**PROJECT REPORT**



**GRADUATE ADMISSION PREDICTION**

**SUBMITTED BY:- Vaishnavi Wasre USN:- 3GN18CS119**

**DEPT:- COMPUTER SCIENCE**

# UNDERTAKING

I declare that the work presented in this project titled

“UNIVERSITY ADMISSION PREDICTION”, submitted to the TIMTS, for the award of the Internship in Data Science, is my original work. I have not plagiarized or submitted the same work for the award of any other Internship. In case this undertaking is found incorrect, I accept that my Project may be unconditionally withdrawn.

October,2021 Vaishnavi Wasre

# CERTIFICATE

This is to certify that the work contained in the project titled “UNIVERSITY ADMISSION PREDICTION”, by Vaishnavi Wasre, has been carried out under my supervision and that this work has not been submitted elsewhere for internship.

Times Institute of Management and Technical Studies

Data science and ML

Bengaluru, Karnataka 560103

# ACKNOWLEDGEMENT

I take upon this opportunity to acknowledge the many people whose prayers and support meant a lot to me. I am deeply indebted to my mentor **Mr. Sumit Chatterjee** who motivated me along the way.

I would like to thank all my teachers who help me in this project.

I further thank my friends.My heartfelt thanks to parents who supported me a lot. I owe my sincere gratitude towards the almighty God.

Finally, I would like to wind up by paying my heartfelt thanks to TIMTS institute who provided me with this great opportunity

# CONTENTS

Topic Page No.

1. Purpose 6
2. Dataset 7
3. Theory of linear Regression 8
4. Implementation 11
5. Import Libraries 12
6. Conslusion 19
7. References 20

## Purpose

To apply for a master's degree is a very expensive and intensive work. With this kernel, students will guess their capacities and they will decide whether to apply for a master's degree or not.

So, basically 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. All those folks who are preparing for MS, might point out this question, from where did you get SOP & LOR scores. These aren’t public figures ? I mean yes, it might not be public, but don’t you think universities might be grading these applications on some scale of rating so that the scores can be standardized. Hence the SOP, LOR scores.

### **Dataset**

This dataset is created for prediction of graduate admissions and the dataset link is below: https://www.kaggle.com/mohansacharya/graduate-admissions 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)

#### THEORY OF LINEAR REGRESSION

In machine learning, any problem can be taken as classification problem or regression problem. Different from classification, the prediction value is always continuous in a certain range, simply to say, target value is the probability of one event happening. For example, a medicine institution wants to diagnose a patient according to known a set of data of patients health records. If we want to know the patient is ill or not, this is a classification problem. But sometimes the result is not absolute. Before getting ill, the patient wants to know the probability of being ill, which can be considered as regression problem.

Linear regression is the simplest statistic model to imply the relation between variables. Firstly, let us only consider two variables relation, then I will extend it to multi variables linear regression .

A. Simple Linear Regression:-

With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.

This requires that you calculate statistical properties from the data such as means, standard deviations, correlations and covariance. All of the data must be available to traverse and calculate statistics.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1\*x

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

B. Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) [regression](https://www.investopedia.com/terms/r/regression.asp) because it involves more than one explanatory variable.

#### *yi*=*β*0+*β*1*xi*1+*β*2*xi*2+...+*βpxip*+*ϵ*

*yi*=dependent variable

*xi*=explanatory variables

*β*0=y-intercept (constant term)

*βp*=slope coefficients for each explanatory variable

*ϵ*=the model’s error term (also known as the residuals)

