PDANA8411 POE PART 1

Brice Derek Agnew (ST10072411)

IIE VARSITY COLLEGE Cape Town

Contents

-Create a linear regression model that 'Provides a sliding scale of charges according to the age, sex, BMI, Number of children, smoking habits, and geographical region.'

Been provided with an 'initial dataset' to train the model : Medical Cost Personal Datasets

Suitability:

The features within this data set use continuous and categorical values, with some binary values, which can be used for linear regression provided that the categorical values are correctly encoded (Hannay, 2025).

The Dataset in question has some quality issues such as duplicates, missing values, outliers, and skewed distributions. The remedial actions taken for all these initial issues will be described below. Additionally, no real multicollinearity can be seen in the dataset(Singh, 2024), which aligns with what is needed for linear regression.

Some common pitfalls of linear regression that should be avoided are in assuming that all relationships are linear, which we can investigate with the use of correlation coefficients and data visualizations, and overfitting the model by training it on uncorrelated features(Grant, Hickey and Head, 2018).

Plan:

The Data Analysis that will be done in this report will follow the steps of the data analytics process:

- Data collection: The data used will be gathered based on relevance and reliability. The dataset provided (Choi, 2018) seems relevant to the case of insurance price prediction.

- Data Cleaning: The Dataset to be used needs to be sifted through to determine any irregularities, missing values, duplcate or outlier values(Codecademy Team, 2025).

    1. The missing values are identified and any found are deleted.
    2. Any duplicate values are identified and deleted.
    3. The categorical data found in some independent features are converted into discrete or binary numeric values via label encoding.
    4. Outliers are identified in the dataset by evaluating the z-scores of each item in each column and transforming all outliers with the use of log transformations and quantile-based clipping.

- Exploratiory Data Analysis: The Dataset will be broken down with key statistics and previously mentioned data errors identified.
    1. Initial key statistics will be identified for each column, these statistics include the mean, the median, the standard deviation, the lower and upper quantile boundaries, the maximum value and the number of items in each column.
    2. The Distributions of each feature or column will be visually represented with the use of histograms for continuous data, and barcharts for the previously encoded data.
    3. The relationships between each independent variable and the dependent feature we are seeking to predict, the continuous ‘charges’ values, with the use of various visualizations. For the relationships between the dependent variable and Binary values, such as ‘smoker’ or ‘sex’, a boxplot will be implemented to display changes in means and quantile boundaries. For the relationships between the dependent variable and Discrete values, such as ‘region’ or ‘children’, a Boxplot will also be implemented for similar reasons. For the relationships between tbhe dependent variable and other continuous values, such as ‘bmi’, we will be using a scatter plot with a regression line to illustrate the trend (Codecademy Team, 2025) . An exception to this was the usage of a

scatterplot with the discrete values of ‘age’, which due to its numerious unique types of answers could be treated as a continuous variable.

4. After assessing the relationships between each independent variable and the dependent variable, the Variance Inflation Factor (VIF) (Singh, 2024) scores of the independent variables will be calculated to assess any multicolinearity between the predictor values.
5. The Correlation Coefficient or other relevant correlation metrics will then be calculated and interpretted for each feature column to determine which independent variables have the most prominent effect on the dependent variable. Particularly, for determining the correlation between Binary values , such as ‘smoker’ and ‘sex’, and the continuous dependent variable, Point-Biserial Correlation will be used (statisticshowto.com, 2016). For determining the correlation between discrete values, such as ‘age’ and ‘children’, and the continuous dependent value, Spearman correlation will be used (statisticshowto.com, 2021). For determining the correlation between continuous variables, particularly ‘bmi’, and the dependent variable, Pearson Correlation will be used. For the correlation of ‘region’ to the dependent variable, ANOVA correlation will be used, since the previously mentioned correlation methods may not be ideal for determining the correlation of categorical variables, or at encoded categorical values(statisticseasily.com, 2024).

- Training the model: After the cleaning and data analysis steps are undertaken, the dataset will then be used to train two models, one standard multiple linear regression model, and one Lasso linear regression model. This will be done in order to measure and compare the performance of our initial linear regression model with an alternative. Before the models are fit the dataset will be split, with 20% of the entries forming the tresting data and 80% forming the training data. The entries selescted for this split is initially random, to ensure that this testing is recreatable at a later stage.

- Model evaluation: The models developed will evaluatied based on their coefficeint of determination, their ‘Root Mean Square Error’ value, and a comparison of their actual and predicted values.

Report:

Data Cleaning:

In the process of cleaning the data, the following insights were found and actions taken.

The program found that there were no missing values in the entie set, as illustrated by this output:

```
Missing Values per column:
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64

 Total number of missing values in set: 0
```

The program found that there was only one duplicated datapoint in the set, and outputted which row it was. The duplicate was then summarily deleted. This can be seen with this output:

```
Total number of Duplicated rows: 1

 The Actual Duplicated rows:
      age   sex    bmi  children smoker     region    charges
581   19  male  30.59         0     no  northwest  1639.5631


Total number of Duplicate columns: 0

Duplicate rows deleted.

Total number of Duplicated rows: 0
```

The program was made to encode any categorical value , namely the values in ‘region’, ‘sex’, and ‘smoker’, into a numeric, discrete value:

```
age           int64
sex           int64
bmi         float64
children      int64
smoker        int64
region        int64
charges     float64
dtype: object

[Testing to see if categorical data has been successfully encoded]
The new region value for the first row:  3
```

The Program was made to identify if any outliers present in ach of the columns, it's results were as follows:

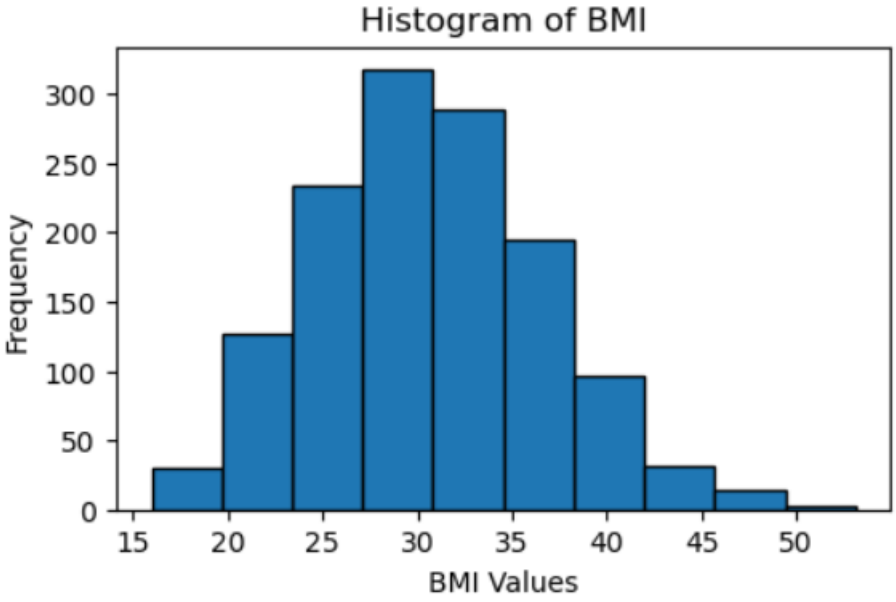
```
Number of outliers per column (original):
 age           0
sex            0
bmi            4
children      18
smoker         0
region         0
charges        7
dtype: int64
```

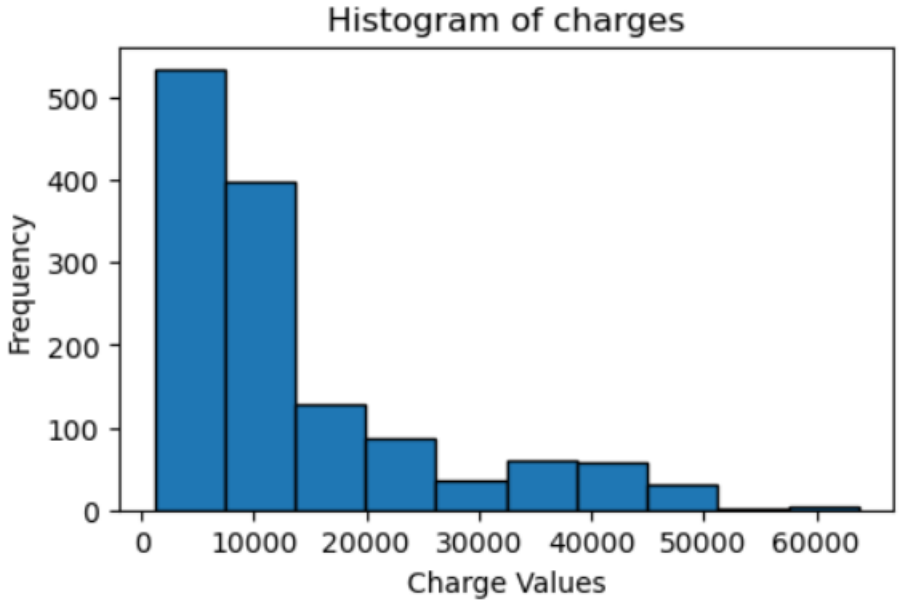
With about 29 outliers identified, the program then displayed which rows were seen as the outliers:

