**Doctor's Consultation Fees Prediction**



**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.

**Data Features-**

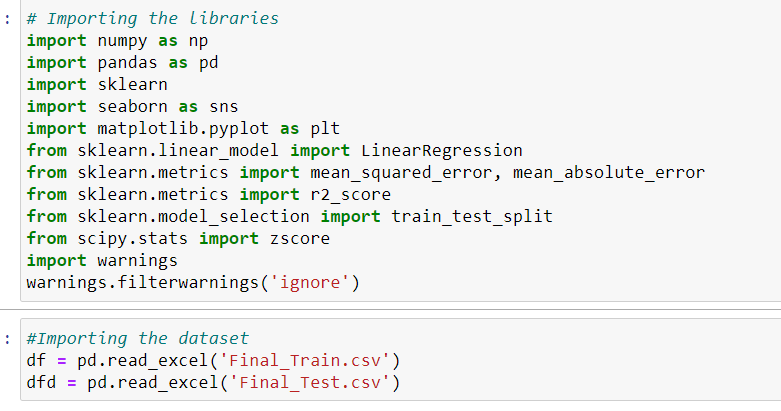
1. Qualification: Qualification and degrees held by the doctor
2. Experience: Experience of the doctor in number of years
3. Rating: Rating given by patients
4. Profile: Type of the doctor
5. Miscellaneous Info: Extra information about the doctor
6. Place: Area and the city where the doctor is located.
7. Fees: Fees charged by the doctor

### Data Pre-processing and Feature creation-

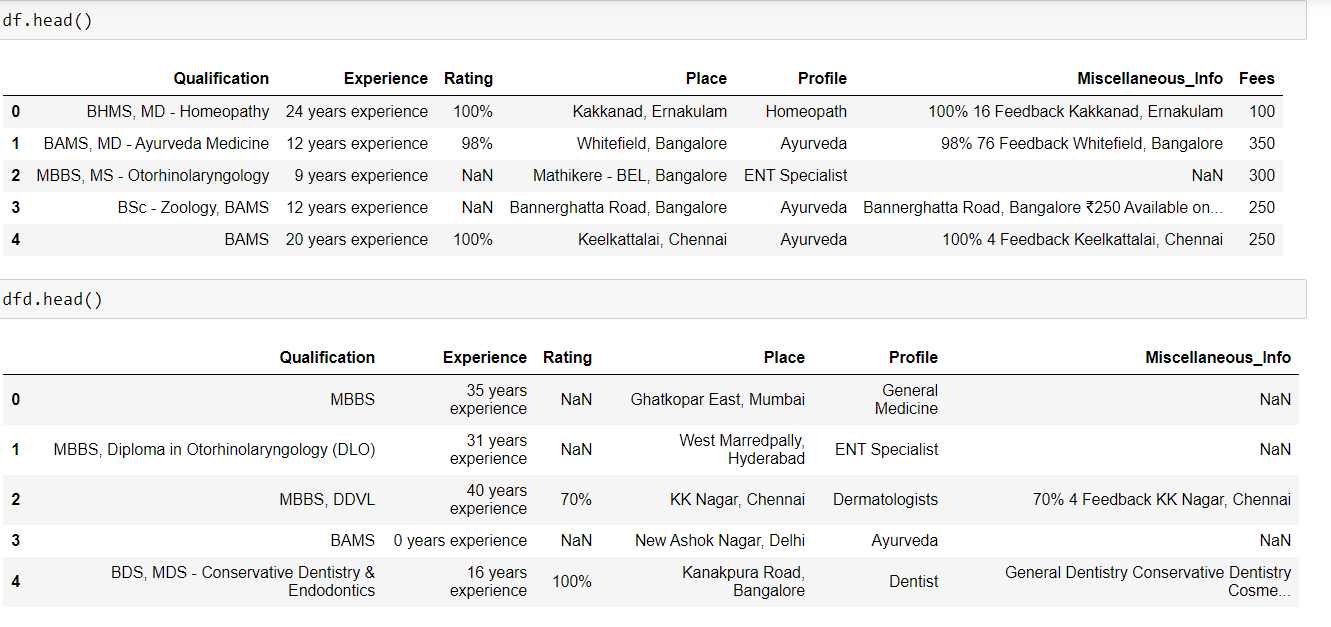
1. Qualification  columns has all the study records of a doctor in a string format by comma separated. E.g. MBBS, MS - Ortho. So, it makes sense to create features out of it like Diploma, Bachelor, Masters, and Extra study, etc., but the problem is we have more than 600 unique study and major combination and we need understanding of all these Medical course to created different features.
2. Extracted just number out of Experience.
3. Removed % from Rating.
4. More than 3000 records doesn't have Miscellaneous Info and in majority case it only consists info about Experience and location, which we already have, so I am not using this column in model.
5. Place column has Area and City info comma separated, so I have created two columns Area & City out of it.

