

**Project Report: INSURANCE CLAIM BILL**

**Report for assessment of Data Analyst Project**

**Submitted by:**

**Nithin S Shetty**

**Academic Year: 2023-24**

**Department of Data Analyst**

**Great Learning Academy**

**Plot No.758-759,1st Floor,19th Main Rd, Sector 2, HSR Layout**

**Bengaluru-560** **102**

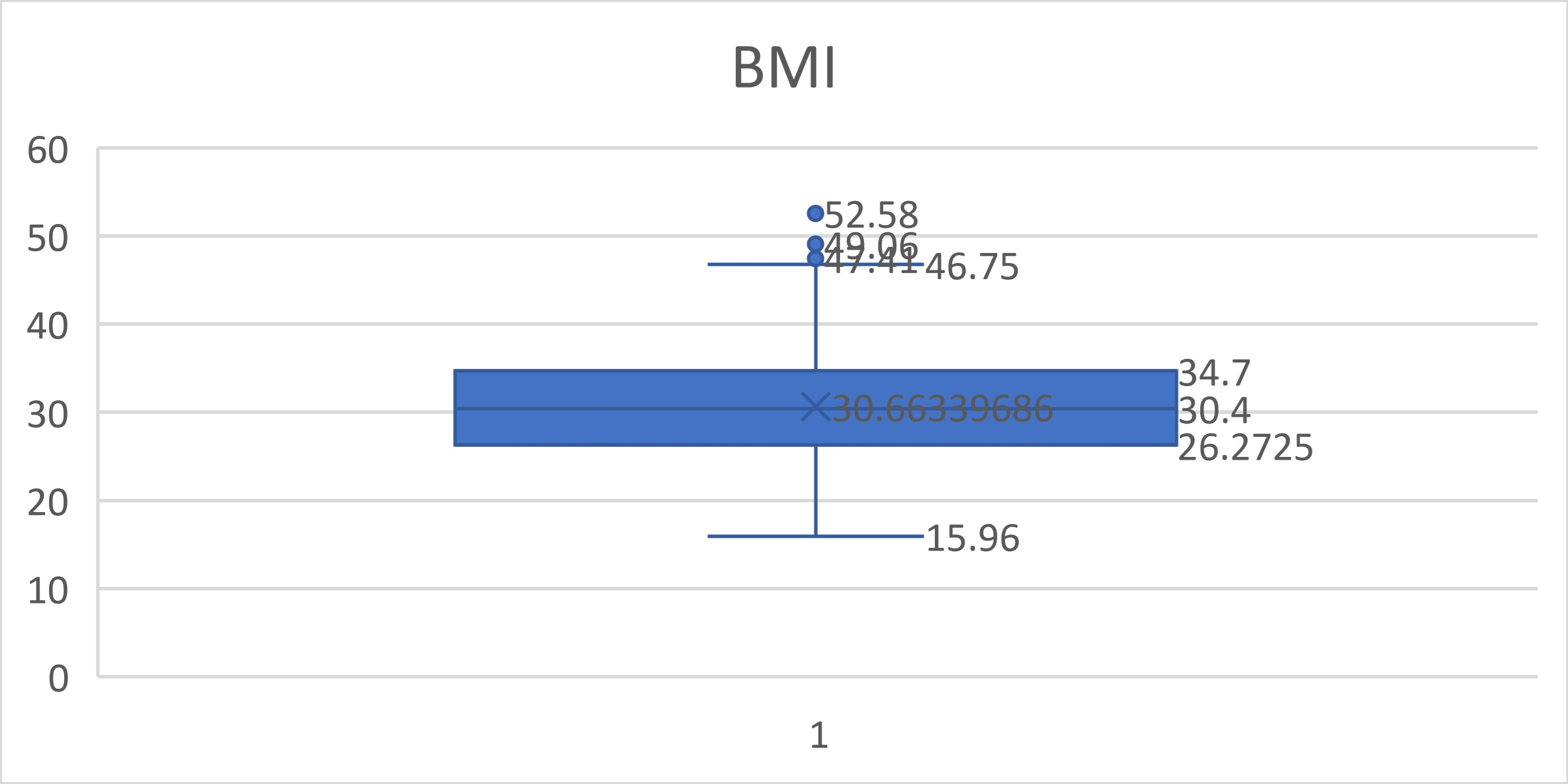
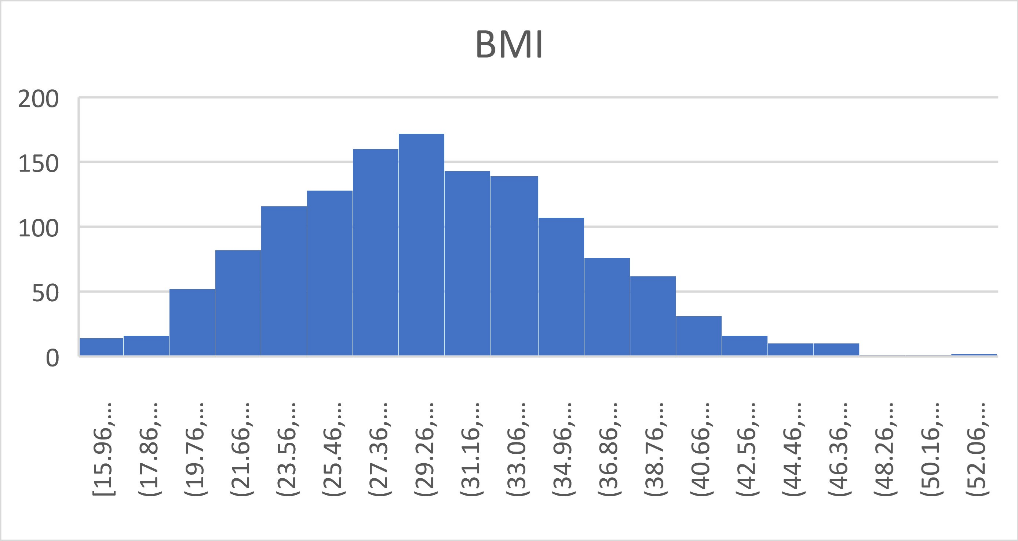
# 1) Perform the Exploratory Data Analysis on the data.