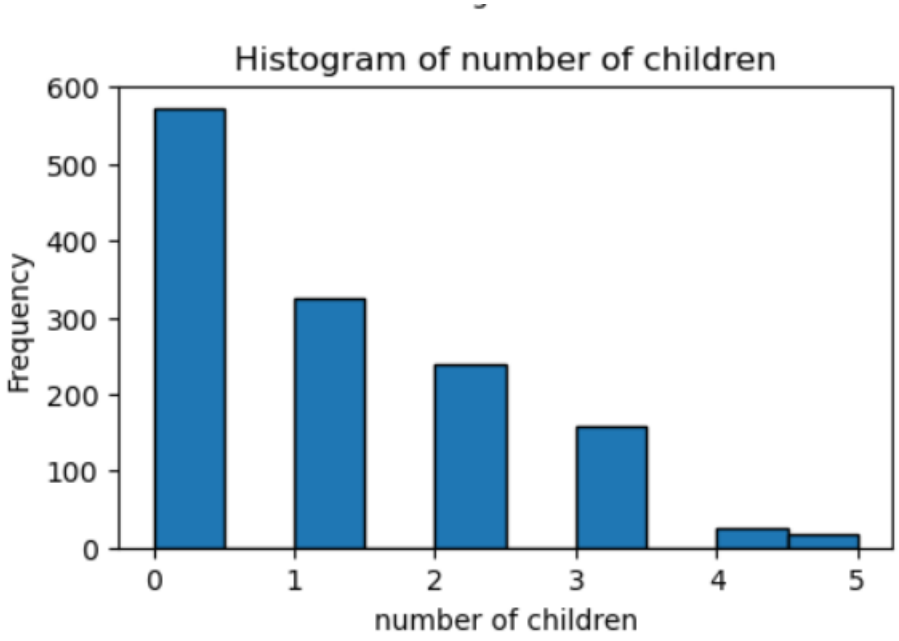
```
The outlier rows:
       age  sex     bmi  children  smoker  region      charges
32      19    0  28.600         5       0       3   4687.79700
34      28    1  36.400         1       1       3  51194.55914
71      31    1  28.500         5       0       0   6799.45800
116     58    1  49.060         0       0       2  11381.32540
166     20    0  37.000         5       0       3   4830.63000
413     25    1  23.900         5       0       3   5080.09600
425     45    1  24.310         5       0       2   9788.86590
438     52    0  46.750         5       0       2  12592.53450
543     54    0  47.410         0       1       2  63770.42801
568     49    0  31.900         5       0       3  11552.90400
577     31    0  38.095         1       1       0  58571.07448
640     33    1  42.400         5       0       3   6666.24300
819     33    0  35.530         0       1       1  55135.40209
847     23    1  50.380         1       0       2   2438.05520
877     33    1  33.440         5       0       2   6653.78860
932     46    1  25.800         5       0       3  10096.97000
937     39    0  24.225         5       0       1   8965.79575
969     39    0  34.320         5       0       2   8596.82780
984     20    1  30.115         5       0       0   4915.05985
1047    22    1  52.580         1       1       2  44501.39820
1085    39    0  18.300         5       1       3  19023.26000
1116    41    1  29.640         5       0       0   9222.40260
1130    39    0  23.870         5       0       2   8582.30230
1146    60    1  32.800         0       1       3  52590.82939
1230    52    1  34.485         3       1       1  60021.39897
1245    28    1  24.300         5       0       3   5615.36900
1272    43    1  25.520         5       0       2  14478.33015
1300    45    1  30.360         0       1       2  62592.87309
1317    18    1  53.130         0       0       2   1163.46270
```

These outliers identified are not unreasonable, and since this is a small dataset, I decided to transform the outliers instead of deleteing them in order to not disrupt the integrity of the set.

The visualizations of the columns with outliers, at least before transformation, are as follows :







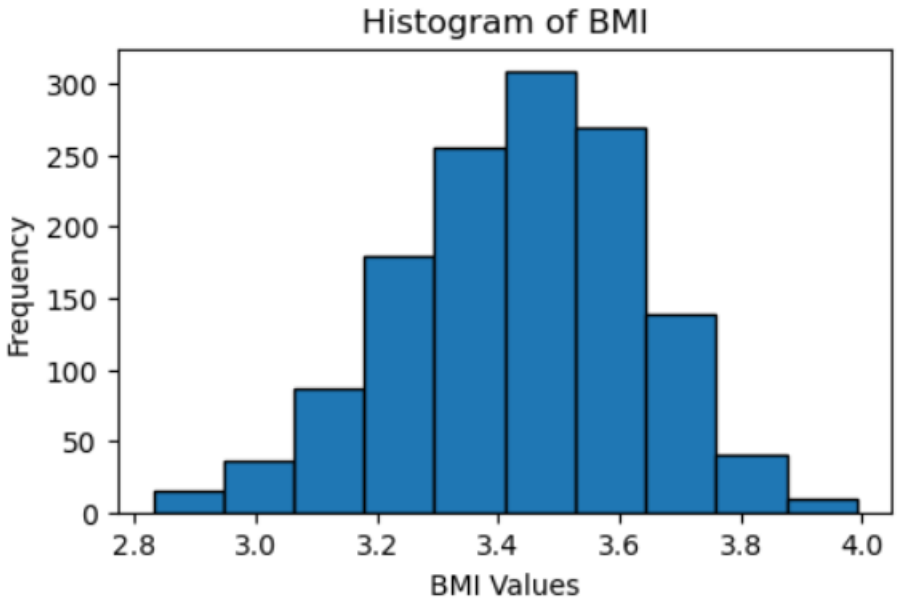
The ‘bmi’ distribution was found to have a left-skewed Gausian distribution, which we would need to normalize in order to optimize the predictive models training.

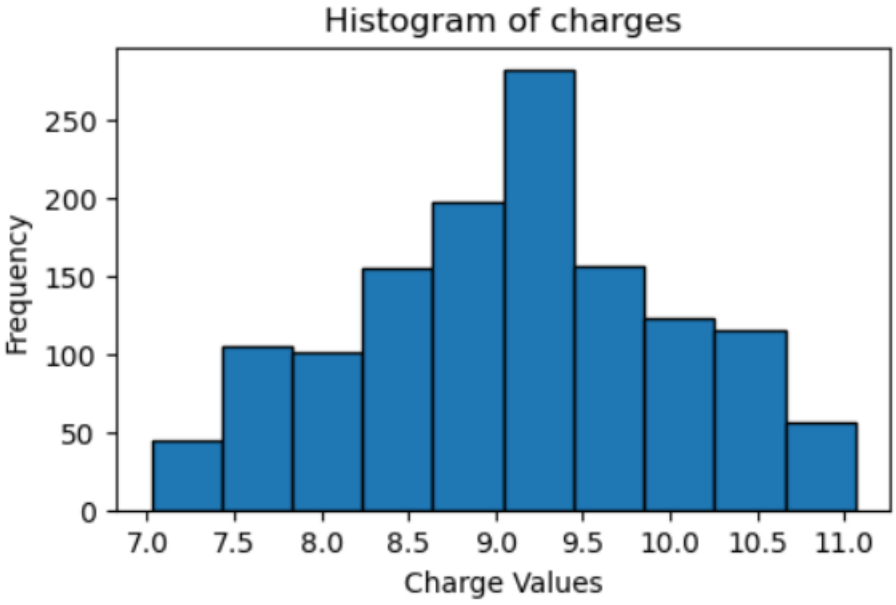
All three of these outlier columns would be transformed to reduce the harmful effects that the extreme outlier values would have on the models ability to generalize data.

