Services_CT_Scan	0.0272	0.038	0.718	0.473	-0.047	0.102
Services_MRI	0.0759	0.065	1.168	0.243	-0.051	0.203
HighBlood	-1.3239	0.025	-52.486	0.000	-1.373	-1.274
Complication_risk	-2.7801	0.017	-163.425	0.000	-2.813	-2.747
Overweight	0.0474	0.027	1.736	0.083	-0.006	0.101
Arthritis	-0.8334	0.026	-32.203	0.000	-0.884	-0.783
Diabetes	-0.9242	0.028	-33.242	0.000	-0.979	-0.870
Hyperlipidemia	-1.1146	0.026	-42.512	0.000	-1.166	-1.063
BackPain	-1.0647	0.025	-42.188	0.000	-1.114	-1.015
Anxiety	-1.0327	0.027	-38.908	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7823	0.025	-30.840	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7373	0.025	-29.276	0.000	-0.787	-0.688
Asthma	0.0135	0.027	0.495	0.620	-0.040	0.067

Omnibus:	37527.594	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1472.903
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.135	Cond. No.	2.83e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.83e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Removal of "Age" with p-value of 0.641:

	Variable	VIF
0	Income	2.732328
1	Marital_Married	1.248239
2	Gender_Male	1.830114
3	Area_Suburban	1.467812
4	TotalCharge	5.630671
5	Initial_admin_Emergency_Admission	2.739359
6	Initial_admin_Observation_Admission	1.787507
7	Services_CT_Scan	1.142768
8	Services_MRI	1.043851
9	HighBlood	1.662911
10	Complication_risk	3.070010
11	Overweight	3.053521
12	Arthritis	1.530357
13	Diabetes	1.356176
14	Hyperlipidemia	1.481460
15	BackPain	1.670373
16	Anxiety	1.450932
17	Allergic_rhinitis	1.607705
18	Reflux_esophagitis	1.661699
19	Asthma	1.384193

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	2.252e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:24	Log-Likelihood:	-16317.
No. Observations:	10000	AIC:	3.268e+04
Df Residuals:	9979	BIC:	3.283e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1315	0.059	-359.607	0.000	-21.247	-21.016
Income	-3.636e-07	4.35e-07	-0.837	0.403	-1.22e-06	4.88e-07
Marital_Married	-0.0182	0.031	-0.588	0.556	-0.079	0.042
Gender_Male	0.0228	0.025	0.917	0.359	-0.026	0.071
Area_Suburban	0.0413	0.026	1.570	0.116	-0.010	0.093
TotalCharge	0.0122	5.74e-06	2119.942	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2328	0.030	-205.080	0.000	-6.292	-6.173
Initial_admin_Observation_Admission	0.0635	0.035	1.800	0.072	-0.006	0.133
Services_CT_Scan	0.0271	0.038	0.714	0.475	-0.047	0.101
Services_MRI	0.0757	0.065	1.164	0.244	-0.052	0.203
HighBlood	-1.3239	0.025	-52.492	0.000	-1.373	-1.275
Complication_risk	-2.7801	0.017	-163.433	0.000	-2.813	-2.747
Overweight	0.0475	0.027	1.741	0.082	-0.006	0.101
Arthritis	-0.8335	0.026	-32.208	0.000	-0.884	-0.783
Diabetes	-0.9242	0.028	-33.245	0.000	-0.979	-0.870
Hyperlipidemia	-1.1147	0.026	-42.516	0.000	-1.166	-1.063
BackPain	-1.0649	0.025	-42.207	0.000	-1.114	-1.015
Anxiety	-1.0328	0.027	-38.912	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7824	0.025	-30.848	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7371	0.025	-29.274	0.000	-0.786	-0.688
Asthma	0.0134	0.027	0.491	0.623	-0.040	0.067

Omnibus:	37525.224	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1473.040
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.135	Cond. No.	2.65e+05

Removal of "Asthma" with p-value of 0.623:

	Variable	VIF
0	Income	2.725262
1	Marital_Married	1.248105
2	Gender_Male	1.827961
3	Area_Suburban	1.467810
4	TotalCharge	5.614654
5	Initial_admin_Emergency_Admission	2.734684
6	Initial_admin_Observation_Admission	1.784298
7	Services_CT_Scan	1.142150
8	Services_MRI	1.043807
9	HighBlood	1.661816
10	Complication_risk	3.066242
11	Overweight	3.041085
12	Arthritis	1.530100
13	Diabetes	1.354444
14	Hyperlipidemia	1.481298
15	BackPain	1.668289
16	Anxiety	1.449571
17	Allergic_rhinitis	1.606522
18	Reflux_esophagitis	1.660888

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	2.370e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:25	Log-Likelihood:	-16317.
No. Observations:	10000	AIC:	3.267e+04
Df Residuals:	9980	BIC:	3.282e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1276	0.058	-362.878	0.000	-21.242	-21.014
Income	-3.623e-07	4.35e-07	-0.834	0.404	-1.21e-06	4.9e-07
Marital_Married	-0.0182	0.031	-0.589	0.556	-0.079	0.042
Gender_Male	0.0228	0.025	0.919	0.358	-0.026	0.071
Area_Suburban	0.0410	0.026	1.560	0.119	-0.011	0.093
TotalCharge	0.0122	5.74e-06	2120.226	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2329	0.030	-205.092	0.000	-6.292	-6.173
Initial_admin_Observation_Admission	0.0635	0.035	1.799	0.072	-0.006	0.133
Services_CT_Scan	0.0273	0.038	0.721	0.471	-0.047	0.102
Services_MRI	0.0757	0.065	1.164	0.244	-0.052	0.203
HighBlood	-1.3239	0.025	-52.492	0.000	-1.373	-1.274
Complication_risk	-2.7802	0.017	-163.448	0.000	-2.814	-2.747
Overweight	0.0477	0.027	1.748	0.081	-0.006	0.101
Arthritis	-0.8336	0.026	-32.213	0.000	-0.884	-0.783
Diabetes	-0.9240	0.028	-33.242	0.000	-0.978	-0.870
Hyperlipidemia	-1.1148	0.026	-42.524	0.000	-1.166	-1.063
BackPain	-1.0648	0.025	-42.206	0.000	-1.114	-1.015
Anxiety	-1.0326	0.027	-38.910	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7824	0.025	-30.848	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7371	0.025	-29.275	0.000	-0.786	-0.688

Omnibus:	37522.167	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1473.211
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.135	Cond. No.	2.65e+05

Removal of "Marital_Married" with p-value of 0.556:

