Eric Yarger

Predictive Modeling using Multiple Regression

A1: Research Question

Multiple regression is a versatile tool in statistical analysis. Put simply, multiple regression is when linear regression is performed with two or more independent/predictor variables. This is in contrast to simple regression, which is linear regression with one predictor variable. Multiple regression can be used to shine light on a wide variety of organizational questions.

In our particular case, we will be using multiple regression to address the question: **What variables best predict the amount of Additional Charges spent by patients in the hospital?** Additional Charges incurred by the patient are represented by the data set feature Additional_charges. Having a model in our organization that can help predict additional charges is relevant, potentially profitable for the organization, and most likely beneficial for our patients. The ability to assess what causes additional charges will help our organization respond more effectively in serving our patients physical and financial health.

A2: Objectives and Goals

The goal for analysis is to find what features we can use to best predict additional charges accrued by our customers. This will be addressed by cleaning, preparing, and analyzing data from the medical_clean dataset provided for this course. A broad net will be cast to include all independent features and their respective dummy variables. This ensures that bias and assumption into potential variable correlation is negated. Initial model features will be statistically identified and analyzed. Model features will be reduced, if viable, and regression analysis will be performed again, resulting in our answer to the research question.

B1: Summary of Assumptions

With any statistical analysis there will be a number of assumptions regarding the data and variables. From WGU's MSDA professor Dr. Sewell's D208 Predictive Modeling Episode 4 presentation (Sewell, n.d.), and from Statology (Zach, 2021), assumptions of a multiple regression model include:

1. Linear relationship: a linear relationship exists between the response variable and each predictor variable.

2. No Multicollinearity: The predictor variables do not correlate highly with each other.

3. Independence: All observations are independent of each other.

4. Homoscedasticity: The residuals have constant variance across the linear model.

5. Multivariate Normality: The models' residuals are distributed normally.

B2: Tool Benefits

JupyterLabs IDE using Python3 as the language was used to perform this analysis. The Python environment is rich in libraries and functions that allow analysts to clean, prepare, and analyze large datasets. Python has a wide variety of open-source tools that analysts can import into their codebase to perform multiple regression analysis. JupyterNotebooks are beneficial for organizing the necessary code to analyze and visualize the dataset. This is a huge benefit over less structured programming environments where altering and organizing code can be a hassle.

The tools used to perform this analysis are both built into Python as well as imported from various open-source libraries. From preparing, cleaning, to analyzing these tools are beneficial across the data analysis lifecycle. These libraries include:

- Matplotlib - free and comprehensive library for creating visualizations in python. Example - Useful for visualizing histograms and the shape of variable data during cleaning.
- Pandas - powerful open source data analysis tool. Example - useful for creating two-dimensional data frames during the preparation stage.
- Seaborn - visualization tool for creating beautiful statistical graphics. Example - useful for visualizing data in the Cleaning and Analysis stages.
- Numpy - Allows Python to process larger arrays than it could normally handle. Example - used in scientific computing using arrays during the analysis stage.
- Missingno - used to visualize missing data in data sets. Example - used to visualize missing data in the cleaning and preparation stages.
- Scipy - Open source Python library for statistical analysis and scientific computing. Example - statistical analysis functions used for multiple regression analysis during the analysis stage.
    - Scipy.stats.zscore - Used for calculating z-scores for independent feature analysis and normalization.

- Sklearn - ML library for Python for a wide variety of analytical processes. Specific tools used in Sklearn library include:
    - preprocessing.StandardScaler - standardizes features by removing the mean and scaling to unit variance.
    - Pipeline.make_pipeline - constructs pipeline from given estimators
    - model_selection.KFold - provides test/ train indices to split data into sets. Splits into specified # of folds, which are then used for analysis validation.
    - neighbors.KNeighborsRegressor - Regression based on k-nearest neighbors. The target is predicted by interpolation of targets associated with the nearest neighbors in the training set.
    - Metrics.mean_squared_error, r2_score - calculation of Root Mean Square Error and R_squared.
    - Model_selection.cross_val_predict - Generates cross-validated estimates for each input data point.
    - linear_model.LinearRegression - Implementation of sklearn's linear regression module.
- Statsmodels.api - Used to initiate OLS for regression analysis
- Statsmodels.stats.outliers_influence.variance_inflation_factor - Used for generating VIF table for feature multicollinearity analysis.
- Patsy.dmatrix - used to construct single design matrices given a specific formula and data.
- Math.sqrt - import square root function.

In contrast, there are other tools and environments that can be used for this type of analysis, such as R or Excel. Python offers the benefit of being widely open source, is a popular programming language used by many analysts, and can be run in many different IDE's. This benefits other analysts in our organization in the future to use the same processes and tools in their preferred environments for analysis. Python is also beneficial for wrangling all of these tools into one place. You just need to import the necessary libraries and tools to use them.

B3: Appropriate Technique

Multiple regression is appropriate for this analysis for a number of reasons. The dependent variable in our analysis, Additional_charges, is a continuous variable. Also, there are multiple independent variables used for this prediction, which rules out simple regression in favor of multiple regression. The independent variables used in this analysis are both categorical and continuous, which multiple regression handles well.

There are continuous dependent variables with normal distributions that will be analyzed in the initial multiple regression model. After performing a VIF analysis the dependent variables also show no multicollinearity, which is optimal for a multiple regression analysis. The assumptions mentioned in section B1 of this document at the onset of this analysis appear to be satisfied. This points to multiple regression being a viable and appropriate technique for this analysis.

