Insurance Premium Prediction

Business Problem:

As data science is booming technology now a days, myself working in insurance industry and data science enthusiast tried to find the overlapping area between this two. In order to calculate premium for clients we generally use static code based on various condition but we can make this more efficient by creating model for calculation on behavioral statistics of Life assured.

The goals of this exercise are to investigate different features and observe their relationships, as well as to plot a multiple linear regression based on several features of an individual, such as age, physical/family condition, and location, against their existing medical expense to be used for predicting future medical expenses and to assist medical insurance companies in determining premium rates.

Data availability:

- Data is available in .xlsx format.
- Dataset contains 1338 observations (rows) and 7 features (columns).
- 7 featured include 3 float data columns (BMI, children and expenses), 1 int type column (age) and 3 object type (smoker, sex, region).

Data has been taken from below URL:

https://www.kaggle.com/noordeen/insurance-premium-prediction

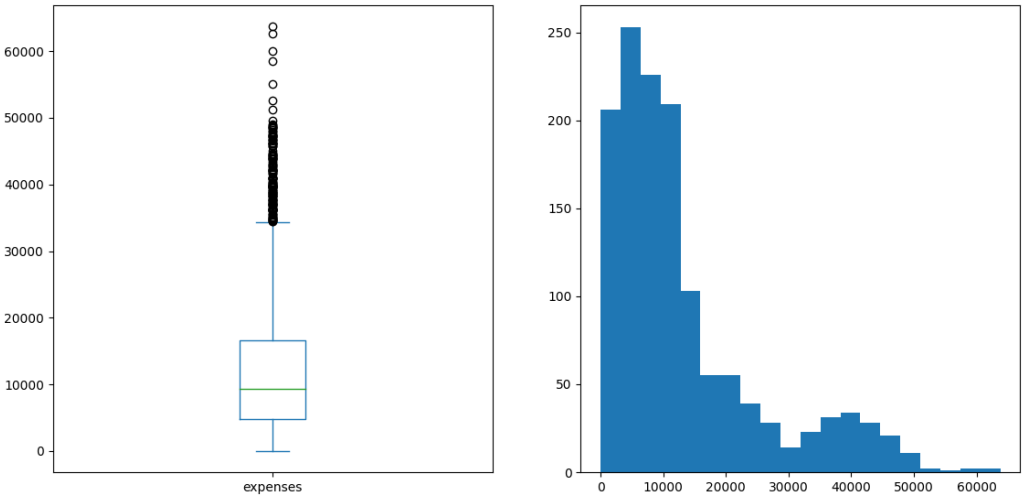
Data cleaning is done to dataset where needed using following methods:

1. Filling region value where it is not mentioned.
2. Filling BMI using median
3. Filling number of children with 0 where it is NULL.

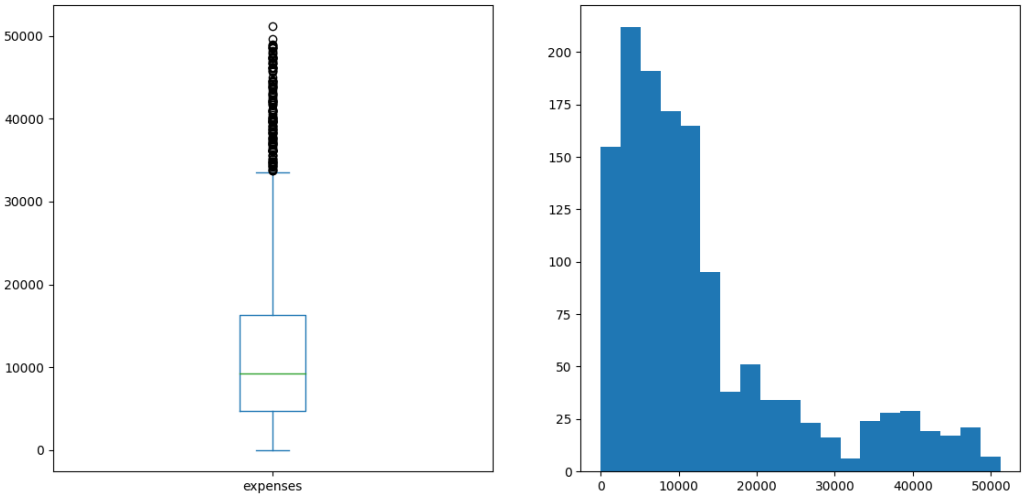
Below is data description:

```
count     1342.000000
mean     13259.992787
std      12094.794413
min       1121.870000
25%       4746.517500
50%       9382.030000
75%      16584.320000
max      63770.430000
```

