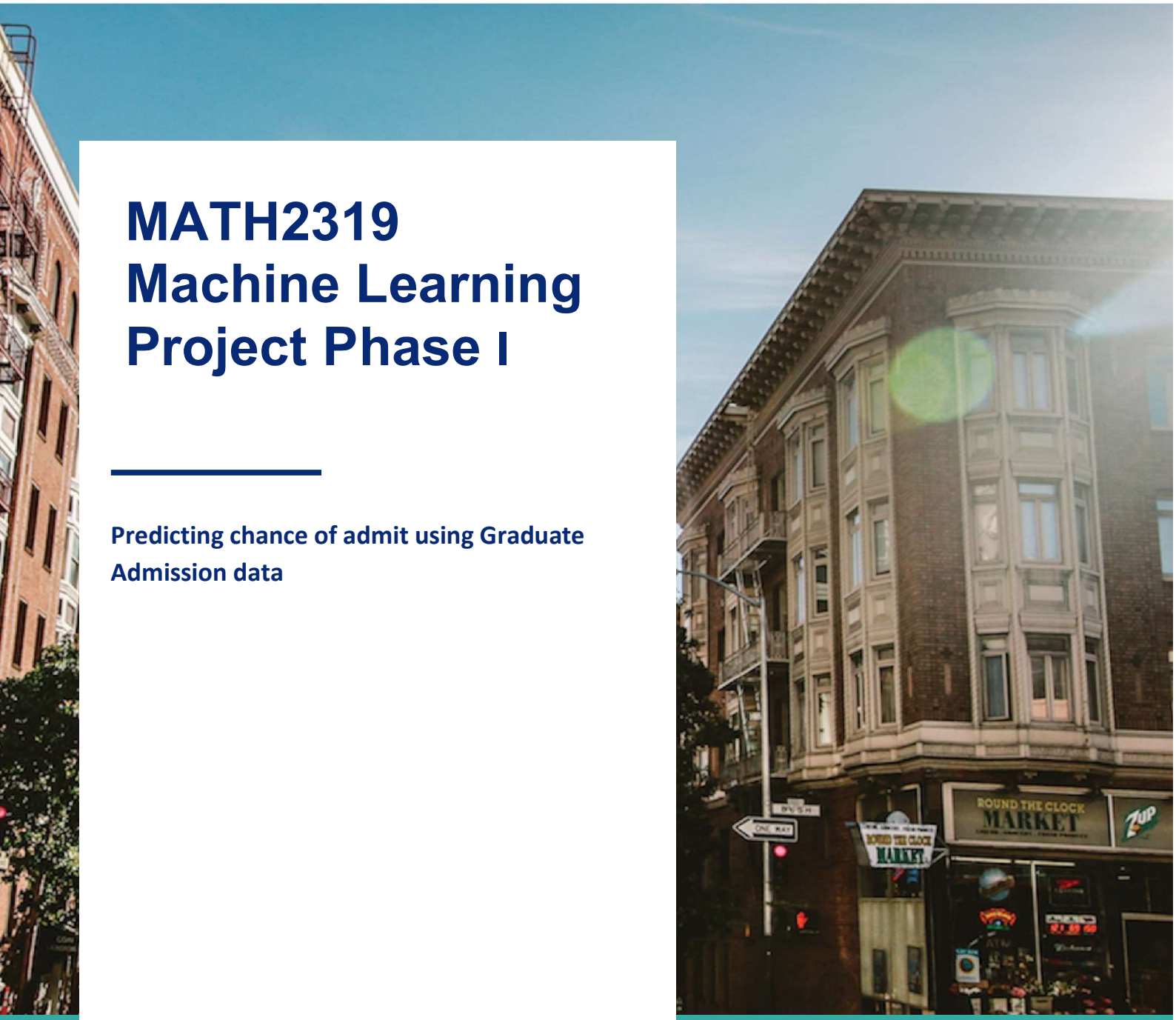
MATH2319 Machine Learning Project Phase I

Predicting chance of admit using Graduate Admission data



APRIL 28, 2019

Contents

PHASE 1

April 28, 2019

1 Introduction

1.1 Objective

The objective of this project is to predict whether a student's admission is dependant on the parameters set out in our dataset. The dataset was sourced from Kaggle at https://www.kaggle.com/mohansacharya/graduate-admissions#Admission_Predict.csv [1]. This project has two phases. Phase I focuses on data pre-processing and exploration, as covered in this report. We shall present model building in Phase II. We have created this report using jupyter notebook.