C1: Data Goals

According to software developer educational site EDUCBA, The goal of data preparation and manipulation efforts is to make interpreting and analyzing the insights from the data more structured and better designed (*Data Manipulation with Python, N.d.)*. Step-by-step code with all data visualizations is provided in the D208_Task_1 PDF of my Jupyter Notebook. The preparation and manipulations were undertaken to align the data with more effective multiple regression analysis. These actions included:

- Examining feature variance
- Renaming applicable features better insight and utility
- Plotting univariate boxplots and histograms to analyze the shape and distribution of features
- Splitting the data set into numerical and categorical feature sets. This allowed me to run Z-score analysis on all numerical data and remove cases with z-scores of more than 3. This removes outlier cases from the dataset, which ensures utility of features for Multiple Regression.
- Plotting bivariate scatterplots for analysis of independent features in their relation to the dependent feature.

Each of these preparation and manipulations of the data set provide a more cohesive, structured, and viable data set for multiple regression. This allows us to ensure that we will return the best possible results from our multiple regression analysis.

C2: Summary Statistics

Summary statistics help us to identify the shape, distribution, correlation, meaning, and size of the data. For instance, data set size provides us with knowledge that our data set will be large enough to provide meaningful and valid insight from analysis. With 9206 observations for each variable, we know our set is large enough for regression analysis. Knowing if a variable is categorical or continuous helps us decide how to clean, prepare, and utilize each feature. We can see that our Target Variable, Additional_charges is continuous, which is required for linear regression. Our predictor variables are made up of 2 continuous and 10 categorical variables.

Mean, Median, and Mode help to visualize the feature's distribution and size. This helps us see the distribution and average observation size for Additional_charges in comparison to all predictor variables. Knowing features percentiles, minimum, and maximum quantities helps to identify shape, distribution, and outliers.

Being able to see correlation between our target variable and each predictor variable helps to identify which features will be viable for analysis. From the correlation table below we can see that predictor variables Age and HighBlood_Yes are by far show the most correlation with our target variable. Also, being able to see correlation between the predictor variables helps to identify independence and potential multicollinearity in conjunction with VIF, which we will calculate later in this assessment.

The table below lists the target variable, Additional_charges, and all predictor variables gathered from the data set to answer the research question and provides the basic details for each. Variable descriptions were provided by the PDF accompanying the course provided data set download, and updated to reflect dummy variable selection and interpretation. The total number of variables used for answering the research question is 13: 1 Target Variable and 12 Predictor Variables. Each variable has 9206 observations.

Variable	**Type**	**Continuous or Categorical**	**# of Observations**	**Example**	**Variable Description**
Additional_charges	float64	Continuous	9206	17939.40342	**Target Variable** for the model. Is the amount charged to the patient for miscellaneous procedures, medicines, treatments, etc.

Age	int64	Continuous	9206	53	Predictor Variable. Age of patient reported in admissions information.
TotalCharge	float64	Continuous	9206	4193.190458	Predictor Variable. The amount charged to patient daily. Does not include charges for specialized treatment.
Marital_Married	uint8	Categorical	9206	Either 1 for selected or 0 for not selected	Predictor Variable. Dummy variable. Represents Marital status of patient who is married if data is 1.
Doc_visits_7	uint8	Categorical	9206	Either 1 for selected or 0 for not selected	Predictor Variable. Dummy variable. Patients whose primary physician visited them 7 times during initial hospitalization if data is 1.
Initial_admin_Emergency_Admission	uint8	Categorical	9206	Either 1 for selected or 0 for not selected	Predictor Variable. Dummy variable. The means by which the patient was admitted to the hospital initially. Represents an emergency admission if data is 1.
Initial_admin_Observations_Admission	uint8	Categorical	9206	Either 1 for selected or 0 for not selected	Predictor Variable. Dummy variable. The means by which the patient was admitted to the hospital initially. Represents an observation admission if data is 1.
HighBlood_Yes	uint8	Categorical	9206	Either 1 for selected or 0 for not selected	Predictor Variable. Dummy variable. Positive state that the patient has high blood pressure if data is 1.
Stroke_Yes	uint8	Categorical	9206	Either 1 for selected or 0 for not selected	Predictor Variable. Dummy variable. Positive state that patient has had a stroke if data is 1.
Complication_risk_Low	uint8	Categorical	9206	Either 1 for selected or 0 for not selected	Predictor Variable. Dummy variable. Level of complication risk for patient as assessed by patient assessment. Positive state that patient complication risk is low if data is 1.
Options_3	uint8	Categorical	9206	Either 1 for selected or 0 for not selected	Predictor Variable. Dummy variable. Positive state that patient answered '3' on survey question regarding Options if data is 1.
Options_4	uint8	Categorical	9206	Either 1 for selected or 0 for not selected	Predictor Variable. Dummy variable. Positive state that patient answered '4' on survey question regarding Options if data is 1.

Courteous_2	uint8	Categorical	9206	Either 1 for selected or 0 for not selected	Predictor Variable. Dummy variable. Positive state that patient answered '2' on survey question regarding Courteous staff if data is 1.

Measures of central tendency are single values that represent the center point of the dataset. There are three common measures of central tendency: the mean, median, and mode (Zach, 2018). Measures of central tendency are detailed for the target variable and all predictor variables in the table below. Variables Additional_charges and TotalCharge had multiple modes, as described below.