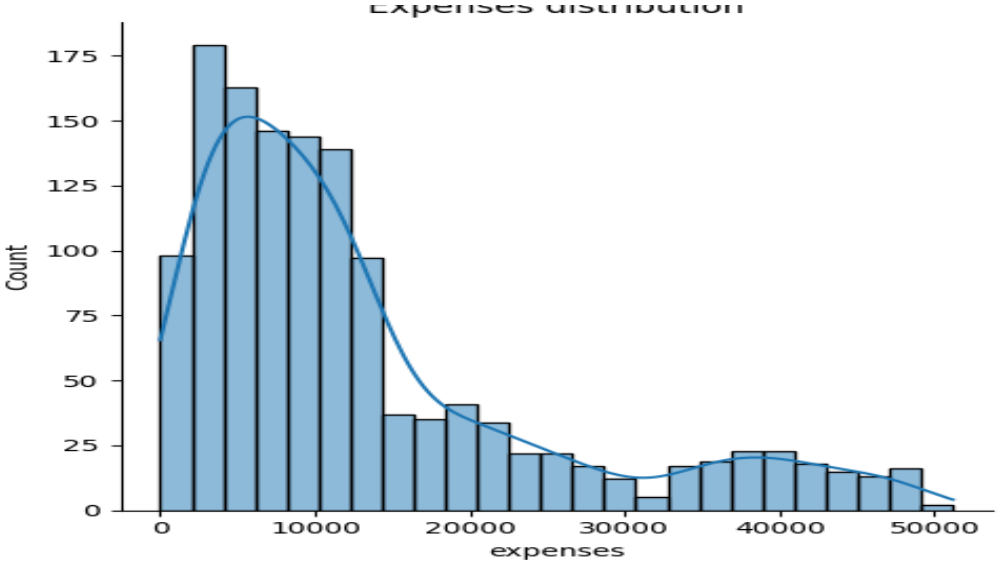
As max is very much higher than minimum we will use box plot and histogram to detect and eliminate outliers.



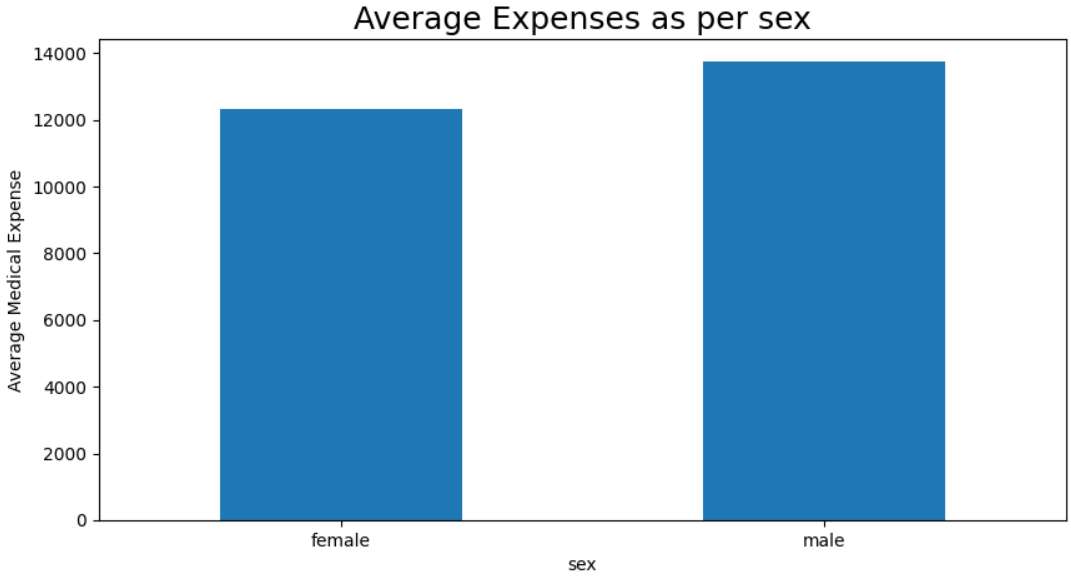
After finding limit of 51240 for removing outliers we proceed further after data cleaning to plot new dataset.

