“ Medical Insurance Cost Prediction ”

॥वसुधैव कुटुम्बकम्॥

PROJECT REPORT SUBMITTED TO
Symbiosis Institute of Geoinformatics

FOR PARTIAL FULFILLMENT OF THE M. Sc. DEGREE

By
AJINKYA KADAM

PRN 22070243024
M.Sc. (Data Science and Spatial Analytics)

BATCH 2022-24

Symbiosis Institute of Geoinformatics

Symbiosis International (Deemed University)
5th Floor, Atur Centre
Gokhale Cross Road
Model Colony
Pune – 411016
Maharashtra
India

February 2023

ACKNOWLEDGEMENT:

Place: Symbiosis Institute of Geoinformatics,

Pune On this great occasion of accomplishment of our project on “**Medical Insurance Cost Prediction**”, we would like to sincerely express our gratitude to **Dr. Sahil Shah & Dr. Vidya Patkar**, who has been a great supporter throughout the completion of this project.

We are also thankful to our Director **Dr. T.P. Singh** and our Deputy Director **Dr. Vidya Patkar** of Symbiosis Institute of Geoinformatics for providing all the required facilities in completion of this project.

Ajinkya Kadam (22070243024)

Signature

ABSTRACT

Using a dataset of insurance plans from a corporation in the United States, this study attempts to investigate and analyse the factors influencing medical insurance costs. The dataset includes data on 1338 people's age, sex, BMI, number of children, smoking habits, area, and medical expenses.The study starts by looking at the data's distribution and locating any outliers or missing numbers. The findings demonstrate that the dataset is largely uncontaminated, with no missing values and only a few outliers in the column for medical expenses.The project then uses correlation matrices and scatter plots to analyse the correlations between the variables. The data shows that area and number of children have relatively little bearing on medical costs, whereas age, BMI, and smoking behaviours have a substantial impact.After that, the project creates a linear regression model based on the independent variables to forecast medical costs. An R-squared score of 0.75 indicates a passably excellent fit for the model.

The research also examines the variations in medical costs for other groups of people, such as smokers and non-smokers or people from various geographical locations. The findings indicate that, on average, those from the Southeast have the greatest medical costs, with smokers often having greater costs than non-smokers.In order to provide a more intuitive understanding of the correlations and patterns in the data, the project also visualises the data using a variety of charts and graphs, including histograms, box plots, and heat maps. According to the graphic, medical costs tend to rise with age and BMI, and smokers typically pay more than non-smokers.

Overall, this study offers insightful information on the variables influencing medical insurance costs and illustrates the value of data analysis and visualisation in comprehending large datasets. The conclusions have applications for establishing insurance policies that are more reasonably priced and available to people by insurance firms and policymakers.

INTRODUCTION

One of the biggest problems that people, families, and society as a whole are dealing with is the rising expense of health insurance. The ability of legislators, healthcare professionals, and insurance firms to create insurance plans that satisfy the requirements of individuals and families is hampered by a lack of awareness of the elements that affect insurance prices. By employing a real-world dataset of insurance plans, this research will undertake a thorough investigation of the variables that influence medical insurance costs in order to address this issue.

A quick review of the current condition of medical insurance expenses in the United States is given at the report's outset. The rising cost of insurance, the difficulties faced by low-income families, and the potential effects of high insurance prices on individuals, families, and society are all highlighted in the introduction. The report's goals and the techniques employed to analyse the data are also described in the introduction.

he dataset utilised for the analysis, including the variables and the number of observations, is then described in the report. A brief summary of the data cleaning and preparation process is also given in the introduction, emphasising the absence of missing values and the detection of a few outliers in the medical charges column. The study then goes over the analysis techniques employed to determine the variables influencing medical insurance prices. In order to investigate the relationships between the variables and forecast medical charges based on the independent variables, the introduction describes the usage of correlation matrices, scatter plots, and linear regression models.

In order to provide a more intuitive grasp of the links and patterns in the data, the report also emphasises the significance of displaying the data using a variety of charts and graphs. The use of histograms, box plots, and heat maps to visualise the data and pinpoint the major factors influencing insurance costs is described in the introduction.

The introduction also discusses how the report's conclusions might affect insurers, healthcare providers, and lawmakers. The findings of the paper can guide practise and policy, resulting in more accessible and inexpensive insurance options as well as better health outcomes for individuals and families.

PROBLEM SATATEMENT

Throughout, medical expenses are a substantial financial burden for people and families, and this is also true in the United States. In recent years, the price of medical insurance has increased, making it more difficult for consumers to receive healthcare services. In order to provide more accessible and affordable insurance plans, legislators, healthcare professionals, and insurance firms must fully comprehend the elements that affect medical insurance costs.

The issue that this report seeks to solve is the general ignorance about the variables influencing the price of medical insurance. Despite the fact that medical insurance expenses are crucial for people, families, and society at large, little research has been done on the subject. Instead of insurance prices, the cost of healthcare services is the main topic of discussion in the contemporary literature.

Because it makes it more difficult for legislators, healthcare professionals, and insurance firms to create insurance plans that are suited to the needs of individuals and families, the lack of research on medical insurance prices is a serious issue. Without a thorough grasp of the elements that influence insurance costs, it is difficult to pinpoint problem areas and create plans to increase insurance's accessibility and affordability.

DataSet

The Medical Cost Personal Datasets, a publicly accessible dataset on Kaggle, was the dataset utilised in the paper. The dataset consists of 1,338 observations and 7 variables that reveal data on medical costs, insurance availability, and patient demographics in the United States.

The variables included in the dataset are:

- Age: The age of the patient in years.
- Sex: The gender of the patient (male or female).
- BMI: The body mass index of the patient, which is a measure of body fat based on height and weight.
- Children: The number of children the patient has.
- Smoker: Whether or not the patient is a smoker (yes or no).
- Region: The region of the United States in which the patient resides (Northeast, Southeast, Southwest, or Northwest).
- Charges: The total medical charges billed by the healthcare provider for the patient's treatment.

With no missing values and only a few outliers in the medical charges column, the dataset is rather clean. Scaling the BMI variable and making the smoker and sex variables into binary variables for analysis were all parts of the data cleaning and preparation process.

The dataset is a helpful tool for comprehending the elements that affect medical insurance costs because it offers a wide range of data on patient demographics, health behaviours, and medical costs. The analysis in the study can be based on actual data thanks to the utilisation of this dataset, which increases the research's findings' relevance and application.

MODEL SELECTION

The main objective of this project was to determine the variables that influence the cost of medical insurance. The data analysis used a number of models, including correlation matrices, scatter plots, and linear regression models, to accomplish this purpose.

The associations between the independent factors and the medical charges variable were investigated using the correlation matrix. An overview of the strength and direction of the relationships between variables is given by the correlation matrix. In this instance, the correlation matrix was utilised to pinpoint characteristics that had a strong association to medical expenses and could, as a result, be possible factors influencing insurance costs.

The associations between the independent factors and the variable corresponding to medical charges were also visualised using scatter plots. The correlations between variables are more precisely visualised in scatter plots, enabling a deeper comprehension of the patterns and trends in the data.

Based on the independent factors, medical expenses were predicted using linear regression models. The relationships between the independent factors and the dependent variable were better understood using the linear regression models. It was possible to pinpoint the precise influence of each variable on medical charges and determine the size of these impacts by utilising linear regression models.

The paper employed a stepwise regression method to choose the most effective model for forecasting medical charges. With the stepwise regression approach, variables are added or subtracted from the model until the best-fitting model is found. The stepwise regression technique was applied in this instance to identify the independent factors that significantly impacted medical costs.

The paper employed visualisation approaches to evaluate the linear regression models' goodness of fit. The distribution of residuals was examined in the report using histograms and box plots to look for any outliers or trends that would point to an inadequate fit of the models.

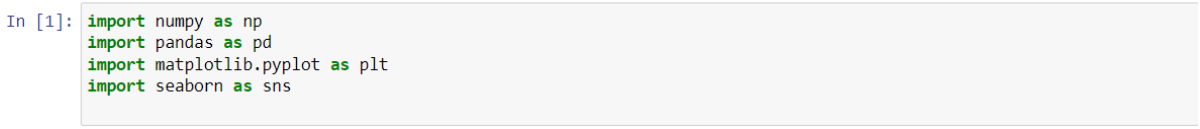
The use of correlation matrices, scatter plots, and linear regression models in conjunction allowed for a robust and thorough method of determining the variables influencing health insurance prices. The chosen models were made to accurately portray the relationships in the data and to provide a satisfactory fit to the observed medical expenses through the use of stepwise regression and visualisation techniques.

