Project Title: - Health Insurance Premium Prediction Using Machine learning

Introduction: -

Health or Medical Insurance is an Insurance Policy that ensures that you get a cashless treatment, in case you fall ill. Here using machine learning for predicting the premium of health insurance in Python.

The amount of the premium for a health insurance policy depends from person to person, so there are many factors that affect the amount of the premium for a health insurance policy.

The factors like age, bmi, sex etc.

Problem Statement: -

Health insurance premium prediction with machine learning using Python**.**

Flow Chart: -

1. Data Collection
2. Data Analysis
3. Data Pre-processing
4. Train and Test Split
5. Prediction of accuracy using Linear Regression Model and Random Forest Regression.

Data Collection: -

The dataset that I am using for the Health insurance premium prediction is collected from Kaggle.

The Dataset Contains

1. the age of the person
2. gender of the person
3. Body Mass Index of the person
4. how many children the person is having
5. whether the person smokes or not
6. the region where the person lives
7. and the charges of the insurance premium

The Data set having 9 Columns like

'age', 'sex', 'bmi', 'children', 'smoker', 'region', 'charges'

Dataset:

	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

Data Cleaning: -

1. After Loading the Data set the shape of the Data set is (1338,7)
   1338 rows and 7 columns
2. We have to Check for the Duplicates, NULL Values and the Data Types
3. Check all the columns are unique

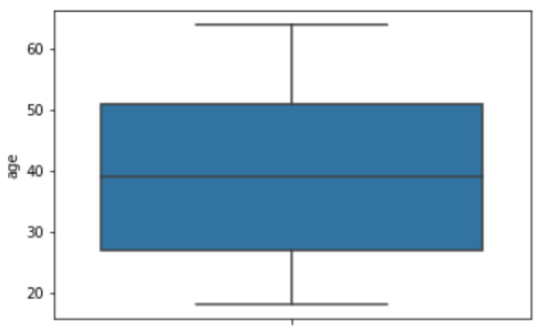
Here I found all the Datatypes, Columns, and also no Null values and Duplicates

```
<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
```

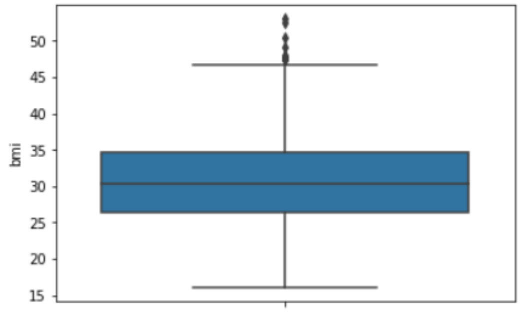
Exploratory Data Analysis:
Data Visualisations

1. Check for the outliers, Here I found there are no outliers in the data

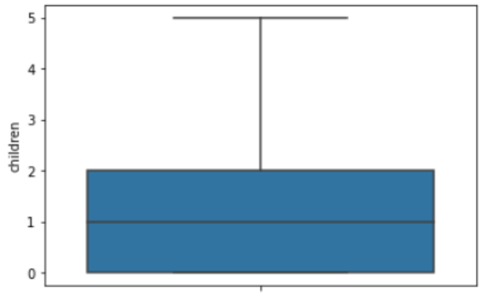
: <AxesSubplot:ylabel='age'>



<AxesSubplot:ylabel='bmi'>

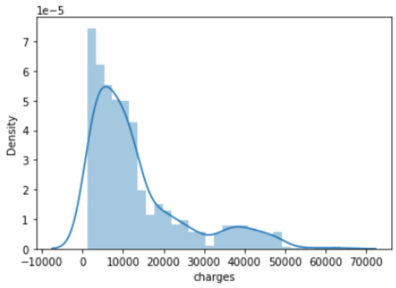


<AxesSubplot:ylabel='children'>

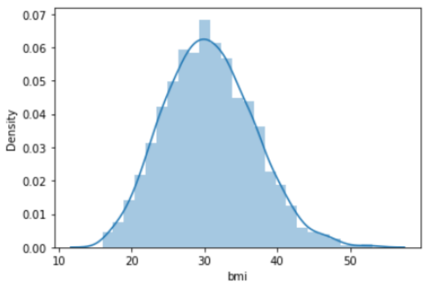


2. After clear analysis of data, it is observed that we have
   - Analysis on Numerical Data

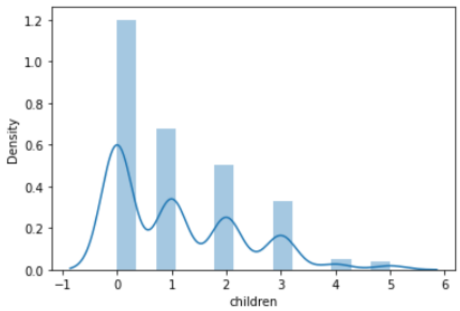
<AxesSubplot:xlabel='charges', ylabel='Density'>



