



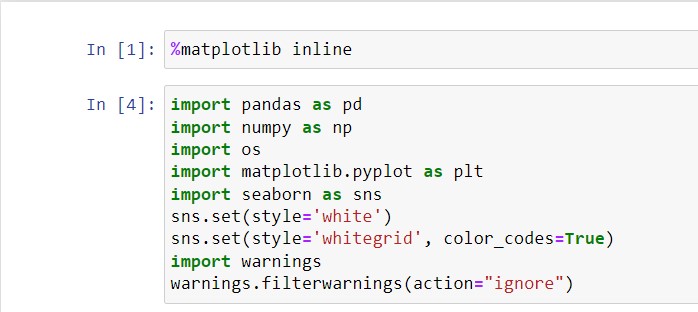
**Project Report**

**Introduction:** Practically this set is about the Graduate Admissions data i.e. Given a set of standardized scores like GRE, TOEFL, SOP standard scores, LOR standard scores, what is probability (basically I have done a YES/NO scenario) of gaining admission into a particular school.

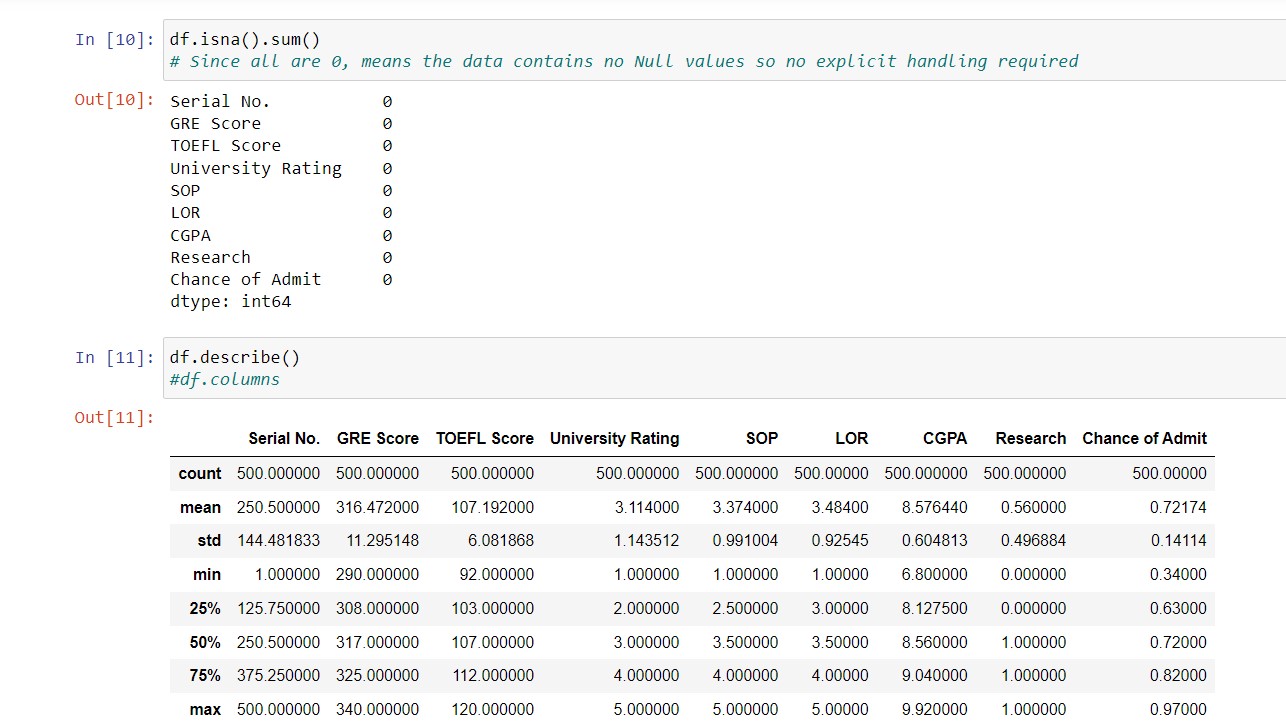
**Dataset:** This dataset is created for prediction of graduate admissions and the dataset link:

* Features in the dataset:
* GRE Scores (290 to 340)
* TOEFL Scores (92 to 120)
* University Rating (1 to 5)
* Statement of Purpose (1 to 5)
* Letter of Recommendation Strength (1 to 5)
* Undergraduate CGPA (6.8 to 9.92)
* Research Experience (0 or 1)
* Chance of Admit (0.34 to 0.97)

