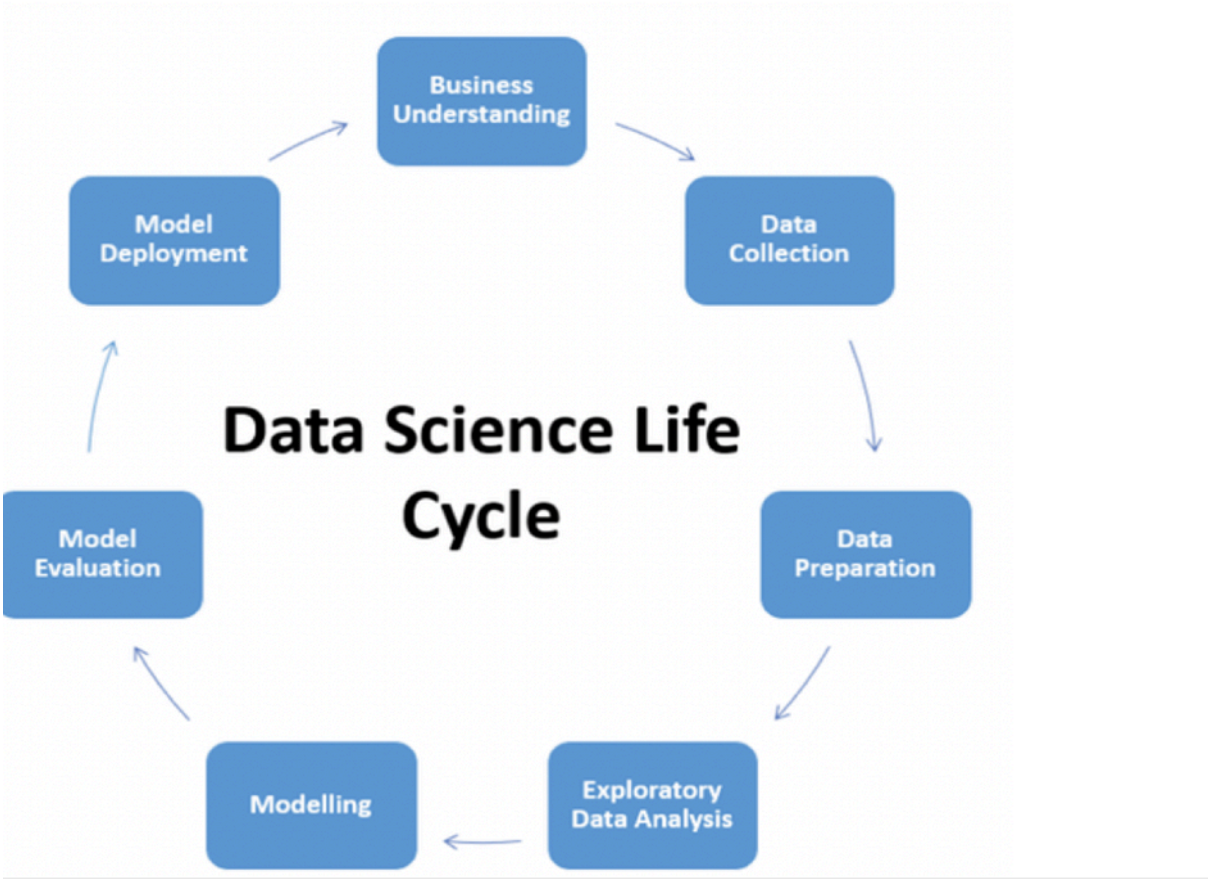
Today was my first day as an intern, and I learned many things from my colleague, Rachit Rijal. First, we discussed which project would be feasible for us. We decided to focus on data wrangling. Together, Rachit and I started searching for datasets on Kaggle and found a graduation-related dataset.

I have also gained a clear understanding of the data lifecycle.



I start to do data wrangling work.
First of all i start to find the input variable of my data:
input Variables

1. Serial No. - Unique row ID
2. GRE Scores - Out of 340
3. TOEFL Scores - Out of 120
4. University Rating - Out of 5
5. Statement of Purpose - Strength Out of 5
6. Letter of Recommendation - Strength Out of 5
7. Undergraduate GPA - Out of 10

Data collection and import data in google colab:

```
df = pd.read_csv('/content/filename.csv')
```

8. Resea rch
   Experience - Either 0 or 1

   We remove the shape and comma error from the column

```
[ ]  df.columns = df.columns.str.strip()
```

   We try to find out the missing value in our data set

```
print(df.isnull().sum())
```

```
Serial No.           0
GRE Score            0
TOEFL Score          0
University Rating    0
SOP                  0
LOR                  0
CGPA                 0
Research             0
Chance of Admit      0
dtype: int64
```

Descriptive Statistics

Descriptive statistics involve a set of summary measures that provide a snapshot of the dataset's characteristics. These measures help us understand the distribution, central tendency, and variability within the data.