## IMPLEMENTATION

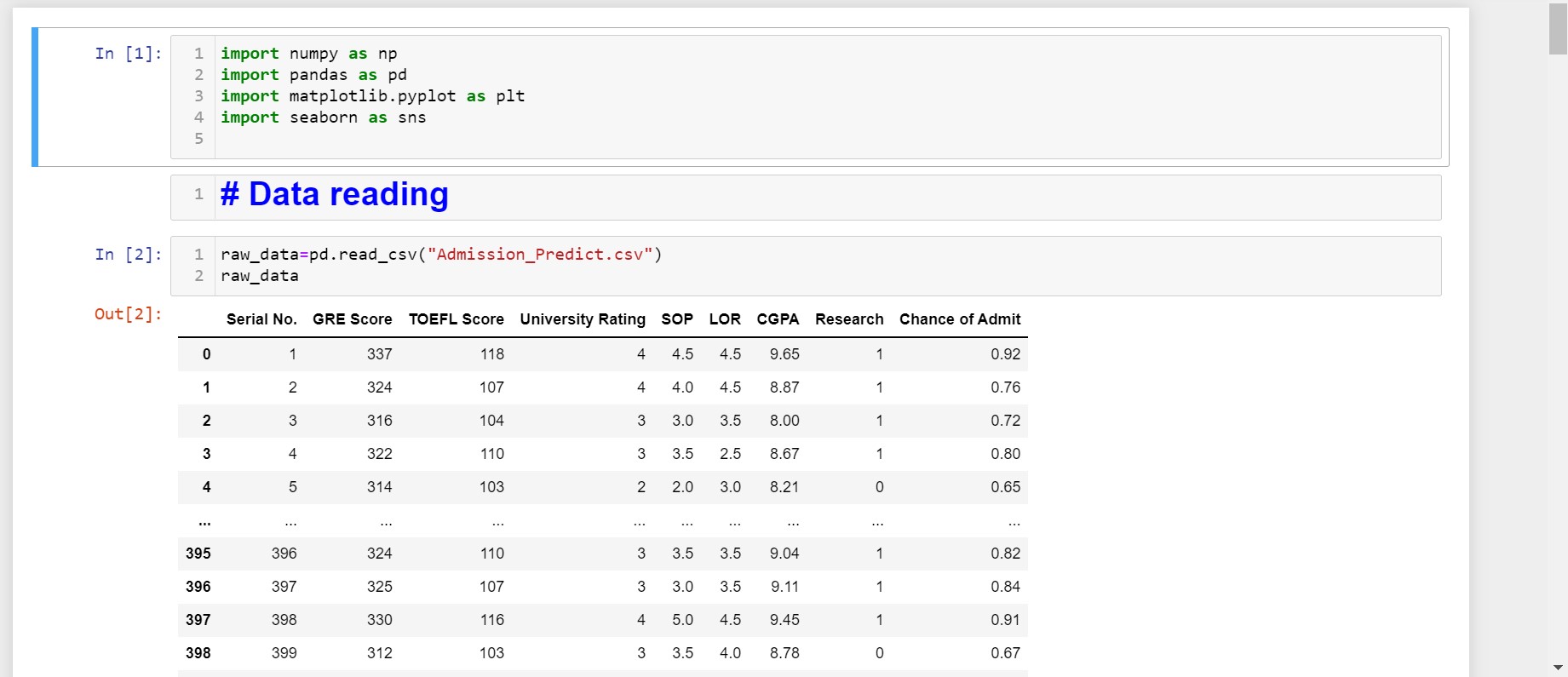
This work used Python programming for this project, as it is a high-level programming language and it has vast libraries and Python automates tasks and makes it efficient. Firstly, we need to install Python then we need to import some libraries, they are:

1) Numpy: Numpy is used for multi-dimensional arrays, it does element to element operations and it also has different methods for processing arrays. 2) Panda: Pandas is one of the highly used python libraries, it provides high performance. It manipulates data and it makes data analysis fast and easy.