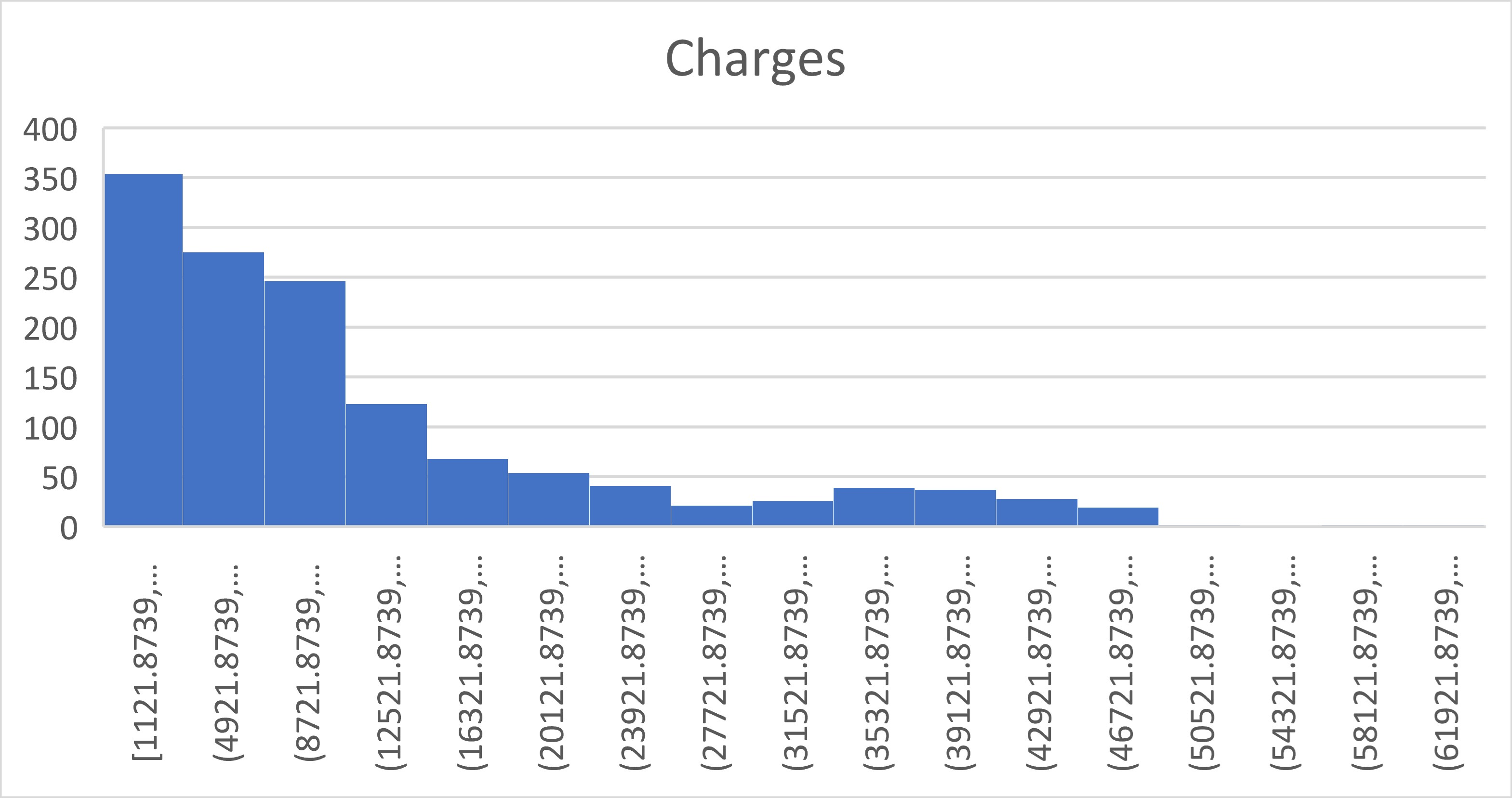
### Identify the categorical and continuous variables