Variable	**Mean**	**Median**	**Mode, Frequency of Mode/Observations**
Additional_charges (The target variable)	12927.980718	11556.7750	22000.06, freq =3 16617.19, freq =3 17874.97, freq =3 12349.45, freq =3 8013.787, freq =3
Age	53.543124	53.0000	47, freq = 152
TotalCharge	5306.435876	5100.2605	7555.452, freq=2 7964.681, freq=2
Marital_Married	.203889	0.0000	0, freq = 7329
Doc_visits_7	.063871	0.0000	0, freq = 8618
Initial_admin_Emergency_Admission	.505323	1.0000	1, freq = 4652
Initial_admin_Observations_Admission	.244189	0.0000	0, freq= 6958
HighBlood_Yes	.407886	0.0000	0, freq = 5451
Stroke_Yes	.199001	0.0000	0, freq = 7374
Complication_risk_Low	.212036	0.0000	0, freq = 7254
Options_3	.341408	0.0000	0, freq = 6063
Options_4	.344123	0.0000	0, freq = 6038
Courteous_2	.132848	0.0000	0, freq = 7983

Pandas Dataframe.describe() method is used to output statistical details for count of observations, mean value, standard deviation, min & max values, and percentiles (25th, 50th, and 75th) for target variable and all predictor variables. Dataframe.corr() method is used to output the correlation between each variable. The screenshot below displays this information for our target variable and all predictor variables.

```
dfc.describe()
```

	Additional_charges	Age	TotalCharge	Marital_Married	Doc_visits_7	Initial_admin_Emergency_Admission	Initial_admin_Observation_Admission	HighBlood_Yes	Stroke_Yes	Complication_risk_Low	Options_3	Options_4	Courteous_2
count	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000	9206.000000
mean	12927.980718	53.543124	5306.435876	0.203889	0.063871	0.505323	0.244189	0.407886	0.199001	0.212036	0.341408	0.344123	0.132848
std	6540.592828	20.609439	2181.251460	0.402909	0.244537	0.499999	0.429628	0.491468	0.399270	0.408772	0.474208	0.475107	0.339429
min	3125.703000	18.000000	1938.312067	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	7991.171750	36.000000	3178.291852	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	11556.775000	53.000000	5100.260500	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	15602.158960	71.000000	7458.542500	0.000000	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000	1.000000	1.000000	0.000000
max	30566.070000	89.000000	9180.728000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

```
dfc.corr()
```

	Additional_charges	Age	TotalCharge	Marital_Married	Doc_visits_7	Initial_admin_Emergency_Admission	Initial_admin_Observation_Admission	HighBlood_Yes	Stroke_Yes	Complication_risk_Low	Options_3	Options_4	Courteous_2
Additional_charges	1.000000	0.716409	0.022020	0.025245	0.022486	0.036228	-0.034389	0.655680	0.033301	-0.038131	0.026605	-0.027679	-0.022164
Age	0.716409	1.000000	0.010785	0.012580	0.005877	-0.004498	-0.010404	0.008265	0.011657	0.000604	0.014939	-0.019079	-0.023345
TotalCharge	0.022020	0.010785	1.000000	0.000992	-0.000072	0.107284	-0.069032	0.015240	-0.007641	-0.014872	-0.001122	0.000477	0.006867
Marital_Married	0.025245	0.012580	0.000992	1.000000	0.005638	0.016453	-0.012139	0.022163	-0.006432	-0.001313	-0.006154	0.013666	-0.008226
Doc_visits_7	0.022486	0.005877	-0.000072	0.005638	1.000000	0.015878	0.000431	0.026361	0.012226	-0.005083	-0.019437	0.010899	-0.006694
Initial_admin_Emergency_Admission	0.036228	-0.004498	0.107284	0.016453	0.015878	1.000000	-0.574486	0.000227	-0.014557	0.007766	-0.017974	0.019728	0.009596
Initial_admin_Observation_Admission	-0.034389	-0.010404	-0.069032	-0.012139	0.000431	-0.574486	1.000000	0.000037	0.008009	-0.001024	0.022137	-0.027989	-0.012398
HighBlood_Yes	0.655680	0.008265	0.015240	0.022163	0.026361	0.000227	0.000037	1.000000	0.005953	-0.031468	0.013524	-0.015439	-0.008365
Stroke_Yes	0.033301	0.011657	-0.007641	-0.006432	0.012226	-0.014557	0.008009	0.005953	1.000000	-0.006955	0.006622	-0.016857	0.024547
Complication_risk_Low	-0.038131	0.000604	-0.014872	-0.001313	-0.005083	0.007766	-0.001024	-0.031468	-0.006955	1.000000	-0.001921	-0.009917	0.005231
Options_3	0.026605	0.014939	-0.001122	-0.006154	-0.019437	-0.017974	0.022137	0.013524	0.006622	-0.001921	1.000000	-0.521524	-0.073933
Options_4	-0.027679	-0.019079	0.000477	0.013666	0.010899	0.019728	-0.027989	-0.015439	-0.016857	-0.009917	-0.521524	1.000000	0.051964
Courteous_2	-0.022164	-0.023345	0.006867	-0.008226	-0.006694	0.009596	-0.012398	-0.008365	0.024547	0.005231	-0.073933	0.051964	1.000000

C3: Steps to Prepare Data

All code used to clean and prepare the data for analysis can be found in the included PDF 'D208_Task_1'. All coded inputs and their generated outputs are shown executed without errors.