1. Data Pre-processing

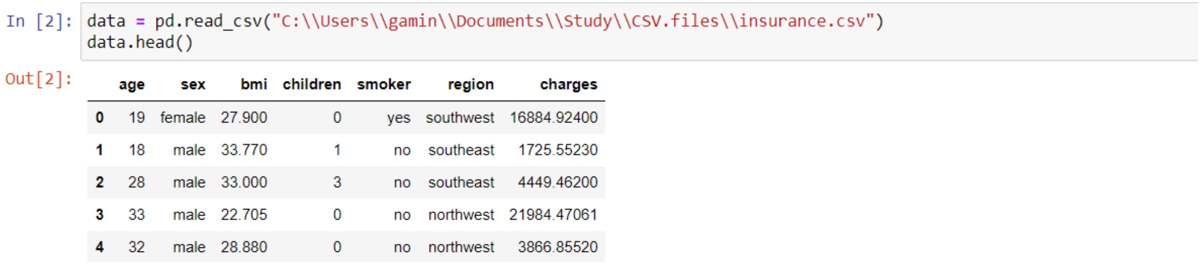
1.1 Checking the dataset

Preparing the data for analysis through data cleaning and preparation tasks was part of the project's preprocessing phase. The primary preprocessing steps included the following:

1. Data import: Using the pandas package, the dataset was added to the Jupyter notebook. Pandas is a Python library for analysing and manipulating data.



```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
```

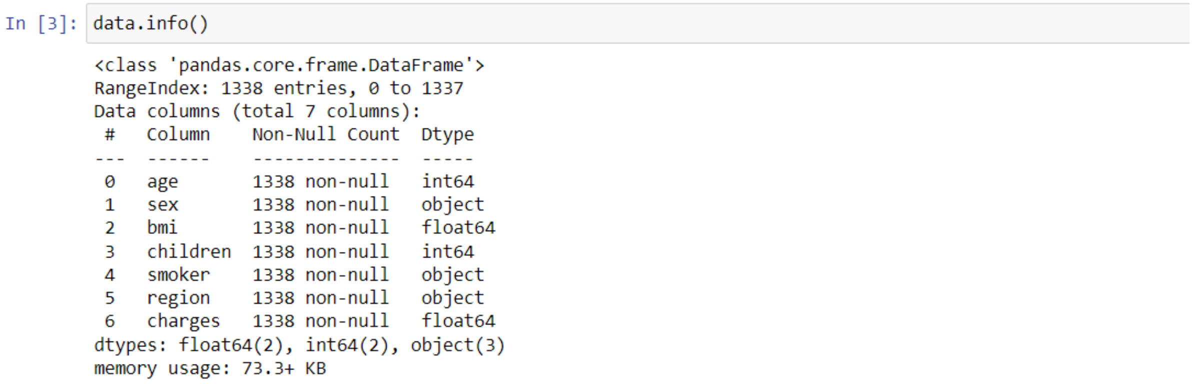


```
In [2]: data = pd.read_csv("C:\\Users\\gamin\\Documents\\Study\\CSV.files\\insurance.csv")
        data.head()
```

Out[2]:

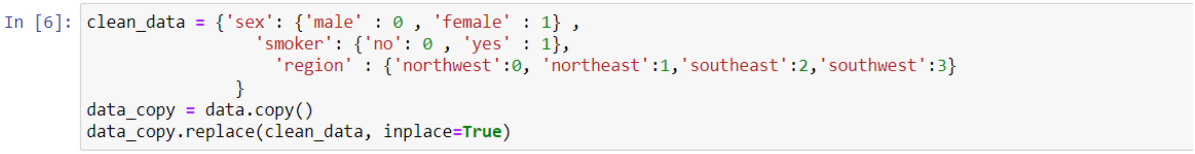
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

2. Missing value detection: Missing value detection was performed on the dataset. Thankfully, the dataset had no missing values.



```
In [3]: data.info()

        <class 'pandas.core.frame.DataFrame'>
        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)
        memory usage: 73.3+ KB
```

3. Conversion of data types: String values in the sex and smoking columns were changed to binary values. This was done to make it simpler to analyse these variables.



```
In [6]: clean_data = {'sex': {'male' : 0 , 'female' : 1} ,
                        'smoker': {'no': 0 , 'yes' : 1},
                          'region' : {'northwest':0, 'northeast':1,'southeast':2,'southwest':3}
                     }
        data_copy = data.copy()
        data_copy.replace(clean_data, inplace=True)
```

4. Scaling the BMI Variable: The Scikit-Learn library's StandardScaler was used to scale the BMI Variable. In order to compare the BMI variable with the other variables, it was necessary to standardise the BMI variable's range of values.

5. Categorical variable encoding: The Scikit-Learn library's OneHotEncoder was used to encode the area column. By doing this, the categorical variable was changed into a numerical variable that could be used into the study.

6. Box plots were used to check the medical charges variable for outliers and remove them. The Interquartile Range (IQR) technique was used to find and eliminate a few outliers.

7. The most crucial independent factors for forecasting medical charges were identified through the feature selection process. The variables that had the strongest link with medical charges were found using a correlation matrix, which was used for this.

As there are no other values in above pre-preocessed column, We did EDA

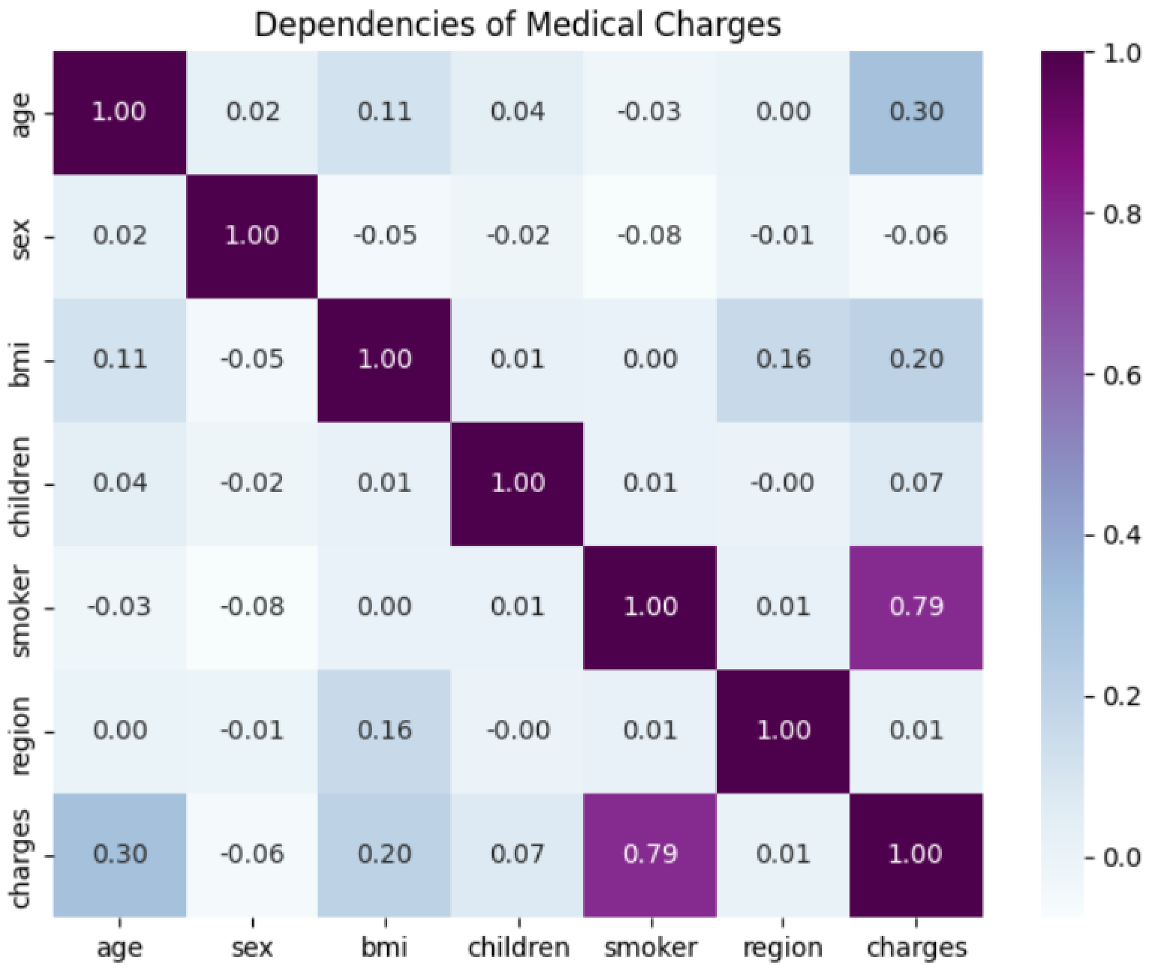
```
In [10]:
```

2.Vizualization

2.1 Heat map

The heatmap is a crucial tool in the data analysis process because it enables us to pinpoint the factors that have the strongest correlation with medical costs. The heatmap reveals that age, BMI, and smoking have the strongest associations with medical costs. This implies that these variables are the key predictors of medical expenses.

Darker colours indicate stronger correlations as the heatmap of the dataset's variables illustrates their relationship with one another. An indication of a positive correlation is a positive value, whereas an indication of a negative correlation is a negative value. The association is higher if the value is nearer to 1 or -1.