	Variable	VIF
0	Income	2.724361
1	Gender_Male	1.825412
2	Area_Suburban	1.467010
3	TotalCharge	5.603008
4	Initial_admin_Emergency_Admission	2.730241
5	Initial_admin_Observation_Admission	1.783120
6	Services_CT_Scan	1.142096
7	Services_MRI	1.043505
8	HighBlood	1.659561
9	Complication_risk	3.061879
10	Overweight	3.037378
11	Arthritis	1.528201
12	Diabetes	1.354444
13	Hyperlipidemia	1.481291
14	BackPain	1.663651
15	Anxiety	1.449138
16	Allergic_rhinitis	1.605680
17	Reflux_esophagitis	1.660881

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	2.502e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:25	Log-Likelihood:	-16317.
No. Observations:	10000	AIC:	3.267e+04
Df Residuals:	9981	BIC:	3.281e+04
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1308	0.058	-364.469	0.000	-21.244	-21.017
Income	-3.594e-07	4.35e-07	-0.827	0.408	-1.21e-06	4.92e-07
Gender_Male	0.0226	0.025	0.909	0.363	-0.026	0.071
Area_Suburban	0.0409	0.026	1.554	0.120	-0.011	0.092
TotalCharge	0.0122	5.74e-06	2120.298	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2330	0.030	-205.112	0.000	-6.293	-6.173
Initial_admin_Observation_Admission	0.0636	0.035	1.801	0.072	-0.006	0.133
Services_CT_Scan	0.0273	0.038	0.721	0.471	-0.047	0.102
Services_MRI	0.0752	0.065	1.157	0.247	-0.052	0.203
HighBlood	-1.3242	0.025	-52.522	0.000	-1.374	-1.275
Complication_risk	-2.7803	0.017	-163.463	0.000	-2.814	-2.747
Overweight	0.0477	0.027	1.747	0.081	-0.006	0.101
Arthritis	-0.8339	0.026	-32.234	0.000	-0.885	-0.783
Diabetes	-0.9238	0.028	-33.239	0.000	-0.978	-0.869
Hyperlipidemia	-1.1146	0.026	-42.521	0.000	-1.166	-1.063
BackPain	-1.0653	0.025	-42.261	0.000	-1.115	-1.016
Anxiety	-1.0327	0.027	-38.915	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7825	0.025	-30.854	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7369	0.025	-29.271	0.000	-0.786	-0.688

Omnibus:	37518.929	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1473.384
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.135	Cond. No.	2.64e+05

Removal of "Services_CT_Scan" with p-value of 0.471:

	Variable	VIF
0	Income	2.723684
1	Gender_Male	1.825319
2	Area_Suburban	1.466906
3	TotalCharge	5.591047
4	Initial_admin_Emergency_Admission	2.729195
5	Initial_admin_Observation_Admission	1.783090
6	Services_MRI	1.038184
7	HighBlood	1.658723
8	Complication_risk	3.057543
9	Overweight	3.034921
10	Arthritis	1.528024
11	Diabetes	1.353540
12	Hyperlipidemia	1.481125
13	BackPain	1.662300
14	Anxiety	1.449123
15	Allergic_rhinitis	1.605113
16	Reflux_esophagitis	1.659397

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	2.649e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:25	Log-Likelihood:	-16317.
No. Observations:	10000	AIC:	3.267e+04
Df Residuals:	9982	BIC:	3.280e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1278	0.058	-365.383	0.000	-21.241	-21.014
Income	-3.618e-07	4.35e-07	-0.833	0.405	-1.21e-06	4.9e-07
Gender_Male	0.0224	0.025	0.903	0.366	-0.026	0.071
Area_Suburban	0.0408	0.026	1.553	0.121	-0.011	0.092
TotalCharge	0.0122	5.74e-06	2120.421	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2331	0.030	-205.123	0.000	-6.293	-6.174
Initial_admin_Observation_Admission	0.0631	0.035	1.788	0.074	-0.006	0.132
Services_MRI	0.0717	0.065	1.106	0.269	-0.055	0.199
HighBlood	-1.3240	0.025	-52.518	0.000	-1.373	-1.275
Complication_risk	-2.7801	0.017	-163.472	0.000	-2.813	-2.747
Overweight	0.0477	0.027	1.748	0.080	-0.006	0.101
Arthritis	-0.8339	0.026	-32.235	0.000	-0.885	-0.783
Diabetes	-0.9235	0.028	-33.232	0.000	-0.978	-0.869
Hyperlipidemia	-1.1146	0.026	-42.522	0.000	-1.166	-1.063
BackPain	-1.0650	0.025	-42.256	0.000	-1.114	-1.016
Anxiety	-1.0328	0.027	-38.921	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7824	0.025	-30.851	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7366	0.025	-29.264	0.000	-0.786	-0.687

Omnibus:	37514.272	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1473.654
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.135	Cond. No.	2.63e+05

Removal of "Income" with p-value of 0.405:

	Variable	VIF
0	Gender_Male	1.814645
1	Area_Suburban	1.462810
2	TotalCharge	5.455755
3	Initial_admin_Emergency_Admission	2.701198
4	Initial_admin_Observation_Admission	1.756617
5	Services_MRI	1.038001
6	HighBlood	1.654839
7	Complication_risk	3.022006
8	Overweight	3.004337
9	Arthritis	1.525280
10	Diabetes	1.351914
11	Hyperlipidemia	1.476368
12	BackPain	1.655638
13	Anxiety	1.446125
14	Allergic_rhinitis	1.599954
15	Reflux_esophagitis	1.650145

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	2.815e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:25	Log-Likelihood:	-16318.
No. Observations:	10000	AIC:	3.267e+04
Df Residuals:	9983	BIC:	3.279e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1429	0.055	-385.084	0.000	-21.250	-21.035
Gender_Male	0.0223	0.025	0.901	0.368	-0.026	0.071
Area_Suburban	0.0407	0.026	1.550	0.121	-0.011	0.092
TotalCharge	0.0122	5.74e-06	2120.638	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2329	0.030	-205.128	0.000	-6.292	-6.173
Initial_admin_Observation_Admission	0.0627	0.035	1.779	0.075	-0.006	0.132
Services_MRI	0.0718	0.065	1.108	0.268	-0.055	0.199
HighBlood	-1.3240	0.025	-52.518	0.000	-1.373	-1.275
Complication_risk	-2.7801	0.017	-163.474	0.000	-2.813	-2.747
Overweight	0.0482	0.027	1.765	0.078	-0.005	0.102
Arthritis	-0.8338	0.026	-32.232	0.000	-0.885	-0.783
Diabetes	-0.9233	0.028	-33.226	0.000	-0.978	-0.869
Hyperlipidemia	-1.1148	0.026	-42.532	0.000	-1.166	-1.063
BackPain	-1.0652	0.025	-42.267	0.000	-1.115	-1.016
Anxiety	-1.0328	0.027	-38.922	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7823	0.025	-30.851	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7369	0.025	-29.282	0.000	-0.786	-0.688

Omnibus:	37507.467	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1474.052
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.135	Cond. No.	3.02e+04

Removal of "Gender_Male" with p-value of 0.368:

