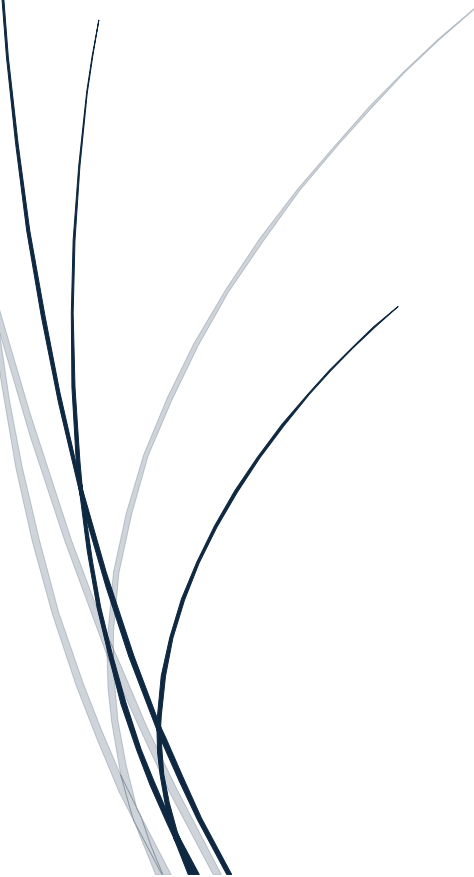


4/25/2025

# PDAN8411

POE Part 1



Sven Kimi Masche  
ST10030798

## Contents

|  |    |
|--|----|
| Exploratory data analysis and linear regression model creation report: ..... | 2  |
| Introduction: .....  | 2  |
| Suitability of dataset: .....  | 2  |
| Planning: .....  | 2  |
| Exploratory data analysis results: .....                                     | 5  |
| Data overview/cleaning: .....  | 5  |
| Correlation analysis: .....  | 6  |
| Univariate analysis: .....   | 9  |
| Bivariate analysis: .....  | 15 |
| Multivariate analysis: .....   | 25 |
| Feature selection: .....   | 27 |
| Model training: .....  | 29 |
| Conclusion: .....  | 32 |
| References .....   | 33 |

# Exploratory data analysis and linear regression model creation report:

## Introduction:

This report covers the steps taken to explore and create a linear regression model based on a medical cost per personal details dataset. The goal is to explore the dataset and uncover any patterns, trends or information about the data that can help with the training of an accurate linear regression model.

## Suitability of dataset:

Before creating a linear regression model it's best to check if the data is actually suitable for linear regression. To conclude this, the following has been checked:

- Linearity between dependent and independent variables. For data to be suitable for a linear regression model, there must be an existing linear relationship between the independent and dependent variables. Using a pairplot it's clear to see most features of the dataset have a linear relationship between each independent variable and the dependent variable(Anon., 2025c).
- No multicollinearity. This occurs when independent variables are too correlated with one another and can cause negative results in a linear regression model. This has been tested for with a correlation matrix where none of the independent variables had a correlation coefficient over 0.8 with one another(Anon., 2025c).

## Planning:

### Exploratory data analysis:

#### Data Overview/cleaning:

Before starting the full exploratory analysis, we need to examine the data. By using panda's library, the plan is to load the csv file and check the various data types used in the data as well as to check for null values. For linear regression and data analysis in general it's best to ensure all values are numerical, for instance where data is non-numerical, we can use label encoding or one-hot encoding to convert the data to numerical values(Anon., 2025a).

#### Correlation analysis:

By using heatmaps and pair plots from the seaborn library, we can see how every feature in the data set interacts with one another. The pair plot will give a visual representation of how the features interact, and the heatmap will give us the numerical number representing

the Pearson co-efficient. This tells us by how much each pair of features affect one another ((Anon., 2025b). The purpose of this correlation analysis is to find the features that make the most impact on one another. For this case, we'll mostly be interested in how each feature interacts with the 'charges' feature, since that's what we are predicting.

### **Univariate analysis:**

We will be looking at each feature individually to try and discern potential characteristics or patterns, such as outliers, skewedness and any other critical information that may help us in the bivariate analysis (Anon., 2024). The techniques to use may differ depending on the nature of the feature:

Categorical data and discrete numerical data:

For these types of data, we will have to make use of count plots to check the distribution of the data for each category.

Numerical data discrete:

Here we can make use of boxplots and histplots to try and uncover more about the data.

### **Bivariate analysis:**

Comparisons between data features will start to be made to see how each feature impacts one another and perform a relationship analysis (Anon., 2024). The graphs that will primarily be used are histplots, box charts, count plots, and regplots. The type of graph to be used as well as the combination of features to be checked against one another is dependent on the findings we gain as we perform the analysis.

### **Multivariate analysis:**

Like bivariate analysis, multiple features will be explored at the same time to explore further patterns based on the findings of the bivariate analysis on the features that seem most promising to the goal of the analysis (Anon., 2024). The main forms of visualization for this analysis are multi-linear regression plots and 3d scatter plots.

### **Feature selection:**

After exploratory analysis we will need to ensure only the most relevant features are used within the model. The form of feature selection to be used will be checking the p values of every feature and their effect on the feature that's desired to be predicted (in this case charges). This will be done using sklearn's standard scaler import to create a stats model that can perform linear regression using all features. The results are then displayed showing the p values of every independent variable. If the p-value is above 0.05 we don't use it (Banerjee Chandradip, 2023).

## **Model creation/evaluation:**

With our selected features all that's left is the creation and evaluation of linear regression models. For this report multiple regression models are going to be created and have their outcomes checked against one another. It's planned to use a standard linear regression model as well as a lasso, ridge and elastic model. After creation of the models their  $R^2$  and RMSE values will be checked against each other as well as graphs visualizing each model's actual vs predicted results as well as predicted results vs error.

## Exploratory data analysis results:

### Data overview/cleaning:

Using the panda's library, we generate a brief overview of the data, how many null values as well as the data types of each feature.

First five rows:

|   | age | sex    | bmi    | children | smoker | region    | charges     |
|---|-----|--------|--------|----------|--------|-----------|-------------|
| 0 | 19  | female | 27.900 | 0        | yes    | southwest | 16884.92400 |
| 1 | 18  | male   | 33.770 | 1        | no     | southeast | 1725.55230  |
| 2 | 28  | male   | 33.000 | 3        | no     | southeast | 4449.46200  |
| 3 | 33  | male   | 22.705 | 0        | no     | northwest | 21984.47061 |
| 4 | 32  | male   | 28.880 | 0        | no     | northwest | 3866.85520  |

Null value count:

```
age      0
sex      0
bmi      0
children 0
smoker   0
region   0
charges  0
dtype: int64
```

Data types:

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

From the first few rows we can see all the features, with age, bmi, and charges being numerical continuous data, sex, smoker and region being categorical, and children being discrete continuous data. We also have the pleasure of seeing no null values. Our first issue is the categorical data in the dataset. For analysis we're going to need to convert this data to numerical values using label encoding.

When we get the first five rows again our data looks like this:

|   | age | sex | bmi    | children | smoker | region | charges     |
|---|-----|-----|--------|----------|--------|--------|-------------|
| 0 | 19  | 0   | 27.900 | 0        | 1      | 3      | 16884.92400 |
| 1 | 18  | 1   | 33.770 | 1        | 0      | 2      | 1725.55230  |
| 2 | 28  | 1   | 33.000 | 3        | 0      | 2      | 4449.46200  |
| 3 | 33  | 1   | 22.705 | 0        | 0      | 1      | 21984.47061 |
| 4 | 32  | 1   | 28.880 | 0        | 0      | 1      | 3866.85520  |

Categorical data is now numerical, and our data overview / cleaning is complete. We can move onto the correlation analysis:

### Correlation analysis:

Before we can individually analyze the features and their effects with one another we should see how every feature is affected by one another. This will give us an idea of what patterns to look out for and where we should focus the most. To check this a heatmap and a pair plot have been generated.

Heatmap:



Within a heatmap each value represents the Pearson coefficient of the effect two features have on one another. Values between 0 and 1 usually show significant affects, while values between  $-1$  and 0 show non-significant effects ((Anon., 2025b). The objective of this analysis is to explore any patterns that may give indication of how we can predict the cost of insurance charges. So, it's best to look at how the features affect the charges variable, here are the values per feature:

age: 0.30

sex: 0.06

bmi: 0.20

children: 0.7

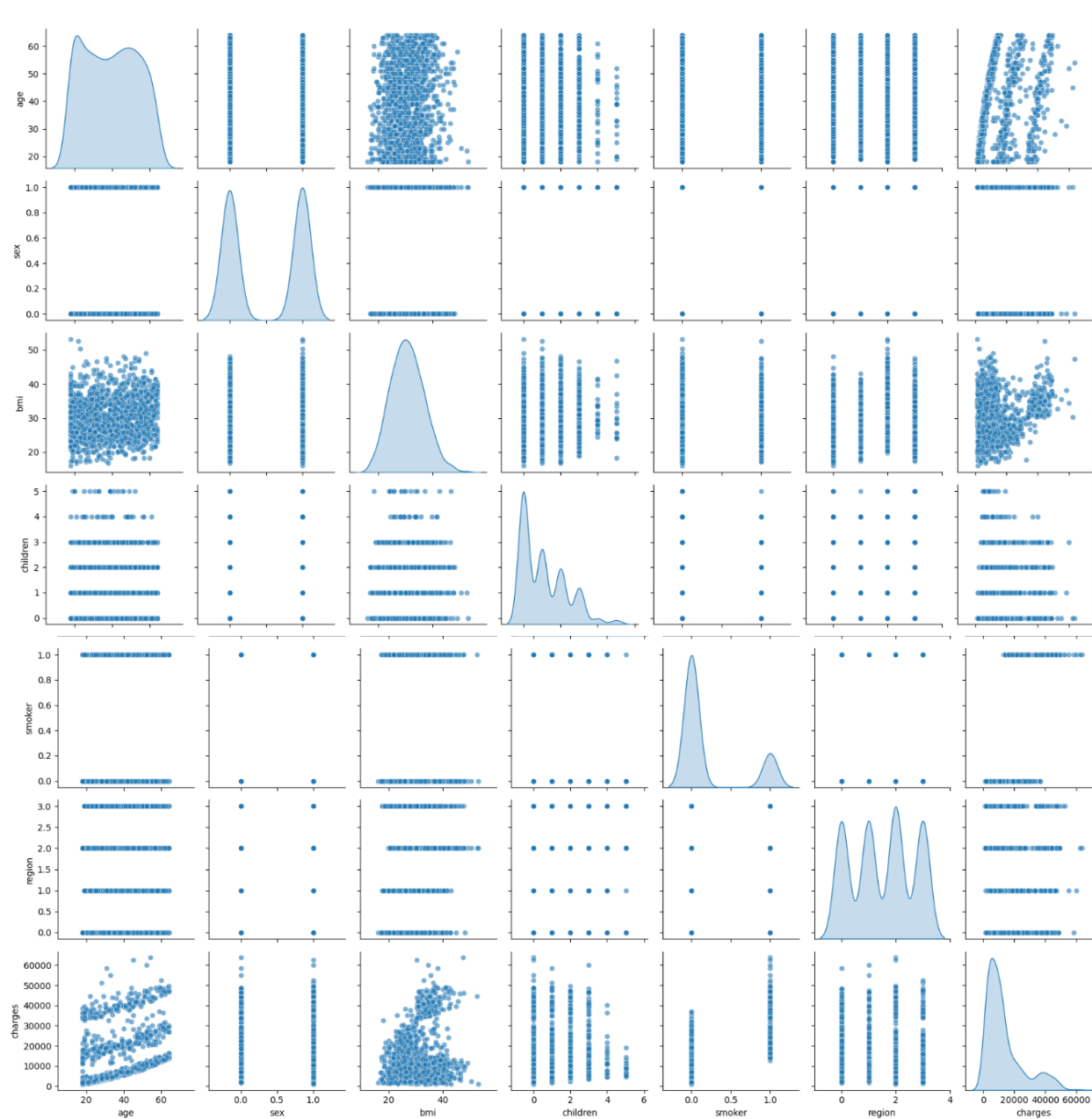
smoker: 0.79

region: -0.01

We can see that smoker, age, bmi, sex and children are the features that are most likely to influence the charge's variable. With smoker, bmi and age being the most indicative.



Pair plot:



The full pair plot only further provides as a visualizer for how the features interact with one another and show any visual correlations. Each of the distributions are to be considered in the next section, univariate analysis. What we can conclude this analysis is that we should look at smoker, bmi and age the most and how they affect one another.

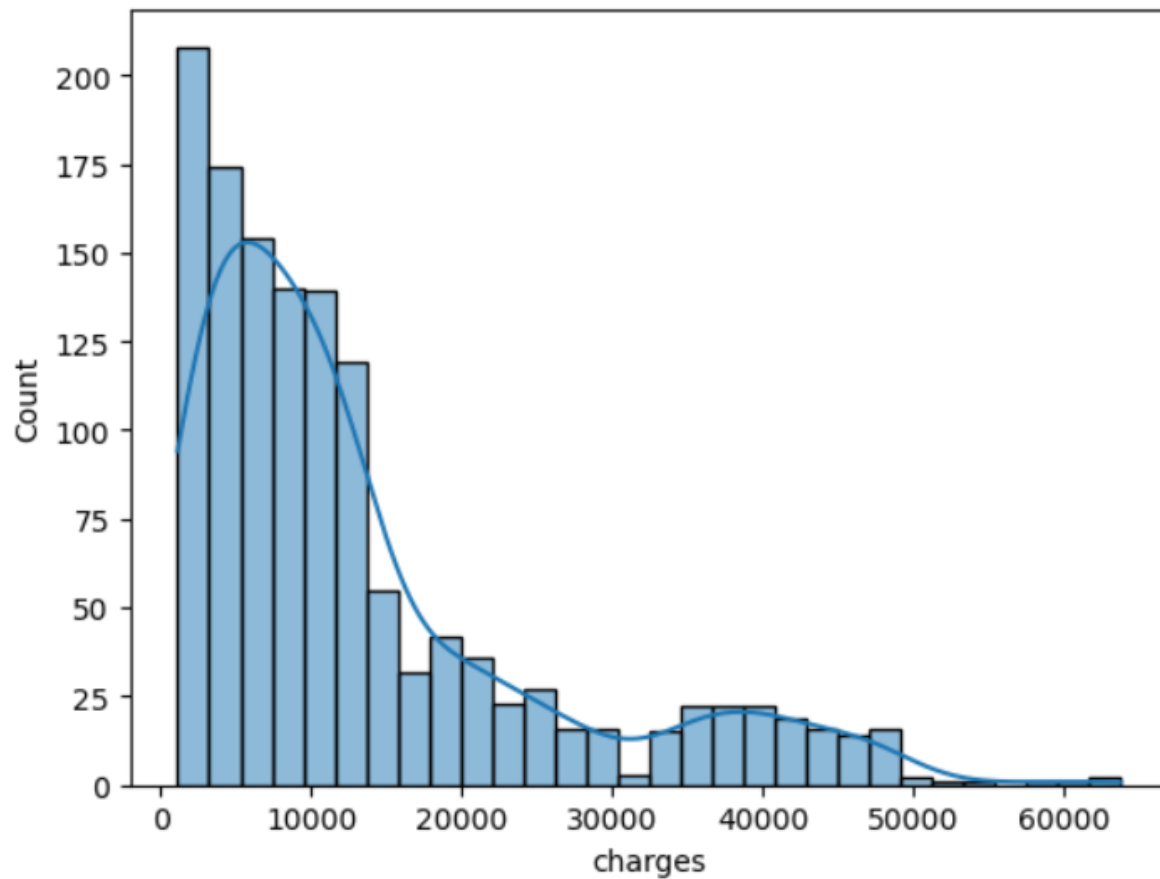
## Univariate analysis:

Here we look at each individual feature, starting with...

### Charges:

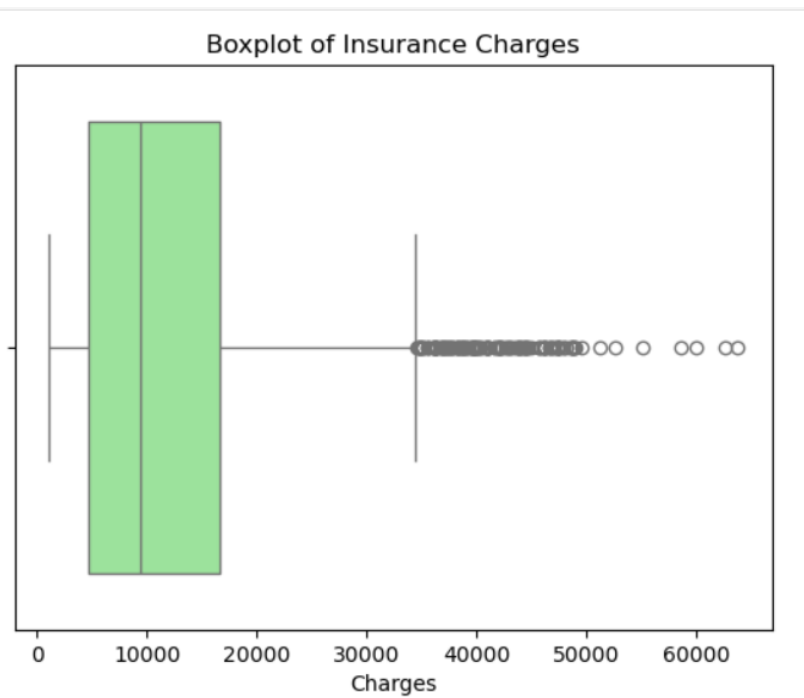
Since this data is continuous, we can use a histplot and a boxplot to analyze.

Histplot:



This hisplot visualizes how the charges are distributed. More specifically we can see that many charges fall under 10000 in costs meaning the data is skewed to the right. We still need to check for outliers of course.