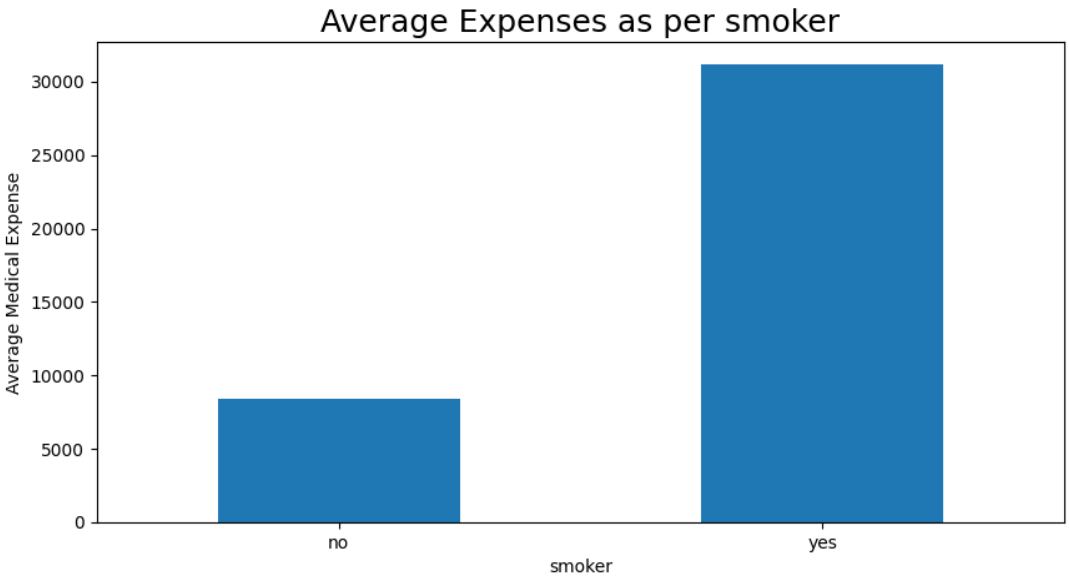


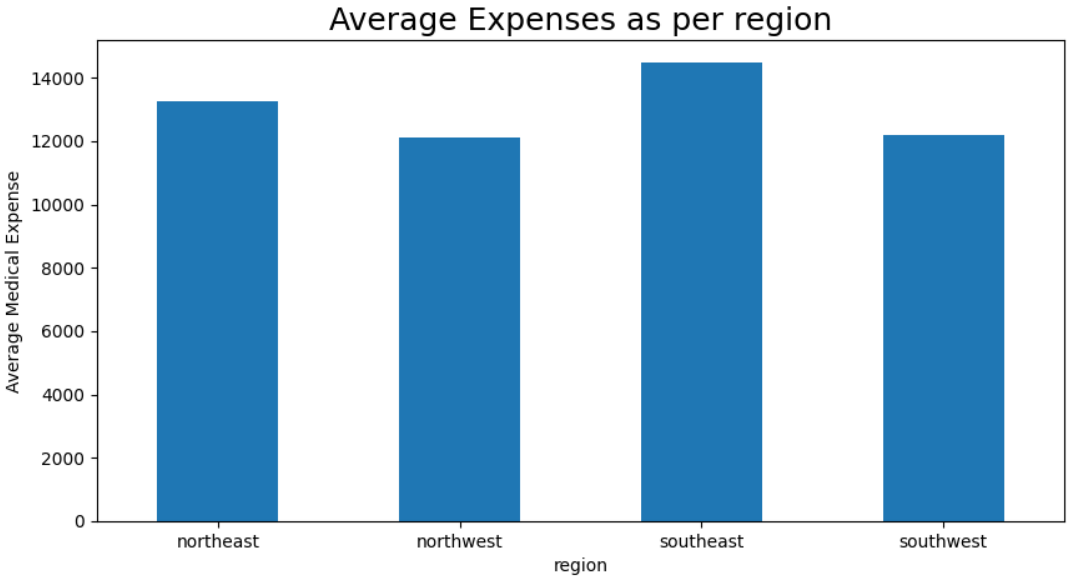
Below is distribution of dataset:

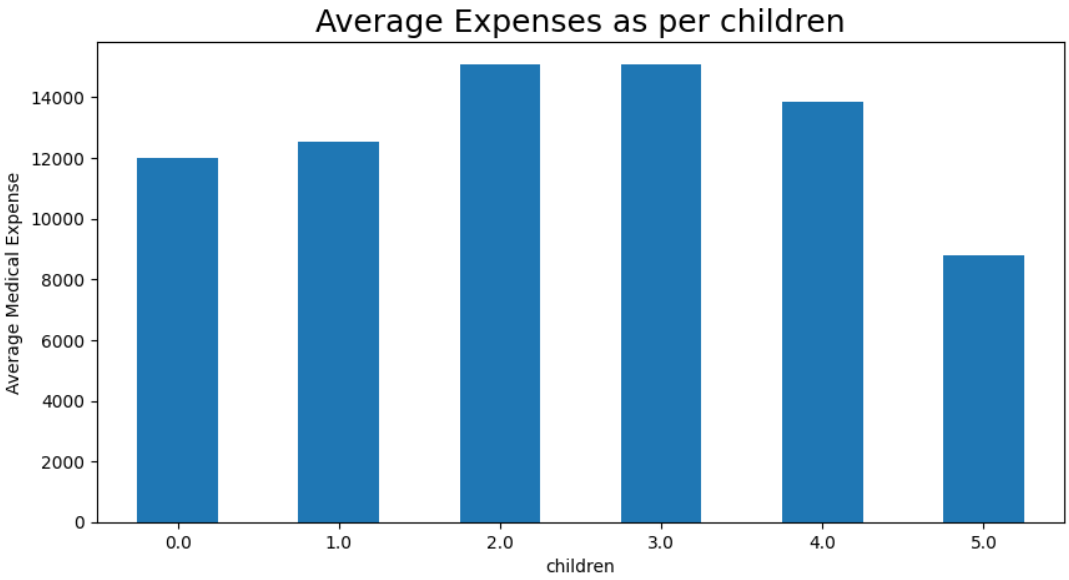


Bar plot for the dependent variables (sex, smoker, region and children) are also as below:

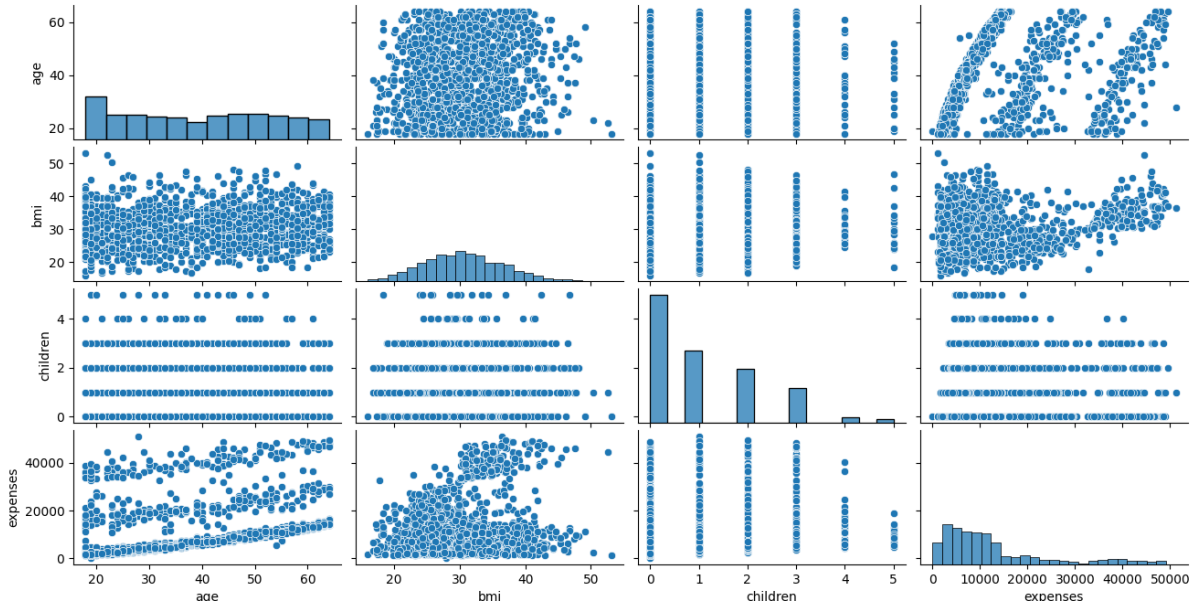



Average Expenses as per smoker
Average Medical Expense
30000
25000
20000
15000
10000
5000
0
no
yes
smoker

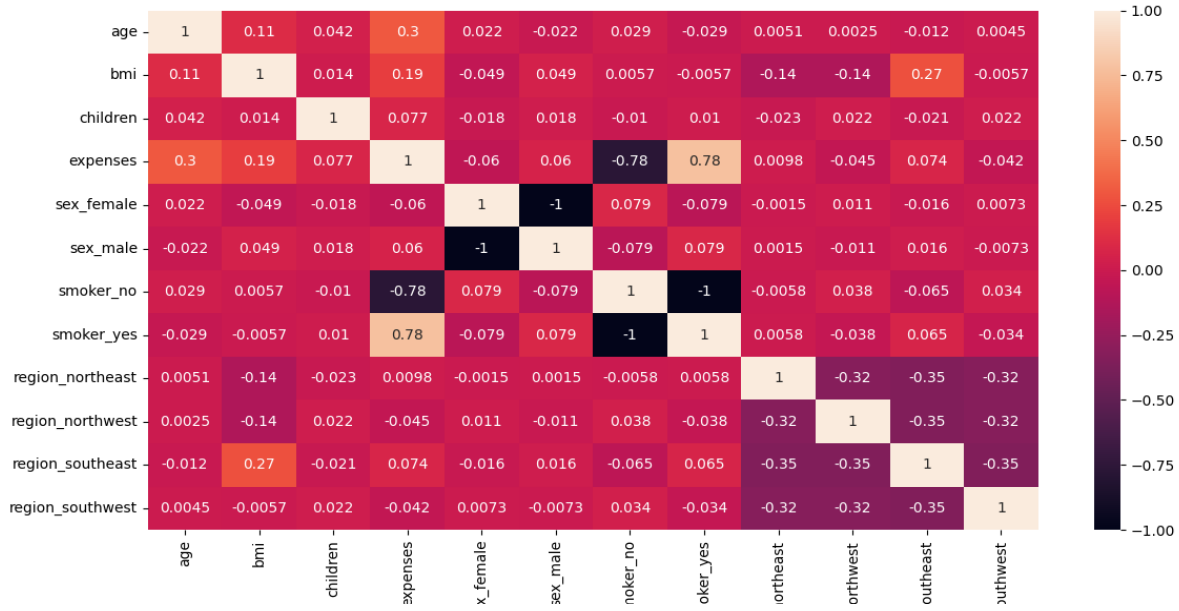

Average Expenses as per region
Average Medical Expense
14000
12000
10000
8000
6000
4000
2000
0
northeast
northwest
southeast
southwest
region



Pair plot with various features:



Heat map to verify highest correlative variable:



Maximum correlation of dependent variable of expense can be seen with smoker feature only.

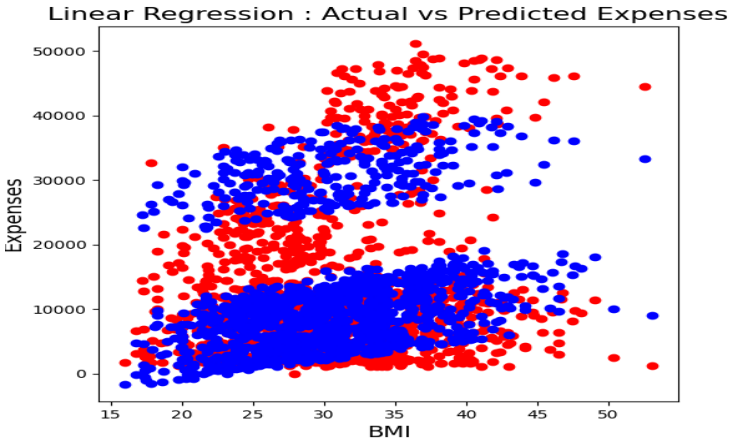
Model Analysis and Visualization:

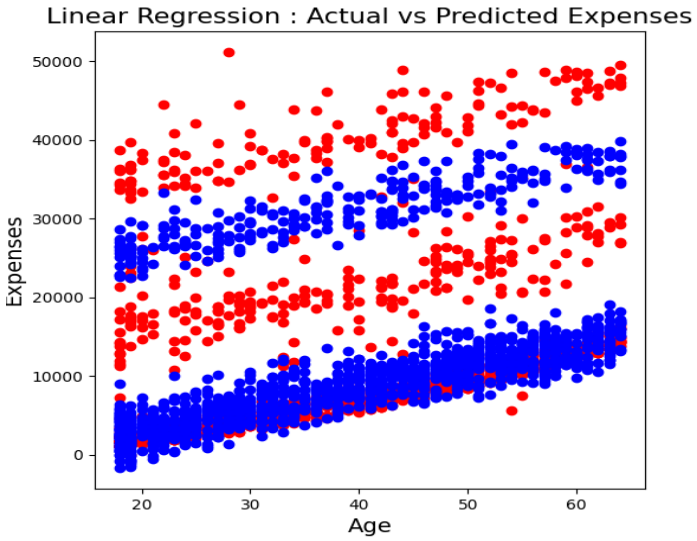
Two types of regression model used

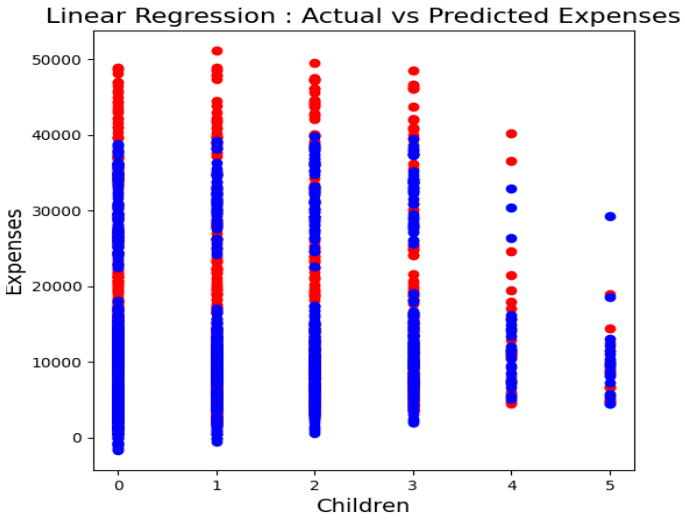
1. Linear Regression
2. Polynomial Regression

Analyzing linear regression model of our data set:

Plotting scatter graph after training model with split of 70 and 30 percent will help to visualize the accuracy of prediction model in line with three features (BMI, age and children)

