**Georgia State University**

**J. Mack Robinson College of Business**



**CIS 8005 Group Project Team 1**

**Medical Expenses Prediction**

**Gopi Sandeep Guntuku:** [**gguntuku1@student.gsu.edu**](mailto:gguntuku1@student.gsu.edu)

**Sukumar Reddy Kottala:** [**skottala1@student.gsu.edu**](mailto:skottala1@student.gsu.edu)

**Project Summary**

**Project description:**

Everyone’s life revolves around their health. Good health is essential to all aspects of our lives. Health refers to a person’s ability to cope up with the environment on a physical, emotional, mental, and social level.

Because of the quick speed of our lives, we are adopting many habits that are harming our health. One spends a lot of money to be healthy by participating in physical activities or having frequent health check-ups to avoid being unfit and get rid of health disorders. When we become ill, we tend to spend a lot of money, resulting in a lot of medical expenses.

So, an application can be made which can make people understand the factors which are making them unfit, and creating a lot of medical expenses, and it could identify and estimate medical expense if someone has such factors.

**Project description:**

This project aims at building Machine Learning models which can predict a patient's medical expenses based on specific features and identifying the factors affecting the medical expenses of the subjects based on the model output.

Factors affecting the medical expenses of the patients :

* **Age**
* **Gender**
* **Body Mass Index**
* **Region**
* **Smoking Behavior**

.

**Business Implications of the Project:**

* Health is the center of everyone’s life.
* Every part of our life relies on good health.
* Health is the extent of an individual’s continuing physical, emotional, mental, and social ability to cope with the environment.

**Data Source**:

* The dataset has been sourced from **kaggle.com** and on **GitHub**.
* This data file includes all needed information to find out more about age, gender, smoking behavior and necessary metrics to make predictions and draw conclusions.
* **Link to the dataset:** [**https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv**](https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv)

The dataset contains information on 1,338 patients.

It includes the following features:

* Age: Patient's age
* Sex: Patient's sex
* BMI: Patient's Body Mass Index
* Children: How many children the patient has
* Smoker: Whether the patient is a smoker or not
* Region: Which region the patient is from (Northeast, Southeast, Southwest, Northwest)
* Expenses: Individual medical costs billed by health insurance

**Target Variable:**

The Expenses column is the target column, and the rest others are independent columns. Independent columns are those which will predict the outcome.

**Independent Variables:**

The first column is Age. Age is an important factor for predicting medical expenses because young people are generally more healthy than old ones and the medical expenses for Young People will be quite less as compared to old people.

The Next column is sex, which has two Categories in this column: Male and Female. The sex of the person can also play a vital role in predicting the medical expenses of a subject.

After that, you have the ‘bmi’ column, then**BMI is Body Mass Index*.*** For most adults, an ideal BMI is in the 18.5 to 24.9 range. For children and young people aged 2 to 18, the BMI calculation considers age and gender as well as height and weight. If your BMI is less than 18.5, you are considered underweight. People with very low or very high ‘bmi’ are more likely to require medical assistance, resulting in higher costs.

The fourth column is the ‘children’ column, which contains information on how many children your patients have. Persons who have children are under more pressure because of their children’s education, and other needs than people who do not have children.

The fifth is the ‘smoker’ column. The Smoking factor is also considered to be one of the Most Important factors as the people who smoke are always at risk when their age reaches 50 to 60.

Next is the ‘region’ column. Some Regions are Hygienic, Clean, Neat, and Prosperous, But some Regions are not, and this information affects health which is related to medical expenses.

**Steps:**

* Data Pre-Processing
* Exploratory Data Analysis
  + Data Visualization
  + Univariate Analysis
  + Data Processing
* Data Modelling
* Evaluation
* Conclusion
* Recommendations
* Future Work

**A snippet of how the data looks like:**

**Table

Description automatically generated**

**We’ll be using the dataset to model and predict the below:**

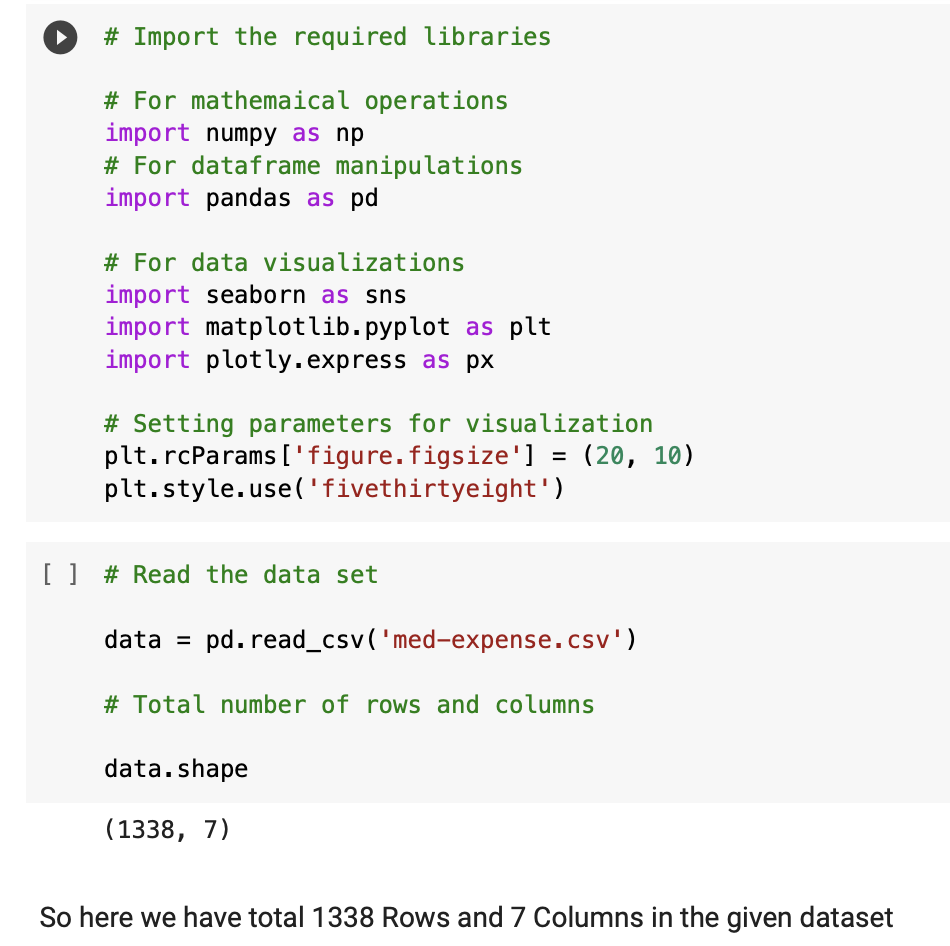
* *Performing EDA on the dataset [Data Pre-Processing and Data Visualization]:*
  + Predict the price of the listings by employing **Linear Regression, K-Nearest Neighbors Regression, Random Forest Regression** after proper data processing, label encoding, normalization & feature engineering.

