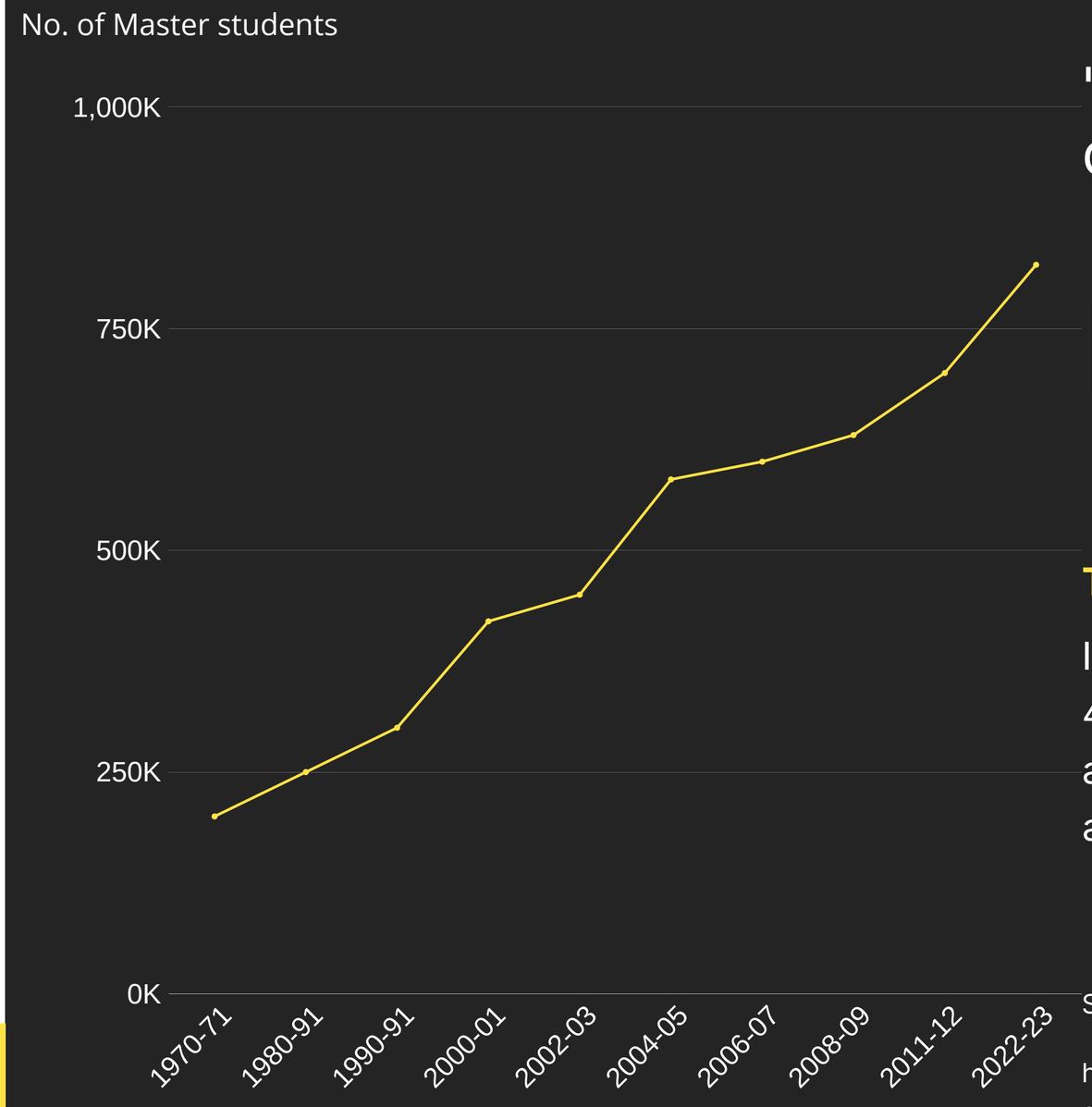
DATA SCIENCE CAPSTONE PROJECT

Analysis of US Graduate Schools Admission + Predictive Modelling

by Kam Wing Sze, Cathy

MAY 30, 2020



"Are undergraduates pursuing further down the road?"

Background

THE GROWING TREND OF MASTER'S DEGREES

In the academic year of 2022-23 it is expected that 493,000 female and 330,000 male students will earn a Master’s degree in the United States. These figures are a significant increase from 1970s.

Source: Erin Duffin (2020)

https://www.statista.com/statistics/185160/number-of-masters-degrees-by-gender-since-1950/

Why all the hype?

'Simple - money, interest, prospect'

ACCORDING TO NYU...

U.S. workers with a master's degree or higher earn an **average annual salary of $55,242**, versus those with a bachelor's degree whose average annual salary is $42,877, according to the United States Census Bureau. That represents nearly a 30% difference in average annual salary—and offers clear evidence that completing a graduate degree can make a positive impact on one's financial situation.

Source: https://gsas.nyu.edu/content/nyu-as/gsas/programs/masters-programs/prospective-students/why-pursue-a-master-s-degree.html

Other factors include...

- To study a **passionate field** and to **explore future employment** in a related area
- To gain **recognition** and credibility
- To **acquire skills** in new technologies and methods that have developed in their fields

Problem Statement

As an undergraduate student myself, I want to find out ...

THIS PROJECT AIMS TO FIND OUT:

1. What contributes to a successful Master admission in the States
2. What the overall distribution of the applicants' academic strength is
3. How likely can one's academic score be sufficient to get into a master programme

WHERE AND WHAT IS THE DATASET?

US GRADUATE SCHOOLS ADMISSION FROM KAGGLE

https://www.kaggle.com/mohansacharya/graduate-admissions

DATA CREDIT

Mohan S Acharya, Asfia Armaan, Aneeta S Antony :

A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

What tools do I used to process and present the data?

1. Dataframe + Organization

2. Visualization

3. Codes running and open sourcing

INTRODUCTION -

DESCRIPTION OF THE DATASET

Target Column:

1. **Chance of Admit (ranging from 0 to 1)**

LET'S TAKE A QUICK GLIMPSE OF OUR DATASET...

The dataset contains several parameters which are considered important for the Masters Programs application. The feature variables and target column are listed at the side:

Feature Variables:

1. **Serial No. (out of 400) *will be omitted**
2. **GRE Scores (out of 340)**
3. **TOEFL Scores (out of 120)**
4. **University Rating (out of 5)**
5. **Statement of Purpose (out of 5)**
6. **Letter of Recommendation (out of 5)**
7. **Cumulative GPA (out of 10)**
8. **Research Experience (either 0 or 1)**

There is a total of...

400

CANDIDATES' DATA

alongside with 8 feature columns and 1 target column

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65
...	...	...	...	...	...	...	...	...	...
395	396	324	110	3	3.5	3.5	9.04	1	0.82
396	397	325	107	3	3.0	3.5	9.11	1	0.84
397	398	330	116	4	5.0	4.5	9.45	1	0.91
398	399	312	103	3	3.5	4.0	8.78	0	0.67
399	400	333	117	4	5.0	4.0	9.66	1	0.95

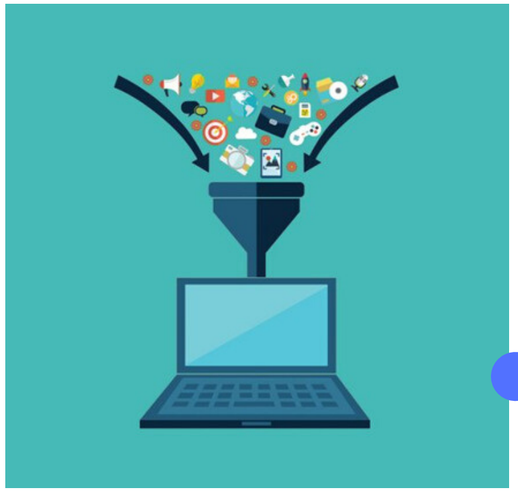
400 rows × 9 columns

LIMITATION ALERT!

