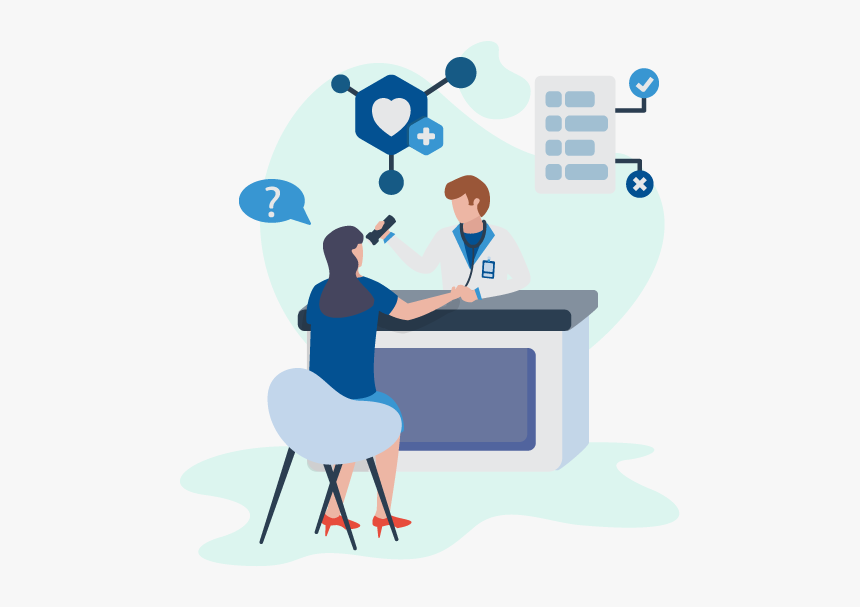
**Doctor Consultation Fees**



In this blog-post, I will go through the whole process of creating a machine learning model on the famous Doctor Consultation Fees. It provides information on the Qualification, Experience, Rating, Profile, Miscellaneous information and Place.

**Problem Statement:**

* We have all been in situation where we go to a doctor in emergency and find that the consultation fees are too high. As a data scientist we all should do better. What if you have data that records important details about a doctor and you get to build a model to predict the doctor’s consulting fee.? This is the use case that lets you do that.

**Steps to Proceed:**

* We’ll start by importing important Liberalises,
* exploratory data analysis,
* then data pre-processing and
* as it is regression problem, we’ll be testing different models such as support vector regressor and decision trees regressor and so on.

**Data Analysis:**

First and foremost, will started with importing the necessary libraries like NumPy, pandas, matplotlib and seaborn, then I moved on to reading the excel files (Train and Test).

The data consists of the following rows:

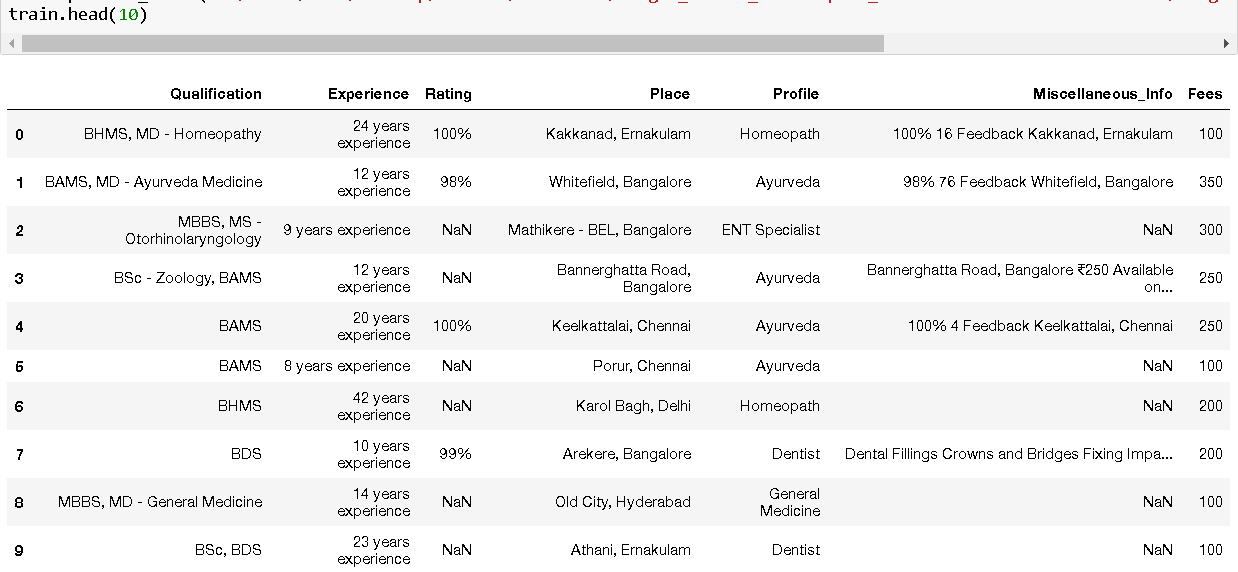
Independent Variable

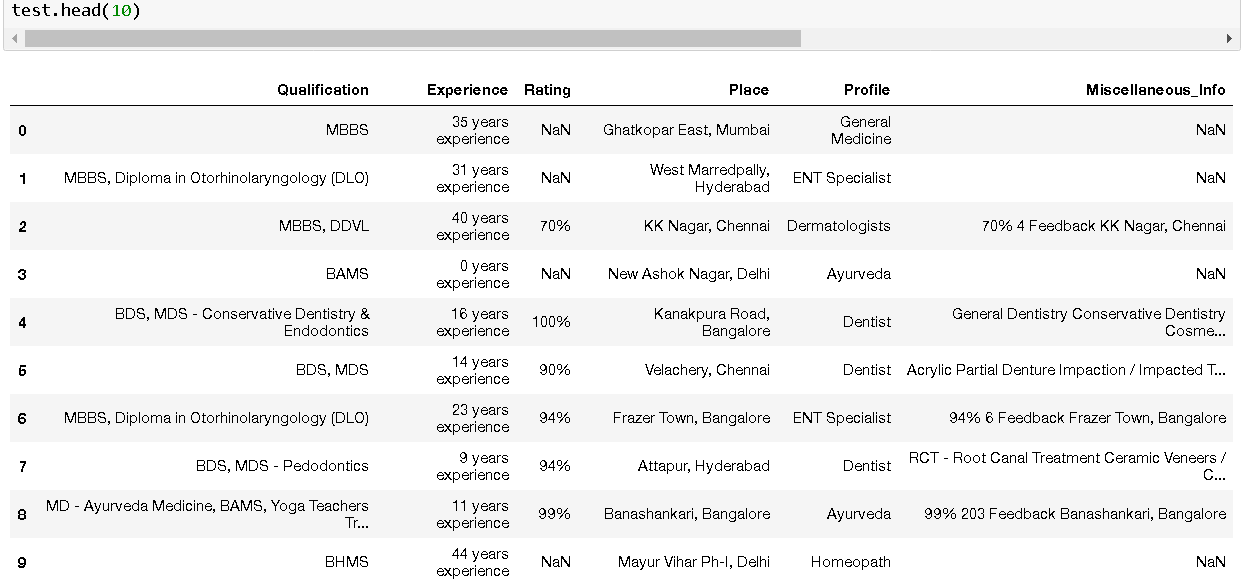
* Qualification: Qualification and degrees held by the doctor
* Experience: Experience of the doctor in number of years
* Rating: Rating given by patients
* Profile: Type of the doctor
* Miscellaneous\_Info: Extra information about the doctor
* Place: Area and the city where the doctor is located

Independent Variable

* Fees: Fees charged by the doctor

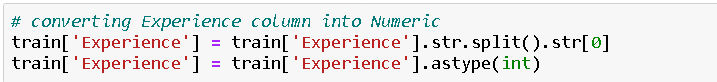
By checking the information of the first 10rows data and we can see that there were few missing values in both the train and test data



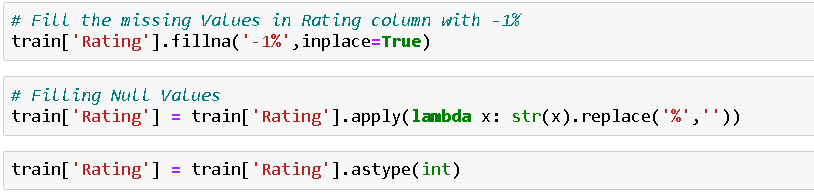


# **Exploratory Data Analysis (EDA):**

Will started by splitting the experience so I can have the exact number of years without the word ‘years experience’ .