**Important Libraries and Data:** Importing libraries and setting the default style.

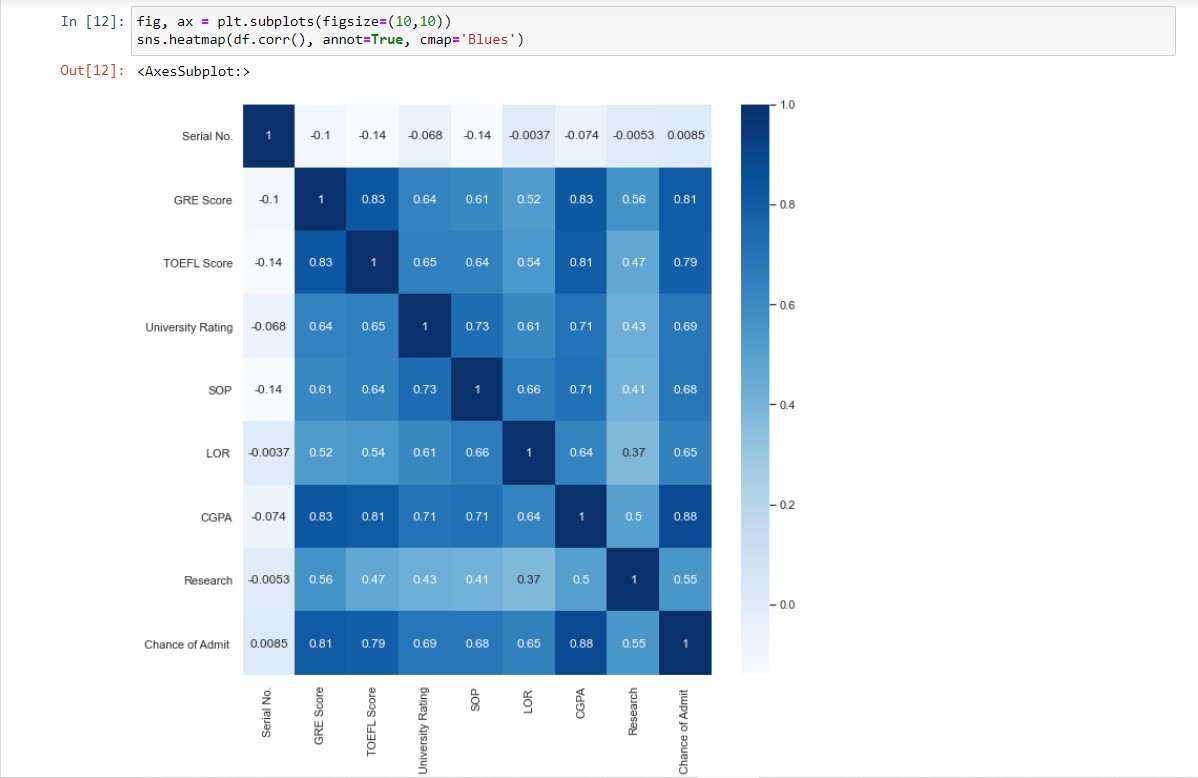


Next, let's import our dataset and see what we're working with.



## **Exploratory Analysis:**

Let's plot a heatmap to see the correlation of all the features compared to Chance to Admit:



The top three features that affect the Chance to Admit are:

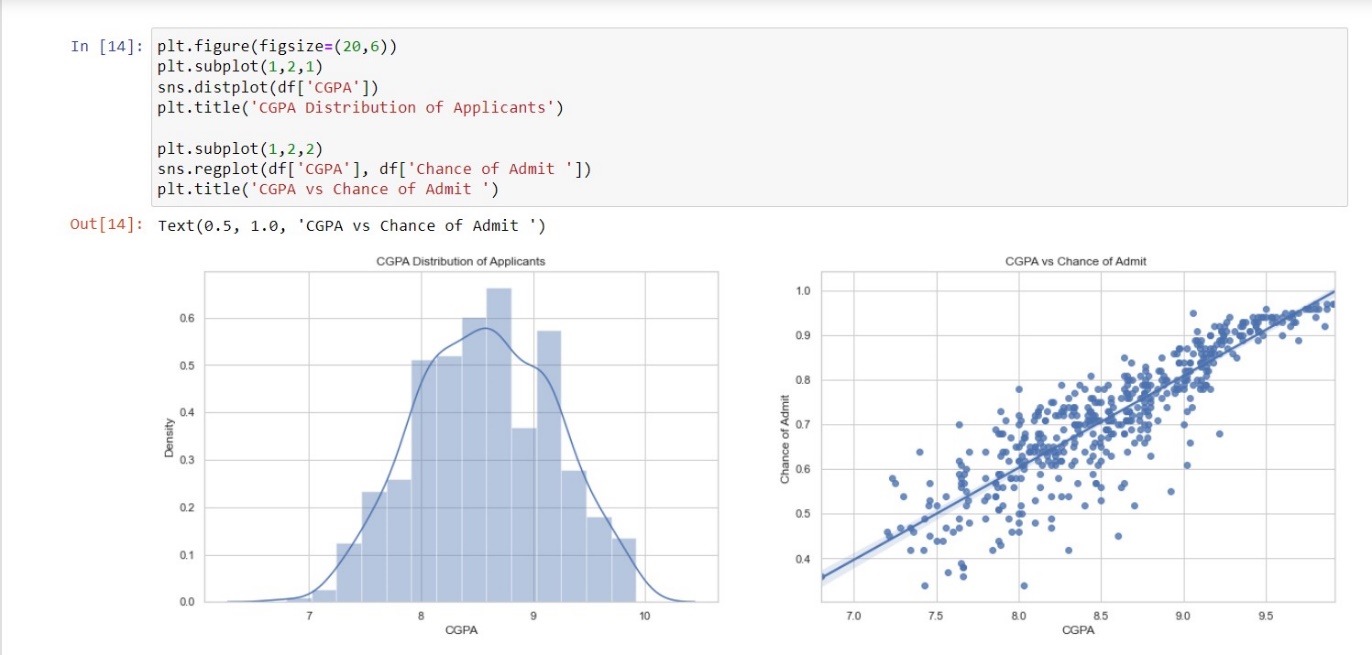
1. CGPA
2. GRE Score
3. TOEFL Score

Let's explore these three features to get a better understanding.

**CGPA**

The Cumulative Grade Point Average is a 10-point grading system. From the data shown below, it appears the submissions are normally distributed. With a mean of 8.6 and standard deviation of 0.6.

**CGPA vs Chance of Admit**

It appears as applicant's CGPA has a strong correlation with their chance of admission.