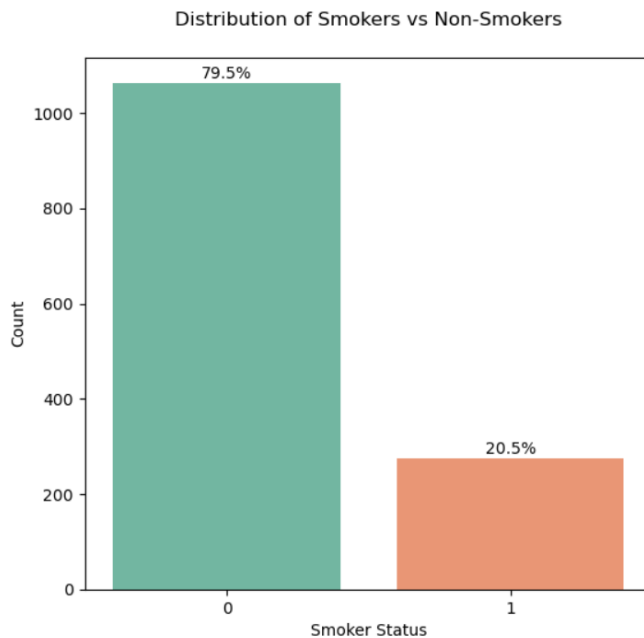
Box plot:



This box plot further justifies the observations in the histplot, as well as it shows us that there are a lot of outliers. That falls out of the IQR range, it's not hard to imagine these outliers consist of high-risk individuals.

## Smoker analysis:

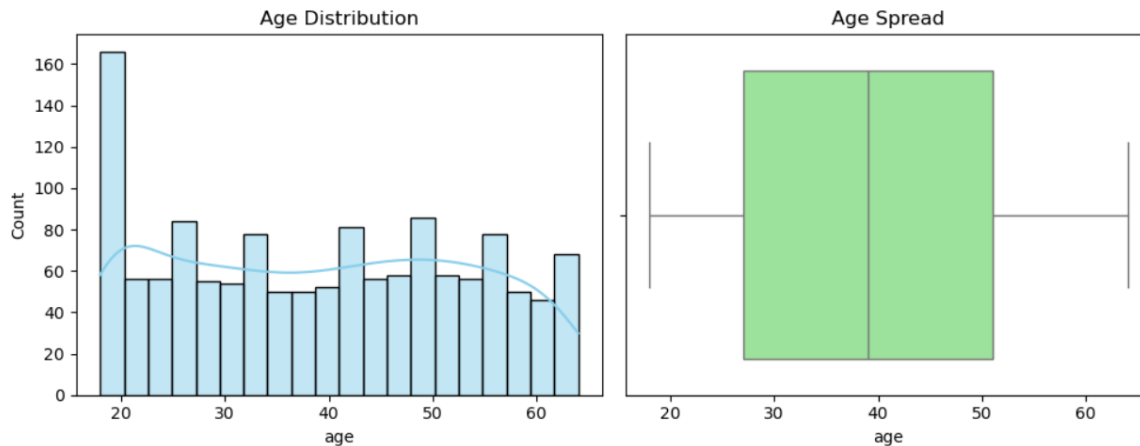
Since smoker data is categorical, it's best to use a count plot:



From this count plot we can see that only a significantly small percentage of individuals are smokers. Since smoking is a health risk it's generally understood that those individuals who smoke are more susceptible to health complications which could be a reason for being charged more for health insurance ((Barendregt, Bonneux and van der Maas, 1997). It's necessary to check this against charges in the bivariate analysis.

## Age analysis:

Since age is continuous, we can use a histplot and box plot again

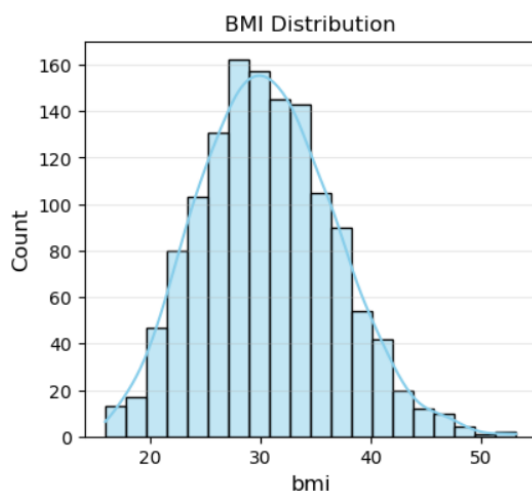


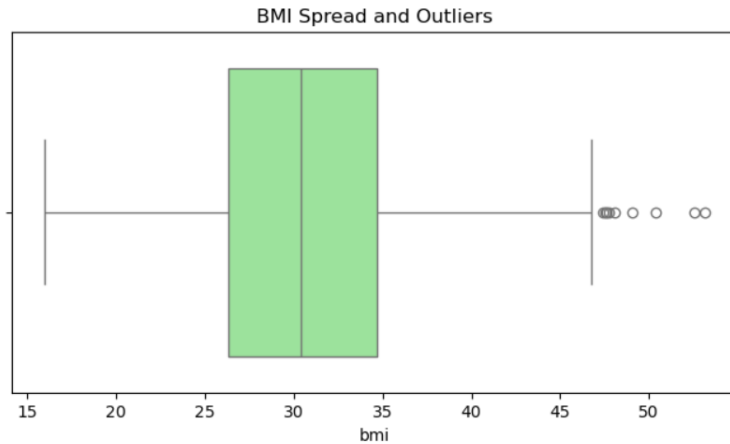
From the histplot we can see that there are more individuals below the age of 20, however the spread is much more even beyond that age until around 50. The histplot almost peaks again showing that the two majority ages within the data are 20 and 50. In the boxplot that spread is shown as being between 30 and 50. No visible outliers can be seen.

Older individuals are more susceptible to health risks which in turn cause insurance premiums to be higher, so we do have reason to check age against charges in the bivariate analysis(Anon., 2023).

## BMI analysis:

Again, with continuous data we use a hisplot and a box plot





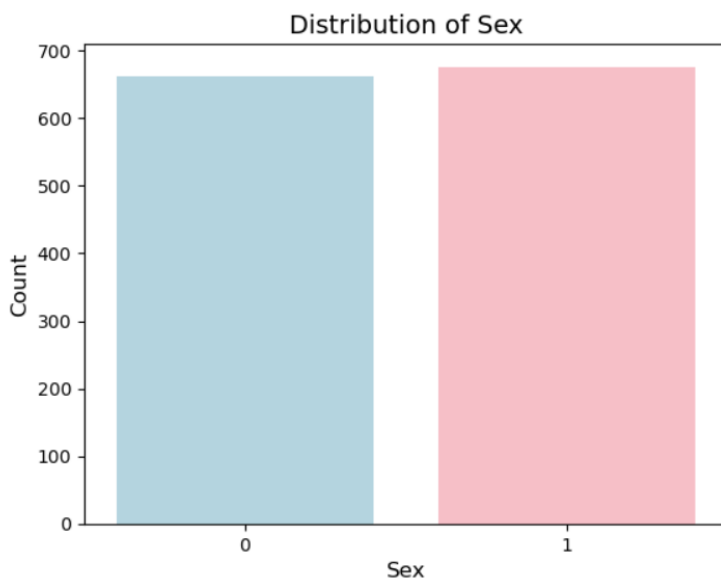
We can see from the histplot that bmi averages out around thirty, and is very slightly skewed to the right, meaning that there are more people with a bmi above 30 than below it.

The boxplot also reciprocates this and shows the existence of outliers beyond bmi of 50. Increased BMI is associated with more health risks, it's very possible that insurance costs for individuals with high BMI may be charged more (Hansen Edwards et al., 2024). BMI will be an important feature to check against in the bivariate analysis.

### Sex analysis:

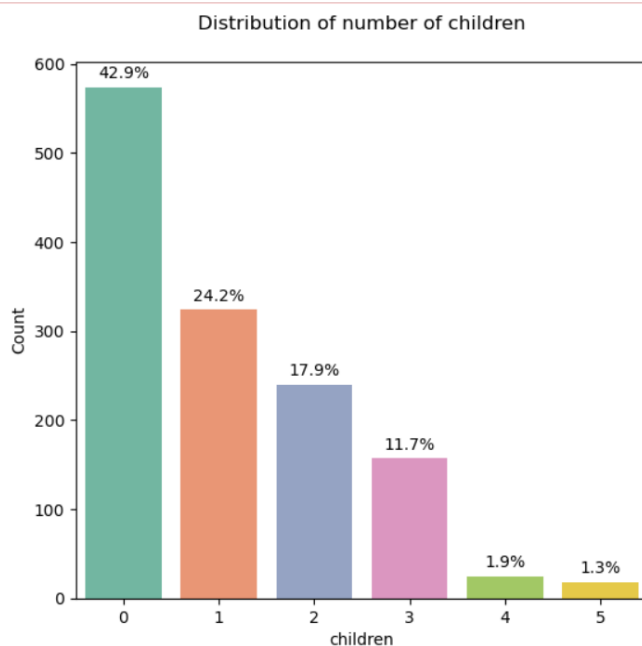
As with categorical data we use a count plot to visualize the distribution.

Note: 0 refers to men, 1 refers to woman.



This graph shows us that over all the number of men compared to woman is very similar with only a slight increase in the amount of woman.

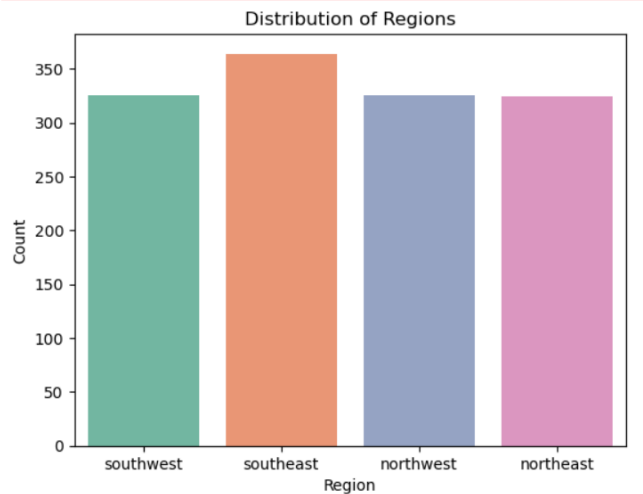
## Children analysis:



From this plot we can see that most individuals don't have children. The data is vastly skewed to the right with almost half of the population not having children. Subsequently each increase in children shows a decrease in count. Since more children equals more people that are liable to health risks it makes sense that having higher amounts of children could correlate to an increase in insurance costs.

