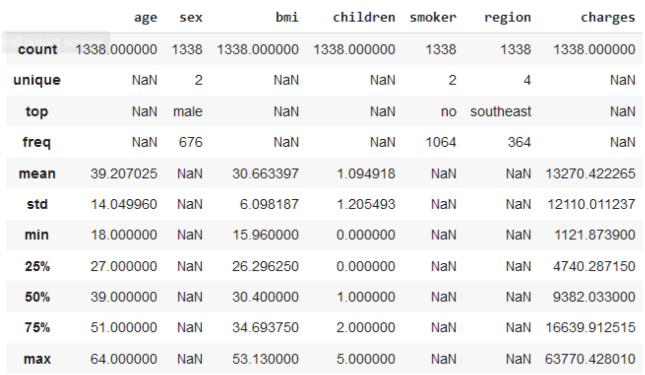
Main objective of the analysis is interpretation, where the impact of the following parameters are being assessed on the charges.

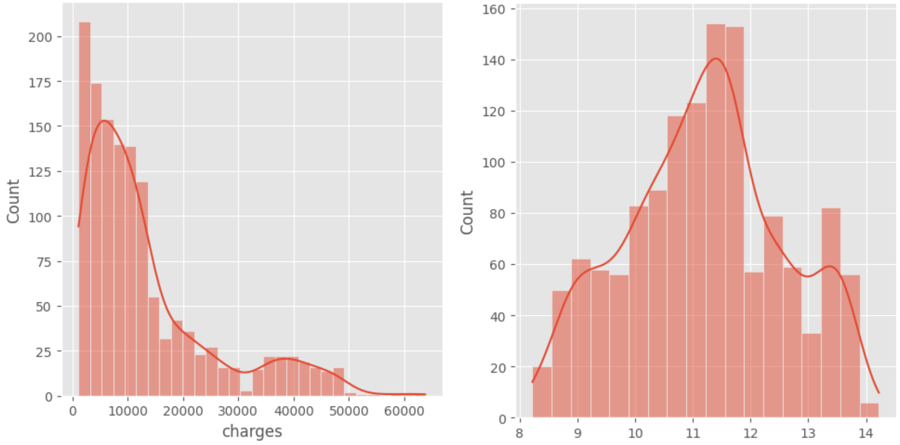
1. age 1338 non-null int64: the members' age from 18 to 64.
2. sex 1338 non-null object
3. bmi 1338 non-null float64: Body Mass Index is listed as a feature.
4. children 1338 non-null int64
5. smoker 1338 non-null object
6. region 1338 non-null object
7. charges 1338 non-null float64

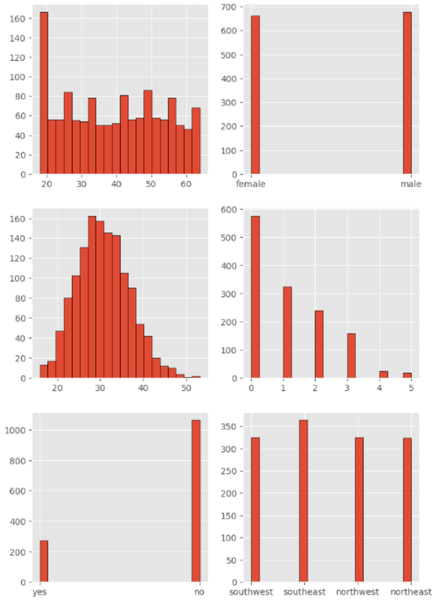
the analyse tests the checks the data statistics.