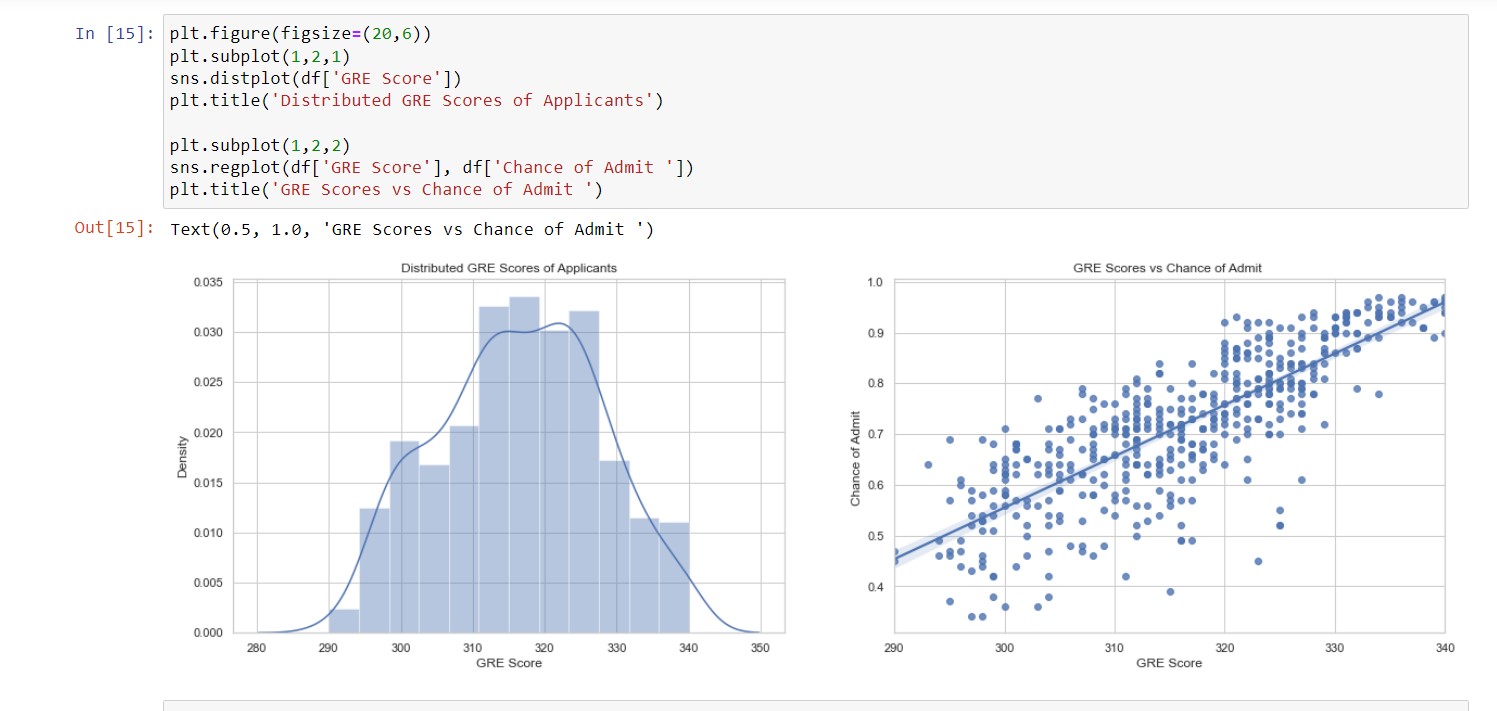
**GRE Score**

The Graduate Record Examination is a standardized exam, often required for admission to graduate and MBA programs globally. It's made up of three components:

1. Analytical Writing (Scored on a 0-6 scale in half-point increments)
2. Verbal Reasoning (Scored on a 130-170 scale)
3. Quantitative Reasoning (Scored on a 130-170 scale)

In this dataset, the GRE Score is based on a maximum of 340 points. The mean is 317 with a standard deviation of 11.5.

**GRE Score vs Chance of Admit**

GRE scores have a strong correlation with the chance of admission however not as strong as one's CGPA.

**TOEFL Score**

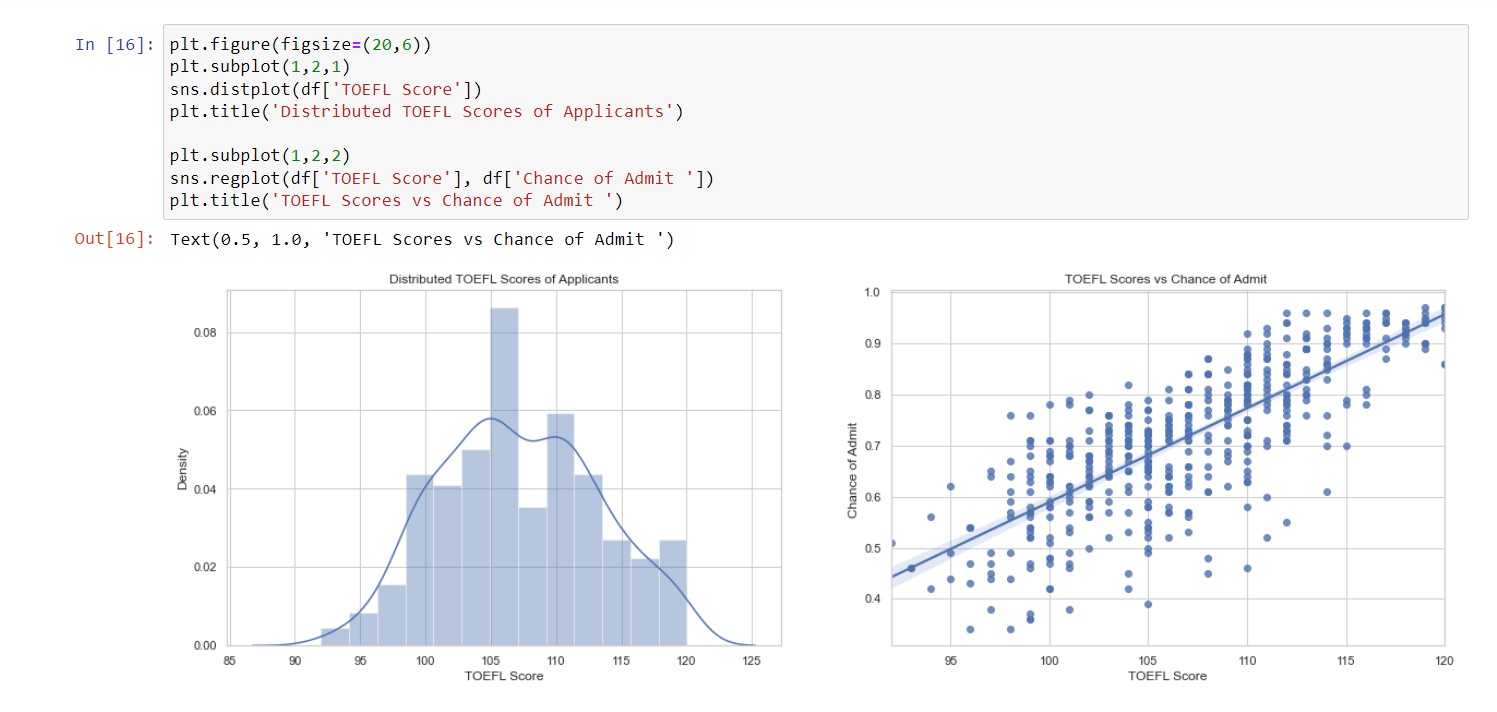
The Test of English as a Foreign Language is a standarized test for non-native English speakers that are choosing to enroll in English-speaking universities.