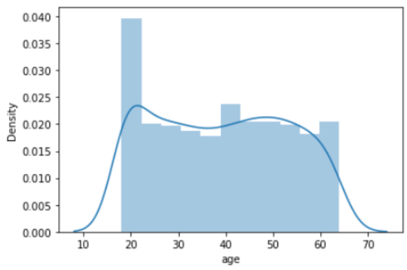
<AxesSubplot:xlabel='bmi', ylabel='Density'>



<AxesSubplot:xlabel='children', ylabel='Density'>

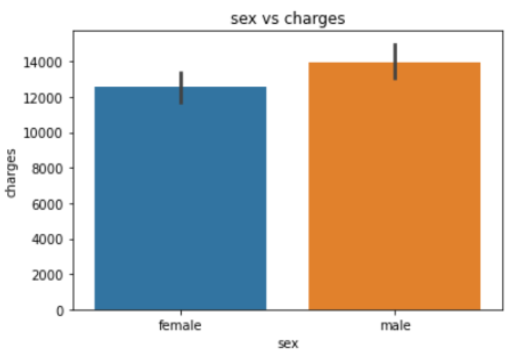


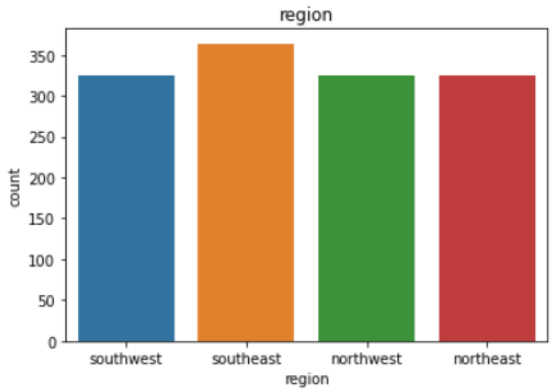
<AxesSubplot:xlabel='age', ylabel='Density'>

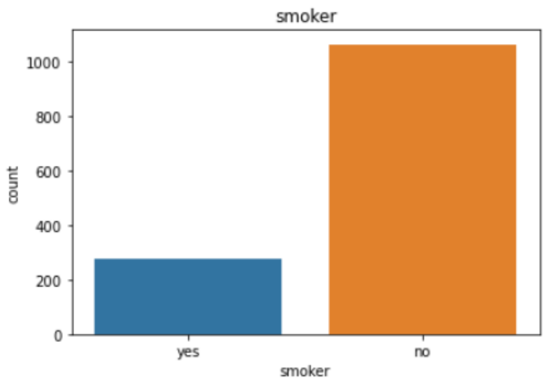


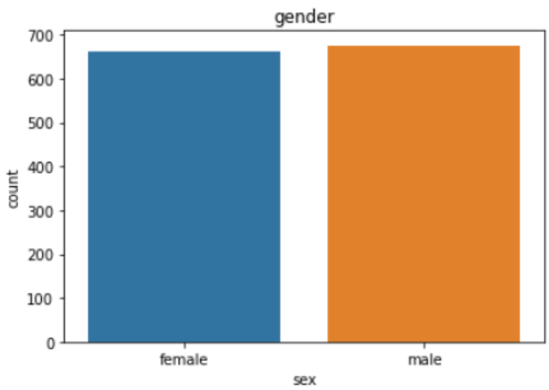
From the above analysis, we can observe that
Age: It is following almost uniform distribution and it seems like there are more customers of age between 18 to 20.
BMI: It following normal distribution approximately mean=30 and there are very few outliers present in this feature that can be ignored.
Children: Here, most customers have no children
Charges: It following Power Law Distribution and highly right skewed and Also, for most customers the annual charges are under 10k