## Region Analysis:

Another instance categorical data calls for another count plot:

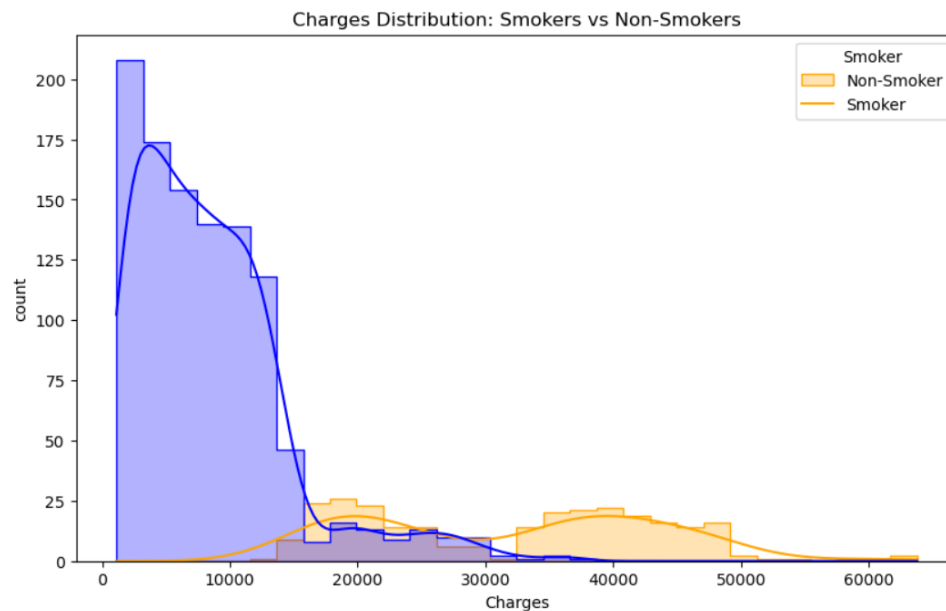


We can see here that each region is decently balanced in terms of distribution.

### Bivariate analysis:

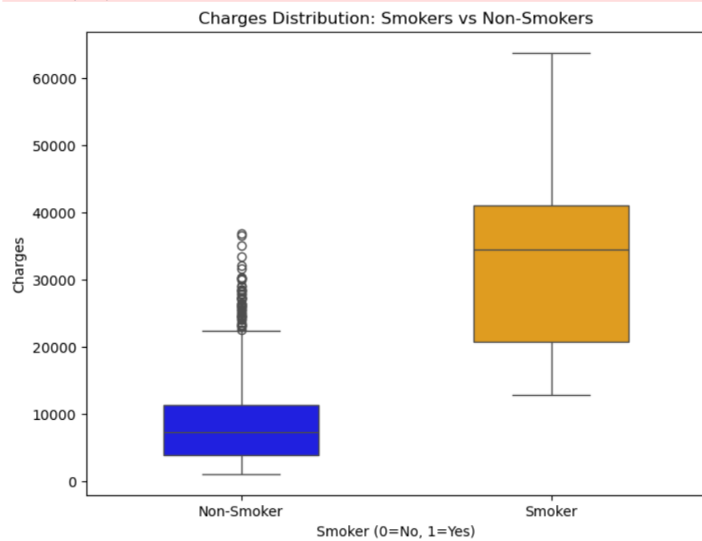
With our univariate analysis concluded, we can now look deeper into the relationship between features based on the findings of the univariate analysis. Key areas to investigate include how charges are affected by bmi, age, smoker and children.

### Smokers and charges:



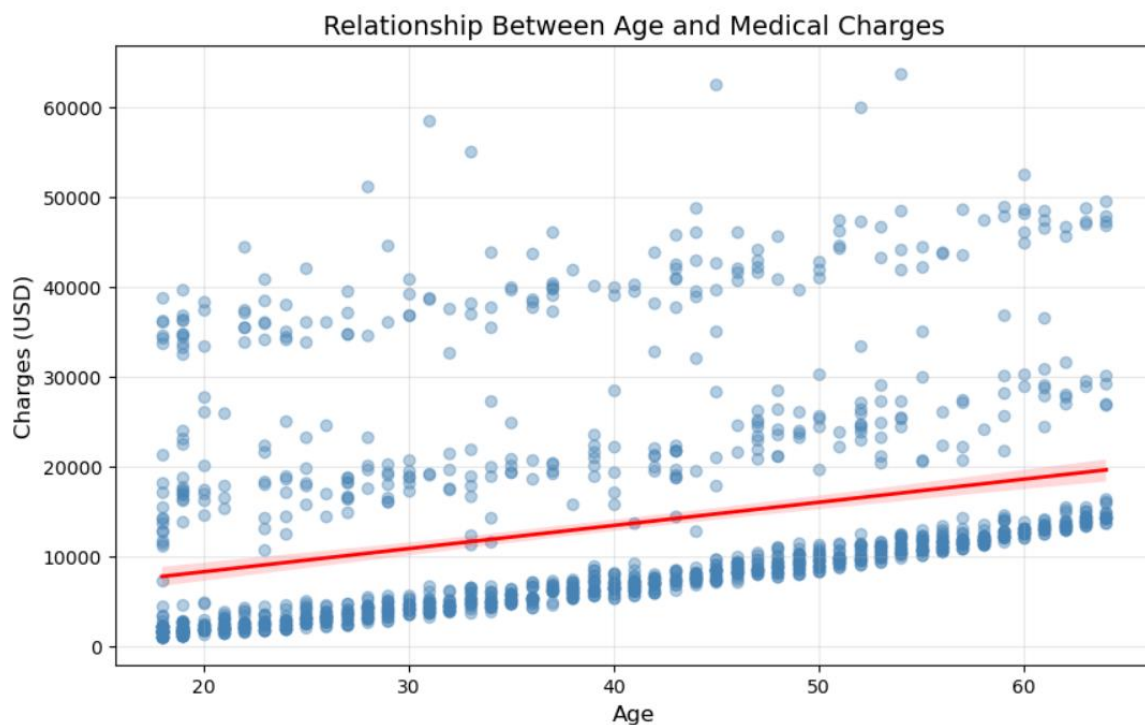
Starting with a histplot we can see that as expected from the univariate analysis, higher insurance charges are associated with smokers.





The box plot further shows this result with a number of outliers for the non-smoking count, indicating that some non-smoking individuals may have other risk factors worth investigating.

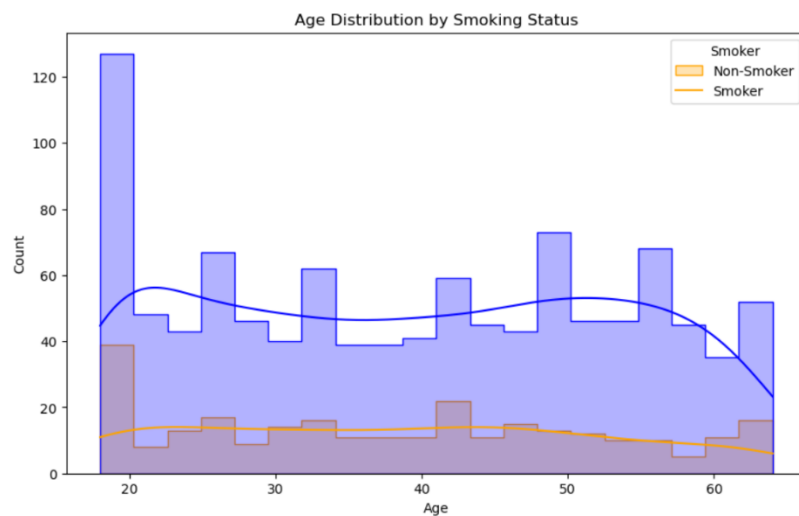
### Age and charges:



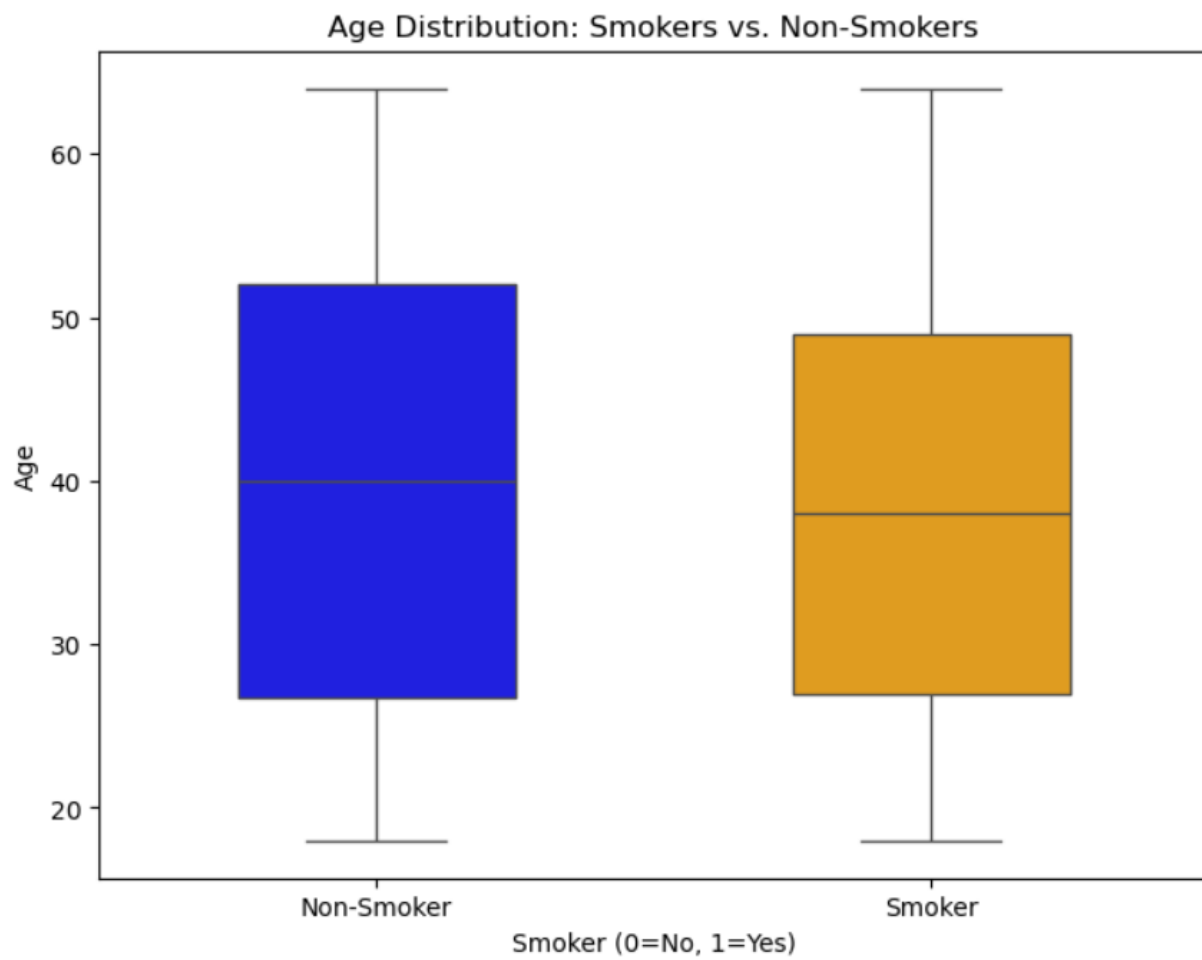
Using this regression plot we can see that there is a linear relationship between age and charges. More specifically we can clearly see that as age increases so do insurance

charges. Since Age is clearly correlated to charges, we can evaluate the relationship between smokers and age.

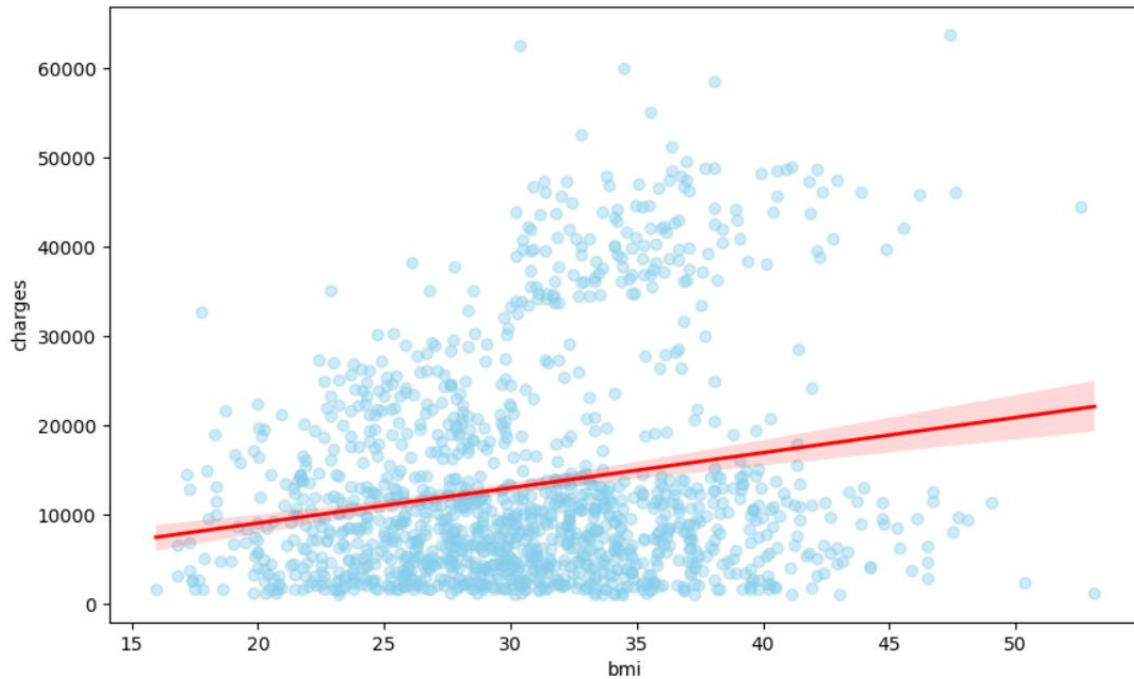
## Smoking and Age:



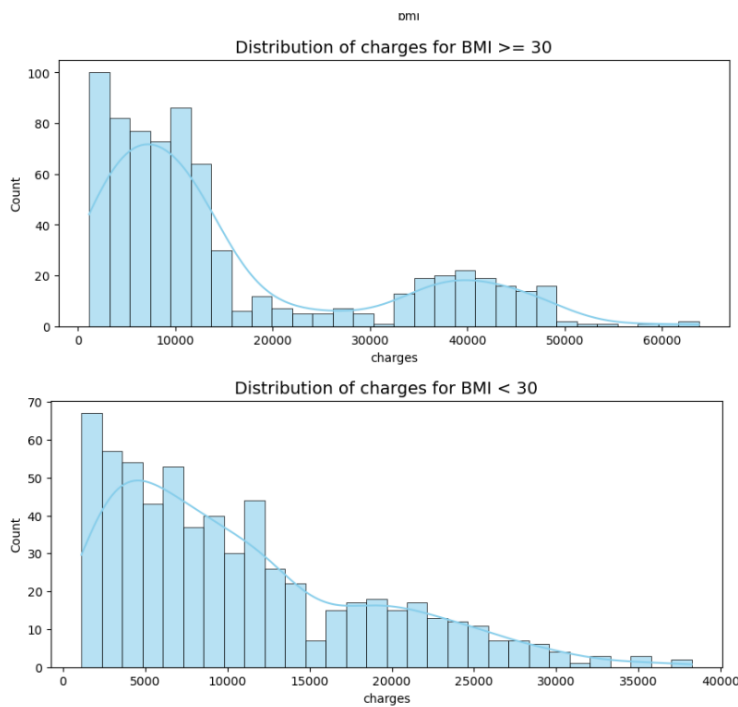
We can see the relationship here that



**BMI and charges:**

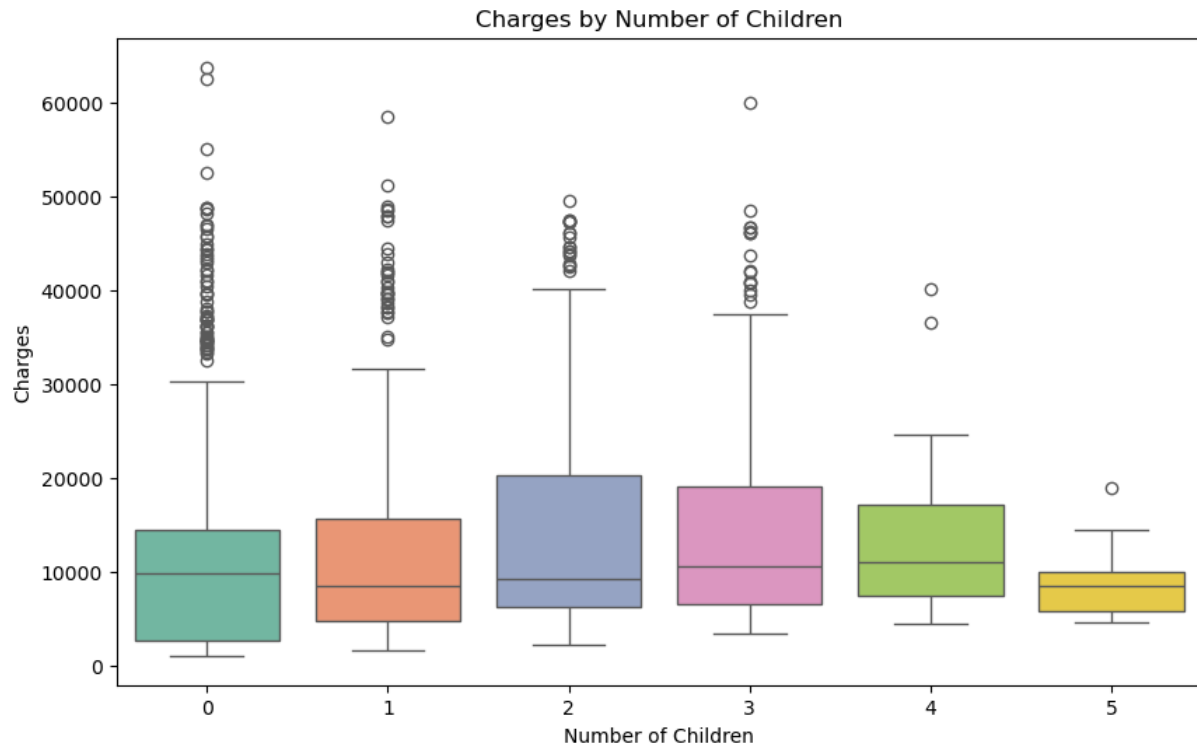


This regression plot shows us how increase in bmi causes an increase in insurance charges, however there are a lot of outliers. Further research shows that an average healthy adult will have a bmi below 30, with this in mind we can check the distribution of charges for bmi above and below 30 (Hansen Edwards et al., 2024).