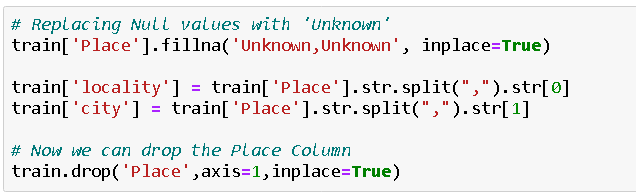


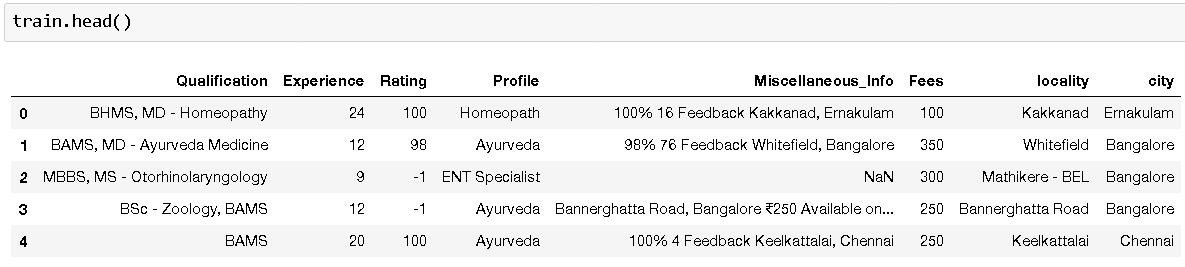
Then separate the sign ‘%’ from the column rating just to have integers alone in the column.



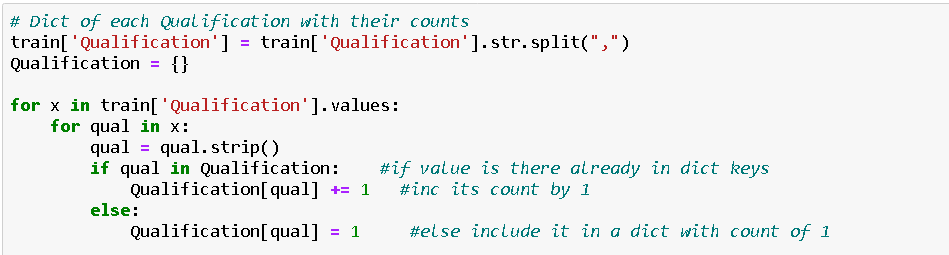
Before extraction, Will replaced all missing values in this column with the string ‘Unknown, Unknown’ to represent them. Side note, sometimes it is a good idea to give missing values a separate class instead of relying on missing values imputation technique like mean/median/mode.

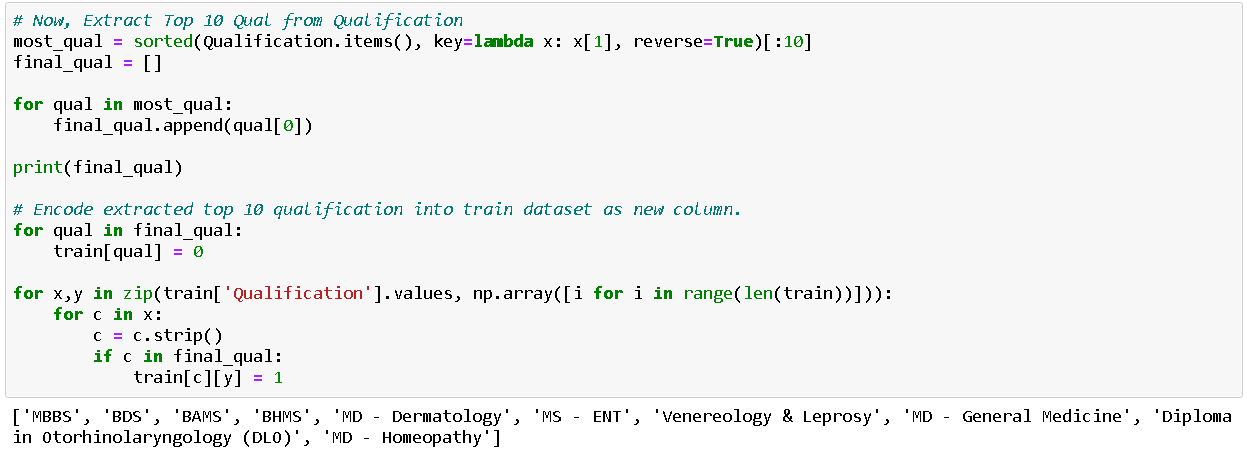
Next, splitting the string at ‘,’ and creating a new column ‘Locality’ & ‘City’ using the last element of the list.

