Exploratory Data Analysis Project Report

By Abhishek Singh

In this project I analyzed a sample insurance data about the individuals age ,sex ,smoking status , region and the charges with the insurance company and analyzed and verified some rational correlations expected between them using python

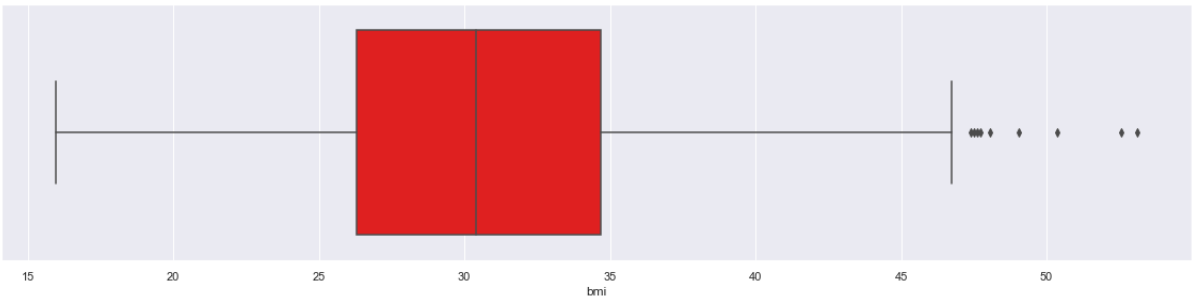
1)Summary of Data Set:

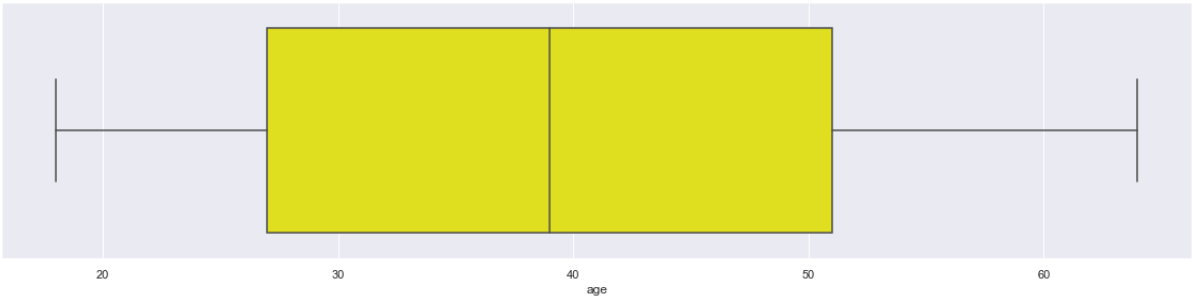
Range Index: 1338 entries, 0 to 1337
Data columns (total 7 columns):

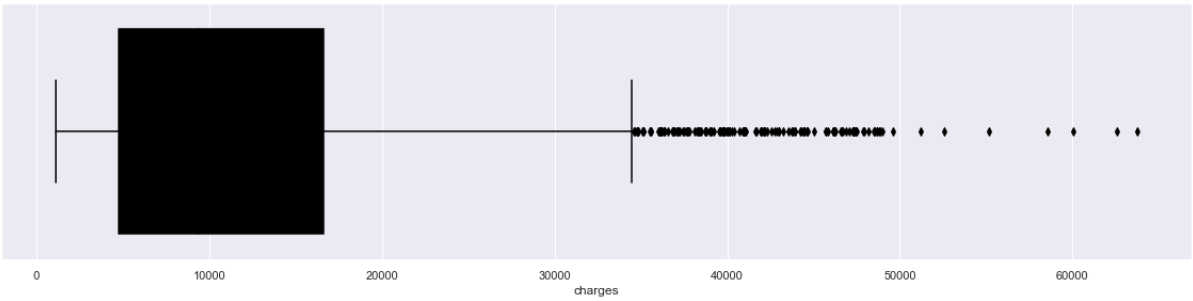
#	Column	Non-Null Count	Dtype
0	age	1338 non-null	int64
1	sex	1338 non-null	object
2	bmi	1338 non-null	float64
3	children	1338 non-null	int64
4	smoker	1338 non-null	object
5	region	1338 non-null	object
6	charges	1338 non-null	float64

dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB

2)Outliers Data Analysis





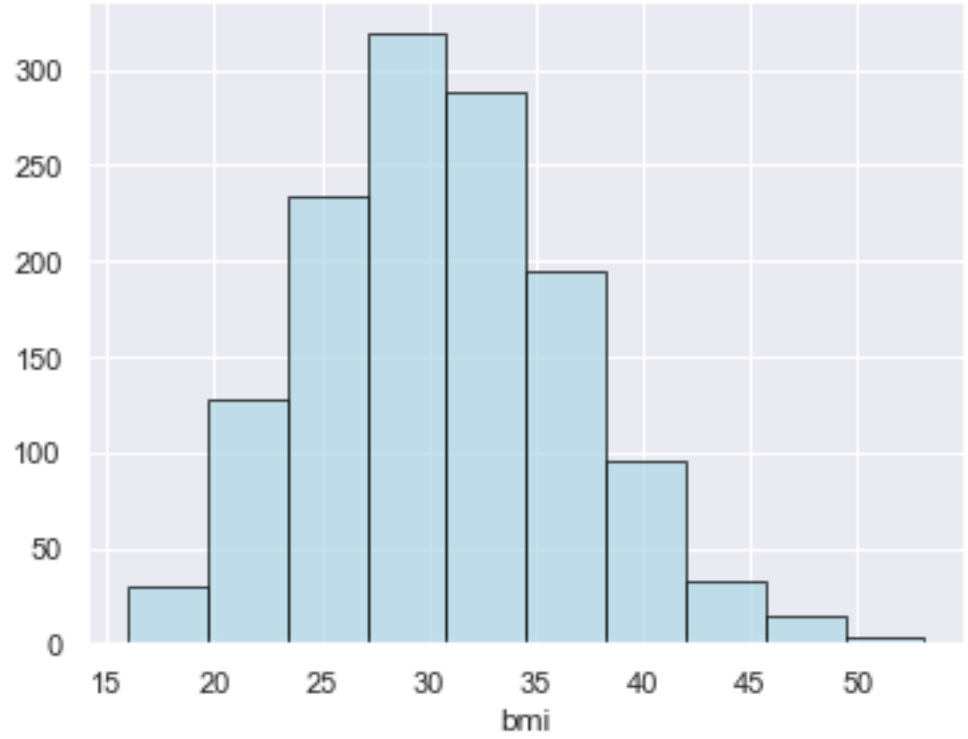


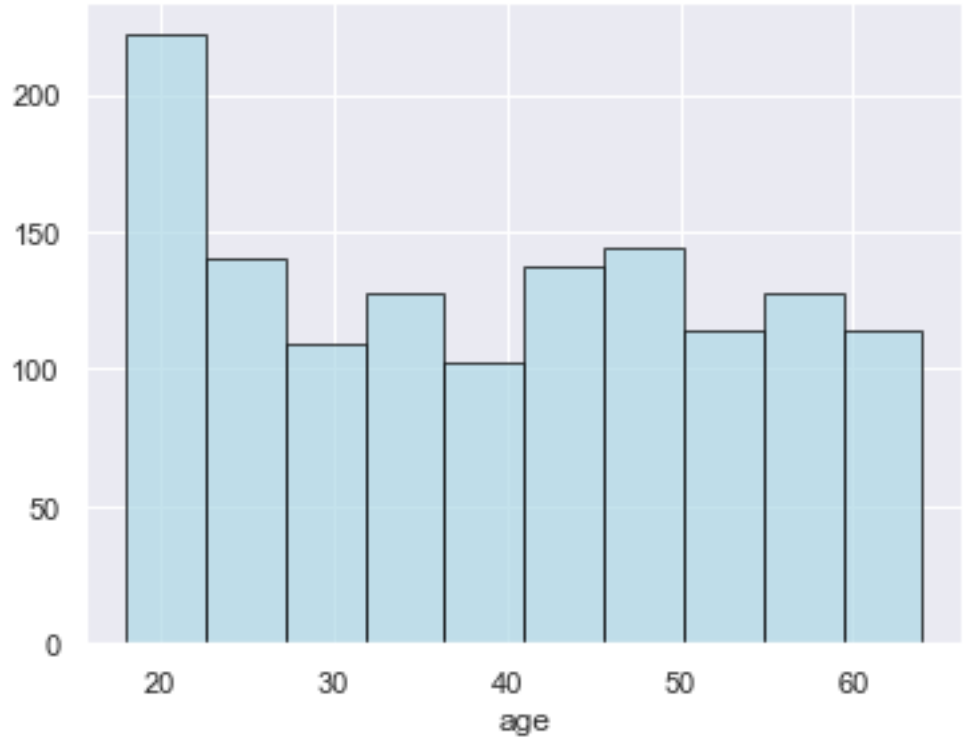
Inferences:

