Regression Analysis Course Project

Insurance Premium Prediction

Guided By:
Monika Bhattacharjee
Assistant Professor, IIT Bombay

Problem Statement

Objective:

To build a MLRM that can accurately predict the Insurance Premium(charges) based on various predictors.

Goal:

Understand the impact of various predictors (age, sex, BMI, smoking status, number of children, and region) on the monthly insurance charge.

Emphasis:

Focus on model diagnostics and validation to draw robust conclusions about the relationship between these factors and insurance charges.

About the Dataset

The dataset, sourced from Kaggle, contains data on **1,338 individuals**, collected randomly to ensure statistical relevance.

Features:

Age: Age of individuals (in years)

Sex: Gender of individuals (1 for male, 0 for female)

BMI: Body Mass Index (kg/m^2)

Smoking Status: Smoking status (1 for smoker, 0 for non-smoker)

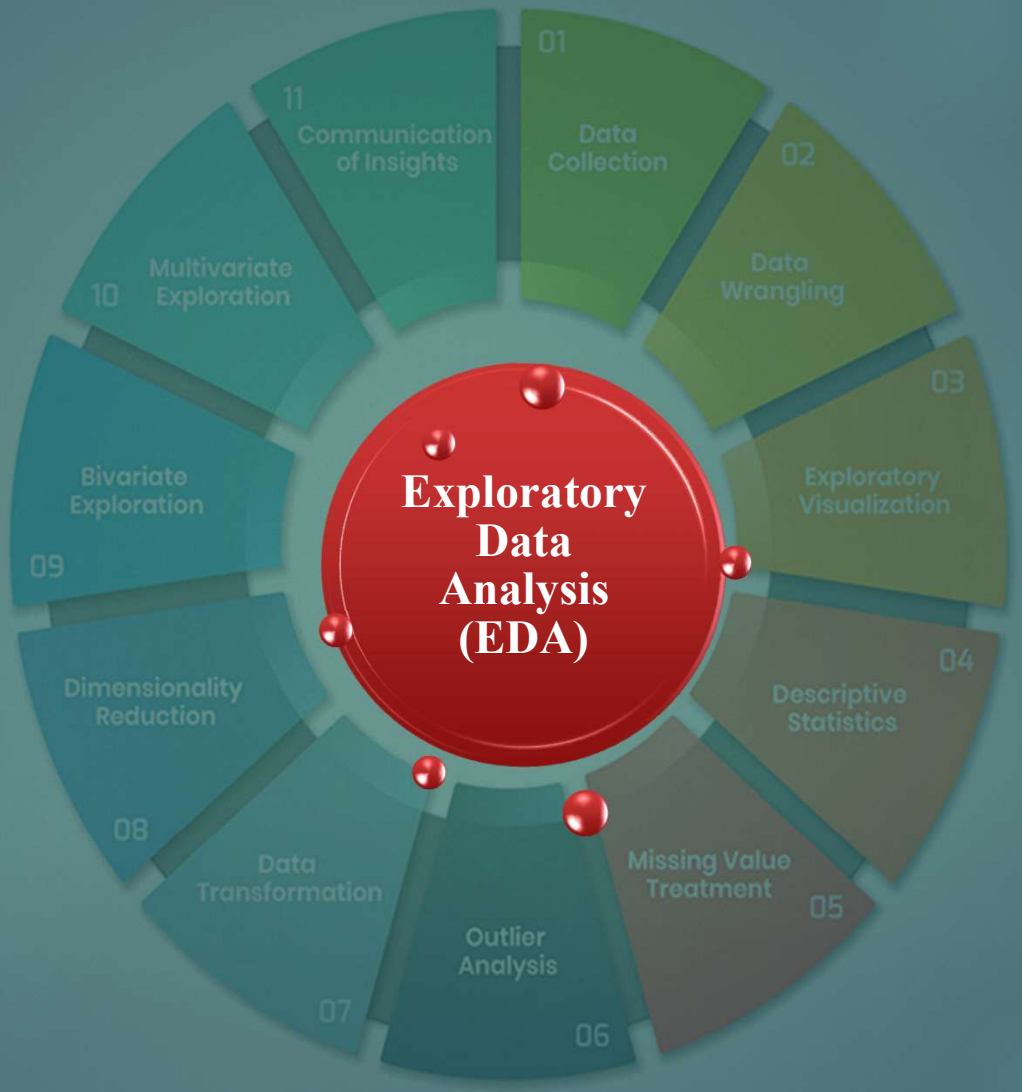
Number of Children: Number of children (0-5)

Region: Geographic region (northeast, northwest, southeast, southwest)

Monthly Insurance Charges: Monthly insurance charges paid (in rupees)

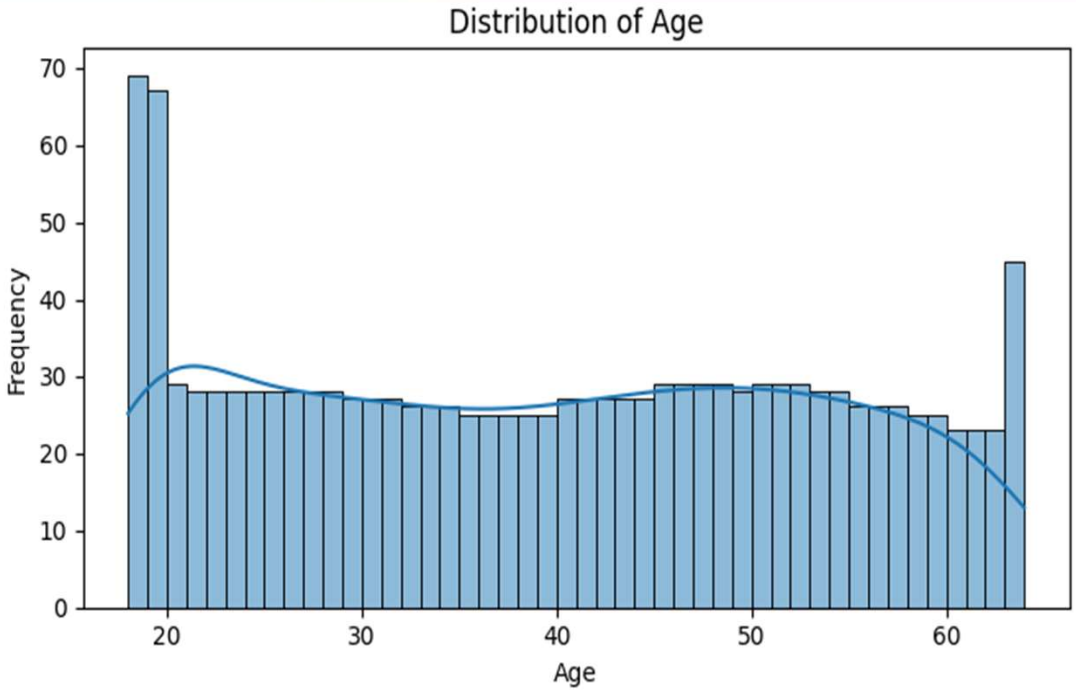
Glimpse of the Data

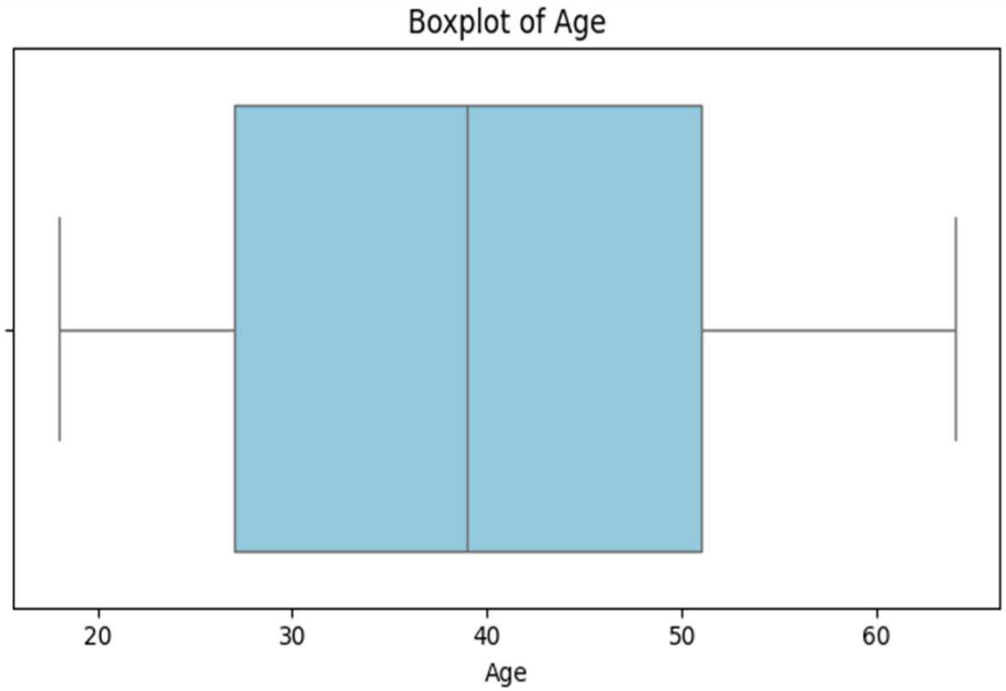
age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	1688.492
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	12984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southeast	3756.622
46	female	33.44	1	no	southeast	13240.59
37	female	27.74	3	no	northwest	7281.506
37	male	29.83	2	no	northeast	6406.411


Exploratory Data Analysis (EDA)
01
Data Collection
02
Data Wrangling
03
Exploratory Visualization
04
Descriptive Statistics
05
Missing Value Treatment
06
Outlier Analysis
07
Data Transformation
08
Dimensionality Reduction
09
Bivariate Exploration
10
Multivariate Exploration
11
Communication of Insights