Precaution: There are some areas you may want to watch out before examining this project

- Every program and university is **subjected to** its own **unique evaluation criteria** and varied across a wide spectrum of disciplines
- **Subjective qualitative skills** are not included within the dataset (e.g. musical, sports and artistic talent)
- Soft skills - such as **personalities** and performance of admission **interviews** - were not taken into account
- Most of the **data are quantitative** which can potentially be an oversight to candidate's inner qualities
- The name of the **universities are not specified**

3-Step Flow of My Analysis

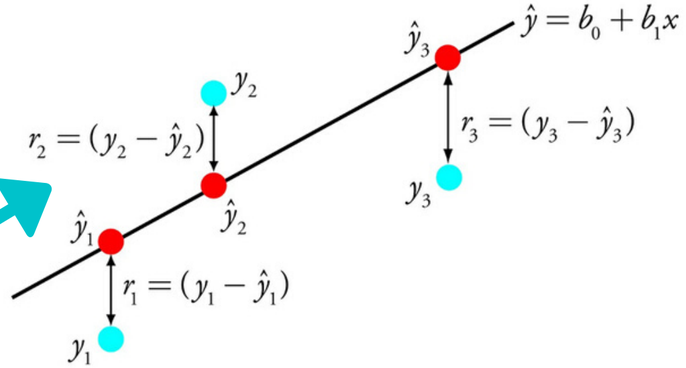


Data Mining + Preprocessing

- To examine whether the dataset is **complete** without null values
- **Remove irrelevant columns** which could possibly affect the accuracy of the prediction

Data Exploration + Visualization

- To draw **insights** from the dataset
- To find out the **correlation** between different academic achievements and chance of admission



Predictive Modelling + Evaluation

- To use **linear regression** and **XGBoost** as predictive models to predict successful admission rate
- A pseudo dummy student's profile will be used to evaluate the chance of admit

Let's start with

Data Mining + Preprocessing

Data Mining -> Import libraries + read the data

```
import pandas as pd
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
matplotlib.rcParams['figure.figsize'] = (20, 10)
import numpy as np
```

```
df = pd.read_csv("US_graduate_schools_admission_parameters_dataset.csv")
df.describe()
df
```

Data mining + Preprocessing → Data Exploration + Visualization → Predictive Modelling + Evaluation

Data Preprocessing - Cleaning

Delete the first column (Serial No.) since it is irrelevant to our target column, and there is already a pre-assigned ID no. for each entry by default. This dataset is good because there are no missing values.

```
new_df = df.drop(columns=['Serial No.'], axis =1)
new_df.head()
new_df["GRE Score"] = new_df["GRE Score"].astype(int)
new_df["Chance of Admit "] = new_df["Chance of Admit "].astype(float)
```

```
missing_values = new_df.isnull().sum()
missing_values
```

```
GRE Score            0
TOEFL Score          0
University Rating    0
SOP                  0
LOR                  0
CGPA                 0
Research             0
Chance of Admit      0
dtype: int64
```

```
new_df
```

Data Preprocessing - Cleaning

Delete the first column (Serial No.) since it is irrelevant to our target column, and there is already a pre-assigned row no. for each entry by default.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	
4	5	314	103	2	2.0	3.0	8.21	0	0.65
...	...	...	...	...	...	...	...	...	
395	396	324	110	3	3.5	3.5	9.04	1	0.8
396	397	32	107	3	3.0	3.5	9.11	1	0.84
397	398	330	116	4	5.0	4.5	9.45	1	0.91
398	399	312		3	3.5	4.0	8.78	0	0.67
399	400	333		4	5.0	4.0	9.66	1	0.95

	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admi
0	337	118	4	4.5	4.5	9.65	1	0.92
1	324	107	4	4.0	4.5	8.87	1	0.76
2	316	104	3	3.0	3.5	8.00	1	0.72
3	322	110	3	3.5	2.5	8.67	1	0.80
4	314	103	2	2.0	3.0	8.21	0	0.65
...	...	...	...	...	...	...	...	
395	324	110	3	3.5	3.5	9.04	1	0.82
396	325	107	3	3.0	3.5	9.11	1	0.84
397	330	116	4	5.0	4.5	9.45	1	0.91
398	312	103	3	3.5	4.0	8.	0	0.67
399	333	117	4	5.0	4.0	9.66		0.95

Data mining + Preprocessing → Data Exploration + Visualization → Predictive Modelling + Evaluation

Let's move on to...

Data Exploration + Visualization

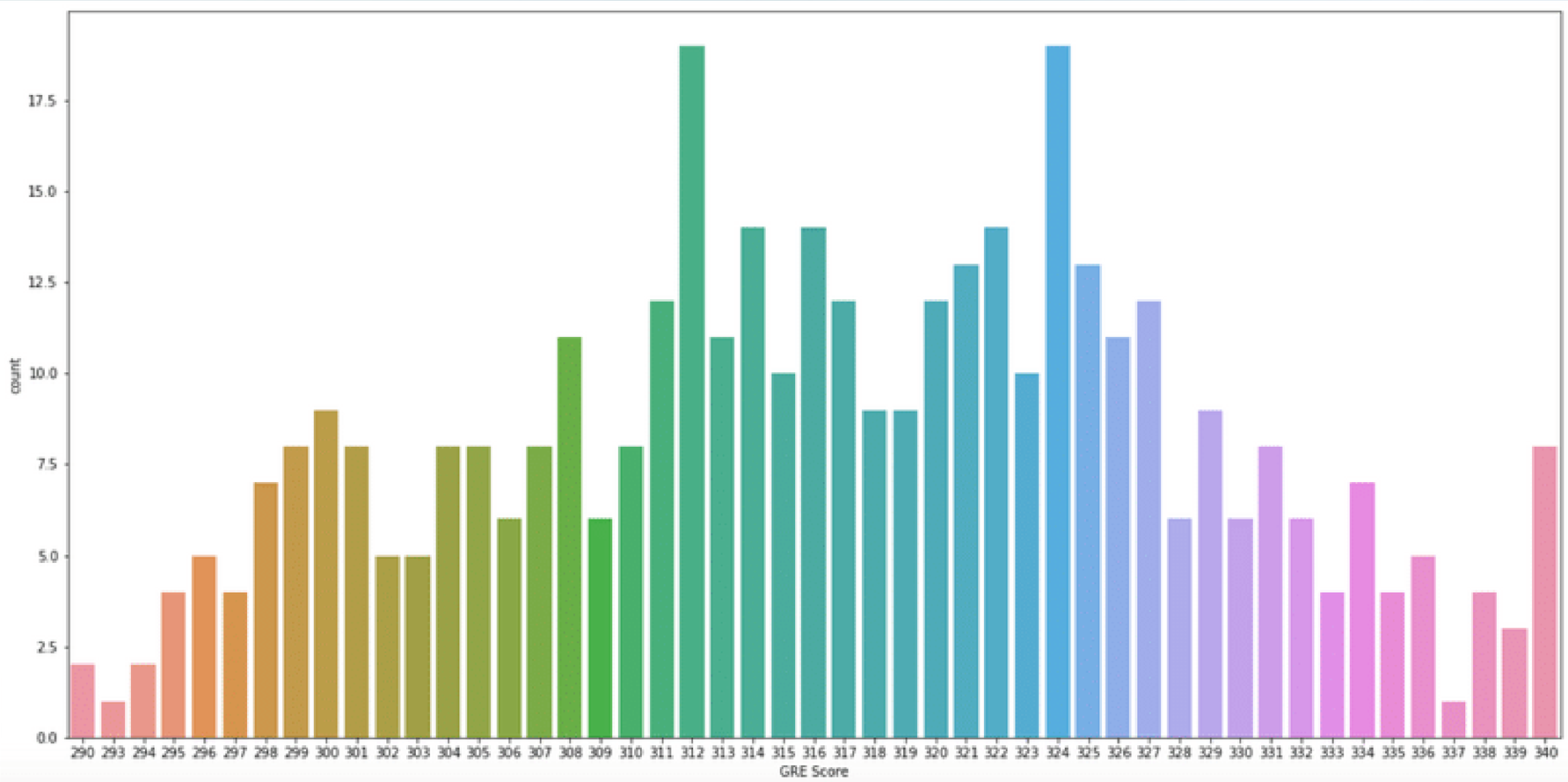
02

Data Visualization - Research

The number of candidates who had done academic research before - either in '1' (YES) or '0' (NO)
It turns out the majority has done a research paper before 57.5%, while 42.5% hasn't.