**Steps-**

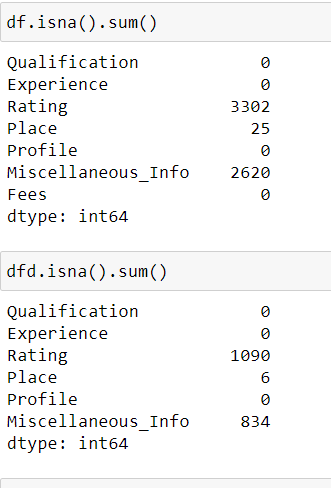
Firstly, importing all the libraries and the dataset in the Jupyter notebook.



View the data,

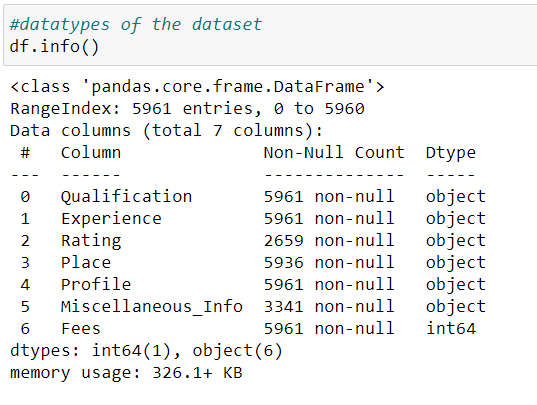


Now, let’s check the missing values from the data,



In the above image we can see that there are some missing values in both dataset and we are going to deal with all the missing values present in the dataset.

Data types of the dataset,



The Qualification column, has the wide range of the qualification .So, I selected the top 10 important qualification.





\*Most of the Doctors have the MBBS qualification(Bachelor of Medicine and Bachelor of Surgery),So these the surgeons.

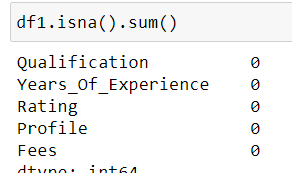
\*The next are the doctors with the BDS degree(Bachelor of Dental Surgery), So these are the Dentists.

\*The next are the doctors with the BAMS degree(Bachelor of Ayurvedic Medicine and Surgery),There are the aurvedic doctor.

**Filling Missing Values**-

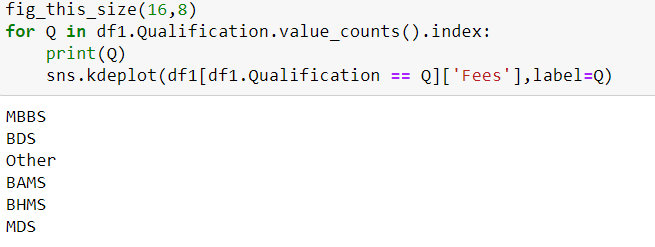
Rating has the highest missing values.So, we will replace the rating values with mean method.