Step 1: Environment setup

Libraries are imported into the development environment in preparation for the data preparation and analysis. These include analytical libraries such as pandas, scipy, and numpy, as well as visual generation libraries like seaborn, pyplot, and missingno. The data is loaded into a JupyterNotebook.

Step 2: Cleaning and Preparation

The dataset is analyzed as a whole to look at data types, if there's missing values, and if any values need to be altered or deleted. Correlation tables are mapped. Z-scores are calculated and outliers z-score >3 are removed. Visualization tools such as missingno's matrix, seaborn's boxplots, and pyplot's histograms are used to plot Univariate and Bivariate graphs.

Dummy variables are created using pd.get_dummies(), with drop_first parameter set = True to ensure that the first variable is dropped. This ensures that each dummy set has k-1 number of features. All applicable variables are renamed for utility and ease of use. Heatmaps and correlation tables are drawn looking for which variables most closely correlate with our target dependent variable, Additional_charges.

At this point the data set is cleaned, dummied, and ready for initial feature selection for Multiple Regression. The original dataset contained 50 columns/variables, our prepared data set contains 98 columns/features. The increase in size is due to the creation of dummy variables for categorical features that have multiple responses, such as 'Doc_visits' and all of the survey response variables.

Initial independent features are chosen from this data set through statistical feature selection. This was performed by selecting all independent features that have a correlation > .02 with our dependent feature, Additional_charges. This left us with the 12 features from section C2 for our initial independent variables. Our data is prepared for Multiple Regression.

C4: Visualizations

Both univariate and bivariate visualizations of the distributions of variables have been created to analyze this data set. Visualizations are included in PDF 'D208_Task_1' of the Jupyter Notebook used for all code for this PA. All bivariate visualizations (pairplots) are mapped with the target variable on the X-Axis, and the independent variables on the Y-Axis.

C5: Prepared Dataset

I was unsure if the PA rubric requirement read as providing a copy of the fully prepared data set (all variables, pre-selection of variables for multiple regression), or a copy of the fully prepared data set (post-variable selection preparation). I've included both in the submission.

D1: Initial Model

The table below provides the results of the initial multiple regression model containing all independent variables statistically selected from section C2.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Additional_charges   R-squared:                       0.938
Model:                            OLS   Adj. R-squared:                  0.938
Method:                 Least Squares   F-statistic:                 1.262e+04
Date:                Sat, 25 Jun 2022   Prob (F-statistic):               0.00
Time:                        15:23:26   Log-Likelihood:                -81153.
No. Observations:                9206   AIC:                         1.623e+05
Df Residuals:                    9194   BIC:                         1.624e+05
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================================================
                                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                -2927.1017     76.268    -38.379      0.000   -3076.604   -2777.600
Age                                    225.5603      0.825    273.245      0.000     223.942     227.178
TotalCharge                              0.0001      0.008      0.016      0.987      -0.015       0.015
Marital_Married                         23.0179     42.216      0.545      0.586     -59.735     105.771
Initial_admin_Emergency_Admission      471.1003     41.691     11.300      0.000     389.376     552.825
Initial_admin_Observation_Admission   -101.9210     48.360     -2.108      0.035    -196.717      -7.125
HighBlood_Yes                        8636.3317     34.625    249.428      0.000    8568.460    8704.204
Stroke_Yes                            353.1468     42.605      8.289      0.000     269.631     436.663
Complication_risk_Low                -292.2811     41.617     -7.023      0.000    -373.860    -210.702
Options_3                             100.3532     42.081      2.385      0.017      17.866     182.840
Options_4                             -14.0323     41.964     -0.334      0.738     -96.290      68.226
Courteous_2                            -7.7754     50.261     -0.155      0.877    -106.298      90.747
==============================================================================
Omnibus:                     1310.043   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              316.783
Skew:                          -0.020   Prob(JB):                     1.63e-69
Kurtosis:                       2.092   Cond. No.                     2.83e+04
```

Multiple regression was performed using Ordinary Least Squares (OLS) regression analysis.

D2: Justification of Model Reduction

Variable selection for model reduction is performed by:

- Looking for multicollinearity between variables by calculating VIF for our independent variables.
- Calculating p-values for each variable to assess statistical significance.
- Using K-neighbors Regression Analysis at different levels of correlation with our dependent variable and cross validated and analyzed for contribution to R_squared and Root Mean Square Error (RMSE) using sklearn's cross_val_predict function. This method allows us to assess overfitting by cross validating our subsets, and addresses variable's residuals by reducing features to lower model RMSE.

VIF

Any variables showing VIF > 5 will be removed from the independent variables. The table below displays VIF measures for each independent feature.

:

	VIF	variable
0	20.132471	Intercept
1	1.001660	Age
2	1.012415	TotalCharge
3	1.001234	Marital_Married
4	1.503821	Initial_admin_Emergency_Admission
5	1.493896	Initial_admin_Observation_Admission
6	1.002135	HighBlood_Yes
7	1.001448	Stroke_Yes
8	1.001558	Complication_risk_Low
9	1.378054	Options_3
10	1.375611	Options_4
11	1.007220	Courteous_2

VIF shows no breach of multicollinearity between independent variables. All variables are kept from these results.

P-VALUES

Next we'll calculate the P-value for each variable, based on their T-score. Variables will be removed from the model if the P-value is > .05. Degrees of Freedom (DOF) for the calculation of P-values is 11, from the OLS model.

Variable	T-Score	P-Value	Action Taken
Age	273.245	<.00001	Keep
TotalCharge	.016	.987521	Remove
Married_Married	42.216	<.00001	Keep
Initial_admin_Emergency_Admission	41.691	<.00001	Keep
Initial_admin_Observation_Admission	48.360	<.00001	Keep
HighBlood_Yes	34.625	<.00001	Keep
Stroke_Yes	42.605	<.00001	Keep

Complication_risk_Low	41.617	<.00001	Keep
Options_3	42.081	<.00001	Keep
Options_4	41.964	<.00001	Keep

All variables but TotalCharge show a high level of statistical significance, and are kept in the reduced model so far.

K-NEAREST NEIGHBORS (KNN) AND CROSS VALIDATION

K-nearest neighbor regression and cross validation is performed and looped at increasing levels of variable correlation to select features (Feeley, 2020). This classification technique uses proximity to classify and predict groupings of data points. Sklearn's KFold splits data into testing and training sets, in our case 10 sets. R_squared, and RMSE are returned. Initial model analysis returns RMSE of 1838.19 and R_squared of .92. When only Age and HighBlood_Yes are included as independent features, KNN returns RMSE of 408.78 and R_Squared of 1. Below is the code used for analysis and resulting outputs.