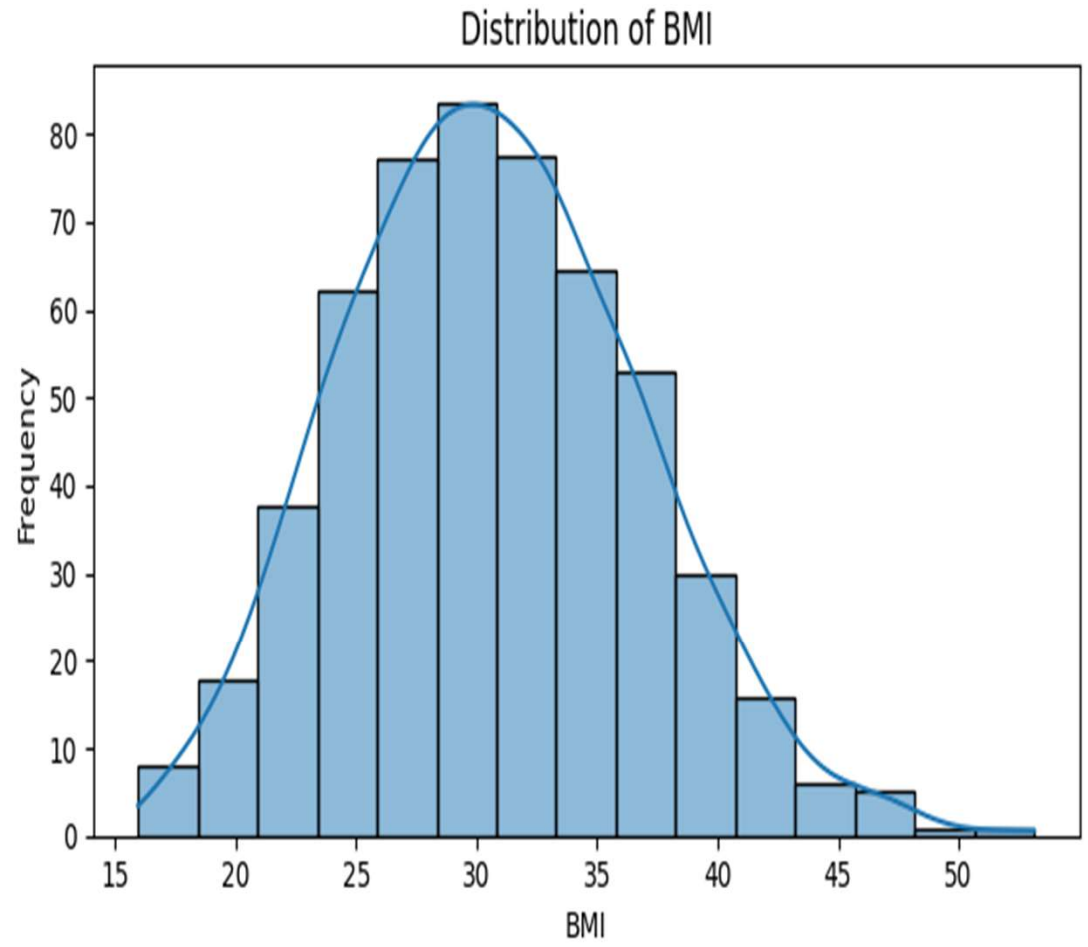
Univariate Data Analysis

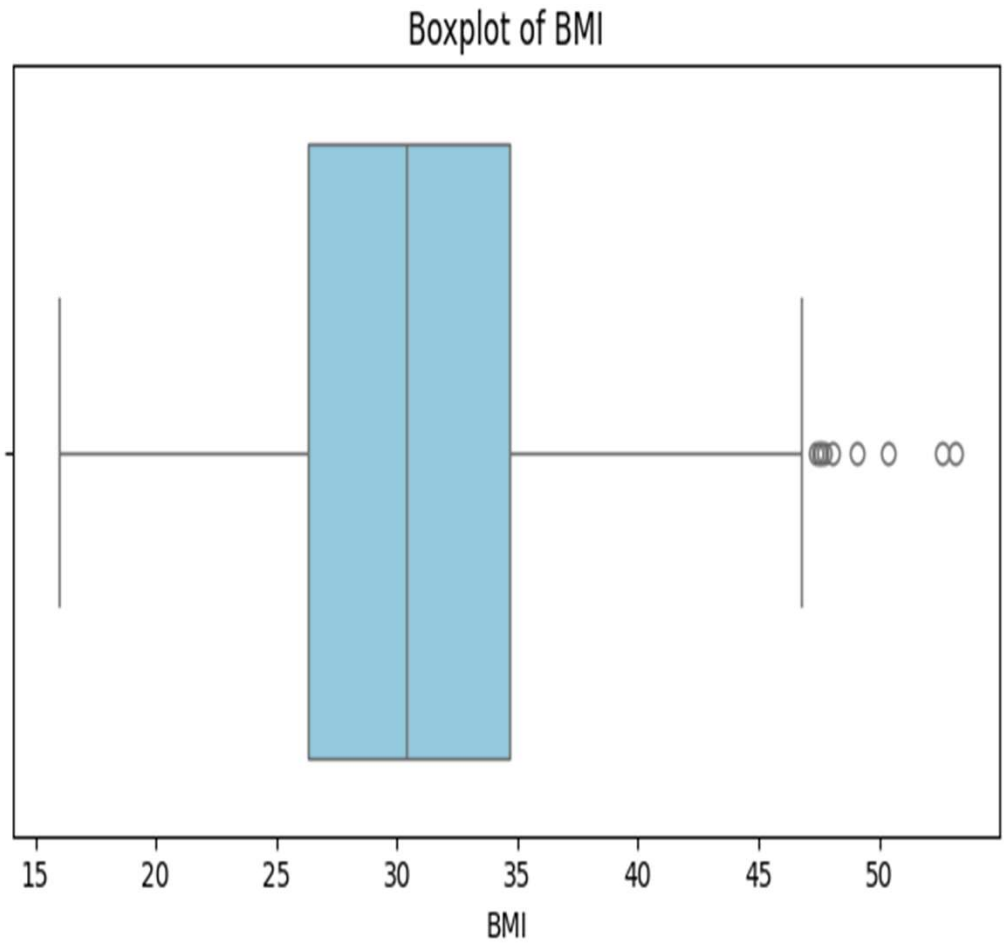
Age



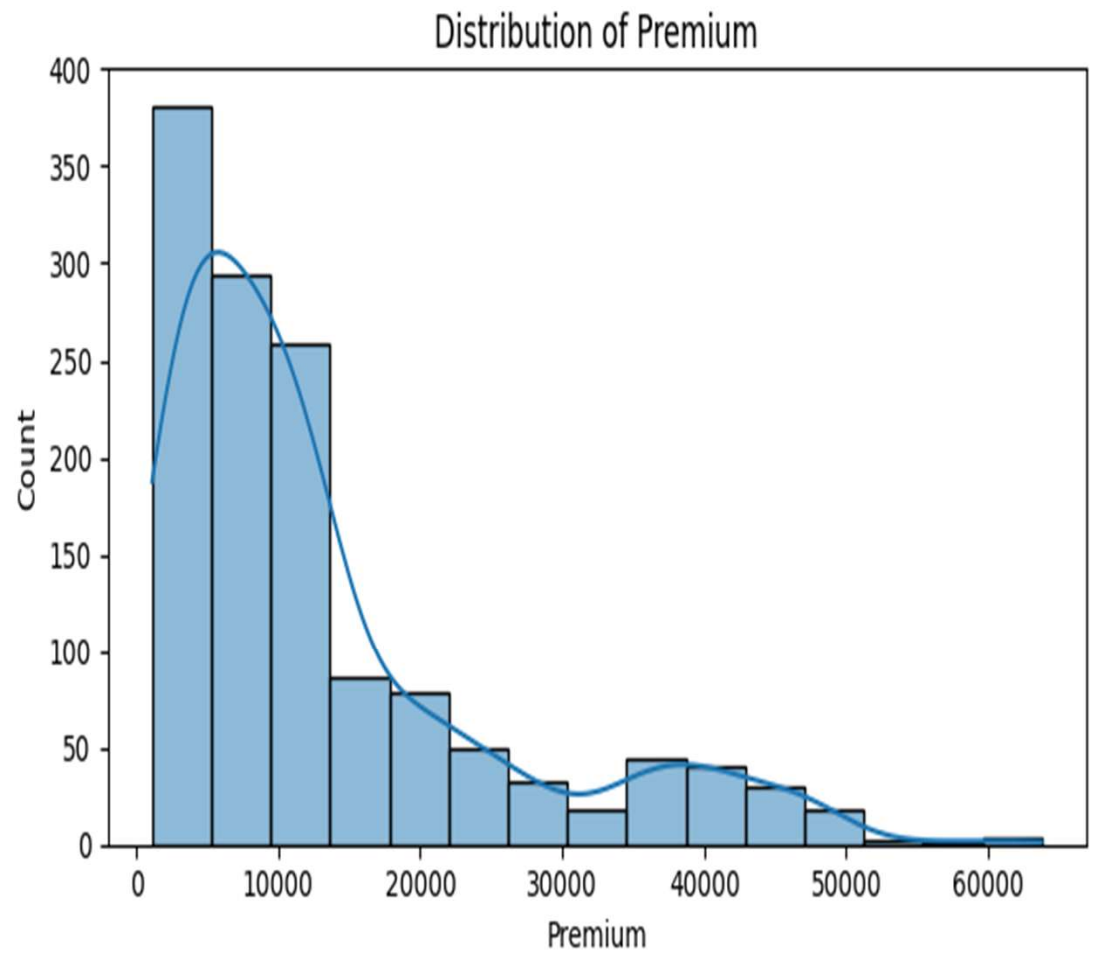


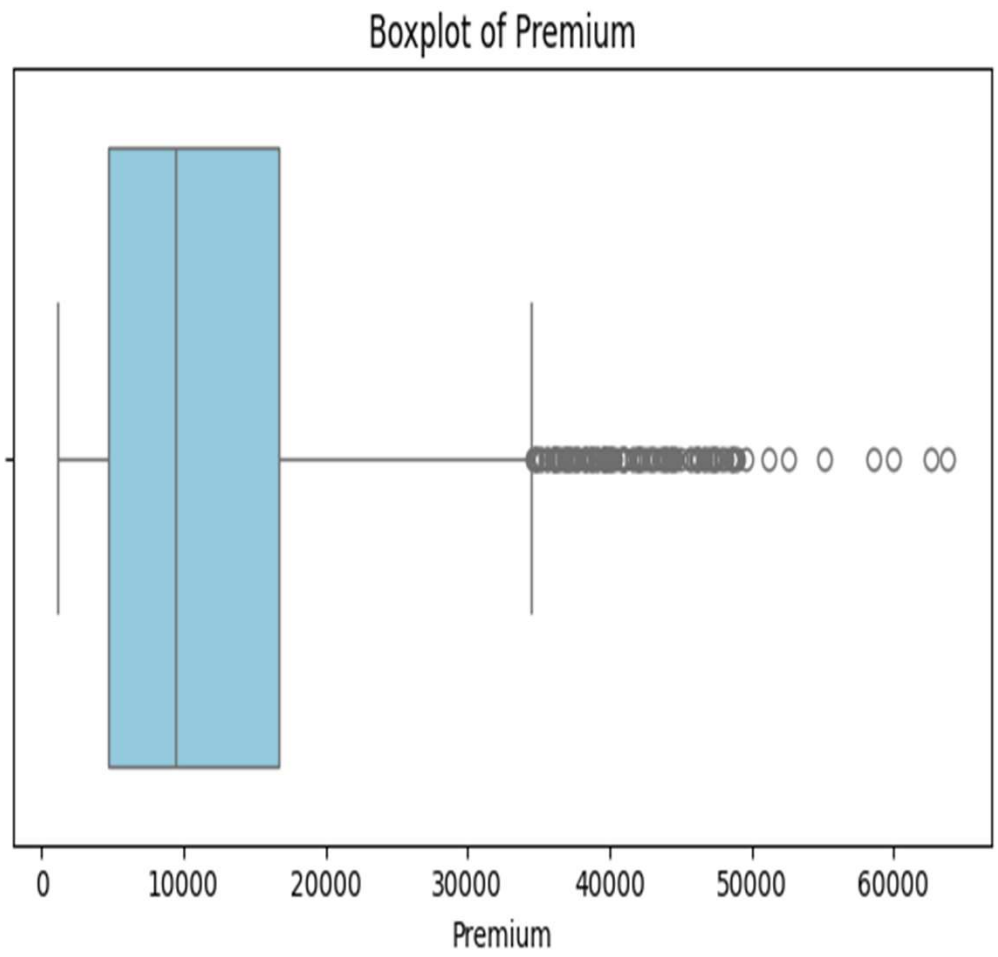
BMI


Distribution of BMI
Frequency
0
10
20
30
40
50
60
70
80
15
20
25
30
35
40
45
50
BMI


Boxplot of BMI
15
20
25
30
35
40
45
50
BMI

Charges


Distribution of Premium
Count
Premium
0
50
100
150
200
250
300
350
400
0
10000
20000
30000
40000
50000
60000