Data Visualization - GRE Score

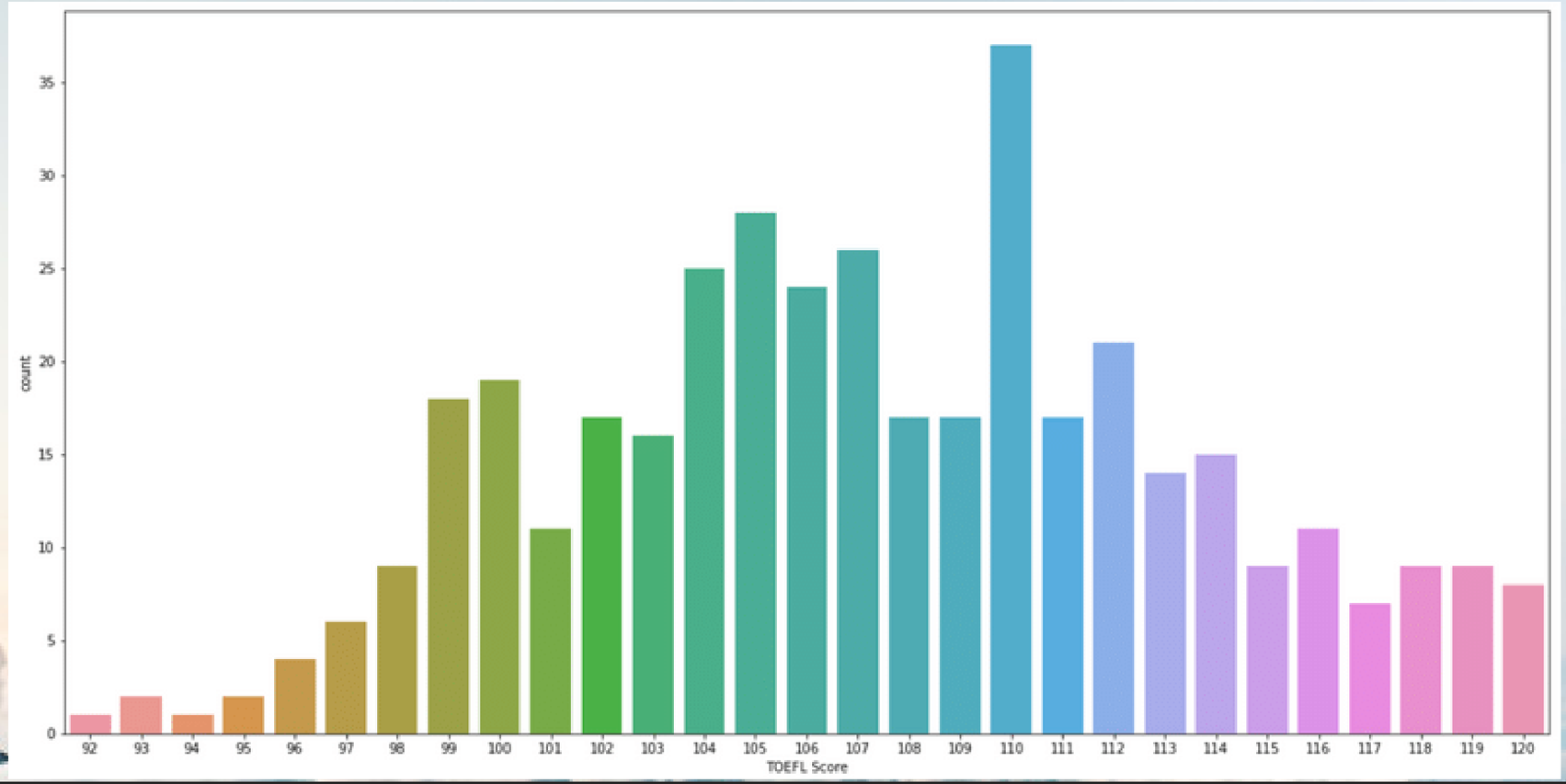
A majority of candidates scored between 310 to 324, which renders the overall distribution a bell curve.



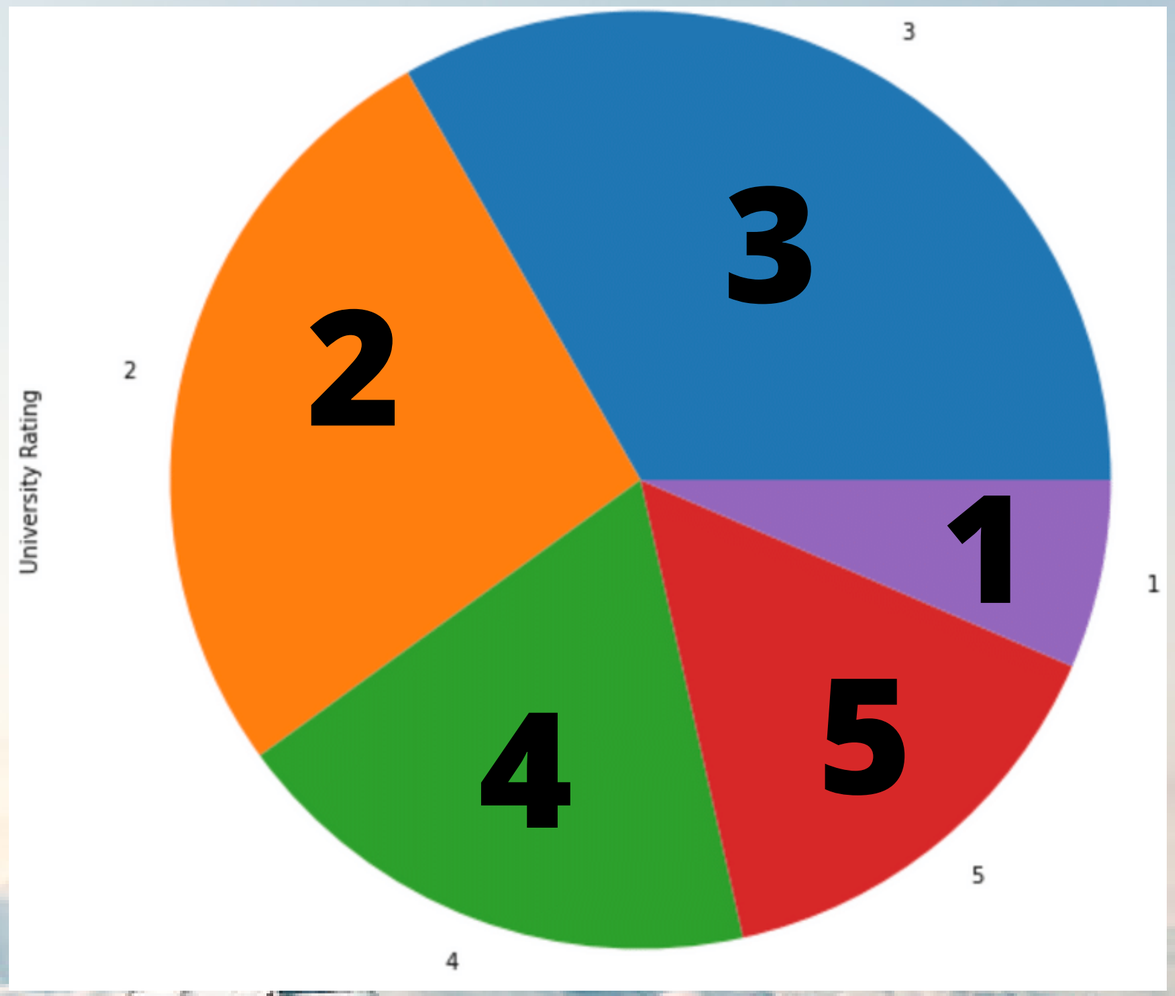
Data mining + Preprocessing → **Data Exploration + Visualization** → Predictive Modelling + Evaluation

Data Visualization - TOEFL Score

Similar to the GRE Score, it resembles a bell curve where most candidates are situating across the range of 100-110. 110 is the most frequent score candidates got.



