Lab -6

PRML
AY 2020-21 Trimester - III
Linear Regression

Name: Akshaykumar Kanani(B19EE008)
Question:

1. Load the dataset and do exploratory data analysis. [3 Marks]
2. Plot correlation between different variables and analyze whether there is a correlation between any pairs of variables or not. [3 marks]
3. Plot the distribution of the dependent variable and check for skewness (right or left skewed) in the distribution. [3 marks]
4. Convert this distribution into normal by applying natural log and plot it. (If the distribution is normal then skip this). [3 marks]
5. Convert categorical data into numbers. (You may choose one hot encoding or label encoding for that). [3 marks]
6. Split the data into training and testing sets with ratio 0.3. [2 marks]
7. Build a model using linear regression equation $\theta=(X^TX)^{-1}X^Ty$. (First add a feature X_0 =1 to the original dataset). [5 marks]
8. Build a linear regression model using the sklearn library. (No need to add X_0 =1, sklearn will take care of it.) [3 marks]
9. Get the parameters of the models you built in step 7 and 8, compare them, and print comparisons in a tabular form. If the parameters do not match, analyze the reason(s) for this (they should match in the ideal case). [5 marks]
10. Get predictions from both the models (step 7 and step 8). [5 marks]
11. Perform evaluation using the MSE of both models (step 7 and step 8). (Write down the MSE equation for the model in step 7 and use the inbuilt MSE for the model in step 8). [5 marks]
12. Plot the actual and the predicted values to check the relationship between the dependent and independent variables. (for both the models)

Ans :

1. We load the data and the basic information is given as:

From describe method:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

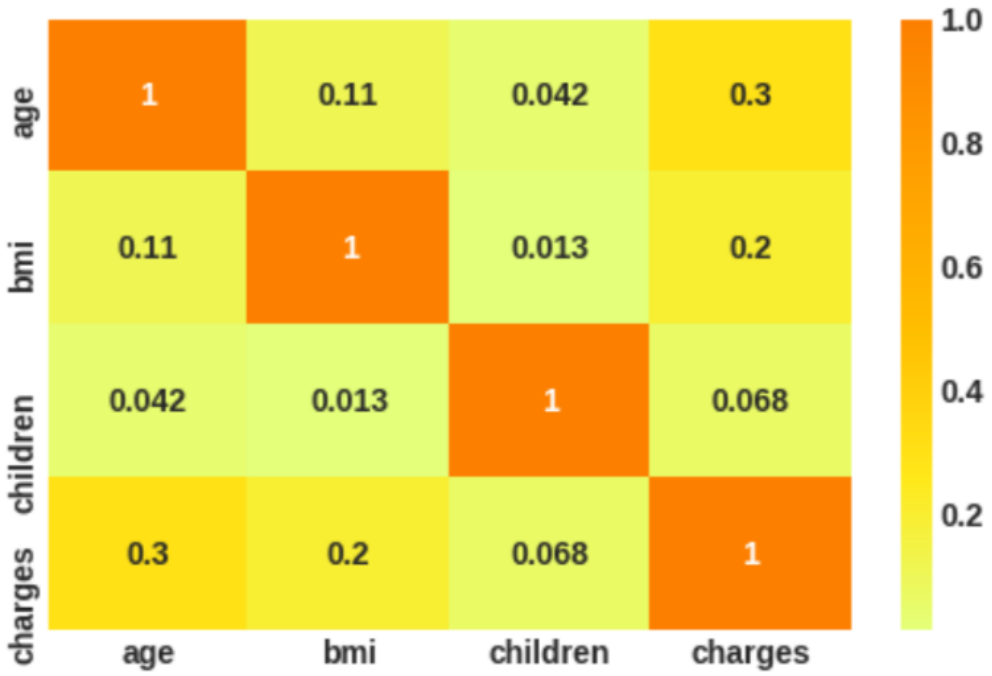
From info method

```
<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
```

2. correlation:

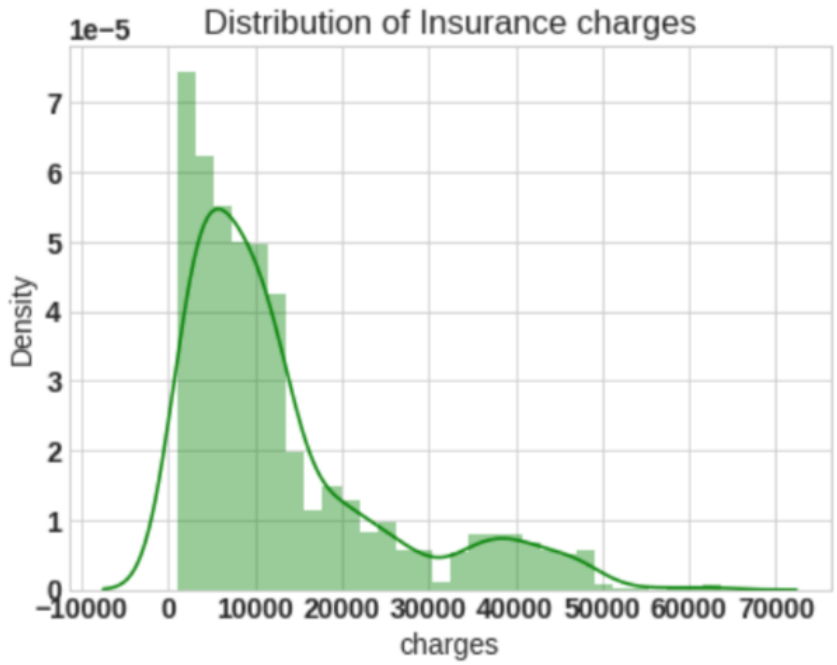
	age	bmi	children	charges
age	1.000000	0.109272	0.042469	0.299008
bmi	0.109272	1.000000	0.012759	0.198341
children	0.042469	0.012759	1.000000	0.067998
charges	0.299008	0.198341	0.067998	1.000000

And graphically:



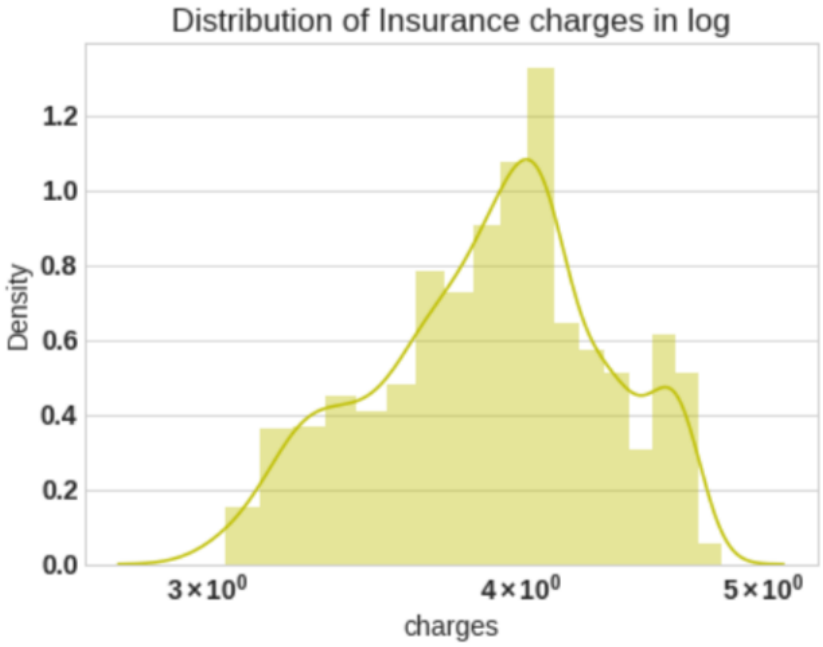
From above we can clearly say that there is NO correlation between and features and the max correlation we get is 0.3 from charges and age

3. Skewness:



From the distribution graph, we can clearly see that the graph is RIGHT SKEW

4. so to nullify this skewness we have to normalize our data. After normalisation we get:



5.

We did label encoding to change the categorical data into integer data in the code file.

6.

And we slit the data using the sklearn library into 70 and 30 %

7.

We build the model using linear regression equation $\theta=(X^TX)^{-1}X^Ty$.
And we got our theta as:

	All_thetas	feachers	theta
0	theta-0	intersect:x_0=1	7.057936
1	theta-1	age	0.034390
2	theta-2	sex	-0.087236
3	theta-3	bmi	0.011996
4	theta-4	children	0.108142
5	theta-5	smoker	1.551346
6	theta-6	region	-0.047787

8.