The test is split up into 4 sections:

1. Reading
2. Listening
3. Speaking
4. Writing

All sections are scored out of 30, giving the exam a total score of 120 marks. In this dataset, the TOEFL scores have a mean of 107 and a standard deviation of 6.

**TOEFL Score vs Chance of Admit**

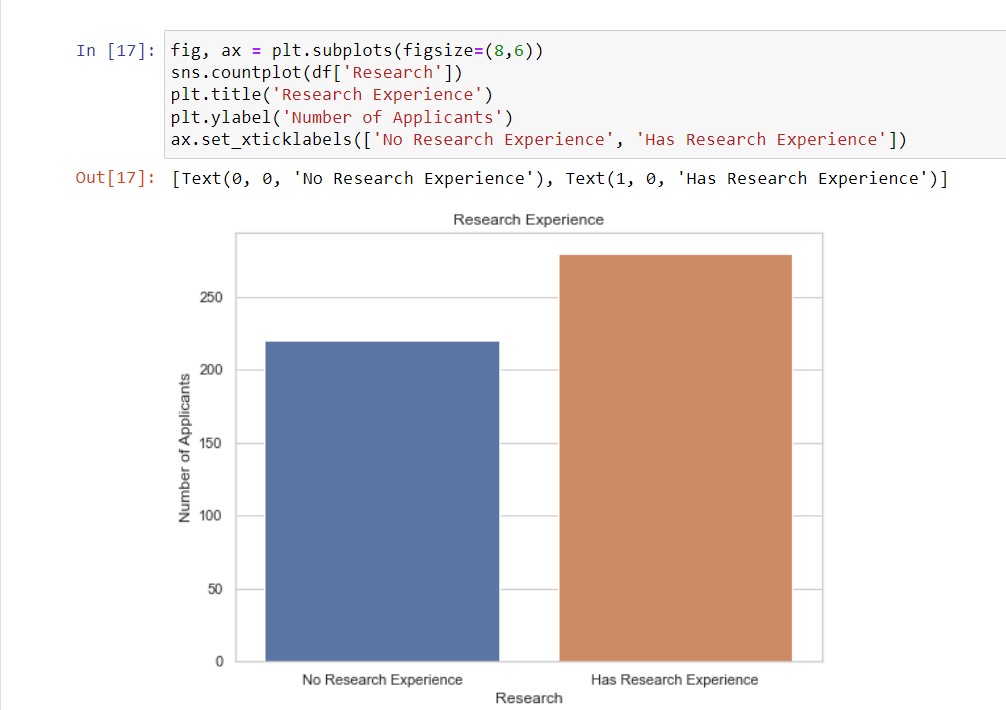
Like GRE scores, the scores received for the TOEFL strongly correlate to an applicant’s chance of admission.

For my curiosity, I want to explore the data a little bit further regarding research and university rankings. Even though they hold a lower importance in the chance of admission, it would be nice to understand their characteristics in the dataset.

### ****Research****

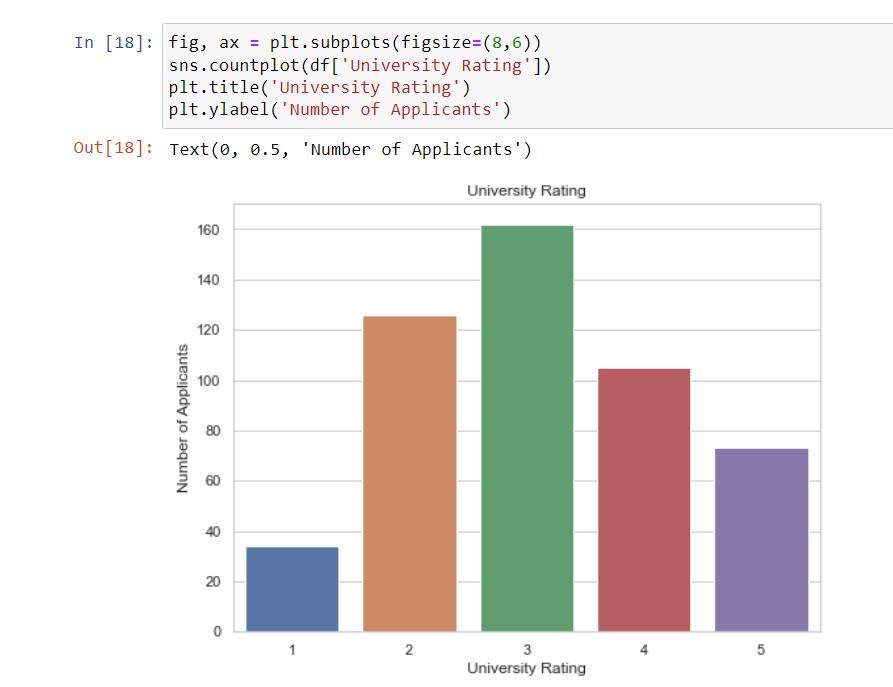
Let's explore how many applicants have research experience.

