PROBLEM STATEMENT

To predict the insurance policy charges based on the demographic data

Independent Variables

Age, Sex , BMI , Children, Smoker, Region

Dependent Variables

Charges

NEED

To predicting future medical expenses of individuals that help medical insurance to make decision on charging the premium

DATA UNDERSTANDING

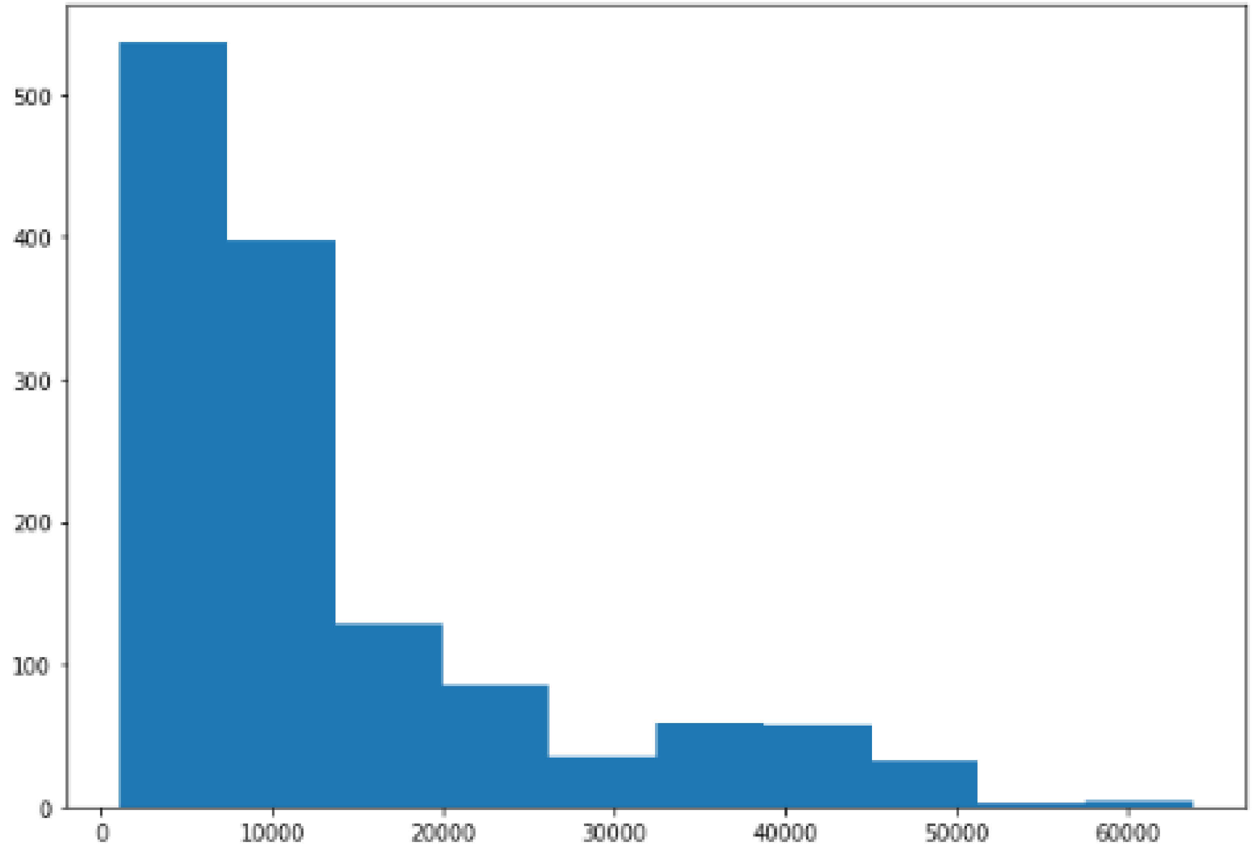
Variableage	Type	Description
Age	Numeric	age of primary beneficiary (in years)
Gender	Category	gender of insurance contractor, either female or male
BMI	Numeric	Body Mass Index which provides an understanding of a body by using a number expressing the ratio of body weight (in kilograms) to height squared (in meters). The value of bmi is ideally between 18.5 and 24.9
Children	Numeric	number of children/dependents covered by health insurance
Smoker	Category	: whether the primary beneficiary smoking or not
Region	Category	the beneficiary's residential area in the US, either northeast, southeast, southwest, or northwest
Charges	Numeric	Individual medical costs billed by health insurance We will use charges as our target variable and the rest as the candidate predictors.