We use sklearn to build our model and we get:

	All_thetas	feachers	theta	sk_theta
0	theta-0	intersect:x_0=1	7.057936	7.057936
1	theta-1	age	0.034390	0.034390
2	theta-2	sex	-0.087236	-0.087236
3	theta-3	bmi	0.011996	0.011996
4	theta-4	children	0.108142	0.108142
5	theta-5	smoker	1.551346	1.551346
6	theta-6	region	-0.047787	-0.047787

9. From the above table we can clearly see the two columns or THETA and SK_THETA And by comparing we can see that both are **exactly the same.**

10.

We got the predicted values for both the model in code file.

11.

For the first model:

The Mean Square Error(MSE) or J(theta) is: 0.19924245957248024

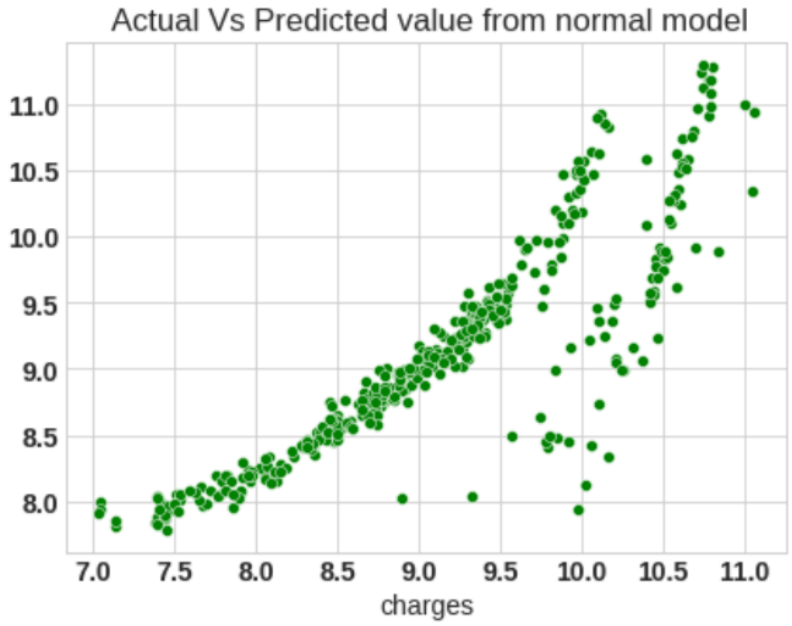
Fro the second model:

The Mean Square Error(MSE) or J(theta) is: 0.1992424595724774

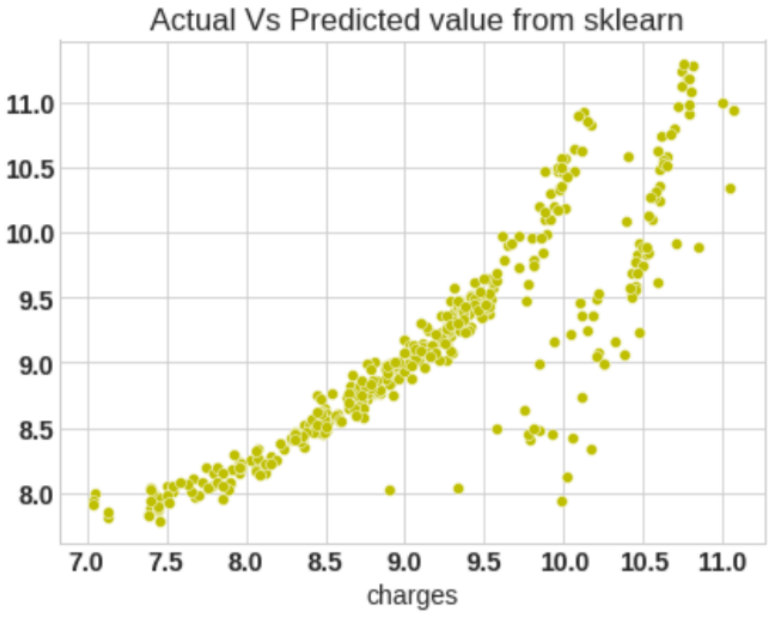
We can clearly see that both mean Square errors are equal and that's what we expect.

12.

From first model we get:



And from the second model(sklearn):



We can clearly see that both the plots are exactly the same so both the models are the same.