1.2 Dataset

The dataset is obtained from kaggle and contains 9 variables. All these variables fall under nominal/ordinal category. This dataset contains 500 observations. These variables measure different aspects of a student to obtain admissions for their masters degree in various universities. This data set contains 400 training observations and a 100 test observations.However, in this phase we combine these datasets into one.

1.2.1 Target Variable

In this data set we are considering the Chance of Admit as the target variable. It is an ordinal variable which states the probability of a student's admission into a unversity. The range of this variable is 0 - 1. The target variable is numeric in nature and our goal is to predict the probability of a student's admission. Hence this is a regression problem.

1.2.2 Descriptive Variables

This Data Set contains 7 descriptive features. Each feature is a metric that measures a student's profile and contributes towards their admission into the Master's program: The features are:

1.GRE Score (max value is 340) : continuous 2.TOEFL Score (max value is 120) : continuous 3.University Rating: 1,2,3,4,5 4.Statement of Purpose: 1,2,3,4,5 5.Letter of Recommendation: 1,2,3,4,5 6.Undergrad GPA (max value is 10): continuous 7.Research Experience: 1(Yes),0(No)

2 Data Pre-processing

We import pandas which will aid us to read the data from the Admission_Predict_Ver1.1.csv dataset file. We further clean the data by performing pre-processing on the dataset such as iden-

tifying missing values, check if there are any outliers, remove whitespaces (if any), look for any unwanted columns.

```
In [ ]: import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns

In [2]: Admission_file = 'Admission_Predict_Ver1.1.csv'

In [3]: import os
        os.getcwd()
        os.chdir('C:\Users\Mohammed\Desktop\Sem 2\Machine Learning\Assignment 1')

In [4]: Admissions = pd.read_csv(Admission_file,decimal='.')

In [5]: Admissions.head()
```

Out[5]:

	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

2.1 Stripping WhiteSpaces

Using the str.strip() function, we strip the whitespaces from the column names of the data set.

```
In [6]: Admissions.columns=Admissions.columns.str.strip()

In [7]: Admissions.columns.values

Out[7]: array(['Serial No.', 'GRE Score', 'TOEFL Score', 'University Rating',
               'SOP', 'LOR', 'CGPA', 'Research', 'Chance of Admit'], dtype=object)
```

2.2 Validating DataTypes of Variables

Using the dtypes function, we cross verify the data types of all the variables. This dataset maintains the appropriate variable data types hence no further action is necessary.

```
In [8]: Admissions.dtypes
```

```
Out[8]: Serial No.             int64
        GRE Score              int64
        TOEFL Score            int64
        University Rating      int64
        SOP                  float64
        LOR                  float64
        CGPA                 float64
        Research               int64
        Chance of Admit      float64
        dtype: object
```

2.3 Checking for Missing Values

This data set, as shown below does not contain any missing values.

```
In [9]: Admissions.isnull().sum()
```

```
Out[9]: 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
```

2.4 Dropping unwanted columns

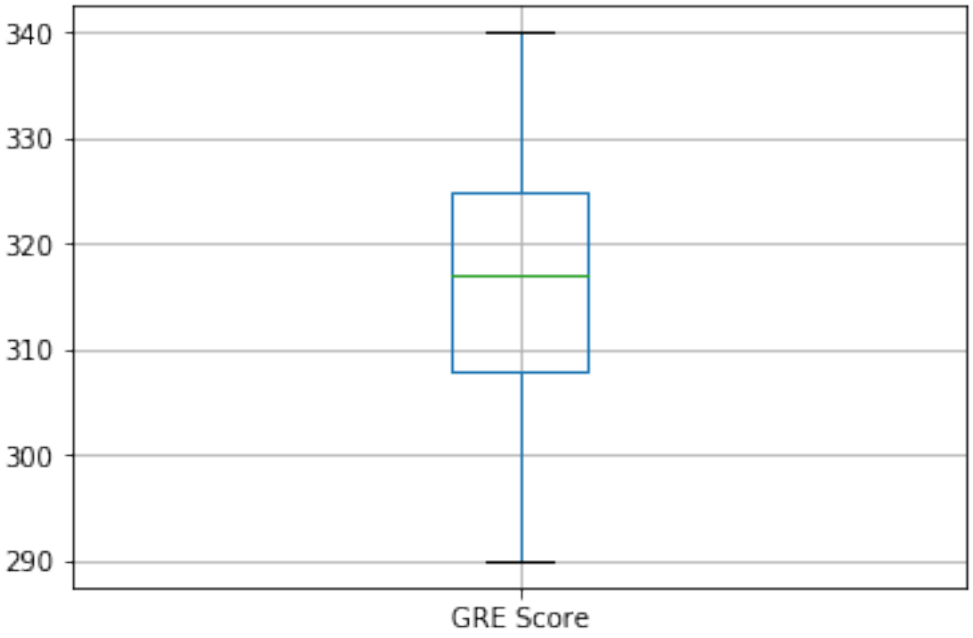
We drop the 'Serial No' column as it does not aid in the forthcoming analysis.