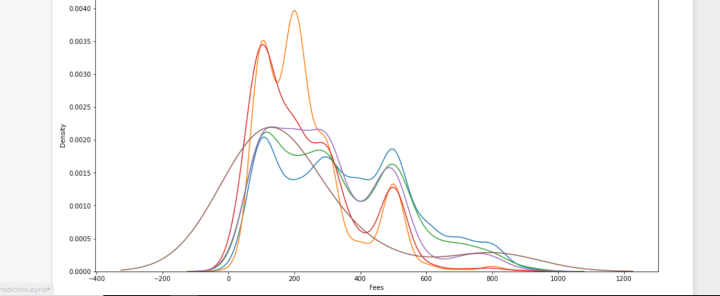




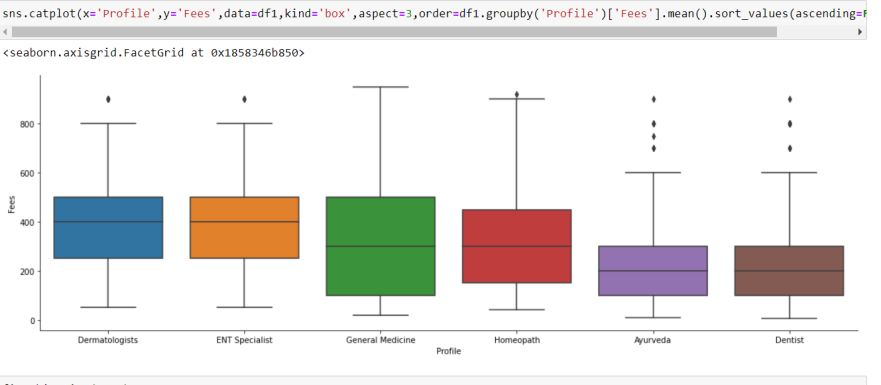
We have removed all the missing values from our dataset.

The next step is the Data Visualization

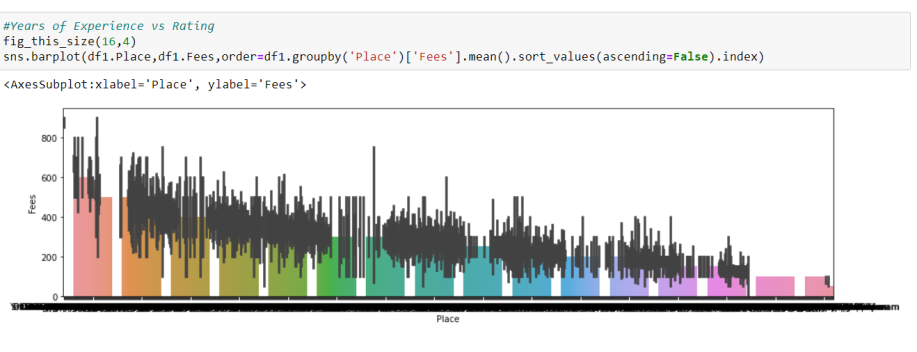




The doctors with BDS & MDS qualification charge high fees. We can see that the fees or a dentist is much higher than that of an MBBS.



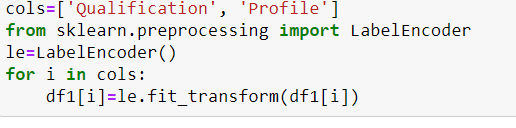




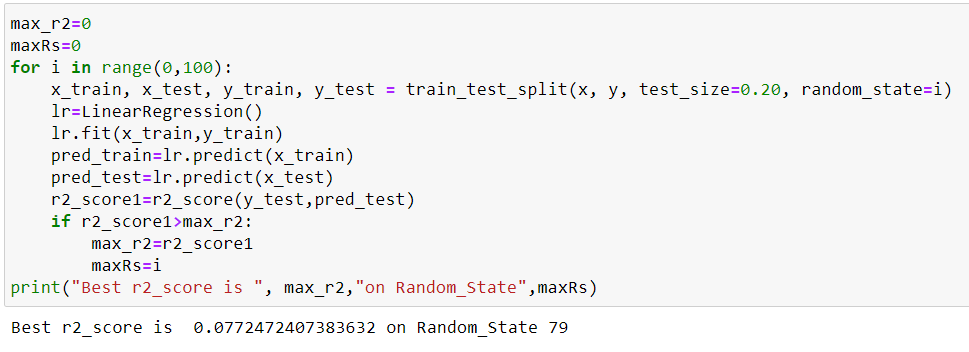
Observations in Data Visualization-

1. Bangalore has the highest average Doctor's Consultancy Fees compare to other cities.
2. Fees charged by the doctor is different from city to city and according to their profession.