It seems the majority of applicants have research experience. However, this is the least important feature, so it doesn't matter all too much if an applicant has the experience or not.

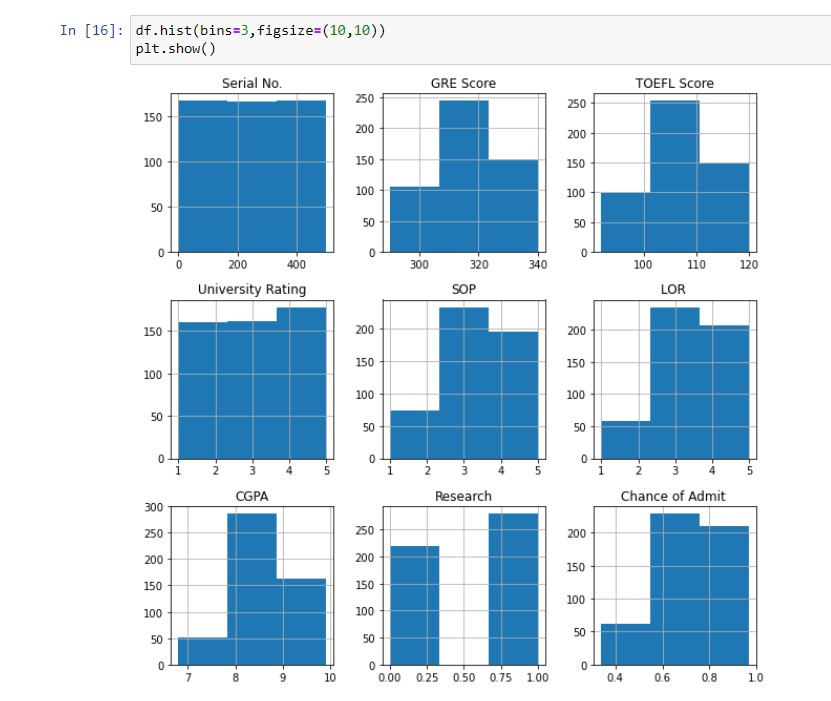


### ****University Rating****

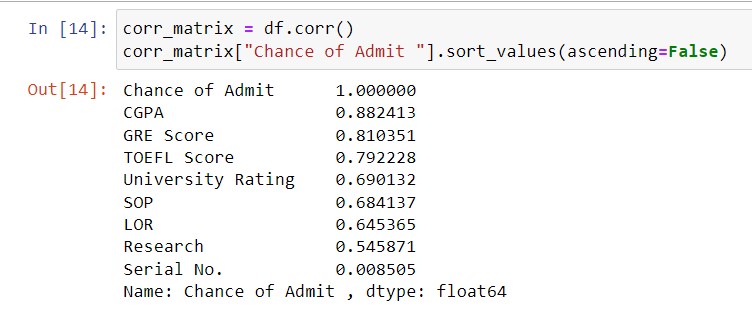
Let's see the distribution of applicants coming from each kind of university. Most applicants come from a tier 3 and tier 2 university.



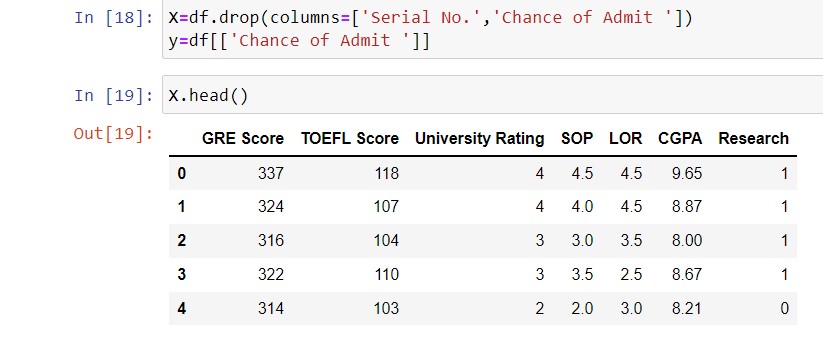
## **Distribution of variables in the data**

Looking at the distribution of variables in the data to see whether certain values in excess don't affect the model too much. We can actually use 500 values each from Research column but in the current model I am passing by this intervention.