**Platform used: Google Collab**

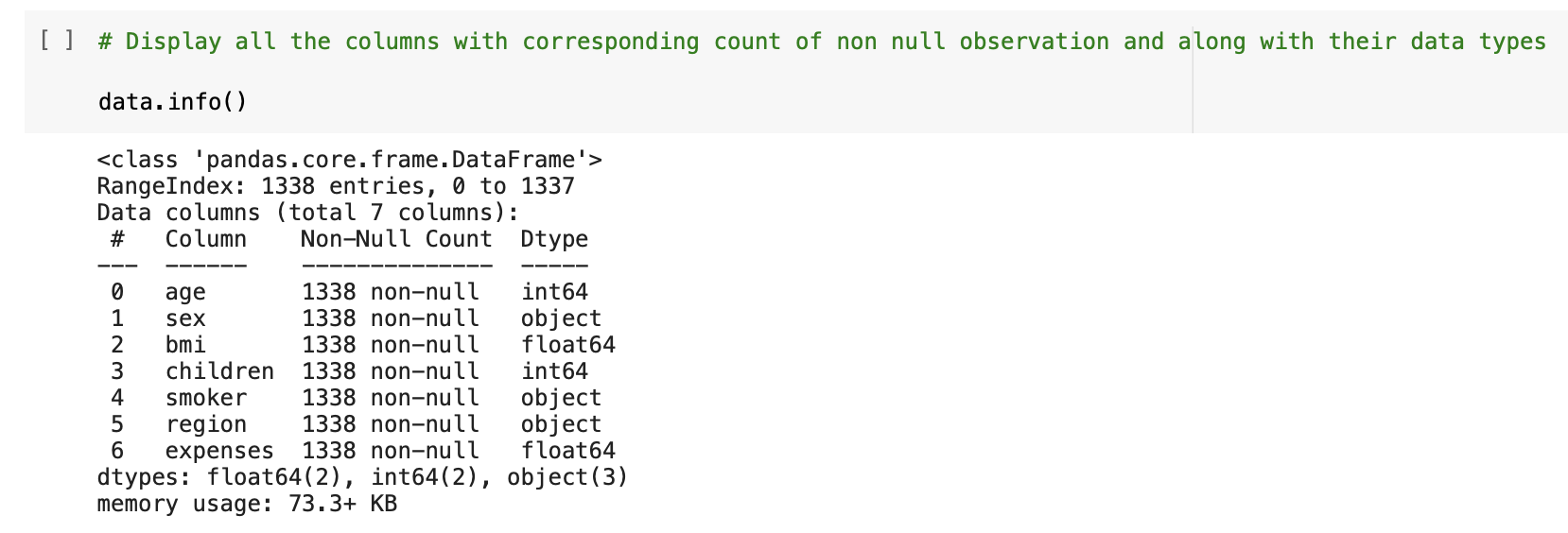
**Data Preparation**

**Data Pre-Processing:**

* Importing the required libraries for mathematical operations, data frame manipulations and data visualizations. Read the dataset.

****

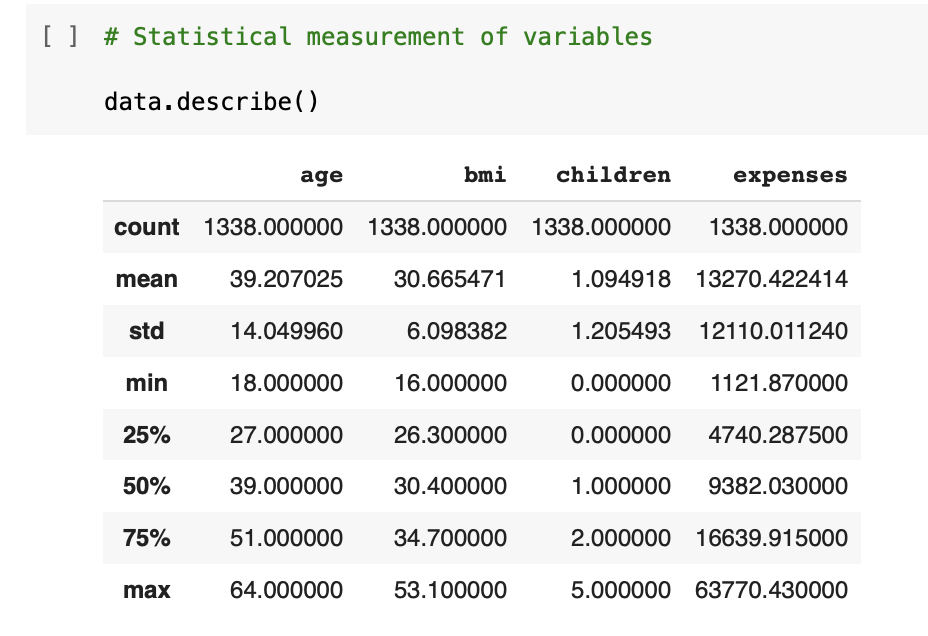
* Display all the columns with corresponding count of non-null observation and along with their data types

****

We observed that "age", "children", "bmi" and "expenses" are numbers, whereas "gender", "smoker" and "region" are strings (possibly categories).

None of the columns contain any missing values, which saves us a fair bit of work.

* Statistical Measurement of Variables



From the above dataset we can see that Age, BMI, Children, Expenses are only Numeric Values and get their corresponding statistical measurement.

Here from the above information, we can conclude that Age variable is symmetrically distributed as Mean = Median.

BMI is also symmetrically distributed.

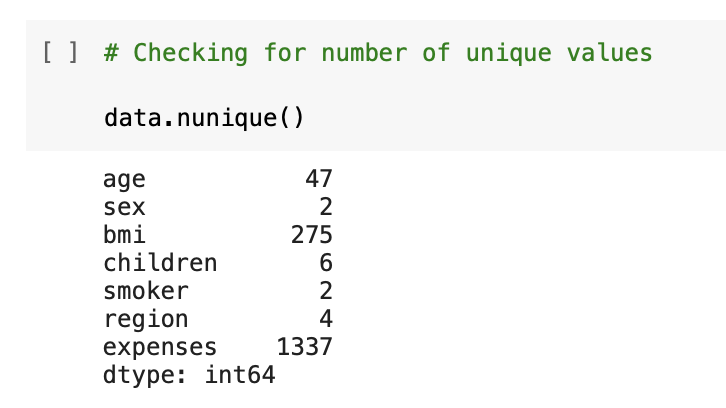
But Expenses variable is positively skewed as Mean > Median.

* Removing the Null values from the dataset using isnull().sum() & use dropna() if there are any null values.

Graphical user interface, text, application

Description automatically generated

* Check for number of unique values.



* Check for duplicates using duplicated() and drop the second duplicated row and check the count after dropping duplicate.