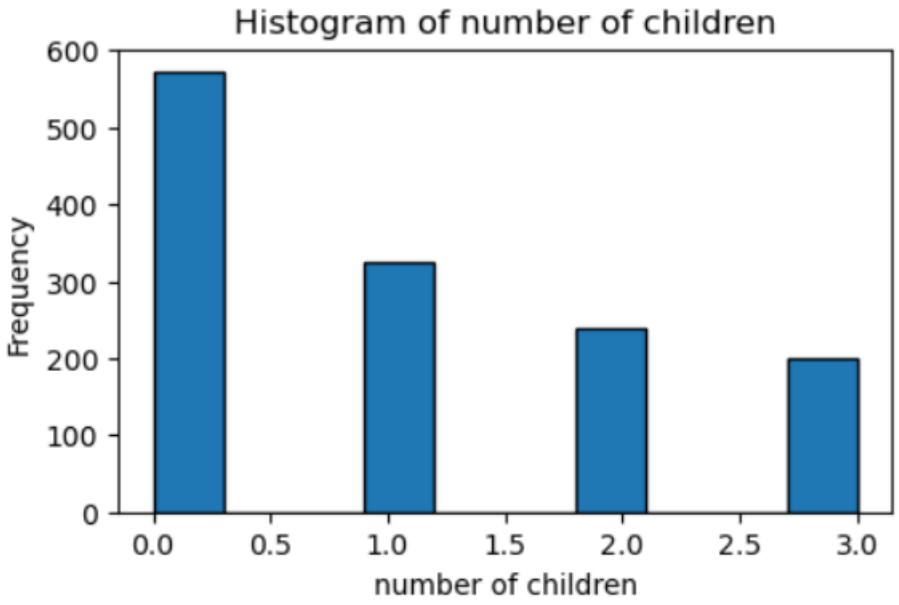
The ‘bmi’ and the ‘charges’ columns were transformed using a natural logorithm function ['numpy.log1p(...)'] in order to normalize the skewed distributions, and to scale the ‘charges’ values in order to make the visualization more legible and other EDA processes easier, without disrupting the ratio of the trends.

The ‘children’ column was transformed using a winsorization method (statisticshowto.com, 2025), involving replacing the values below or above the lower and upper quantiles of the column respectively with the values near those quantiles in the acceptable range(statisticshowto.com, 2025). This was done to reduce the effect of these outliers without shrinking the set, or deleting the corresponding datapoints(statisticshowto.com, 2025).

The results of these transformations were represented as follows:







After these transformations, we checked to see how many outliers remained in the dataset:

```
Number of outliers per column (transformed):
 age         0
sex          0
bmi          1
children     0
smoker       0
charges      0
dtype: int64
```

We also checked the difference in size between the original dataset and the cleaned dataset:

```
Original dataset size: 1337
Transformed dataset size: 1337
```

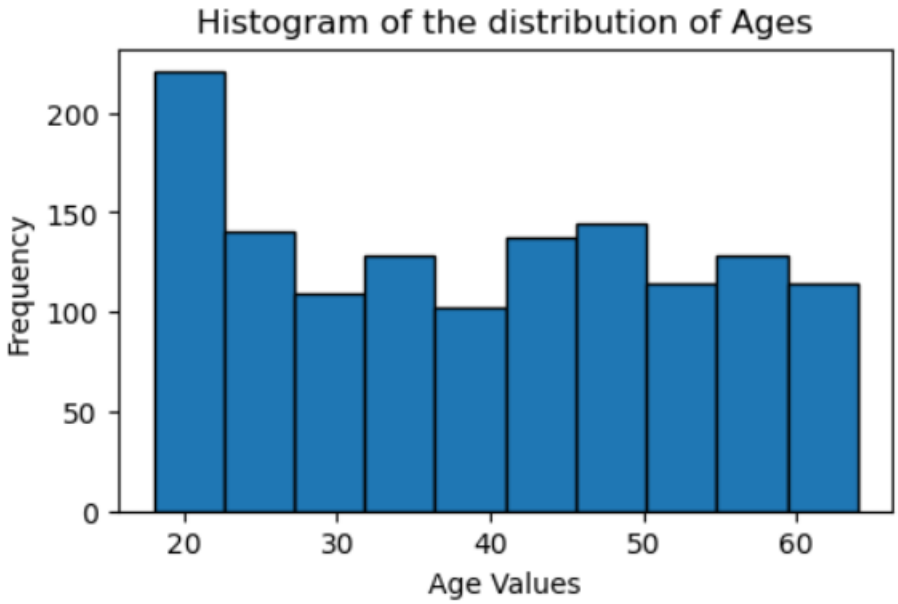
Exploratory Data Analysis:

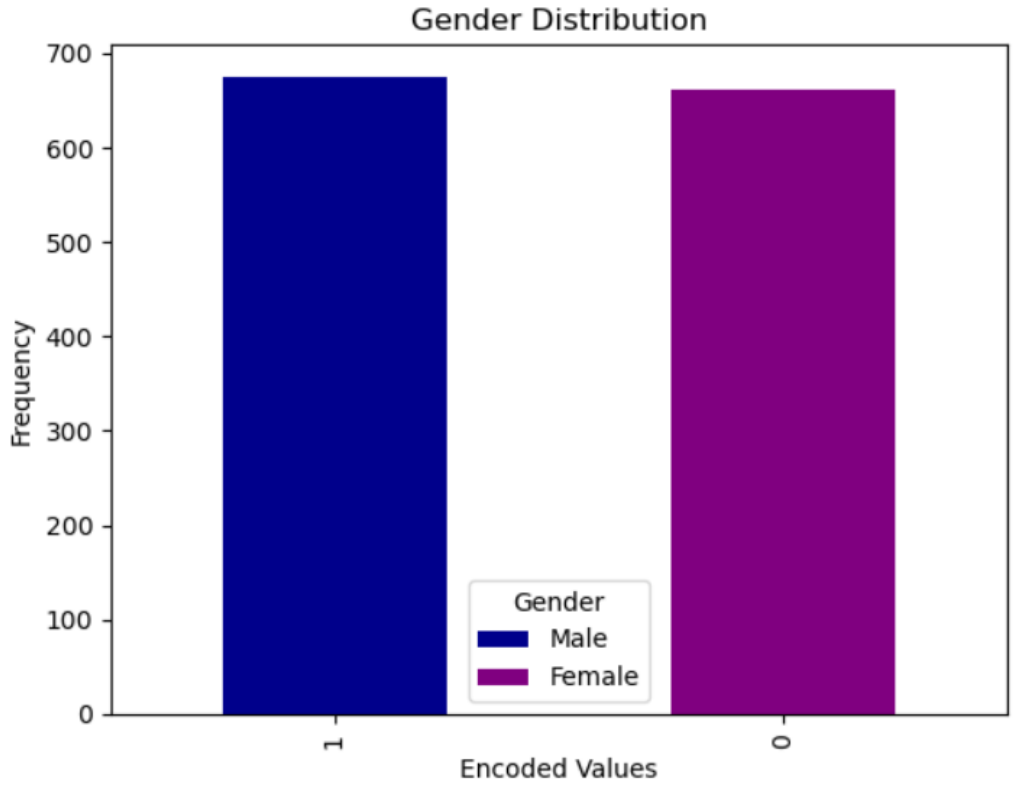
In this stage of the project, we first programmed the software to display the key statistics of every column in the cleaned dataset, the output given was as follows:

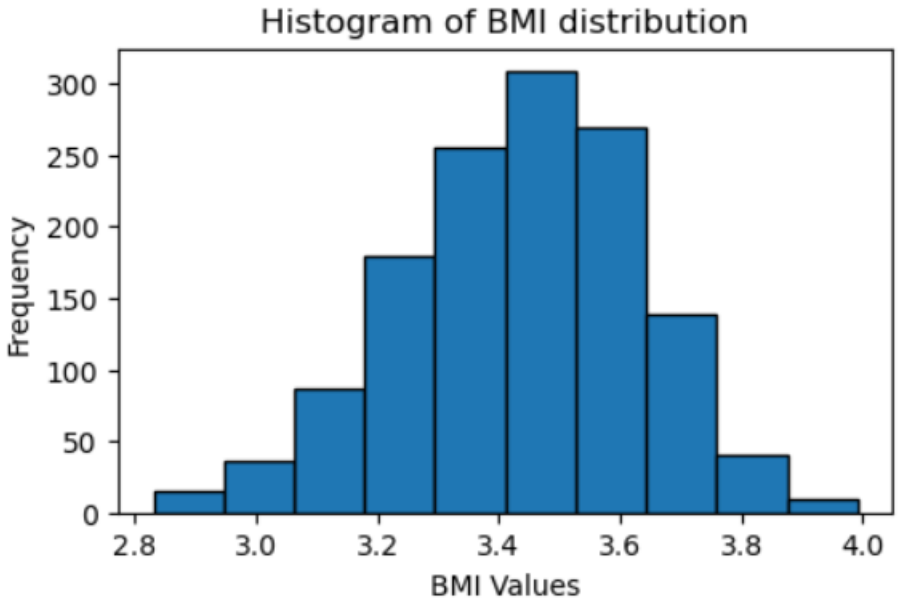
```
Here is a basic description of several key metrics for each column, post data-cleaning.

Description of the insurance dataset:
                age          sex          bmi     children       smoker  \
count  1337.000000  1337.000000  1337.000000  1337.000000  1337.000000
mean     39.222139     0.504862     3.436321     1.050112     0.204936
std      14.044333     0.500163     0.195762     1.097644     0.403806
min      18.000000     0.000000     2.830858     0.000000     0.000000
25%      27.000000     0.000000     3.306520     0.000000     0.000000
50%      39.000000     1.000000     3.446808     1.000000     0.000000
75%      51.000000     1.000000     3.575151     2.000000     0.000000
max      64.000000     1.000000     3.991389     3.000000     1.000000

            region      charges
count  1337.000000  1337.000000
mean      1.516081     9.100097
std       1.105208     0.918551
min       0.000000     7.023647
25%       1.000000     8.465341
50%       2.000000     9.147098
75%       2.000000     9.720689
max       3.000000    11.063061
```