**Encoding the data**



**Finding the best Random State**-

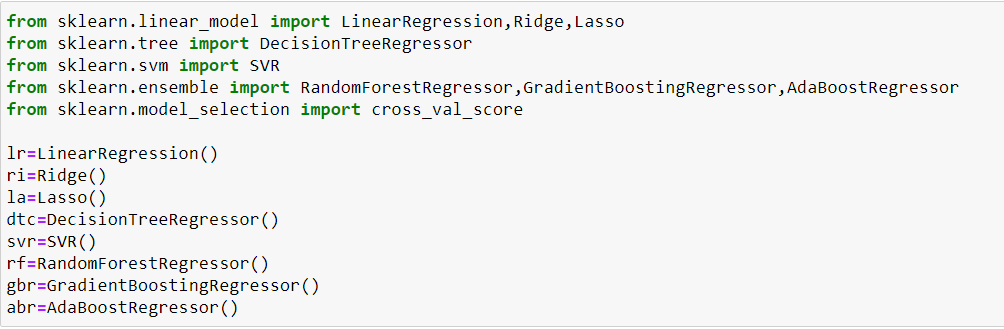


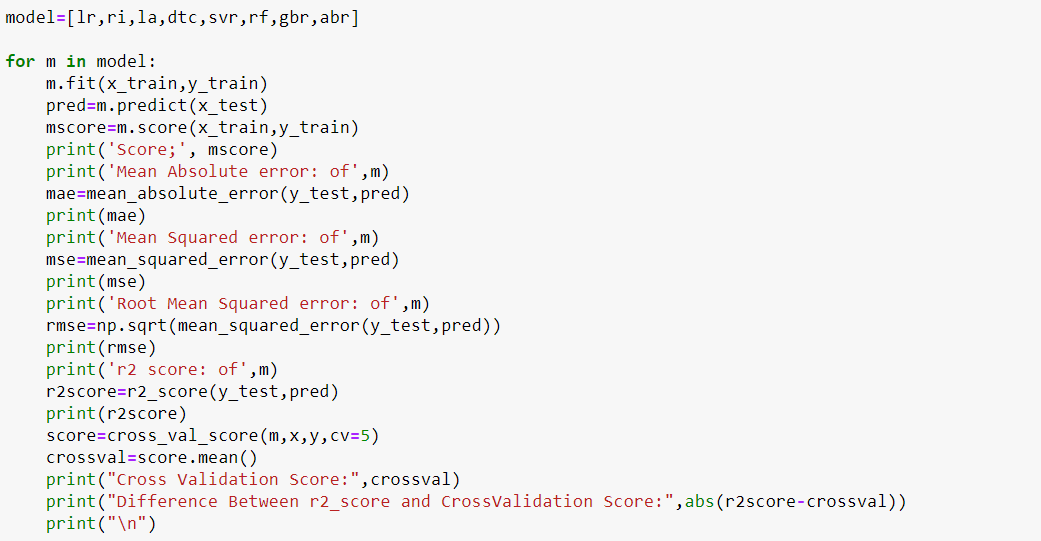
So we will use Random State Value of 79.