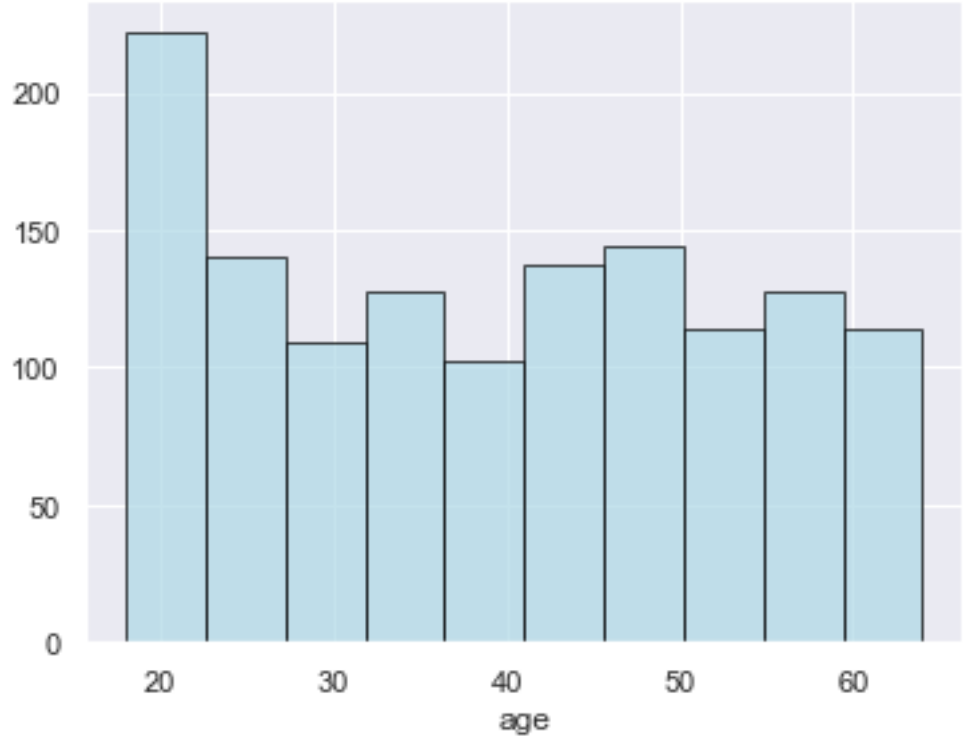
->BMI has a few extreme values.

->Charges as it is highly skewed, there are quiet a lot of extreme values.

