Name = GAUTAM Kumar

roll =B19EE031

Lab 6 report

Head of dataset

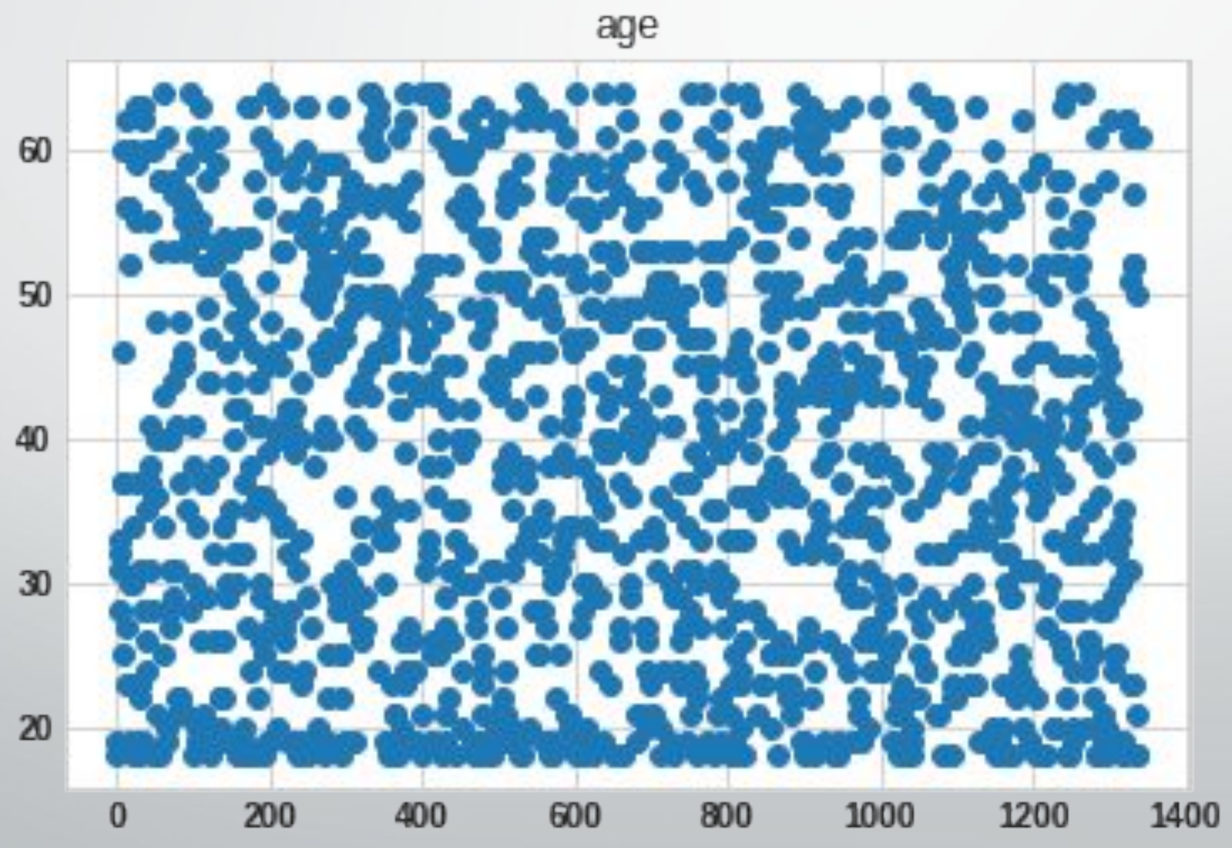
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Description of dataset

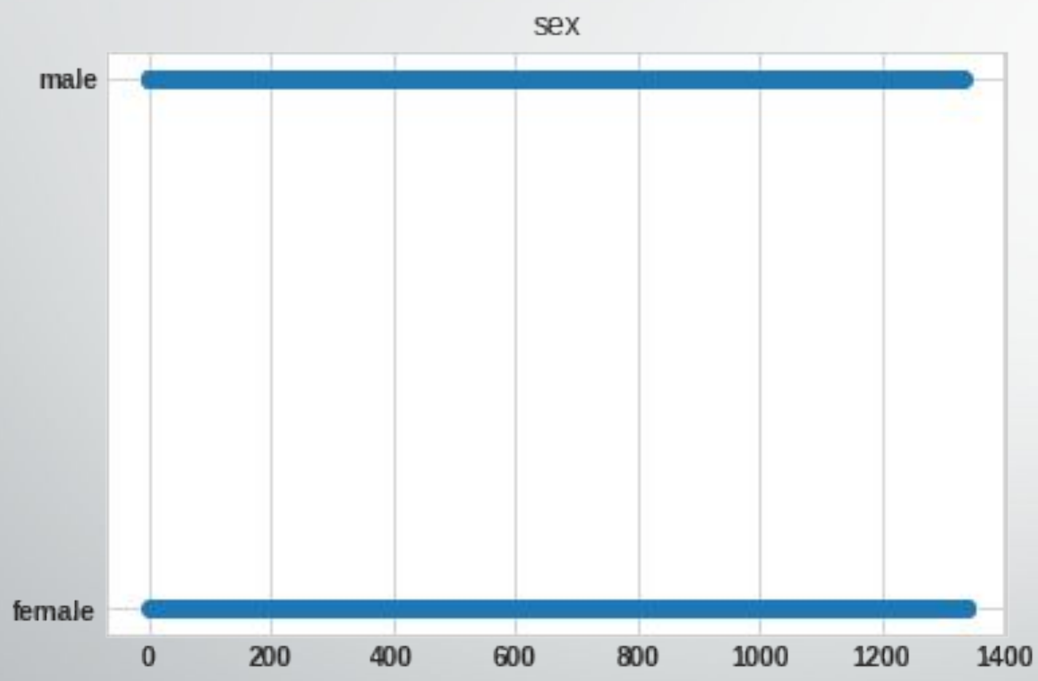
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000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