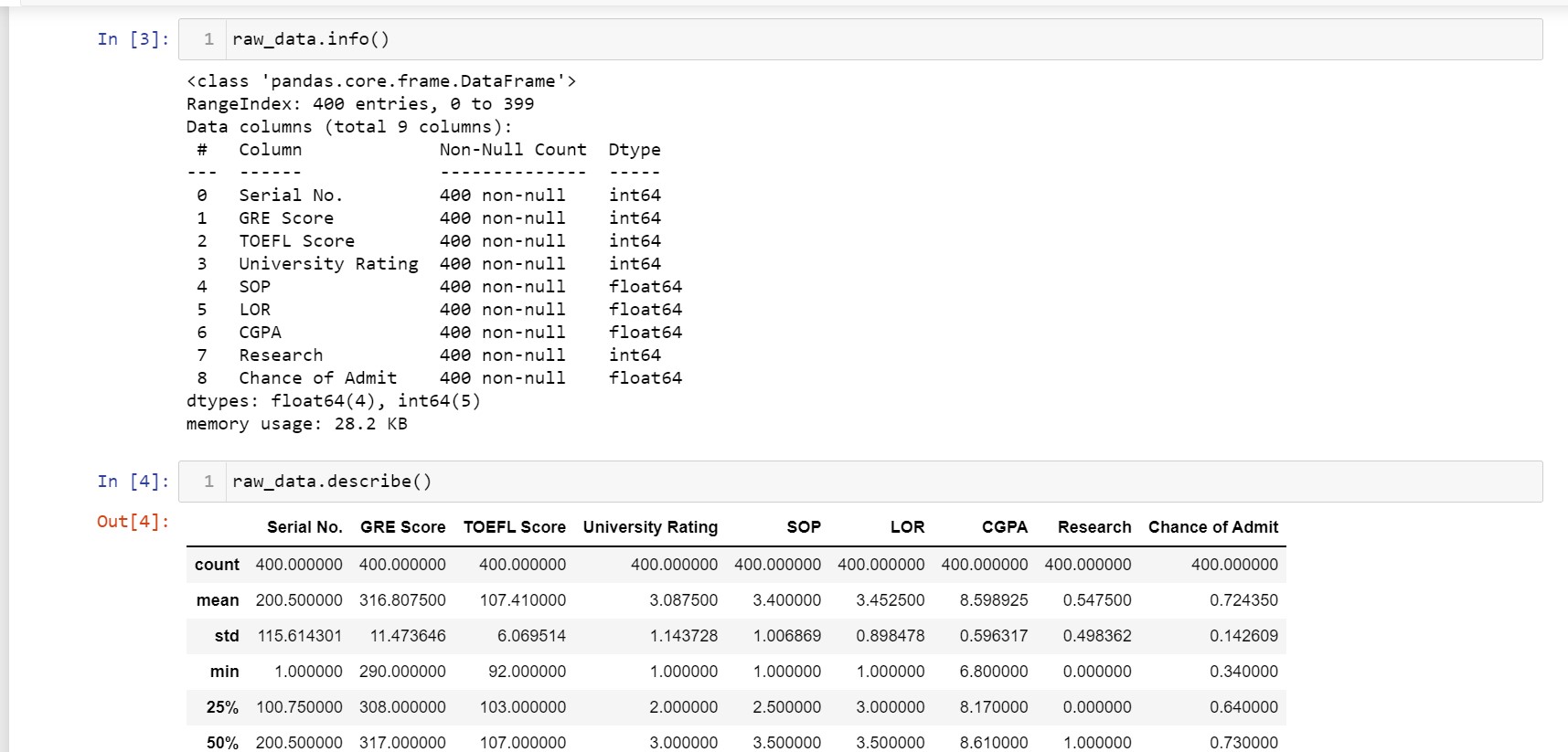
3) Sklearn: It is most useful library, this library contains lot of efficient tools, it is used to build models like statistical modelling including classification, regression, clustering. After loading required packages, we divide dataset as training and testing as follows, here 80 % of dataset is taken as training and remaining 20 % as to perform test.