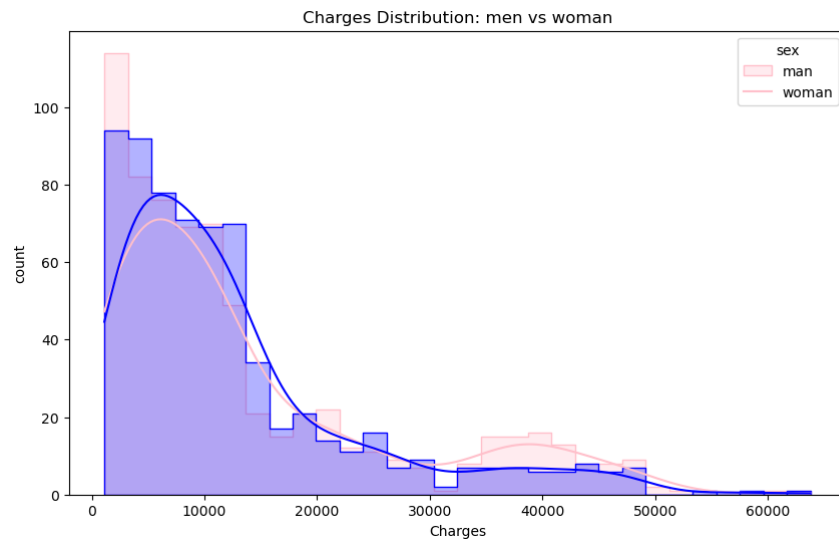
The two histplots show that higher charges are often issued to individuals with a bmi over 30 and charges are less for individuals with bmi below 30. Charges for bmi under thirty only go up to almost 40000, where above thirty bmi goes past 60000

### Children and charges:

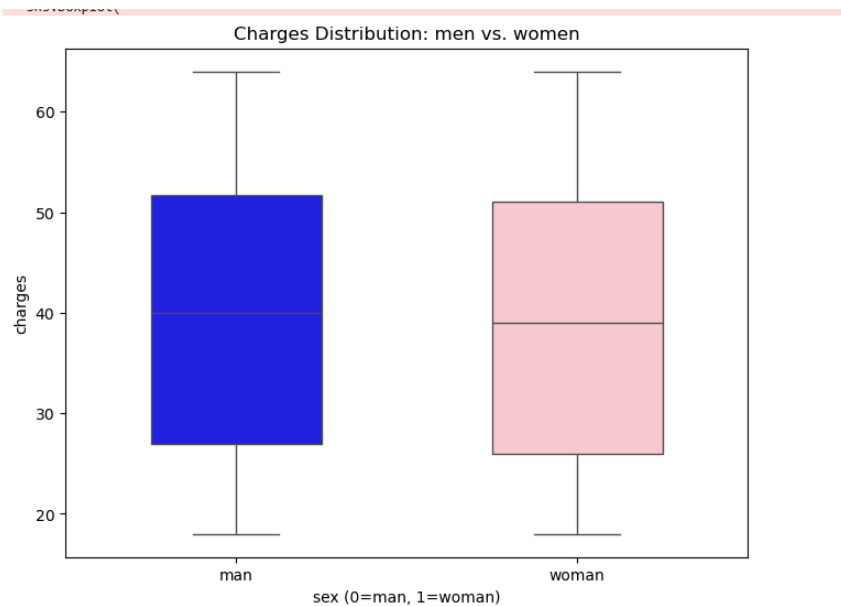


Using a box plot we can see how individuals with 2 children pay the most in terms of charges. Oddly enough individuals with no children seem to pay more charges than those with 5 children. Plenty of outliers seem to be present. Overall, the children's feature seems to have an effect on charges.

## Sex and charges:



From this hist plot we can see that men and woman are charged very similarly. Women seem to average higher costs, however that may just be due to the higher count. Sex doesn't seem to affect charges all that much.



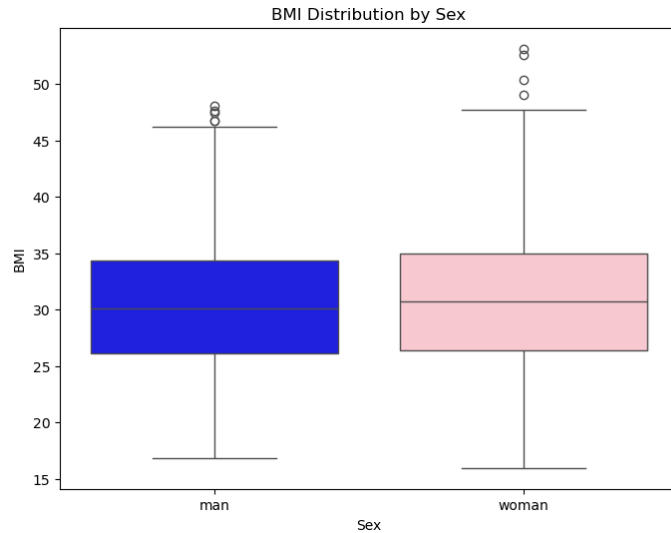
This is further seen in the box plot that shows almost no distinct difference between the distribution of costs.

## Sex and Age



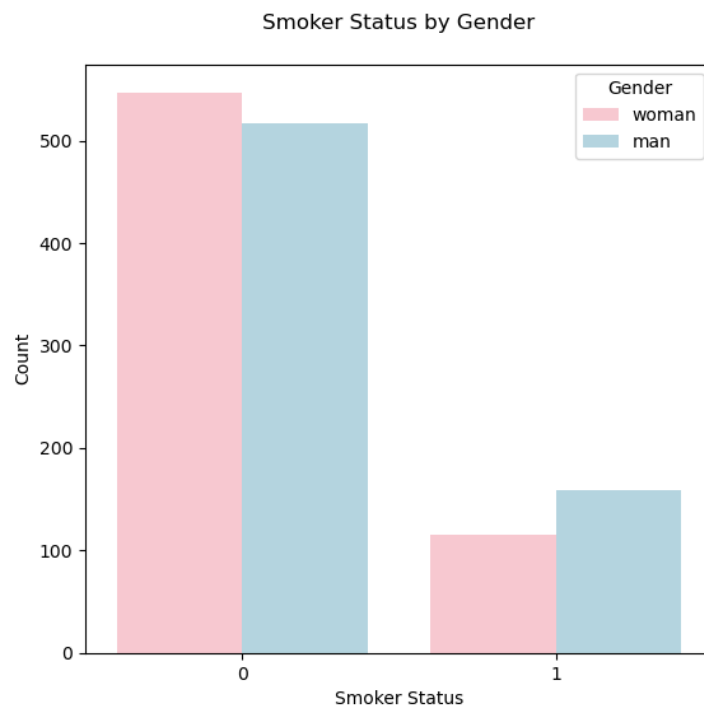
From this histplot we can see again that there is very little difference in the distribution of men and women in age. Only a minor increase in the amount of woman, most likely due to there being more woman in the data then men.

## Sex and BMI



With this boxplot we can see that there is again little difference between men and woman regarding the distribution of bmi. There are of course a few outliers for both sexes with woman seeming to have very high outliers.

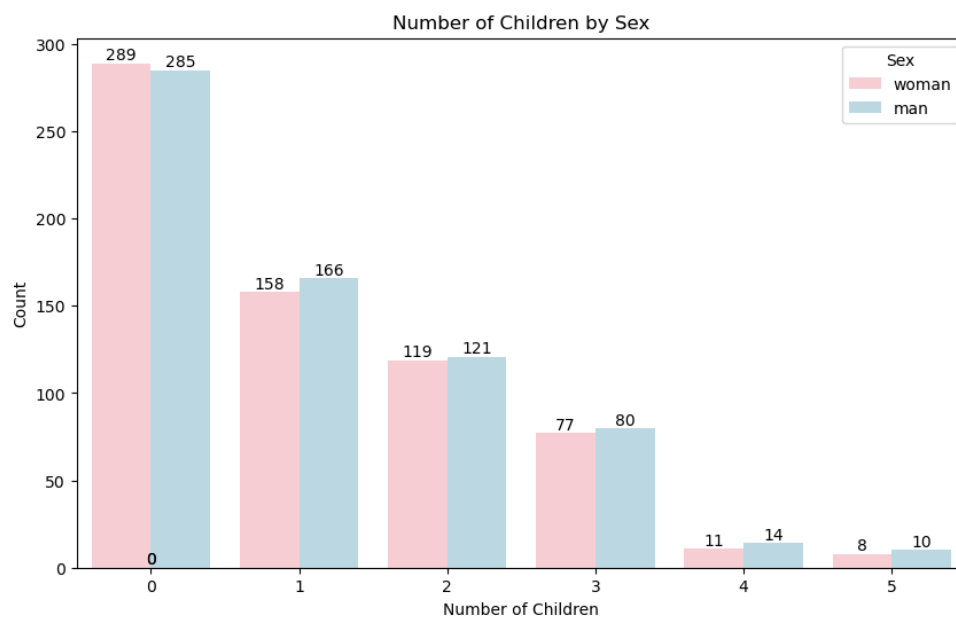
## Sex and smoker



The count plot indicates that the distribution of smokers by sex is similar with more men being smokers than woman. Maybe sex influences charges after all, since almost none of the other bivariate analysis really proved any distinct connection between sex and charges.