DISTRIBUTION OF INPUT FEATURES

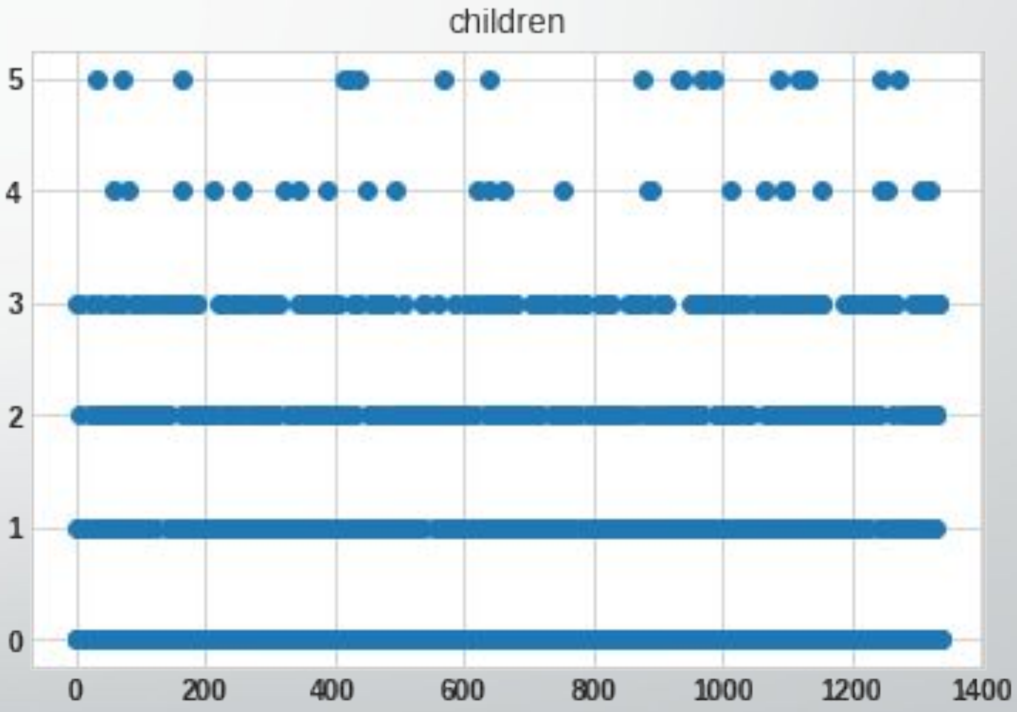
Age Distribution



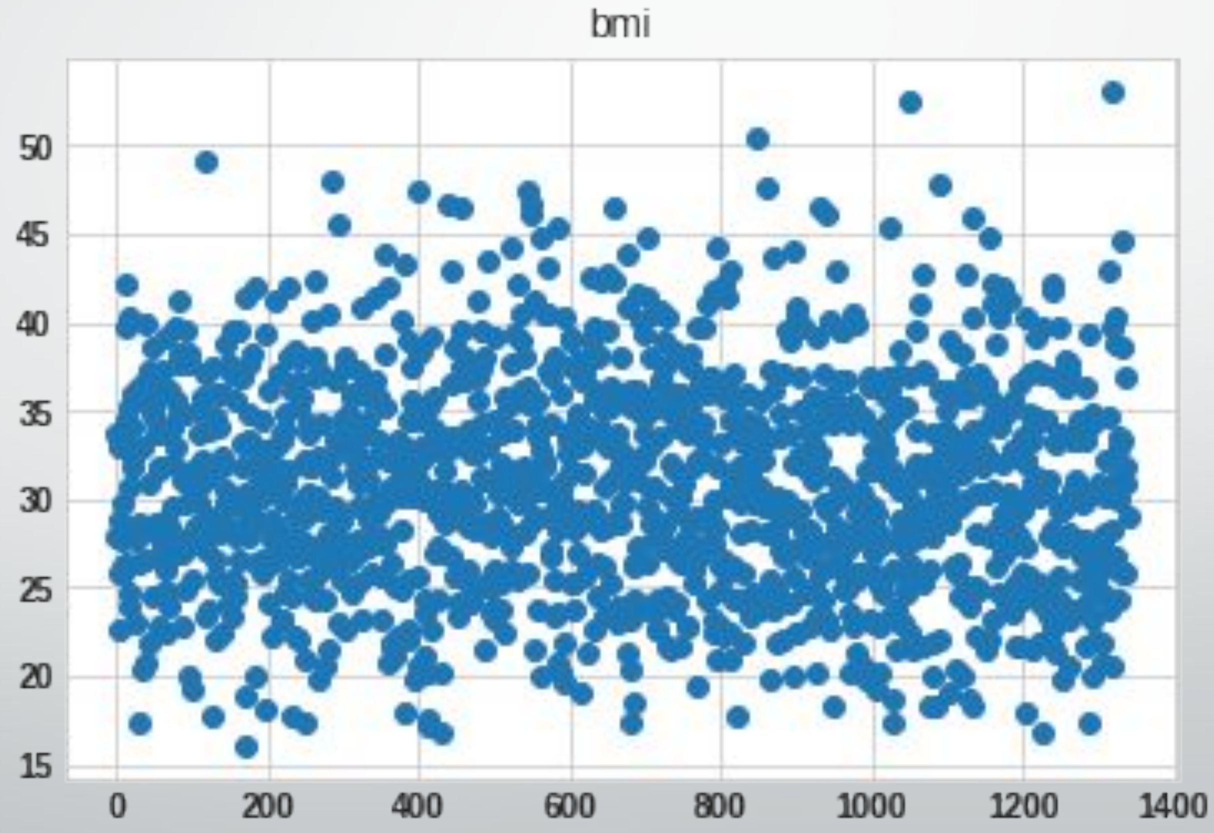
Gender



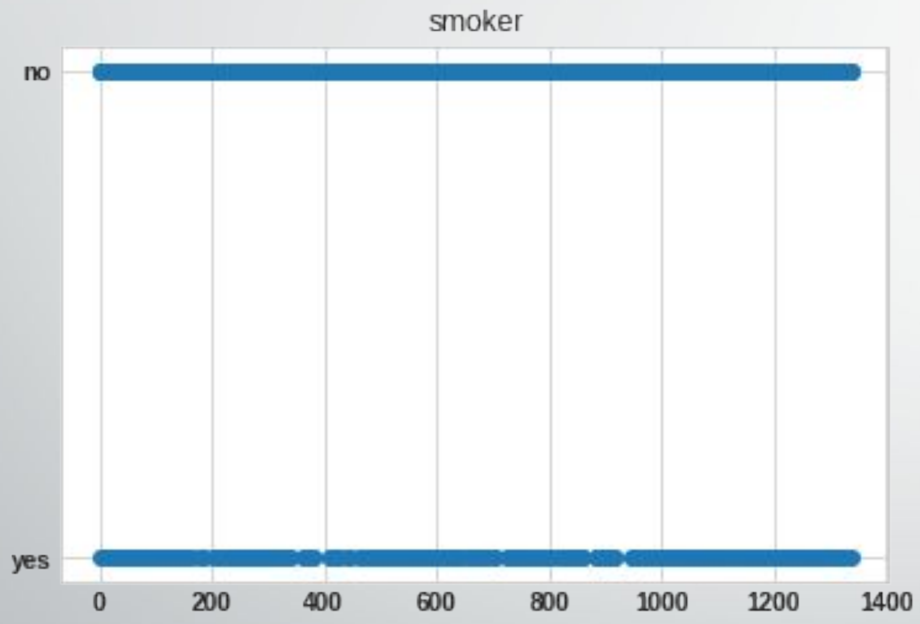