Graphical user interface, text, application

Description automatically generated

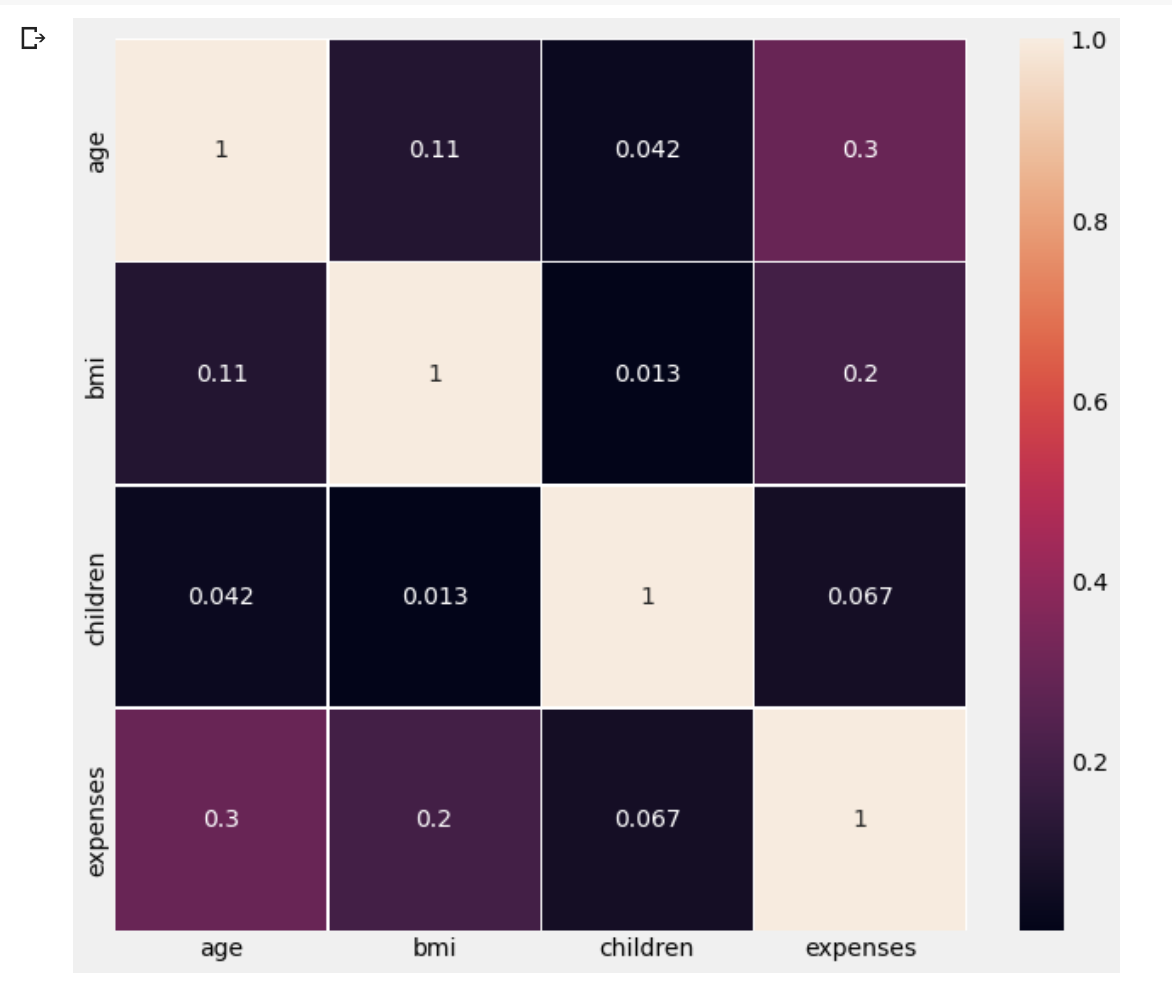
**Exploratory Data Analysis:**

**Data Visualization:**

**Chart

Description automatically generated**

**Visualizing data using Pair Plot**

****

**Correlation Matrix on Raw Data**

**Chart, bar chart, histogram

Description automatically generated**

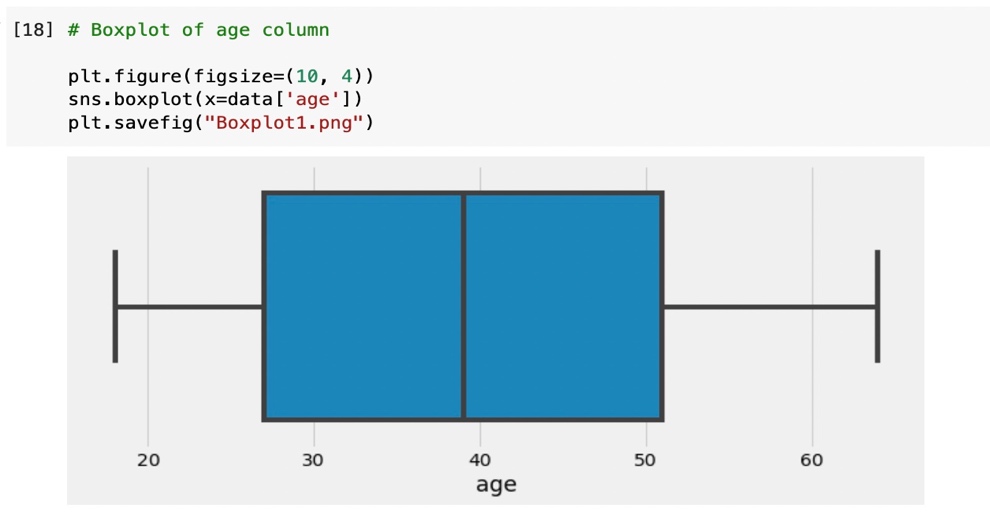
**Visualizing data using Histogram**

**Graphical user interface, application

Description automatically generated**

**Visualizing data using Box Plot on all numerical variables**

**Box Plots of individual variables:**

****

**Visualizing data using Box Plot on age column**

**Chart, box and whisker chart

Description automatically generated**

**Visualizing data using Box Plot on children column**

**Chart, histogram

Description automatically generated**

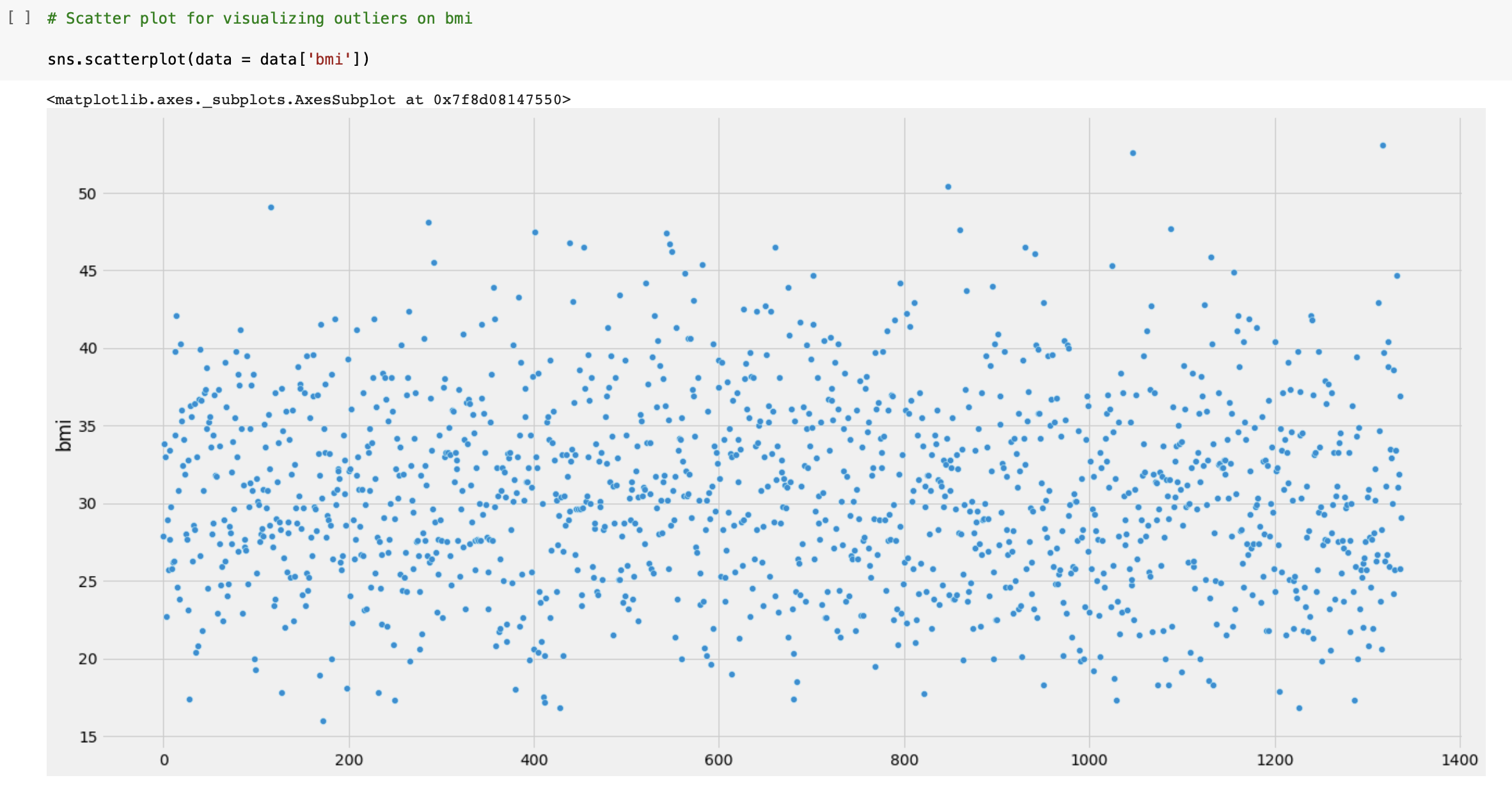
**Visualizing data using Box Plot on bmi column**