	Variable	VIF
0	Area_Suburban	1.461463
1	TotalCharge	5.355003
2	Initial_admin_Emergency_Admission	2.692106
3	Initial_admin_Observation_Admission	1.751054
4	Services_MRI	1.037999
5	HighBlood	1.651952
6	Complication_risk	3.002756
7	Overweight	2.981711
8	Arthritis	1.522387
9	Diabetes	1.350626
10	Hyperlipidemia	1.472573
11	BackPain	1.654638
12	Anxiety	1.445154
13	Allergic_rhinitis	1.597254
14	Reflux_esophagitis	1.648884

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	3.003e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:25	Log-Likelihood:	-16318.
No. Observations:	10000	AIC:	3.267e+04
Df Residuals:	9984	BIC:	3.278e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1316	0.053	-395.344	0.000	-21.236	-21.027
Area_Suburban	0.0406	0.026	1.544	0.123	-0.011	0.092
TotalCharge	0.0122	5.74e-06	2120.709	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2335	0.030	-205.210	0.000	-6.293	-6.174
Initial_admin_Observation_Admission	0.0620	0.035	1.757	0.079	-0.007	0.131
Services_MRI	0.0712	0.065	1.099	0.272	-0.056	0.198
HighBlood	-1.3239	0.025	-52.515	0.000	-1.373	-1.274
Complication_risk	-2.7801	0.017	-163.477	0.000	-2.813	-2.747
Overweight	0.0481	0.027	1.762	0.078	-0.005	0.102
Arthritis	-0.8336	0.026	-32.226	0.000	-0.884	-0.783
Diabetes	-0.9233	0.028	-33.229	0.000	-0.978	-0.869
Hyperlipidemia	-1.1144	0.026	-42.524	0.000	-1.166	-1.063
BackPain	-1.0656	0.025	-42.285	0.000	-1.115	-1.016
Anxiety	-1.0330	0.027	-38.931	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7824	0.025	-30.853	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7373	0.025	-29.300	0.000	-0.787	-0.688

Omnibus:	37498.974	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1474.546
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.134	Cond. No.	3.02e+04

Removal of "Services_MRI" with p-value of 0.272:

	Variable	VIF
0	Area_Suburban	1.461415
1	TotalCharge	5.348366
2	Initial_admin_Emergency_Admission	2.690960
3	Initial_admin_Observation_Admission	1.750875
4	HighBlood	1.651792
5	Complication_risk	3.002734
6	Overweight	2.980974
7	Arthritis	1.522376
8	Diabetes	1.349734
9	Hyperlipidemia	1.472542
10	BackPain	1.654541
11	Anxiety	1.445153
12	Allergic_rhinitis	1.597136
13	Reflux_esophagitis	1.648689

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	3.217e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:26	Log-Likelihood:	-16319.
No. Observations:	10000	AIC:	3.267e+04
Df Residuals:	9985	BIC:	3.278e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1288	0.053	-395.730	0.000	-21.233	-21.024
Area_Suburban	0.0405	0.026	1.542	0.123	-0.011	0.092
TotalCharge	0.0122	5.74e-06	2120.751	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2334	0.030	-205.205	0.000	-6.293	-6.174
Initial_admin_Observation_Admission	0.0617	0.035	1.750	0.080	-0.007	0.131
HighBlood	-1.3238	0.025	-52.512	0.000	-1.373	-1.274
Complication_risk	-2.7804	0.017	-163.505	0.000	-2.814	-2.747
Overweight	0.0480	0.027	1.759	0.079	-0.005	0.102
Arthritis	-0.8338	0.026	-32.231	0.000	-0.884	-0.783
Diabetes	-0.9227	0.028	-33.213	0.000	-0.977	-0.868
Hyperlipidemia	-1.1145	0.026	-42.527	0.000	-1.166	-1.063
BackPain	-1.0656	0.025	-42.286	0.000	-1.115	-1.016
Anxiety	-1.0332	0.027	-38.939	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7824	0.025	-30.853	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7372	0.025	-29.298	0.000	-0.787	-0.688

Omnibus:	37487.214	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1475.240
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.134	Cond. No.	2.76e+04

Removal of "Area_Suburban" with p-value of 0.123:

	Variable	VIF
0	TotalCharge	5.304598
1	Initial_admin_Emergency_Admission	2.680869
2	Initial_admin_Observation_Admission	1.745207
3	HighBlood	1.650095
4	Complication_risk	2.994282
5	Overweight	2.967114
6	Arthritis	1.521060
7	Diabetes	1.348274
8	Hyperlipidemia	1.471274
9	BackPain	1.651785
10	Anxiety	1.444630
11	Allergic_rhinitis	1.596320
12	Reflux_esophagitis	1.646751

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	3.464e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:26	Log-Likelihood:	-16320.
No. Observations:	10000	AIC:	3.267e+04
Df Residuals:	9986	BIC:	3.277e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.1157	0.053	-400.570	0.000	-21.219	-21.012
TotalCharge	0.0122	5.74e-06	2120.604	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2333	0.030	-205.188	0.000	-6.293	-6.174
Initial_admin_Observation_Admission	0.0617	0.035	1.749	0.080	-0.007	0.131
HighBlood	-1.3236	0.025	-52.501	0.000	-1.373	-1.274
Complication_risk	-2.7805	0.017	-163.501	0.000	-2.814	-2.747
Overweight	0.0482	0.027	1.765	0.078	-0.005	0.102
Arthritis	-0.8336	0.026	-32.223	0.000	-0.884	-0.783
Diabetes	-0.9223	0.028	-33.198	0.000	-0.977	-0.868
Hyperlipidemia	-1.1144	0.026	-42.519	0.000	-1.166	-1.063
BackPain	-1.0651	0.025	-42.267	0.000	-1.114	-1.016
Anxiety	-1.0334	0.027	-38.944	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7827	0.025	-30.863	0.000	-0.832	-0.733
Reflux_esophagitis	-0.7371	0.025	-29.290	0.000	-0.786	-0.688

Omnibus:	37463.582	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1476.648
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.133	Cond. No.	2.73e+04

Removal of "Initial_admin_Observation_Admission" with p-value of 0.080:

	Variable	VIF
0	TotalCharge	5.083797
1	Initial_admin_Emergency_Admission	1.970951
2	HighBlood	1.642528
3	Complication_risk	2.958324
4	Overweight	2.899371
5	Arthritis	1.515869
6	Diabetes	1.345049
7	Hyperlipidemia	1.465654
8	BackPain	1.641757
9	Anxiety	1.440249
10	Allergic_rhinitis	1.593832
11	Reflux_esophagitis	1.641170