Children



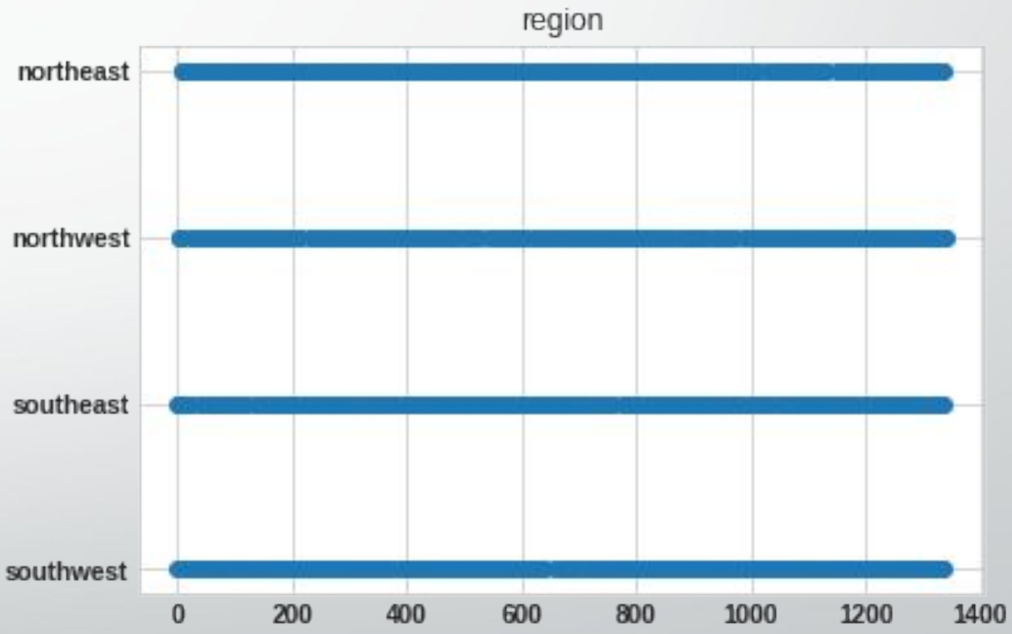
BMI



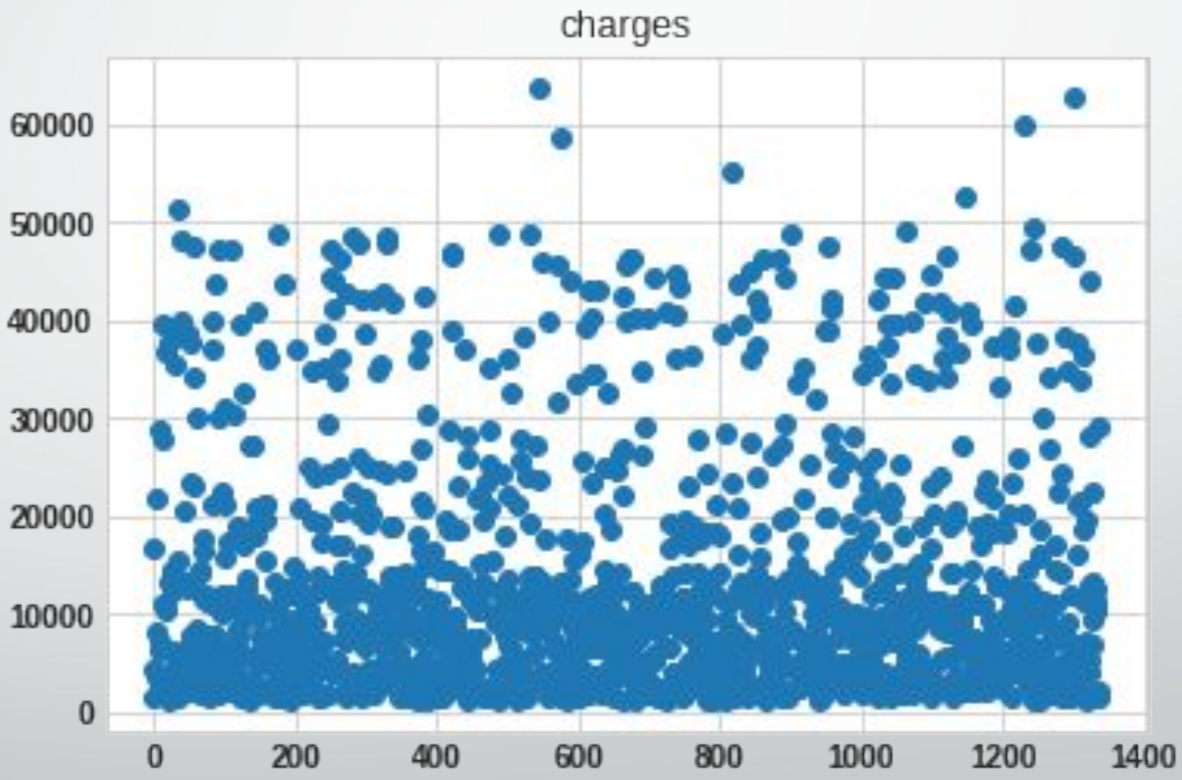
Smoker



Region



Target Variable : Charges