3)Plots to see the distribution of the continuous features individually







Inferences:
->Skewness of bmi is very low as seen in the previous step

->Age is uniformly distributed and thus not skewed

->Charges are highly skewed

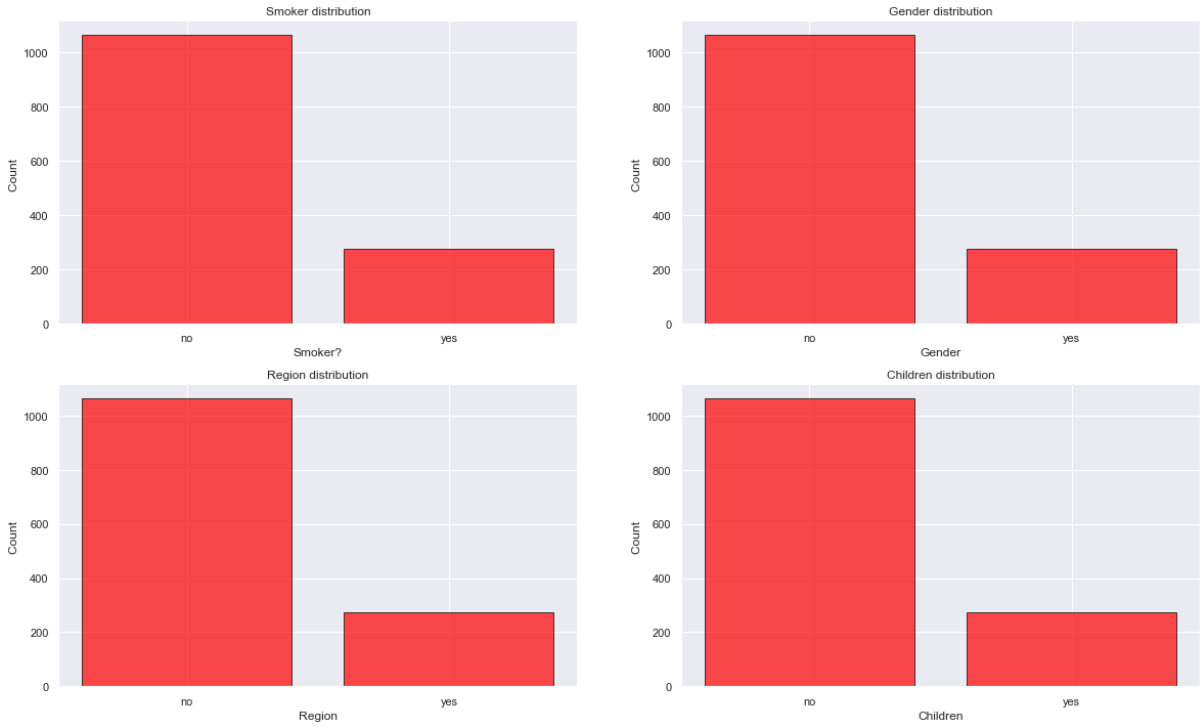
4)Skewness of the Variables

Bmi - 0.283

Age. - 0.055

Charges - 1.514

5)Visual Data Analysis from graphs



Inferences

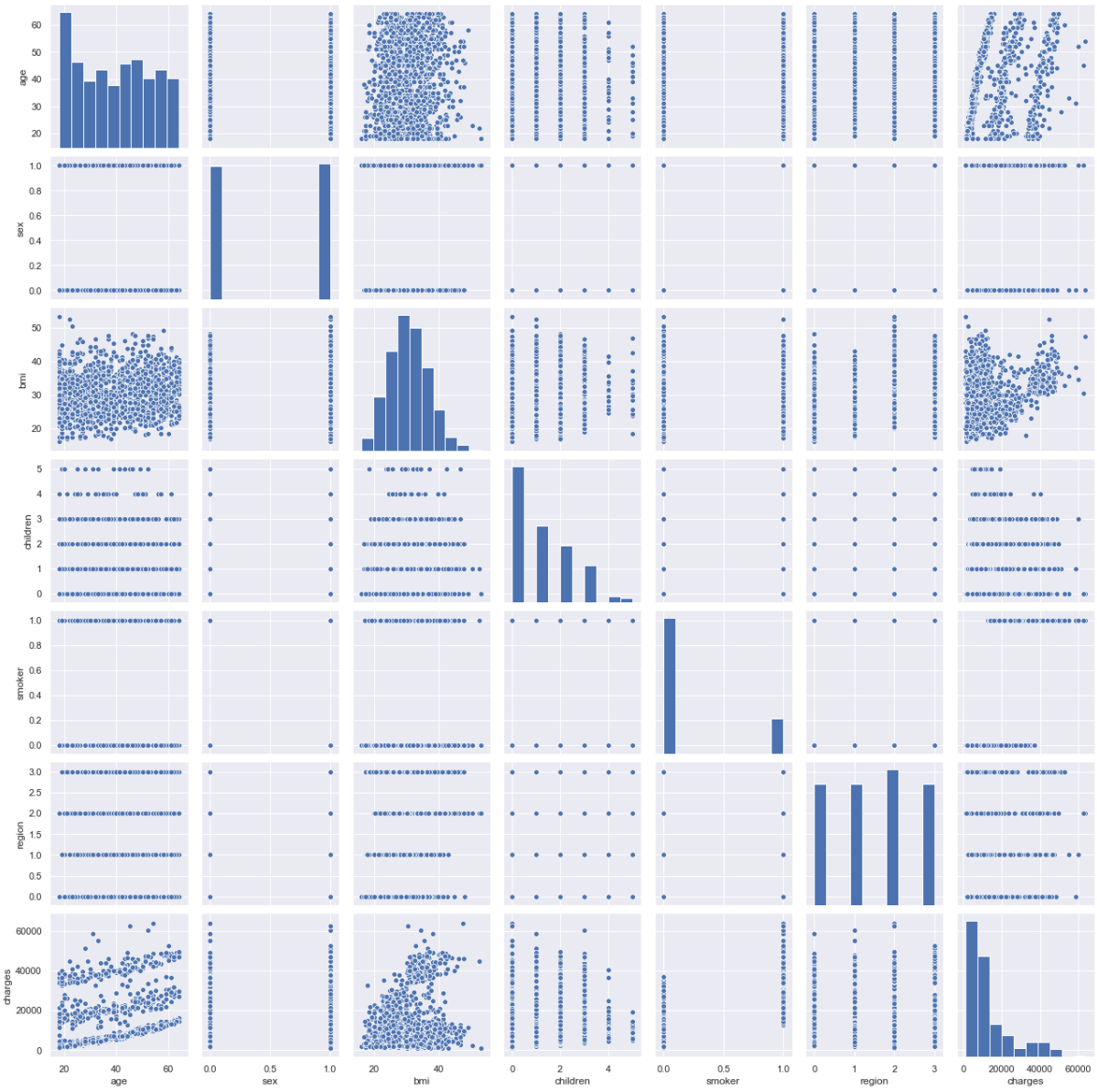
->There are lot more non-smokers than smokers.

->Instances are distributed evenly accross all regions.

->Gender is also distributed evenly.

->Most instances have less than 3 children and very few have 4 or 5 children.

6)Pair Plots to observe correlations



->There is an obvious correlation between 'charges' and 'smoker'

->Looks like smokers claimed more money than non-smokers

->There's an interesting pattern between 'age' and 'charges'. Noticing that older people are charged more than the younger ones

7)T- Test to check dependency of smoking on charges

Result:
Charges of smokers and non-smokers are not the same as p value (0)<0.05
Thus, Smokers seem to claim significantly more money than non-smokers

8)T-test to check dependency of bmi on gender

Result:

Gender has no effect on bmi as p value (0.812) > 0.05
Thus both the genders have more or less the same bmi

9)Chi Square Test to check if smoking habits are different for people of different regions
Result:

Gender has an effect on smoking as p value (0.007) <0.05
Proportion of smokers in males is significantly different from that of the females

10)Test to see Effect of children on bmi of females
Result:
No. of children had no effect on bmi as p value (0.716) >0.05