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	3.752e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:26	Log-Likelihood:	-16321.
No. Observations:	10000	AIC:	3.267e+04
Df Residuals:	9987	BIC:	3.276e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.0842	0.050	-425.533	0.000	-21.181	-20.987
TotalCharge	0.0122	5.74e-06	2120.405	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2637	0.025	-251.335	0.000	-6.313	-6.215
HighBlood	-1.3233	0.025	-52.485	0.000	-1.373	-1.274
Complication_risk	-2.7811	0.017	-163.548	0.000	-2.814	-2.748
Overweight	0.0484	0.027	1.773	0.076	-0.005	0.102
Arthritis	-0.8336	0.026	-32.219	0.000	-0.884	-0.783
Diabetes	-0.9226	0.028	-33.204	0.000	-0.977	-0.868
Hyperlipidemia	-1.1143	0.026	-42.511	0.000	-1.166	-1.063
BackPain	-1.0646	0.025	-42.244	0.000	-1.114	-1.015
Anxiety	-1.0333	0.027	-38.938	0.000	-1.085	-0.981
Allergic_rhinitis	-0.7839	0.025	-30.917	0.000	-0.834	-0.734
Reflux_esophagitis	-0.7376	0.025	-29.308	0.000	-0.787	-0.688

Omnibus:	37431.325	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1478.495
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.132	Cond. No.	2.47e+04

Final reduced regression model:

	Variable	VIF
0	TotalCharge	4.748255
1	Initial_admin_Emergency_Admission	1.959246
2	HighBlood	1.623714
3	Complication_risk	2.879580
4	Arthritis	1.506364
5	Diabetes	1.340189
6	Hyperlipidemia	1.458477
7	BackPain	1.626610
8	Anxiety	1.435697
9	Allergic_rhinitis	1.582119
10	Reflux_esophagitis	1.632262

OLS Regression Results

Dep. Variable:	Initial_days	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	4.092e+05
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	17:53:26	Log-Likelihood:	-16323.
No. Observations:	10000	AIC:	3.267e+04
Df Residuals:	9988	BIC:	3.276e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.0492	0.045	-463.102	0.000	-21.138	-20.960
TotalCharge	0.0122	5.74e-06	2120.308	0.000	0.012	0.012
Initial_admin_Emergency_Admission	-6.2640	0.025	-251.332	0.000	-6.313	-6.215
HighBlood	-1.3221	0.025	-52.450	0.000	-1.372	-1.273
Complication_risk	-2.7811	0.017	-163.533	0.000	-2.814	-2.748
Arthritis	-0.8334	0.026	-32.208	0.000	-0.884	-0.783
Diabetes	-0.9229	0.028	-33.214	0.000	-0.977	-0.868
Hyperlipidemia	-1.1145	0.026	-42.516	0.000	-1.166	-1.063
BackPain	-1.0641	0.025	-42.223	0.000	-1.113	-1.015
Anxiety	-1.0339	0.027	-38.956	0.000	-1.086	-0.982
Allergic_rhinitis	-0.7837	0.025	-30.909	0.000	-0.833	-0.734
Reflux_esophagitis	-0.7381	0.025	-29.329	0.000	-0.787	-0.689

Omnibus:	37398.001	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1480.410
Skew:	0.121	Prob(JB):	0.00
Kurtosis:	1.131	Cond. No.	2.26e+04

Section E1) Model Comparison

The initial regression model began with 31 predictor variables. However, through the backward stepwise feature elimination process, the final model was reduced to 11 predictors: “TotalCharge,” “Initial_admin_Emergency_Admission,” “HighBlood,” “Complication_risk,” “Arthritis,” “Diabetes,” “Hyperlipidemia,” “BackPain,” “Anxiety,” “Allergic_rhinitis,” and “Reflux_esophagitis.”

Adjusted R^2 is a metric to evaluate how well a regression model fits the data (IBM, 2024). Unlike regular R^2, it accounts for the number of predictors by penalizing the inclusion of irrelevant variables that do not improve the model’s performance (Ouko, 2024). In the initial model, the Adjusted R^2 value was 0.997783, indicating that 99.7783% of the variability in the dependent variable, “Initial_days,” was explained by the predictor variables. In the final model, the Adjusted R^2 value increased slightly to 0.997784, meaning 99.7784% of the variability was explained by the reduced set of predictors. Although the increase is slight (0.0001%), it reflects the removal of irrelevant features, resulting in a model that explains the variability in the dependent variable more effectively per predictor.

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are additional metrics used to estimate prediction error and the quality of a regression model in relation to each other while penalizing for overfitting (Bayesian Information Criterion, n.d., Akaike Information Criterion, n.d.). The initial model’s AIC was 32694.806, compared to the final model’s AIC of 32669.769. For BIC, the initial value was 32925.537, while the final model achieved a lower value of 32756.293. Since lower AIC and BIC values indicate better model performance, the final model is preferred over the initial one.

The Mean Squared Error (MSE) is another evaluation metric that measures the average squared difference between predicted and actual values (Padhma, 2024), indicating how well a model makes predictions (Day, 2023). For the final model, the MSE is 1.532193. A lower MSE reflects a model that makes predictions closer to the actual values, showing that the reduced model better fits the data.

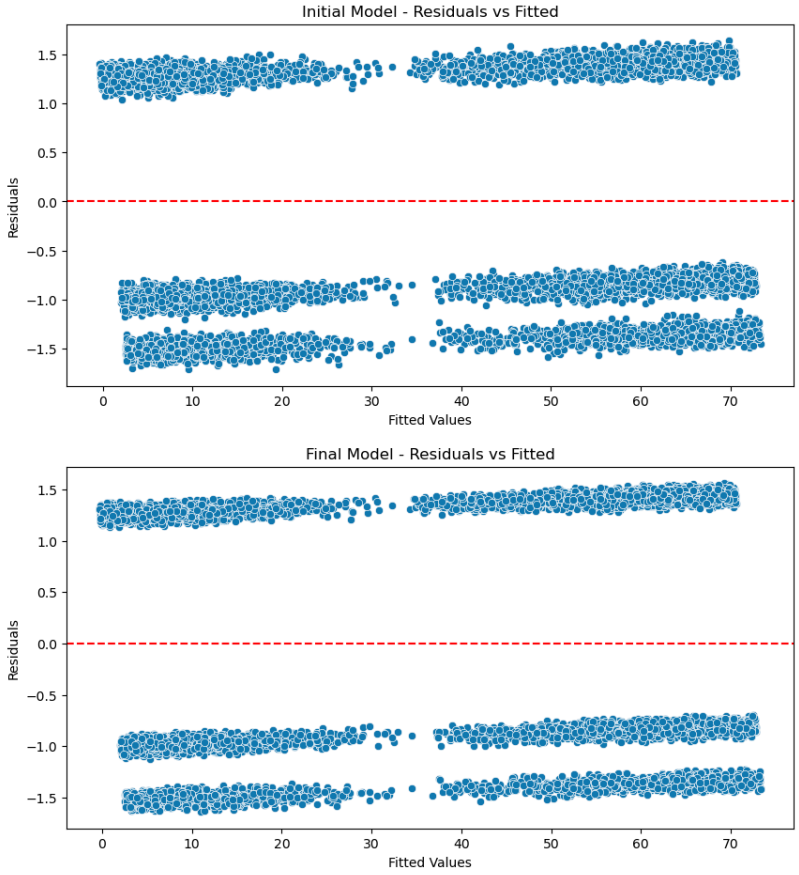
Here is annotated code comparing the metrics between the initial regression model and the reduced model:

```
# Compare initial vs final model

# Metrics for Initial Model
initial_adj_r2 = initial_model.rsquared_adj
initial_aic = initial_model.aic
initial_bic = initial_model.bic
initial_mse = np.mean(initial_model.resid**2)

# Metrics for Final Model
final_adj_r2 = final_model.rsquared_adj
final_aic = final_model.aic
final_bic = final_model.bic
final_mse = np.mean(final_model.resid**2)

# Create a comparison DataFrame
metrics_comparison = pd.DataFrame({
        “Metric”: [”Adjusted R²”, “AIC”, “BIC”, “MSE”],
        “Initial Model”: [initial_adj_r2, initial_aic, initial_bic, initial_mse],
        “Final Model”: [final_adj_r2, final_aic, final_bic, final_mse]
})

print(metrics_comparison)
```

```
         Metric  Initial Model   Final Model
0  Adjusted R²       0.997783      0.997784
1          AIC   32694.805926  32669.769304
2          BIC   32925.536818  32756.293389
3          MSE       1.529902      1.532193
```