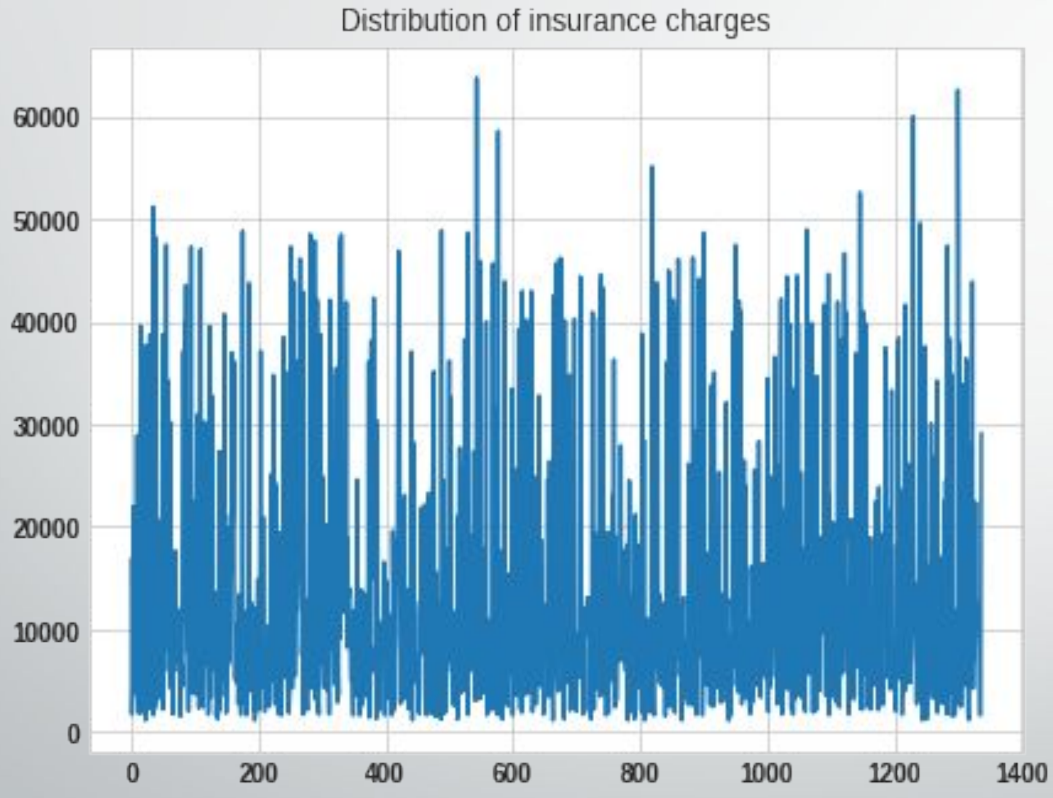
Correlation Plot

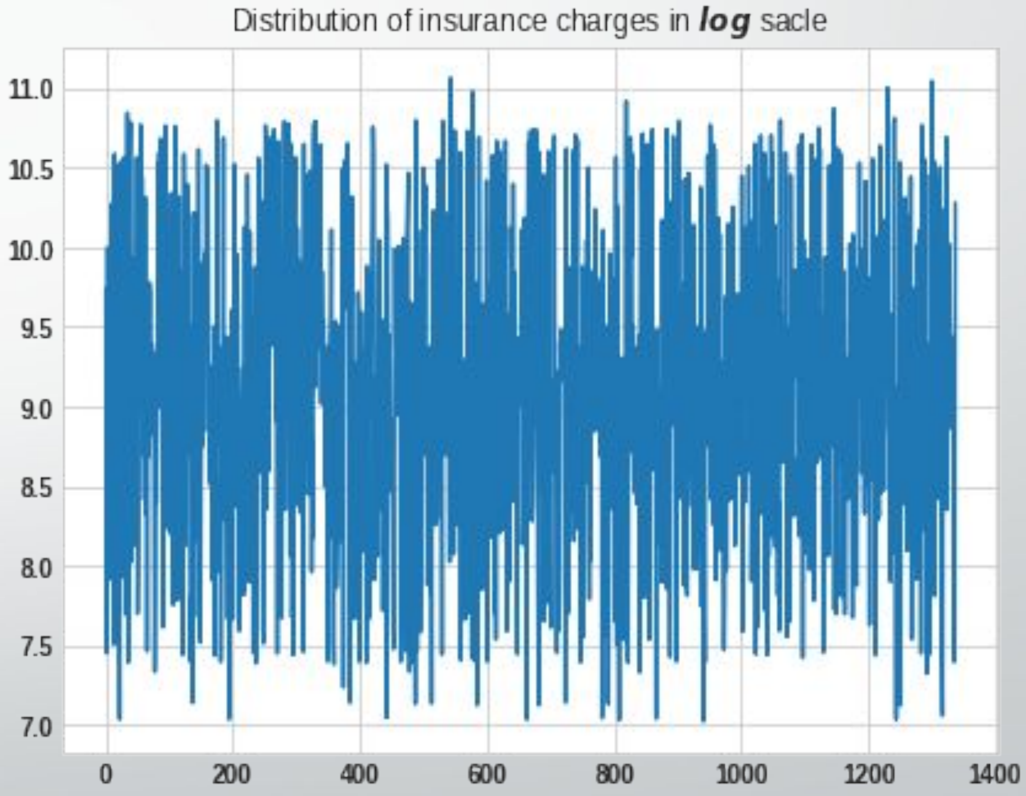
	age	bmi	children	charges
age	1.000000	0.109272	0.042469	0.299008
bmi	0.109272	1.000000	0.012759	0.198341
children	0.042469	0.012759	1.000000	0.067998
charges	0.299008	0.198341	0.067998	1.000000

```
corr = df.corr()
corr.style.background_gradient(cmap='coolwarm')
```