```
X=df[['Age','Marital_Married','Initial_admin_Emergency_Admission','Initial_admin_Observation_Admission','HighBlood_Yes','Stroke_Yes','Complication_risk_Low','Options_3','Options_4','Courteous_2']]
y = df.Additional_charges
```

```
cv = KFold(n_splits=10, random_state=0, shuffle=True)
classifier_pipeline = make_pipeline(StandardScaler(), KNeighborsRegressor(n_neighbors=10))
y_pred = cross_val_predict(classifier_pipeline, X, y, cv=cv)
print("RMSE: " + str(round(sqrt(mean_squared_error(y,y_pred)),2)))
print("R_squared: " + str(round(r2_score(y,y_pred),2)))
```

```
RMSE: 1395.79
R_squared: 0.95
```

```
vals = [0.02,0.05,0.08,0.1,0.2]
for val in vals:
    features = abs(df.corr()["Additional_charges"][abs(df.corr()["Additional_charges"])>val].drop('Additional_charges')).index.tolist()

    X = df.drop(columns='Additional_charges')
    X=X[features]

    print(features)

    y_pred = cross_val_predict(classifier_pipeline, X, y, cv=cv)
    print("RMSE: " + str(round(sqrt(mean_squared_error(y,y_pred)),2)))
    print("R_squared: " + str(round(r2_score(y,y_pred),2)))
```

```
['Age', 'TotalCharge', 'Marital_Married', 'Doc_visits_7', 'Initial_admin_Emergency_Admission', 'Initial_admin_Observation_Admission', 'HighBlood_Yes', 'Stroke_Yes', 'Complication_risk_Low', 'Options_3', 'Options_4'
Courteous_2']
RMSE: 1838.19
R_squared: 0.92
['Age', 'HighBlood_Yes']
RMSE: 408.78
R_squared: 1.0
['Age', 'HighBlood_Yes']
RMSE: 408.78
R_squared: 1.0
['Age', 'HighBlood_Yes']
RMSE: 408.78
R_squared: 1.0
['Age', 'HighBlood_Yes']
RMSE: 408.78
R_squared: 1.0
```

By combining the results of these three statistical selection and analysis techniques our reduced regression model will only include independent variables 'Age' and 'HighBlood'_Yes.

D3: Reduced Multiple Regression Model

Results from the statistical analysis performed in section D2 reduce our multiple regression model to having 'Age' and 'HighBlood_Yes' for independent features. 'Age' is a continuous variable and 'HighBlood_Yes' is categorical. The below table are the results from OLS analysis with these two features.

OLS Regression Results

Dep. Variable:	Additional_charges	R-squared:	0.935
Model:	OLS	Adj. R-squared:	0.935
Method:	Least Squares	F-statistic:	6.669e+04
Date:	Sat, 25 Jun 2022	Prob (F-statistic):	0.00
Time:	15:25:20	Log-Likelihood:	-81330.
No. Observations:	9206	AIC:	1.627e+05
Df Residuals:	9203	BIC:	1.627e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2681.5619	50.211	-53.406	0.000	-2779.986	-2583.137
Age	225.6544	0.840	268.483	0.000	224.007	227.302
HighBlood_Yes	8647.7532	35.245	245.361	0.000	8578.665	8716.841

Omnibus:	931.630	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	269.728
Skew:	-0.021	Prob(JB):	2.69e-59
Kurtosis:	2.163	Cond. No.	172.

E1: Model Comparison

Below is a copy of the Initial Mode summary.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Additional_charges   R-squared:                       0.938
Model:                            OLS   Adj. R-squared:                  0.938
Method:                 Least Squares   F-statistic:                 1.262e+04
Date:                Sat, 25 Jun 2022   Prob (F-statistic):               0.00
Time:                        15:23:26   Log-Likelihood:                -81153.
No. Observations:                9206   AIC:                         1.623e+05
Df Residuals:                    9194   BIC:                         1.624e+05
Df Model:                          11
Covariance Type:            nonrobust
=============================================================================================================
                                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
const                                   -2927.1017     76.268    -38.379      0.000   -3076.604   -2777.600
Age                                       225.5603      0.825    273.245      0.000     223.942     227.178
TotalCharge                                 0.0001      0.008      0.016      0.987      -0.015       0.015
Marital_Married                            23.0179     42.216      0.545      0.586     -59.735     105.771
Initial_admin_Emergency_Admission         471.1003     41.691     11.300      0.000     389.376     552.825
Initial_admin_Observation_Admission      -101.9210     48.360     -2.108      0.035    -196.717      -7.125
HighBlood_Yes                           8636.3317     34.625    249.428      0.000    8568.460    8704.204
Stroke_Yes                                353.1468     42.605      8.289      0.000     269.631     436.663
Complication_risk_Low                    -292.2811     41.617     -7.023      0.000    -373.860    -210.702
Options_3                                 100.3532     42.081      2.385      0.017      17.866     182.840
Options_4                                 -14.0323     41.964     -0.334      0.738     -96.290      68.226
Courteous_2                                -7.7754     50.261     -0.155      0.877    -106.298      90.747
==============================================================================
Omnibus:                     1310.043   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              316.783
Skew:                          -0.020   Prob(JB):                     1.63e-69
Kurtosis:                       2.092   Cond. No.                     2.83e+04
```