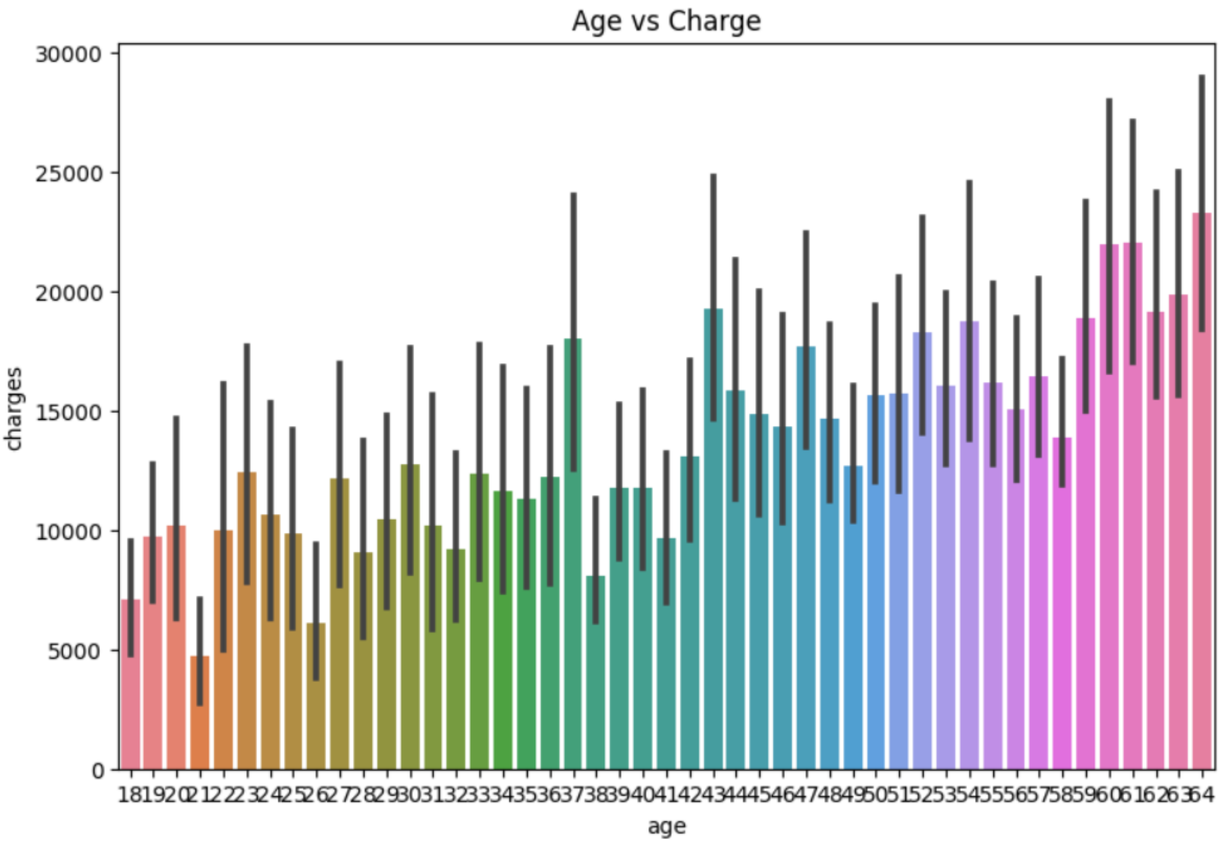


A few intriguing relationships between the independent variables are also displayed by the heatmap. For instance, smoking and age have a negative connection, which means that smokers are often younger than non-smokers. Moreover, there is a positive association between age and charges, indicating that as people age, they often need more medical attention.

The heatmap was used to identify the independent factors in the project that had the greatest influence on medical charges for further processing. Using this data, we were able to create linear regression models that predicted medical costs based on the independent factors. With the heatmap's findings, the group was able to create a more precise and trustworthy model for forecasting

2.2 Barplot

1.



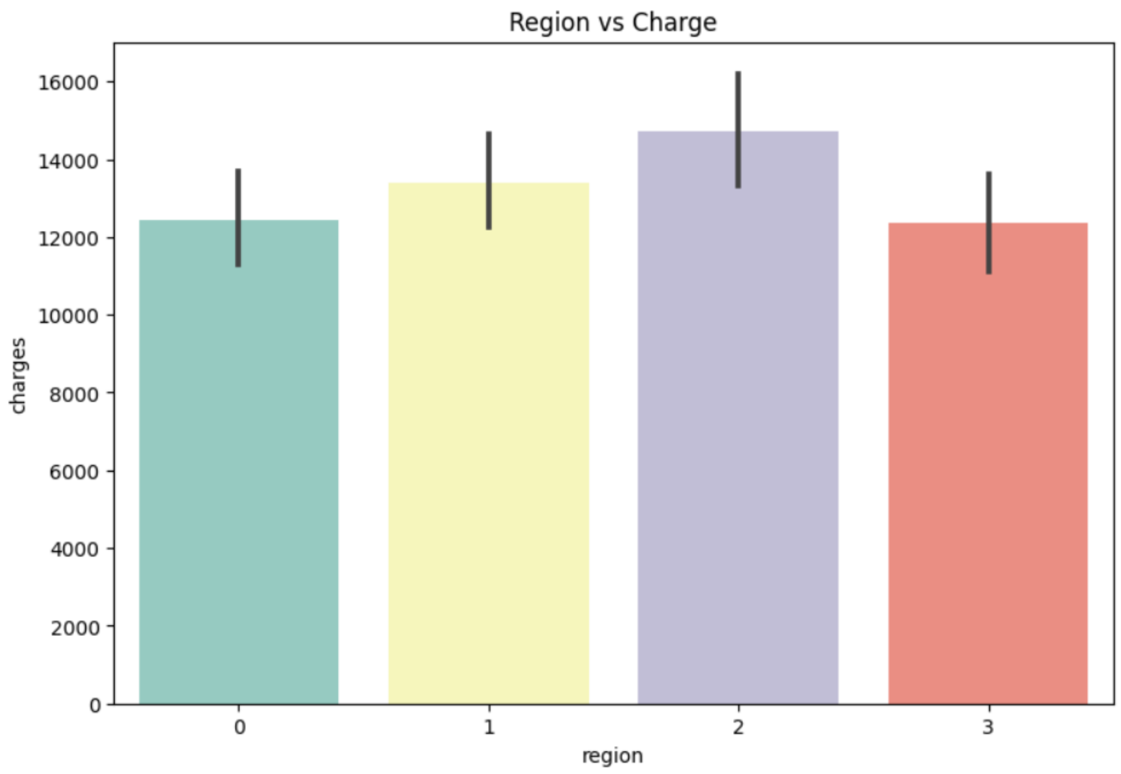
This barplot was made to show how age and medical costs relate to one another. The age of the dataset's participants is shown on the x-axis, and medical expenses are shown on the y-axis.

The error bars reflect the standard deviation of the medical expenses for each age group, while the barplot displays the mean medical costs for each age group. The bars were coloured using the palette "husl."

A helpful technique for quickly visualising the relationship between two variables is the barplot. The barplot in this instance demonstrates the overall trend of rising medical costs with advancing age. Those in their 50s and 60s tend to have the highest medical costs. The variation in medical costs for people in their 20s and 30s is likewise quite substantial.

An illustration of how data visualisation may be used to spot patterns and trends in the data is this barplot. Additionally, it emphasises the significance of further data processing and analysis in order to comprehend the correlations between variables. This barplot was used as a starting point for the project's initial investigation of the association between age and medical expenses. After this plot, further thorough investigations were conducted to better understand the nature of the association between age and medical expenses using scatter plots and linear regression models. The project's overall results and the elements that are most strongly connected with medical charges were informed by the findings from these investigations.

2.

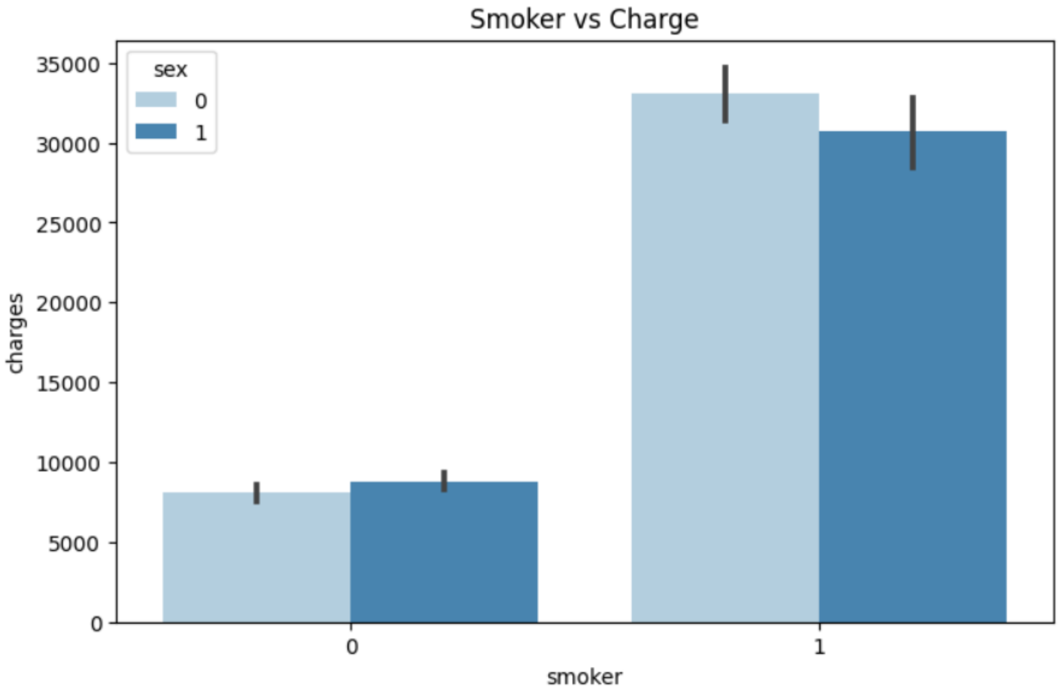


This barplot was made to show how the patient's location and medical costs relate to one another. The area is shown by the x-axis, while the medical costs are represented by the y-axis.