EDA

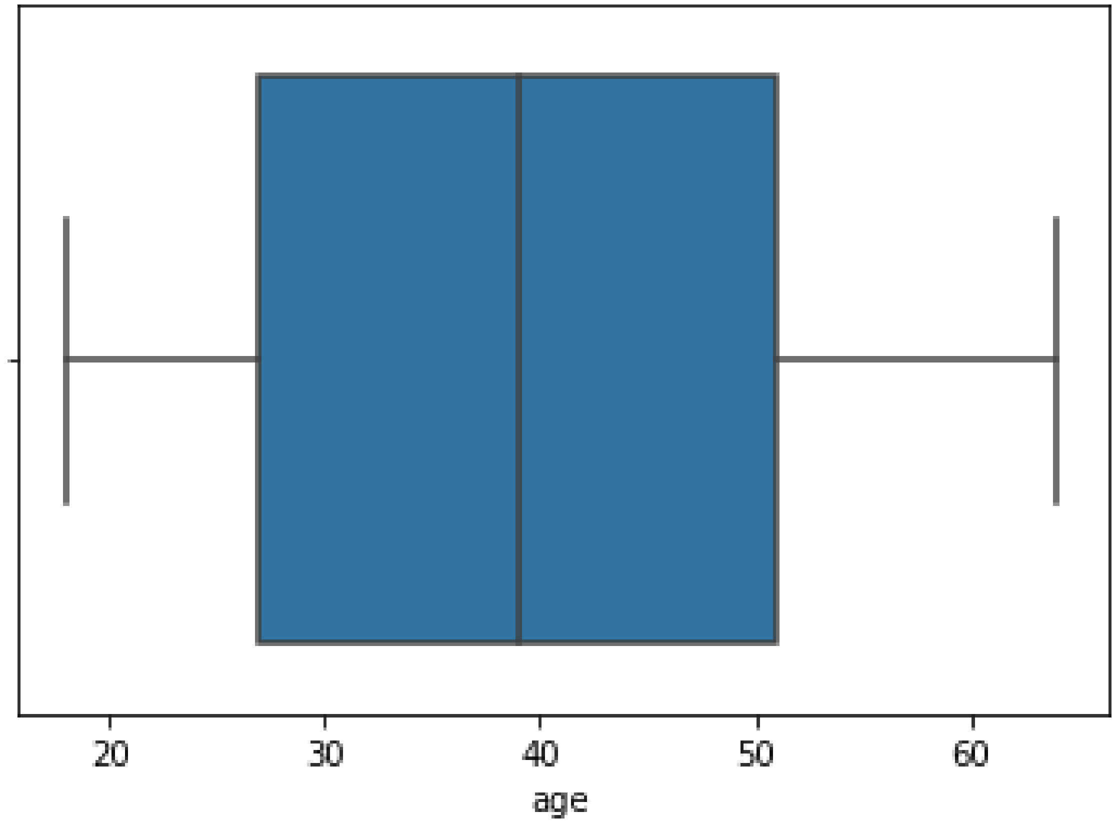
- Complete data of 1338 Entries with 7 columns
- No missing values
- Average age is 39
- Overweight (Avaerage BMI~30)
- Max insurancecharges 63770
- Target variable not normally distributed

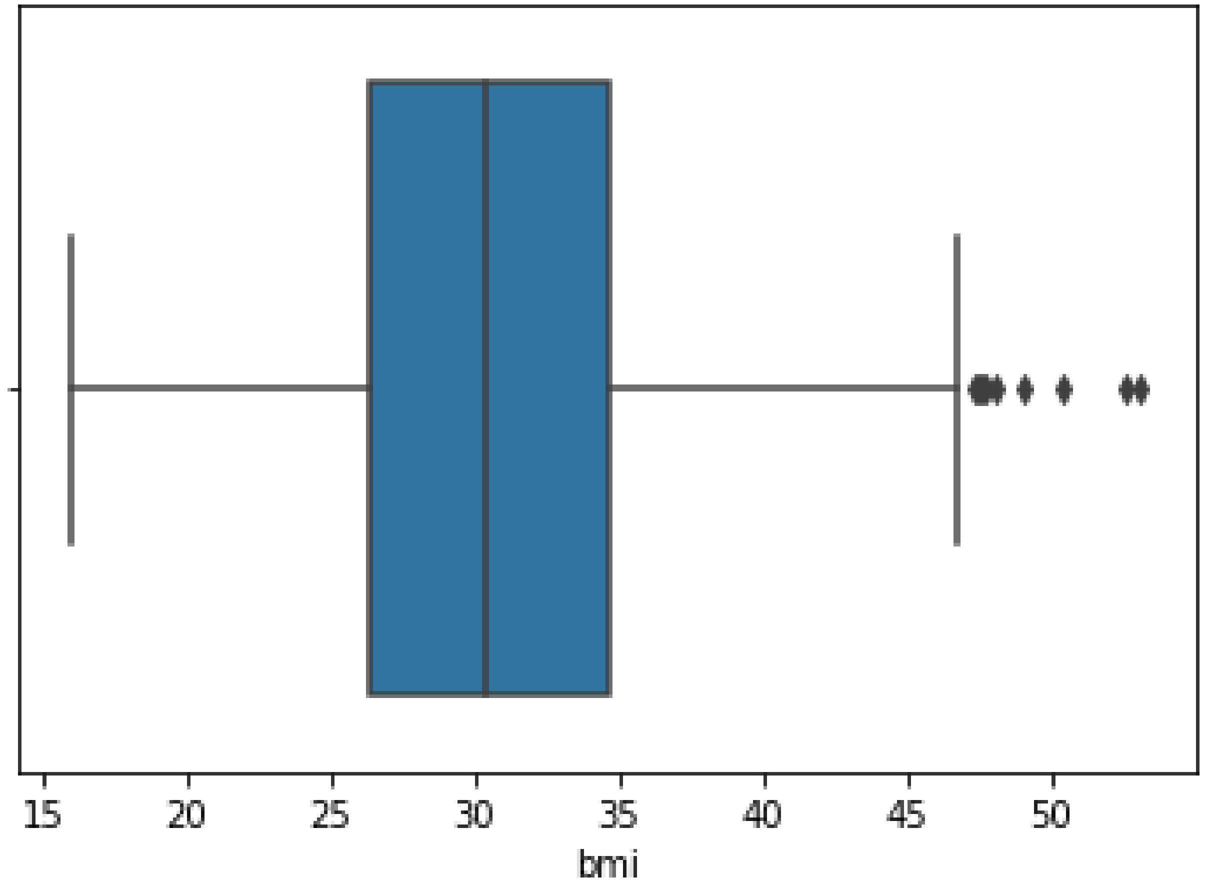
```
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
```