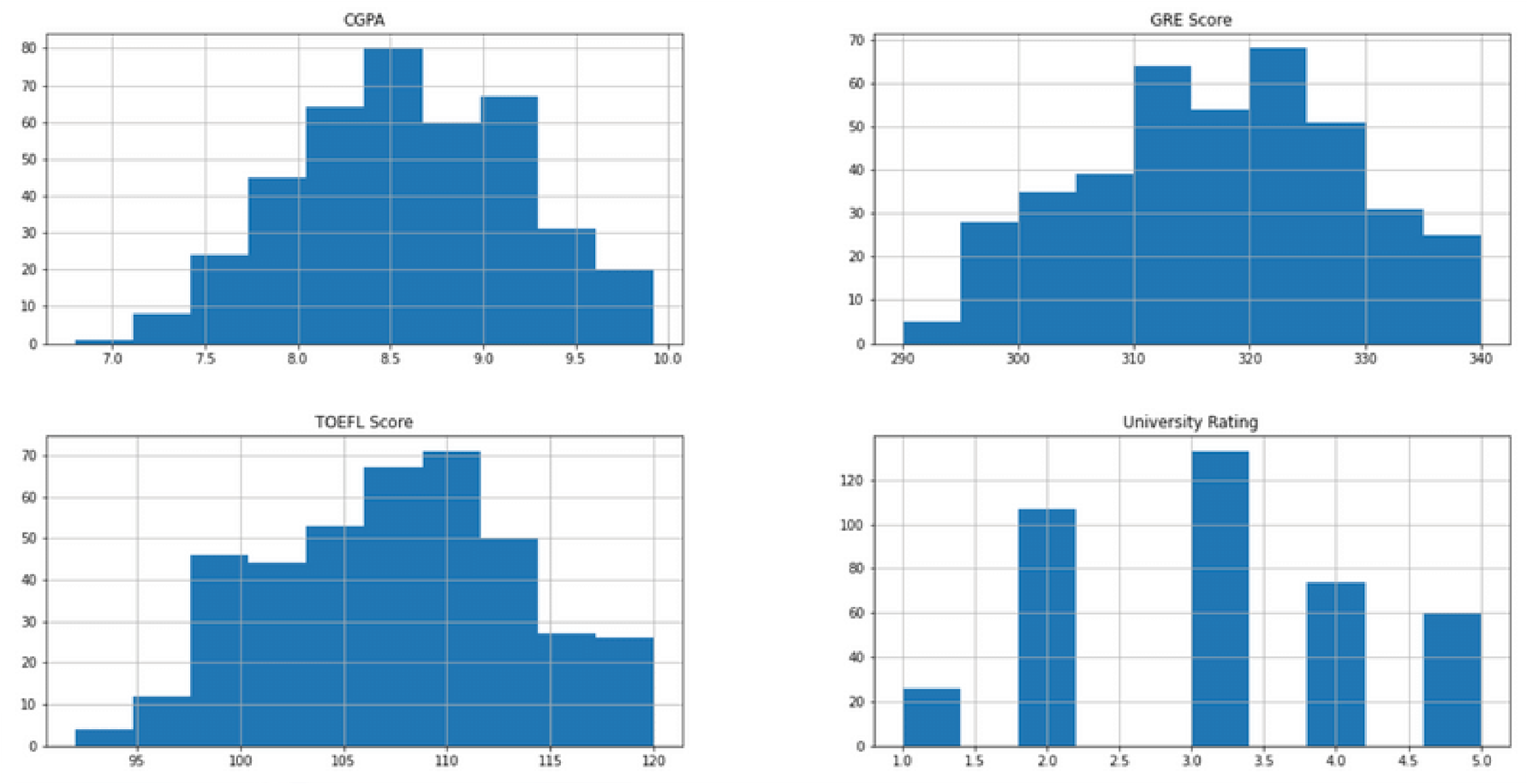
Data Visualization - University Rating



Surprisingly, most of the undergraduates were not coming from 'Ivy Leagues'. A large proportion of them come from 3rd and 4th class university accordingly.

*Remember **5** is the highest score to represent 1st class university.

Data Visualization- Detailed Data Breakdown



Histograms for 4 main academic achievement to show the overall distribution of the applicants' academic strength. All of them resembled a bell shape.

Data Visualization - Bivariate Analysis

The **correlation of different academic achievements and Chance of admit** is shown by heat map.

CGPA crowns the most influential factor which could make a significant impact on admission.

*The lighter the color is, the more related two factors are.

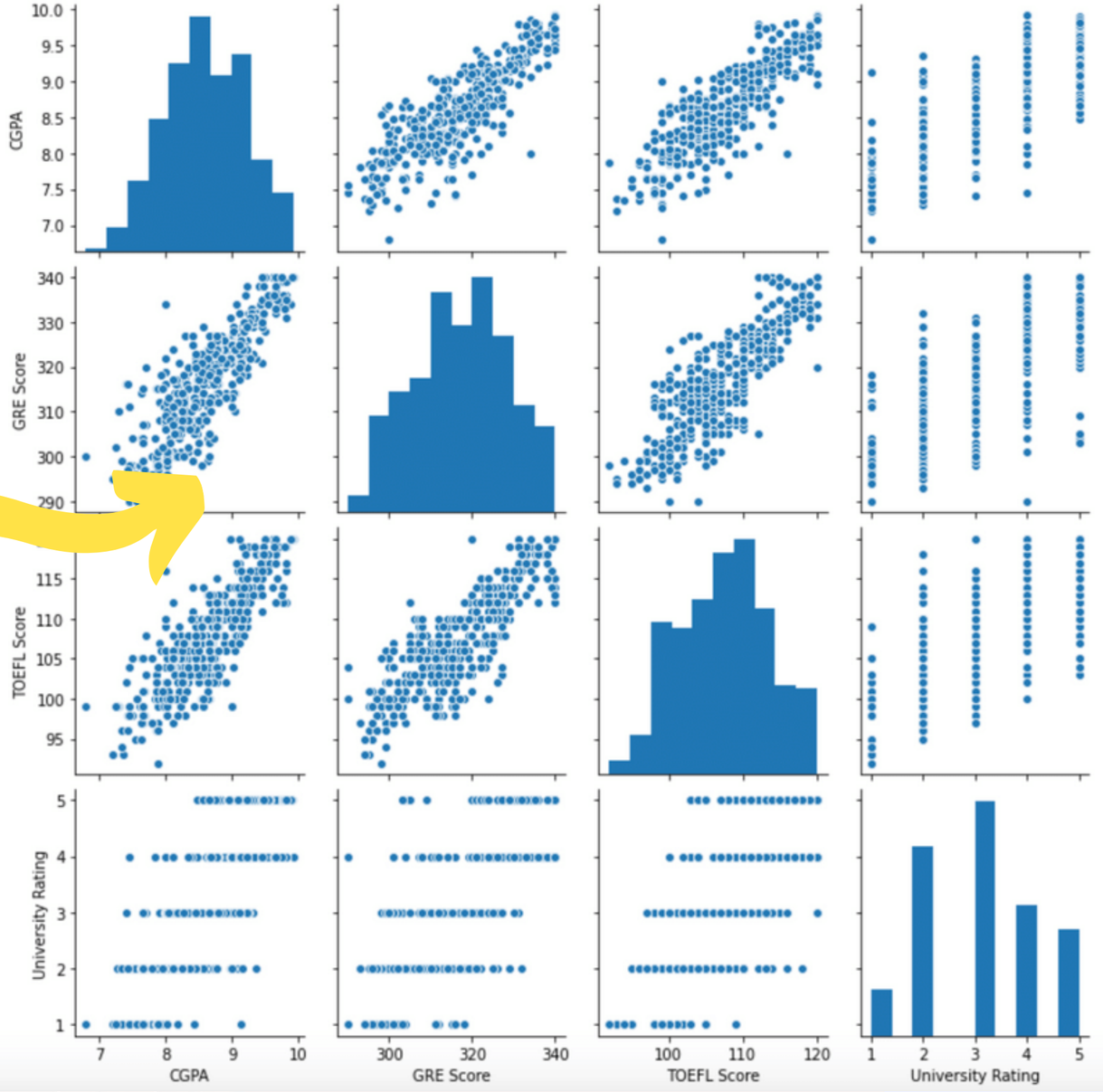
Data mining + Preprocessing → **Data Exploration + Visualization** → Predictive Modelling + Evaluation