### IMPORT LIBRARIES and LOAD DATA

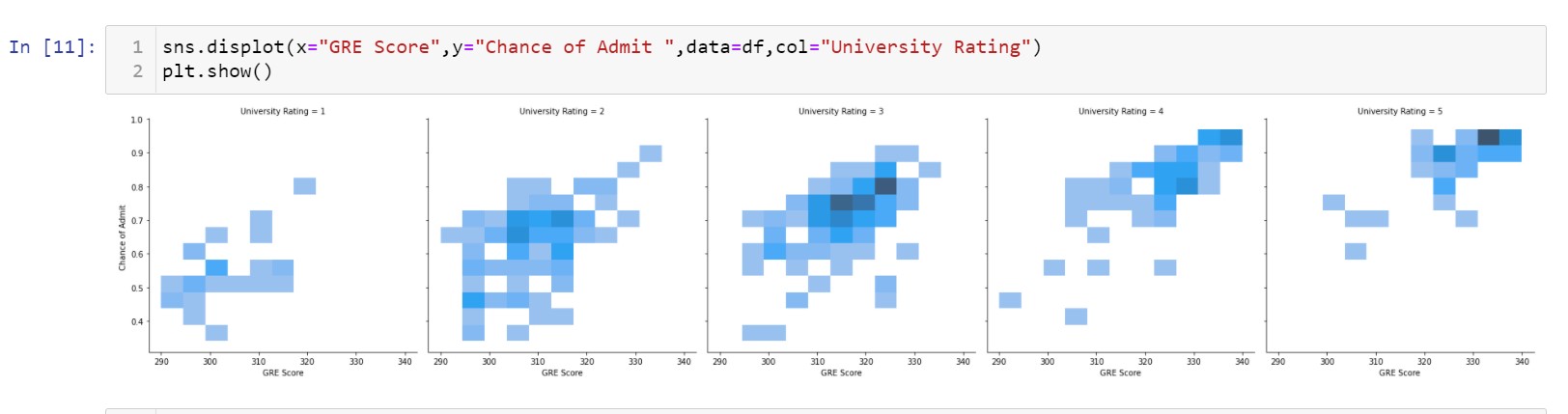
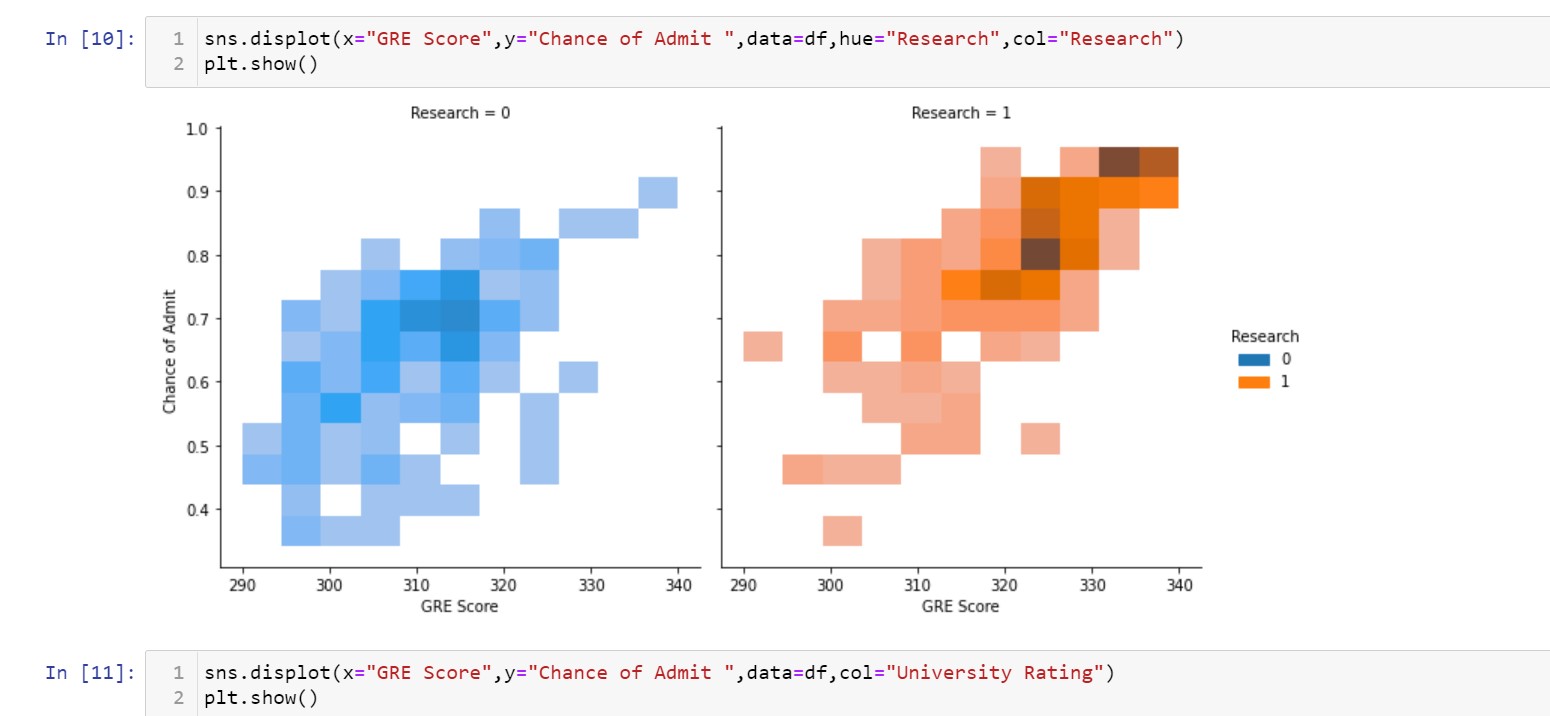
First, let’s import all the modules, functions and objects we are going to use.We can load the data directly from the UCI Machine Learning repository. We are using pandas to load the data. We will also use pandas next to explore the data both with descriptive statistics and data visualization.