**Chart, bar chart

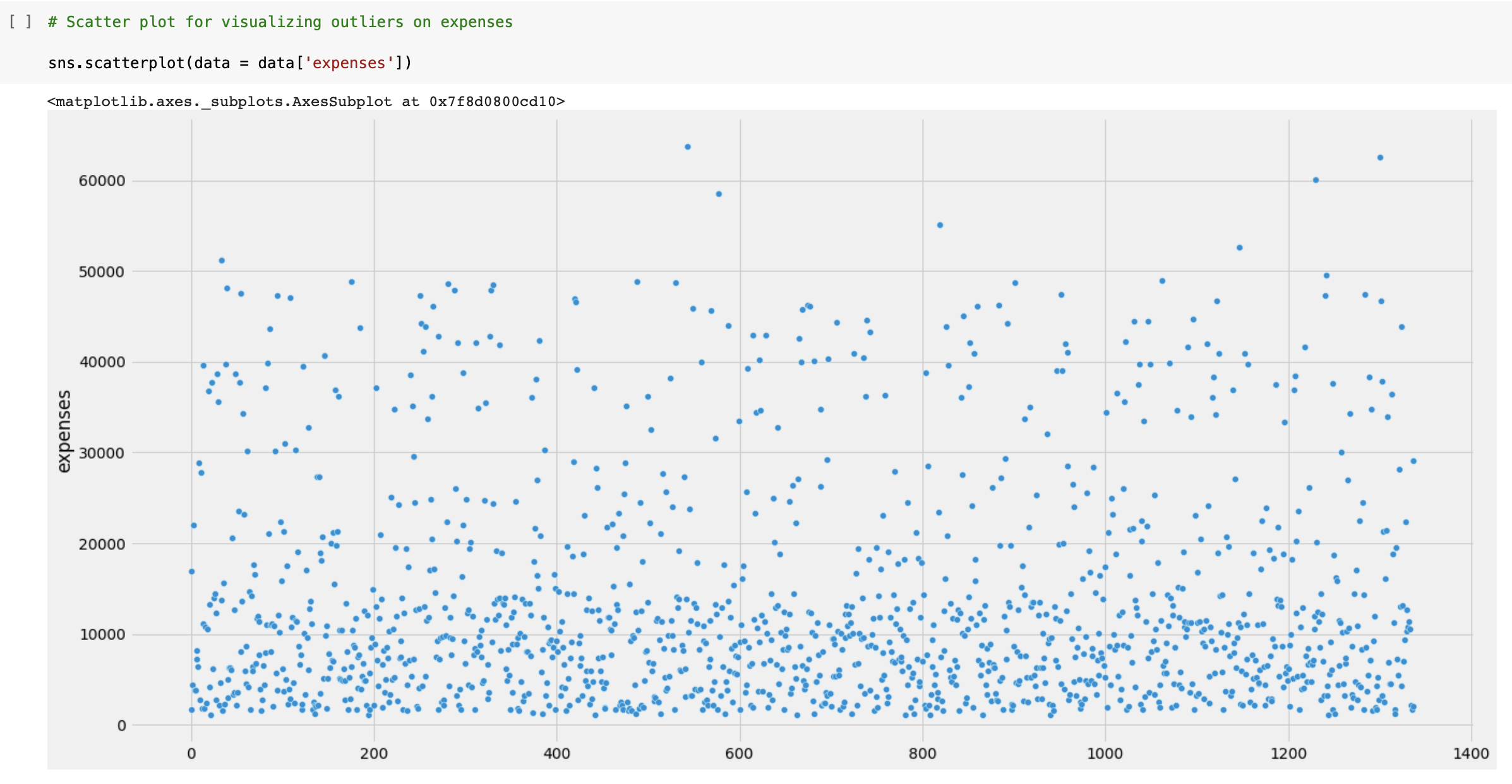
Description automatically generated**

**Visualizing data using Box Plot on expenses column**

**Using Scatter Plots to show outliers on bmi and expenses:**

****

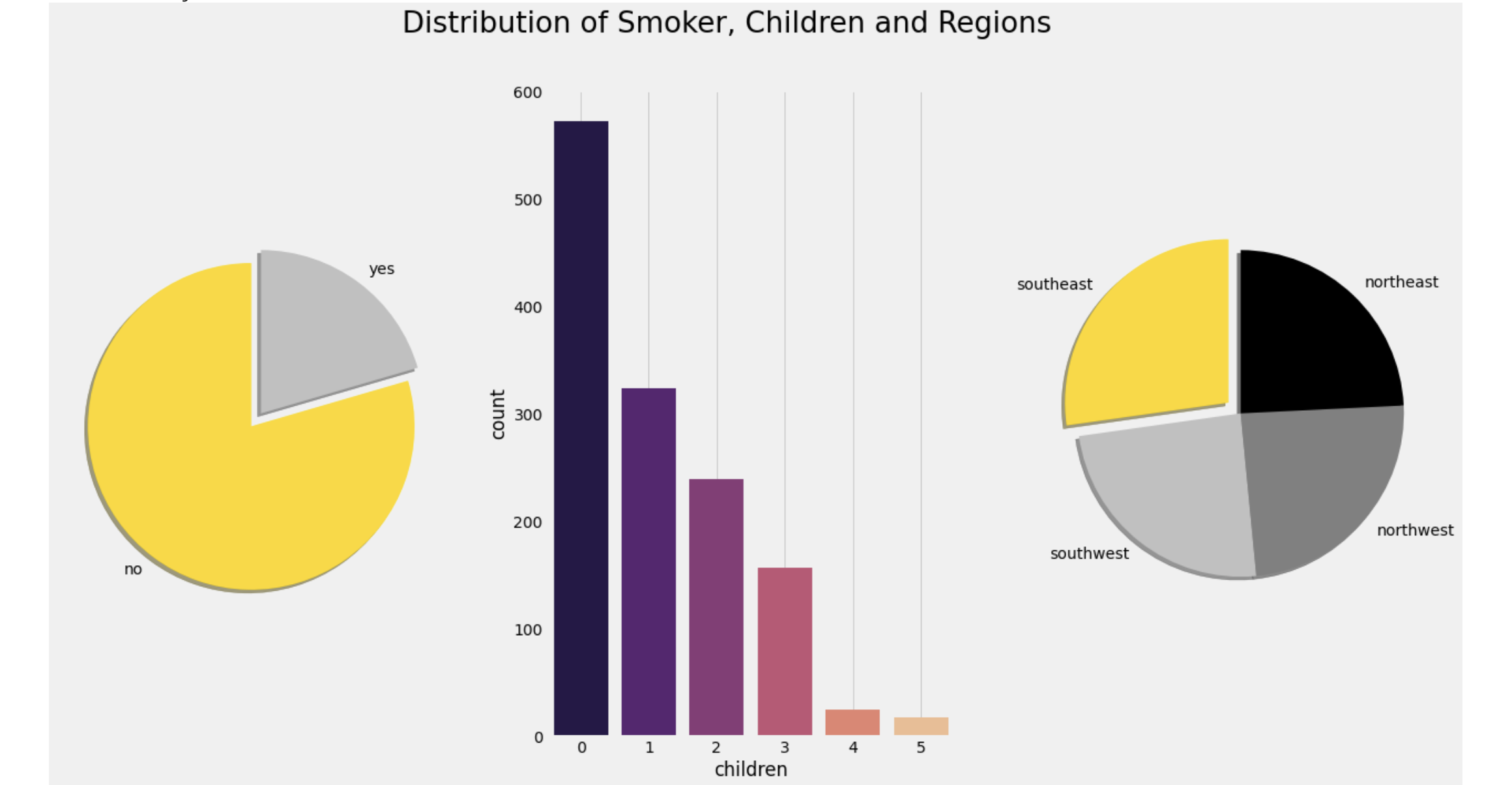
**Visualizing outliers using Scatter Plot on bmi column**

****

**Visualizing outliers using Scatter Plot on expenses column**

**Univariate Analysis:**

* Only one variable is involved.
* It's used to figure out how the variables in the dataset are distributed and to extract useful data from them.
* It can be used to examine both numerical and categorical variable distributions.