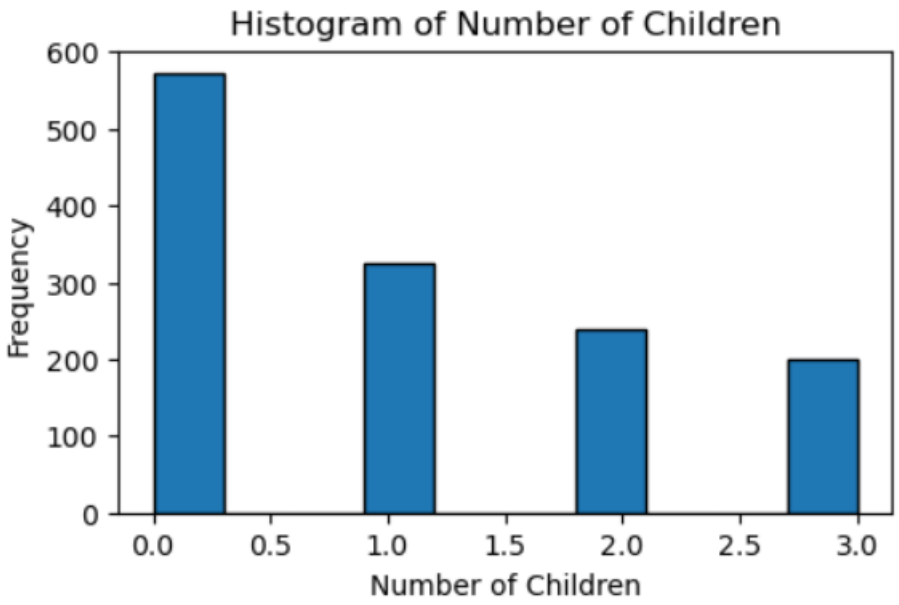
These key statistics from the entire dataset give us a better idea as to how it is distributed, and form useful metrics for the next steps of analysis.

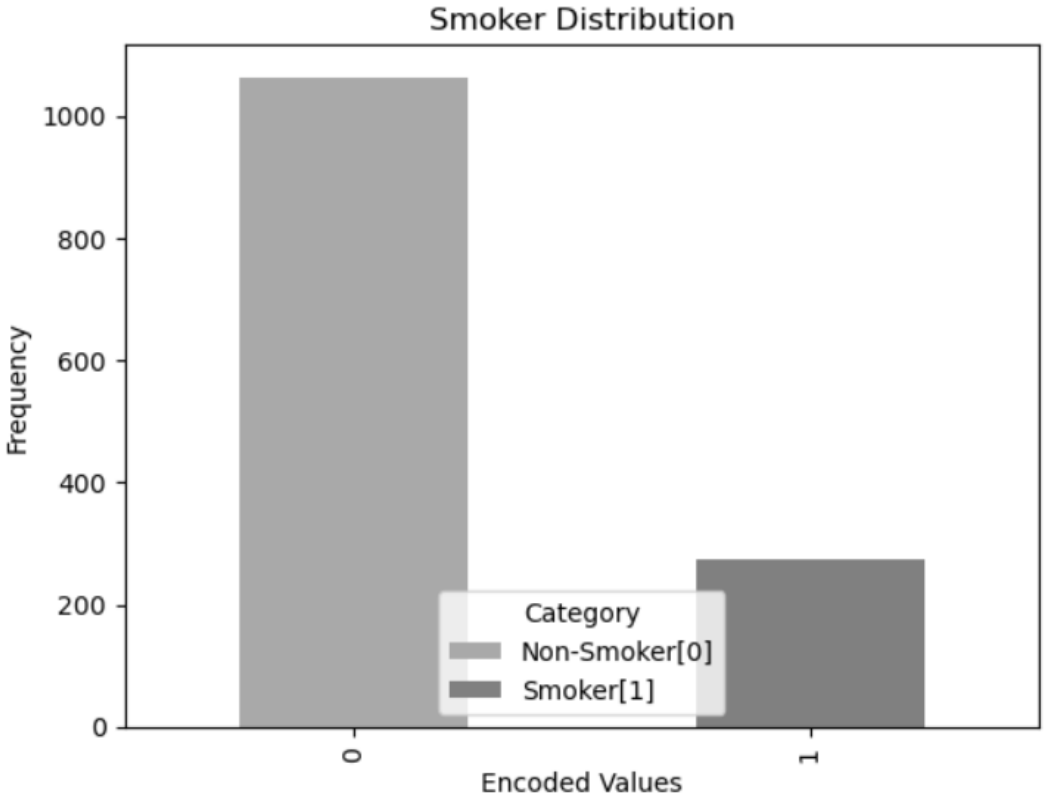
With these key metrics identified, we then programmed the software to output graphical representations of how the data in each column was distributed.

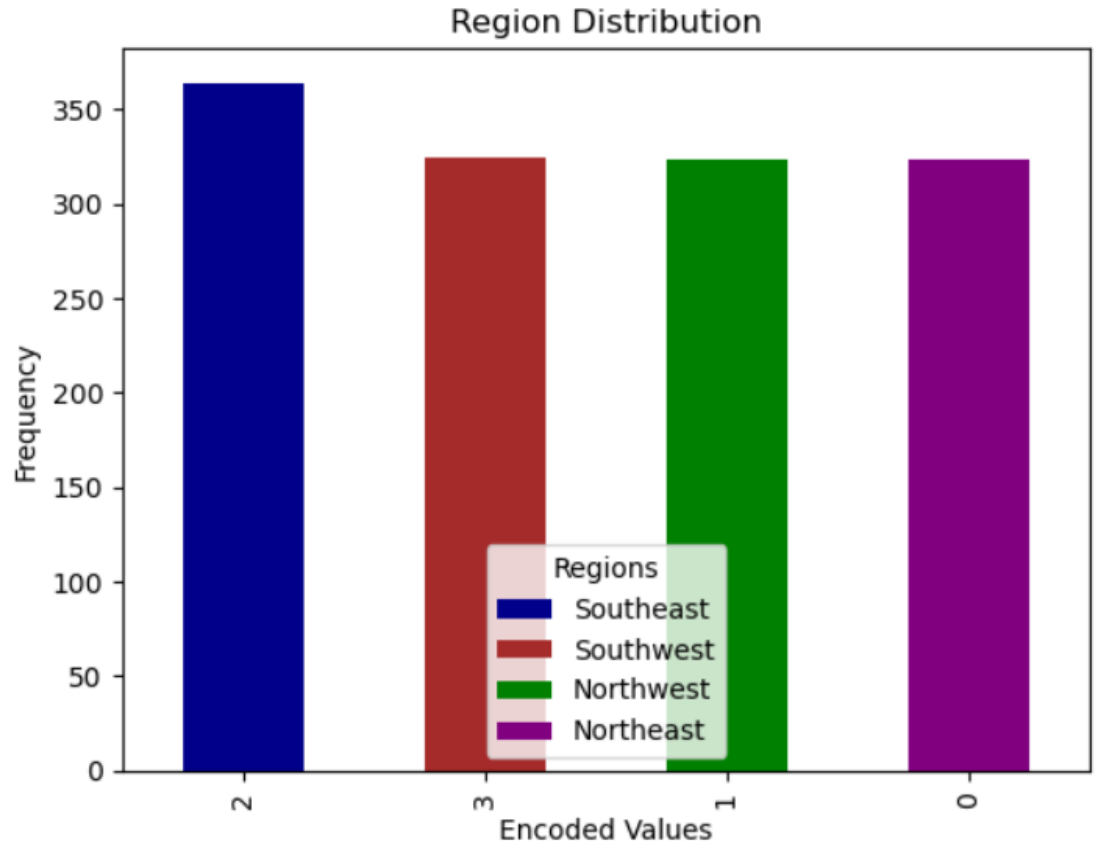


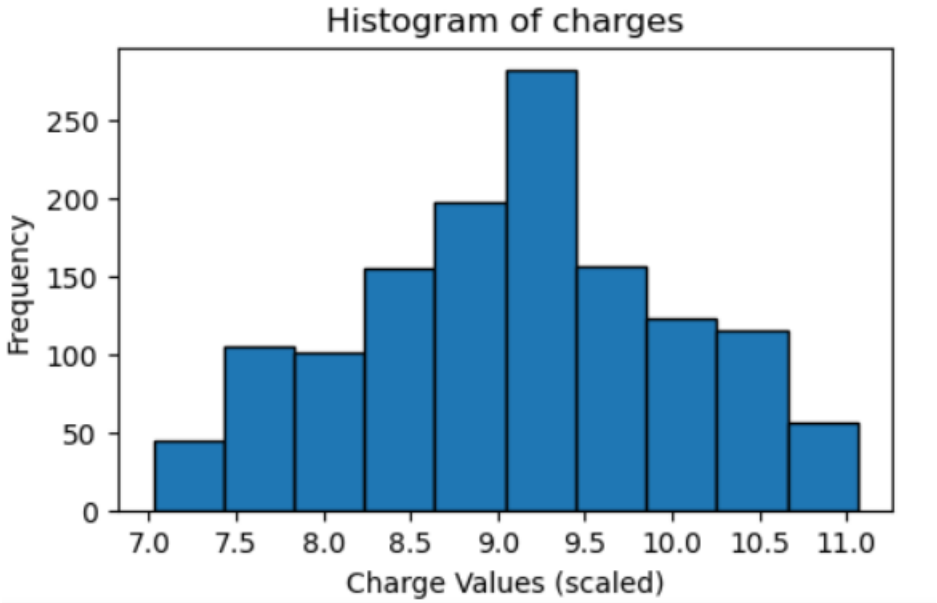

Gender Distribution
700
600
500
400
300
200
100
0
Frequency
Gender
Male
Female
1
0
Encoded Values