Boxplot of Premium
Premium
0
10000
20000
30000
40000
50000
60000

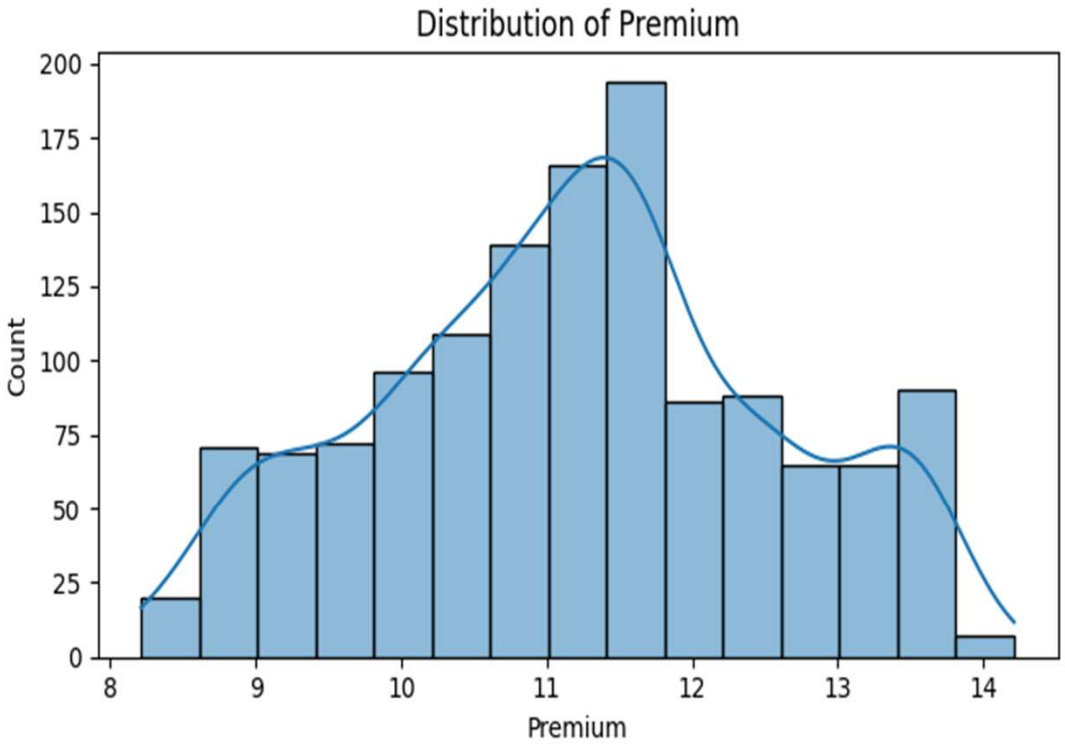
Box-Cox Transformation

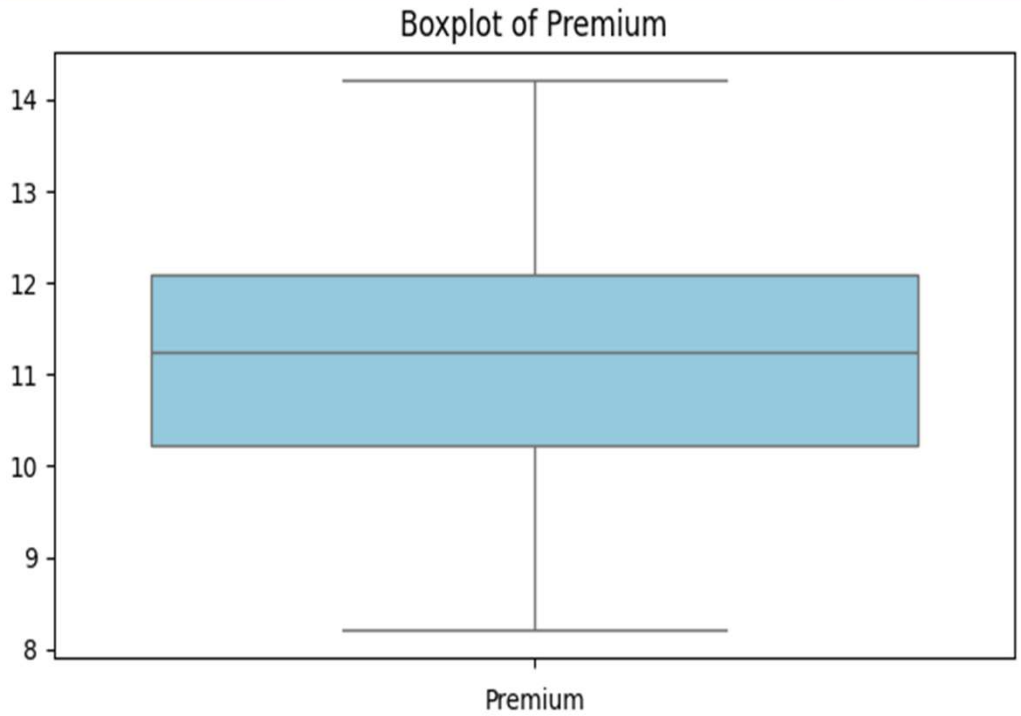
- The response variable is highly positively skewed with numerous outliers.We applied the Box-Cox transformation, aiming to make the response variable bell-shaped and symmetric.
- The subsequent slide displays the transformed response variable using the Box-Cox method.

If w is our transformed variable and "y" is our target variable, then the Box-Cox transformation equation looks like this:

$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ (y_t^{\lambda} - 1)/\lambda & \text{otherwise.} \end{cases}$$

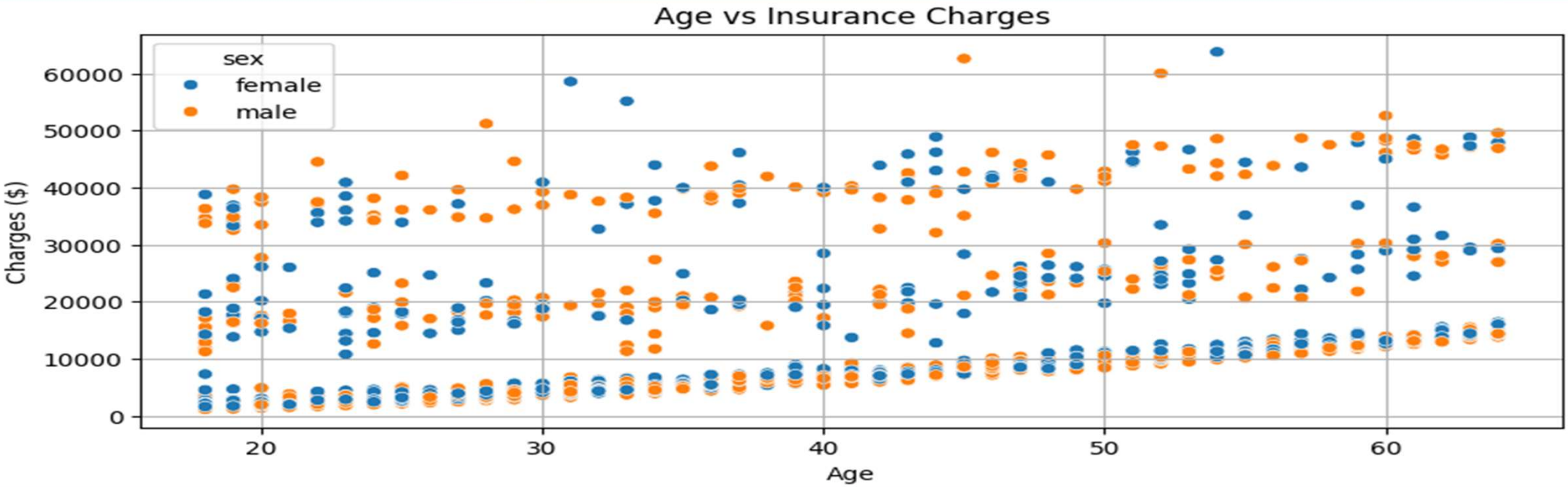
- After using Box-Cox transformation we got transformed charges with approximately 0 skewness.