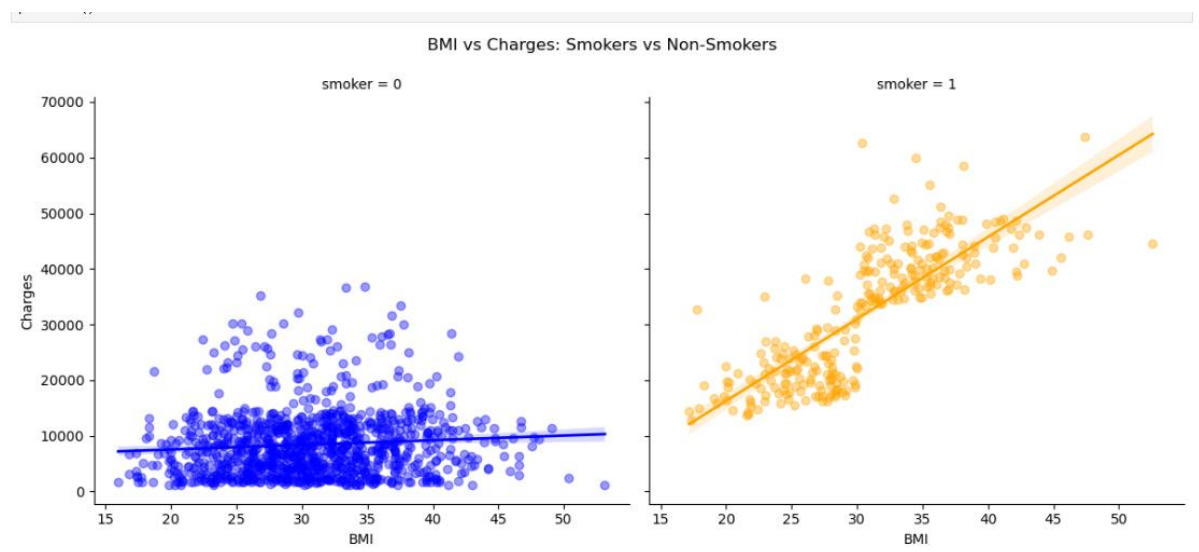
## Sex and children:



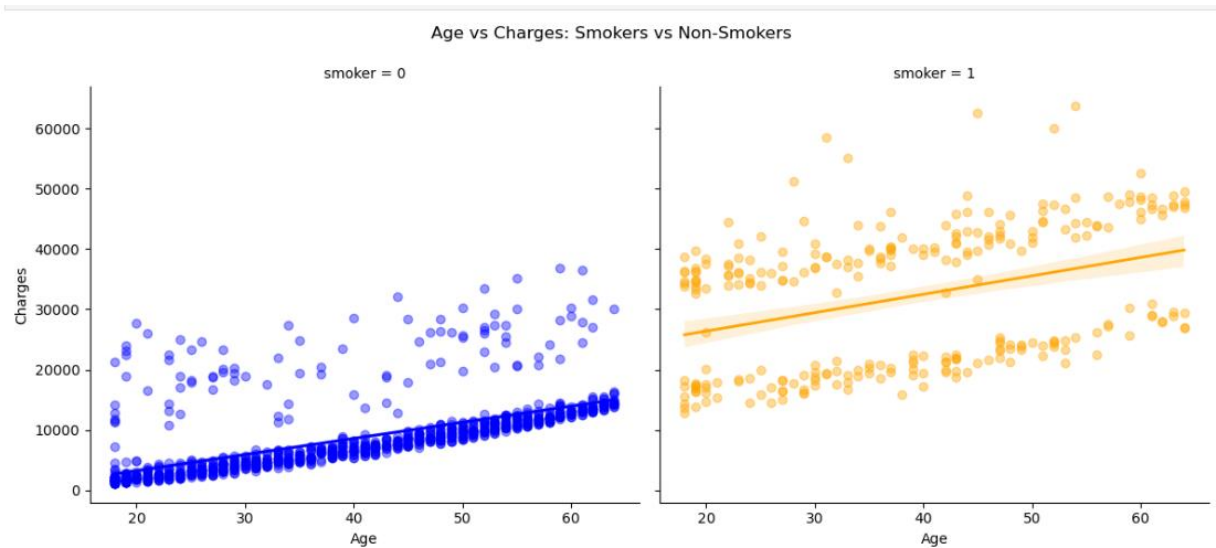
The following count plot indicates that distribution of children by sex show no significant difference or patterns. Sex so far is the weakest possible feature in terms of its effect on all independent variables, and the dependent variable.

### Multivariate analysis:

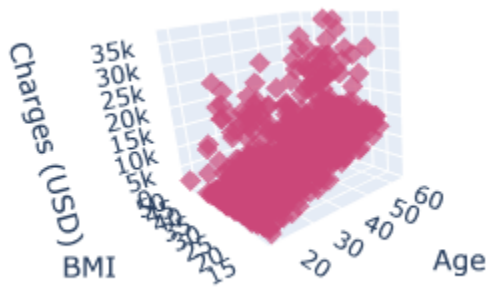
With our findings from bivariate analysis, we can see that there are a few strong patterns and trends we can follow up on. The main features that seem to have an effect on charges and one another are age, bmi, and smoker. With this uncovered we can look at all three at once.



Here we use a multi-linear regression plot. Similar to the regression plot we used earlier, this one can show multiple features. Here we can see that bmi alone affects a light increase in charges, however individuals with increased bmi that are also smokers rise in charges greatly.



This multi-linear regression plot replaces bmi with age, from this we can again see that age on its own causes an increase in charges, however with the added property of being a smoker, the increase is much greater.



Using a 3d scatter plot we can see how these three features all affect one another and how they all contribute to higher charges.

## Feature selection:

With our exploratory analysis complete, we've seen all possible trends and patterns to be found in the data. Now before we can train a model, we need to decide which features are to be included in the model to ensure the highest possible accuracy. The method used to perform feature selection will be using p values.

P values:

P values are a statistical value that indicates the likelihood of the observed data falling under the null hypothesis of a statistical test. In this case we're going to calculate the p values of each variable. The null hypothesis in this case would be if an independent variable affects the dependent variable which in this case is charges. If the p value > 0.05 it's not a significant variable and we won't include it in our model ((Banerjee Chandradip, 2023).

Using sklearn library, more specifically standard scaler we can build a small model and retrieve the p values of all the variables regarding how they affect the charges variable.

These were the results:

```
=====
                        OLS Regression Results
=====
Dep. Variable:          charges    R-squared:                0.751
Model:                  OLS        Adj. R-squared:            0.750
Method:                 Least Squares    F-statistic:           668.1
Date:                  Fri, 25 Apr 2025    Prob (F-statistic):      0.00
Time:                  06:50:52          Log-Likelihood:         -13548.
No. Observations:      1338             AIC:                  2.711e+04
Df Residuals:          1331             BIC:                  2.715e+04
Df Model:               6
Covariance Type:       nonrobust
=====
                        coef      std err          t      P>|t|      [0.025      0.975]
-----
const          1.327e+04      165.662      80.106      0.000      1.29e+04      1.36e+04
x1              3613.5362      166.932      21.647      0.000      3286.058      3941.014
x2             2027.3168      168.992      11.997      0.000      1695.798      2358.836
x3              577.6603      165.867       3.483      0.001      252.271      903.050
x4             -65.5517      166.396      -0.394      0.694     -391.979      260.876
x5             9612.5731      166.196      57.839      0.000      9286.538      9938.608
x6            -390.5855      167.799      -2.328      0.020     -719.764     -61.407
=====
Omnibus:                299.003      Durbin-Watson:           2.088
Prob(Omnibus):           0.000      Jarque-Bera (JB):        713.975
Skew:                    1.207      Prob(JB):                 9.17e-156
Kurtosis:                 5.642      Cond. No.                  1.23
=====

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

Feature p-values:
x1      2.848607e-89
x2      1.471633e-31
x3      5.125266e-04
x4      6.936815e-01
x5      0.000000e+00
x6      2.007715e-02
dtype: float64
```

X1 –x6 represents the variables 'age', 'bmi', 'children', 'sex', 'smoker', 'region' respectively.

The only variable not to include in our model is sex with the only p value greater than 0.05. The removal of this variable is understood seeing as in the bivariate analysis there was no real effect of sex on any of the other variables.

## Model training:

Now that we have our features we can train our models. I've set to train 4 separate models and compare to see which one is the most accurate.

Linear regression models make predictions of a dependent variable based on it having a linear relationship with its independent variables. To evaluate each model, we'll be looking at the following:

R<sup>2</sup> score:

This is the coefficient of determination and represents how well the data is fitted to the model by how well the model explains the variance in the data. A good score would be close to the value 1 (Anon., 2025d).