Below is a copy of the Reduced Model summary.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Additional_charges   R-squared:                       0.935
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                 6.669e+04
Date:                Sat, 25 Jun 2022   Prob (F-statistic):               0.00
Time:                        15:25:20   Log-Likelihood:                -81330.
No. Observations:                9206   AIC:                         1.627e+05
Df Residuals:                    9203   BIC:                         1.627e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -2681.5619     50.211    -53.406      0.000   -2779.986   -2583.137
Age          225.6544      0.840    268.483      0.000     224.007     227.302
HighBlood_Yes  8647.7532     35.245    245.361      0.000    8578.665    8716.841
==============================================================================
Omnibus:                      931.630   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              269.728
Skew:                          -0.021   Prob(JB):                     2.69e-59
Kurtosis:                       2.163   Cond. No.                         172.
```

After reduction only Age and HighBlood_Yes are left in the model. With the reduction of 10 features the coefficient (coef) for has changed, as compared below.

Feature	Initial Model	Reduced Model	Result
const	-2927.1017	-2681.5619	Slightly lower in reduced model .
Age	225.5603	225.6544	Slightly higher in the reduced model.
HighBlood_Yes	8636.3317	8647.7532	Slightly higher in the reduced model.

Comparing the initial and reduced models yields a number of valuable insights, presented below.

Result	Initial Model	Reduced Model	Comparison Analysis
Method	OLS	OLS	Same technique
No. Observations	9206	9206	Same result
DF, Model	11	2	Result of feature reduction
Covariance Type	Nonrobust	Nonrobust	Same result
R-Squared	0.938	0.935	.003 less in reduced model
Adj. R-Squared	.938	.935	.003 less in reduced model
F-Statistic	1.262e+04	6.669e+04	Higher F-stat of the reduced model indicates that reducing variables resulted in all variables in the model being significant.
Log-likelihood	-81153	-81330	Initial model shows a very **slight** better fit of model to the data. At this level of difference it is not significant to our analysis.
AIC	1.623e+05	1.627e+05	Initial model displays a slightly better fit. At this level not significant to our analysis
BIC	1.624e+05	1.627e+05	Initial model displays a slightly better fit. At this level not significant to our analysis.
Durbin-Watson	1.996	1.998	Both models indicate there is no autocorrelation in the data set.
Jarque-Bera	316.783	269.728	Results indicate data is not normally distributed for both models. Reduced model slightly more normal. At 9206 observations this is less relevant than if the data set were smaller (e.g. < ~40).

The analysis process differences between the initial and reduced model are modest. Even though the initial model has 10 more features, most of the significance from both models comes from just Age and HighBlood_Yes. This means that with feature reduction the models' summarized calculations change minimally. Feature reduction in this case has been successful and a sensible statistical method in simplifying the results to answer our research question.

E2: Output and Calculations

Please refer to 'D208_Task_1' PDF attached to submission for residual error visualization and calculations of all analysis performed.

E3: Code

Please refer to 'D208_Task_1' PDF attached to submission for a complete and accurate copy of all code used in this research analysis.

F1: Results

To discuss the result of the analysis, let's begin with restating the research question: What variables best predict the amount of Additional Charges spent by patients in the hospital? The results of the initial model, reduced feature model, and the limitations of the analysis are discussed below.

Initial Regression Model results and analysis

The screenshot below shows the summary OLS results for the initial model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Additional_charges   R-squared:                       0.938
Model:                            OLS   Adj. R-squared:                  0.938
Method:                 Least Squares   F-statistic:                 1.157e+04
Date:                Thu, 30 Jun 2022   Prob (F-statistic):               0.00
Time:                        11:42:37   Log-Likelihood:                -81153.
No. Observations:                9206   AIC:                         1.623e+05
Df Residuals:                    9193   BIC:                         1.624e+05
Df Model:                          12
Covariance Type:            nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                 -2927.6667     76.352    -38.344      0.000   -3077.334   -2777.999
Age                                     225.5595      0.826    273.225      0.000     223.941     227.178
TotalCharge                               0.0001      0.008      0.017      0.987      -0.015       0.016
Marital_Married                          22.9864     42.219      0.544      0.586     -59.772     105.745
Doc_visits_7                             11.2221     69.579      0.161      0.872    -125.169     147.613
Initial_admin_Emergency_Admission       470.9664     41.702     11.294      0.000     389.222     552.711
Initial_admin_Observation_Admission    -102.0149     48.366     -2.109      0.035    -196.823      -7.207
HighBlood_Yes                         8636.1851     34.638    249.325      0.000    8568.286    8704.084
Stroke_Yes                             353.0603     42.611      8.286      0.000     269.533     436.587
Complication_risk_Low                 -292.2519     41.620     -7.022      0.000    -373.836    -210.668
Options_3                              100.4664     42.089      2.387      0.017      17.963     182.969
Options_4                              -14.0420     41.966     -0.335      0.738     -96.304      68.221
Courteous_2                             -7.7094     50.265     -0.153      0.878    -106.240      90.822
==============================================================================
Omnibus:                     1309.636   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              316.742
Skew:                          -0.020   Prob(JB):                     1.66e-69
Kurtosis:                       2.092   Cond. No.                     2.84e+04
==============================================================================
```