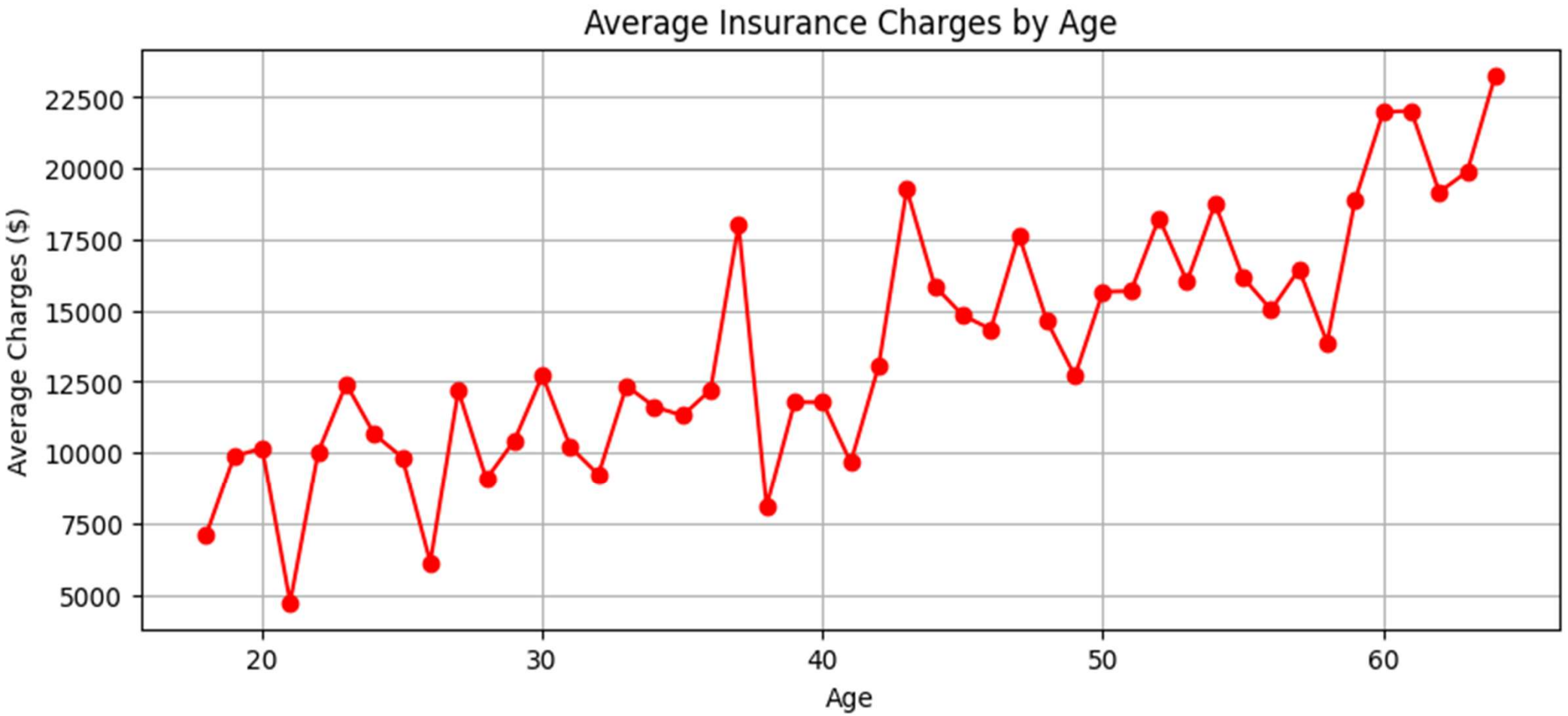
Bivariate Exploratory Data Analysis

Age vs Charges

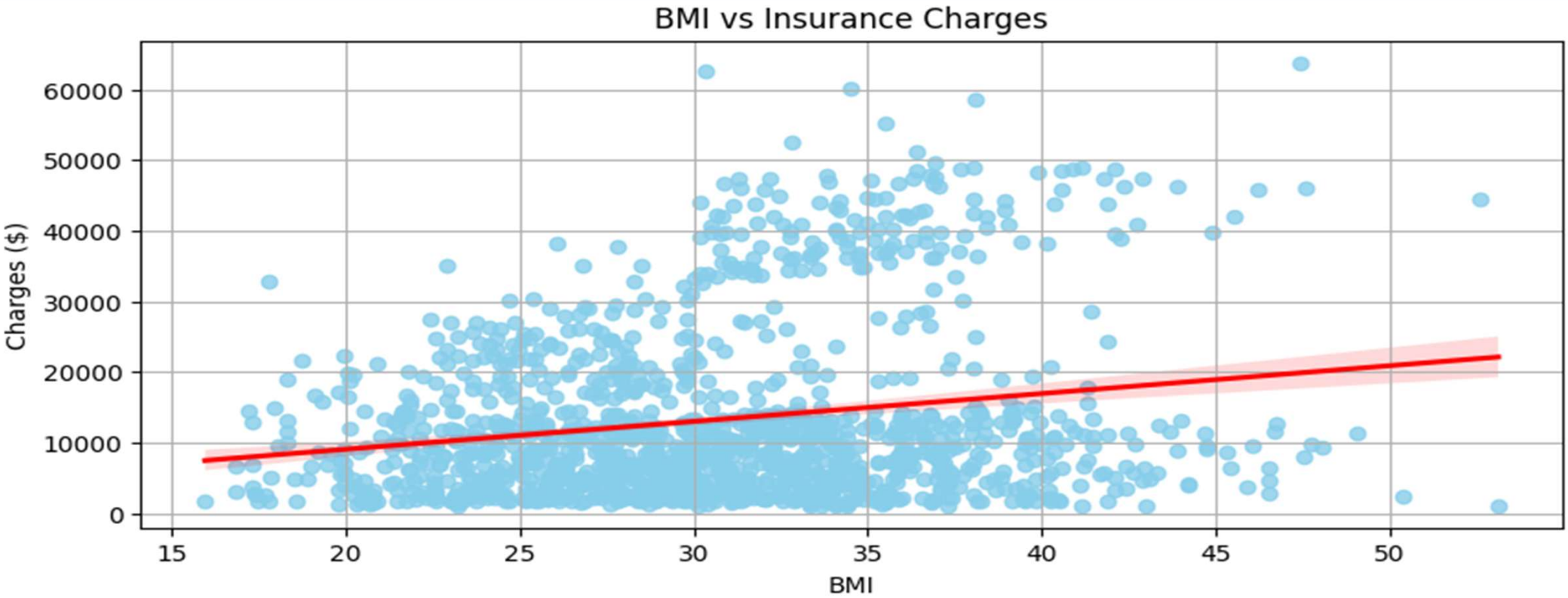


- The plot indicates a weak positive linear relationship between age and the response variable

Average Premium vs Age

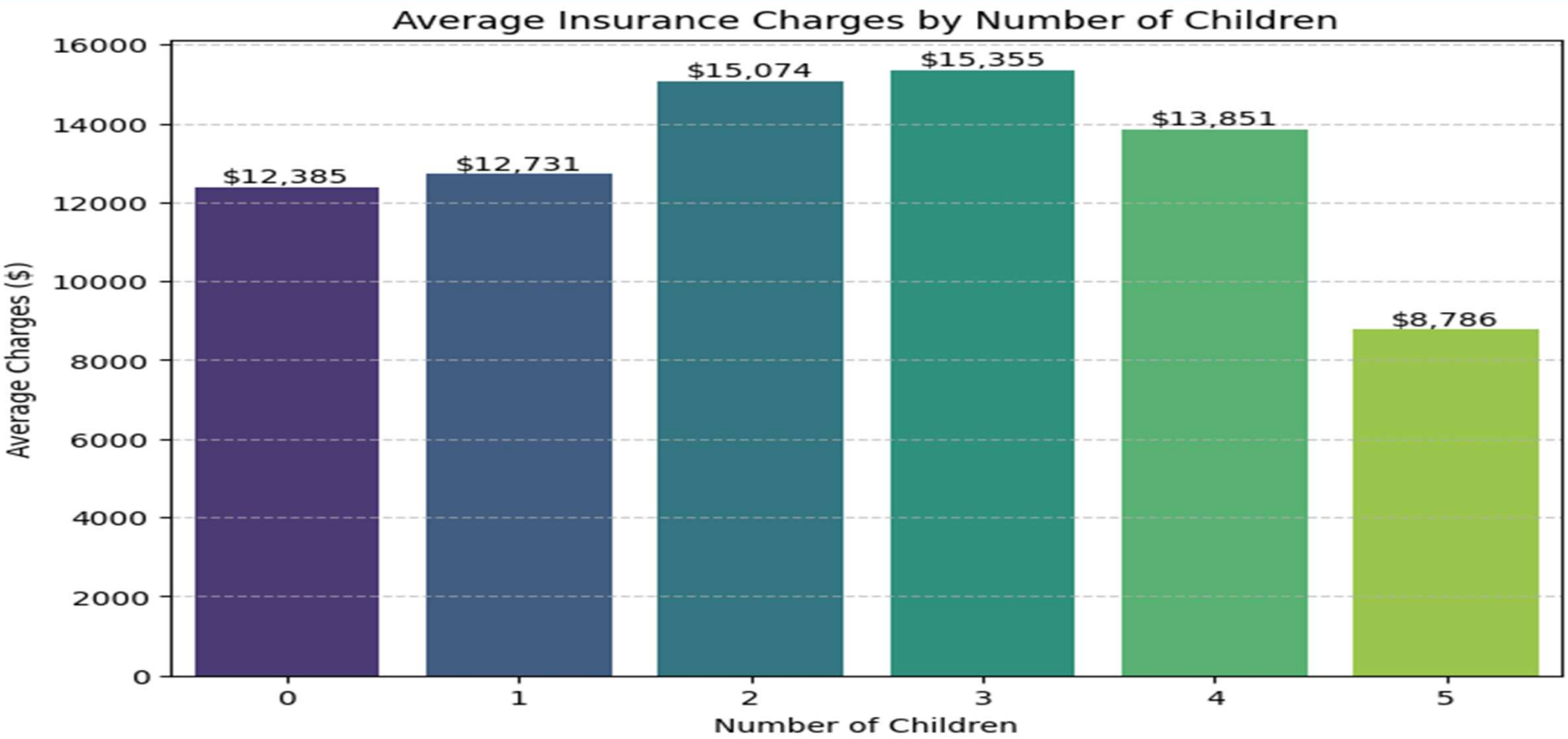


BMI vs Charges

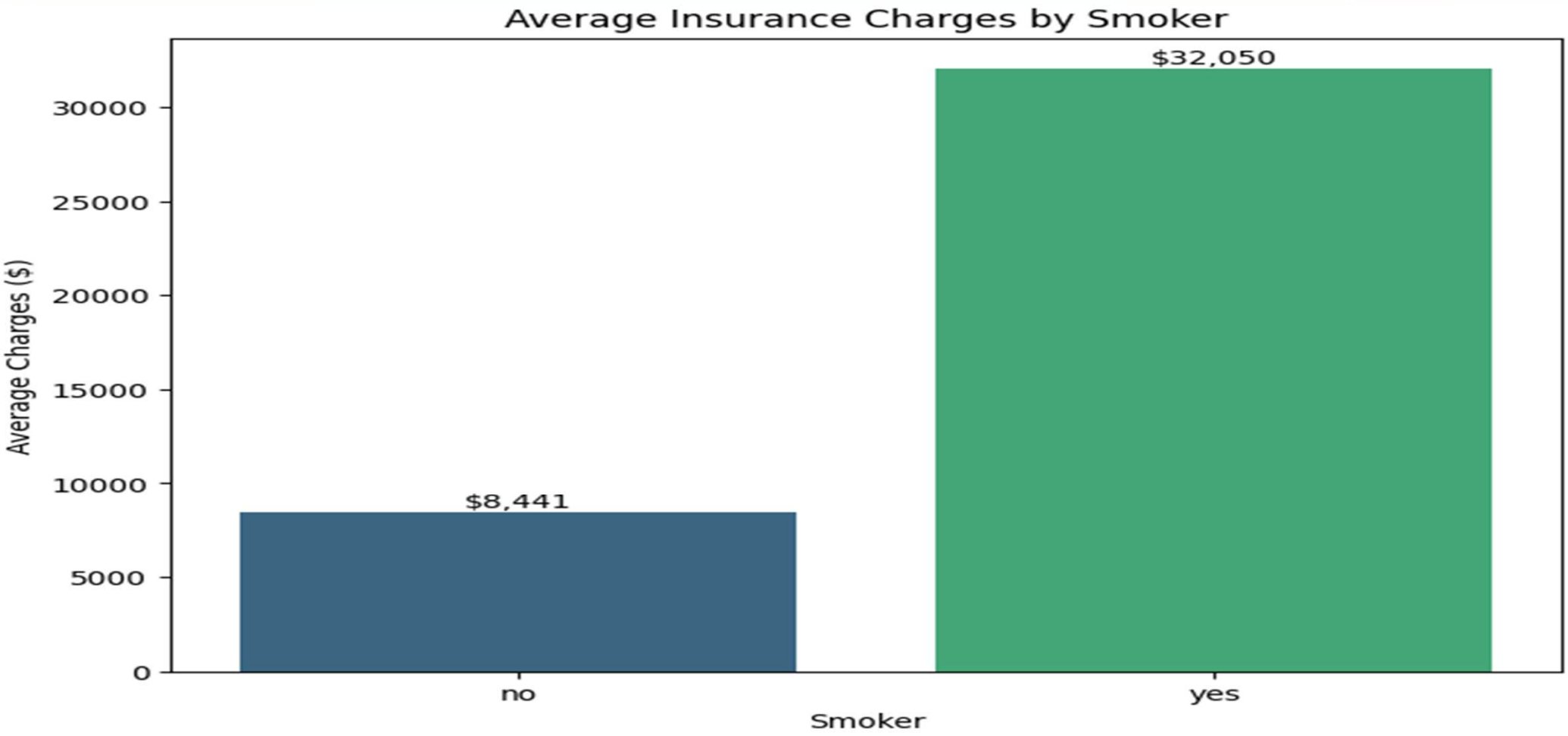


- The plot indicates a weak positive linear relationship between BMI and the response variable