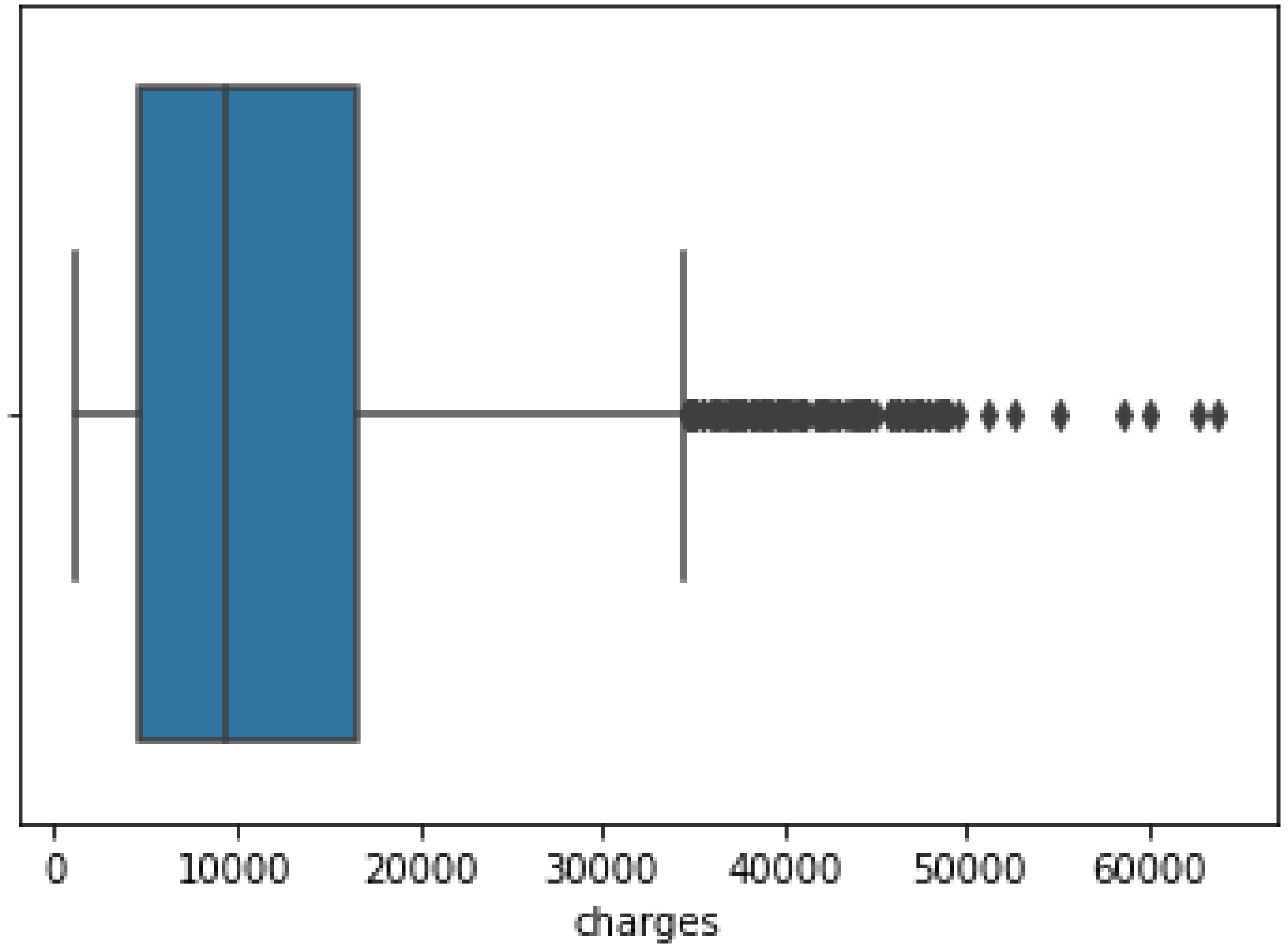
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010



OUTLIER DETECTION AND REMOVAL







OUTLIER DETECTION AND REMOVAL

- Outliers present in BMI, Charges
- Lower and upper ranges are identified
- Entries reduced from 1338 to 1191
- 10.9 percent reduction of records

```
def dropout(df,col):
  for i in col:
    q25,q75 = np.percentile(a = df[i],q=[25,75])
    IQR = q75 - q25
    lowrange=q25-(1.5*IQR)
    uprange=q75+(1.5*IQR)
    print (i," lower = ",lowrange," upper = ",uprange)
    df=df[(df[i]>=lowrange) & (df[i]<=uprange) ]
  return df
```

```
col = ["bmi","charges"]
newdf = dropout(data,col)
print(newdf.shape)
```