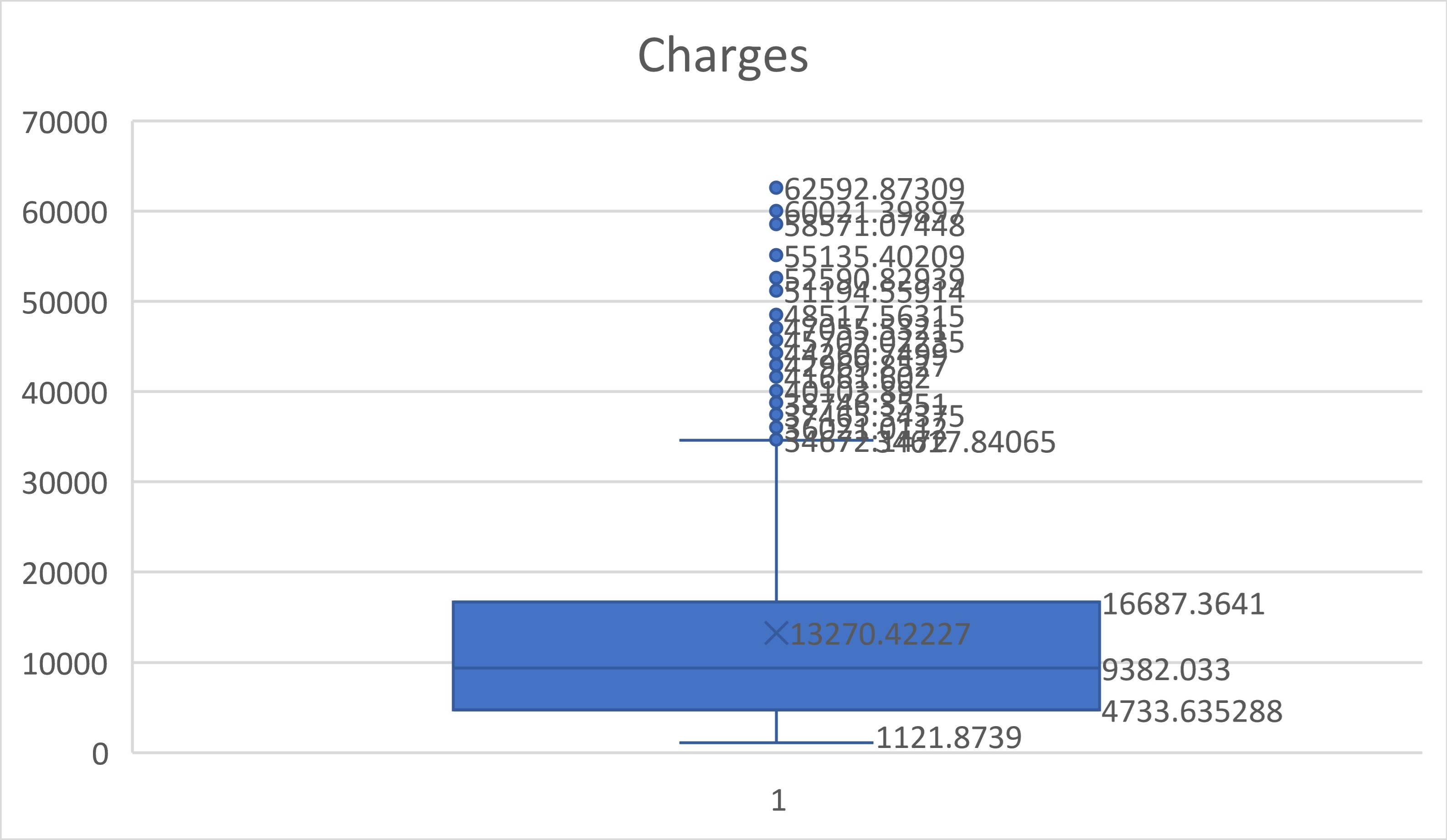
|  |
| --- |
| **Categorical variables Continuous variables** |
| Sex Bmi |
| Smoker Charges |
| Region |

