Assignment 2

Formulate an interesting regression problem and solve it.

Nitesh Yadav(35), Rohit Rana(40), Rohit Shakya(41)

M.Sc. Computer Science
Department of Computer Science
University of Delhi

Overview

The problem is to find out the cost of treatment of different patients.

About Data

- Data set dedicated to the cost of treatment of different patients.
- The cost of treatment depends on many factors: diagnosis, type of clinic, city of residence, age and so on.
- We have no data on the diagnosis of patients.

Specifications

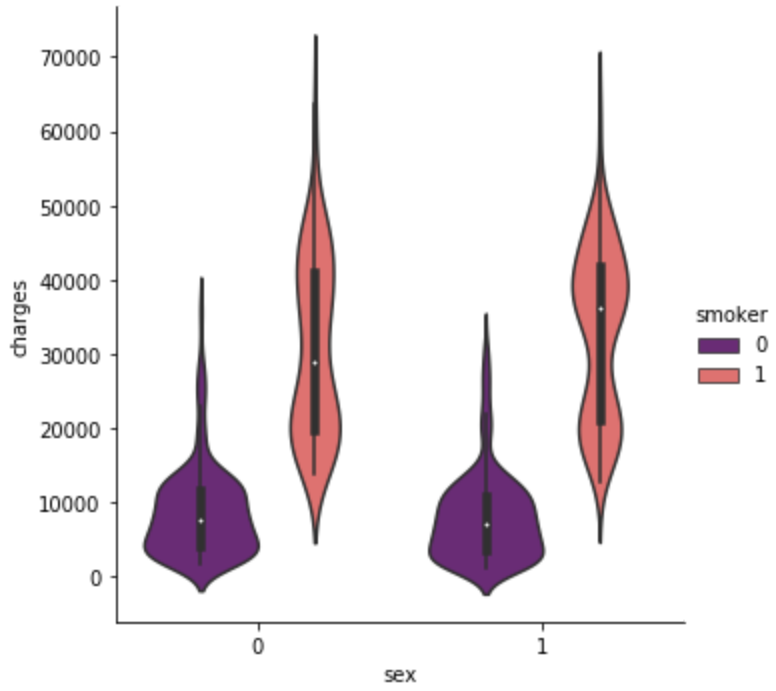
- **Number of Instances:** 1338**.**
- **Number of Attributes :**
    - **age:** age of the primary beneficiary
    - **sex:** insurance contractor gender, female, male
    - **bmi:** Body mass index, providing an understanding of the body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
    - **children:** Number of children covered by health insurance / Number of dependents
    - **smoker:** Smoking
    - **charges:** Individual medical costs billed by health insurance

- **Missing Attribute Values**: No missing value
- **Correlation of attributes with charges**

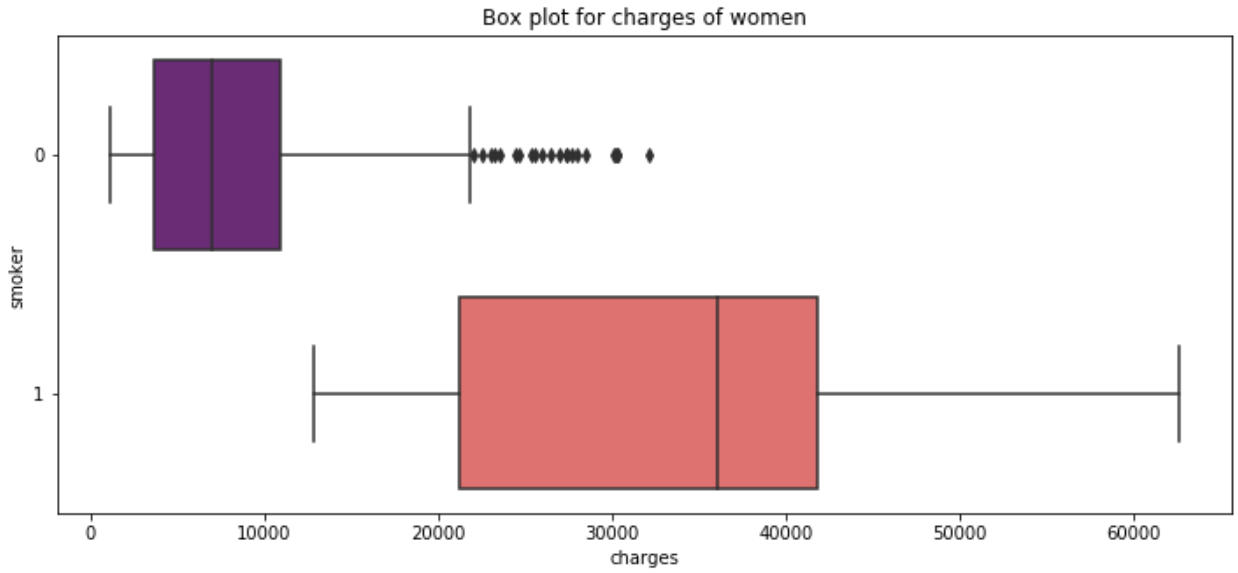
sex	0.057292
children	0.067998
bmi	0.198341
age	0.299008
smoker	0.787251
charges	1.000000