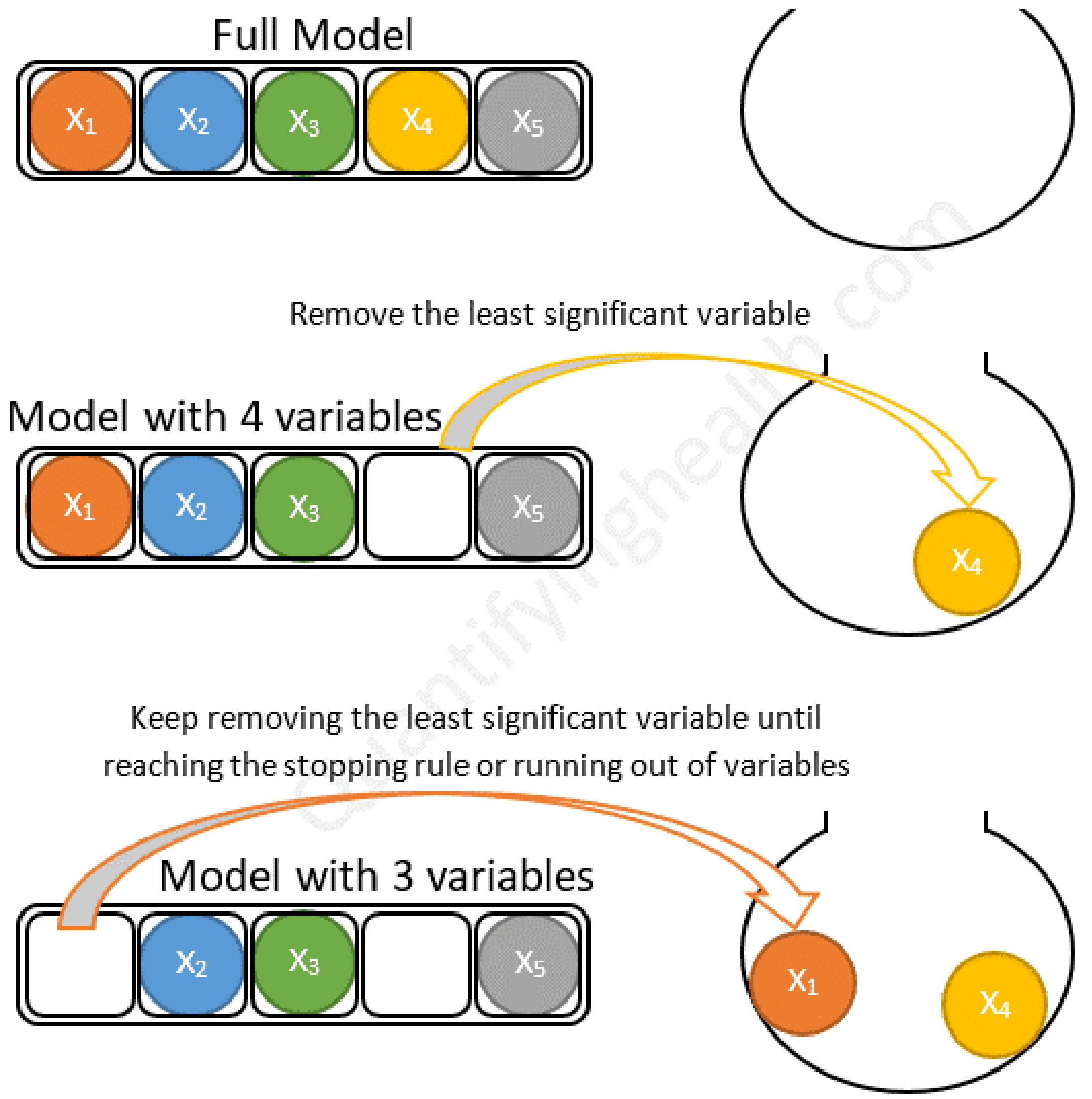
```
bmi  lower =  13.7  upper =  47.290000000000006
charges  lower =  -13034.076065  upper =  34358.841975
(1191, 7)
```

DUMMY VARIABLES

- Dummay variables for categorical variables - Gender, Smoker,Region

	age	bmi	children	charges	sex_male	smoker_yes	region_northwest	region_southeast	region_southwest
0	19	27.900	0	16884.92400	0	1	0	0	1
1	18	33.770	1	1725.55230	1	0	0	1	0
2	28	33.000	3	4449.46200	1	0	0	1	0
3	33	22.705	0	21984.47061	1	0	1	0	0
4	32	28.880	0	3866.85520	1	0	1	0	0

STEPWISE REGRESSION



SOME IMPORTANT TERMS:

TOP SECTION:

R-SQUARED TELLS ABOUT THE GOODNESS OF THE FIT, RANGES BETWEEN 0 AND 1. THE CLOSER THE VALUE TO 1, THE BETTER IT EXPLAINS THE DEPENDENT VARIABLES VARIATION IN THE MODEL. HOWEVER, IT IS BIASED IN A WAY THAT IT NEVER DECREASES WHEN WE ADD NEW VARIABLES.

ADJ. R-SQUARED HAS A PENALIZING FACTOR. IT DECREASES OR STAYS IDENTICAL TO THE PREVIOUS VALUE AS THE NUMBER OF PREDICTORS INCREASES. IF THE VALUE KEEPS INCREASING ON REMOVING THE UNNECESSARY PARAMETERS GO AHEAD WITH THE MODEL OR STOP AND REVERT.

F-STATISTIC USED TO COMPARE TWO VARIANCES AND THE VALUE IS ALWAYS GREATER THAN 0. IN REGRESSION, IT IS THE RATIO OF THE EXPLAINED TO THE UNEXPLAINED VARIANCE OF THE MODEL.

MID SECTION

COEF IS THE COEFFICIENT/ESTIMATE VALUE OF INTERCEPT AND SLOPE.

P>|T| REFERS TO THE P-VALUE OF PARTIAL TESTS WITH THE NULL HYPOTHESIS H0 THAT THE COEFFICIENT IS EQUAL TO ZERO (NO EFFECT).