**Age** and **Children** is discrete so we separately place it in the category discrete.

* 1. Make Histograms and box plots (univariate analysis) for continuous variables and do a correlation analysis (multivariate analysis)







|  |  |  |
| --- | --- | --- |
| **Correlation analysis** | | |
|  |  |  |
|  | *bmi* | *charges($)* |
| bmi | 1 |  |
| charges($) | 0.198340969 | 1 |

* 1. **Make relevant Pivot tables and charts for:**
     1. Male/Female ratio and share information on which gender has more smokers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Count of smoker**  **Sex** | **Column Labels no** |  | **yes** |  |
| female |  | 547 |  | 115 |
| male |  | 517 |  | 159 |

**we can conclude that males has more smokers.**

600

500

400

300

no

yes

200

100

0

female

male

* + 1. Charges vs Age

|  |  |
| --- | --- |
| **Age** | **Average of**  **charges($)** |
| 15-24 | 9011.340 |
| 25-34 | 10352.393 |
| 35-44 | 13134.169 |
| 45-54 | 15853.928 |
| 55-65 | 18513.276 |
| **Grand**  **Total** | **13270.422** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | |  |  |  |
|  |  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

* + 1. **Charges vs BMI**

20000.000

18000.000

16000.000

14000.000

12000.000

10000.000

Total

8000.000

6000.000

4000.000

2000.000

0.000

15-24

25-34

35-44

45-54

55-65

**BMI**

**Average of**

**charges($)**

15-25 10282.22447

25-35 12714.63543

35-45 16913.68151

45-55 17547.92675

20000

18000

16000

14000

12000

10000

Total

8000

6000

4000

2000

0

15-25

25-35

35-45

45-55

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |
|  |  |
|  |  |
|  |  |

### Charges for Smokers vs Non-smokers

**Average of Smokers charges($)**

yes 32050.23183

no 8434.268298

35000

30000

25000

20000

15000

Total

10000

5000

0

yes

no

## Region-wise smokers vs Non-smokers analysis with one or more pivot table and charts

|  |  |  |  |
| --- | --- | --- | --- |
| **No of smoker**  **Row Labels** | **Region yes** | **no** |  |
| northeast | 67 |  | 257 |
| northwest | 58 |  | 267 |
| southeast | 91 |  | 273 |
| southwest | 58 |  | 267 |

