* WOMart follows a multi-channel distribution strategy with 350+ retail stores spread across 100+ cities. Our task is to predict the store sales for each store in the test set for the next two months.
* We have given sales data for 18 months from 365 stores of WOMart is available along with information on Store Type, Location Type for each store, Region Code for every store, Discount provided by the store on every day, Number of Orders everyday etc.
* I started from importing relevant libraries, after it I had imported train and test dataset that has been provided and visualize its features.
* I had checked for missing values present in both datasets but fortunately, there no missing values present in our data.
* There is date column in our data, I dealt with it and convert the datatype of it to datetime from object. Later segregated it in day, month and year so model can recognize it.
* After that I segregate our columns to numerical and categorical column. Firstly, I had to handle categorical variable. Before that with the help of seaborn and matplotlib library, I had visualized data and see some insights into it.
* There were nominal categorical data present , so with the help of one hot encoding method , I had encoded categorical columns. After, that I saw there were outliers present in our dependent numerical column, I had replaced all entries that were greater than 200000 with median, so that prediction won’t get affected by those large values.
* Later, I separated dependent and independent data, I split the data in 80 / 20 ratio for training the model.
* I had used function so that I could test and try various regression models. After hit and trial method, I found that Random forest regression model has performed better than rest of regression models, so to increase the performance of regression.
* After that , I test my model on test set and create submission file.