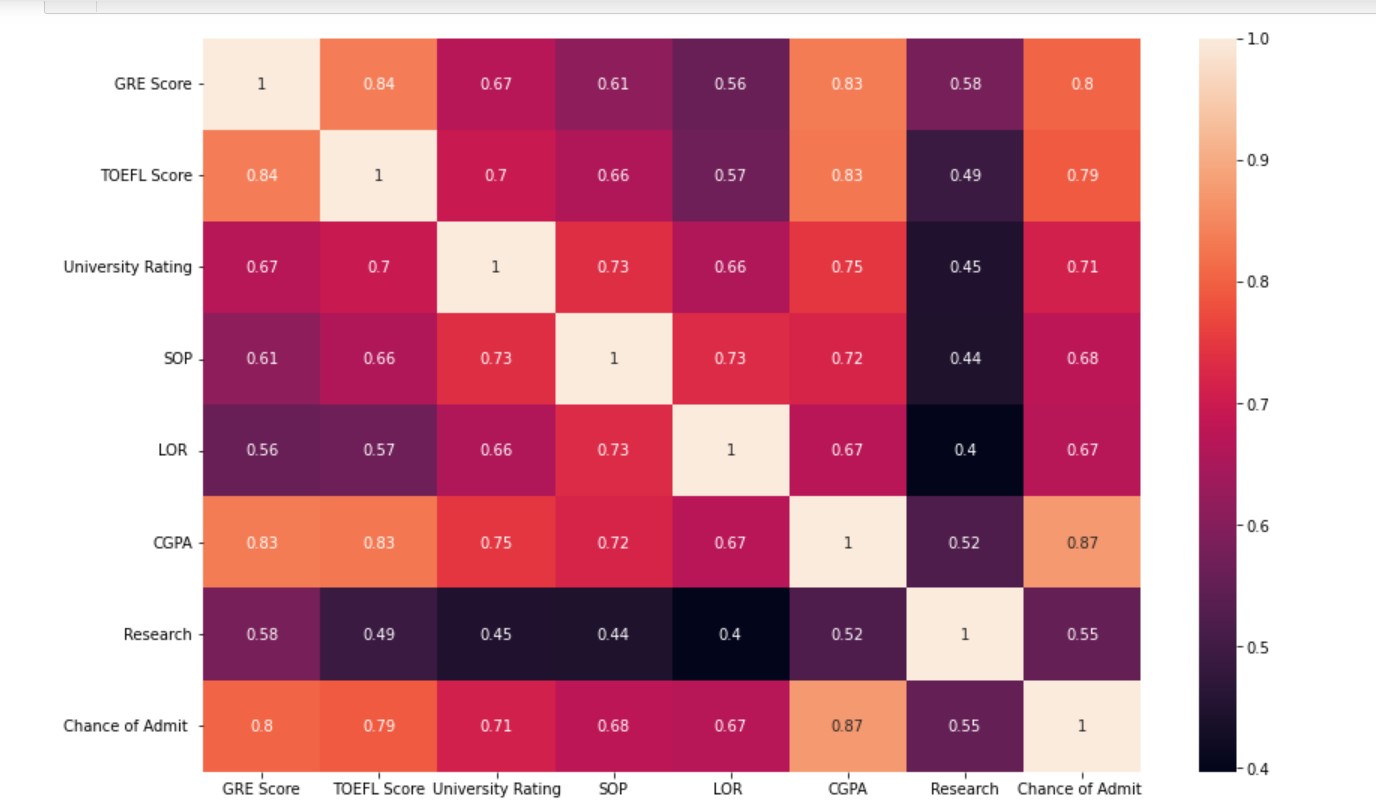
### Analysis the Data