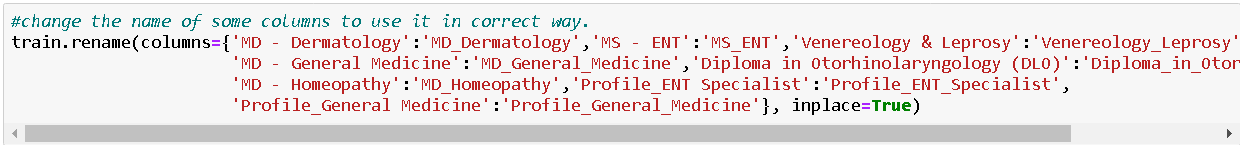




Now, for the ‘Qualification’ column, it consists of various qualification of the doctor without any standardized reporting method. Will start by doing the normal split and then will separate some words from the columns ‘Qualification’ so I can have the exact degree e.g. - 'MBBS', 'BDS', 'BAMS', 'BHMS', 'MD - Dermatology', 'MS - ENT', 'Venereology & Leprosy', 'MD - General Medicine', 'Diploma in Otorhinolaryngology (DLO)', 'MD - Homeopathy'





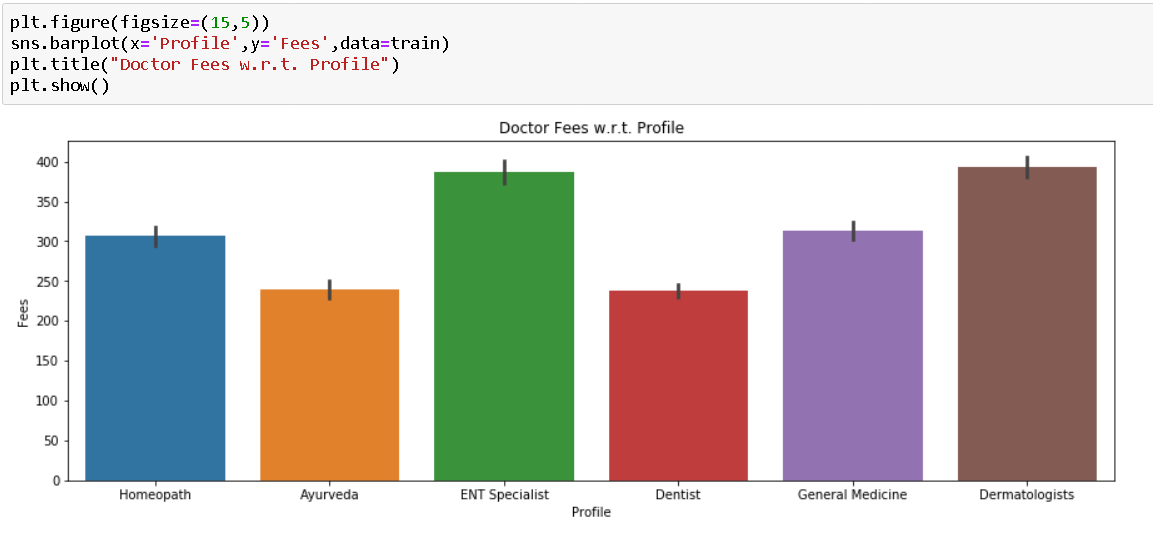




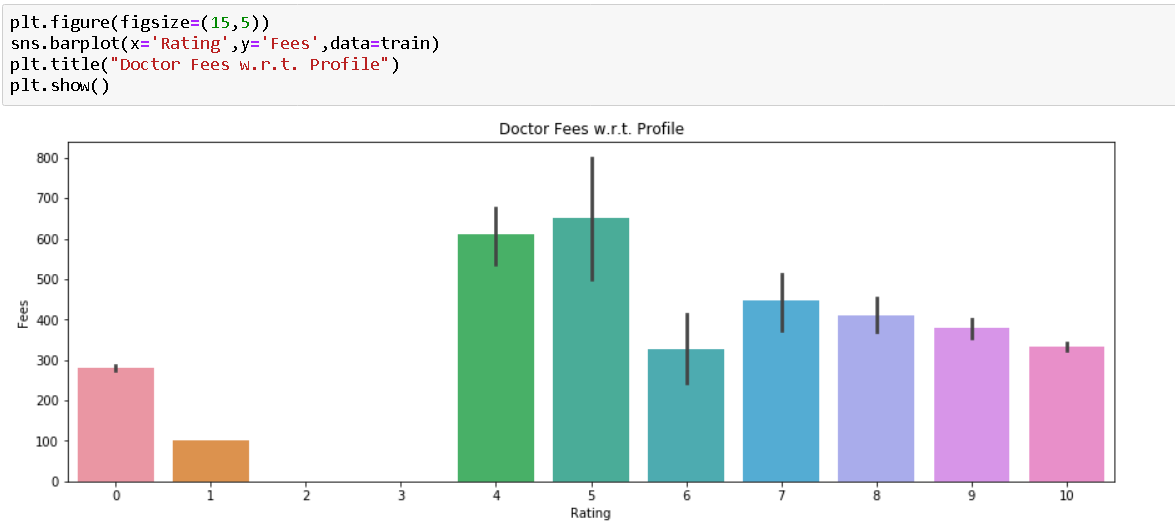
**Key Visualizations:**

Now will look at through Some insight about features

Doctor Fees w.r.t. Profile:



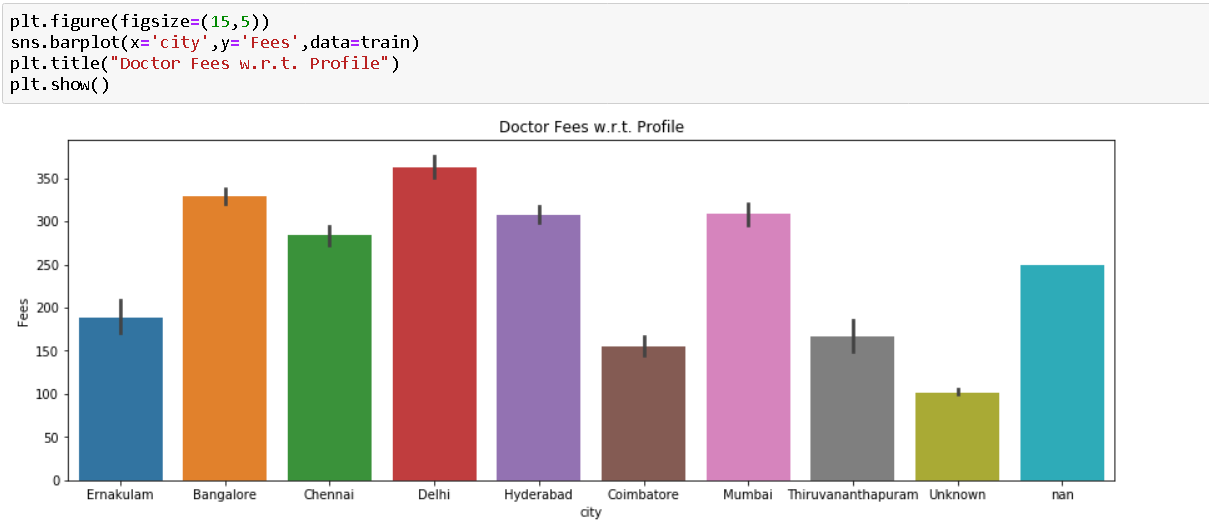
Doctor Fees w.r.t. Rating:



Doctor Fees w.r.t. City:

Fees charged by doctors differ between Cities

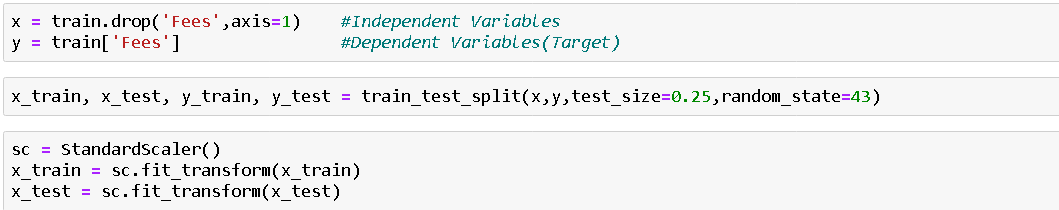
Distribution of the different doctor profile within each city is similar for most cities

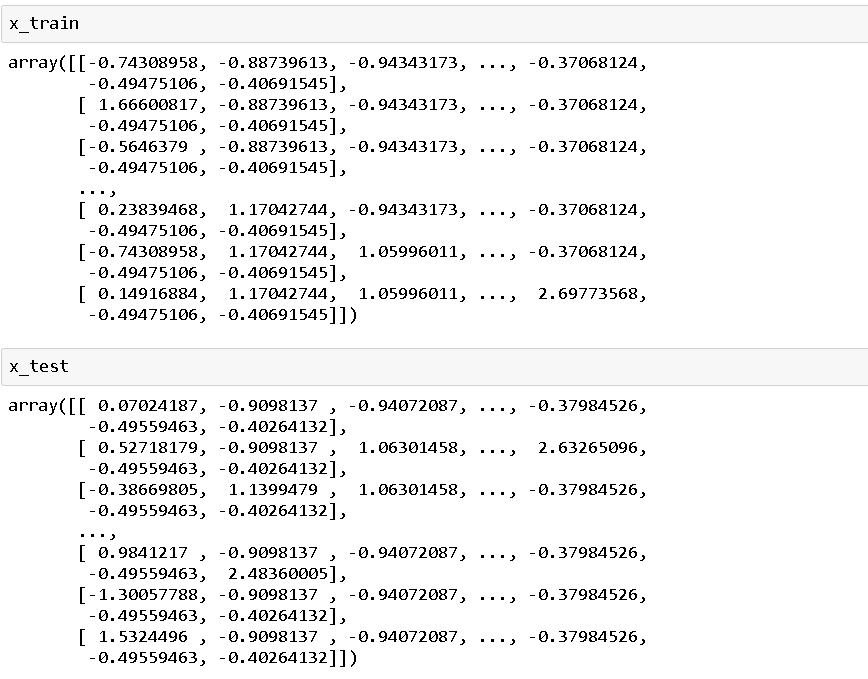


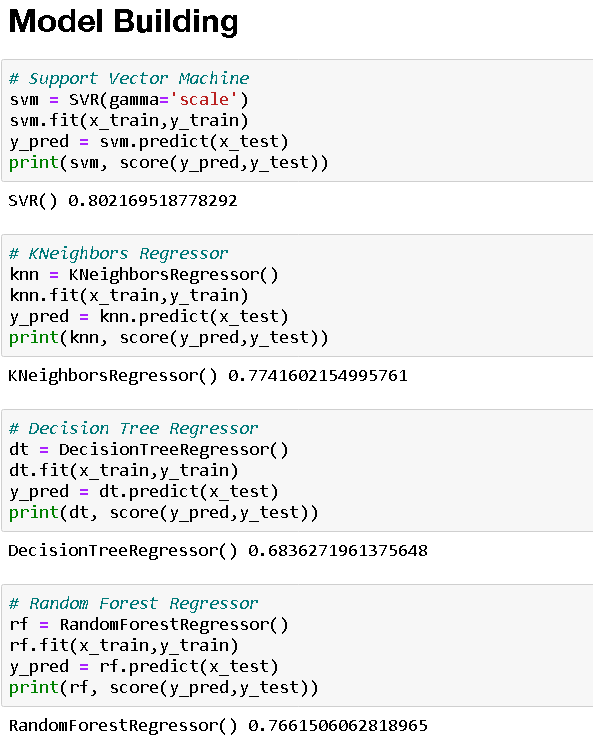
**Building Machine Learning Models**

Now we will train several Machine Learning models and compare their results. we need to use the predictions on the training set to compare the algorithms with each other.