A LOW P-VALUE (< 0.05) INDICATES THAT THE PREDICTOR HAS SIGNIFICANT EFFECT TO THE TARGET VARIABLE.

STEPWISE REGRESSION

- Step 1 : First take all the independent variables and run the Annova Model to check whether the model is significant or not by looking at the probability of F-Statistics. If the probability of F-Statistics is smaller than 0.05 then the model is significant otherwise not.
- Step 2 : From the OLS table, the feature sex_male is not significant and hence is to be removed.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.606
Model:                            OLS   Adj. R-squared:                  0.603
Method:                 Least Squares   F-statistic:                     226.9
Date:                Wed, 02 Nov 2022   Prob (F-statistic):          1.19e-232
Time:                        07:06:58   Log-Likelihood:                -11712.
No. Observations:                1191   AIC:                         2.344e+04
Df Residuals:                    1182   BIC:                         2.349e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -2971.4610    818.000     -3.633      0.000   -4576.356   -1366.566
x[0]         244.6000      9.435     25.924      0.000     226.088     263.112
x[1]          69.1163     24.123      2.865      0.004      21.788     116.445
x[2]         434.0112    108.084      4.015      0.000     221.953     646.070
x[3]        -348.6055    262.750     -1.327      0.185    -864.114     166.903
x[4]        1.439e+04    428.642     33.578      0.000    1.36e+04    1.52e+04
x[5]        -293.5151    369.968     -0.793      0.428   -1019.383     432.353
x[6]       -1082.8273    381.272     -2.840      0.005   -1830.872    -334.782
x[7]       -1392.6524    376.421     -3.700      0.000   -2131.180    -654.124
==============================================================================
Omnibus:                      755.733   Durbin-Watson:                   2.054
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5449.081
Skew:                           3.046   Prob(JB):                         0.00
Kurtosis:                      11.527   Cond. No.                         325.
==============================================================================
```

STEPWISE REGRESSION

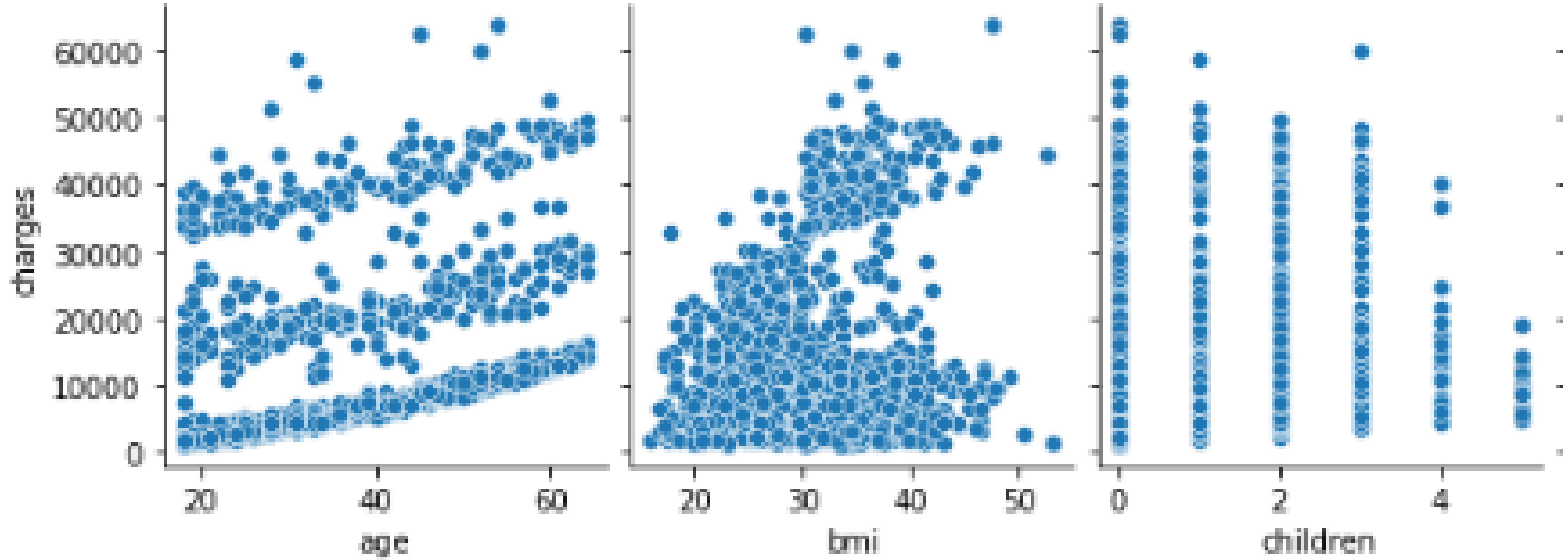
- **Step 3 : The region_northwest also has an insignificant p-value but region_southeast & region_southwest are significant and hence retained.**
- **Step 4: Run the model after removing gender variable**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.605
Model:                            OLS   Adj. R-squared:                  0.603
Method:                 Least Squares   F-statistic:                     258.9
Date:                Thu, 03 Nov 2022   Prob (F-statistic):          1.71e-233
Time:                        08:00:42   Log-Likelihood:                -11712.
No. Observations:                1191   AIC:                         2.344e+04
Df Residuals:                    1183   BIC:                         2.348e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -3127.4022    809.772     -3.862      0.000   -4716.152   -1538.652
x[0]         244.9016      9.436     25.955      0.000     226.389     263.414
x[1]          68.3008     24.123      2.831      0.005      20.973     115.629
x[2]         431.6618    108.105      3.993      0.000     219.564     643.760
x[3]        1.438e+04    428.706     33.548      0.000    1.35e+04    1.52e+04
x[4]        -293.3025    370.087     -0.793      0.428   -1019.403     432.798
x[5]       -1077.5957    381.374     -2.826      0.005   -1825.841    -329.351
x[6]       -1388.7256    376.530     -3.688      0.000   -2127.467    -649.984
==============================================================================
Omnibus:                      753.305   Durbin-Watson:                   2.052
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5398.767
Skew:                           3.035   Prob(JB):                         0.00
Kurtosis:                      11.482   Cond. No.                         322.
==============================================================================
```