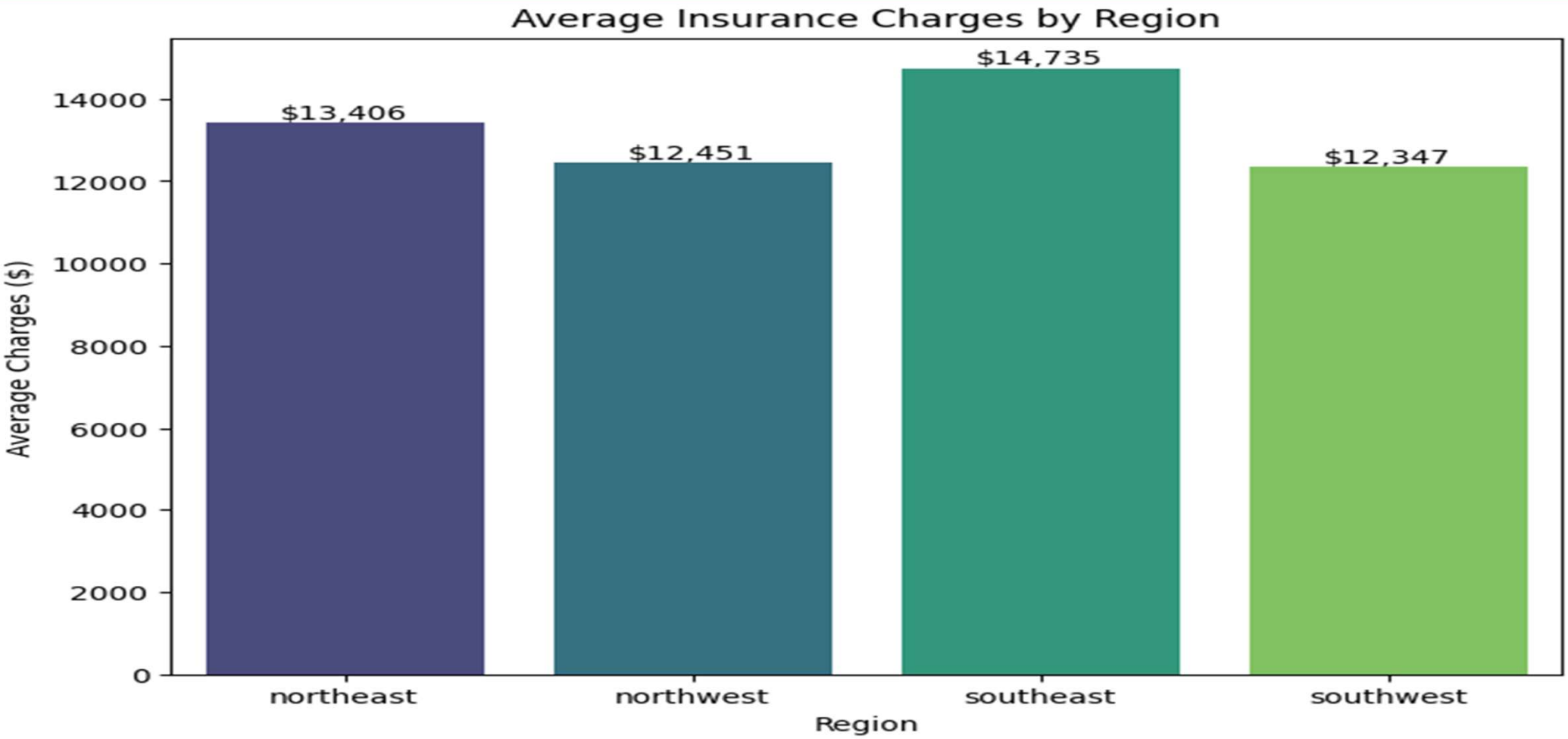
Number of children vs Charges



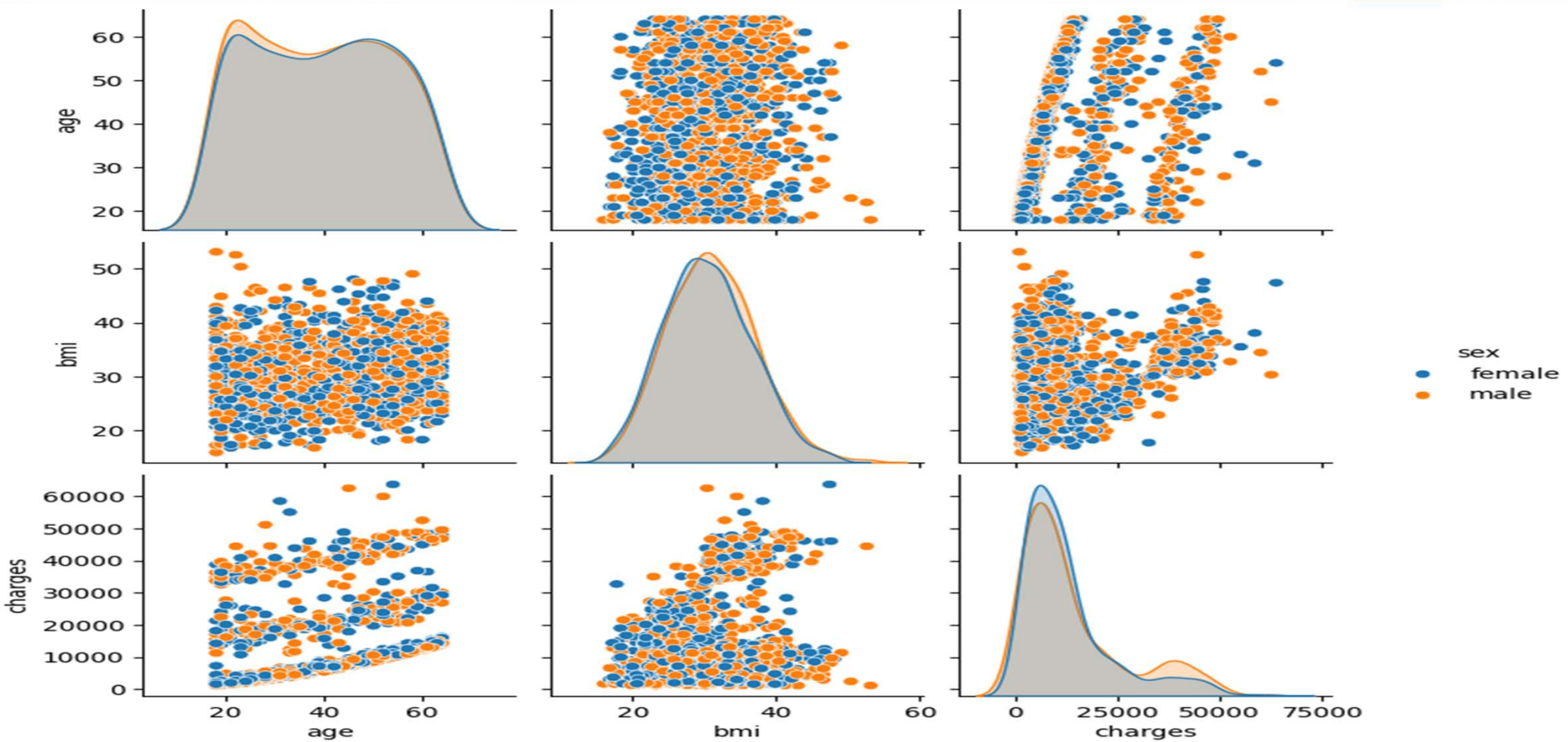
Smoker vs Charges



Region vs Charges



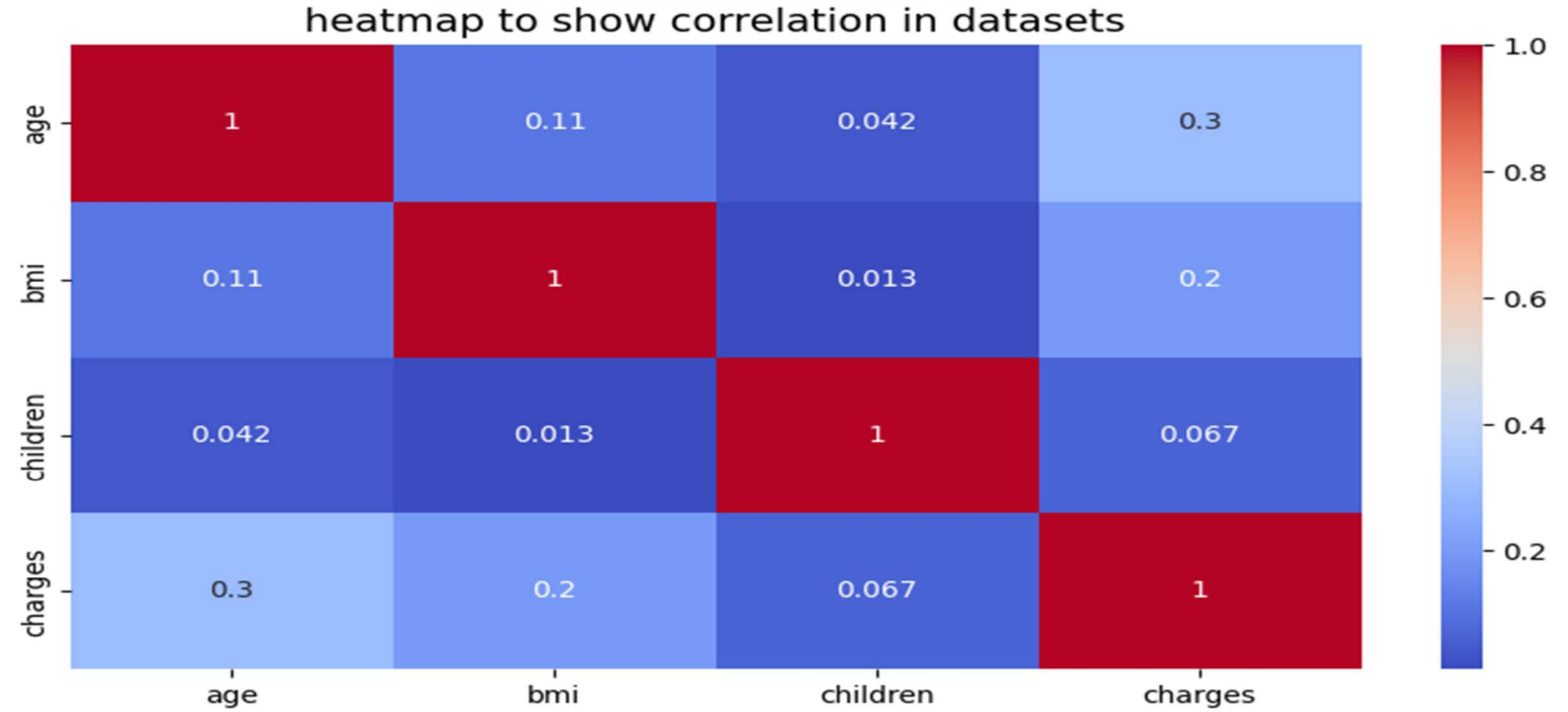
Pairplot


age
bmi
charges
sex
female
male
20
30
40
50
60
0
10000
20000
30000
40000
50000
60000
25000
75000

Pairplot Insights

The **pairplot** generated using Seaborn, visualizes the pairwise relationships between the variables **age**, **BMI**, and **charges**, with points colored by **sex**. This helps identify potential correlations, patterns, or group-wise differences across these variables. Diagonal plots show the distribution of each variable, while off-diagonal plots depict scatter plots between variable pairs, allowing for a quick visual inspection of linear or non-linear trends and how they may differ by gender.

Heatmap


heatmap to show correlation in datasets
age
bmi
children
charges
1
0.11
0.042
0.3
0.11
1
0.013
0.2
0.042
0.013
1
0.067
0.3
0.2
0.067
1
age
bmi
children
charges
1.0
0.8
0.6
0.4
0.2

- The highest correlation is between "age" and "charges" with a coefficient of **0.3**, indicating a moderate positive linear relationship. As age increases, insurance charges tend to increase as well.
- The correlation between "bmi" and "charges" is **0.2**, showing a weak positive linear relationship. Higher BMI values are associated with significantly higher charges.
- "Age" and "bmi" have a very weak correlation (**0.11**), suggesting that age does not strongly influence BMI in this dataset
- "Children" and "charges" show a negligible correlation (**0.067**), implying that the number of children has almost no linear impact on insurance charges


Data Preprocessing

Handling Categorical Features

Discrete numerical values of ‘children’ column are grouped into 3 levels :

- 0 children : **"No Children"** (Baseline)
- 1,2 children : **"Few Children"**
- >2 children : **"Many Children”**

- ’sex’ is encoded as a binary variable, where **1** indicates **male** and **0** indicates **female (**baseline).
- ’region has 4 levels: Northeast (baseline), Northwest,Southeast, Southwest .
- ’smoker’ is also encoded as a binary variable, where **1** indicates **smoker** and **0** indicated **non-smoker** .
- Then sex, smoker, region and new children column are one hot encoded and the baseline level for each variable is dropped to prevent multicollinearity, ensure meaning analysis and interpretation **.**

Handling Continuous Features

- The **numeric features** (including the **scaled response** column) are standardized using StandardScaler from sklearn.
- Standardization transforms numeric features to have a **mean** of **0** and a **standard deviation** of **1**, which not only **improves model performance** but also allows for **easier interpretation of regression coefficients** by putting **all features on a comparable scale**.
- Our final X contains scaled numeric features and one hot encoded categorical features, ready for modelling our final response column.