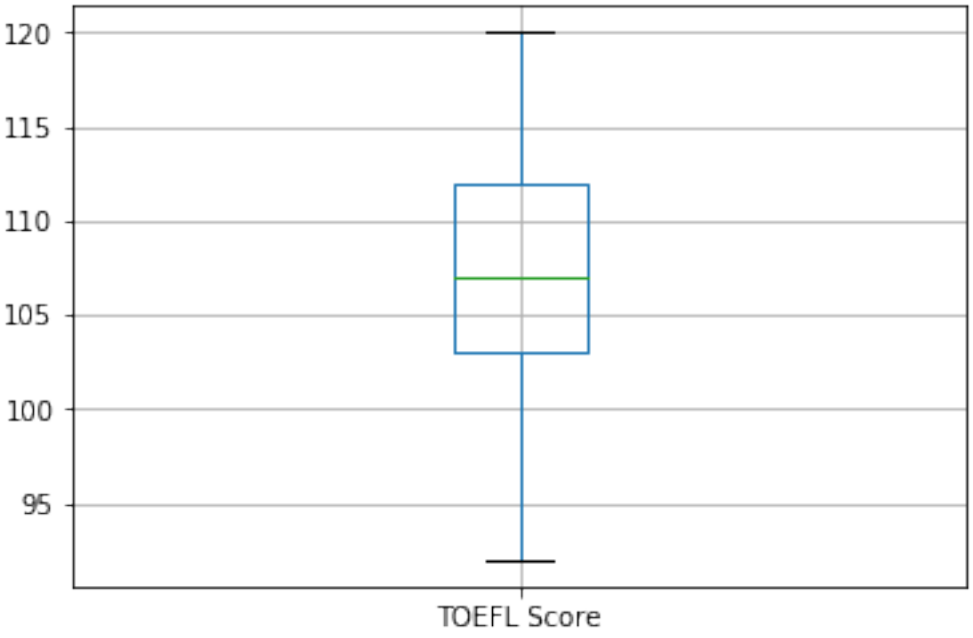
```
In [10]: Admissions=Admissions.drop(['Serial No.'],axis=1)
```