Histogram of BMI distribution
300
250
200
150
100
50
0
Frequency
2.8
3.0
3.2
3.4
3.6
3.8
4.0
BMI Values


Histogram of Number of Children
Frequency
Number of Children
0
100
200
300
400
500
600
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Smoker Distribution
Frequency
Encoded Values
Category
Non-Smoker[0]
Smoker[1]
0
200
400
600
800
1000
0
1

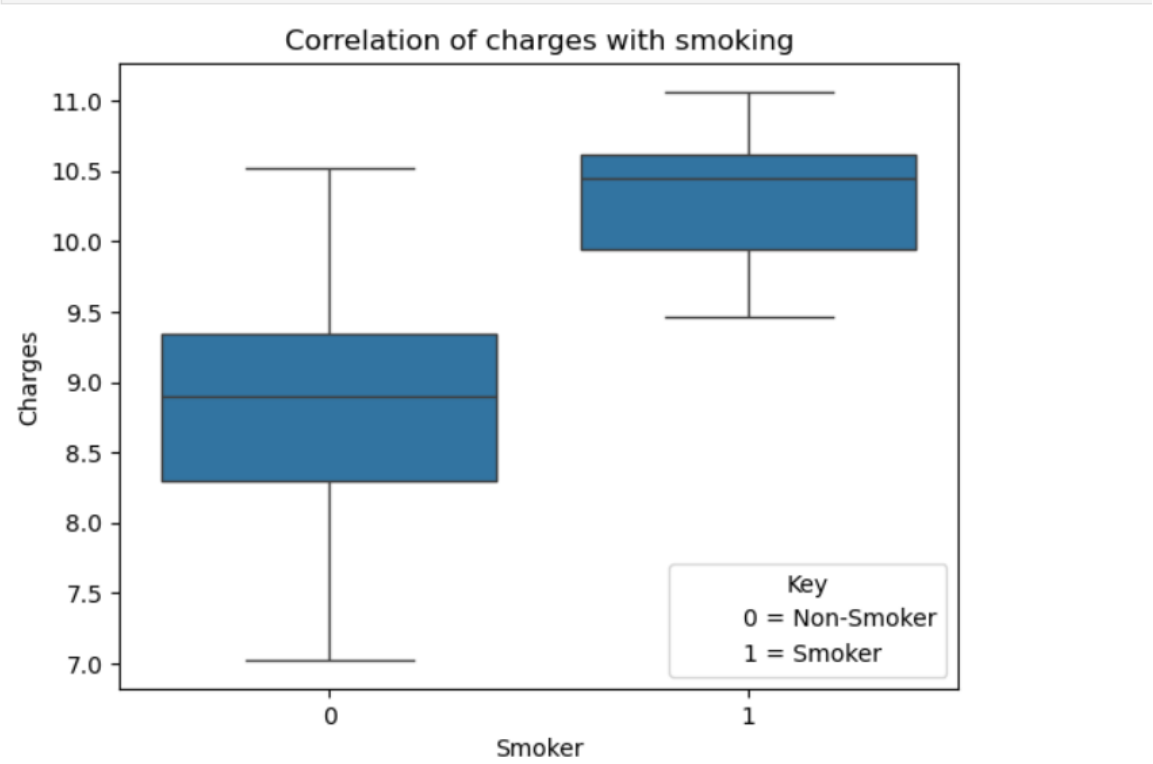




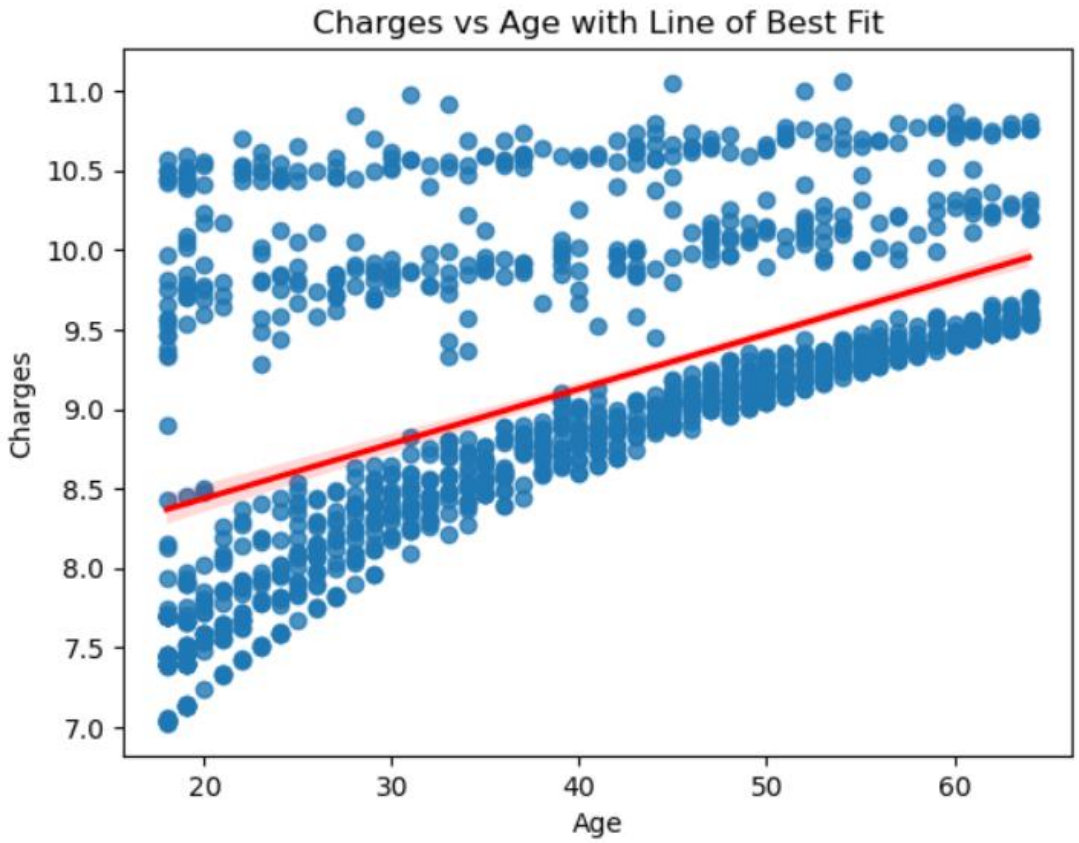
These charts solely represent the frequency of certain values within their respective distributions, sand give us further insights as to how these values are distributed.

The next step after initial visualization is to determine how each independent variable was related to the dependent variable of ‘charges’.