****

**Visualizing distribution of Smoker, Children and Regions**

Here we have used a pie chart to plot the Smoker Column, as the Smoker column has only two values: **Yes and No**. We have found 20.48% of the subjects are smokers and 79.52 % are non-smoker.

Using a Count plot, we have shown the subjects having children ranging from 0 to 5 and it has been computed and observed from the count plot also that those who are having no children are highest in number.

* Number of Subjects having no children- 574
* Number of Subjects having one child- 324
* Number of Subjects having two children- 240
* Number of Subjects having three children- 157
* Number of Subjects having four children- 25
* Number of Subjects having five children-18
* We have again used a pie chart to plot the number of inhabitants in the region column which consists of four segments: Northeast, Northwest, Southeast, Southwest

The number of Southwest and Northwest are the same and the value is 324, but the number of inhabitants in Northeast and Southeast are respectively 324 and 364.

Similarly, we plotted the distribution of Age, BMI and Expenses as below.

We have an equal number of people of all ages.

The BMI of the patients seems to be normally distributed where maximum people have BMI around 30 and very few people have less BMI around 10, similarly very few people have high BMI around 60.

****

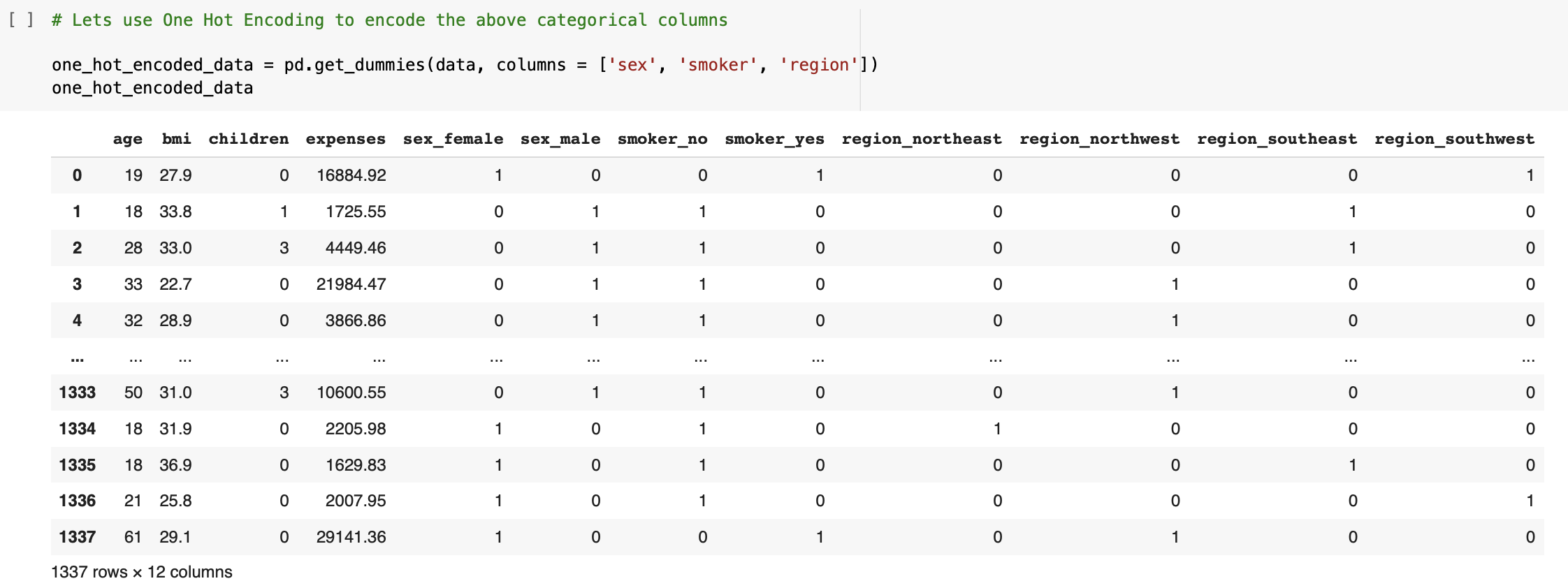
**Visualizing distribution of Age, BMI and Expenses**

**Data Processing:**

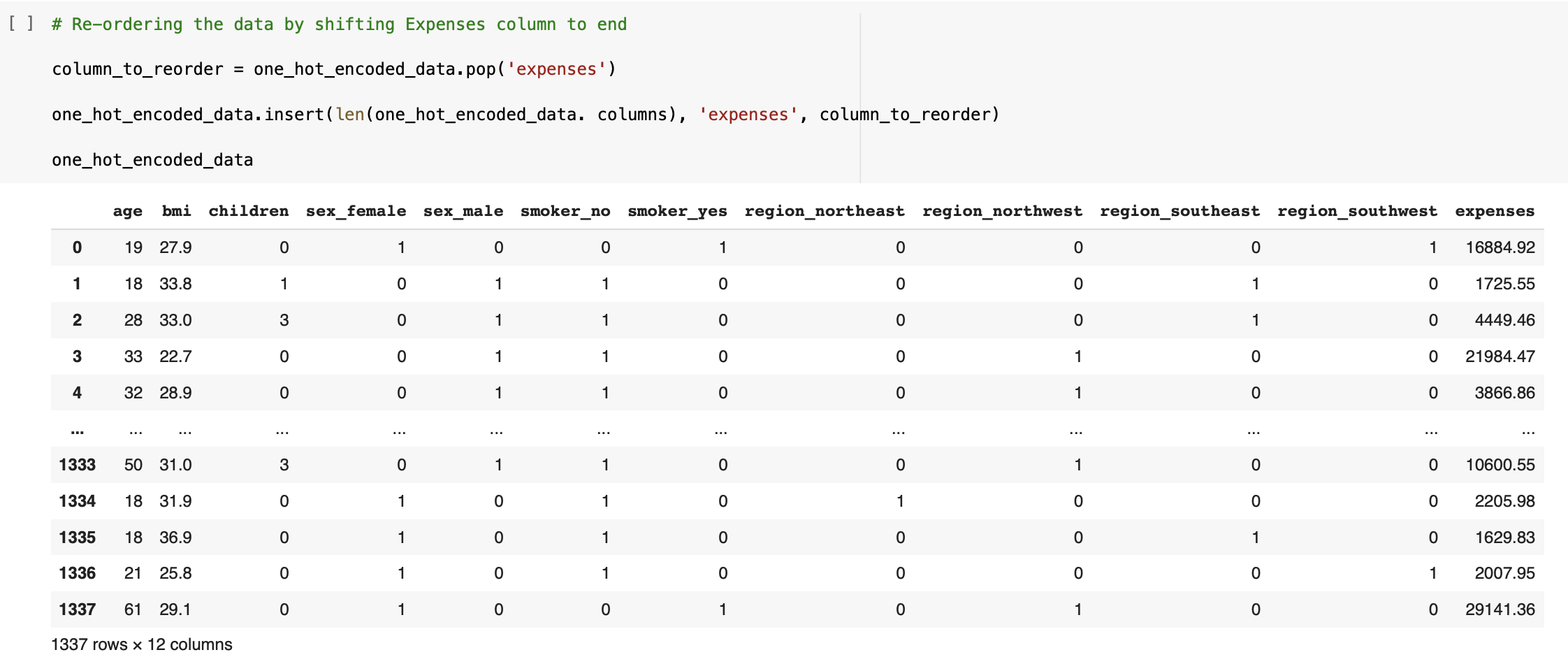
* Check for the categorical columns.