TESTING FOR LINEARITY

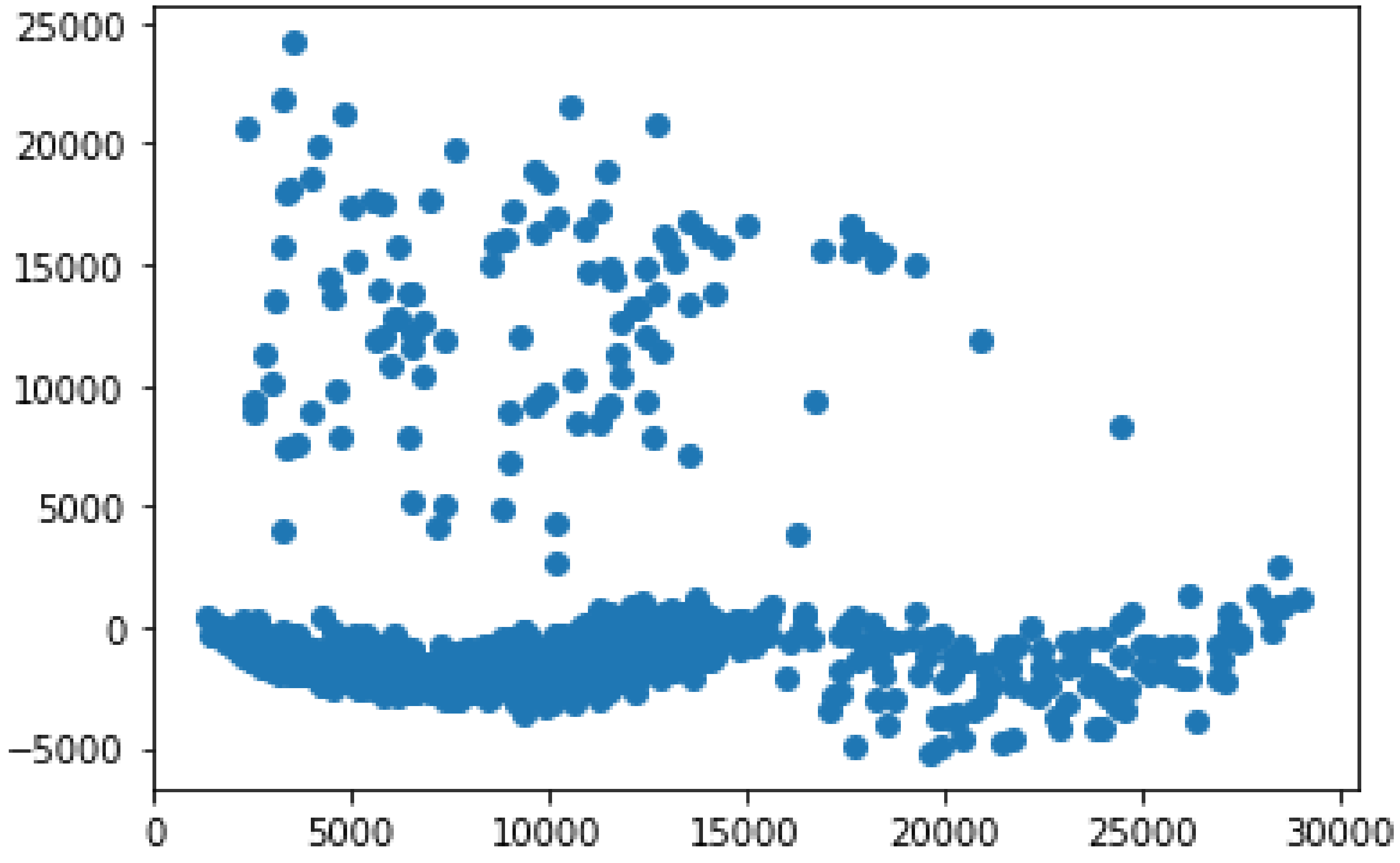
- **Checking of linearity between independent variables and dependent variable**
- **Scatter plot is created**