Data Visualization - Bivariate Analysis

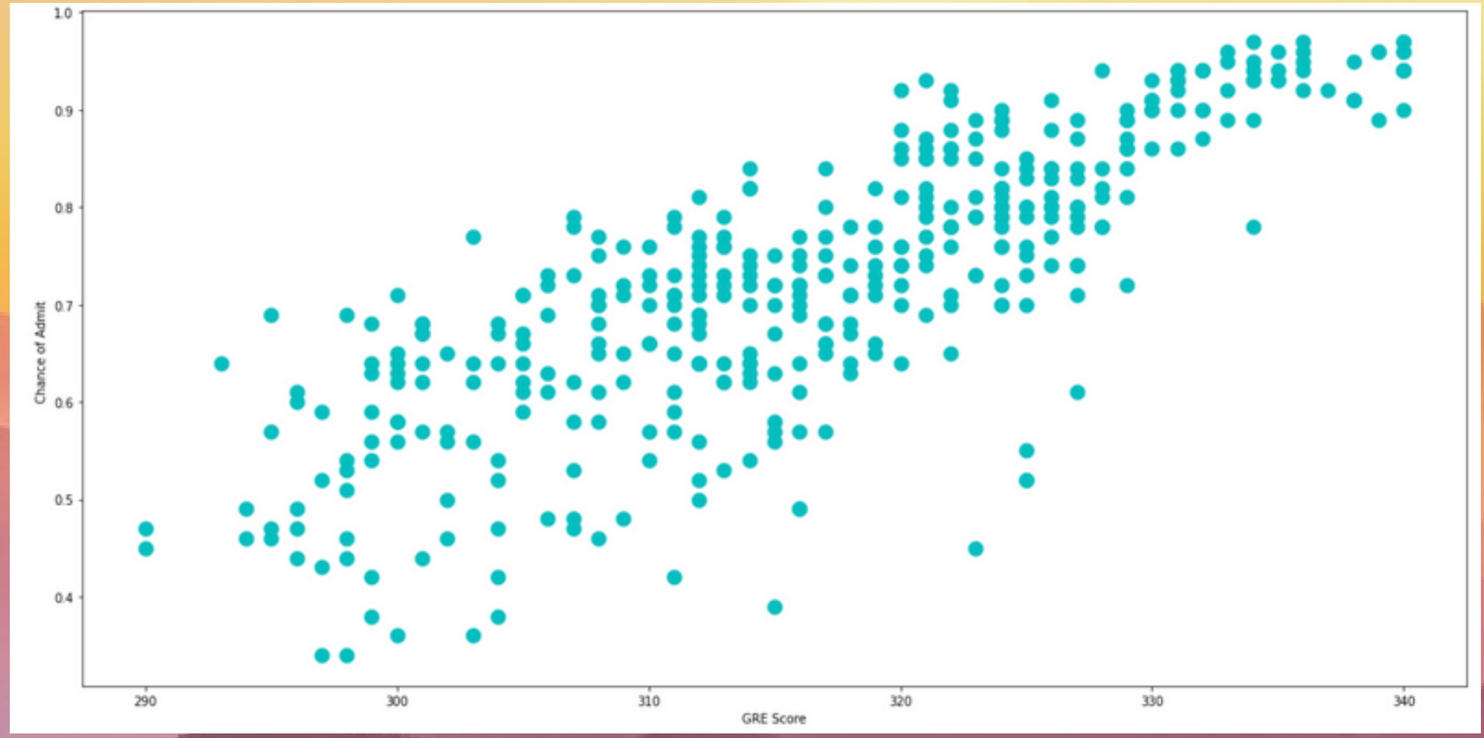
Instead of just focusing on the extent of correlation with the chance of admit.

Quantifiable comparison is also drawn between different academic achievements by pair scatter plots and bar charts.

All of them **shared positive correlation.** For instance, the higher marks you got in CGPA, the higher mark that people tend to get in GRE Score.

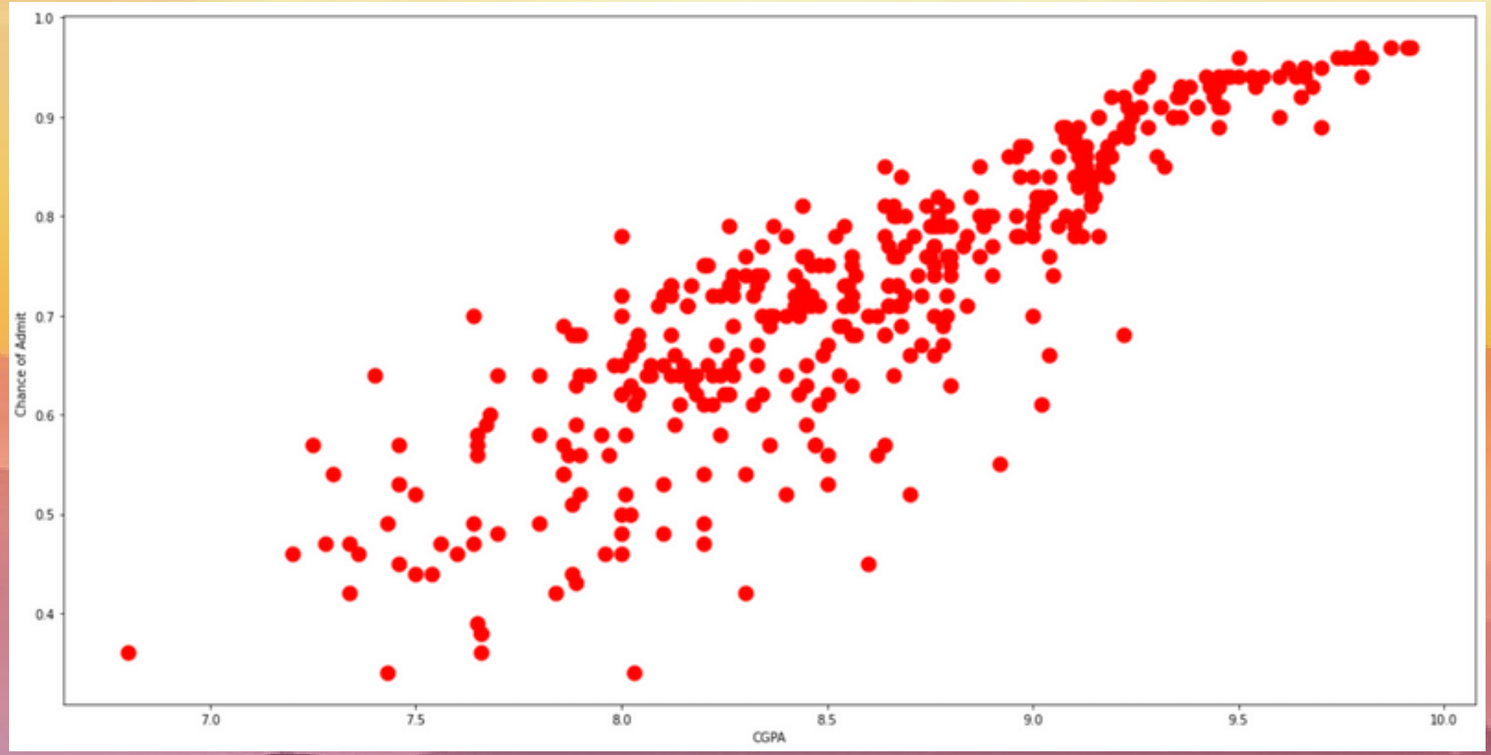


Data Visualization - Bivariate Analysis (GRE)



Data mining + Preprocessing → **Data Exploration + Visualization** → Predictive Modelling + Evaluation

Data Visualization - Bivariate Analysis (CGPA)