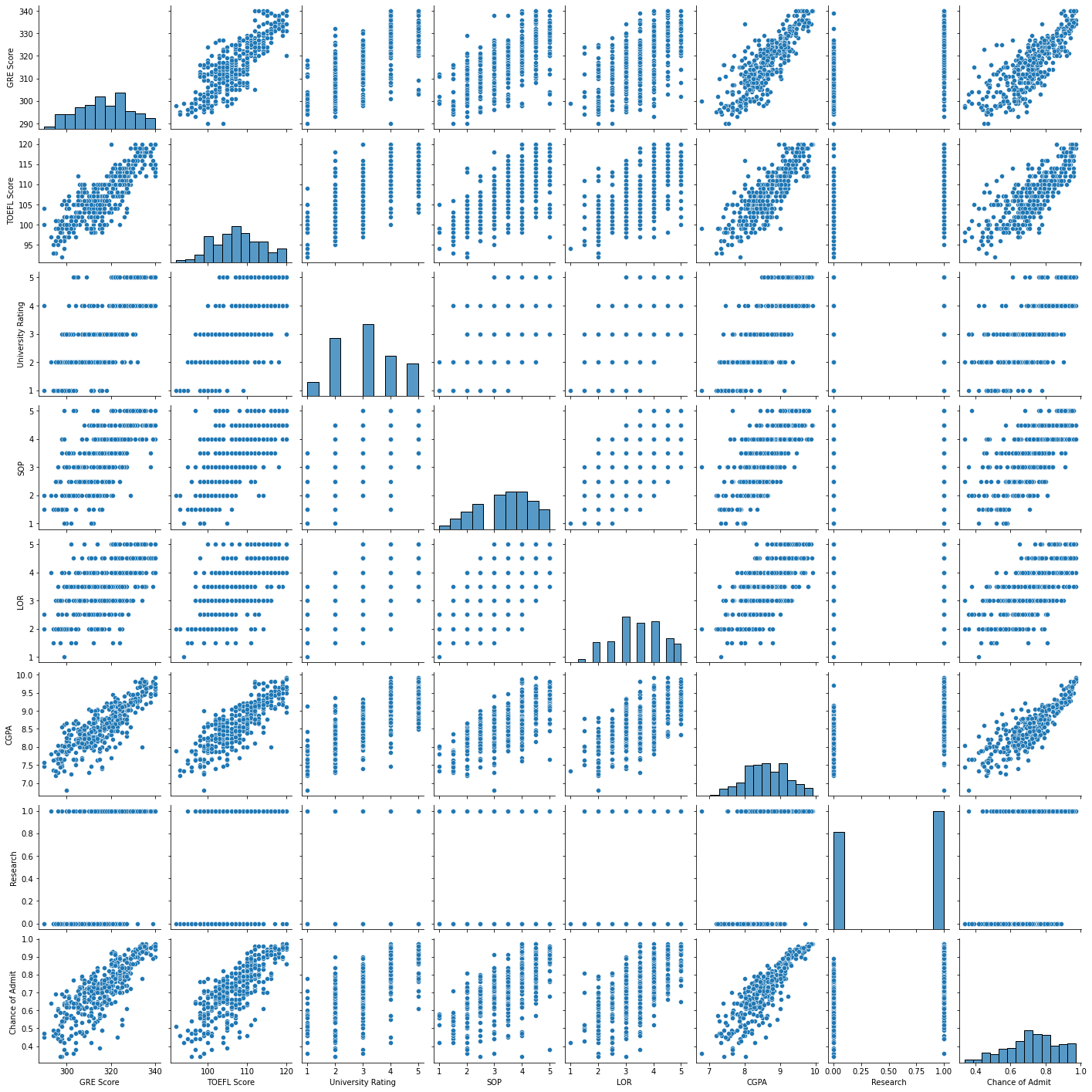
### DATA VISUALIZATION



Heat Map:

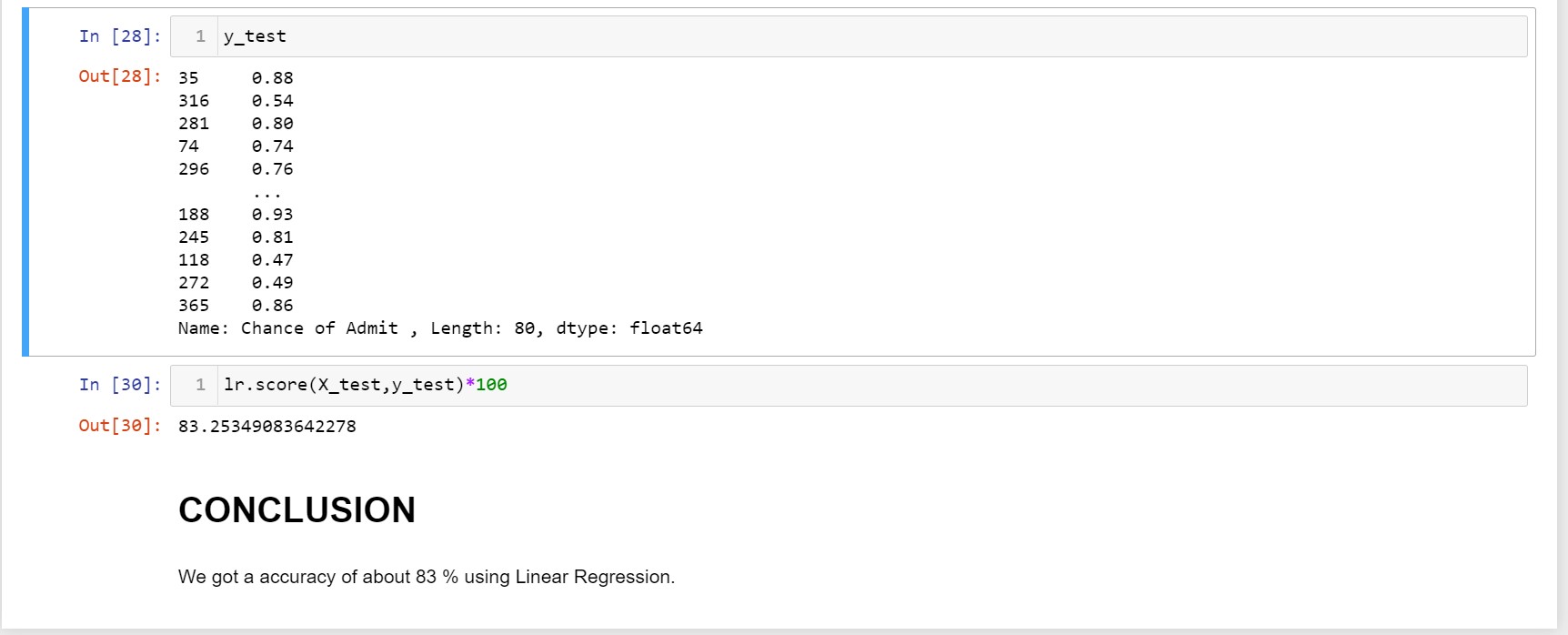


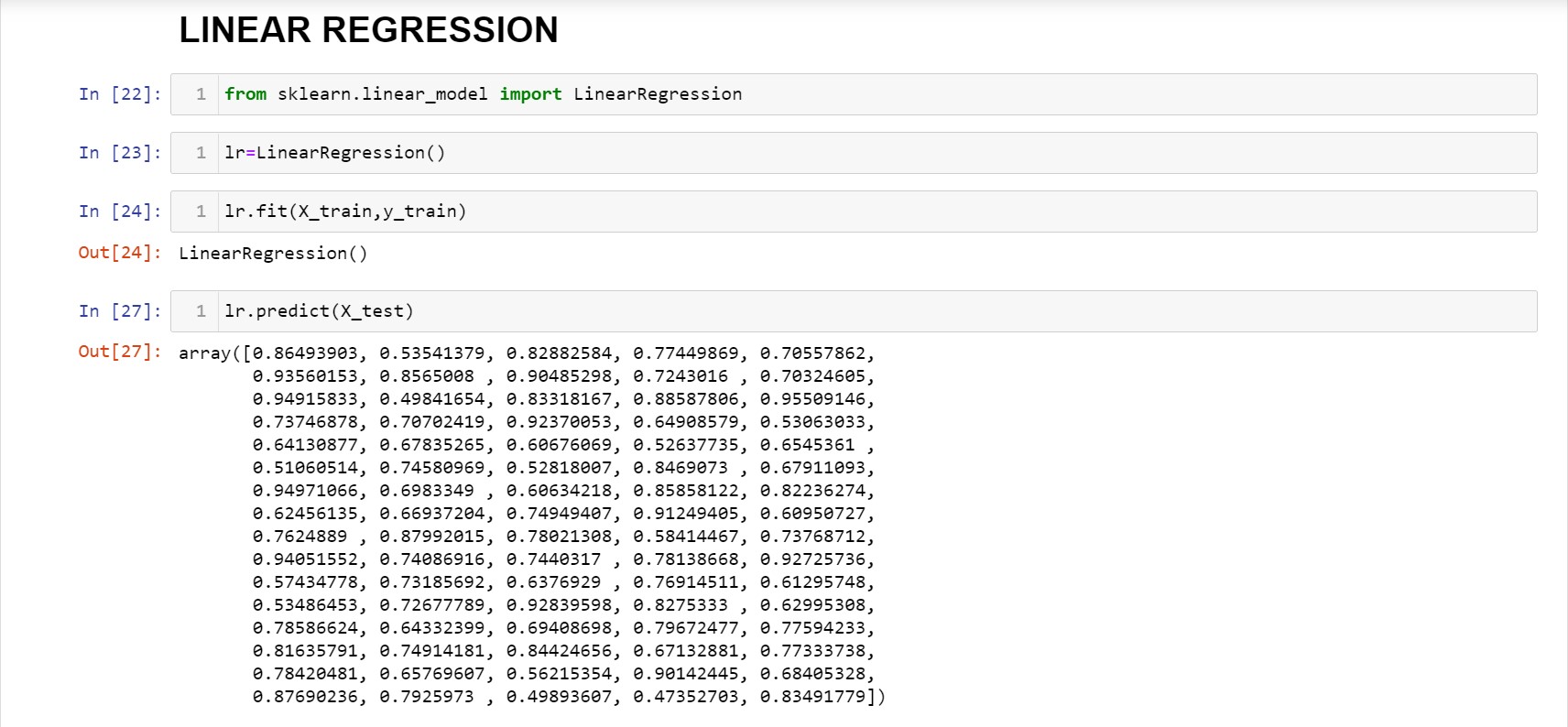
Pari Plot



TRAINING AND TESTING DATA



MODEL BUILDING



### **CONCLUSION**

My project mainly includes three parts: data preprocessing, base model selection, modeling and stacking. Because the original data is completed and clean without too many features, most energy I spent on my project is the base model selection and stacking. Selecting base model is important so as to I save more energy in result analysis and have base line to compare with following models.

The Analysis done on the data helps both student and college to choose according to the grades.

Without proper analysis student will not be able to choose college, which affect the decision making to which they apply. That’s where Data Analysis help them to select

college according to their marks.

#### REFERENCES

* [https://machinelearningmastery.com/linear-regression-f](https://machinelearningmastery.com/linear-regression-for-machine-learning) [or-machine-learning](https://machinelearningmastery.com/linear-regression-for-machine-learning)
* https://www.kaggle.com/mohansacharya/graduate-

admissions