	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

Distribution of dependent data





Label Encoding and log of target variable

	age	sex	bmi	children	smoker	region	charges	log_transform
0	19	0	27.900	0	1	3	16884.92400	9.734176
1	18	1	33.770	1	0	2	1725.55230	7.453302
2	28	1	33.000	3	0	2	4449.46200	8.400538
3	33	1	22.705	0	0	1	21984.47061	9.998092
4	32	1	28.880	0	0	1	3866.85520	8.260197

Addition of xo = 1 in input features

	age	sex	bmi	children	smoker	region	xo
436	22	1	31.730	0	0	0	1
886	57	1	28.975	0	1	0	1
514	39	1	28.300	1	1	3	1
928	62	0	39.160	0	0	2	1
417	36	0	22.600	2	1	3	1

Parameters comparison

- *Using* $\Theta = (X^T X)^{-1} X^T Y$, *PARAMETERS* = [0.03386144 -0.06066302 0.01140949 0.10016429 1.54669051 -0.03834556 7.0917214]

- *Using sklearn linear model:* **WEIGHTS** : [0.03386144 -0.06066302 0.01140949 0.10016429 1.54669051 -0.03834556]
  INTERCEPT : 7.091721404756623

ANALYSIS

- *Hence we can see that we got the same set of weights with very minor differences.*
- *By adding $x_0 = 1$, we got the last value of weights equal to intercept.*

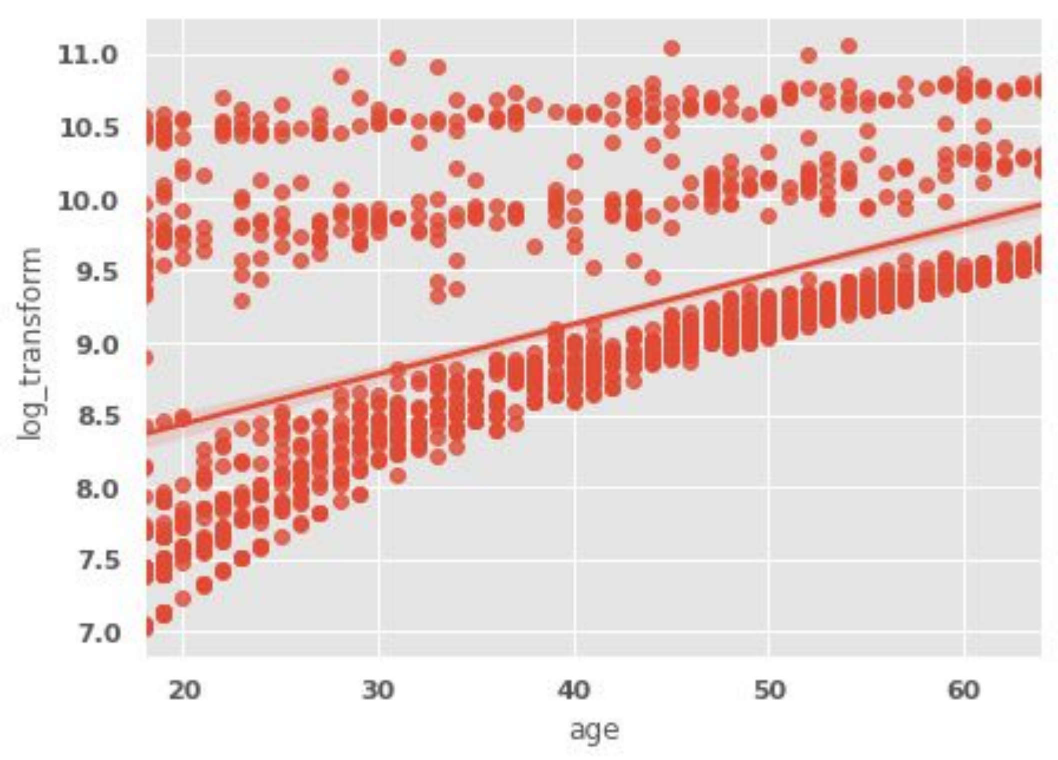
- *The values of y will be calculated by linear combination of x and weights.*
- $Y = w_0x_0 + w_1x_1 + w_2x_2 + \ldots W_nX_n + \text{intercept}$