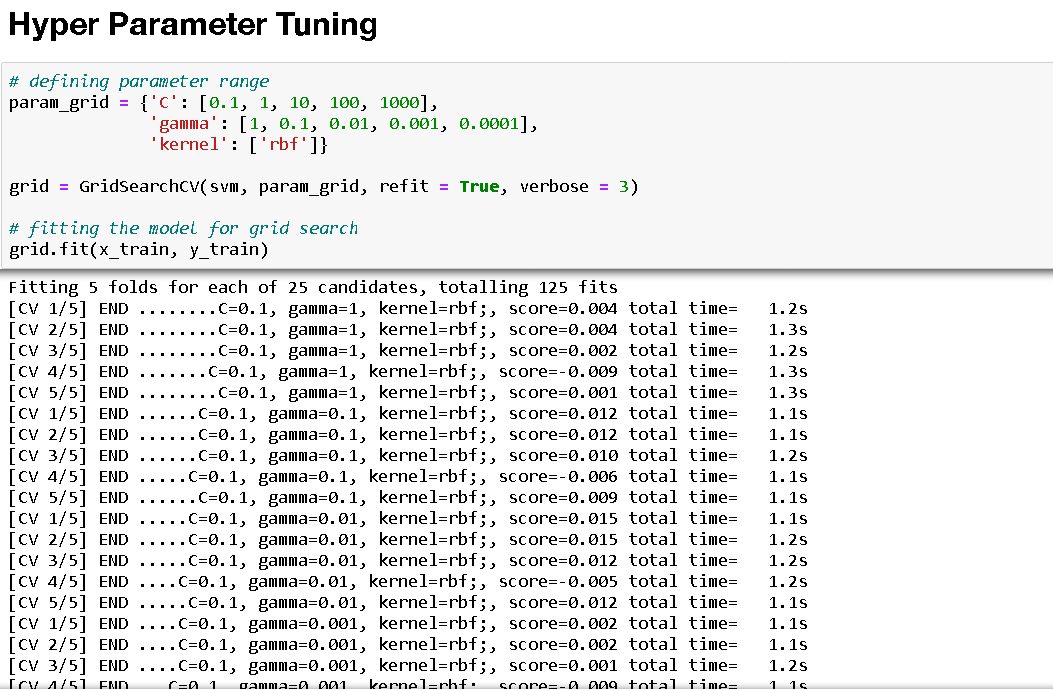
But before that we need to scale our data

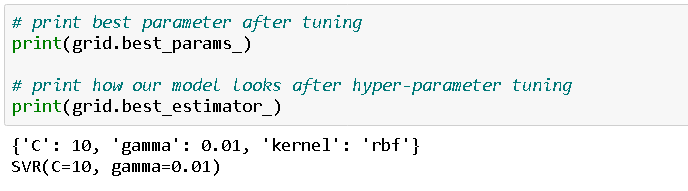




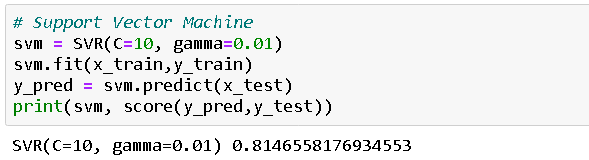


As we are getting good score svc. So, we select it for hyper parameter tuningand print best parameter after tuning.





Finally, we going to take those parameters and finalize our model.



# **Conclusion:**

# We’ve gone through a good portion of the data science pipeline in this article, mainly EDA, Preprocessing and Modeling and we’ve used essential regression models such as Decision Tree Regressor, SVM and Random Forest Regressor.

**Summary**

We started with the data exploration, checked about missing data and learned which features are important. During this process we used seaborn and matplotlib to do the visualizations. During the data preprocessing part, we computed missing values, grouped values into categories and created a few new features. Afterwards we started training 4 different machine learning models, picked SVM (support vector machine) and applied cross validation on it and tuned its performance through optimizing it’s hyperparameter values.

There is still room for improvement, like doing a more extensive feature engineering, by comparing and plotting the features against each other and identifying and removing the noisy features. Another thing that can improve the overall result.