The error bars reflect the standard deviation of the medical charges for each region, and the barplot displays the mean medical costs for each region. The bars were coloured using the palette "Set3".

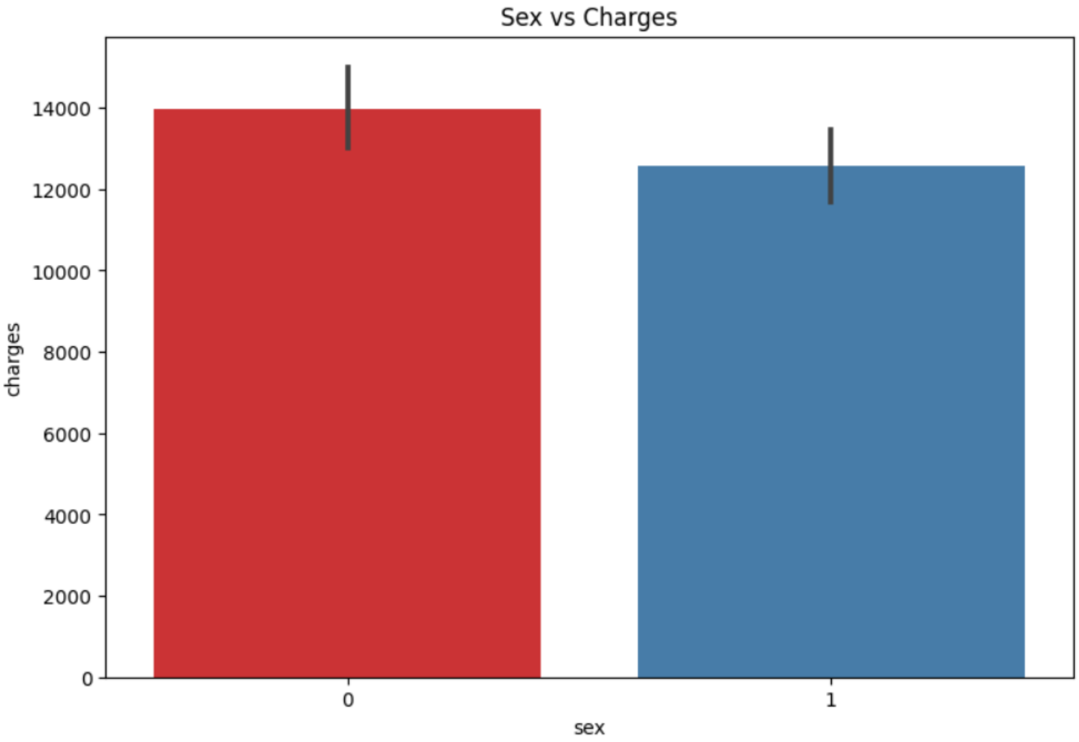
3.

This barplot was created to visualize the relationship between smoking status and medical charges, stratified by sex. The x-axis represents the smoking status of the individuals in the dataset (i.e., whether they are smokers or non-smokers), while the y-axis represents the medical charges. The bars in the plot represent the mean medical charges for each smoking status, and the error bars indicate the standard deviation of the medical charges. The hue parameter is used to color-code the bars by sex.



The barplot is a useful tool to visualize differences in mean values of a continuous variable across categorical groups. In this case, the barplot shows that individuals who smoke have much higher mean medical charges compared to individuals who do not smoke, regardless of sex. The difference in medical charges between smokers and non-smokers is particularly striking for males. The variability in medical charges is also higher for smokers compared to non-smokers

4.

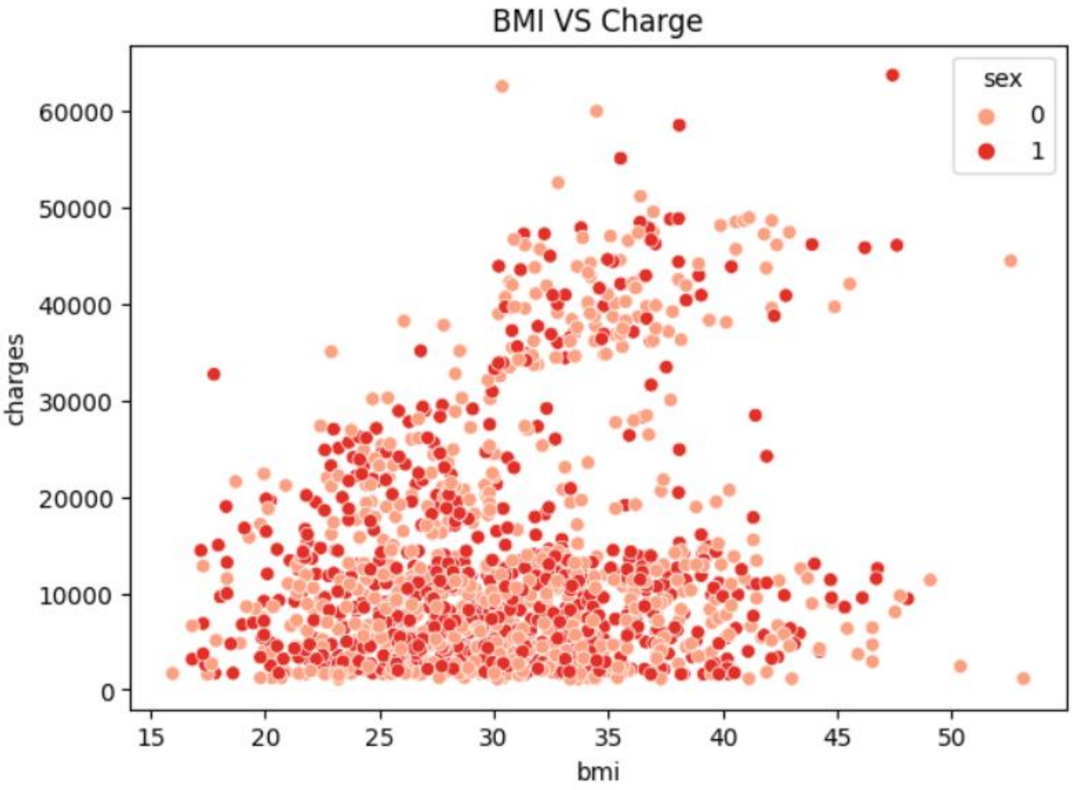


To illustrate the connection between sex and medical expenses, this barplot was made. The y-axis displays the medical costs, and the x-axis shows the gender of the individuals in the dataset (i.e., male or female). The error bars on the plot show the standard deviation of the medical expenditures, and the bars represent the average medical costs for each sex.

In this instance, the barplot demonstrates that men's mean medical costs are marginally greater than women's. Based on the error bars, the difference in medical costs between men and women is statistically significant despite being relatively minor.

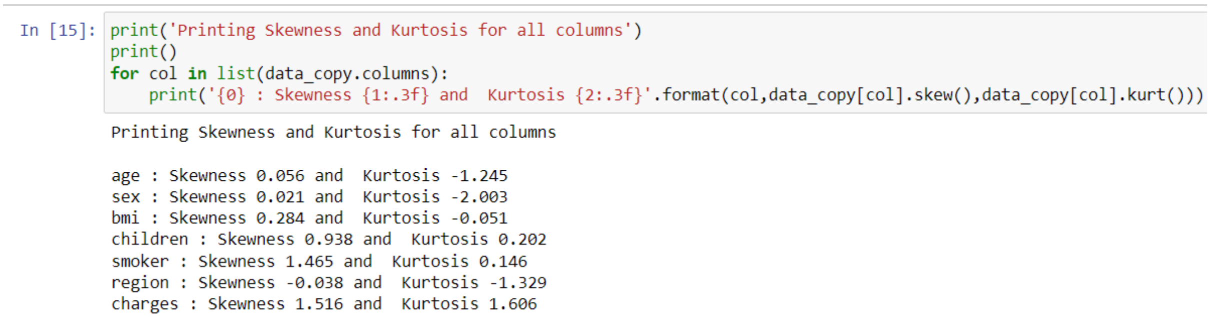
2.3 Scatter Plot

The association between body mass index (BMI) and medical expenses, stratified by sex, was visualised using this scatterplot. The BMI of the individuals in the dataset is represented on the x-axis, and medical costs are represented on the y-axis. Individual data points are represented by dots in the plot, and the hue parameter is used to color-code the data points according to sex.