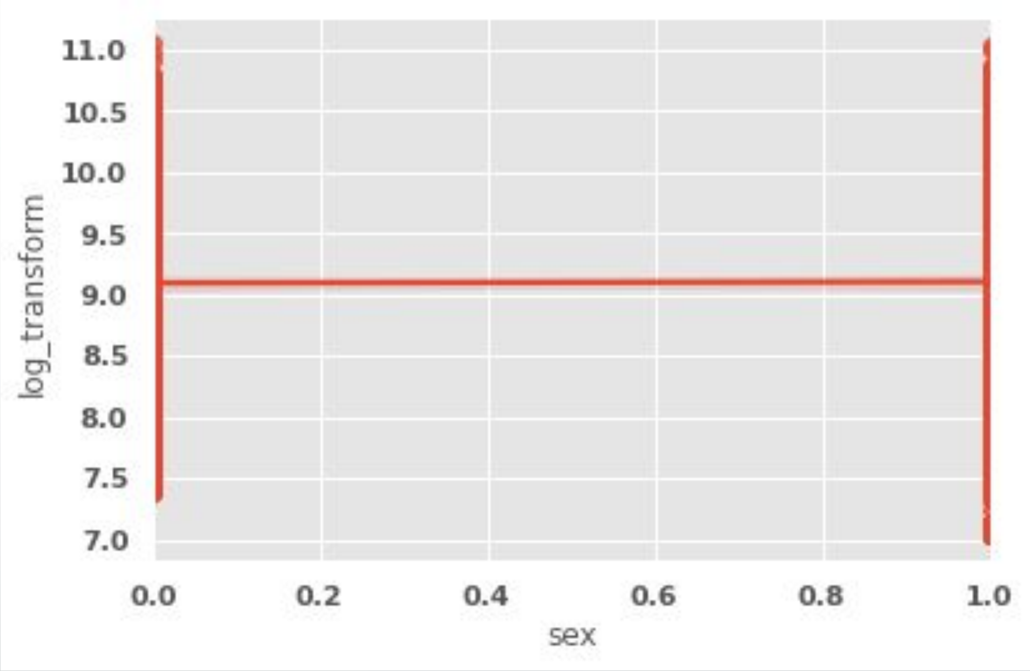
- *For our built model, we have already accounted for intercept by adding the feature of x_0.*

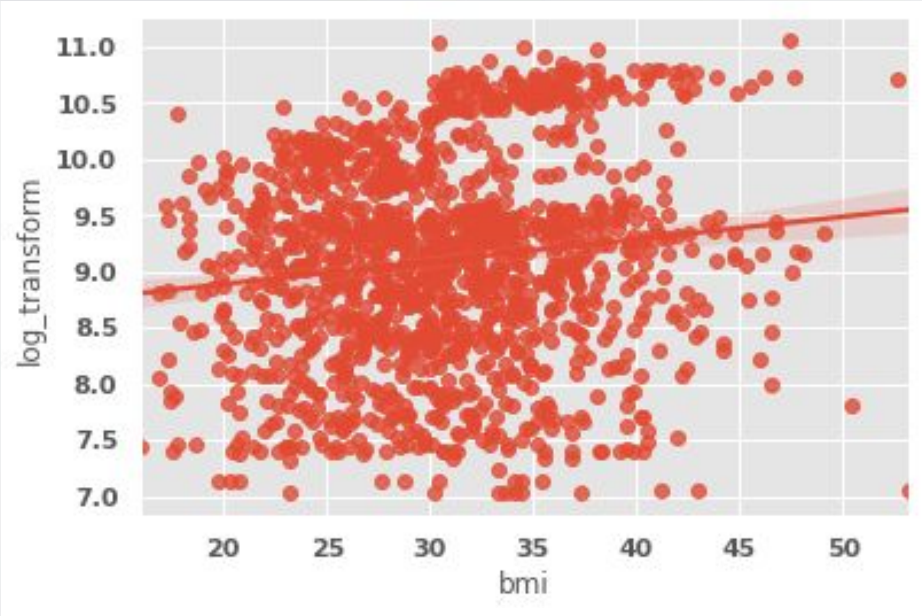
ANALYSIS

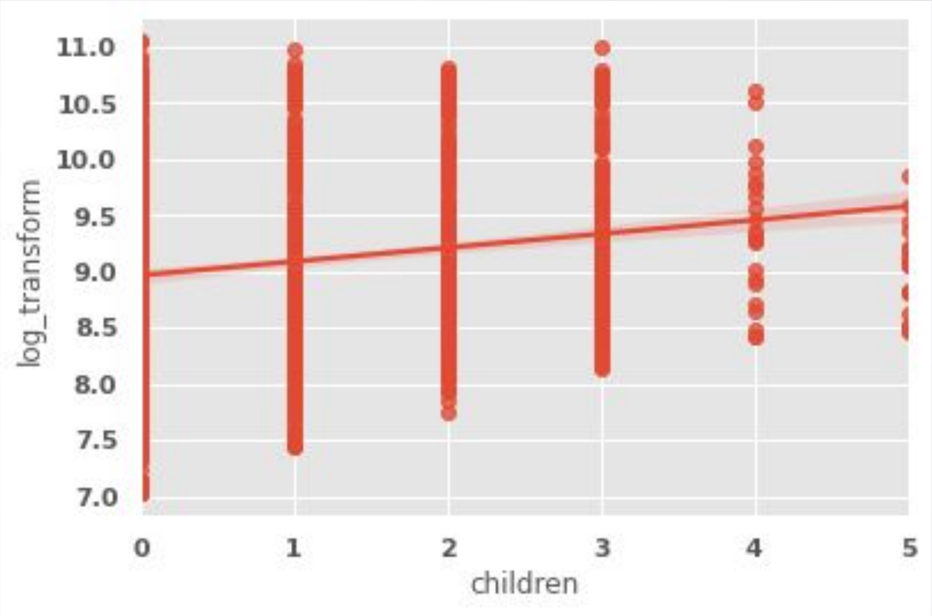
- *Mean Squared Error through self build function:* 0.1987110556807042
- *Mean Squared Error through sklearn library:* 0.19871105568070455
- *Hence both the values are nearly same.*
- *Both models are giving accurate results as their errors are low.*