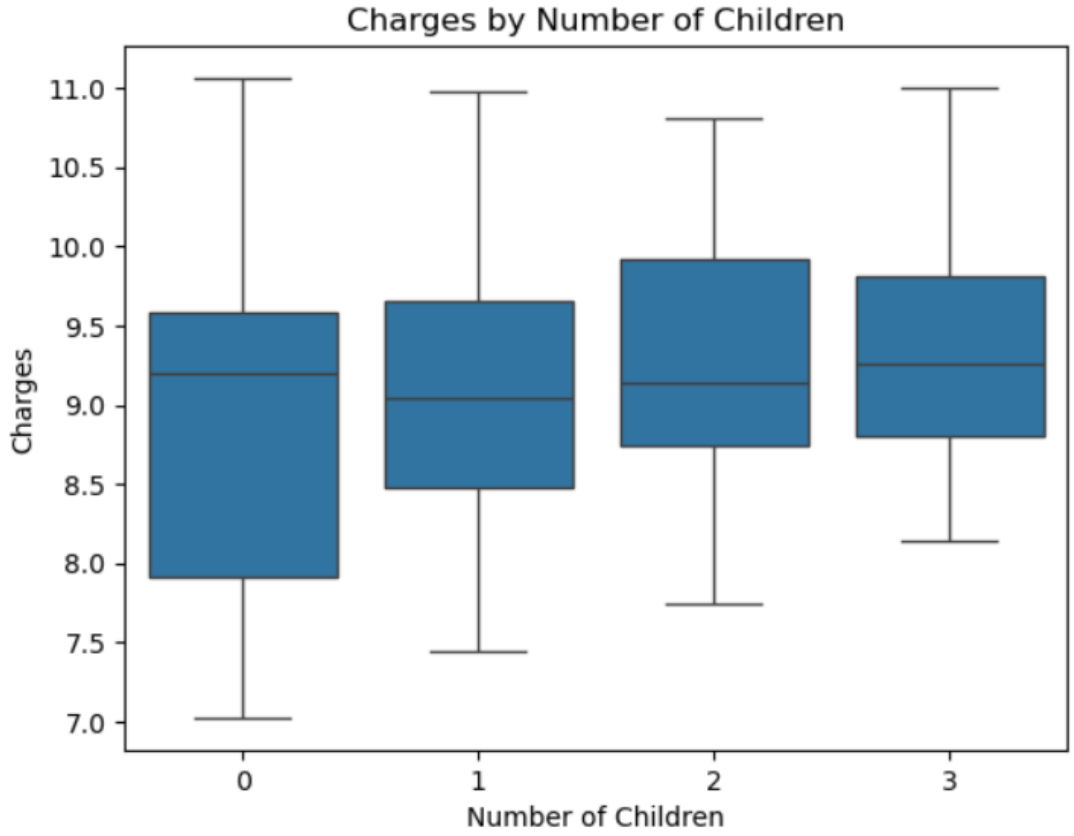
A decent amount of consideration was given in this step, to ensure that the relationships between these different data types were approached in the most suitable way possible. The following charts were developed to represent these relationships:



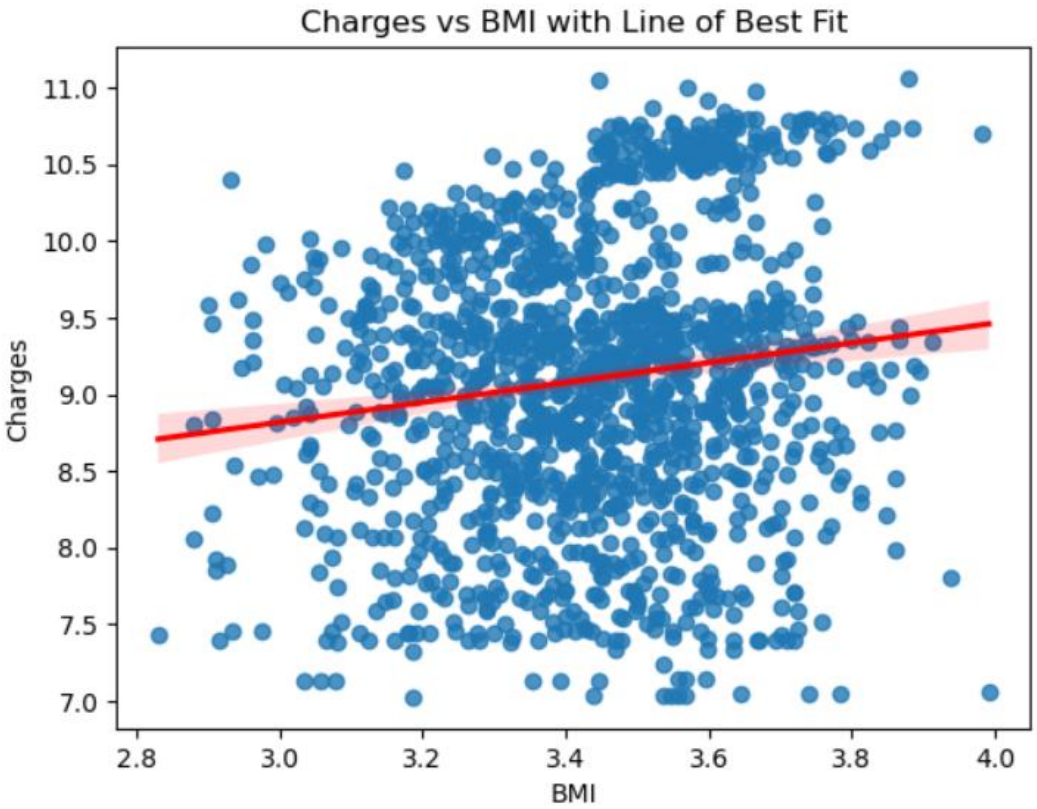
This boxplot shows that there is a great difference between the mean charges assigned to non-smokers and smokers. We can observe that the minimum value, lower quartile, median, upper quartile, and maximum values for charges given to the smokers in the dataset is much greater than that of non-smokers in the dataset. This describes that, by every metric, smokers are generally being charged more than non-smokers for insurance.



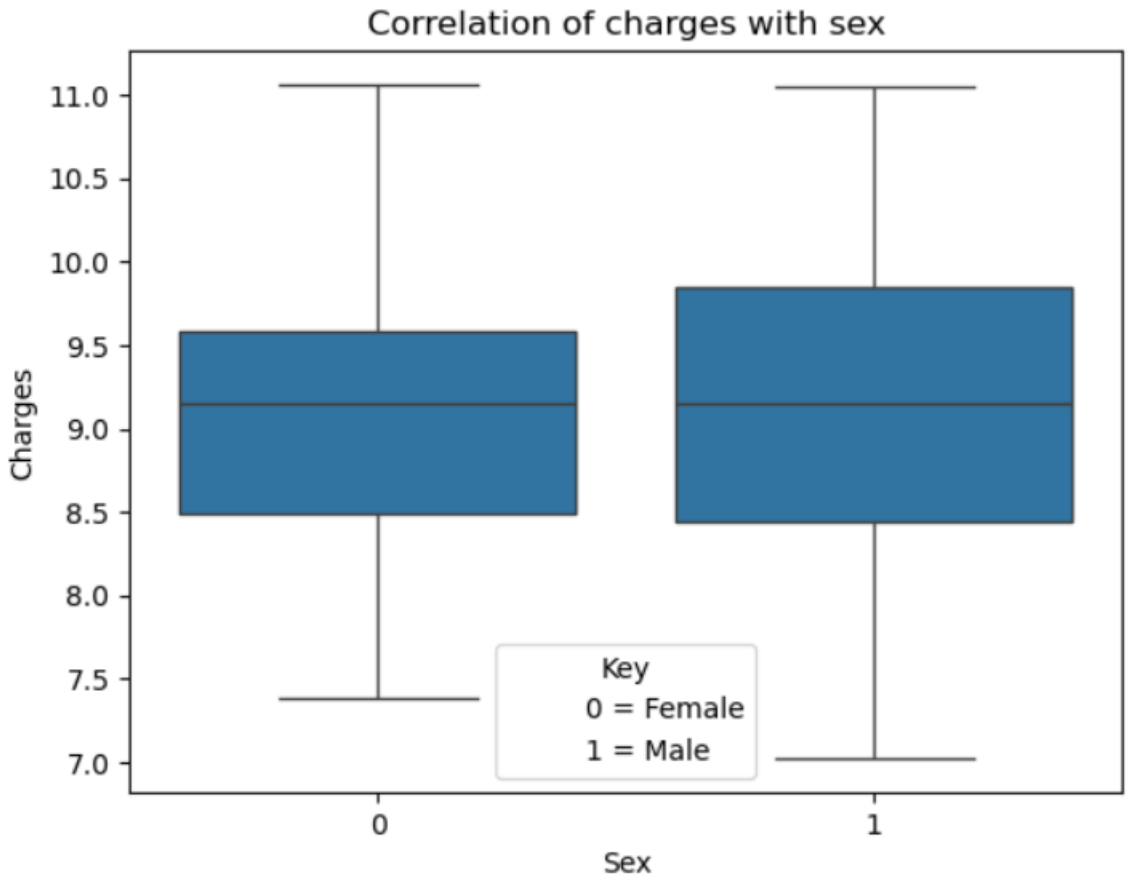
This scatter plot represents the relationship between the ‘charges’ and ‘age’ columns and their values, with the general trend being represented with the regression line. This graph indicates that there is a positive correlation, charges can be observed to be greater as age becomes greater. However the vertical gaps seen between the groups could indicate that this relationship might be affected by other factors, this will be explored further later.