RMSE:

The average sum of the errors made in the predictions of the model. In other words, It tells you how far off your model is on average. This score needs to be low for a good model (Anon., 2025d).

Coefficients:

These are the impact of each variable on the model. It lets you know the type of change you should expect in the prediction as that coefficient goes up or down (Anon., 2025d).

## Regular linear regression:

The results of using a regular linear regression model from sklearn library:

```
R2 Score: 0.7962793642414525
RMSE: 5663.272820202942
```

Coefficients:

|   | Feature  | Coefficient |
|---|----------|-------------|
| 0 | smoker   | 9473.400815 |
| 1 | age      | 3516.995718 |
| 2 | bmi      | 2009.994068 |
| 3 | children | 543.295506  |
| 4 | region   | -376.697530 |

R<sup>2</sup> score is good; we can assume that the model can explain about 80% of the variance of insurance charges.

RMSE score indicates that on average the predicted charge is off by +- 5663 which still isnt bad.

## Ridge model:

These are the results of using a ridge model from sklearn library.

```
Ridge Regression Results:  
R2 Score: 0.7962131351694951  
RMSE: 5664.193303350936
```

```
Coefficients:  
  Feature  Coefficient  
0   smoker  9473.400815  
1     age   3516.995718  
2     bmi   2009.994068  
3 children   543.295506  
4   region  -376.697530
```

The R<sup>2</sup> score is still good and only slightly worse than the regular regression model.

The RMSE score is also only slightly worse by one degree.

## Lasso Model:

These are the results of the lasso regression model from sklearn library:

```
Lasso Regression Results:  
R2 Score: 0.7962779681837642  
RMSE: 5663.292224820417
```

```
Coefficients:  
  Feature  Coefficient  Selected  
0   smoker  9482.966143    True  
1     age   3520.727173    True  
2     bmi   2011.642004    True  
3 children   543.240850    True  
4   region  -376.885883    True
```

R<sup>2</sup> is better than the ridge model, however its still worse than the original regular regression model.

REMSE is also better than the ridge model, but still worse than the original regular regression model.

## Elastic regression model:

---

ElasticNet Results:

$R^2$  Score: 0.791

RMSE: 5730.61

Coefficients:

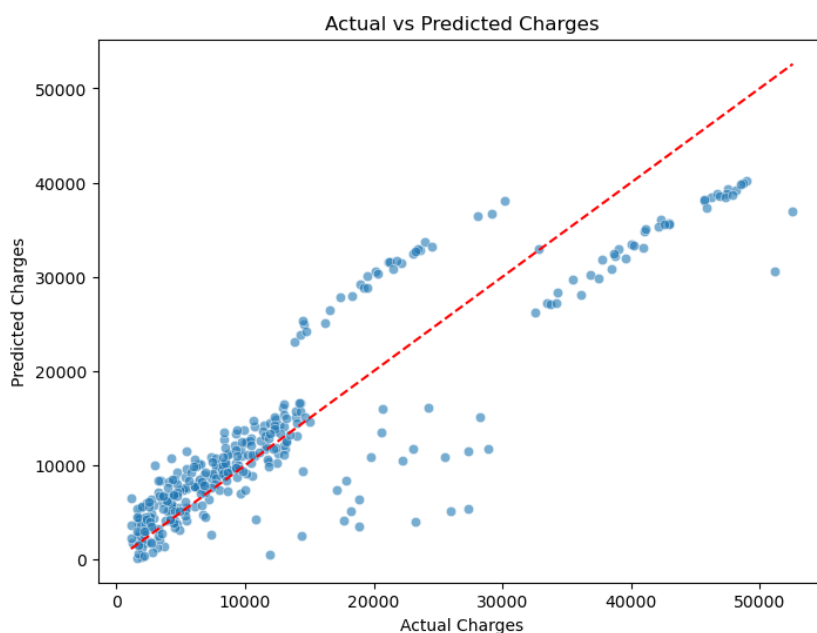
|   | Feature  | Coefficient | Selected |
|---|----------|-------------|----------|
| 0 | smoker   | 9021.495332 | True     |
| 1 | age      | 3339.273101 | True     |
| 2 | bmi      | 1927.143028 | True     |
| 3 | children | 540.209439  | True     |
| 4 | region   | -362.165407 | True     |

---

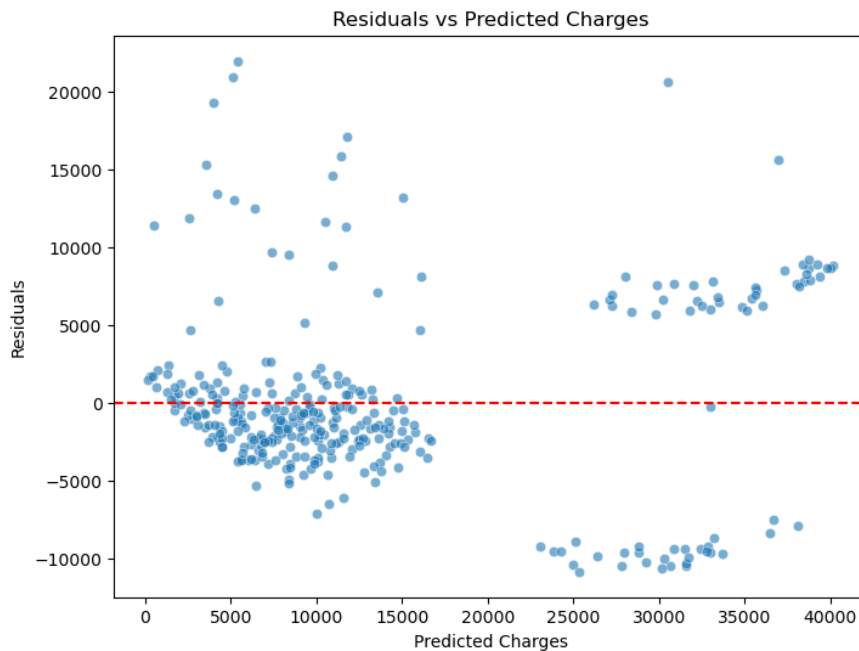
The  $R^2$  score is still not bad but is significantly worse than all the other trained models.

The RMSE score is also worse than all the other models

## Linear regression graphs:







These are the actual vs predicted, and residuals vs predicted graphs of the normal linear regression model, also known as the most accurate model.

Looking at the actual vs predicted graph, we can see that the model is more accurate in predicting the lower charges, and losses accuracy as the charges increase.

The residuals vs predicted charges also indicate the same thing, that errors are much more prominent among higher charges predictions.

## Conclusion:

After a thorough exploratory data analysis covering univariate, bivariate and multivariate analysis, we've uncovered many patterns and trends to give us an idea of how the data in this model influences one another. With this we found that BMI, age, and smoker we're the strongest features, while sex was the weakest. The feature selection further emphasized the results of the exploration as the sex variable was as expectedly dropped. Furthermore, after training 4 linear regression models it's clear that the simplest and most basic regular linear regression model is the best for predicting the insurance charges since its RMSE and  $R^2$  score indicate that the model had the most accurate results.

## References

- Anon. 2023. *How Does Age Affect Life Insurance Premiums* | *Hollard*. [online] Available at: <<https://www.hollard.co.za/media-centre/news-and-articles/life/how-does-age-affect-life-insurance>> [Accessed 25 April 2025].
- Anon. 2024. *Univariate, Bivariate and Multivariate data and its analysis* | *GeeksforGeeks*. [online] Available at: <<https://www.geeksforgeeks.org/univariate-bivariate-and-multivariate-data-and-its-analysis/>> [Accessed 25 April 2025].
- Anon. 2025a. *Linear Regression with sklearn using categorical variables* | *Saturn Cloud Blog*. [online] Available at: <<https://saturncloud.io/blog/linear-regression-with-sklearn-using-categorical-variables/>> [Accessed 25 April 2025].
- Anon. 2025b. *Pearson Correlation and Linear Regression*. [online] Available at: <<https://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/>> [Accessed 25 April 2025].
- Anon. 2025c. *Understanding the Assumptions of Linear Regression Analysis*. [online] Available at: <<https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/>> [Accessed 25 April 2025].
- Anon. 2025d. *What are  $R^2$  and RMSE?* [online] Available at: <<https://click.clarity.io/knowledge/r2-rmse>> [Accessed 25 April 2025].
- Banerjee Chandradip, 2023. *P value and Feature Selection. P value and Feature Selection* | *by Chandradip Banerjee* | *Medium*. [online] Available at: <<https://medium.com/@chandradip93/p-value-and-feature-selection-629bec71d828>> [Accessed 25 April 2025].
- Barendregt, J.J., Bonneux, L. and van der Maas, P.J., 1997. The Health Care Costs of Smoking. *New England Journal of Medicine*, [online] 337(15), pp.1052–1057. <https://doi.org/10.1056/NEJM199710093371506>,.
- Hansen Edwards, C., Håkon Bjørngaard, J., Minet Kinge, J., Åberge Vie, G., Halsteinli, V., Ødegård, R., Kulseng, B. and Waaler Bjørnelv, G., 2024. The healthcare costs of increased body mass index—evidence from The Trøndelag Health Study. *Health Economics Review*, [online] 14(1), pp.1–11. <https://doi.org/10.1186/S13561-024-00512-8/FIGURES/5>.