The residual scatter plots compare the initial and final models by showing the differences between actual and predicted values. In the initial model, the residuals are scattered around the zero line. However, there is visible clustering at specific fitted value ranges, particularly near lower fitted values (around 0–20) and higher fitted values (around 50–70). The clustering suggests that the model could have a potential issue with non-linearity.

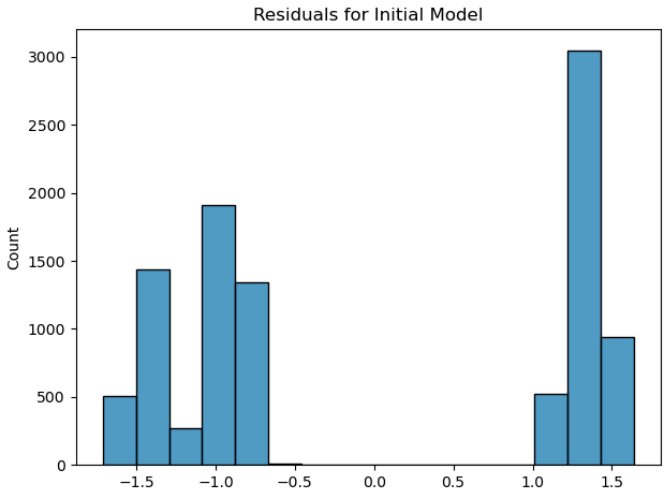
On the other hand, the residuals in the final model have less clustering or visible patterns. This scatter plot shows that the final model better fits the data and effectively captures the variables' relationships.

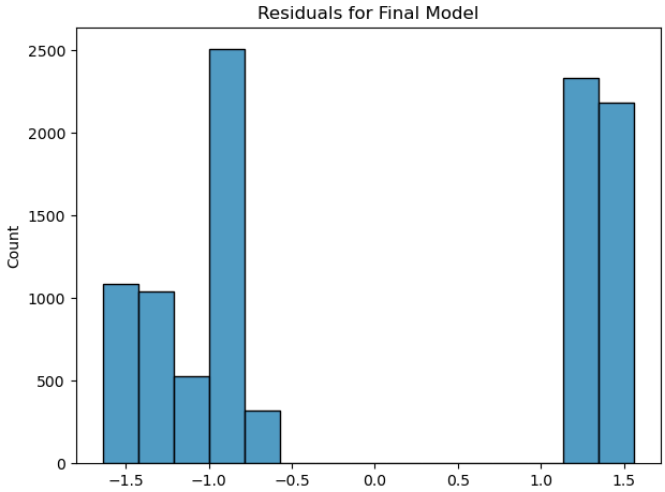


The residual histograms for the initial and final models show the distribution of residuals (differences between actual and predicted values) and whether they meet the assumptions of normality.

The histogram shows a bimodal distribution in the initial model with two distinct peaks around -1 and between 1.0 and 1.5. This indicates that the residuals are not normally distributed, suggesting the model does not fully explain the variation in the data.

In the final model, the histogram is still bimodal, but the peaks are slightly less pronounced. While the residuals are still not perfectly normal, this represents a slight improvement over the initial model.

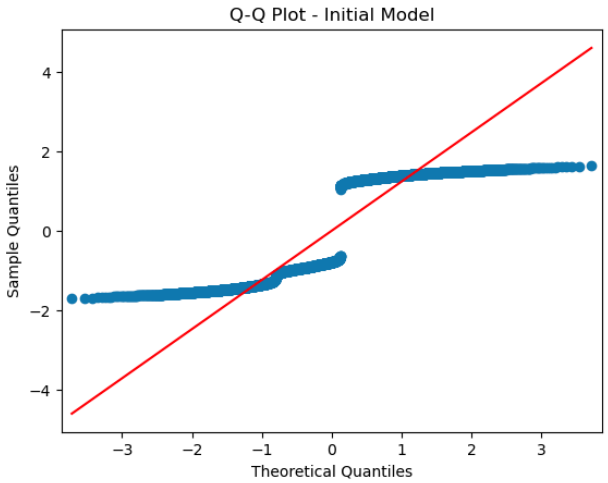


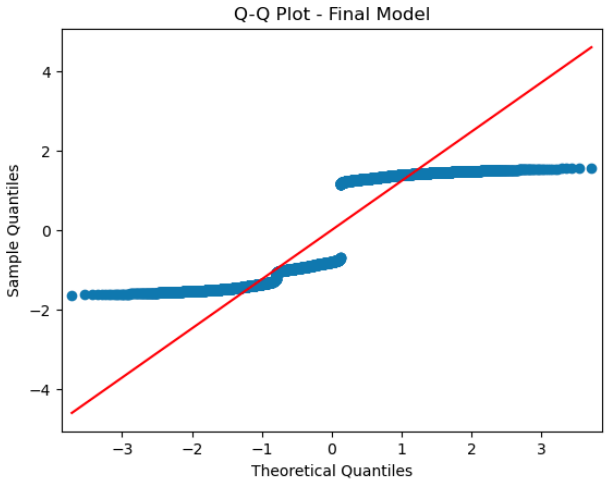


The Q-Q plots for the initial and final models help assess whether the residuals follow a normal distribution by comparing sample quantiles (actual residuals) to theoretical quantiles (expected residuals under normality).