Model Building

- **Fitting the MLR model with transformed response and final predictors**

- Final predictors

- The final predictors are given by : Age, BMI, sex_male, smoker_yes, region_northwest, region_southeast,region_southwest,children_category_few_children and children_category_many children

- Final response

- Transformed & scaled ‘charge’ variable

Model Description

Formula and Calculation of Multiple Linear Regression (MLR)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Assumptions of MLRM

- **Linearity**:
  The relationship between the independent variables and the dependent variable is linear.
- **Independence**:
  The errors are independent. This implies no autocorrelation.
- **Homoscedasticity**:
  The residuals have constant variance across all levels of the independent variables (i.e., no heteroscedasticity).
- **Normality of Errors**:
  The errors are normally distributed. This is important for valid hypothesis testing and confidence intervals.
- **No Multicollinearity**:
  The independent variables are not highly correlated with each other. High multicollinearity can inflate standard errors and make estimates unreliable.

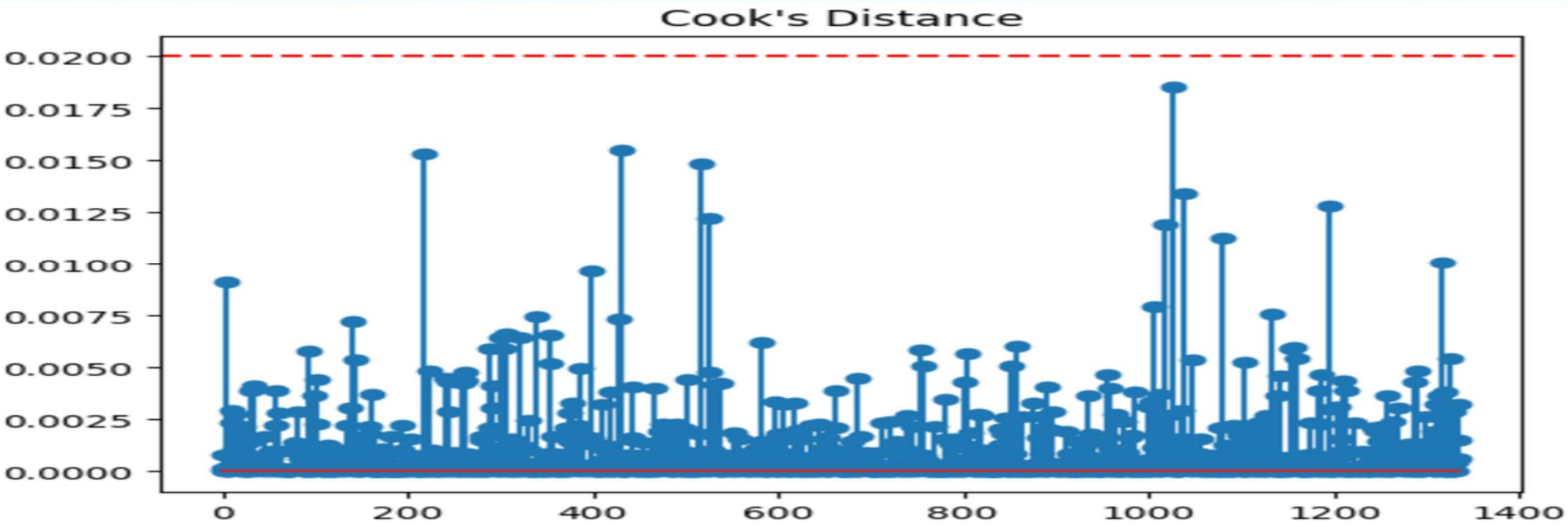
Initial Model Building

- Initially fitted MLR model on entire data to assess the HAT Matrix for finding out Leverage Points(points having unusual X coordinate).
- Found three Leverage Points, now proceeded for checking possible influential points in the entire dataset using Cook D.

- Cook Distance : $$D_i = \frac{r_i^2}{p} \cdot \frac{h_{ii}}{1 - h_{ii}}$$

- Where r_i is the studentized or Standardized residuals.

- h_{ii} is the Leverage of observation i (from the Hat Matrix).
- p is the number of parameters(including intercept).
- **Rule Of Thumb :** We consider a point as influential point if D_i >1
- For the entire dataset we got all the D_i < 0.02<<1 so no influential points.



- The dataset is splitted into **training (80%)** and **testing (20%)** sets with a fixed random state for reproducibility.
- A **constant 1 vector** is added to X_train to include the intercept in the regression model.
- An **OLS regression model** is fitted using the training data.
- The model is trained to **minimize the sum of squared errors**.
- The key metrics are displayed like **coefficients**, **R-squared**, **p-values**, **confidence intervals etc** for model evaluation.

- Also added a **constant 1** vector to X_test to match the training model.
- Predicted values (y_pred) are computed using the test set.
- **Residuals** are calculated as the difference between actual and predicted values.
- Evaluated model performance using:
  - **MSE** – average squared prediction error
  - **RMSE** – error in the same unit as the target
  - **R-squared** – variance explained by the model
  - **Adjusted R-squared** – accounts for number of predictors
- These metrics indicate how well the model performs on **unseen data**.

Performance on Train set

OLS Regression Results

Dep. Variable:	y	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.748
Method:	Least Squares	F-statistic:	353.2
Date:	Tue, 15 Apr 2025	Prob (F-statistic):	1.48e-311
Time:	10:07:45	Log-Likelihood:	-761.19
No. Observations:	1069	AIC:	1542.
Df Residuals:	1059	BIC:	1592.
Df Model:	9		
Covariance Type:	nonrobust		

Fitted Model Parameters

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.3584	0.040	-9.005	0.000	-0.437	-0.280
age	0.5099	0.015	33.104	0.000	0.480	0.540
bmi	0.0889	0.016	5.519	0.000	0.057	0.121
sex_male	-0.0796	0.030	-2.611	0.009	-0.139	-0.020
smoker_yes	1.6976	0.038	44.420	0.000	1.623	1.773
region_northwest	-0.0488	0.044	-1.118	0.264	-0.135	0.037
region_southeast	-0.1387	0.044	-3.128	0.002	-0.226	-0.052
region_southwest	-0.1075	0.044	-2.455	0.014	-0.193	-0.022
children_category_Few Children	0.2005	0.033	6.105	0.000	0.136	0.265
children_category_Many Children	0.2966	0.046	6.417	0.000	0.206	0.387

10-Fold Cross Validation

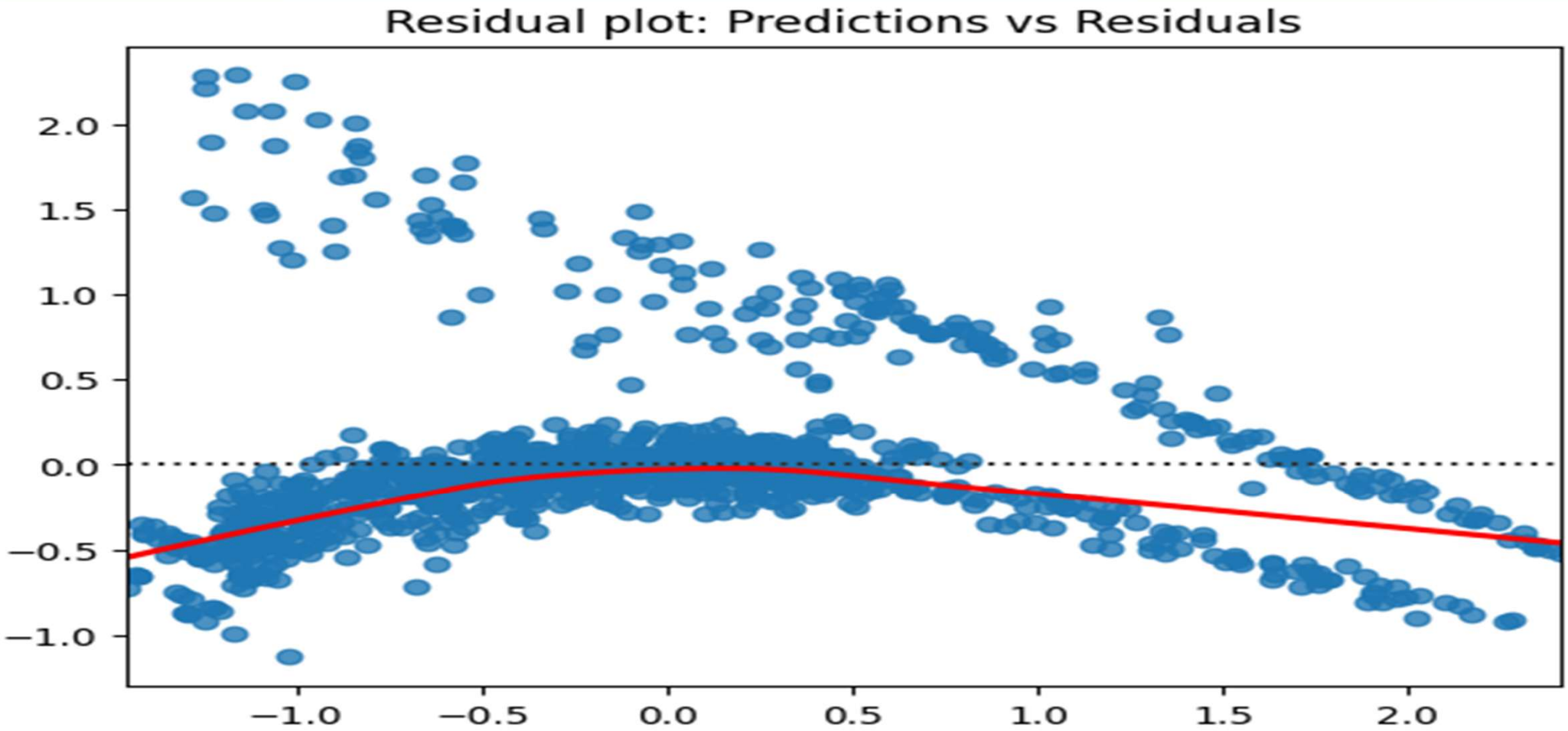
- Used **10-fold cross-validation** to evaluate model stability across different data splits.
- **KFold** ensures data is shuffled and splitted into 10 equal parts.
- Computed **R^2 scores** for each fold to assess model performance.
- **Adjusted R^2** is calculated for each fold to account for number of predictors.
- Printed **R^2 scores**, **average R^2**, and **average Adjusted R^2** to summarize model performance.