****

* We used **One Hot Encoding** method and encoded the categorical variables “sex”, “region”, and “expenses” to convert them to numerical variables for executing the correlation b/w the features

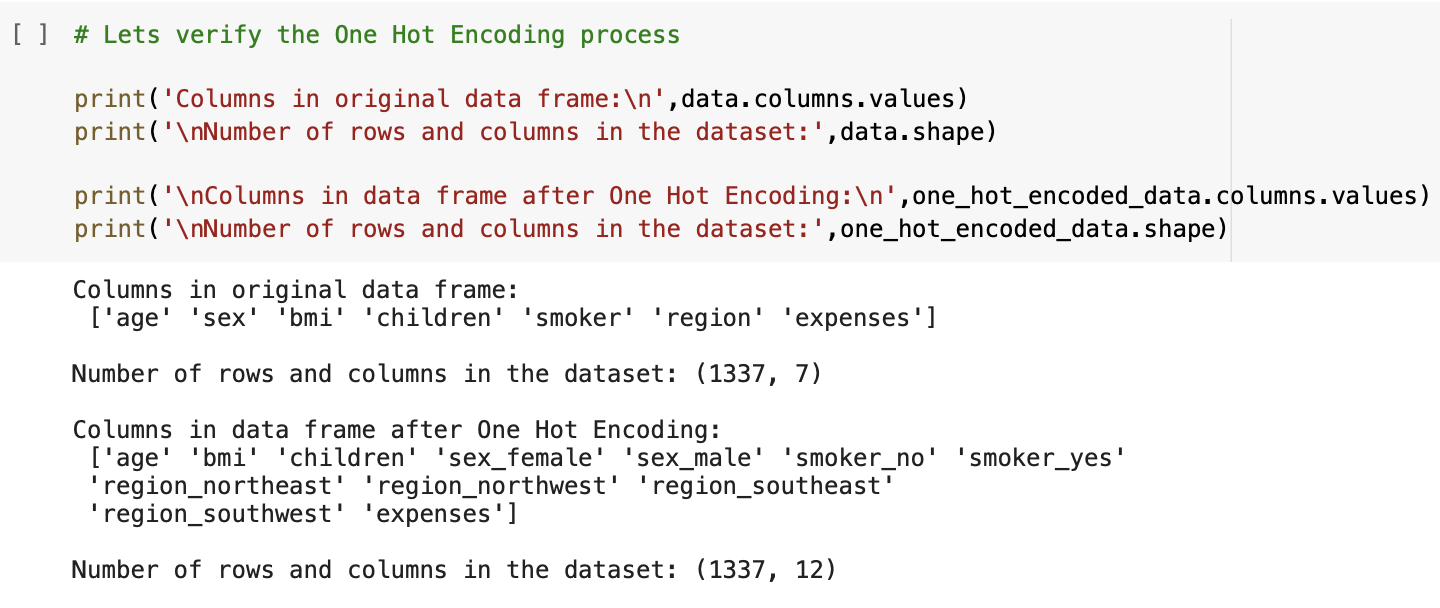


* After encoding, we reordered the data by shifting target variable to end of the columns

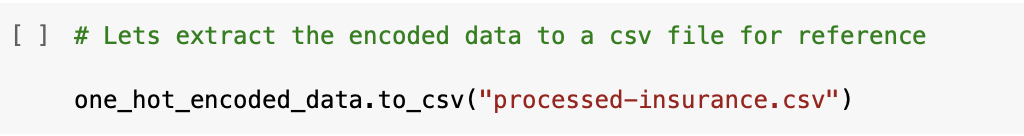


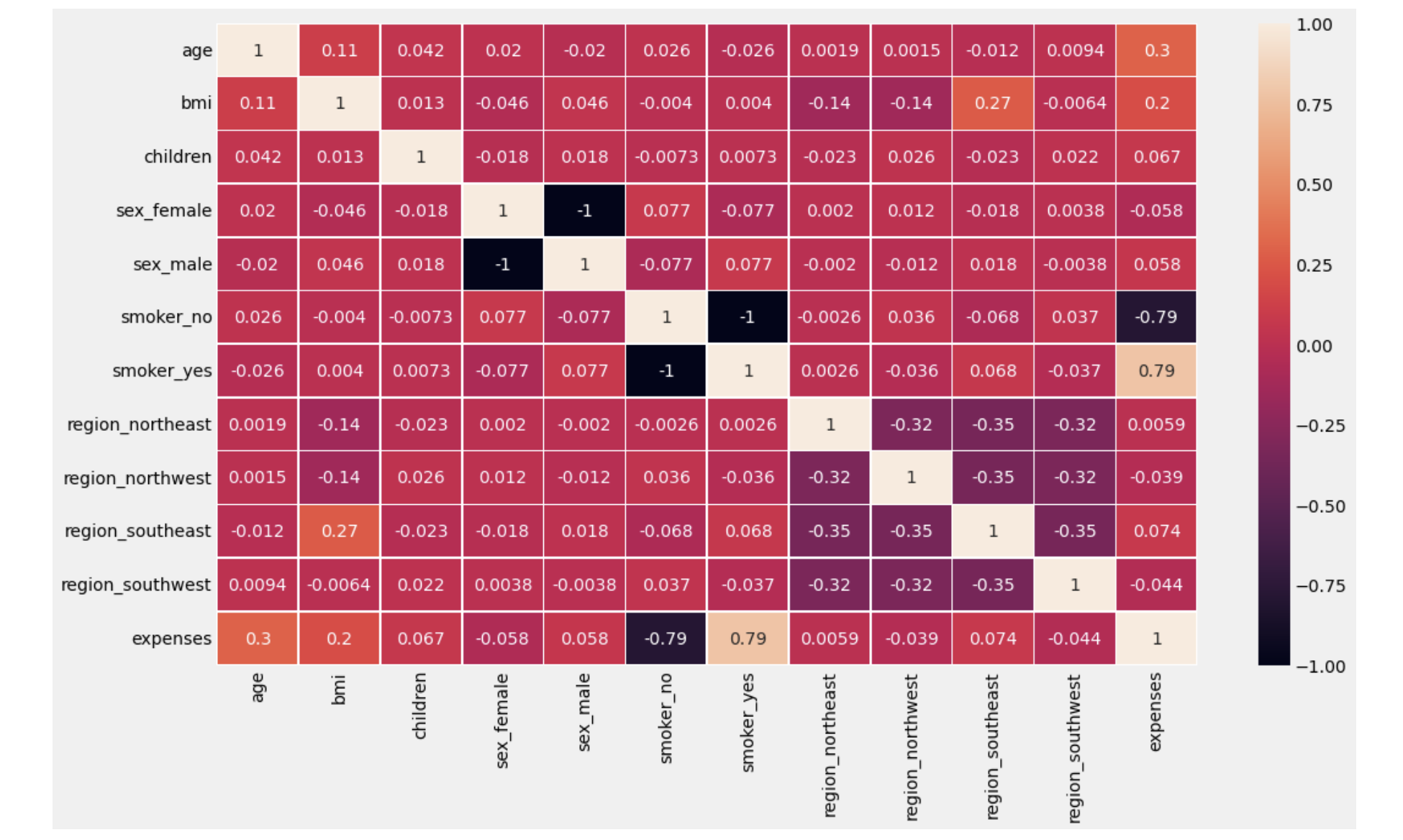
**A Snippet of Data after encoding categorical variables**