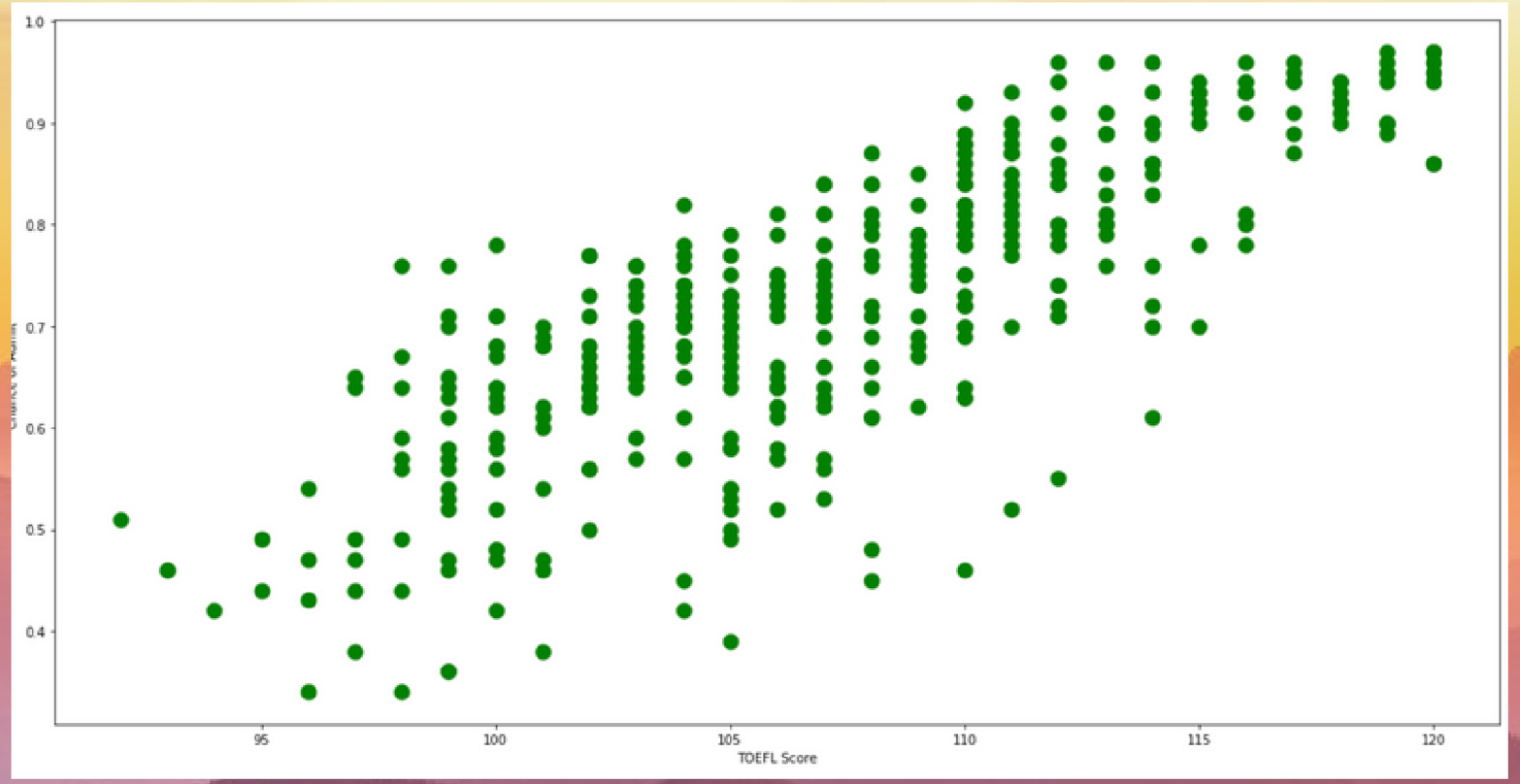


Data mining + Preprocessing

Data Exploration + Visualization

Predictive Modelling + Evaluation

Data Visualization - Bivariate Analysis (TOEFL)