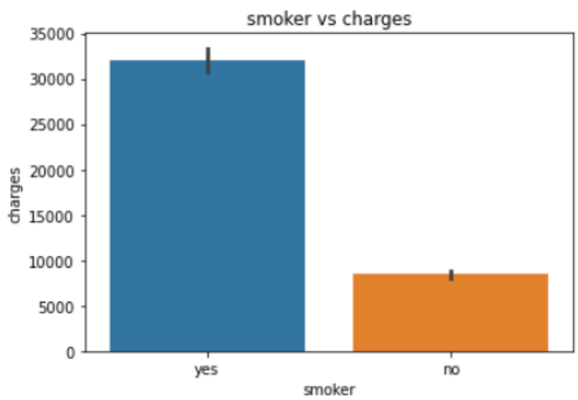
➢ Analysis on Categorical Data







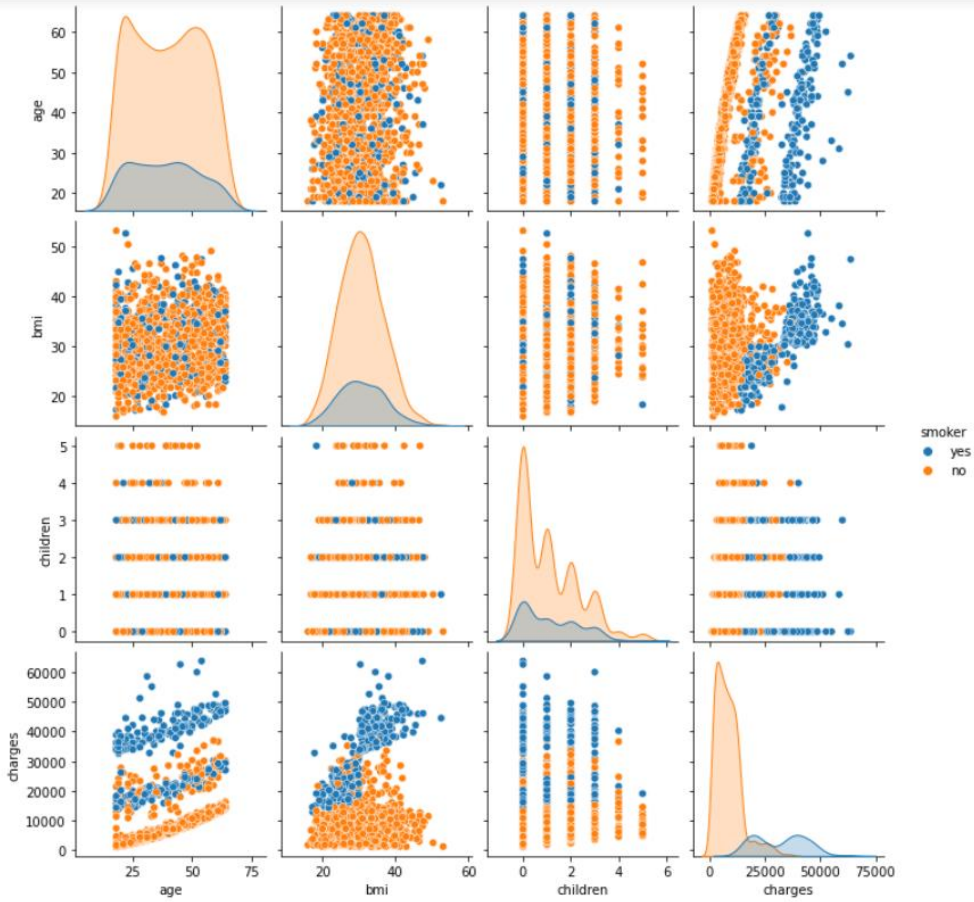




From the above analysis, we can observe that

- The Data is having same number of males and females
- Here we can see only 20% of the customers having smoking habit
- Here, the data is almost same for all regions

- In this , we can observe females contains less charges compared to male and Non-smoker contains less charges than smoker