- Mean: The average value of the data.
- Median: The middle value when the data is sorted.
- Mode: The most frequently occurring value.
- Range: The difference between the maximum and minimum values.
- Standard Deviation: A more interpretable measure of data spread. These statistics provide a preliminary understanding of the dataset, which is valuable for subsequent analysis and decision-making.

```
df.describe()
```

	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

Remove any null value in data set

```
df.duplicated().sum()
```

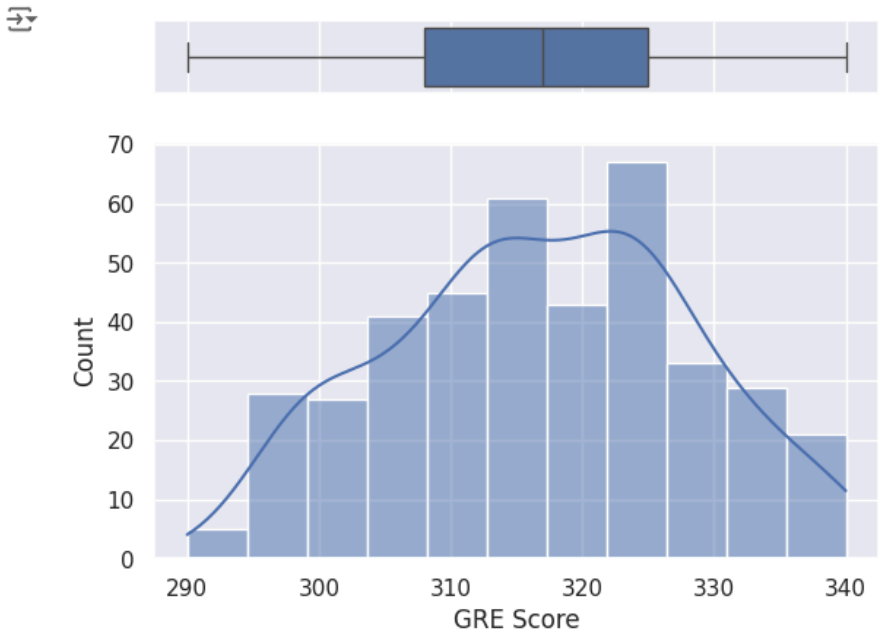
Data Wrangling

- Data Inspection
    - Checking Duplicate Entries
    - Checking Missing Values
    - Checking standard format
    - Checking data entry typos and errors
- Data Cleaning
    - Removing Duplicates
    - Handling Missing Values
    - Standardising Formats
    - Correcting Errors
- Data Transformation
    - Feature Engineering
    - Normalisation/Scaling
    - One-Hot Encoding
- Data Integration
- Data Reduction
- Data Formatting
- Data Enrichment
- Data Validation
- Documentation
- Exploratory Data Analysis (EDA)

I have learn this through this session in data wrangling process:

After that I have learn the univarient analysis in data

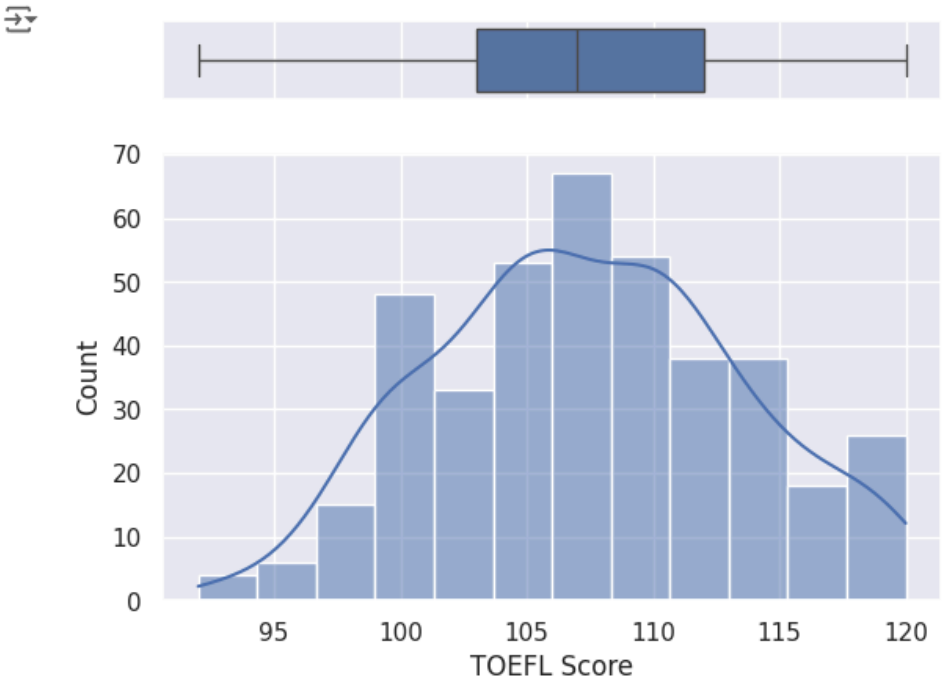
```
[ ]  sns.set(style="darkgrid")
     f, (ax_box, ax_hist) = plt.subplots(2, sharex=True, gridspec_kw={"height_ratios": (.15, .85)})
     sns.boxplot(data=df,x='GRE Score', ax=ax_box,)
     sns.histplot(data=df, x="GRE Score", ax=ax_hist,kde=True)
     ax_box.set(xlabel='')
     plt.show()
```