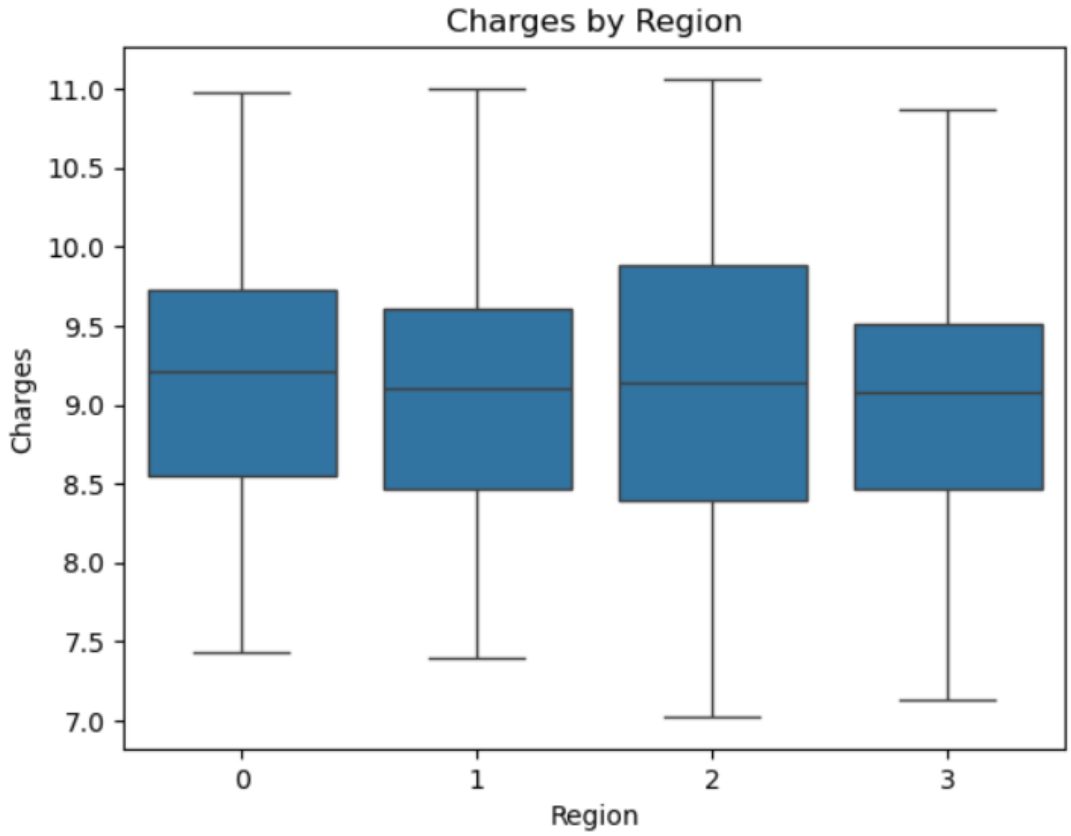
This plot, showing the relationship between ‘charges’ and ‘children’, seems to stay similar across each level. Notably the minimum value charged increasing as the number of children increases, which does make sense, since the families with more children would have to pay for more coverage.



This graphs shows that, while spread out, there is still an observable linear relationship between ‘charges’ and ‘bmi’, with the regression line showing a somewhat weak but still positive correlation.



In this graph we can see some subtle differences between the charges given to men and women. The minimum value of charges applied to women is higher than men, which means that the lowest charged woman is still paying more that the lowest charged man. Men exhibit a higher Q3 indicating that 25% of men paid more than most women did, while the slighlty lower Q1 indicates that the lower 25% of men were charged slightly less than women. Both had approximately the same median value.



The graph representing the relationship between ‘charges’ and ‘region’ shows some differences between the how much was paid between different regions, but the relationship is still a bit ambiguous.

Before going further, we need to ensure that multicolinearity between features is minimal, as notable inter correlations can harm the predictive accuracy of our model(Singh, 2024). To test this I calculated the Variance Inflation Factor (VIF)(Singh, 2024) scores for each of the features. The results were as follows.

```
VIF score for Age:
  Feature       VIF
1     age  1.017916

VIF scores for the other features:
  Feature       VIF
2     bmi  1.042808

    Feature       VIF
3  children  1.003557

  Feature       VIF
4     sex  1.008813

  Feature       VIF
5  smoker  1.006937

  Feature       VIF
6  region  1.026288

The VIF score of '1.017916' for age indicates that, while there is some multicollinearity for this feature, the extent of which is very low, and thus we
can still use it for the linear regression models without further intervention.
```

In interpreting VIF scores, we can say the following:

If 'VIF = 1' then there is absolutely no multicollinearity between the feature and the other columns(Singh, 2024). Whereas If VIF was between 1 and 5 then the multicollinearily of

the feature is low to moderate, which is acceptable for linear regression. However if 'VIF'm is greater than 5 then the multicollinearity is high, meaning remedial action must be taken for those features before applying the rergression model (Singh, 2024).

What this test shows us is that there is major multicollinearity found between the features, and thus the set does not need further intervention in this regard.

The next step was to determine the correlation scores of each of the features in regard to the dependent variable 'charges'.

While all features contained numeric data, not all data was continuous, like 'charges', and thus each feature needed the relevant and appropriate method for comparing its own data and data type to the data of the continuous 'charges' values. The following is the result of that process:

```
Correlation scores for each of the independent variables:

Smoker - Point-Biserial Correlation: r = 0.666, p = 0.000
[A strong r-value of '0.666' and a p-value < 0.05 suggest a significant, strong, positive correlation. Being a smoker increases the likelihood of higher
charges.]

Age - Spearman Correlation: r = 0.534
[For 'age', the r-value of '0.534' indicates a moderate, positive correlation which suggests that charges tend to increase with age.]

Children - Spearman Correlation: r = 0.134
[For 'children', the weak positive r-value of '0.134' suggests a minimal correlation with charges. This implies number of children has little...
..impact on charges.]

BMI - Pearson Correlation: r = 0.138
[For 'bmi', a low r-value of '0.138' suggests a very weak correlation with charges, indicating it likely has minimal predictive power.]

Sex - Point-Biserial Correlation: r = 0.007, p = 0.798
[An extremely low r-value and a high p-value indicates no significant correlation between sex and charges.]

Region - ANOVA: F = 1.369, p = 0.251
[For region, the p-value > 0.05, which means the f-value is not significant and the variations of charges in relation to region are likely...
... caused by chance, while...
...the F-value is 1.369, meaning the 'region' column does not have a stong impact on or correlation with the 'charges' column anyway.]
Region - ANOVA: F = 1.369, p = 0.251
```

After investigating every column and its types, the correlation methods as shown above were applied.

For assessing the correlation between binary values and continuous values, Point-Biserial Correlation was used(statisticshowto.com, 2016).

For the correlation between discrete and continuous values, Spearman correlation was used (statisticshowto.com, 2021).

For determining the correlation between continous values, the Pearson correlation method was used(metwarbio.com, 2024), since the relationship between bmi and charges was also proven to be linear and both columns were examples of normal distributions(metwarbio.com, 2024) .

For the correlation between encoded categorical values and the continous dependent values, ANOVA correlation was used(statisticseasily.com, 2024).

The result of this process is that we determined that the columns ‘region’, ‘sex’, and ‘children’ had the lowest correlation the ‘charges’ column out of all the other independent variables. Conversely ‘smoker’, ‘age’, and ‘bmi’ seem to have the highest correlation out of the independent variables.

Model Training:

The cleaned dataset was broken into two subsets. ‘x’ , which contained the independent variables to be used in training the models, and ‘y’, which contained the dependent variable ‘charges’.

Informed by the previous section, the two columns that had the lowest statistical significance and correlation score where droppend form the training data, in order to minimize model complexity to improve model genralization, and to reduce the chances of the model overfitting to statistically insignificant data points, liken the ones from ‘region’, and ‘sex’. And even though the ‘children’ column has a similar correlation value to the columns that were dropped, it is being purposefully included in the training data to help with performance evaluation at a later stage.

After the dataset is broken into the ‘x’ and ‘y’ subsets, the data is further split into training and testing subsets. The test and training data for both ‘y’ and ‘x’ is split so that 80% of the respective set is randomly selected to be a part of the training subset and 20% is randomly selected to be a part of the testing subset.

The training and testing data was then fit to a basic muliple linear regression model, and was also independently fit to a Lasso Linear regression model, a model with improved data regularization, in order to test the key performance metrics between them.

Evaluation:

In the evaluation of these model, we will be calculating and comparing the 'coefficient of determination' (R^2), a measure of how well a regression model predicts variance and trends(LibreTexts, 2025), The 'Root Mean Square Error' (RMSE), a measure of of the distance between predicted and actual values in a set, as well as visual depictions of actual and predicted values by both models, in order to determine which model has the highest accuracy in which metrics.