## **Correlation between Variables**



## **Looks like all variables are important variables for predicting Admit chances**

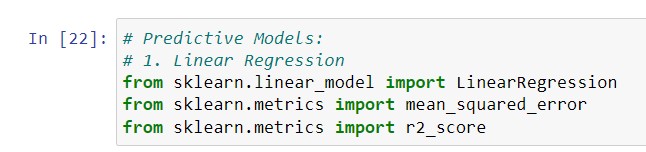
Hence, taking all columns in independent variables(predictors) and Chance of Admit as Dependent variable (to be predicted).

### Splitting Data into Testing and Training Set

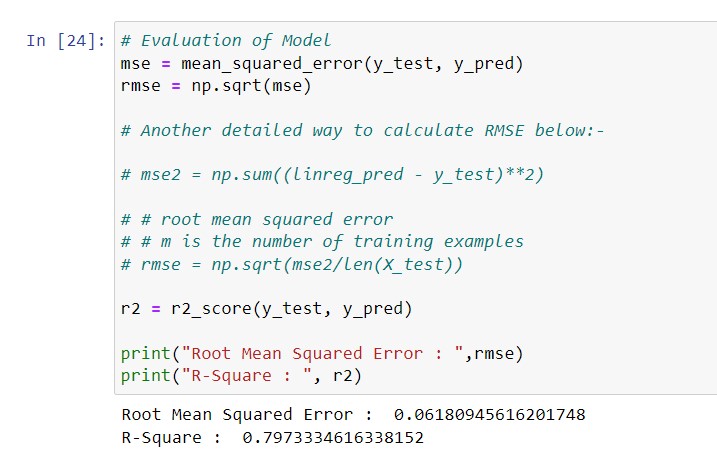
### 

# **Predictive Models**

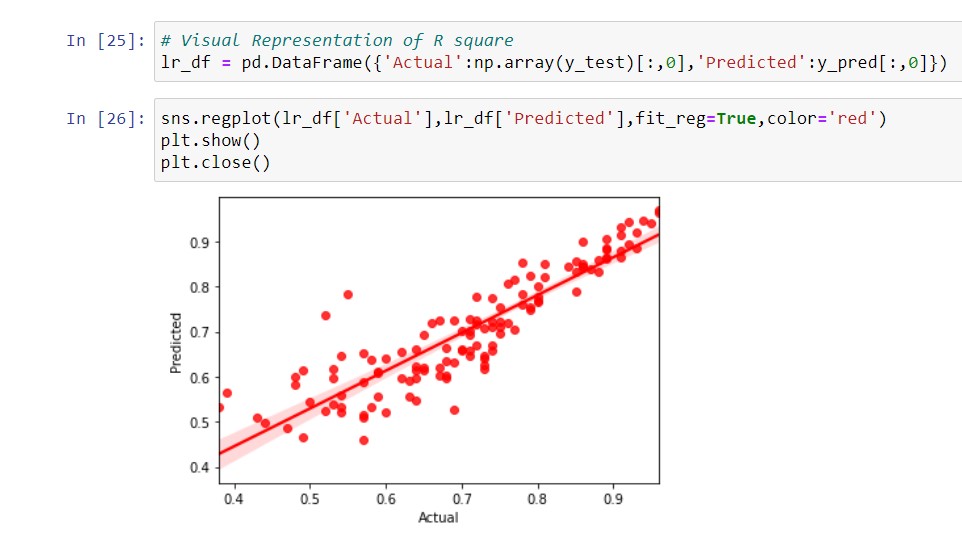
## **Linear Regression**



### Building the model and Predicting on Test Set

Evaluation of Model

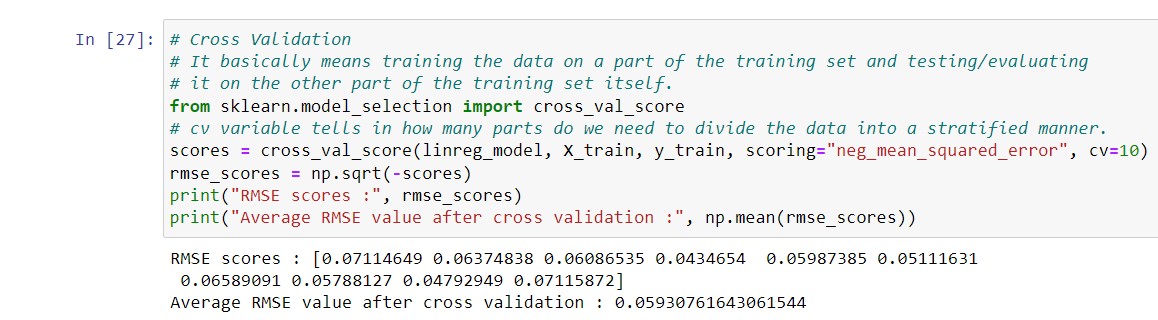
### Visual Representation of R square

R² score or the coefficient of determination explains how much the total variance of the dependent variable can be reduced by using the least square regression.

## **Fine Tuning Model**

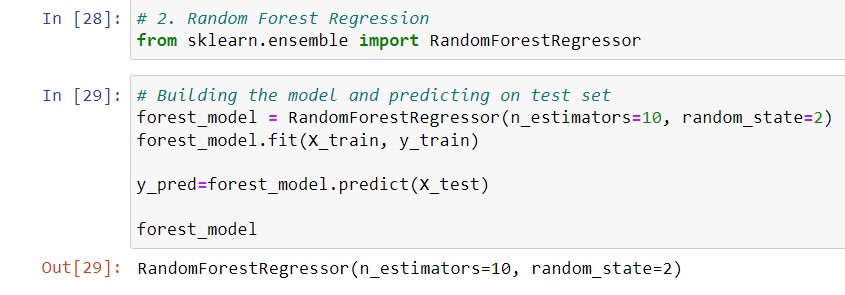
## **Cross Validation**

#### It basically means training the data on a part of the training set and testing/evaluating it on the other part of the training set itself. There is a possibility that we may be overfitting our