In the above Pair Plot

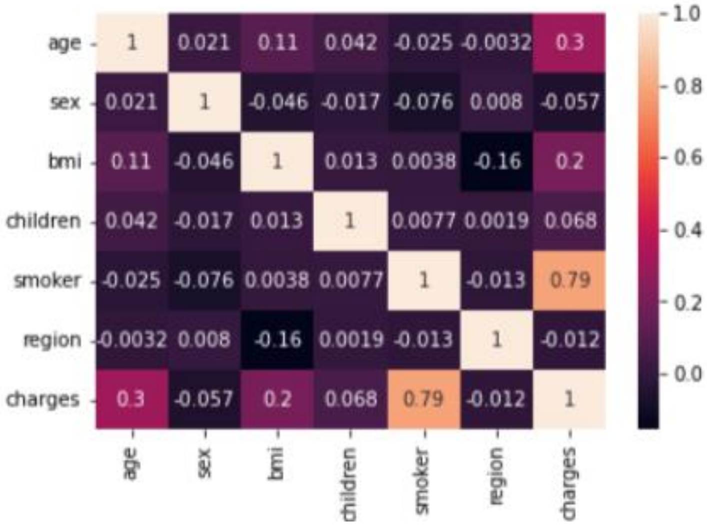
- We can observe that, here person who has more bmi and also having smoking habit is paying more charges.
- Person with less age having less charges compare to more age

Data Pre-processing:

Converting Categorical Variable into Numerical Variable

	age	sex	bmi	children	smoker	region	charges
0	19	1	27.900	0	1	1	16884.92400
1	18	0	33.770	1	0	2	1725.55230
2	28	0	33.000	3	0	2	4449.46200
3	33	0	22.705	0	0	4	21984.47061
4	32	0	28.880	0	0	4	3866.85520

```
<AxesSubplot:>
```



- In the above Heat Map here, we can see smoker, bmi and age have more correlation to charges

Model Building:

Train and Test Split:

- The test set is 20% of overall dataset

Prediction of accuracy using Linear Regression Model and Random Forest Regression:

1. The Performance metrics after running the model with Linear Regression Method

```
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

```
mse=mean_squared_error(y_test,predictions)
np.sqrt(mse)
```

```
6199.25545970555
```

```
mae=mean_absolute_error(y_test,predictions)
mae
```

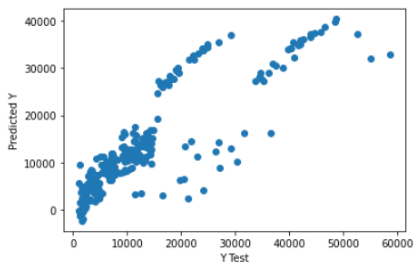
```
4287.000383987176
```

```
r2_score(y_test, predictions)
```

```
0.7441032539760071
```

```
import matplotlib.pyplot as plt
plt.scatter(y_test,predictions)
plt.xlabel('Y Test')
plt.ylabel('Predicted Y')
```

```
Text(0, 0.5, 'Predicted Y')
```



- The above scatter plot is Actual vs Predicted values, in this plot we can see that Non-Linear Correlation

2. The Performance metrics after running the model with Random Forest Regressor Method

```
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

```
mse=mean_squared_error(y_test,prediction)
np.sqrt(mse)
```

```
4954.69444467886
```

```
mae=mean_absolute_error(y_test,prediction)
mae
```

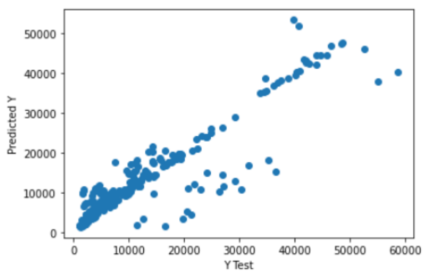
```
2776.4285991264555
```

```
r2_score(y_test, prediction)
```

```
0.8365370054772243
```

```
import matplotlib.pyplot as plt
plt.scatter(y_test,prediction)
plt.xlabel('Y Test')
plt.ylabel('Predicted Y')
```

```
Text(0, 0.5, 'Predicted Y')
```



- The above scatter plot is Actual vs Predicted values; in this plot we can see that Linear Correlation.

3. The Performance metrics r2_score higher in Random Forest regressor method when compared to Linear Regression method.
4. According to this Health Insurance Dataset Random Forest Regressor is the Best fit Model.

Conclusion: -

- As bmi, number of children and age increases the insurance charges also increases.
- The insurance charges for male are little bit more when compared to female.
- The smoker has high correlation with charges and the charges for smoker is more than Non-smoker.
- Among two algorithms Random Forest regressor was the best.