The folllowing shows a calculation and comparrison of thes metrics:

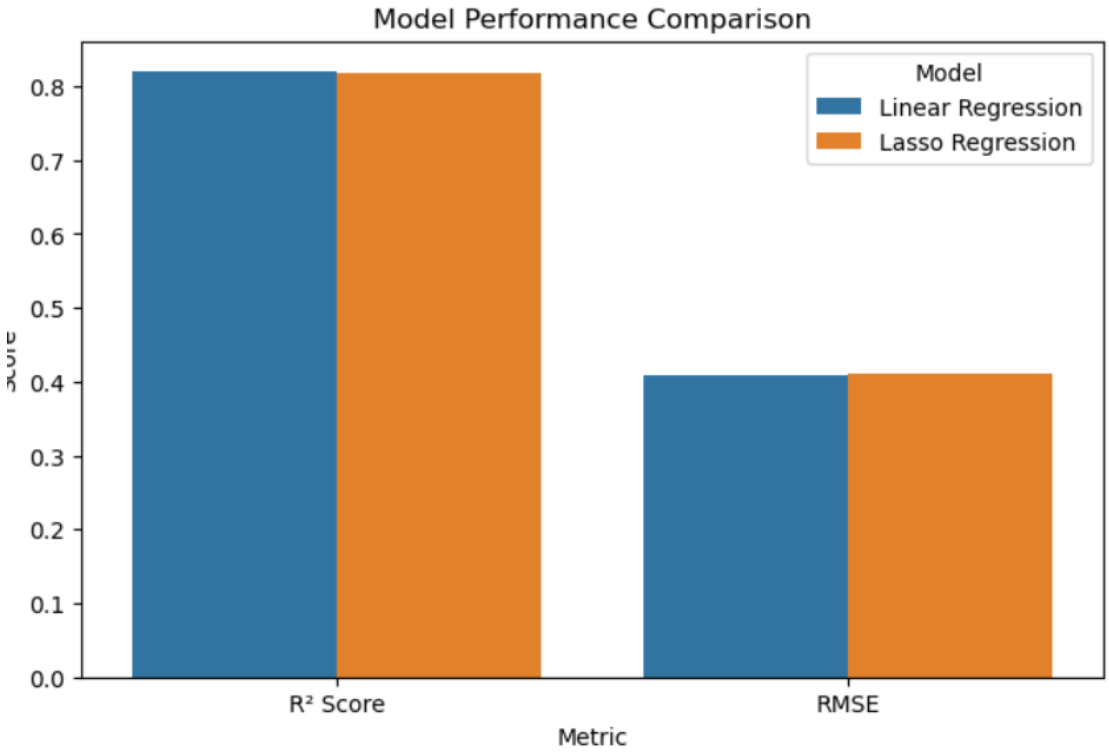
```
---- Linear Regression ----
R^2 Score: 0.8202
RMSE: 0.4084

---- Lasso Regression ----
Optimal Alpha: 0.0066
R^2 Score: 0.8175
RMSE: 0.4115

Linear Coefficients:
age         0.034050
children    0.107936
bmi         0.356455
smoker      1.521101
dtype: float64

Lasso Coefficients:
age         0.034327
children    0.102815
bmi         0.177728
smoker      1.478929
dtype: float64
```

Based on the outputs given here, we can determine that the model with highest R^2 score is the linear regression mode, with a score of 82% indicating a very strong fit to the data, where as the Lasso modle has a R^2 score of 81.75%, a marginally weaker fit than the linear model.

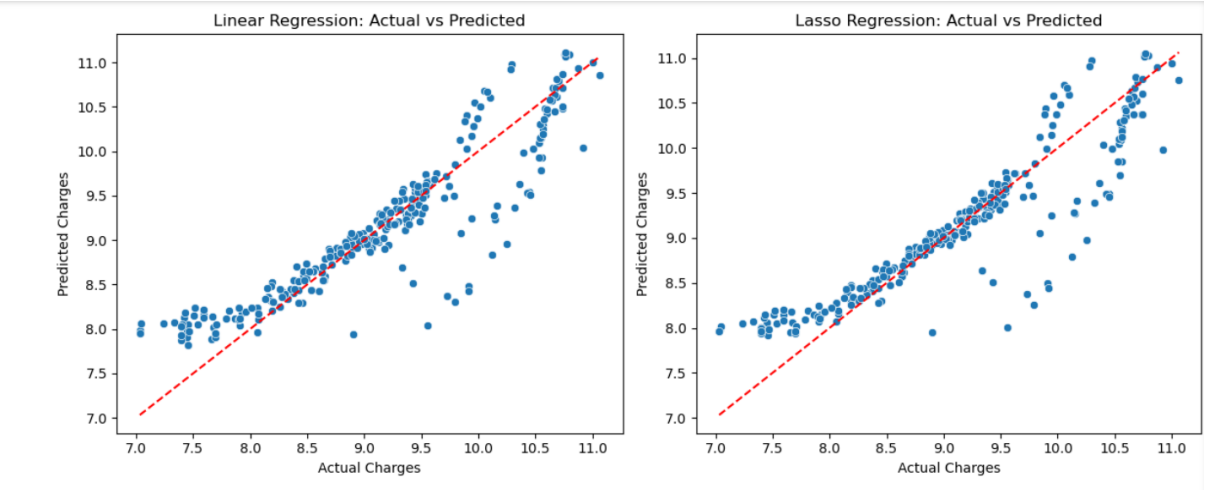
The lowest average distance between predicted and actual values (RMSE) belongs to the Linear model, with an RMSE score of 0.4084. Comparatively the Lasso model has an RMSE score of 0.4115, which indicates a greater distance between actual and predicted values, meaning that the lasso model is comparatively less accurate than the standars linear model with this set.

The value of the coeffiecients for each model indicates how much of an effect each feature has on the prediction values, for example you can see that in the linear model we can see that the ‘smoker’ feature is ‘more impactful’ than the ‘age’ feature. And notably, the impact of certain features differs between models, with ‘smoker’ being marginally less impactful in lasso when compared to its impact in the linear model.

The following is the graphical representation of these metrics that accompanied these figure:



After determining these metrics, we also developed a scatter plot representation to depict the trends between each models predicted and actual values:



Evaluation conclusion:

After evaluating the performance of these models, we can determine that, while the difference is close, the standard Linear regression is the most accurate model for this data set, as It has an 82% accuracy in predictions and an average distance of 0.4084 between predicted and actual values, which is the smallest distance of the two.

References

Choi, M., 2018. *Medical Cost Personal Datasets*. kaggle.com.

Codecademy Team, 2025. *Exploratory Data Analysis with Data Visualization*. [online] codecademy.com. Available at: <https://www.codecademy.com/article/eda-data-visualization> [Accessed 25 April 2025].

Grant, S.W., Hickey, G.L. and Head, S.J., 2018. Statistical primer: multivariable regression considerations and pitfalls. *European Journal of Cardio-Thoracic Surgery*, [online] 55(2), pp.179–185. Available at: <https://academic.oup.com/ejcts/article/55/2/179/5265263> [Accessed 25 April 2025].

Hannay, K., 2025. *Chapter 12 Regression with Categorical Variables*. [online] Introduction to Statistics and Data Science. Available at: <https://faculty.nps.edu/rbassett/_book/regression-with-categorical-variables.html> [Accessed 25 April 2025].

LibreTexts, 2025. *10.2: Validating Your Model*. [online] eng.libretexts.org. Available at: <https://eng.libretexts.org/Bookshelves/Data_Science/Principles_of_Data_Science_(OpenStax)/10%3A_Reporting_Results/10.02%3A_Validating_Your_Model> [Accessed 25 April 2025].

metwarbio.com, 2024. *Mastering Pearson Correlation: A Step-by-Step Guide to Analyzing Data Relationships*. [online] metwarebio.com. Available at: <https://www.metwarebio.com/pearson-correlation-analysis-step-by-step-guide/#:~:text=Pearson%20correlation%20analysis%20requires%20the%20following%20five%20assumptions,Both%20variables%20should%20follow%20an%20approximately%20normal%20distribution.> [Accessed 25 April 2025].

Singh, V., 2024. *Variance Inflation Factor (VIF): Addressing Multicollinearity in Regression Analysis*. [online] datacamp.com. Available at: <https://www.datacamp.com/tutorial/variance-inflation-factor> [Accessed 25 April 2025].

statisticseasily.com, 2024. *One-Way ANOVA Statistical Guide: Mastering Analysis of Variance*. [online] statisticseasily.com. Available at: <https://statisticseasily.com/one-way-anova-statistical-guide/#:~:text=Essentially%2C%20One-way%20ANOVA%20examines%20the%20influence%20of%20a,research%20domains%20where%20comparing%20multiple%20groups%20is%20essential.> [Accessed 25 April 2025].

statisticshowto.com, 2016. *Point-Biserial Correlation & Biserial Correlation: Definition, Examples*. [online] statisticshowto.com. Available at:

<https://www.statisticshowto.com/point-biserial-correlation/> [Accessed 25 April 2025].

statisticshowto.com, 2021. *Spearman Rank Correlation (Spearman's Rho): Definition and How to Calculate it*. [online] statisticshowto.com. Available at: <https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/spearman-rank-correlation-definition-calculate/> [Accessed 25 April 2025].

statisticshowto.com, 2025. *Winsorize: Definition, Examples in Easy Steps*. [online] statisticshowto.com. Available at: <https://www.statisticshowto.com/winsorize/> [Accessed 25 April 2025].