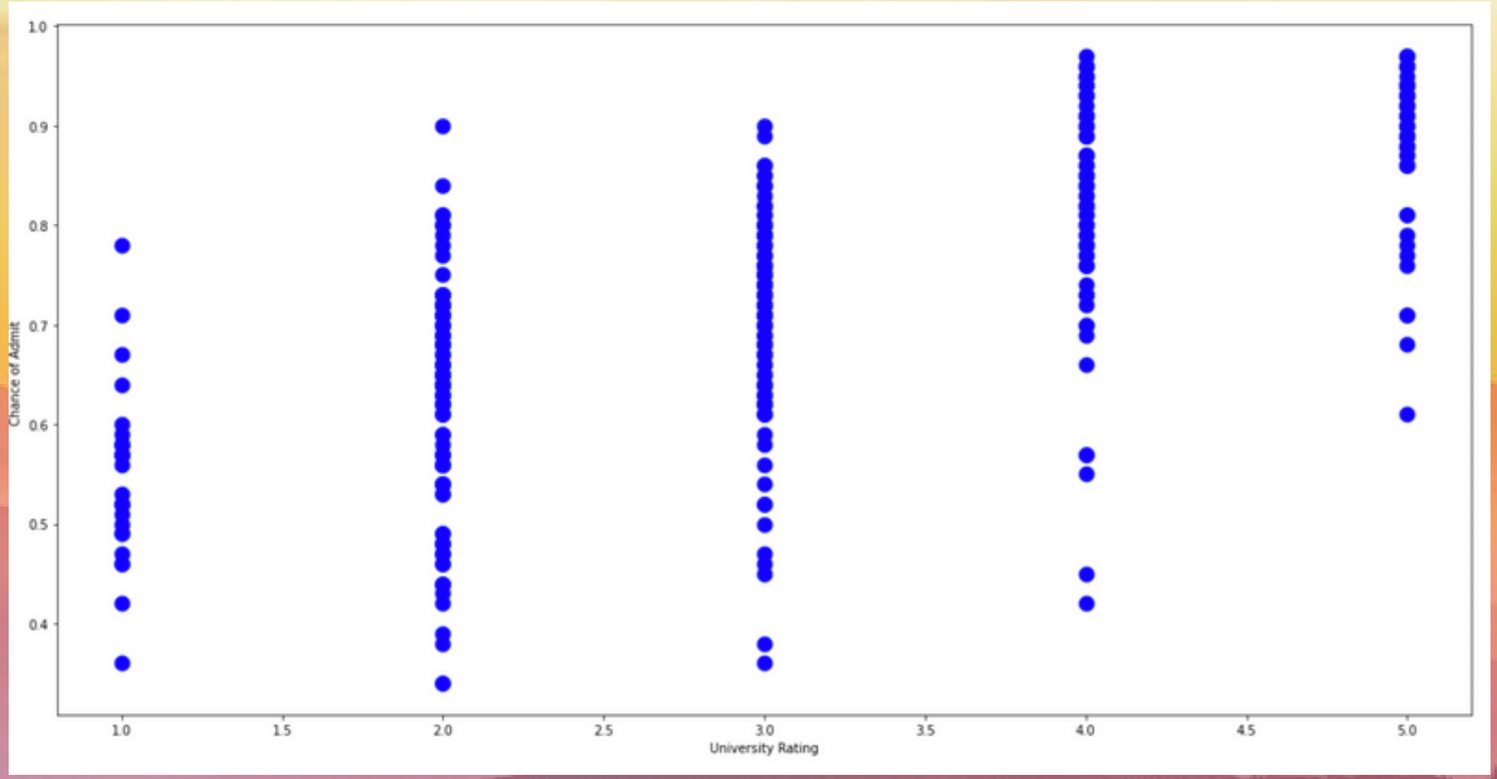


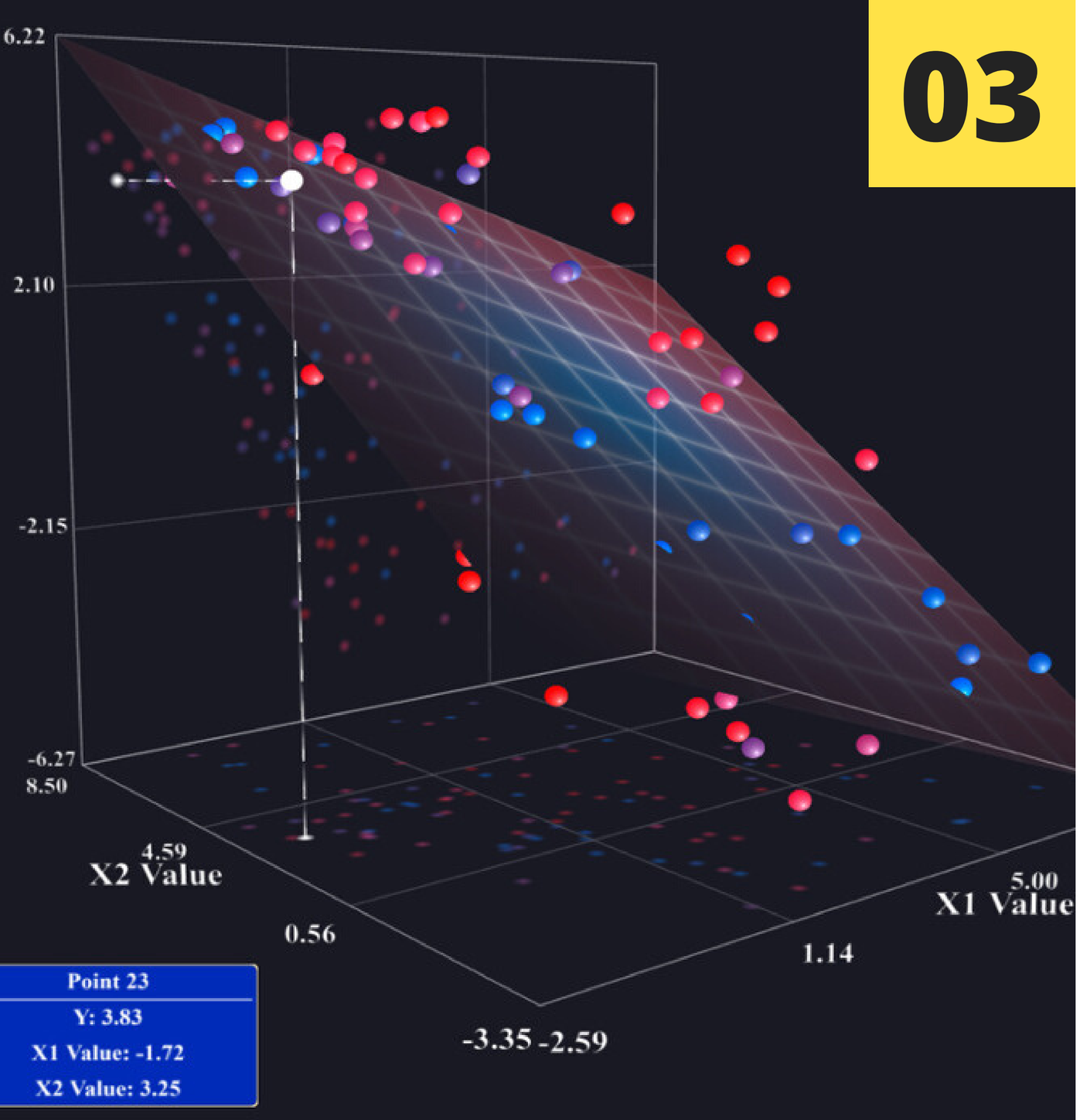
Data mining + Preprocessing

Data Exploration + Visualization

Predictive Modelling + Evaluation

Data Visualization - Bivariate Analysis (Uni Rating)





Here comes the

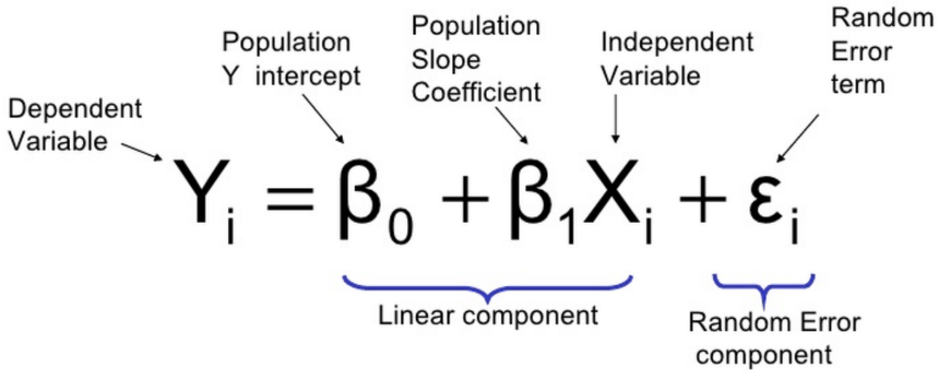
Predictive Modeling + Evaluation

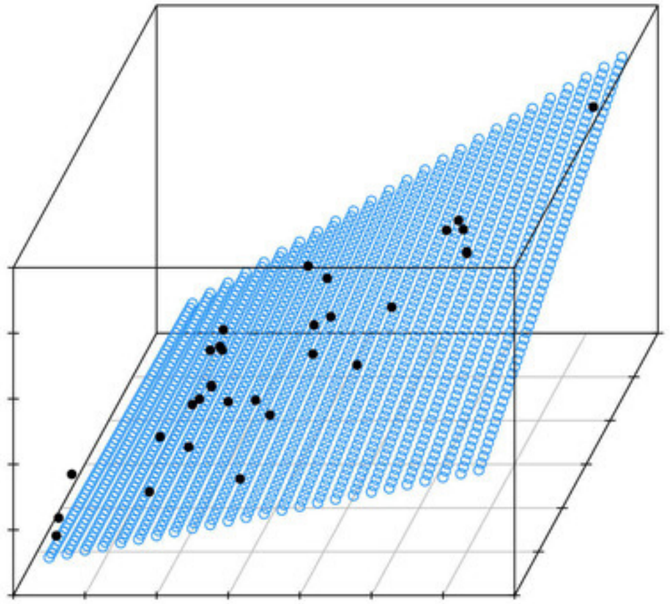
Why Linear Regression?

Linear regression is often used as **a predictive model** to express a c**ontinuous quantifiable value**. It aims to model the linear relationship in particular.

It is a statistical technique that uses several feature variables (CGPA etc.) to predict the outcome of a response variable (Chance of admit).

According to the bivariate plots and visuals, most distributions show **an upward vertical trend.** This is why linear regression is most preferable.





Predictive Modeling - Feature Scaling

```
X = new_df.iloc[:,0:7]
Y = new_df.iloc[:,7]
```

```
X
```

Select the feature columns to be scaled.

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range=(0, 1))
rescaledX = scaler.fit_transform(X)
```

```
rescaledX
```

```
array([[0.94       , 0.92857143, 0.75       , ..., 0.875      , 0.91346154,
        1.        ],
       [0.68       , 0.53571429, 0.75       , ..., 0.875      , 0.66346154,
        1.        ],
       [0.52       , 0.42857143, 0.5        , ..., 0.625      , 0.38461538,
        1.        ],
       ...,
       [0.8        , 0.85714286, 0.75       , ..., 0.875      , 0.84935897,
        1.        ],
       [0.44       , 0.39285714, 0.5        , ..., 0.75       , 0.63461538,
        0.        ],
       [0.86       , 0.89285714, 0.75       , ..., 0.75       , 0.91666667,
        1.        ]])
```

MinMaxScaler function is called to scale large quantitative difference between various feature columns.

The result of scaling is listed on the right.

Predictive Modeling - Spliting into training and testing data

```
from sklearn.model_selection import train_test_split
from sklearn import linear_model
from sklearn.metrics import mean_squared_error, r2_score
import math
```

```
X_train, X_test, y_train, y_test = train_test_split(rescaledX,Y,test_size=0.2,random_state=42 )
```

Predictive Modeling - Employ Linear Regression

Create a linear regression object, train the model using the training sets and make predictions using the testing set.

```
regr = linear_model.LinearRegression()

regr.fit(X_train, y_train)

y_pred = regr.predict(X_test)
```

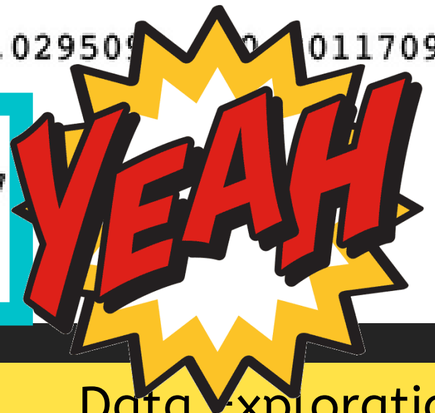
Predictive Modeling - Assess the Accuracy (LR)

Examine the coefficients, mean squared error and variance score based on the linear regression model. It turns out to be extremely accurate!

```
print('Coefficients: \n', regr.coef_)

print(f"Mean squared error:{mean_squared_error(y_test, y_pred): .2f}")

print(f"Root Mean squared error: {math.sqrt(mean_squared_error(y_test, y_pred)) :.2f}")

print(f'Variance score: {r2_score(y_test, y_pred):.2f}')
```

```
Coefficients:
 [ 0.09312548  0.07626324  0.02950[illegible]0117097  0.06308097  0.35776778
```

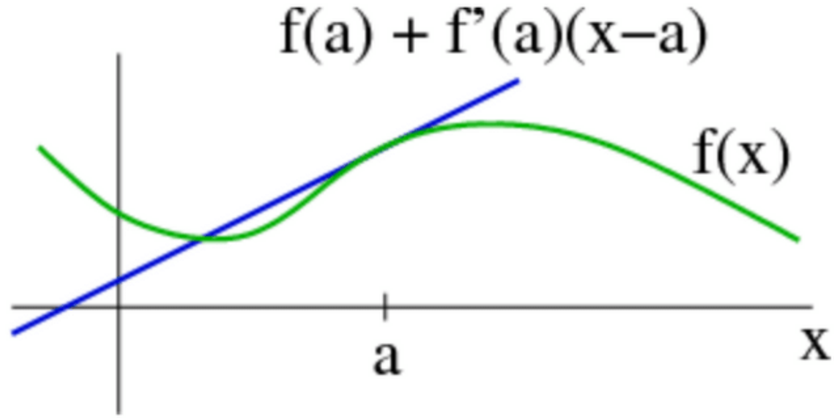
```
Mean squared error: 0.00
Root Mean squared error: 0.07
Variance score: 0.82
```



Data mining + Preprocessing → Data Exploration + Visualization → **Predictive Modelling + Evaluation**

Predictive Modeling - How about XGBoost?

XGBoost is a supervised learning method that is based on function approximation by optimizing specific loss functions as well as applying several regularization techniques. It is commonly regarded as a good model to predict target column accurately.



Taylor linear approximation of a function around a point a

Real value (label) known from the training data-set

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

Can be seen as $f(x + \Delta x)$ where $x = \hat{y_i}^{(t-1)}$

XGBoost objective function analysis

Predictive Modeling - How about XGBoost?

Let us find out if that is truly the case.