Performance on 10-fold Cross Validation

- The average of adjusted **R-squared** came out to be 76% which validates the consistent performance of the model.

```
R^2 scores for each fold: [0.84776642 0.81489595 0.69651047 0.74459442 0.78322051 0.72316062
 0.7192197  0.75361519 0.76542915 0.77844902]
Average R^2: 0.7626861431415791
Average Adjusted R^2: 0.761076629417595 7
```

Residual Plot



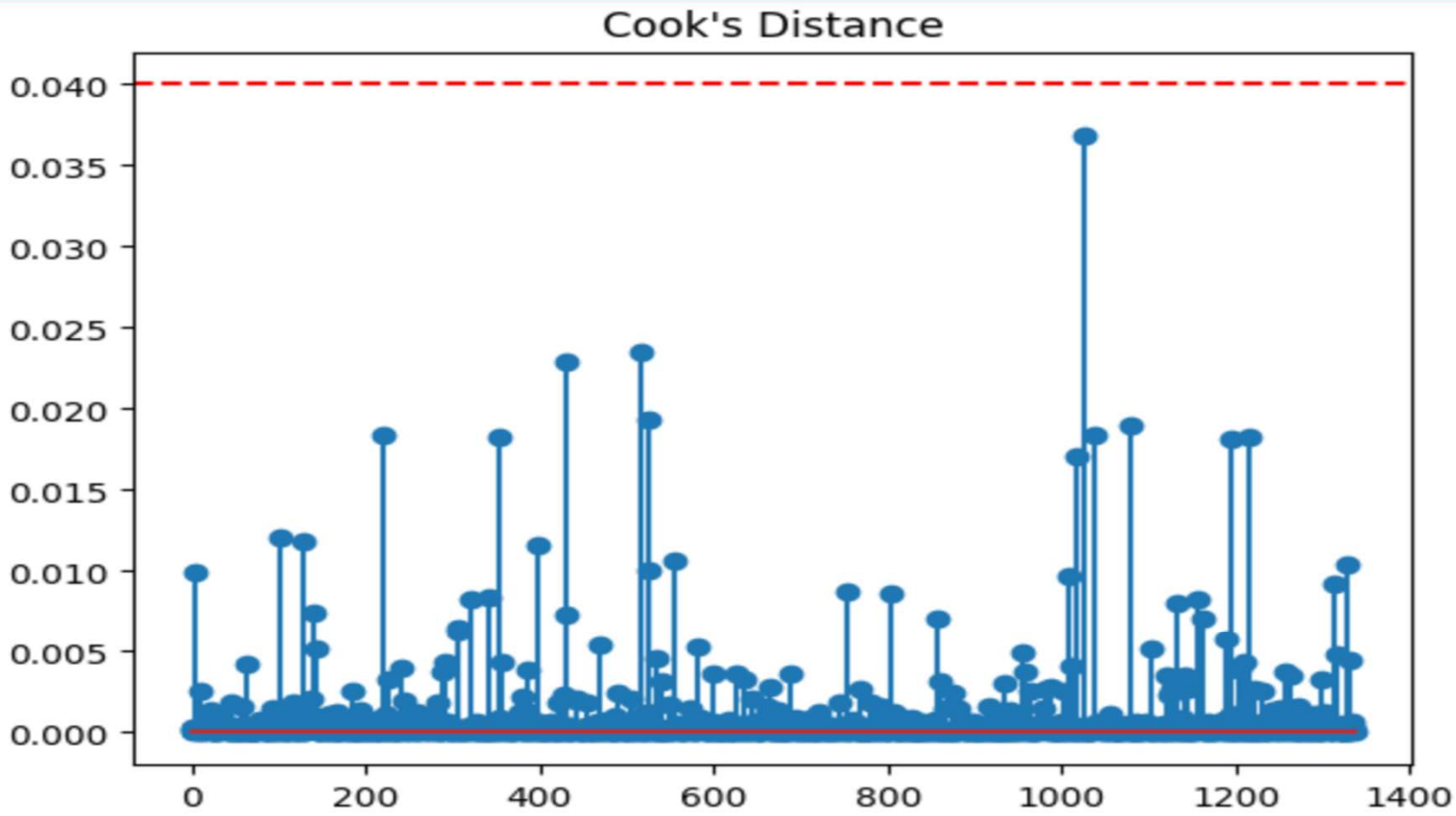
Residual Plot Insights

- The residuals should be **randomly scattered** around zero, without any clear pattern which ensures :
  - Linear relationship between the predictors and the response.
  - Homoscedasticity (constant variance of errors)
- We got a curved pattern in the residual plot which indicates a non-linear relationship, between predictors and response variable, thus we cannot validate the assumptions of MLRM.

Stepping Towards Polynomial Regression

- Now we perform multiple linear regression again but this time we use additional polynomial features '**age^2**', '**bmi^2**', **interactions** between '**age**' and '**bmi'**, '**age**' and '**smoker**' & '**bmi**' and '**smoker**' .
- Thus we have 14 predictors at hand to model the response variable.
- We fit a multiple linear regression model (with intercept) on the entire data available and find Cook's distances (for each 14-dimensional datapoint) to check for influential points, as done previously.

Inquiring for Influential Points



Since all points have Cook's D < 0.04 << 1, no influential point is detected in the dataset.

Training Polynomial Model

- Now we train the new model and test its performance on the same train-test splits to ensure consistency of model improvement with additional features.
- Now, we study the model's performances on train and test sets respectively.

Performance on Train set

OLS Regression Results

Dep. Variable:	y	R-squared:	0.812
Model:	OLS	Adj. R-squared:	0.810
Method:	Least Squares	F-statistic:	325.4
Date:	Tue, 15 Apr 2025	Prob (F-statistic):	0.00
Time:	10:11:47	Log-Likelihood:	-608.74
No. Observations:	1069	AIC:	1247.
Df Residuals:	1054	BIC:	1322.
Df Model:	14		

Model Parameters

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2968	0.042	-7.083	0.000	-0.379	-0.215
A	0.6159	0.015	40.769	0.000	0.586	0.646
B	0.0106	0.016	0.670	0.503	-0.021	0.042
S	1.6859	0.033	50.511	0.000	1.620	1.751
A^2	-0.0309	0.016	-1.882	0.060	-0.063	0.001
AB	0.0037	0.014	0.270	0.787	-0.023	0.030
AS	-0.5011	0.033	-15.056	0.000	-0.566	-0.436
B^2	-0.0325	0.010	-3.262	0.001	-0.052	-0.013
BS	0.3708	0.033	11.199	0.000	0.306	0.436
sex_male	-0.0875	0.027	-3.285	0.001	-0.140	-0.035
region_northwest	-0.0474	0.038	-1.244	0.214	-0.122	0.027
region_southeast	-0.1139	0.039	-2.946	0.003	-0.190	-0.038
region_southwest	-0.1401	0.038	-3.674	0.000	-0.215	-0.065
children_category_Few Children	0.2030	0.030	6.674	0.000	0.143	0.263
children_category_Many Children	0.3109	0.041	7.494	0.000	0.229	0.392

Performance on Test Set

- The **adjusted R-squared** came out to be 87% which validates that model performance is good for unseen data.

```
Testing Set:
Mean Squared Error: 0.13644084060337078
Root Mean Squared Error: 0.3693789931809479
R-squared: 0.8765211645596273
Adj_R-squared: 0.8691712338786527
```

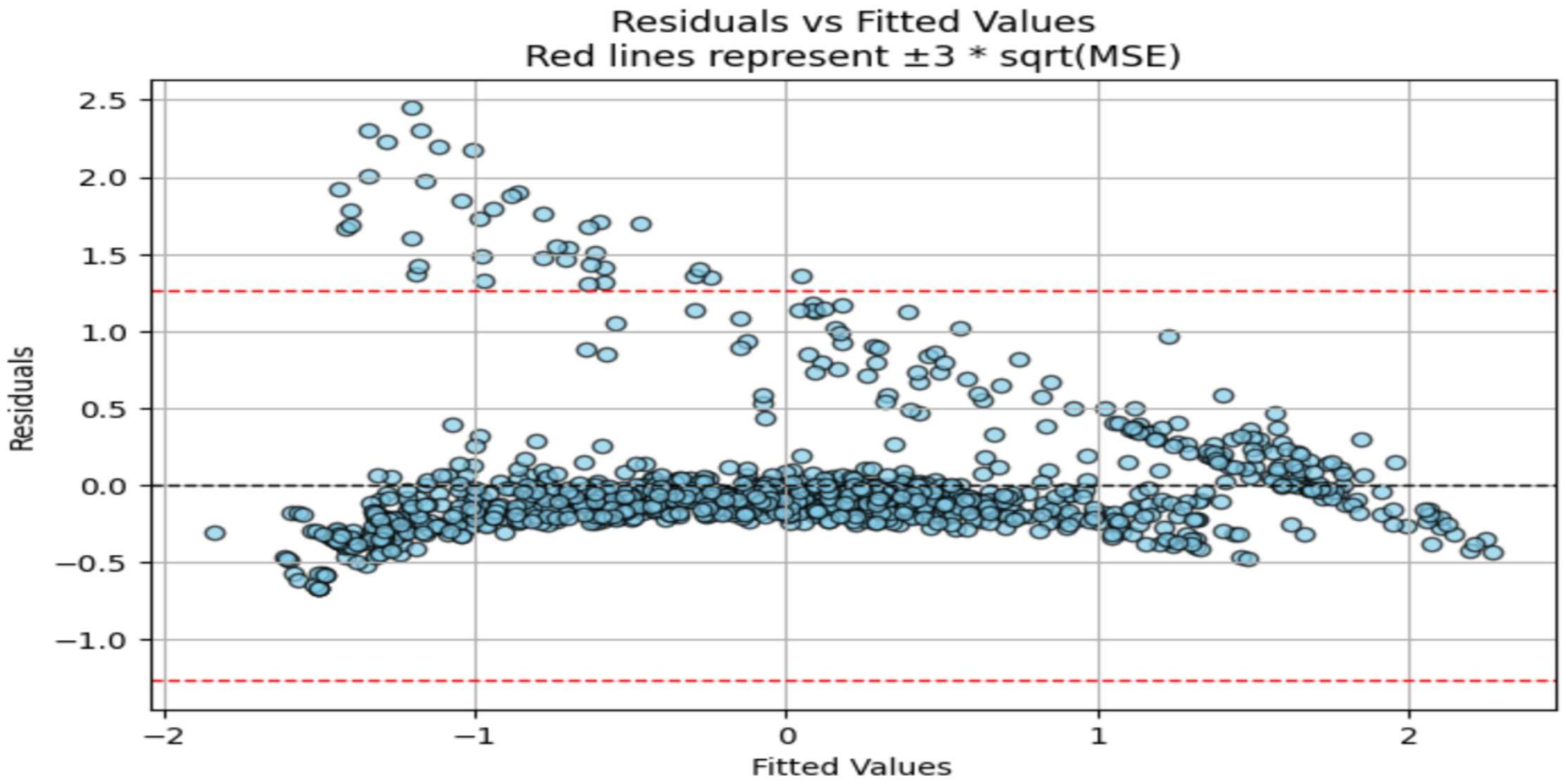
Performance on 20-fold Cross Validation

- The average of **adjusted R-squared** came out to be 82% which validates the consistent performance of the model.

```
R^2 scores for each fold: [0.92956819 0.88455546 0.84688018 0.83982085 0.84273994 0.61164033
 0.71808036 0.93083585 0.84930256 0.86408629 0.78085111 0.73991856
 0.77704718 0.8299653  0.84094719 0.81452668 0.86437806 0.78276364
 0.76309466 0.89532233]
Average R^2: 0.8203162363708021
Average Adjusted R^2: 0.8190975823597524
```

After cross validation, 82% Mean Adj. R^2 is quite good.

Residual Plot



Residual Plot Insights

- The residuals should be **randomly scattered** around zero, without any clear pattern to validate :
  - Linear relationship between the predictors and the response.
  - Homoscedasticity (constant variance of errors)

- We got a curved pattern in the residual plot which indicates a non-linear relationship, thus we cannot validate the assumptions of MLRM.
- We do not have enough evidence to validate homoscedasticity and linearity even in the improved model.