From the above graph, we can say -

- Average GRE score is around 318.
- 25% of the score is less than 310.
- 75% of the score is less than 325.

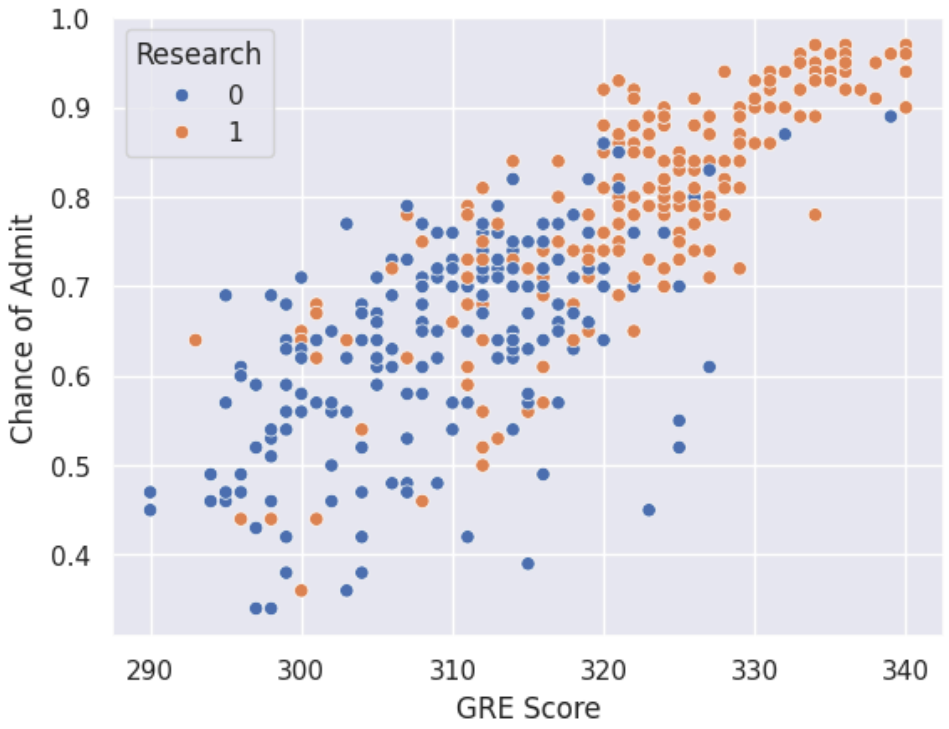
```
sns.set(style="darkgrid")
f, (ax_box, ax_hist) = plt.subplots(2, sharex=True, gridspec_kw={"height_ratios": (.15, .85)})
sns.boxplot(data=df,x='TOEFL Score', ax=ax_box)
sns.histplot(data=df, x="TOEFL Score", ax=ax_hist,kde=True)
ax_box.set(xlabel='')
plt.show()
```



From the above graph, we can say

- Average TOEFL score is around 106.
- 25% of the score is less than 104.
- 75% of the score is less than 112

```
[ ]  sns.scatterplot(x='GRE Score',y='Chance of Admit',data=df,hue='Research')
     plt.show()
```



From this scatter plot we can analysis:

From the above graph we can say that GRE score and Chance of admit has a linear relationship.

Remaining other work is done by my friend Richit Rijal.

Now we have to do many more things in this project. We will continue our project from here.