```
import xgboost as xgb
```

```
regr_xgb = xgb.XGBRegressor()
```

```
regr_xgb = xgb.XGBRegressor()

regr_xgb.fit(X_train, y_train)

y_pred = regr_xgb.predict(X_test)
```

Predictive Modeling - Assess the Accuracy (XGB)

It turns out XGBoost is performing less accurate than Linear Regression by a very small range of error. It is estimated that this is due to the linear characteristic of the dataset.

```
print('Coefficients: \n', regr.coef_)

print(f"Mean squared error:{mean_squared_error(y_test, y_pred): .2f}")

print(f"Root Mean squared error: {math.sqrt(mean_squared_error(y_test, y_pred)) :.2f}")

print(f'Variance score: {r2_score(y_test, y_pred):.2f}')
```

```
Coefficients:
 [ 0.09312548  0.07626324  0.0295098  -0.00117097  0.06308097  0.35776778
  0.02222705]
Mean squared error: 0.01
Root Mean squared error: 0.08
Variance score: 0.78
```

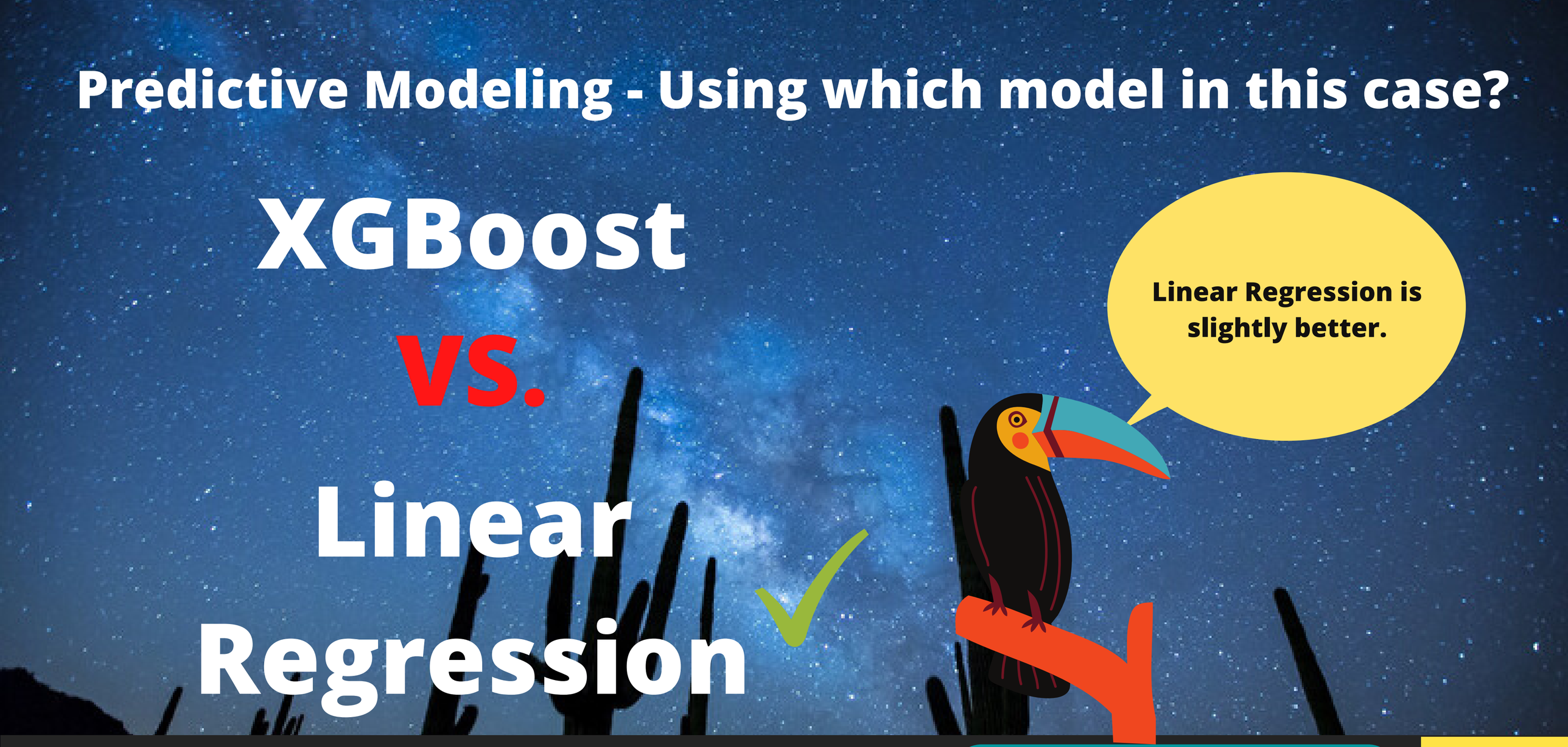
Data mining + Preprocessing → Data Exploration + Visualization → **Predictive Modelling + Evaluation**

Predictive Modeling - Using which model in this case?

XGBoost

VS.

Linear Regression



Data mining + Preprocessing → Data Exploration + Visualization → **Predictive Modelling + Evaluation**

Let's put a dummy student to the test!

Can you get admitted into a US Master's degree successsfully?

If there is a dummy student coming in with scores below...

GRE Score: 330
TOEFL Score: 110
University Rating: 5
SOP: 4
LOR: 4
CGPA: 8
Research: 1

How likely is he/she going to be successful?

Data mining + Preprocessing → Data Exploration + Visualization → **Predictive Modelling + Evaluation**

Predictive Modeling - Dummy Student Testing

Enter the imaginary scores as in array according to the feature columns, and rescale it accordingly.

```
student = [330, 110, 5, 4, 4, 8, 1]
```

```
student = np.asarray(student)
```

```
rescaledX_student = scaler.transform(student.reshape(1,-1))
```

```
rescaledX_student
```

```
array([[0.8       , 0.64285714, 1.        , 0.75      , 0.75      ,
        0.38461538, 1.        ]])
```

Predictive Modeling - Dummy Student 's result

Here is the predicted chance of admit using linear regression.

```
regr.predict(rescaledX_student)
```

```
array([0.72237759])
```

```
student_pred = regr.predict(rescaledX_student)
```

```
if student_pred[0] > 0.72:
    print("Congratulations! You're very likely to be successful!")

else:
    print("Emm...maybe you need to work harder to secure the admission.")
```

```
Congratulations! You're very likely to be successful!
```

Predictive Modeling - Why 0.72 as the threshold?

Because 0.72 is the mean of the chance of admit.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000
mean	200.500000	316.807500	107.410000	3.087500	3.400000	3.452500	8.598925	0.547500	0.724350
std	115.614301	11.473646	6.069514	1.143728	1.006869	0.898478	0.596317	0.498362	0.142609
min	1.000000	290.000000	92.000000	1.000000	1.000000	1.000000	6.800000	0.000000	0.340000
25%	100.750000	308.000000	103.000000	2.000000	2.500000	3.000000	8.170000	0.000000	0.640000
50%	200.500000	317.000000	107.000000	3.000000	3.500000	3.500000	8.610000	1.000000	0.730000
75%	300.250000	325.000000	112.000000	4.000000	4.000000	4.000000	9.062500	1.000000	0.830000
max	400.000000	340.000000	120.000000	5.000000	5.000000	5.000000	9.920000	1.000000	0.970000

The End

Best of luck to those who are pursuing a Master's degree in U.S.!

If you have any questions, feel free to Contact me via:

Mobile Phone: **+852 5161 6382**

E-mail: **nancykam2006@hotmail.com**

Personal website: **https://nancykam2006.wixsite.com/mysite**