* Verifying the One Hot Encoding Process

****

* Extracting the encoded data to a csv file for reference

****



**Correlation Matrix on encoded data (Linear Correlation)**

From the above representation we can say that,

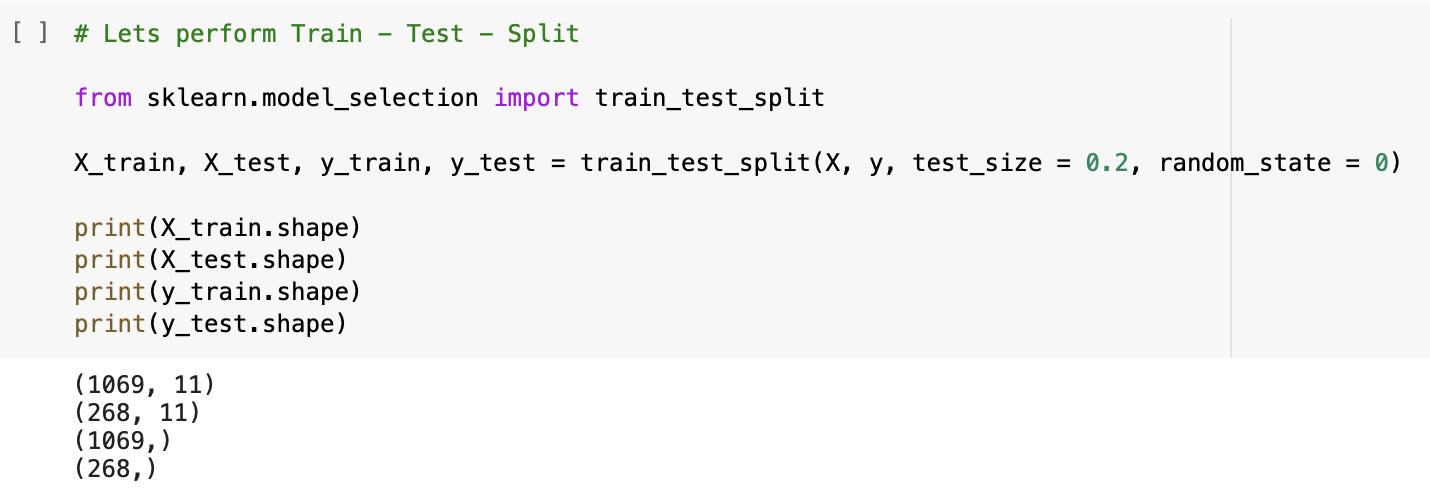
* In between Age and Sex, we have very weak Correlation, age and BMI have weak correlation, age and smoker also have weak correlation. Age and Expenses have a moderate correlation, Age and South-West region also Age and North-west region have a negative Correlation.
* Sex and Expenses have a negative Correlation.
* Correlation between Children and Expenses is very weak.
* Correlation between Smoker and Expenses is strongly Negatively Correlated.
* Correlation between South-East region and Expenses is weak.
* Correlation between South-West region and Expenses is weakly negative.
* Correlation between North-West region and Expenses is weakly negative.
* Correlation between North-East region and Expenses is zero.

**Data Partition:**

* Let’s form dependent and independent sets

****

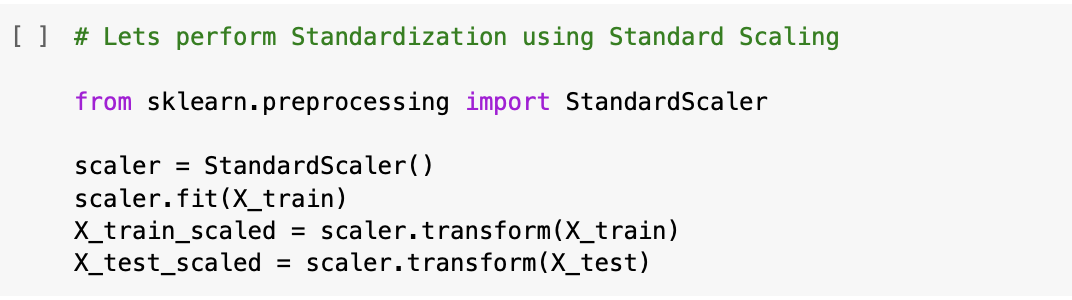
* For the modeling, we split the dataset into train and test using train\_test\_split().



**We have split the data into 80 : 20**

**Feature Scaling:**

* Feature scaling is a technique for normalizing a set of independent variables or data features.
* Data Normalization is another name for feature scaling.
* It helps in the normalization of data within a particular range.
* It also helps in the speeding of algorithmic calculations.
* We performed Standardization using Standard Scaling by using StandardScaler() on the data before building the model as the values for a few features were exorbitantly high.



**Normalizing the data using Standard Scaler Technique**

**Data Modeling**

The first few attempts at model building for this project were performed using regression models from several libraries.

The models were:

* Linear Regression Model
* K-Nearest Neighbors Model
* Random Forest Model
* Ensemble of all above three models
* Weighted Average of all three models
* Cross Validation of all three models
* Comparing performance of all models

**Evaluation Metrics**:

* **R2** **Score**
  + It is generally used to determine the strength of correlation between the independent features and the target column.
* **Root Mean Squared Error(RMSE)**
  + It is the square root of the mean of the differences between actual and the predicted values.

Data used for the model building was divided with 80% used for training and the remaining 20% for testing. The models were then fitted using the training data set and predictions made on the test data set. Accuracy using the R2 scoring metric was recorded.

**Linear Regression**

* A linear approach to modeling the relationship between a dependent and one or more independent variables is known as linear regression.
* The dependent and independent variables must have a linear relationship.
* Two variables must have trend lines that are either increasing or decreasing.
* In statistical terminology, all variables must have the same variance.

As our target variable is continuous dependent variable, Linear Regression is the best choice for predicting the future medical expenses.

* Here we predicted future medical expenses using Linear Regression on unscaled data and scaled data.
* We computed R2 and RMSE values obtained as **0.7530** and 6445.47 for unscaled data.

**Code Snippet for the Linear Regression model depicting the RMSE and R2** **values**

Graphical user interface, text, application

Description automatically generated

**Linear Regression Model – Unscaled Data**

* For scaled data we got **0.7534** and 6440.49 as R2 and RMSE values.

**Code Snippet for the Linear Regression model depicting the RMSE and R2** **values**

Graphical user interface, text, application

Description automatically generated

**Linear Regression Model – Scaled Data**

We observed that for scaled data, the Accuracy has been improved.

**K-Nearest Neighbors Model**

* K-NN is a simple algorithm that stores all available cases and predicts the numerical target variable base on similarity measure.
* This means that the new point is assigned a value based on how closely it resembles the values in the training set.
* Here we predicted future medical expenses using K-NN model on unscaled data and scaled data.
* We computed R2 and RMSE values obtained as **0.3391** and 10543.63 for unscaled data.

**Code Snippet for the K-NN model depicting the RMSE and R2** **values**

Graphical user interface, text, application

Description automatically generated

**K-NN Model – Unscaled Data**

* For scaled data we got **0.7982** and 5825.76 as R2 and RMSE values.

**Code Snippet for the KNN model depicting the RMSE and R2** **values**

Graphical user interface, text, application

Description automatically generated

**K-NN Model – Scaled Data**

We observed that for scaled data, the Accuracy has been improved drastically.