model. Hence, just to verify it is better to cross validate data on your training set itself. This is usually done when we have divided are dataset into 3 main sets: Training data, Test data and Validation Set. Usually, we don't touch the test set until we are ready to launch a model, we are confident about, so we need to use part of the training set for training, and part for model validation.



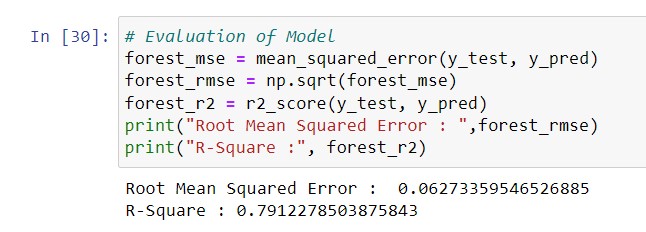
## **Random Forest Regression**

Building the model and predicting on test set



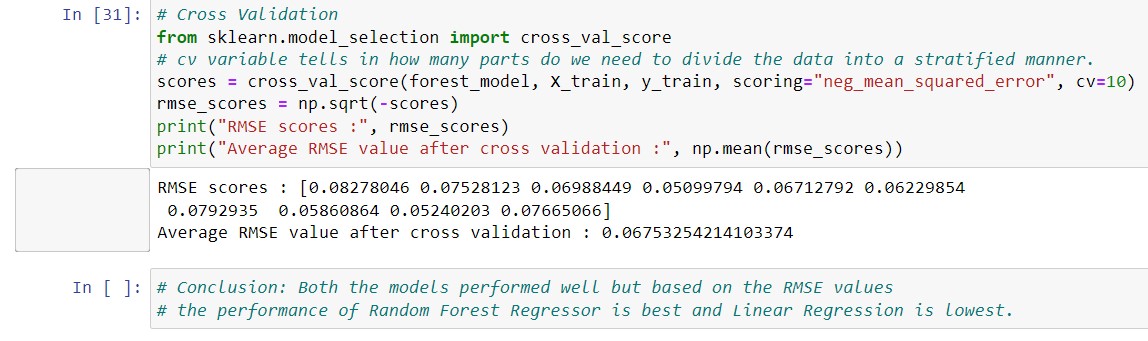
### Evaluation of Model

The Random Forest model has a higher R-square than the Decision Tree model and a lower RMSE. This means that this model performs better than DT model but there is also a possibility of overfitting. Hence, we should cross-validate on this data for all the ML models created for prediction.



## **Fine Tuning Model**

## **Cross Validation**



### Conclusion:

### All models perform well but based on the RMSE values the performance of Random Forest Regressor is best and Linear Regression is lowest.