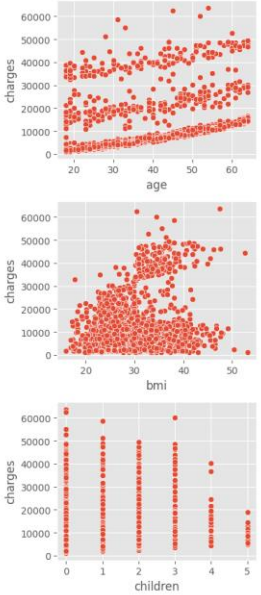
	age	sex	bmi	children	smoker	region	charges
count	1338.000000	1338	1338.000000	1338.000000	1338	1338	1338.000000
unique	NaN	2	NaN	NaN	2	4	NaN
top	NaN	male	NaN	NaN	no	southeast	NaN
freq	NaN	676	NaN	NaN	1064	364	NaN
mean	39.207025	NaN	30.663397	1.094918	NaN	NaN	13270.422265
std	14.049960	NaN	6.098187	1.205493	NaN	NaN	12110.011237
min	18.000000	NaN	15.960000	0.000000	NaN	NaN	1121.873900
25%	27.000000	NaN	26.296250	0.000000	NaN	NaN	4740.287150
50%	39.000000	NaN	30.400000	1.000000	NaN	NaN	9382.033000
75%	51.000000	NaN	34.693750	2.000000	NaN	NaN	16639.912515
max	64.000000	NaN	53.130000	5.000000	NaN	NaN	63770.428010

The histogram of the data is plotted as below. The features' distribution is accepted for linear regression modeling. However, the target variable is not normally distributed. A boxcox transformation is used to make it more normal.



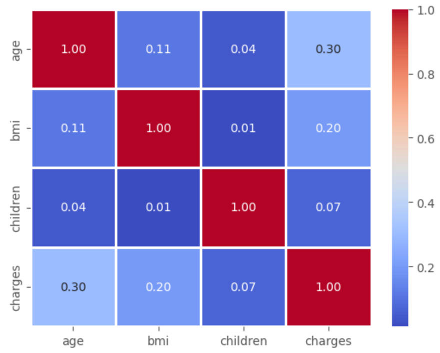


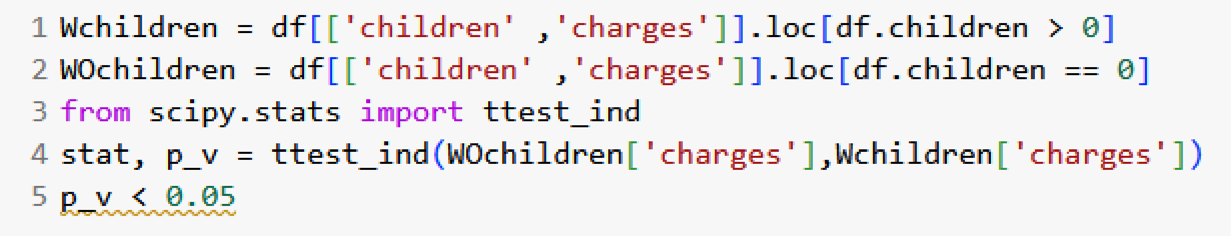
The scatter plot shows how target varies with numerical features, which can be assumed linear.



In order to check Homoscedaticity of data, the residplots are plotted which show Homoscedaticity assumption to be correct.

The correlation plot shows a very small correlation between charges and the number of children. I tested statistically if having children is significant in charges or not. The p_value is less than 0.05, it suggests that the difference in charges between people with children and people without children is statistically significant.

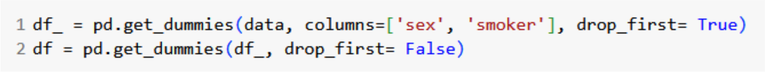




```
Wchildren = df[['children' ,'charges']].loc[df.children > 0]
WOchildren = df[['children' ,'charges']].loc[df.children == 0]
from scipy.stats import ttest_ind
stat, p_v = ttest_ind(WOchildren['charges'],Wchildren['charges'])
p_v < 0.05
```