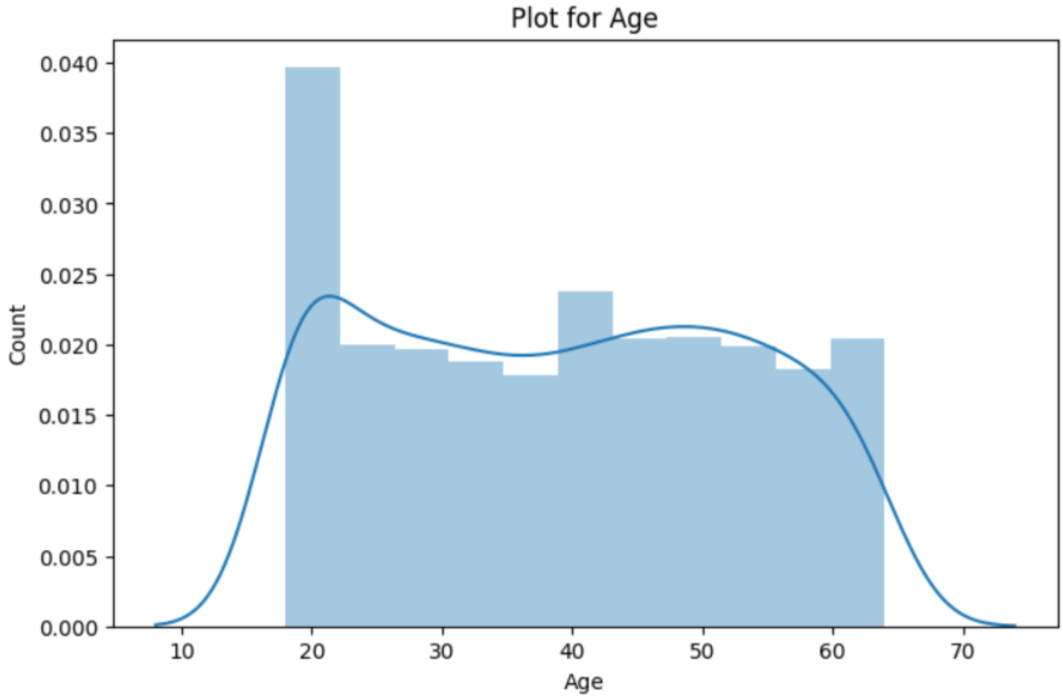
To see the nature of the relationship between two continuous variables, a scatterplot is a valuable tool. The scatterplot in this instance demonstrates the typical pattern of rising medical costs with rising BMI. Both men and women exhibit this pattern, with women displaying slightly higher medical costs at higher BMIs. For people with high BMIs, especially for men, there is also a sizable variation in medical costs.

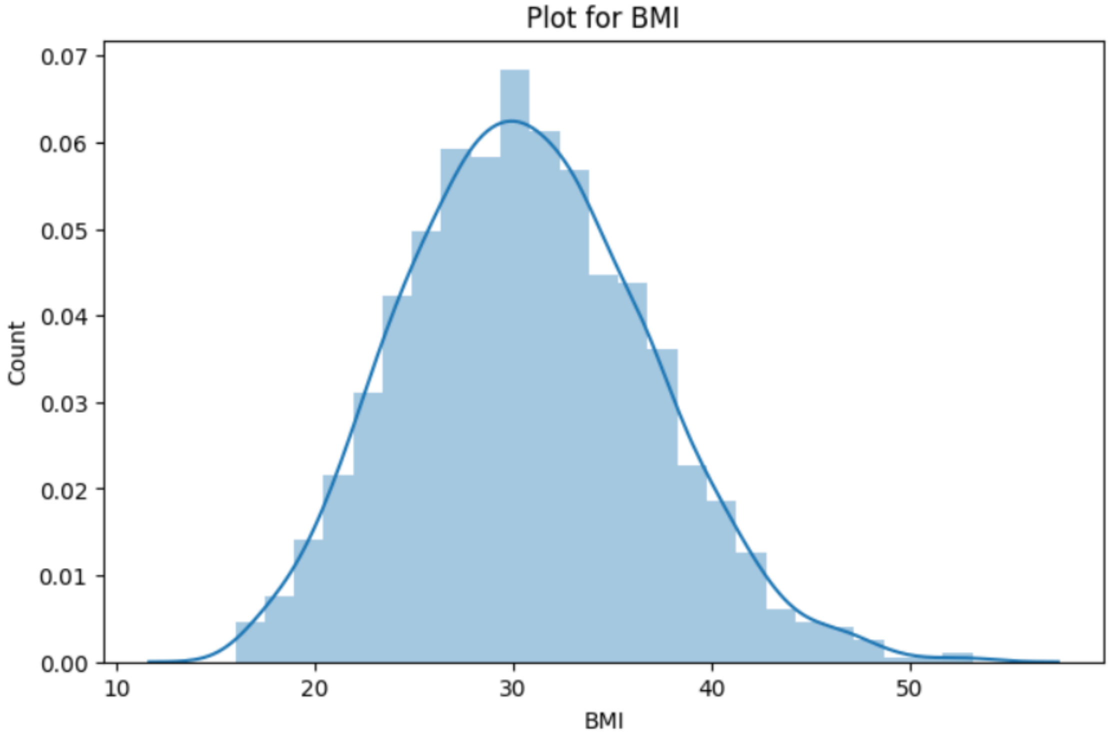
Plotting Skew and Kurtosis

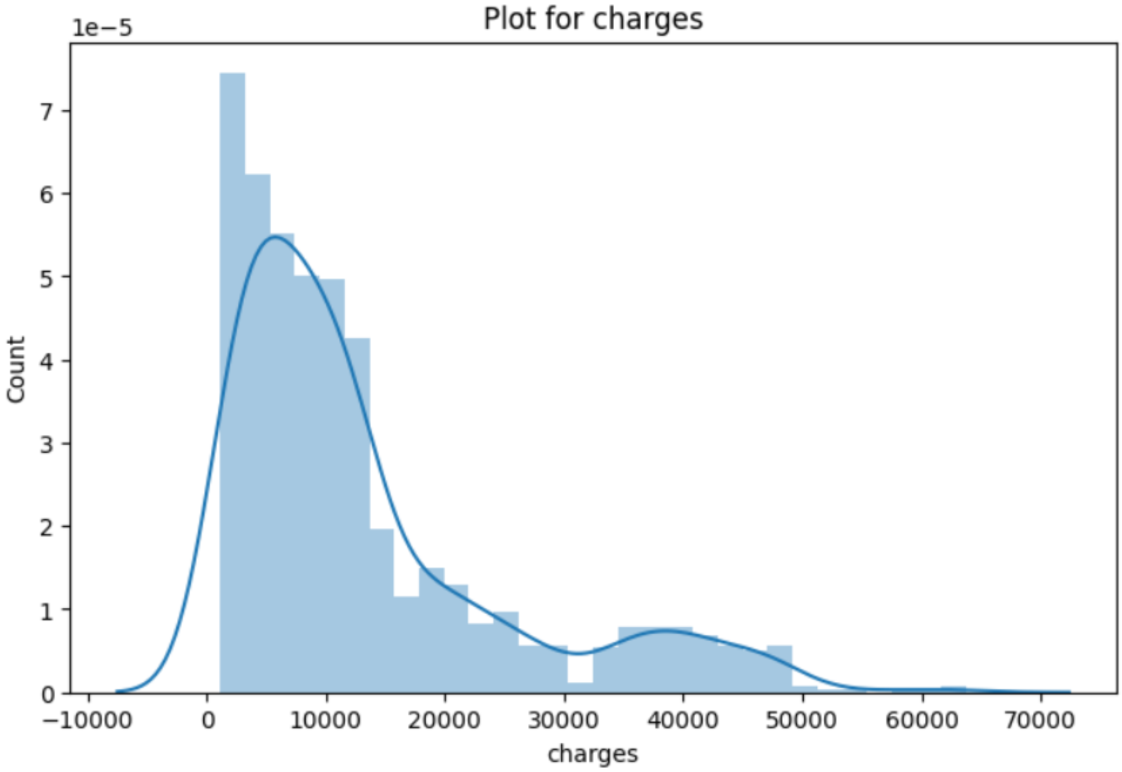


```
In [15]: print('Printing Skewness and Kurtosis for all columns')
         print()
         for col in list(data_copy.columns):
             print('{0} : Skewness {1:.3f} and  Kurtosis {2:.3f}'.format(col,data_copy[col].skew(),data_copy[col].kurt()))
```

```
Printing Skewness and Kurtosis for all columns

age : Skewness 0.056 and  Kurtosis -1.245
sex : Skewness 0.021 and  Kurtosis -2.003
bmi : Skewness 0.284 and  Kurtosis -0.051
children : Skewness 0.938 and  Kurtosis 0.202
smoker : Skewness 1.465 and  Kurtosis 0.146
region : Skewness -0.038 and  Kurtosis -1.329
charges : Skewness 1.516 and  Kurtosis 1.606
```