In the initial model, the Q-Q plot shows significant deviations from the red diagonal line, especially at the tails. The residuals have two horizontal clusters, indicating that they are not normally distributed. This indicates that the initial model does not meet the assumption of normality.

In the final model, the Q-Q plot still shows deviations from the diagonal line and horizontal clustering, but the alignment appears slightly improved compared to the initial model. While the residuals are still not perfectly normal, the final model represents a slight improvement in terms of meeting the normality assumption.





Section E2) Output and Calculations

The reduced regression model evaluates the factors influencing the length of a patient's initial hospital stay ("Initial_days") using metrics like Mean Squared Error (MSE) and Residual Standard Error (RSE) to assess its accuracy.

The Mean Squared Error (MSE) is 1.53219272652417, indicating the average squared difference between predicted and actual hospital stays. The Residual Standard Error (RSE), the square root of MSE, is 1.237817727504405, meaning the model's predictions deviate from actual values by about 1.24 days on average. These metrics show the model is effective in explaining hospital stay durations while leaving room for small errors.

```
# Find MSE & Residual Standard Error for Reduced Model:
from sklearn.metrics import mean_squared_error

X = df[[”TotalCharge”, “Initial_admin_Emergency_Admission”, “HighBlood”,
“Complication_risk”,  “Arthritis”,
        “Diabetes”, “Hyperlipidemia”, “BackPain”, “Anxiety”, “Allergic_rhinitis”,
“Reflux_esophagitis”]]
y = df['Initial_days']

# Fit reduced model
model = sm.OLS(y, sm.add_constant(X[[”TotalCharge”,
“Initial_admin_Emergency_Admission”, “HighBlood”, “Complication_risk”,  “Arthritis”,
        “Diabetes”, “Hyperlipidemia”, “BackPain”, “Anxiety”, “Allergic_rhinitis”,
“Reflux_esophagitis”]])).fit()

# Get predicted values and actual values
y_pred = model.predict(sm.add_constant(X[[”TotalCharge”,
“Initial_admin_Emergency_Admission”, “HighBlood”, “Complication_risk”,  “Arthritis”,
        “Diabetes”, “Hyperlipidemia”, “BackPain”, “Anxiety”, “Allergic_rhinitis”,
“Reflux_esophagitis”]]))
y_true = y

# Calculate MSE
mse = mean_squared_error(y_true, y_pred)

print(”Mean Squared Error: “, mse)
```

```
# Calculate RSE
rse = np.sqrt(mse)
print(”Residual Squared Error: “, rse)
```

```
Mean Squared Error:  1.53219272652417
Residual Squared Error:  1.237817727504405
```

Cross-validation is a method used to test how well a model performs on unseen data (Jain, 2024). It works by splitting the dataset into several parts, called “folds.” The model is trained on some folds and tested on the remaining fold. This process is repeated multiple times so that each fold is used as a test set once.

```
# Perform Cross validation
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score

# Initialize the model
linear_model = LinearRegression()

# Define variables
X = df[[”TotalCharge”, “Initial_admin_Emergency_Admission”, “HighBlood”,
“Complication_risk”,  “Arthritis”,
        “Diabetes”, “Hyperlipidemia”, “BackPain”, “Anxiety”, “Allergic_rhinitis”,
“Reflux_esophagitis”]]
y = df['Initial_days']

# Perform cross-validation
cv_scores = cross_val_score(linear_model, X, y, cv=5)

# Print results
print(”Cross-validation scores:”, cv_scores)
print(”Mean score:”, cv_scores.mean())
```

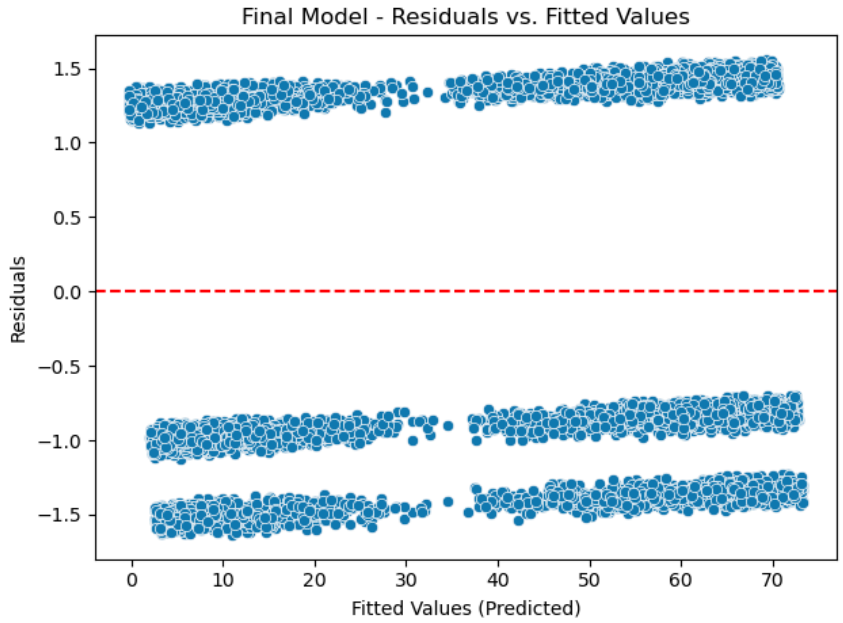
```
Cross-validation scores: [0.9583502  0.95803202 0.99775671 0.97905238 0.97933148]
Mean score: 0.9745045585150376
```

The cross-validation scores represent how well the model predicted during each fold. The scores are close to 1, indicating the model performs well across the folds. The mean score, 0.9745045585150376, reflects the model's overall accuracy, showing it generalizes well to new

data. Cross-validation helps ensure the model isn't overfitted to the training data and works consistently across different subsets of data.