Region wise Smokers

300

250

200

150

100

50

0

yes

no

northeast northwest southeast southwest

Region

No of Smokers

**Southeast has more number of smokers and northwest along with southwest holds the less number of smokers**

### Region-wise charges for smokers vs non-smokers

|  |  |  |
| --- | --- | --- |
| **Average of**  **charges($) Column Labels** | | |
| **Row Labels** | **no** | **yes** |
| northeast | 9165.532 | 29673.536 |
| northwest | 8556.464 | 30192.003 |
| southeast | 8032.216 | 34844.997 |
| southwest | 8019.285 | 32269.063 |

* 1. **Has charges got something to do with the number of dependents ?**

Correlation between number of dependents and charges = 0.067998

**Since we have a positive relation we can say that they are directly related. Thus we can say that as the value of no of dependents increase, charges also increase.**

### Do a similar dependents-charges analysis, Region-wise

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Average of charges($)** | **Number of Students** |  | | | | |
| **Region** | **0** | **1** | **2** | **3** | **4** | **5** |
| northeast | 11626.463 | 16310.206 | 13615.153 | 14409.913 | 14485.193 | 6978.973 |
| northwest | 11324.371 | 10230.256 | 13464.315 | 17786.161 | 11347.019 | 8965.796 |
| southeast | 14309.868 | 13687.042 | 15728.471 | 18449.846 | 14451.024 | 10115.442 |
| southwest | 11938.505 | 10406.485 | 17483.486 | 10402.442 | 14933.261 | 8444.159 |

* 1. **Do at least one more pivot table and chart of your own choice on the remaining variables**

|  |  |  |
| --- | --- | --- |
| **Average of bmi Row Labels** | **Sex female** | **male** |
| no | 30.53952468 | 30.77058027 |
| yes | 29.60826087 | 31.50418239 |

### Give your understanding from the patterns observed in point (b)

* + - * **The data in BMI is normally distributed with a median of 30.4.**
      * **For BMI the first quartile data is under 26.272 and third quartile data is under 34.7.**
      * **The data in Charges are positively skewed with a median of 9382.033.**
      * **The first quartile data is under 4733.635.**

### Give your interpretation for observations made in point (c)

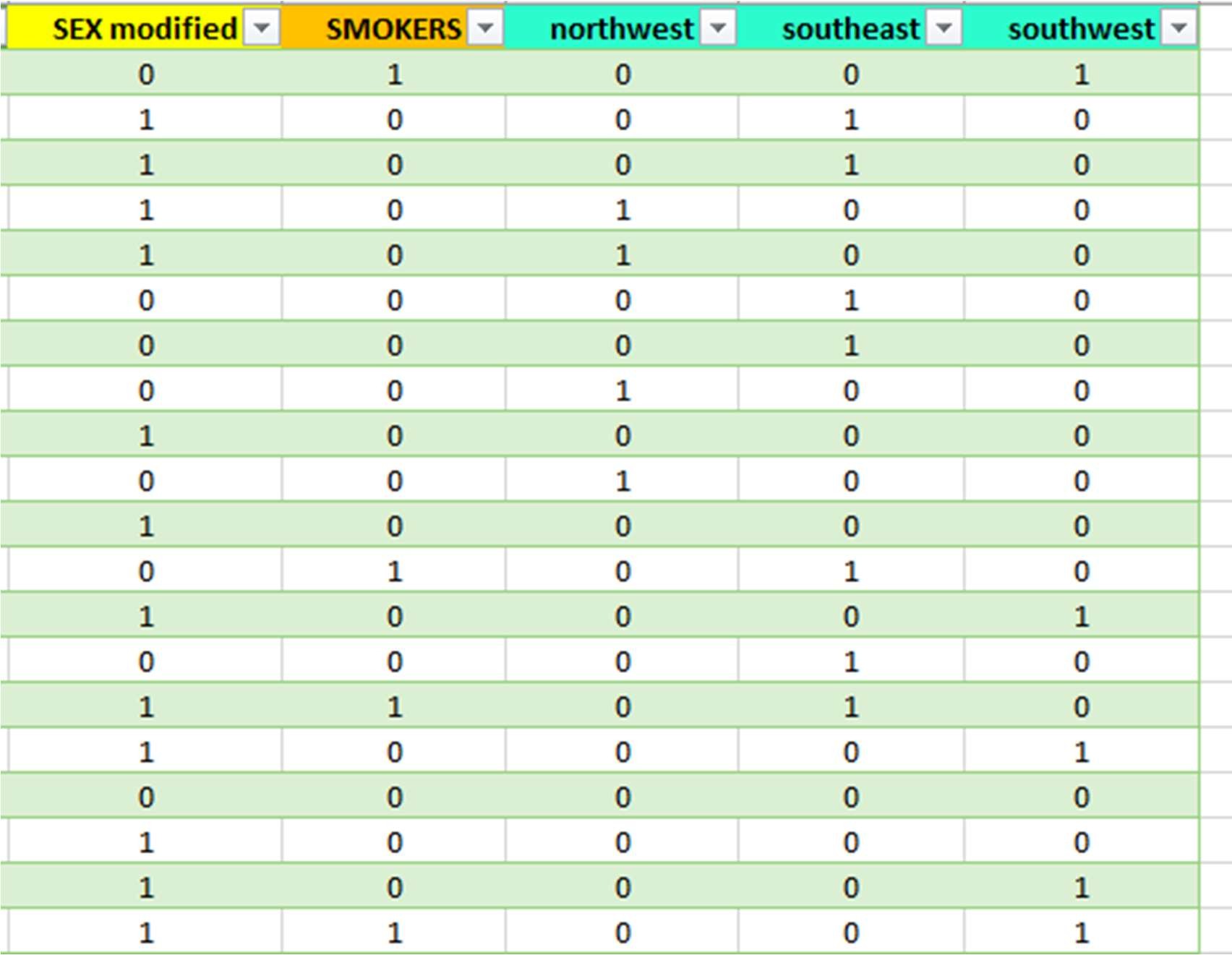
* + - * **Males has more number of smokers.**