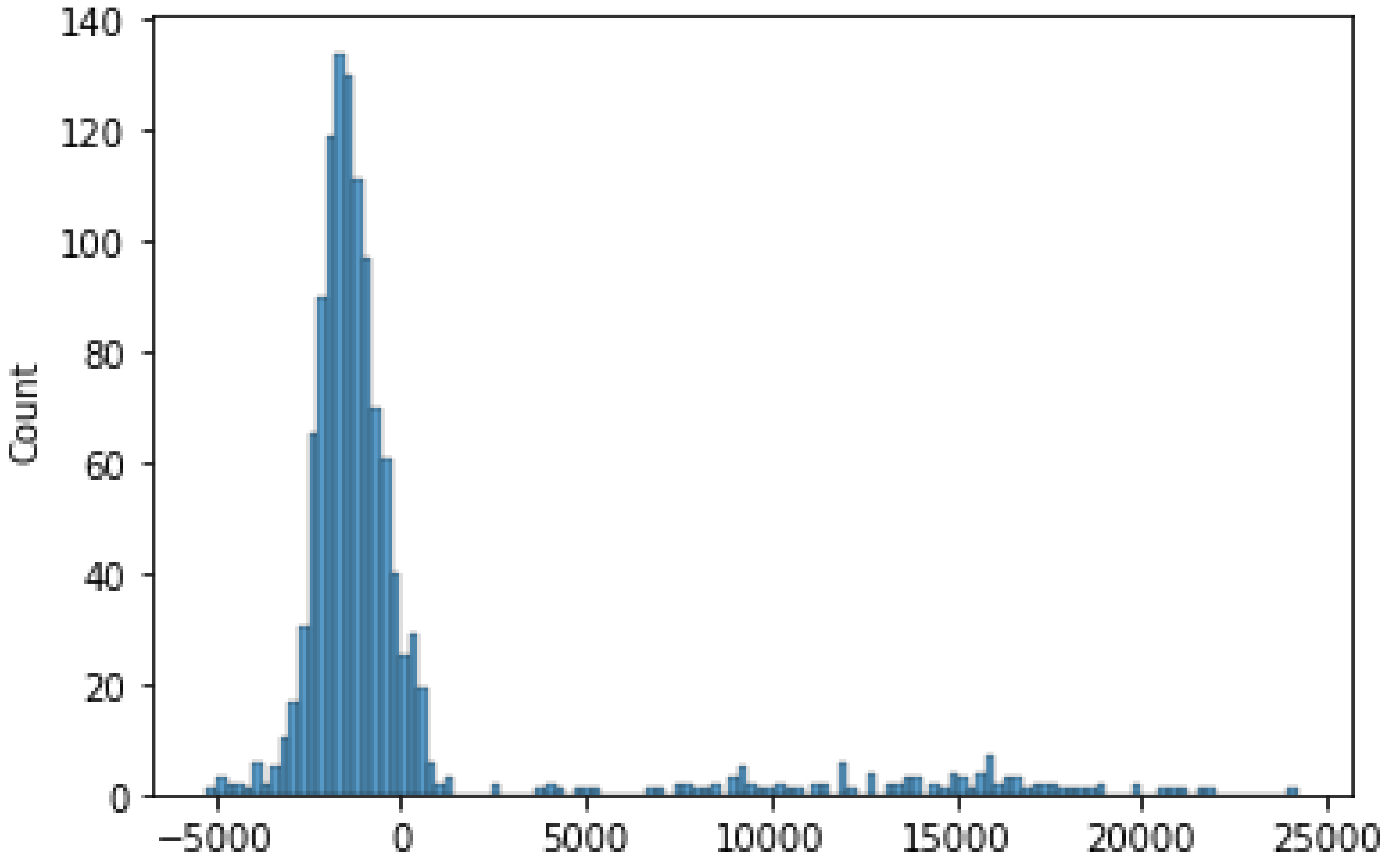
```
ndata = pd.read_csv("Insurance Data - Insurance Data.csv")
import seaborn as sns
sns.pairplot(ndata)
```



TESTING FOR NORMALITY OF RESIDUALS

- Normality of Residual–Multiple regression assumes that the residuals are normally distributed.