This proves that **K-NN model is best when working with scaled data.**

**Random Forest Model**

* Random forest is an ensemble learning method for classification and regression by constructing multiple number of decision trees at training time.
* And it outputs the average prediction of the individual trees in case of regression and mode of the classes in case of classification.
* It is one of the Most powerful Machine Learning algorithms which works well in most cases.
* Here we predicted future medical expenses using Random Forest on unscaled data and scaled data.
* We computed R2 and RMSE values obtained as **0.8375** and 5226.96 for unscaled data.

**Code Snippet for the Random Forest model depicting the RMSE and R2** **values**

Text

Description automatically generated

**Random Forest Model – Unscaled Data**

* For scaled data we got **0.8368** and 5238.56 as R2 and RMSE values.

**Code Snippet for the Random Forest model depicting the RMSE and R2** **values**

Graphical user interface, text, application, email

Description automatically generated

**Random Forest Model – Scaled Data**

**Ensemble of all above three models**

* Ensemble models is a machine learning approach to combine multiple other models in the prediction process.
* Those models are referred to as base estimators.
* It is a solution to overcome the technical challenges of building a single estimator
* Here we predicted future medical expenses using Ensemble by average of all models on unscaled data and scaled data.
* We computed R2 and RMSE values obtained as **0.7617** and 6330.46 for unscaled data.

**Code Snippet for Ensemble of all models depicting the RMSE and R2** **values**

Graphical user interface, text, application

Description automatically generated

**Ensemble by Average – Unscaled Data**

* For scaled data we got **0.8266** and 5399.65 as R2 and RMSE values.

**Code Snippet for Ensemble of all models depicting the RMSE and R2** **values**

Text

Description automatically generated with medium confidence

**Ensemble by Average – Scaled Data**

**Weighted Average of all three models**

* We created a weighted average of all three models with below weightage:
  + Random Forest – 50%
  + KNN – 30%
  + Linear Regression – 20%

**Graphical user interface, text, application

Description automatically generated**

**Cross Validation – 5 Fold**

* Cross Validation is a resampling procedure which is used to evaluate the machine learning models on limited data samples.
* The goal of cross validation is to test the model’s ability to predict new data that was not used in estimating it.
* It has a single parameter called **k**, which indicates the number of groups that the data would be split into.
* Here, we set the value of k = 5.

**A picture containing text

Description automatically generated**

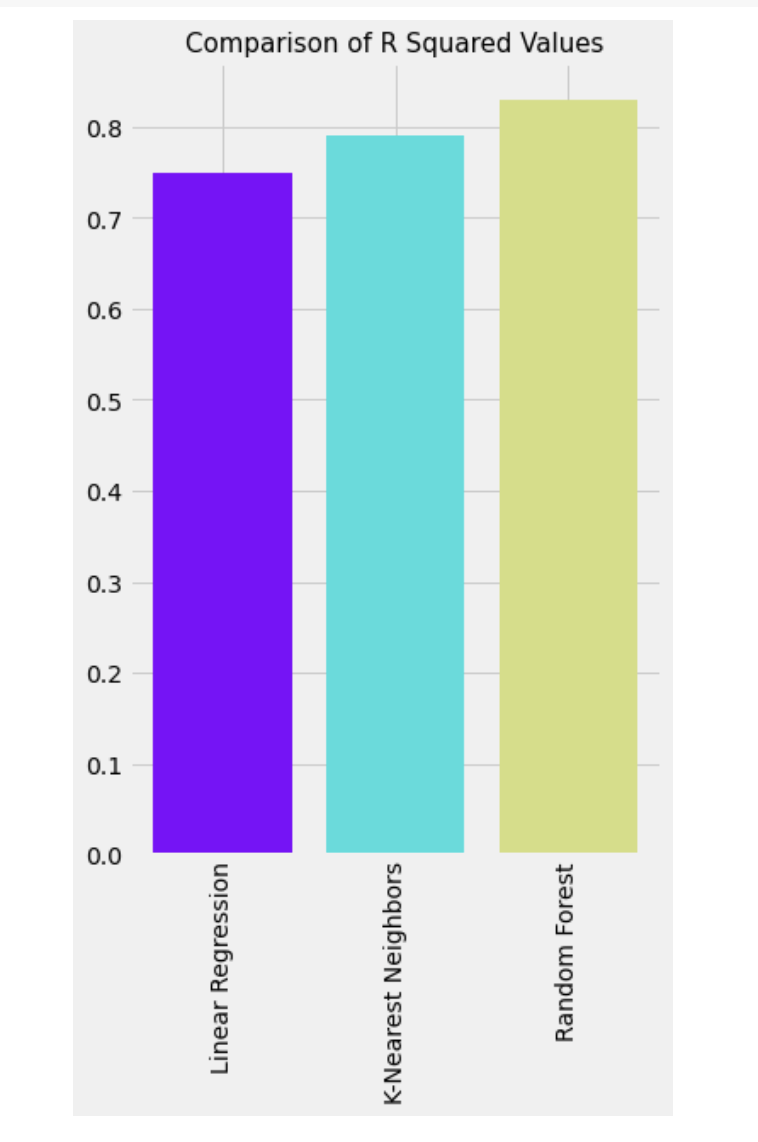
**The 5-fold Cross Validation scores using Random Forest model:** [0.85036998 0.77562962 0.86951228 0.83052304 0.85310877]

**Comparison of all three models**

* We have created a NumPy array of the R2 score of all three models, Linear Regression, Random Forest, and K-NN.
* An array for the labels was also created as well, to compare these Values using bar charts.
* Here a Rainbow palette has been used and the Bar plot built using the seaborn Library shows a higher R2 score value for Random Forest and lowest for Linear Regression.
* That means, the **Random Forest Model is the Best Choice** whereas the Linear Regression Model is the worst for this Case.
* We have successfully built our predictive model and compared these predictive models based on their accuracies and Results.

****

Below is the bar plot to show the performance of the three used models.

****

**Evaluation**

* We came to know that the Most Important Factor to Predict the Medical Expenses of a subject is Smoking Behavior and Age, that means, smoking is Bad for Health, as we already know that and which inevitably increases medical expenses as due to smoking one is likely to fall ill more than the nonsmokers.
* We also found that with increasing of age, one needs to take some more care and precautions for your health as with the increase of age health becomes fragile, so they go for frequent medical check-up, likely to fall ill quickly as with the increase of age immunity falls so they adopt measures to stay healthy by taking medicines and engaging in some physical activities.
* Apart from that we also understood that Gender, Number of Children, the Region also have a good impact on determining Medical Expenses.

**Conclusion**

We have built three models among which the **Random Forest Regressor model** shows the best result through which we can say **83.75%** variability of expenses can well be explained by predictor variables and which yields comparatively low RMSE value so our predicted expense through this model will not vary too much from the actual expense.

**Recommendations**

***The accuracy can be much more enhanced with the following recommendations:***

* The data sample size is relatively small in this case, especially with a perspective of applying this in real world.
* The number of features is also limited in this dataset. With an addition of 5 or more valid/logical features, the predictions can be improved much further.
* We can try converting the Expense column to a normal distribution using log or square root transformation for removing the skewness.
* Future enhancements to this project could include experiments with additional predictive models such as Gradient Boosting or SVM models. Improved accuracy of the linear regression models is likely with additional preprocessing of the data to better conform to the assumptions of linear modeling.

**Future Work**

* We plan to take this project further as health is center of everyone’s life and every part of our life relies on good health.
* We intend to develop an application to predict possible medical costs using PyCharm editor on AWS platform by plugging in basic details such as age, sex, smoker, children, region and then pressing a submit button.
* This application will help people understand the factors that are making them unfit so that they can reduce their medical expenses.
* We can also use this health expense predictor in hospitals so that patients can adjust their medical expenses.