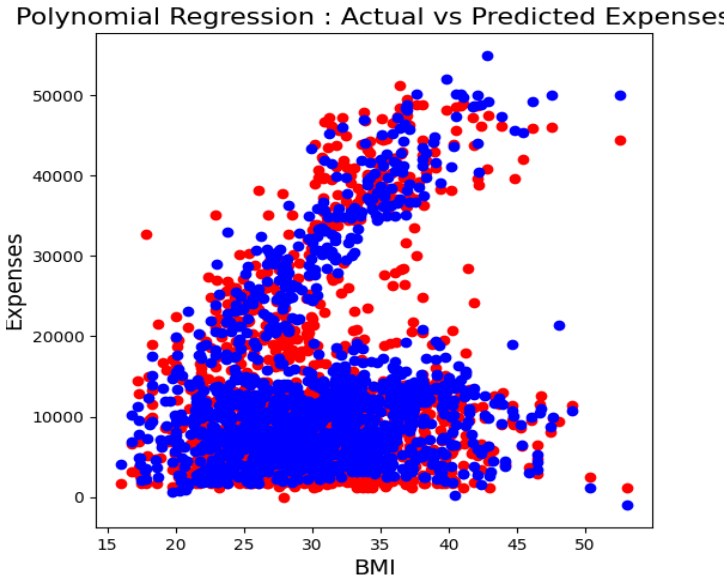


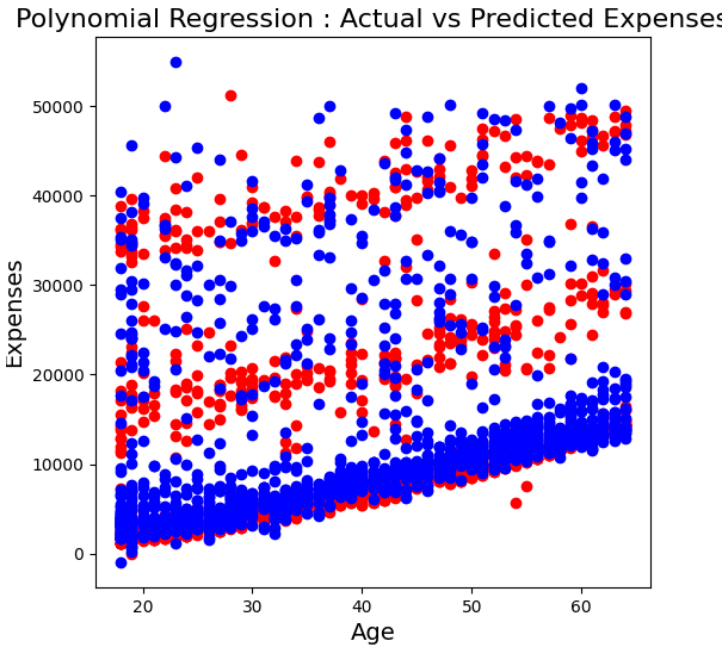


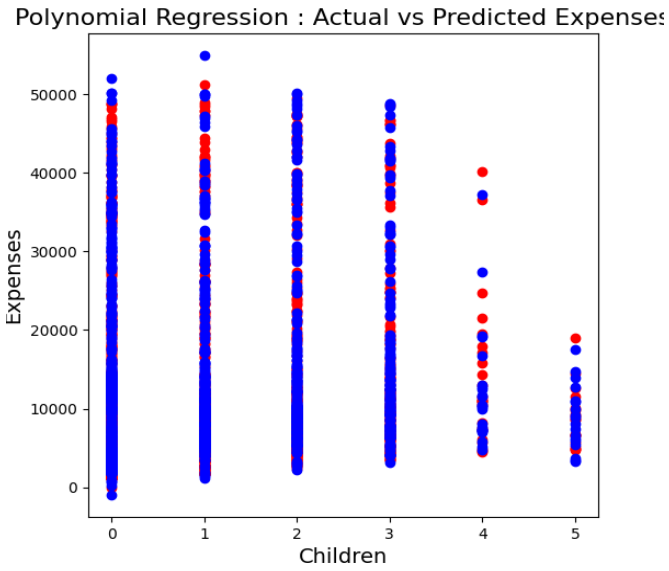


Accuracy of linear regression model for dataset calculated as 74.88% approximately.

Same with polynomial model with scatter graph:


Polynomial Regression : Actual vs Predicted Expenses
Expenses
BMI
0
10000
20000
30000
40000
50000
15
20
25
30
35
40
45
50


Polynomial Regression : Actual vs Predicted Expenses
Expenses
Age
0
10000
20000
30000
40000
50000
20
30
40
50
60



Accuracy of Polynomial regression model for dataset calculated as 80.88% approximately.

As accuracy of polynomial model turns out to be better than linear prediction is done using polynomial model only.

Softwares Used:

1. PyCharm 2021.2
2. Libraries : Pandas, SKLearn, matplotlib.pyplot and seaborn.

Conclusion:

The model generated can be used for predicting medical expenses, using which we can predict our Insurance Premium amount.

Using this model we can predict the medical expenses of a person and hence can predict the Insurance Premium amount according to the predicted medical expenses.

Appendix:

1. Dataset : https://www.kaggle.com/noordeen/insurance-premium-prediction
2. Reference : https://www.edureka.co/blog/data-science-projects/