The initial model's regression equation model is identified by interpreting the coefficients. The constant term 'const' coefficient is the intercept of the regression line, -2927.1017. Column 'coef' displays the coefficient for each predictive variable. The coefficient quantifies the change in Additional_charges for every unit change in the predictive variable. For example, variable 'Age' has a coefficient of 225.5603. This can be interpreted as for every year older the patient is we can predict an increase in Additional_charges by 225.5603. The regression equation is the summation of all the predictor variables multiplied by their coefficient. The full equation can be shown as:

Additional_charges = -2927.1017+Age(225.5603)+TotalCharge(.0001)+Marital_Married(23.0179)+Initial_admin_Emergency_Admission(471.1003)-Initial_admin_Observation_Admission(101.9210)+HighBlood_Yes(8636.3317)+Stroke_Yes(353.1468)-Complication_risk_Low(292.2811)+Options_3(100.3532)-Options_4(14.0323)-Courteous_2(7.7754)

The statistical significance of the initial model is analyzed by looking at the OLS regression results above. At an alpha of .05 we can statistically identify predictive variables that contribute to Additional_charges. Variables const, Age, Initial_admin_Emergency_Admission, Initial_admin_Observation_Admission, HighBlood_Yes, Stroke_Yes, Complication_risk_Low, and Options_3 have P(t) values less than .05. They are statistically significant at this level of confidence, and reject the null hypothesis. Variables TotalCharge, Marital_Married, Doc_visits_7, Options_4, and Courteous_2 all have P(t) values greater than .05, and fail to reject the null hypothesis. This is one statistical method for identifying predictive variables to keep in the reduced model. AIC and BIC, metrics used to compare the fit of our model, are 1.623e+05 and 1.624e+05, respectively. AIC and BIC will be compared with results from the reduced model to compare model fit.

To address the practical significance in relation to the statistical significance of the model we ask if the model results have real-world application and meaning. With an R-squared and Adj. R-squared of .938 this model explains 93.8% of the variance with the chosen predictive variables. This is statistically quite good, but we know that there's five predictive variables in the model that are not significant. In order to have a more practically significant model we would want to eliminate these features, and identify any other variables that are not a good model fit..

Reduced Feature Regression Model Results

Taking what we learned from the initial regression model we know that we have predictive features that can be removed from the model. VIF is calculated for each predictor variable, and K-nearest neighbor regression is performed to statistically evaluate all initial predictor variables. The results indicate that predictor variables Age and HighBlood_Yes account for the vast majority of the model's regression results. Below is the results of OLS regression results from the reduced model.

```
6]: ols = sm.OLS(y, X).fit()

7]: print(ols.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Additional_charges   R-squared:                       0.935
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                 6.669e+04
Date:                Thu, 30 Jun 2022   Prob (F-statistic):               0.00
Time:                        11:43:02   Log-Likelihood:                -81330.
No. Observations:                9206   AIC:                         1.627e+05
Df Residuals:                    9203   BIC:                         1.627e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -2681.5619     50.211    -53.406      0.000   -2779.986   -2583.137
Age          225.6544      0.840    268.483      0.000     224.007     227.302
HighBlood_Yes  8647.7532     35.245    245.361      0.000    8578.665    8716.841
==============================================================================
Omnibus:                      931.630   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              269.728
Skew:                          -0.021   Prob(JB):                     2.69e-59
Kurtosis:                       2.163   Cond. No.                         172.
==============================================================================
```

The reduced multiple regression equation model is identified by interpreting the coefficients. The constant term 'const' coefficient is the intercept of the regression line, -2681.5619. Column 'coef' is the coefficient for each predictive variable. The coefficient quantifies the change in Additional_charges for every unit change in the predictive variable. For example, variable 'Age' has a coefficient of 225.6544, which is up very slightly from the initial model's coefficient for Age. This can be interpreted as for every year older the patient is we can predict an increase in Additional_charges by 225.6544. The regression equation is the summation of 'const' and all predictor variables multiplied by their coefficient. For the reduced multiple regression model the regression equation is:

Additional_charges = -2681.5619+Age(225.6544)+HighBlood_Yes(8647.7532)

The statistical significance of the reduced model is analyzed by looking at the OLS regression results above. At an alpha of .05 we can statistically identify significant predictive variables. All predictive variables in the predictive model have a P(T) of 0.000, rejecting the null hypothesis at alpha =.05. This means that all of our variables are significant in relation to Additional_charges. AIC and BIC are both 1.627e+05, slightly higher than the initial model, but not enough to conclude that the initial model is a better fit.