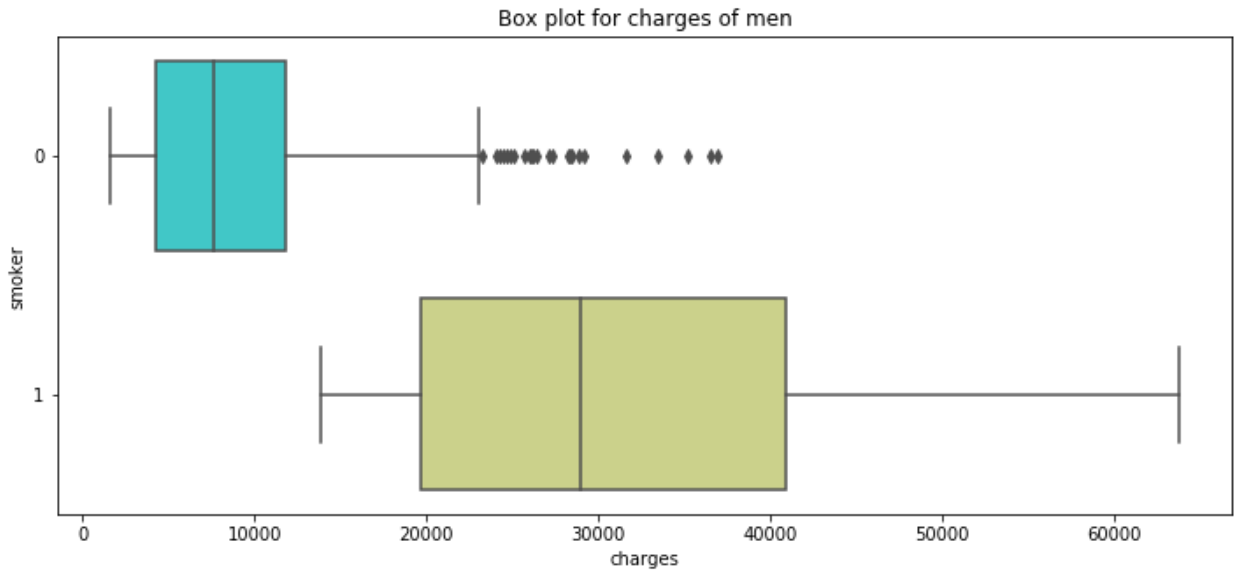
* We can see that charges have an almost linear positive correlation with smokers.