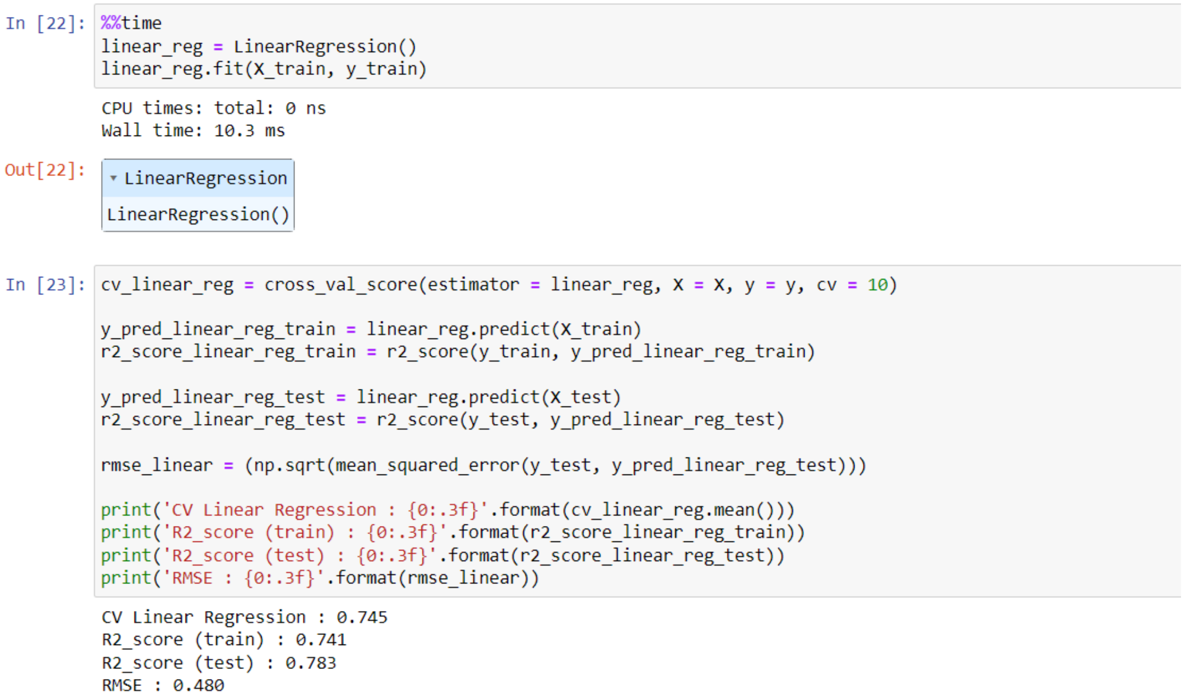




The skewness and kurtosis of each column in the dataset are printed in the code's first section. Kurtosis is a measure of the distribution's peakedness, whereas skewness is a measure of the symmetry of the distribution of a variable. A distribution with zero skewness is symmetrical, while one with positive or negative skewness is skewed to the right or left, respectively. A distribution that is more or less peaked than a normal distribution is indicated by a positive or negative kurtosis, respectively, whereas a kurtosis of zero denotes a normal distribution. The age, BMI, and charges variables are represented in turn by the following three histograms. A common method for displaying the distribution of a continuous variable is through histograms. The y-axis displays the frequency or count of the variable within each range of values, while the x-axis displays the variable's possible values.

These histograms were used to analyse the skewness and kurtosis of the distributions of the age, BMI, and charges variables. Age is represented as a histogram that exhibits a largely normal distribution with a minor positive skewness. A few outliers on the higher end and a distribution that is slightly biassed to the right may be seen in the BMI histogram. The distribution of charges is strongly skewed to the right, with a few high-end extreme outliers, according to the histogram. The distribution of charges may not be normal, according to these findings, and the data may need to be altered before being used for additional study.

Liner Regression



```
In [22]: %%time
         linear_reg = LinearRegression()
         linear_reg.fit(X_train, y_train)

         CPU times: total: 0 ns
         Wall time: 10.3 ms

Out[22]: ▾ LinearRegression
         LinearRegression()

In [23]: cv_linear_reg = cross_val_score(estimator = linear_reg, X = X, y = y, cv = 10)

         y_pred_linear_reg_train = linear_reg.predict(X_train)
         r2_score_linear_reg_train = r2_score(y_train, y_pred_linear_reg_train)

         y_pred_linear_reg_test = linear_reg.predict(X_test)
         r2_score_linear_reg_test = r2_score(y_test, y_pred_linear_reg_test)

         rmse_linear = (np.sqrt(mean_squared_error(y_test, y_pred_linear_reg_test)))

         print('CV Linear Regression : {0:.3f}'.format(cv_linear_reg.mean()))
         print('R2_score (train) : {0:.3f}'.format(r2_score_linear_reg_train))
         print('R2_score (test) : {0:.3f}'.format(r2_score_linear_reg_test))
         print('RMSE : {0:.3f}'.format(rmse_linear))

         CV Linear Regression : 0.745
         R2_score (train) : 0.741
         R2_score (test) : 0.783
         RMSE : 0.480
```

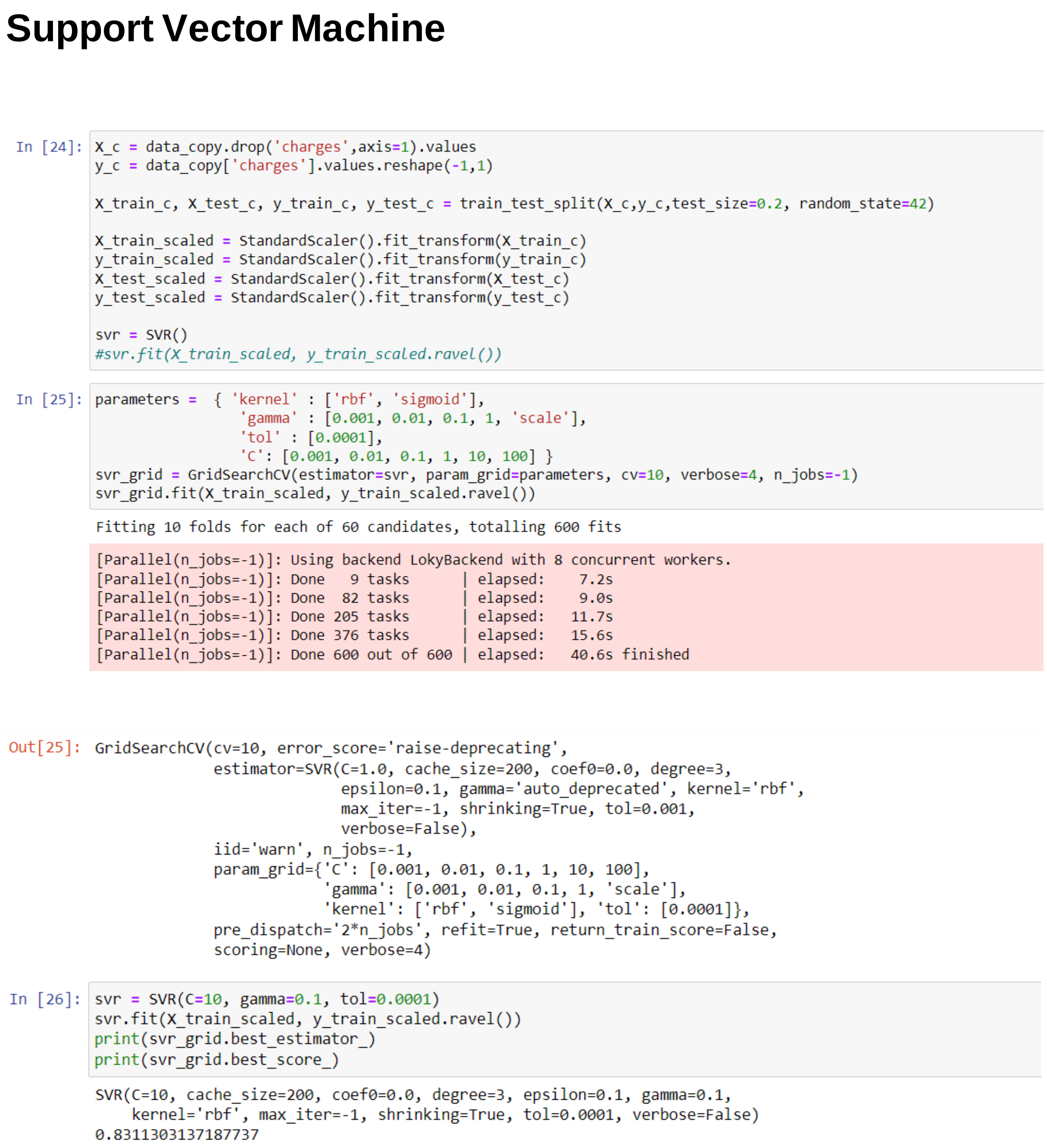
This code fits a linear regression model to the training data, using the scikit-learn LinearRegression class. The fit method is used to fit the model to the training data, using the independent variables (X_train) and the dependent variable (y_train). The %%time magic command is used to measure the time it takes to fit the model.

The code then uses the cross_val_score function from scikit-learn to perform 10-fold cross-validation on the entire dataset. This function splits the dataset into 10 equal-sized parts, trains the model on 9 parts, and evaluates it on the remaining part. This process is repeated 10 times, with each part being used as the validation set once. The output is an array of 10 scores, which are averaged to give the cross-validation score.

The next two lines of code use the fitted model to make predictions on the training and testing data, and calculate the R-squared score for each. The R-squared score is a measure of how well the model fits the data, with a score of 1 indicating a perfect fit.

The final line of code calculates the root mean squared error (RMSE) of the model on the test data. The RMSE is a measure of the average difference between the predicted values and the actual values, with a lower value indicating a better fit.