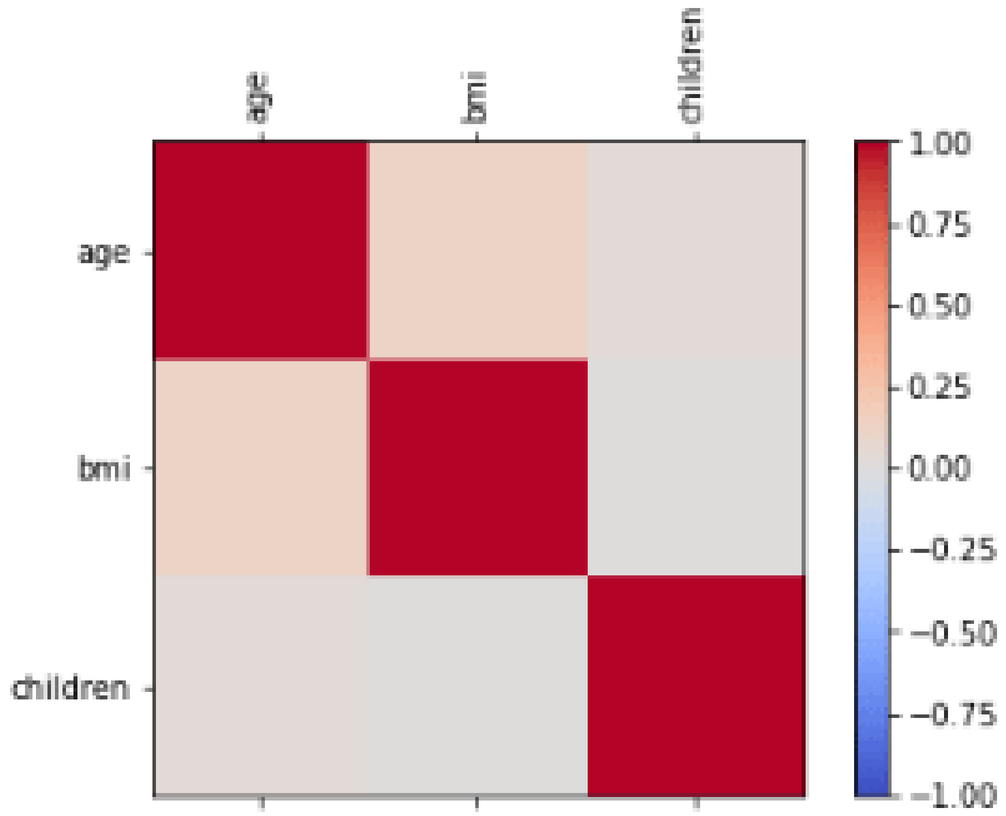




TESTING FOR MULTICOLLINEARITY

- **No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.**

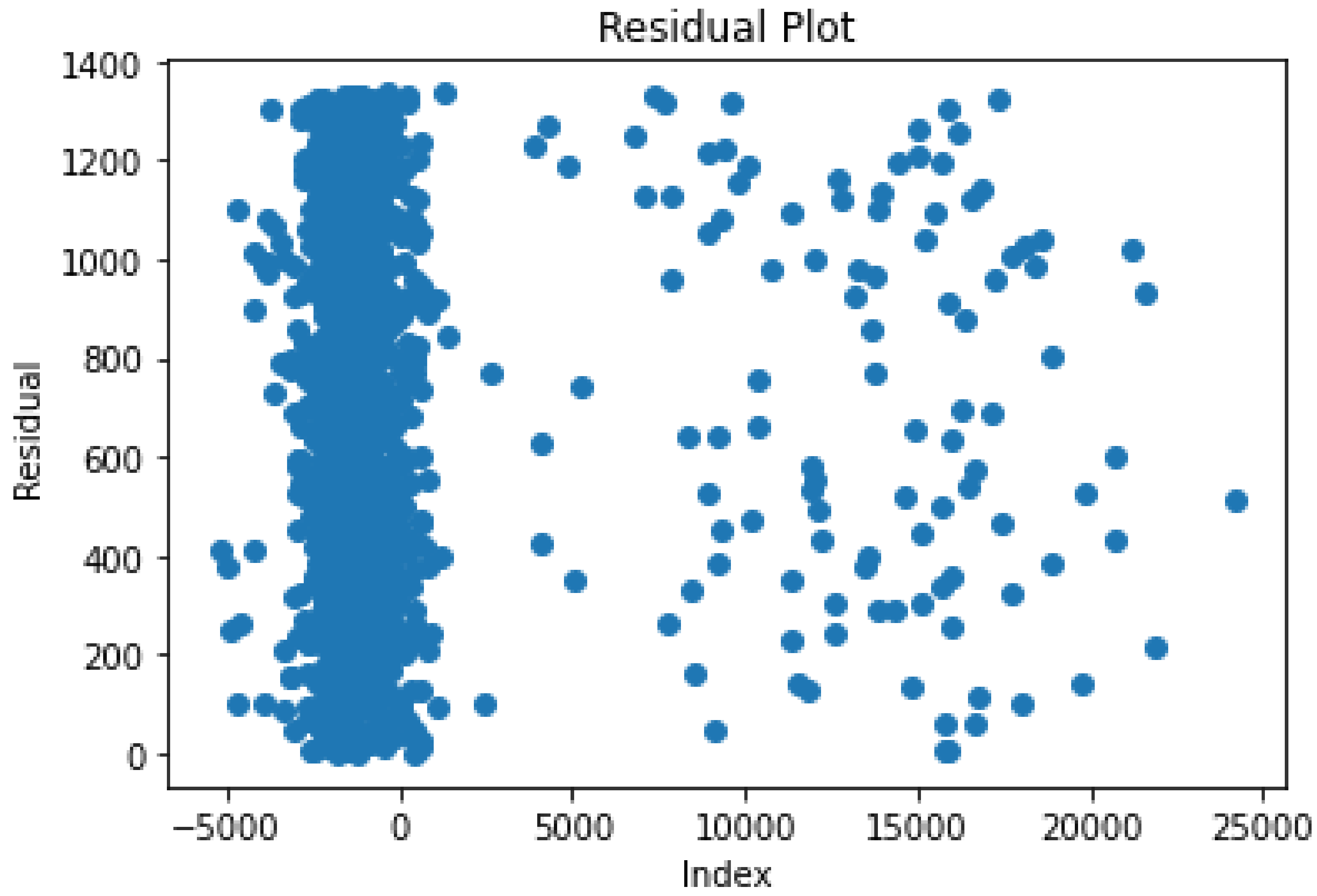
	age	bmi	children
age	1.000000	0.123827	0.038179
bmi	0.123827	1.000000	0.007546
children	0.038179	0.007546	1.000000



	feature	VIF
0	age	7.616749
1	bmi	7.935312
2	children	1.768840

TESTING FOR HOMOSCEDASTICITY

- **Homoscedasticity–This assumption states that the variance of error terms is similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables**



TESTING FOR AUTOCORRELATION

- **Durbin Watson test**

```
 statsmodels.stats.stattools import durbin_watson

form Durbin-Watson test
in_watson(MLR.resid)
ce this is within the range of 1.5 and 2.5, we would consider autocorrelation not to be problematic in this regression model.
```

Autocorrelation means the self relationship of errors

if durbinWatson < 1.5
Signs of positive autocorrelation', '\n')

if durbinWatson > 2.5:
Signs of negative autocorrelation

Omnibus:	753.305	Durbin-Watson:	2.052
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5398.767
Skew:	3.035	Prob(JB):	0.00
Kurtosis:	11.482	Cond. No.	322.