Visualizations

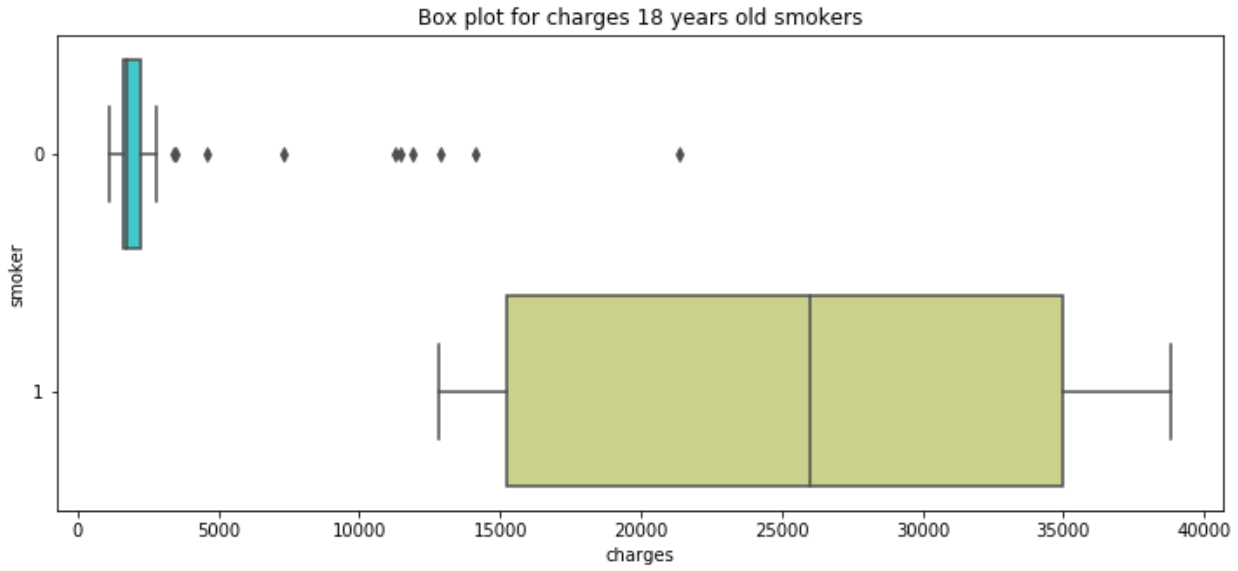


- Women smokers and non-smokers tend to have fewer charges as compared to men smokers and non-smokers respectively.