### The BMI range of 45-55 has highest average charge of 17547.92675.

* + - * **The Age group 55-65 has the highest average charge of 18513.26.**

# Edit the data as following, to obtain dummy variables:

1. Sex : Replace all the “Males” with “1” and “Females” with “0”, creating numerical entries for gender this way will help you do analysis further. You can use the “Replace with Match entire cell content” option. Do a replace all to save time.
2. Smoker: Replace all the “Smokers” with “1” and “Non-smokers” with “0”.
3. Region: We always create one less category column for the dummy data w.r.t the categories available for that original variable. So for Region, we will create three dummy columns, assuming “Northeast” as zero and omit the column for it. Now create three columns for “northwest”, “Southeast”, “Southwest”. Whichever row has “northwest” region as an entry will take “1” as an entry otherwise “0” in “northwest” column. Similarly in the “Southeast” column, whichever row had “southeast” as an entry will take “1” as the new entry and “0” for the rest of the column (Southeast). Do a similar operation on the “Southwest” column. Please refer to the below image for your understanding,
4. We use the if function to edit the data (=IF(Cell="male",1,0))
5. We use the if function to edit the data (=IF(Cell="yes",1,0))
6. We use the if function to edit the data (=IF(Cell=" northwest",1,0)) We use the if function to edit the data (=IF(Cell=" Southeast",1,0)) We use the if function to edit the data (=IF(Cell=" Southwest ",1,0))



### Do a descriptive summary analysis for the edited data. Perform a Multiple Linear Regression analysis to identify which variables decide the insurance charges/billed insurance claim. Give your interpretation for the above analysis, do another set of regression analysis by dropping insignificant variables, if needed.