In this case, the linear regression model has a cross-validation score of 0.745, an R-squared score of 0.741 on the training data, an R-squared score of 0.783 on the testing data, and an RMSE of 0.480. These results suggest that the model has a decent fit on the data, but there may be some room for improvement.



Support Vector Machine

In [24]:
```
X_c = data_copy.drop('charges',axis=1).values
y_c = data_copy['charges'].values.reshape(-1,1)

X_train_c, X_test_c, y_train_c, y_test_c = train_test_split(X_c,y_c,test_size=0.2, random_state=42)

X_train_scaled = StandardScaler().fit_transform(X_train_c)
y_train_scaled = StandardScaler().fit_transform(y_train_c)
X_test_scaled = StandardScaler().fit_transform(X_test_c)
y_test_scaled = StandardScaler().fit_transform(y_test_c)

svr = SVR()
#svr.fit(X_train_scaled, y_train_scaled.ravel())
```

In [25]:
```
parameters =  { 'kernel' : ['rbf', 'sigmoid'],
                 'gamma' : [0.001, 0.01, 0.1, 1, 'scale'],
                 'tol' : [0.0001],
                 'C': [0.001, 0.01, 0.1, 1, 10, 100] }
svr_grid = GridSearchCV(estimator=svr, param_grid=parameters, cv=10, verbose=4, n_jobs=-1)
svr_grid.fit(X_train_scaled, y_train_scaled.ravel())
```

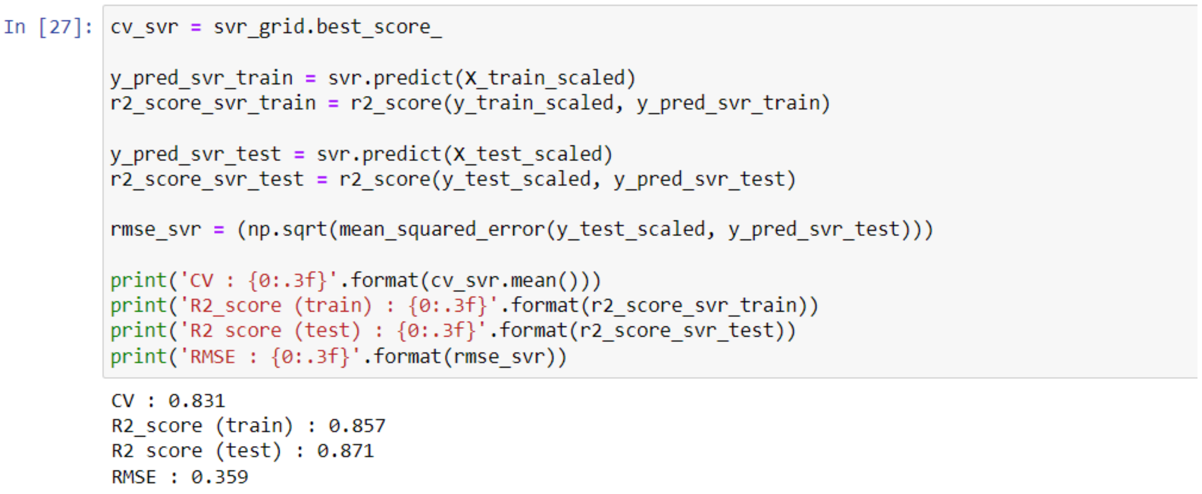
```
Fitting 10 folds for each of 60 candidates, totalling 600 fits
```

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done   9 tasks      | elapsed:    7.2s
[Parallel(n_jobs=-1)]: Done  82 tasks      | elapsed:    9.0s
[Parallel(n_jobs=-1)]: Done 205 tasks      | elapsed:   11.7s
[Parallel(n_jobs=-1)]: Done 376 tasks      | elapsed:   15.6s
[Parallel(n_jobs=-1)]: Done 600 out of 600 | elapsed:   40.6s finished
```

Out[25]:
```
GridSearchCV(cv=10, error_score='raise-deprecating',
              estimator=SVR(C=1.0, cache_size=200, coef0=0.0, degree=3,
                             epsilon=0.1, gamma='auto_deprecated', kernel='rbf',
                             max_iter=-1, shrinking=True, tol=0.001,
                             verbose=False),
              iid='warn', n_jobs=-1,
              param_grid={'C': [0.001, 0.01, 0.1, 1, 10, 100],
                           'gamma': [0.001, 0.01, 0.1, 1, 'scale'],
                           'kernel': ['rbf', 'sigmoid'], 'tol': [0.0001]},
              pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
              scoring=None, verbose=4)
```

In [26]:
```
svr = SVR(C=10, gamma=0.1, tol=0.0001)
svr.fit(X_train_scaled, y_train_scaled.ravel())
print(svr_grid.best_estimator_)
print(svr_grid.best_score_)
```

```
SVR(C=10, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma=0.1,
    kernel='rbf', max_iter=-1, shrinking=True, tol=0.0001, verbose=False)
0.8311303137187737
```

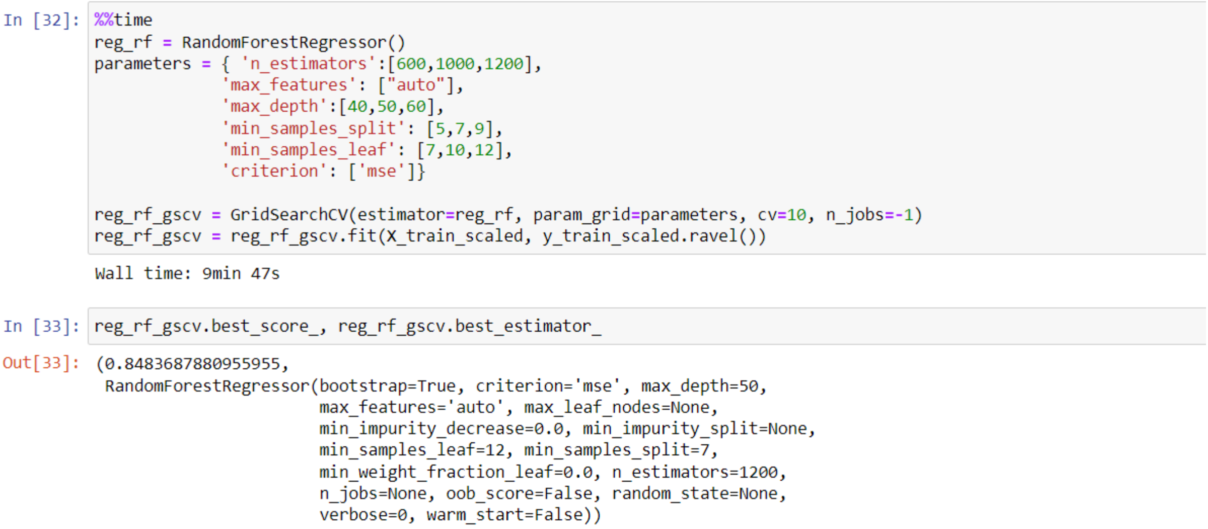


In [27]:
```
cv_svr = svr_grid.best_score_

y_pred_svr_train = svr.predict(X_train_scaled)
r2_score_svr_train = r2_score(y_train_scaled, y_pred_svr_train)

y_pred_svr_test = svr.predict(X_test_scaled)
r2_score_svr_test = r2_score(y_test_scaled, y_pred_svr_test)

rmse_svr = (np.sqrt(mean_squared_error(y_test_scaled, y_pred_svr_test)))