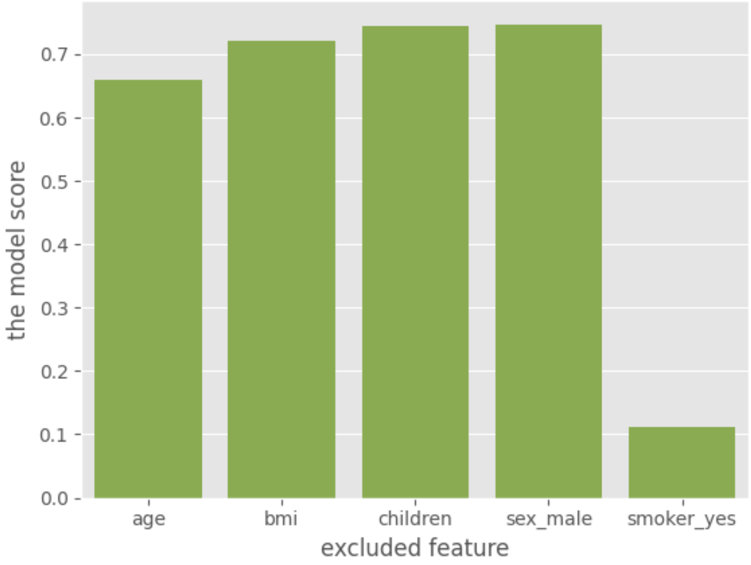
```
True
```

get_dummies action is taken for feature engineering and transforming object data to numeric data.



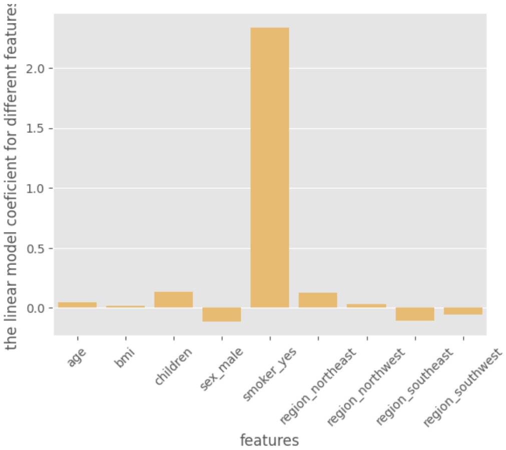
```
df_ = pd.get_dummies(data, columns=['sex', 'smoker'], drop_first= True)
df = pd.get_dummies(df_, drop_first= False)
```

I leverage the cross validation capabilities to exclude each column and evaluate the model performance without that feature. As we see removing smoker feature drastically reduces the score and this shows how important is this feature.

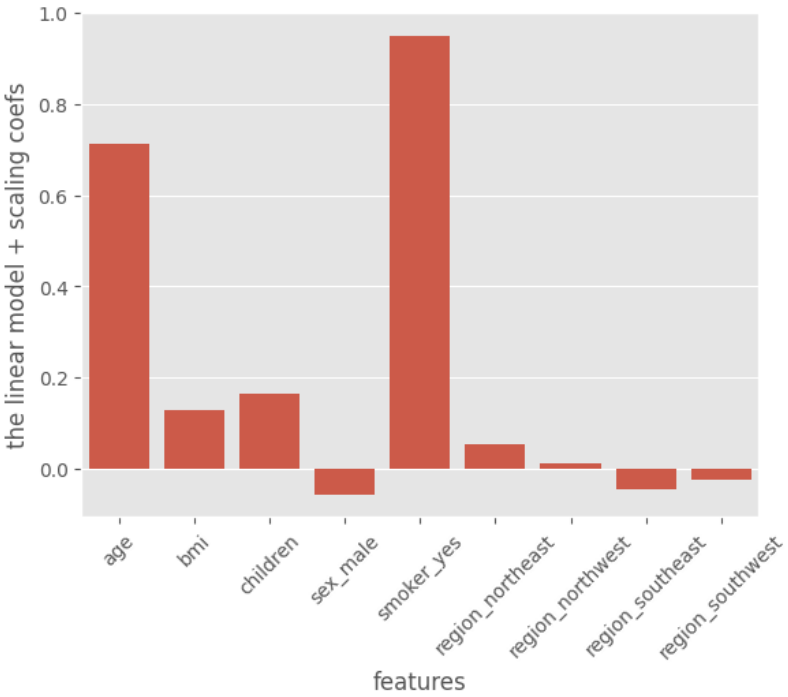


Regression analysis:

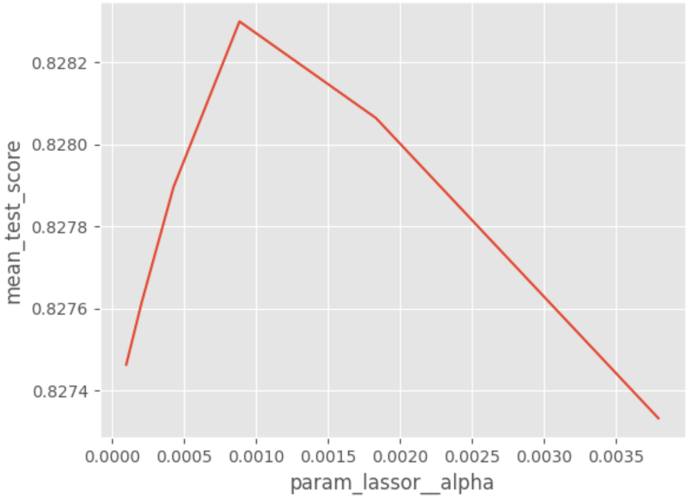
The simple LR analysis was performed where the data was not scaled and polynomial features were not applied. The result shows the large coefficient for smoker feature compared to rest of features. The features were not scaled (i.e., no standardization or normalization). In many cases, especially when features have different units or large disparities in scale, scaling (like StandardScaler or MinMaxScaler) is important. Without scaling, features with larger numerical ranges can dominate the model, leading to distorted coefficients. **Polynomial features** are not used, meaning the model only considers linear relationships between features and the target. Higher-order interactions or curves in the data are not captured.



When I apply scaling, the coefficient distribution is being more evenly distributed. For example “age” coefficient is being larger compared to amoker_yes column. When scaling is applied, the features are transformed to a similar scale, reducing the impact of differences in the original ranges. This allows the model to focus more on the relationships between the features and the target, rather than being influenced by their numerical range. When all features are scaled (e.g., via StandardScaler), each feature now has a similar range (mean of 0 and standard deviation of 1). Age is no longer dwarfed by the range of other features, and its coefficient reflects its true relationship with the target variable (e.g., insurance charges). The magnitude of the coefficient is now directly related to how strongly "age" correlates with the target variable compared to other features. If "age" has a strong relationship with the target (after scaling), its coefficient will be larger. However, the r2 score does not change after scaling! **Scaling** the features (e.g., using StandardScaler) **does not directly change the relationships** between the features and the target variable. It only alters the scale of the features themselves. Therefore, in terms of the overall fit of the model (how well it predicts the target variable), **scaling does not affect the R^2 score** because the underlying relationships between the variables stay the same. When I apply polynomical features, the r2 score increases from 0.79 to 0.84.



As a regulization technique, Lasso regression is applied on scaled polynomial features where alpha varies in np.geomspace(0.0001,100,20) range. Using GridSearchCV the best alpha is selected (0.00088) and r2_score is calculated which is 0.84. the best score for Lasso is equal to the OLS regression. When the regularization parameter α is very small (like selected value of 0.00088), the Lasso model will behave almost like OLS regression because the penalty for large coefficients is negligible. A small alpha in Lasso regression can lead to the best regularization fit in this case, particularly when the original model does not suffer from overfitting or when the data is already well-behaved (i.e., not overly complex or noisy). So for the data selected, the simple regression model with polynomial feature does not overfit and works well.



The smoker feature has a large positive coefficient, meaning that smokers have significantly higher insurance charges compared to non-smokers. This is a major driver of the target variable. The age feature has a moderate positive coefficient, indicating that older individuals tend to have higher insurance charges, which is a common finding in insurance datasets. The children feature shows a smaller effect, suggesting that the number of children may have a less significant influence on the charges when compared to other factors like smoking or age. The inclusion of polynomial features (e.g., age^2) helped capture more complex relationships between the features and the target variable, although the regularization prevented overfitting, as evidenced by the similar R^2 score to OLS. The smoker feature is clearly one of the most influential variables in predicting insurance charges. This suggests that health-related factors, such as smoking, play a critical role in determining insurance premiums.