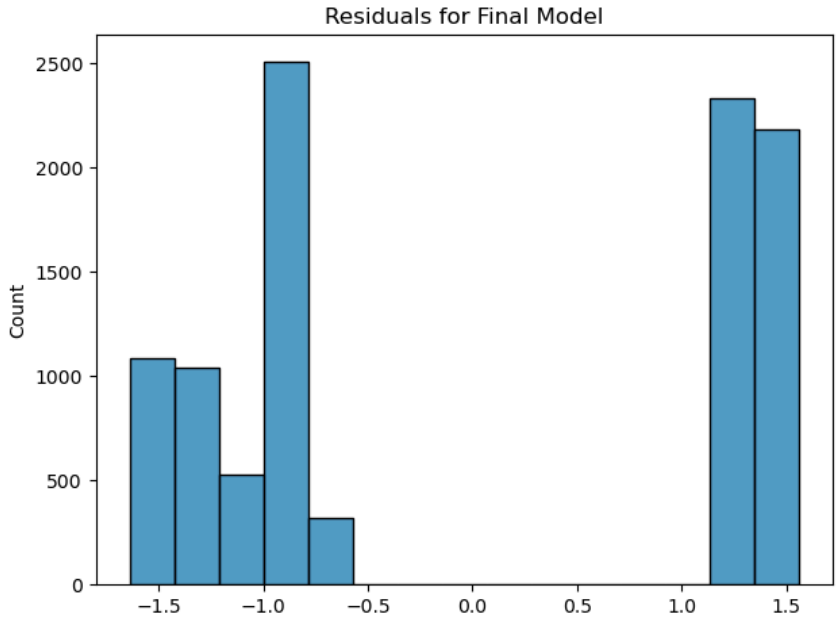
A residual plot shows the differences between the actual and predicted values (How to Interpret a Residual Plot | Algebra, 2021) It helps assess how well a regression model fits the data and checks key assumptions, such as linearity and homoscedasticity (StatsNotebook, 2020). In this scatterplot, the residuals are evenly distributed around the horizontal line at zero, indicating that the model satisfies the assumption of linearity. Additionally, the spread of residuals is consistent, suggesting equal variance.

```
# Scatter plot of residuals vs fitted values for final model

sns.scatterplot(x=final_model.fittedvalues, y=final_model.resid)
plt.axhline(y=0, color="red", linestyle="--")  # Add a horizontal line at 0
plt.title("Final Model - Residuals vs. Fitted Values")
plt.xlabel("Fitted Values (Predicted)")
plt.ylabel("Residuals")
plt.tight_layout()
plt.show()
```



A residual histogram shows the distribution of residuals (the differences between actual and predicted values) to check if they follow a normal distribution, which is an important assumption in linear regression. In this histogram, the residuals form two distinct clusters around -1 and 1 instead of a smooth bell curve centered around zero. This suggests that the residuals are not normally distributed, which could mean that one of the model's assumptions could be violated (*Help Online - Origin Help - Residual Plot Analysis*, n.d.).

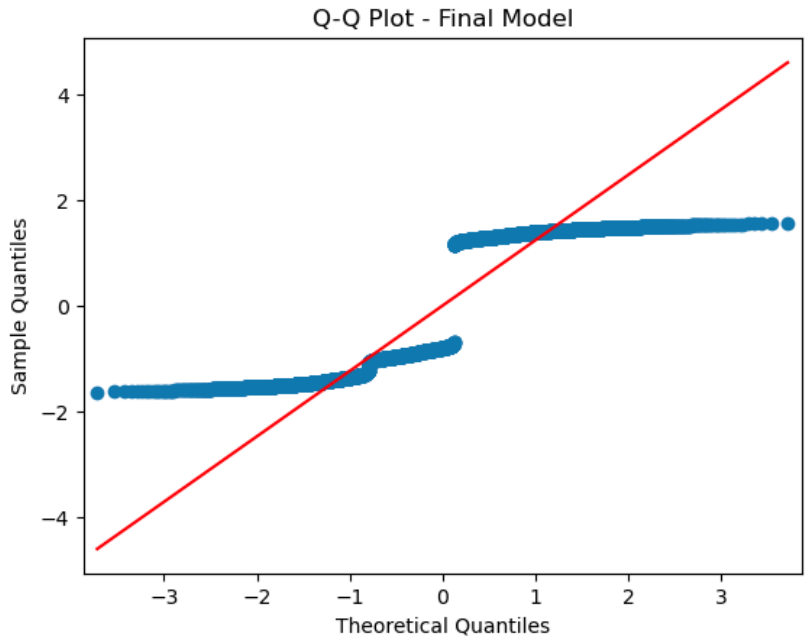
```
# Histogram of residuals for final model

sns.histplot(final_model.resid)
plt.title("Residuals for Final Model")
plt.tight_layout()
plt.show()
```



A Q-Q (Quantile-Quantile) plot is used to assess whether the residuals of a regression model follow a normal distribution. It compares the observed residuals (sample quantiles) to a theoretical normal distribution (theoretical quantiles). If the residuals are normally distributed, the points in the plot will closely follow the red diagonal line.

In this Q-Q plot, the residuals deviate significantly from the diagonal line, forming two distinct horizontal lines. This indicates that the residuals do not follow a normal distribution. The clear separation between clusters suggests potential issues with the model, such as unaccounted categorical variables, nonlinearity, or other violations of regression assumptions.

```
# Q-Q Plot for Final Model
sm.qqplot(final_model.resid, line='s')
plt.title("Q-Q Plot - Final Model")
plt.show()
```



Section E3) Code

Error-free code is attached, titled "d208-pa-task-1-revised.ipynb"

Part V: Data Summary and Implications

Section F1) Results

The regression equation represents the relationship between the dependent variable, "Initial_days" and the predictors in the reduced model (Regression Equation: What It Is and How to Use It, n.d.). The equation consists of two main components: the intercept and the coefficients of the predictors. The intercept represents the baseline value of "Initial_days" when all predictors are zero.

Each coefficient represents the strength and direction of a predictor's relationship with "Initial_days" (Goyal, 2024) as well as the change in the dependent variable for a one-unit change in the predictor variable while keeping all the others constant. This allows the impact of each predictor to be isolated from the influence of the other independent variables in the model. For example, the coefficient for "TotalCharge" (0.0122) means that for every additional dollar in charges, the length of the hospital stay increases by 0.0122 days, assuming all other predictors like "HighBlood" and "Stroke" remain unchanged. A positive coefficient indicates that as the predictor increases, the length of stay increases, while a negative coefficient suggests shorter hospital stays as the predictor increases. (Frost, n.d.)

The formula is constructed by combining the intercept and coefficients, using the final_model.params() function:

```
# Equation for final model

# Extract the coefficients
coefficients = final_model.params

# Extract the intercept and coefficients
intercept = coefficients['const']
predictors = coefficients.index[1:]  # Exclude the intercept ('const')
coeff_values = coefficients.values[1:]  # Exclude the intercept value

# Create the formula
formula = f"Initial_days = {intercept}"
for predictor, coeff in zip(predictors, coeff_values):
        formula += f" + ({coeff:.4f}) * {predictor}"

print("Final Model Formula:")
print(formula)
```

The above code results in the following equation:
Initial_days = -21.04919073322405 + (0.0122) * TotalCharge + (-6.2640) * Initial_admin_Emergency_Admission + (-1.3221) * HighBlood + (-2.7811) * Complication_risk + (-0.8334) * Arthritis + (-0.9229) * Diabetes + (-1.1145) * Hyperlipidemia + (-1.0641) * BackPain + (-1.0339) * Anxiety + (-0.7837) * Allergic_rhinitis + (-0.7381) * Reflux_esophagitis

The intercept, −21.0492, represents the baseline value of "Initial_days" when all predictors are zero. For the variable "TotalCharge," the coefficient 0.0122 indicates that for every additional unit of charge (e.g., one dollar), the hospital stay increases by 0.0122 days. This suggests higher charges are linked to slightly longer stays, likely reflecting greater resource use or more intensive care. On the other hand, patients admitted through emergency services ("Initial_admin_Emergency_Admission") have shorter stays by an average of 6.2640 days.