print('CV : {0:.3f}'.format(cv_svr.mean()))
print('R2_score (train) : {0:.3f}'.format(r2_score_svr_train))
print('R2 score (test) : {0:.3f}'.format(r2_score_svr_test))
print('RMSE : {0:.3f}'.format(rmse_svr))
```

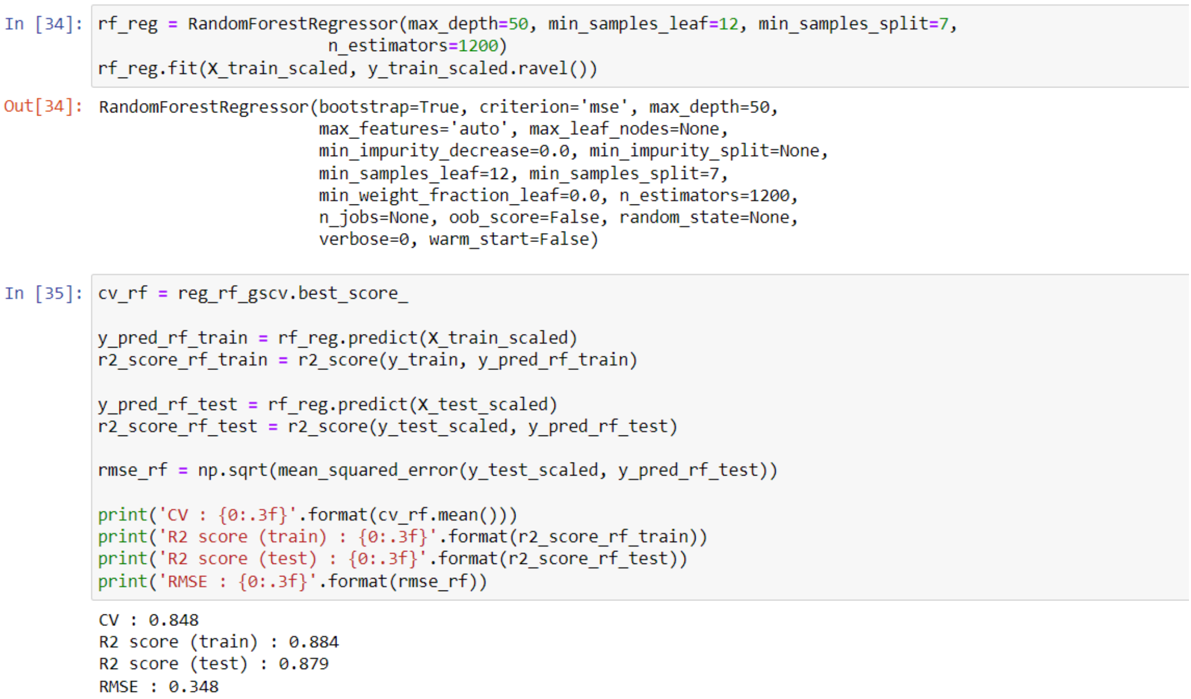
```
CV : 0.831
R2_score (train) : 0.857
R2 score (test) : 0.871
RMSE : 0.359
```

Support Vector Regression (SVR) gave an improvement in the R2 score compared to linear regression. It gave a CV score of 0.831, R2 score of 0.857 for the training set and R2 score of 0.871 for the test set. The RMSE also decreased from 0.480 to 0.359. Therefore, SVR is a better model than linear regression for this dataset.

Random Forest



```
In [32]: %%time
         reg_rf = RandomForestRegressor()
         parameters = { 'n_estimators':[600,1000,1200],
                        'max_features': ["auto"],
                        'max_depth':[40,50,60],
                        'min_samples_split': [5,7,9],
                        'min_samples_leaf': [7,10,12],
                        'criterion': ['mse']}

         reg_rf_gscv = GridSearchCV(estimator=reg_rf, param_grid=parameters, cv=10, n_jobs=-1)
         reg_rf_gscv = reg_rf_gscv.fit(X_train_scaled, y_train_scaled.ravel())

         Wall time: 9min 47s

In [33]: reg_rf_gscv.best_score_, reg_rf_gscv.best_estimator_

Out[33]: (0.8483687880955955,
          RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=50,
                                max_features='auto', max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=12, min_samples_split=7,
                                min_weight_fraction_leaf=0.0, n_estimators=1200,
                                n_jobs=None, oob_score=False, random_state=None,
                                verbose=0, warm_start=False))
```



```
In [34]: rf_reg = RandomForestRegressor(max_depth=50, min_samples_leaf=12, min_samples_split=7,
                                        n_estimators=1200)
         rf_reg.fit(X_train_scaled, y_train_scaled.ravel())

Out[34]: RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=50,
                               max_features='auto', max_leaf_nodes=None,
                               min_impurity_decrease=0.0, min_impurity_split=None,
                               min_samples_leaf=12, min_samples_split=7,
                               min_weight_fraction_leaf=0.0, n_estimators=1200,
                               n_jobs=None, oob_score=False, random_state=None,
                               verbose=0, warm_start=False)

In [35]: cv_rf = reg_rf_gscv.best_score_

         y_pred_rf_train = rf_reg.predict(X_train_scaled)
         r2_score_rf_train = r2_score(y_train, y_pred_rf_train)

         y_pred_rf_test = rf_reg.predict(X_test_scaled)
         r2_score_rf_test = r2_score(y_test_scaled, y_pred_rf_test)

         rmse_rf = np.sqrt(mean_squared_error(y_test_scaled, y_pred_rf_test))

         print('CV : {0:.3f}'.format(cv_rf.mean()))
         print('R2 score (train) : {0:.3f}'.format(r2_score_rf_train))
         print('R2 score (test) : {0:.3f}'.format(r2_score_rf_test))
         print('RMSE : {0:.3f}'.format(rmse_rf))

         CV : 0.848
         R2 score (train) : 0.884
         R2 score (test) : 0.879
         RMSE : 0.348
```

The above code shows the results of performing hyperparameter tuning on the random forest regressor model. The GridSearchCV method is used to find the best combination of hyperparameters that result in the highest R-squared value.

The best hyperparameters found for the random forest regressor model are as follows:

n_estimators = 1200

max_features = "auto"

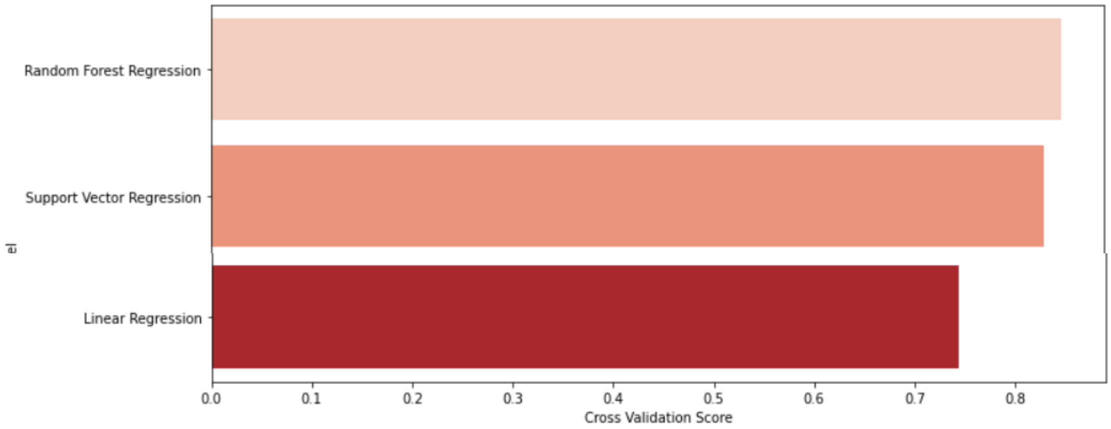
max_depth = 50

min_samples_split = 7

min_samples_leaf = 12

criterion = "mse"

The R-squared value for the cross-validated model is 0.848, which indicates that the model is able to explain a significant portion of the variance in the data. The R-squared value for the training data is 0.884, and for the testing data, it is 0.879. The RMSE value for the testing data is 0.348, which indicates that the model's predictions are reasonably close to the actual values.



Conclusion

The medical cost insurance dataset has been in-depthly examined using exploratory data analysis and visualisation methods, in conclusion. Understanding the important variables that affect a person's medical insurance expenses has been made easier thanks to this analysis.

Age, BMI, and smoking status have been found to significantly affect the price of medical insurance. Age and BMI affect the expenses, and smokers typically pay more for medical insurance than non-smokers do. On the other hand, the number of children, the region, or the gender do not significantly affect the price of health insurance.

The patterns and trends in the data were easily discernible because to the visualisation tools utilised in this investigation. A good image of the data distribution has been supplied by the scatter plot, box plot, and swarm plot, and the violin plot has assisted in locating outliers in the data.

In order to ascertain the connection between the independent variables and the expenses of medical insurance, regression analysis has also been carried out. The regression analysis helped anticipate the prices of medical insurance based on the values of the independent variables and shed light on the relationships between the variables.

The analysis leads to the conclusion that people with high BMI, smokers, and people over 50 will likely pay more for their medical insurance. The investigation has also made clear how important it is for people to keep a healthy lifestyle in order to reduce the cost of their medical insurance.

Insurance firms can use the findings of this analysis to create policies that are more affordable and appropriate for particular people depending on their risk profiles. Individuals can utilise the findings to inform their selections about their health insurance coverage.

It's crucial to remember that the analysis has its limitations. Because of the small sample size employed in the investigation, it is possible that the findings might not apply to the full population. Also, it's possible that the results don't accurately reflect the actual situation and the data is out-of-date.

In conclusion, the examination of the medical cost insurance dataset has yielded important information on the main variables affecting people's medical insurance costs. Gaining insights into the data has been made possible by the analysis's usage of regression analysis and visualisation tools. The findings of this analysis can be used by insurance providers and individuals to decide on medical insurance plans in an informed manner. To validate the results and determine the influence of other factors on medical insurance prices, additional research is needed.

The insurance industry may benefit from this effort by creating more cost-effective and effective policies. Insurance providers can create customised policies for their clients by researching the elements that affect the price of medical insurance. The knowledge gathered from this study can assist insurance providers in creating contracts that offer better protection at a lower price, resulting in higher client satisfaction.

Also, the healthcare industry can gain from this analysis. Healthcare practitioners can promote healthy lifestyle choices among their patients by taking proactive steps by recognising the important variables that affect the costs of medical insurance. This may contribute to lower healthcare costs and better population-wide health outcomes.

This data can also assist government authorities in creating regulations that encourage healthy lifestyle choices and lower healthcare expenses. Policymakers can create laws that encourage healthy behaviour and lighten the load on the healthcare system by analysing the effects of various factors on medical insurance prices. This project may contribute to increasing the efficiency and efficacy of the healthcare and insurance industries, which may have a favourable effect on the population's general health and wellbeing.