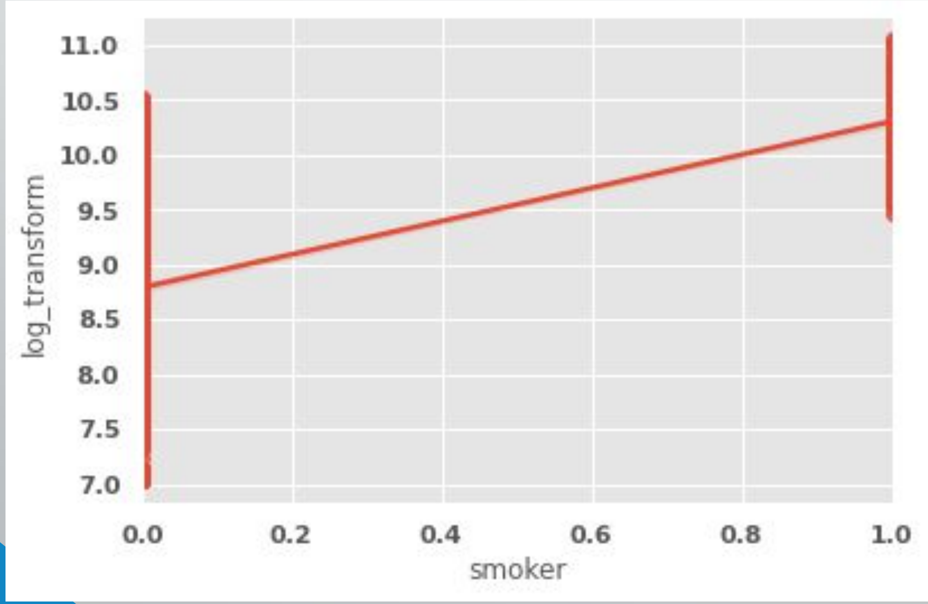
Checking Linearity of features with target

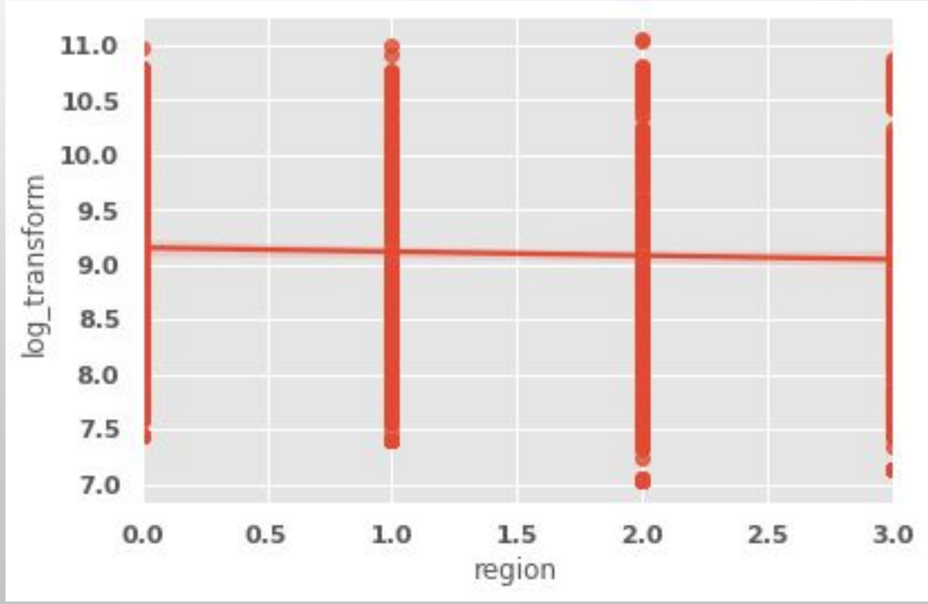





11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
log_transform
20
25
30
35
40
45
50
bmi


11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
log_transform
0
1
2
3
4
5
children


11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
log_transform
0.0
0.2
0.4
0.6
0.8
1.0
smoker


11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
log_transform
0.0
0.5
1.0
1.5
2.0
2.5
3.0
region

Conclusion

- *Hence we learnt about linear regressor model.*
- *We learnt the method to calculate weights and intercept of the model.*
- *We found the correlation of output with different input features.*
- *We learnt to write a function that calculates mean square error.*
- *We compared error and accuracy of both the models.*