Most coefficients for patients with specific conditions are negative, indicating shorter hospital stays. For instance, patients with high blood pressure have stays that are 1.3221 days shorter. Similarly, complication risks reduce hospital stays by 2.7811 days. Chronic conditions like arthritis (coefficient: −0.8334), diabetes (coefficient: −0.9229), and hyperlipidemia (coefficient: −1.1145) are also associated with shorter stays, suggesting that these conditions do not require prolonged hospitalization.

Similarly, other conditions such as back pain (−1.0641), anxiety (−1.0339), allergic rhinitis (−0.7837), and reflux esophagitis (−0.7381) similarly result in shorter hospital stays. This suggests that these medical conditions may not need extended inpatient care.

Overall, the model shows that higher charges are associated with more extended hospital stays, and medical conditions are linked to shorter durations. This can help inform resource allocation strategies and improve patient care delivery.

The model has a high Adjusted R^2 value of 0.998, explaining 99.8% of the variation in "Initial_days." This shows that the model does an excellent job of capturing nearly all the factors that influence hospital stays. Additionally, all the predictors in the model have p-values below 0.05, showing that each variable has a meaningful relationship with the length of the hospital stay. The F-statistic (4.092e+05) and its p-values of 0.000 shows that the model is highly significant overall in explaining the variation in hospital stays.

Furthermore, the Variance Inflation Factor (VIF) values for all predictors are below 5, which shows no issue with multicollinearity. This combination of high Adjusted R^2 value, significant predictors, and low multicollinearity demonstrates that the reduced model is statistically significant.

The analysis of the data has some limitations. The assumption of normality in residuals is not met, as shown by the Q-Q plot and residual histogram. Additionally, clustered residuals suggest potential violation of assumptions.

Section F2) Recommendations

Healthcare facilities can benefit from the findings of this data analysis to enhance their strategies, streamline operations, and improve patient outcomes. Higher total charges correlate with longer hospital stays, emphasizing the importance of directing resources effectively. Designing tailored care plans for patients with longer hospital stays can help control costs while providing quality patient treatment.

Emergency admissions can still benefit from well-organized discharge processes and adequate follow-up care plans. These approaches can mitigate the risk of readmissions.

Based on the connection between medical conditions and shorter hospital stays, creating initiatives focused on chronic condition management and post-discharge care can help reduce hospital admissions.

In addition, managing patients with varying levels of complication risk is crucial. Hospitals can create specialized teams to monitor high-risk patients and adopt early intervention strategies to prevent complications from turning into prolonged hospital stays.

Part VI: Demonstration

Section G) Panopto Demonstration

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=550bbb07-aeb4-46f2-98d7-b25a0039da0a

Section H) Sources of Third-Party Code

GeeksforGeeks (n.d.). *How to perform ordinal encoding using sklearn?* GeeksforGeeks. Retrieved January 2, 2025, from https://www.statology.org/multiple-linear-regression-assumptions/

Section I) Sources

Akaike information criterion. (n.d.). Wikipedia. Retrieved January 2, 2025, from https://en.wikipedia.org/wiki/Akaike_information_criterion

Bayesian information criterion. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/Bayesian_information_criterion

Bobbit, Z. (2021, November 16). *The Five Assumptions of Multiple Linear Regression*. Statology. https://www.statology.org/multiple-linear-regression-assumptions/

Day, U. (2023, September 21). *Understanding Mean Squared Error (MSE) in Regression Models*. Medium. https://medium.com/@wl8380/understanding-mean-squared-error-mse-in-regression-models-9ade100c9627

Frost, J. (n.d.). *How to Interpret P-values and Coefficients in Regression Analysis*. Statistics By Jim. Retrieved January 2, 2025, from https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

Frost, J. (n.d.). *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions*. Statistics By Jim. Retrieved January 2, 2025, from https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Goyal, C. (2024, December 24). *Regression Coefficients: Definition, Formula and Examples*. Analytics Vidhya. Retrieved January 2, 2025, from https://www.analyticsvidhya.com/blog/2021/03/standardized-vs-unstandardized-regression-coefficient/

Hayes, A. (2022, January 10). *Stepwise Regression: Definition, Uses, Example, and Limitations*. Investopedia. Retrieved January 2, 2025, from https://www.investopedia.com/terms/s/stepwise-regression.asp.

Help Online - Origin Help - Residual Plot Analysis. (n.d.). OriginLab. Retrieved January 2, 2025, from https://www.originlab.com/doc/origin-help/residual-plot-analysis

How to Interpret a Residual Plot | Algebra. (2021, April 23). Study.com. Retrieved January 2, 2025, from https://study.com/skill/learn/how-to-interpret-a-residual-plot-explanation.html

IBM. (2024, January 18). *Adjusted R squared*. IBM. https://www.ibm.com/docs/en/cognos-analytics/12.0.0?topic=terms-adjusted-r-squared

IBM. (2024, June 19). *What is Data Transformation?* IBM. Retrieved January 1, 2025, from https://www.ibm.com/think/topics/data-transformation

Jain, S. (2024, December 30). *Cross Validation in Machine Learning*. GeeksforGeeks. Retrieved January 2, 2025, from https://www.geeksforgeeks.org/cross-validation-machine-learning/

JMP. (n.d.). *Multiple Linear Regression | Introduction to Statistics*. JMP. Retrieved January 2, 2025, from https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-multiple-regression.html

Lu, T. (2024, September 2). *What is Data Wrangling? A Practical Guide With Examples*. https://www.datacamp.com/blog/what-is-data-wrangling

Medical Data Considerations and Dictionary. (n.d.). Western Governors University.

Ouko, A. (2024, September 22). *Adjusted R-Squared: A Clear Explanation with Examples*. DataCamp. Retrieved January 2, 2025, from https://www.datacamp.com/tutorial/adjusted-r-squared

Padhma. (2024, November 21). *A Comprehensive Introduction to Evaluating Regression Models*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/

Regression Equation: What it is and How to use it. (n.d.). Statistics How To. Retrieved January 2, 2025, from https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-is-a-regression-equation/

Singh, V. (2024, November 18). *Variance Inflation Factor: How to Detect Multicollinearity*. DataCamp. Retrieved January 2, 2025, from https://www.datacamp.com/tutorial/variance-inflation-factor

Soetewey, A. (2021, October 4). *Multiple linear regression made simple*. Stats and R. Retrieved January 2, 2025, from https://statsandr.com/blog/multiple-linear-regression-made-simple/

StatsNotebook. (2020, October 16). *Residual Plots and Assumption Checking*. StatsNotebook. Retrieved January 2, 2025, from https://statsnotebook.io/blog/analysis/linearity_homoscedasticity/