**Descriptive Summary Analysis of edited data**

**We use the summary statistics in the data analytics function**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***age*** |  | | ***bmi*** | |  | |
|  |  | |  | |  | |
| **Mean** | **39.2070** | | **Mean** | | **30.6634** | |
| **Standard Error** | **0.3841** | | **Standard Error** | | **0.1667** | |
| **Median** | **39.0000** | | **Median** | | **30.4000** | |
| **Mode** | **18.0000** | | **Mode** | | **32.3000** | |
| **Standard**  **Deviation** | **14.0500** | | **Standard**  **Deviation** | | **6.0982** | |
| **Sample Variance** | **197.4014** | | **Sample Variance** | | **37.1879** | |
| **Kurtosis** | **-1.2451** | | **Kurtosis** | | **-0.0507** | |
| **Skewness** | **0.0557** | | **Skewness** | | **0.2840** | |
| **Range** | **46.0000** | | **Range** | | **37.1700** | |
| **Minimum** | **18.0000** | | **Minimum** | | **15.9600** | |
| **Maximum** | **64.0000** | | **Maximum** | | **53.1300** | |
| **Sum** | **52459.0000** | | **Sum** | | **41027.6250** | |
| **Count** | **1338.0000** | | **Count** | | **1338.0000** | |
|  |  | |  | |  | |
| *children* |  | *SEX* | |  | |  |
|  |  |  | |  | |
| Mean | 1.0949 | Mean | | 0.5052 | |
| Standard Error | 0.0330 | Standard Error | | 0.0137 | |
| Median | 1.0000 | Median | | 1.0000 | |
| Mode | 0.0000 | Mode | | 1.0000 | |
| Standard  Deviation | 1.2055 | Standard  Deviation | | 0.5002 | |
| Sample Variance | 1.4532 | Sample Variance | | 0.2502 | |
| Kurtosis | 0.2025 | Kurtosis | | -2.0026 | |
| Skewness | 0.9384 | Skewness | | -0.0210 | |
| Range | 5.0000 | Range | | 1.0000 | |
| Minimum | 0.0000 | Minimum | | 0.0000 | |
| Maximum | 5.0000 | Maximum | | 1.0000 | |
| Sum | 1465.0000 | Sum | | 676.0000 | |
| Count | 1338.0000 | Count | | 1338.0000 | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *SMOKERS* |  | *northwest* |  | *southeast* |  |
|  |  |  |  |  |  |
| Mean | 0.2048 | Mean | 0.2429 | Mean | 0.2720 |
| Standard Error | 0.0110 | Standard Error | 0.0117 | Standard Error | 0.0122 |
| Median | 0.0000 | Median | 0.0000 | Median | 0.0000 |
| Mode | 0.0000 | Mode | 0.0000 | Mode | 0.0000 |
| Standard Deviation | 0.4037 | Standard Deviation | 0.4290 | Standard Deviation | 0.4452 |
| Sample Variance | 0.1630 | Sample Variance | 0.1840 | Sample Variance | 0.1982 |
| Kurtosis | 0.1458 | Kurtosis | -0.5599 | Kurtosis | -0.9495 |
| Skewness | 1.4648 | Skewness | 1.2004 | Skewness | 1.0256 |
| Range | 1.0000 | Range | 1.0000 | Range | 1.0000 |
| Minimum | 0.0000 | Minimum | 0.0000 | Minimum | 0.0000 |
| Maximum | 1.0000 | Maximum | 1.0000 | Maximum | 1.0000 |
| Sum | 274.0000 | Sum | 325.0000 | Sum | 364.0000 |
| Count | 1338.0000 | Count | 1338.0000 | Count | 1338.0000 |

|  |  |  |  |
| --- | --- | --- | --- |
| *southwest* |  | *charges($)* |  |
|  |  |  |  |
| Mean | 0.2429 | Mean | 13270.4223 |
| Standard Error | 0.0117 | Standard Error | 331.0675 |
| Median | 0.0000 | Median | 9382.0330 |
| Mode | 0.0000 | Mode | 1639.5631 |
| Standard Deviation | 0.4290 | Standard Deviation | 12110.0112 |
| Sample Variance | 0.1840 | Sample Variance | 146652372.1529 |
| Kurtosis | -0.5599 | Kurtosis | 1.6063 |
| Skewness | 1.2004 | Skewness | 1.5159 |
| Range | 1.0000 | Range | 62648.5541 |
| Minimum | 0.0000 | Minimum | 1121.8739 |
| Maximum | 1.0000 | Maximum | 63770.4280 |
| Sum | 325.0000 | Sum | 17755824.9908 |
| Count | 1338.0000 | Count | 1338.0000 |

### We use the regression analysis in data analytics function from the data tab for Multiple Linear Regression analysis