Best Model Selection

- Used AIC,BIC and Mallows Cp for selection of best subset of predictors.
- **Formula:**

$$AIC = 2k - 2\ln(L)$$

$$BIC = k\ln(n) - 2\ln(L)$$

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p)$$

- Where k = number of parameters (including intercept),L = likelihood of the model,n = number of observations
- SSE_p = Sum of Squared Errors for the candidate model with p predictors

Finding the Best Model

- Fitted all the possible models($2^{14} = 16{,}384$ models)and calculated the values of AIC,BIC and Mallows Cp.
- Investigated the models having lowest AIC,BIC and Mallows Cp values.
- BIC reduces the model complexity from 14 predictors to 10 predictors.
- AIC and Mallows Cp reduces the model complexity from 14 predictors to 12 predictors.
- Selected the model having 10 predictors as the final model.
- Final predictors selected are:
- age,bmi,smoker,bmi^2,interaction terms of smoker with age and bmi,sex_male,region_southeast,region_southwest,few_children and many_children.

Predictors Selected by AIC,BIC & Mallow's Cp

	BIC	Cp	AIC
0	const	const	const
1	A	A	A
2	S	S	S
3	AS	A^2	A^2
4	B^2	AS	AS
5	BS	B^2	B^2
6	sex_male	BS	BS
7	region_southeast	sex_male	sex_male
8	region_southwest	region_northwest	region_northwest
9	children_category_Few Children	region_southeast	region_southeast
10	children_category_Many Children	region_southwest	region_southwest
11	NaN	children_category_Few Children	children_category_Few Children
12	NaN	children_category_Many Children	children_category_Many Children

Fitting Best Model Obtained using BIC

- We use the same train test split for fitting and evaluating model performance, but we use the 10 predictors selected using the mentioned criterion to account for tradeoff between model complexity and goodness of fit.

- **Model Performance on train set :**

OLS Regression Results

Dep. Variable:	y	R-squared:	0.811
Model:	OLS	Adj. R-squared:	0.809
Method:	Least Squares	F-statistic:	454.3
Date:	Tue, 15 Apr 2025	Prob (F-statistic):	0.00
Time:	10:17:13	Log-Likelihood:	-611.54
No. Observations:	1069	AIC:	1245.
Df Residuals:	1058	BIC:	1300.
Df Model:	10		

Final Model Parameters

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.3655	0.030	-12.364	0.000	-0.424	-0.308
A	0.6155	0.015	41.423	0.000	0.586	0.645
S	1.6859	0.033	50.576	0.000	1.620	1.751
AS	-0.5035	0.033	-15.156	0.000	-0.569	-0.438
B^2	-0.0311	0.010	-3.185	0.001	-0.050	-0.012
BS	0.3758	0.030	12.661	0.000	0.318	0.434
sex_male	-0.0873	0.027	-3.277	0.001	-0.140	-0.035
region_southeast	-0.0846	0.032	-2.609	0.009	-0.148	-0.021
region_southwest	-0.1126	0.033	-3.446	0.001	-0.177	-0.049
children_category_Few Children	0.2210	0.029	7.719	0.000	0.165	0.277
children_category_Many Children	0.3300	0.040	8.199	0.000	0.251	0.409

Performance of Final Model on Test Set

- The **adjusted R-squared** came out to be 86% which validates that model is extremely good with unseen data.

```
Testing Set:
Mean Squared Error: 0.1402365545845199
Root Mean Squared Error: 0.3744817146197126 4
R-squared: 0.8730860468926268
Adj_R-squared: 0.8676327129700444
```

Performance of final model on 20-fold Cross Validation

```
R^2 scores for each fold: [0.92877084 0.87840295 0.84234614 0.83706403 0.83990351 0.61401578
 0.72051593 0.92866639 0.84028672 0.86746981 0.78002361 0.74090971
 0.7781069  0.83620388 0.83296503 0.81339894 0.86623489 0.79055031
 0.76654704 0.89410531]
Average R^2: 0.8198243865576101
Average Adjusted R^2: 0.8177784863292711
```

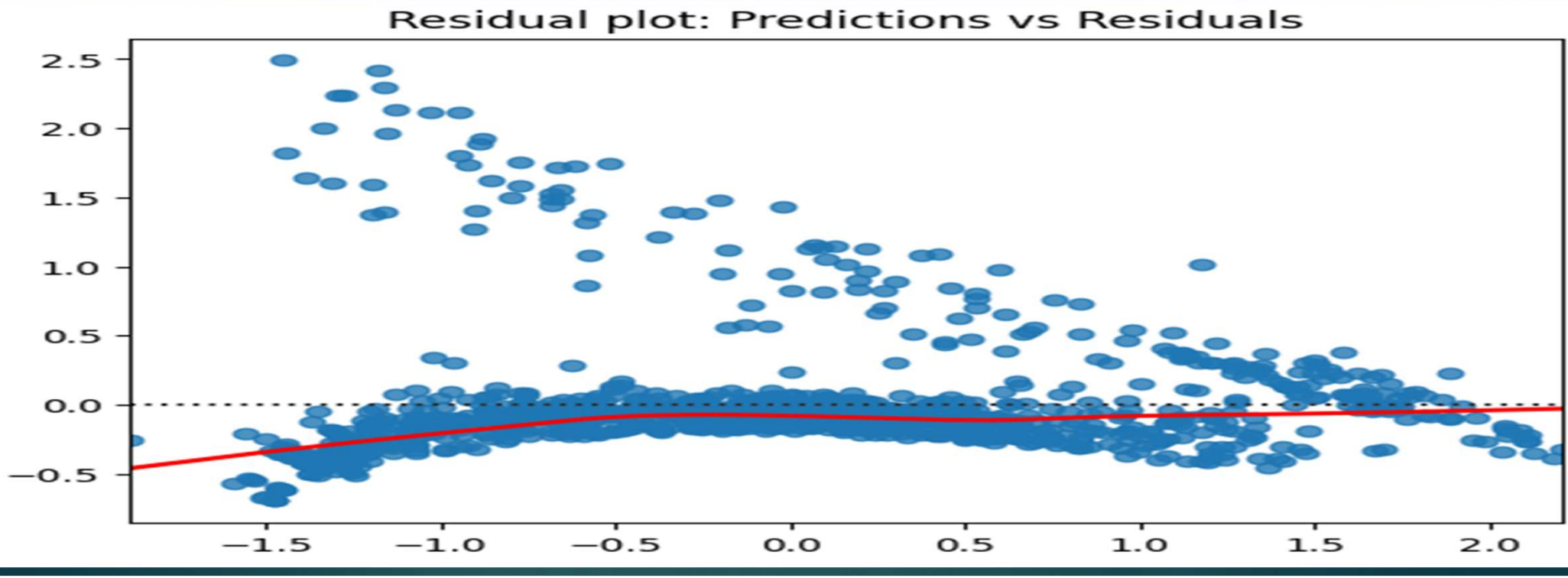
Double-click (or enter) to edit

Upon 20-fold Cross Validation, 81% Adj. R^2 is obtained on average by our final model with only 10 predictors, which is a decent level of model complexity reduction.


Best
Model
Diagnostics

Linearity

- The curve pattern in the residual vs fitted plot for our final model does not allow us to validate linear relationship between response and predictors and homoscedasticity of errors.



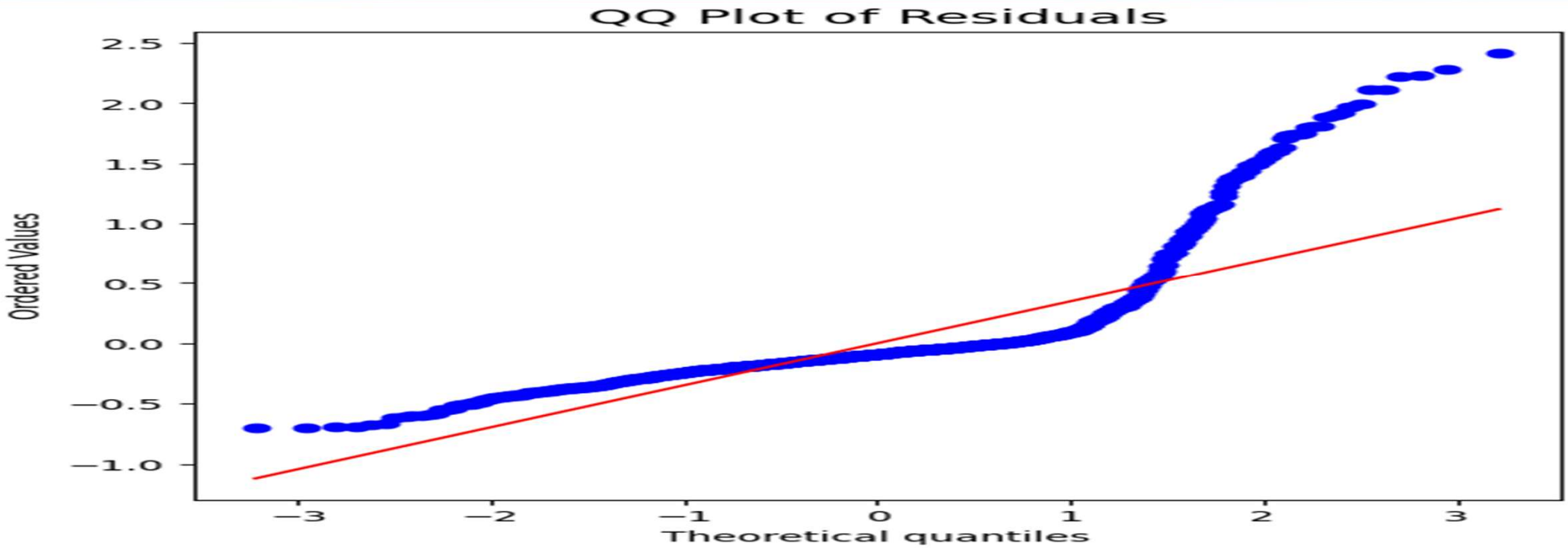
Multicollinearity

	Feature	VIF
0	A	1.256972
1	S	1.013398
2	AS	1.262840
3	B^2	1.033046
4	BS	1.038399
5	sex_male	1.011113
6	region_southeast	1.180846
7	region_southwest	1.142706
8	children_category_Few Children	1.150800
9	children_category_Many Children	1.152839

- Since all VIF values of predictors in our final model is very small (<<5), we can conclude absence of multicollinearity in our final design matrix.

Normality of Errors

- **Shapiro Wilk Test** yielded a p-value of approximately **8.28×10^{-15}**, which is significantly less than the typical significance level of 0.05.
- Thus, we **reject the null hypothesis** that the residuals are normally distributed, indicating a **violation of the normality assumption** of errors in our model.



Independence of Errors & Homoscedasticity

- Since value of Durbin-Watson Test Statistic comes out to be 1.9374 (close to 2) , we can accept the null hypothesis that errors are not autocorrelated.
- P-value for the Breusch-Pagan test comes out to be 1.5289038546450486e-09 << 0.05 hence we reject the null hypothesis that errors are homoscedastic.

Conclusion

- Our final Multiple Linear Regression model showed good predictive performance with a test set adj R^2 of 0.873 and RMSE of 0.3745. However, key assumptions like normality of errors and homoscedasticity were violated. Despite an average cross-validated R^2 of 0.82, these violations suggest that exploring advanced non-parametric methods may lead to more robust and reliable predictions.

Presented By:

Tathagat Pandey
Neha Bawdekar
Deepshikha Sarda
Siboham Pattanayak
Brian Wanzama
Jitesh Yadav


Thank
You