The Practical significance of the reduced model becomes apparent when compared to the initial model. With an R-squared and Adj. R-squared of .938, the reduced model is only .003 less on both metrics than the initial model with 10 less predictive variables. The reduction in complexity of the reduced model lends itself to much greater practical significance with nearly the same explanation of model variance.

Limitations of the analysis

A major limitation to the analysis is that it is quite possible linear analysis is not the best analytical technique for addressing our research question with this data set. There were very few potential predictor variables that showed any significant correlation with Additional_charges. This limits the number of predictor variables that can be included in the initial, and ultimately, reduced multiple regression models.

Predictive variable problems are another limitation. With so few predictive variables that showed correlation with Additional_charges to choose for the initial multiple regression model their normality and distribution can affect the analysis integrity. Another limitation of the analysis is implying that correlation = causation. For example, HighBlood_Yes was the second most predictive variable in the analysis. It also has a very high coefficient of 8647.7532. The CDC estimates that 47% of the adult American population has high blood pressure (Facts About Hypertension, N.d.). To imply causation for 8647.7532 of additional charges incurred by patients having it would need further analysis to be useful in a professional and practical manner.

F2: Recommendations

The overall result of the analysis is that multiple regression is not a great tool used to assess this question with this data set. The features can be analyzed using multiple regression, but the results are less than optimal and lack viability. This is in large part due to the assumption that residuals should be normally distributed with a mean of zero is not able to be adhered to in this analysis.

Recommendations for how our organization applies the results of this analysis to find an appropriate course of action are:

- Utilize the reduced multiple regression model to calculate additional charges but only to provide a rough, non legally binding estimate. The model has the potential to give a reliable back-of-hand estimate for in-house collection and revenue models,, but in no way should it be relied on exclusively.

- Our organization should decide how critical it is to be able to accurately predict additional charges. If an extremely accurate and reliable calculation is necessary, we should look at alternative modeling processes that can fit the data better, such as a logistic model.
- The CDC estimates that 47% of United States adults have high blood pressure (Facts About Hypertension, N.d.). Our model heavily weighs on this feature as a prediction of the additional charges billed to patients. My recommendation is that although the model predicts and weighs this feature heavily, it is too common in the population of our patients to predict such a high amount of the additional charges. This would need to be further analyzed to provide real-world utility.
- Ultimately, if being able to predict additional charges beforehand is a critical item to predict based off of other gathered features in the data set, I recommend looking at different analytical techniques. Our organization would need to allocate more time and resources to generating an accurate viable model.

G: Panopto Demonstration

Included in this submission is a Panopto recording demonstrating the execution of all code, recording both myself and my display.

H: Sources of Third-Party Code

Bushmanov, Sergey. (January 28, 2019). *How to remove Outliers in Python?.* Stackoverflow. https://stackoverflow.com/questions/54398554/how-to-remove-outliers-in-python

Feeley, Ciara. [Ph.D. and Productivity]. (May 21, 2020). *Feature Selection in Python | Machine Learning Basics | Boston Housing Data* [Video]. YouTube. https://www.youtube.com/watch?v=iJ5c-XoHPFo&ab_channel=PhDandProductivity

Seaborn.residplot. (N.d.). seaborn. https://seaborn.pydata.org/generated/seaborn.residplot.html

Zach. (July 20, 2020). *How to Calculate VIF in Python.* Statology.org. https://www.statology.org/how-to-calculate-vif-in-python/

Ordinary Least Squares (OLS) using statsmodels. (Mar 10, 2022). Geeksforgeeks.org. https://www.geeksforgeeks.org/ordinary-least-squares-ols-using-statsmodels/

Seaborn.heatmap. (N.d.). seaborn. https://seaborn.pydata.org/generated/seaborn.heatmap.html

Pandas.get_dummies. (N.d.). pandas.

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html

I: Sources

Sewell, William. D208 Predictive Modeling - Episode 4 [Minute 2:00]. College of IT. WGU.

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=5328e088-48c2-42d7-a5e3-ad27015d138f

Zach. (November 16, 2021). *The Five Assumptions of Multiple Linear Regression.* Statology.org.

https://www.statology.org/multiple-linear-regression-assumptions/

Zach. (September 18, 2018). *Measures of Central Tendency: Definition and Examples.* Statology.org.

https://www.statology.org/measures-central-tendency/

*Data Manipulation with Python. (*N.d.). EDUCBA.com.

https://www.educba.com/data-manipulation-with-python/

Feeley, Ciara. [Ph.D. and Productivity]. (May 21, 2020). *Feature Selection in Python* | *Machine Learning Basics* | *Boston Housing Data* [Video]. YouTube.

https://www.youtube.com/watch?v=iJ5c-XoHPFo&ab_channel=PhDandProductivity

Facts About Hypertension. (N.d.). Centers for Disease Control and Prevention (CDC).

https://www.cdc.gov/bloodpressure/facts.htm