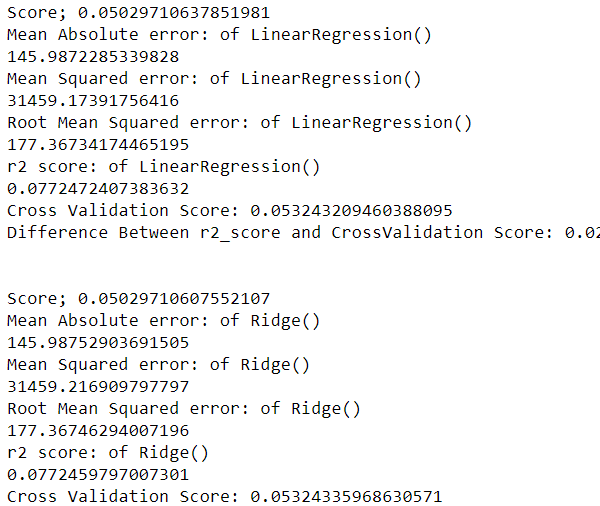
**Performing Train Test Split-**

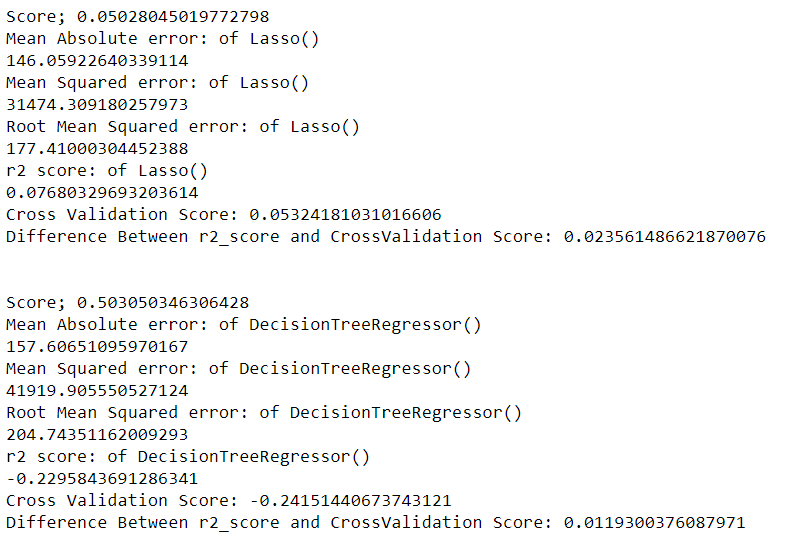


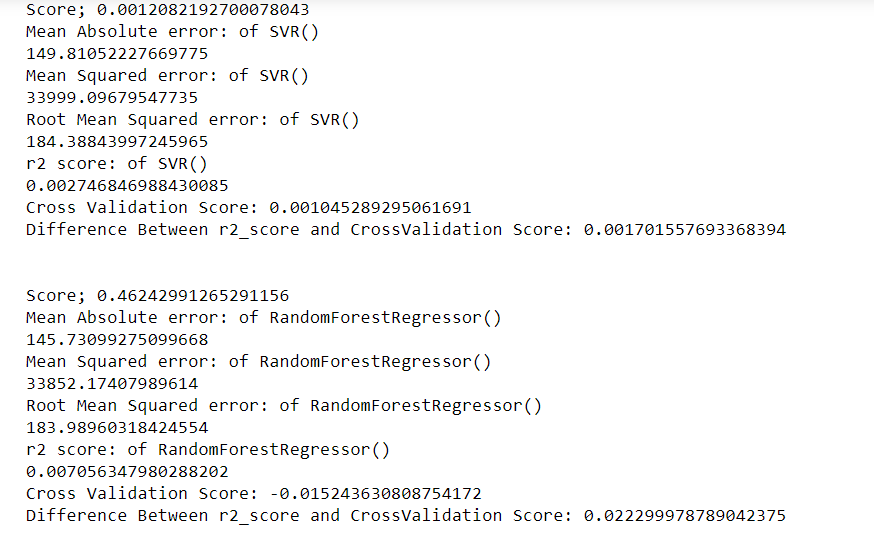
**Training our Model-**

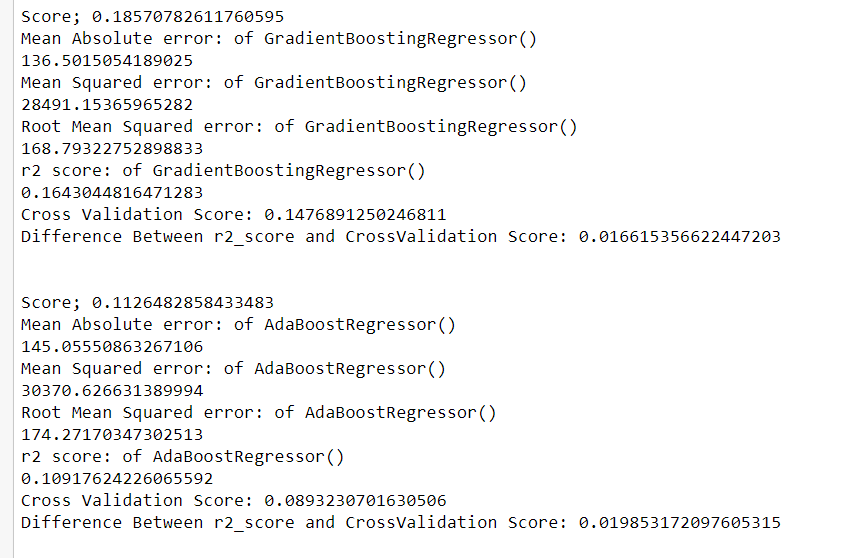
****

****

****

****

****

****

**Analysis-**

* After doing data analysis it is clear that Rating and Experience has less correlation with target variable Fees.
* It was clear from Box plots of Fees for Profile and City that these columns are more important for segregation.
* But without adding Education information we can still not find good separation in data groups.

### We can see that Random Forest Regressor is our best model.