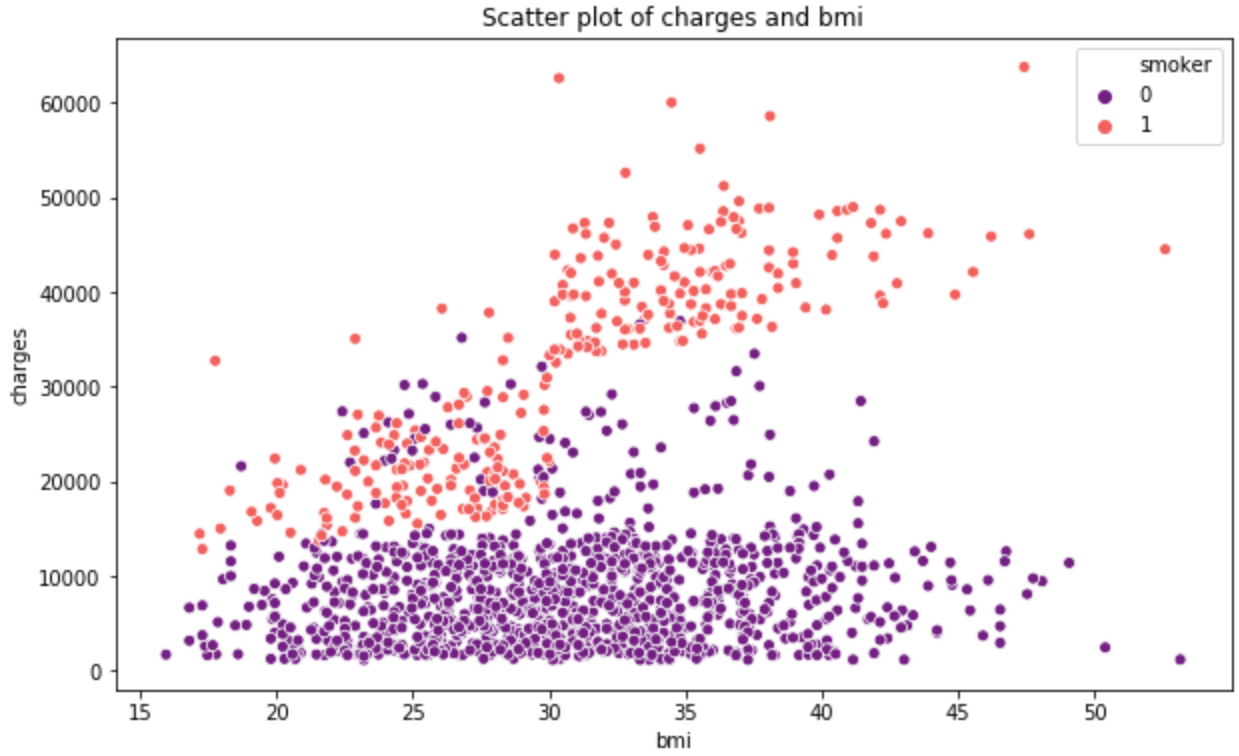




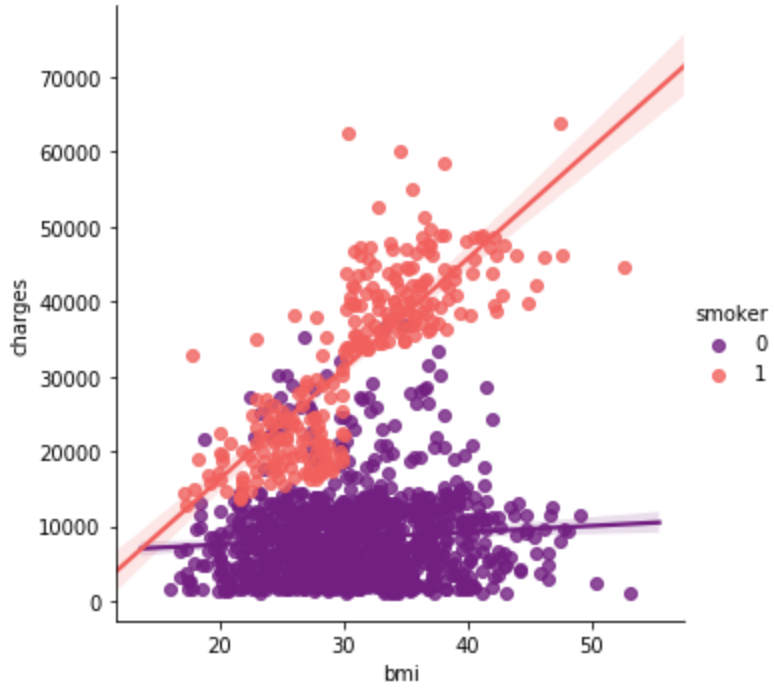
- Both men and women smokers and non-smokers tend to have similar charges but the average charges(mean) are more in case of women smokers.



- The 18-year-old smokers spend much more on treatment than non-smokers.



- Smokers with bmi>30 have very high charges.



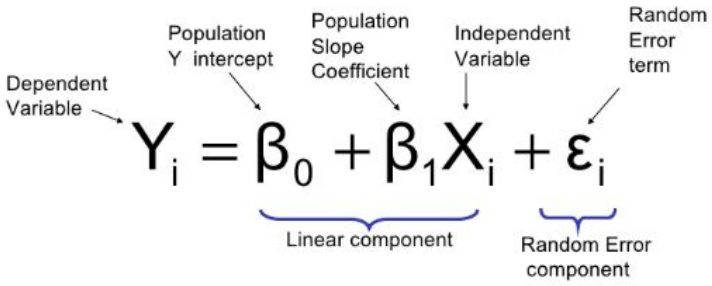
- For smokers, charges have a positive linear relation with bmi.

Methods

Linear Regression

Definition: Linear regression is a linear model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, y can be calculated from a linear combination of the input variables (x).

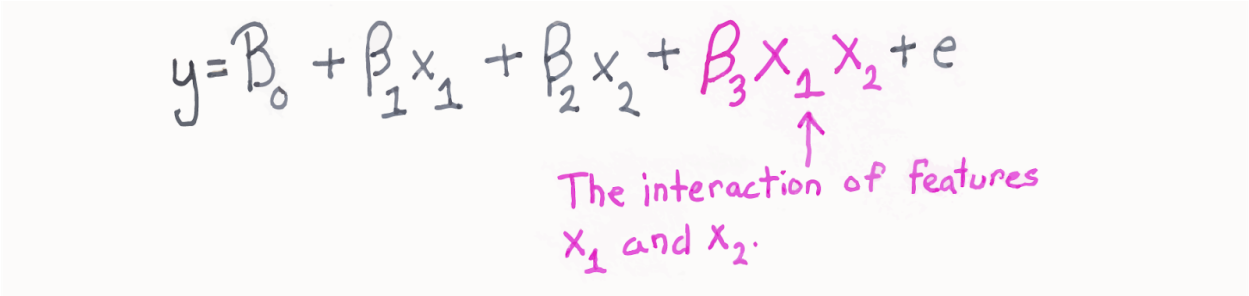
Formula



Linear Regression with polynomial interaction

Interaction terms allow us to model relationships when the effects of a feature on the target are influenced by another feature.

Formula



The model with interaction is more flexible (i.e., we've added a parameter).

The Random Forest Classifier

Definition: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size.

Metrics Used to compare models :

Root Mean Square Error (**RMSE**) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Findings

Accuracy wise ordering of all three methods from highest to lowest is:

1. Random Forest Classifier (RMSE - 2830.45)
2. Linear Regression with polynomial interaction (RMSE - 5439.77)
3. Linear Regression (RMSE - 6320.04)