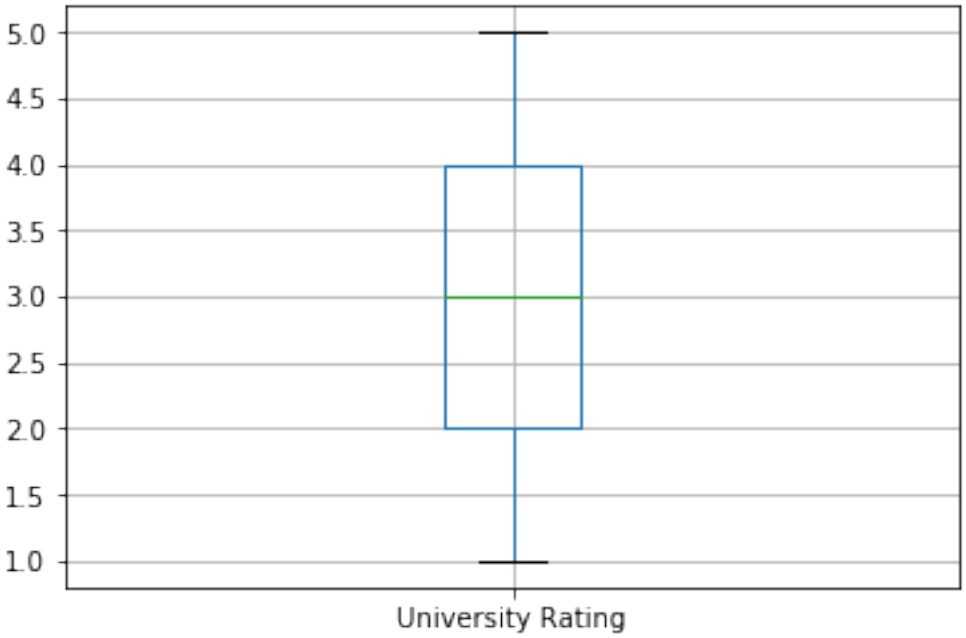
2.5 Checking for Outliers

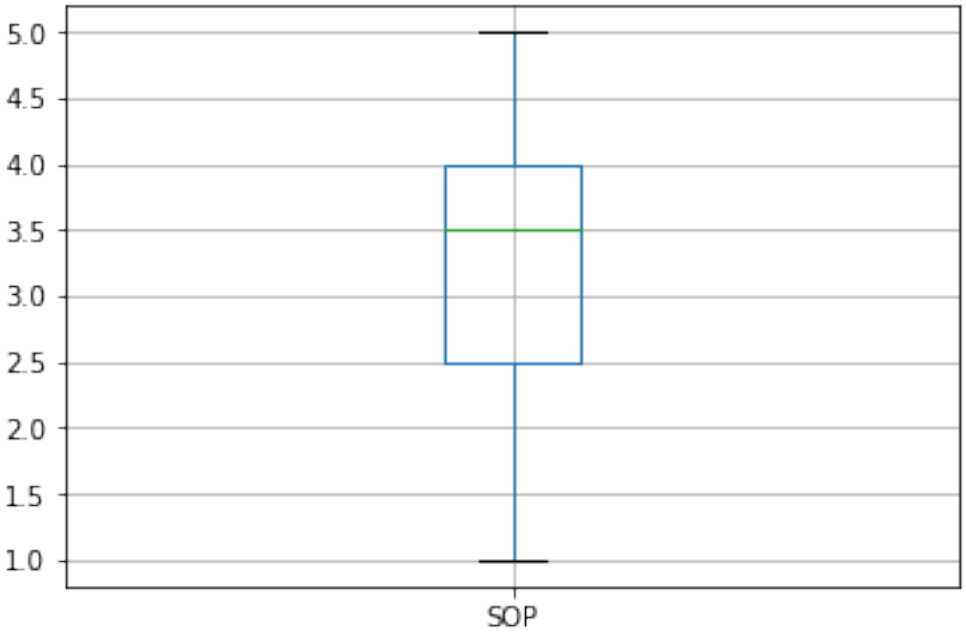
We have created boxplots for all the variables in the dataset. Looking at the plots, we can infer that there are no outliers.

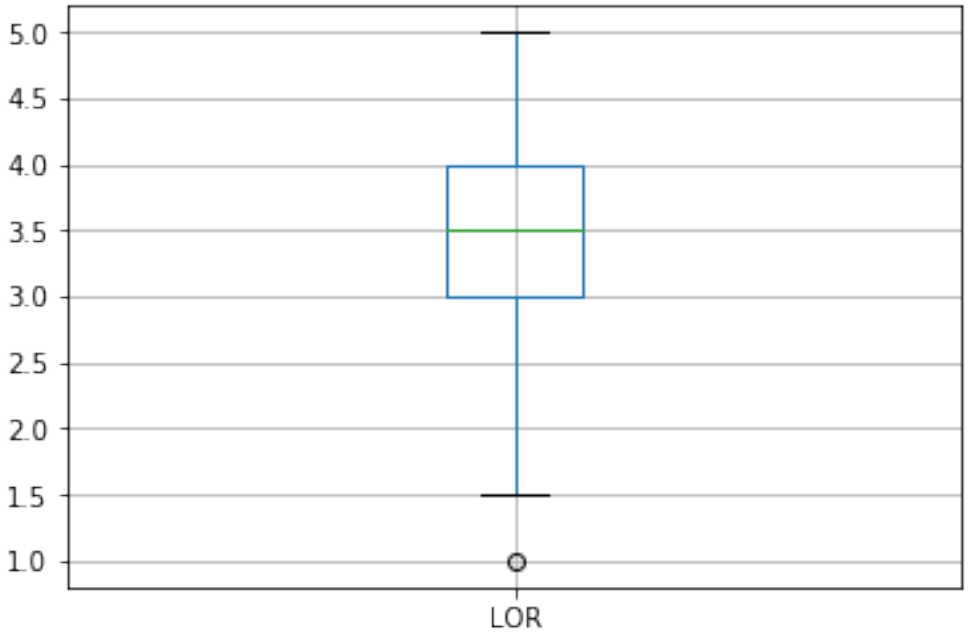
```
In [27]: for column in Admissions:
             plt.figure()
             Admissions.boxplot([column])
```

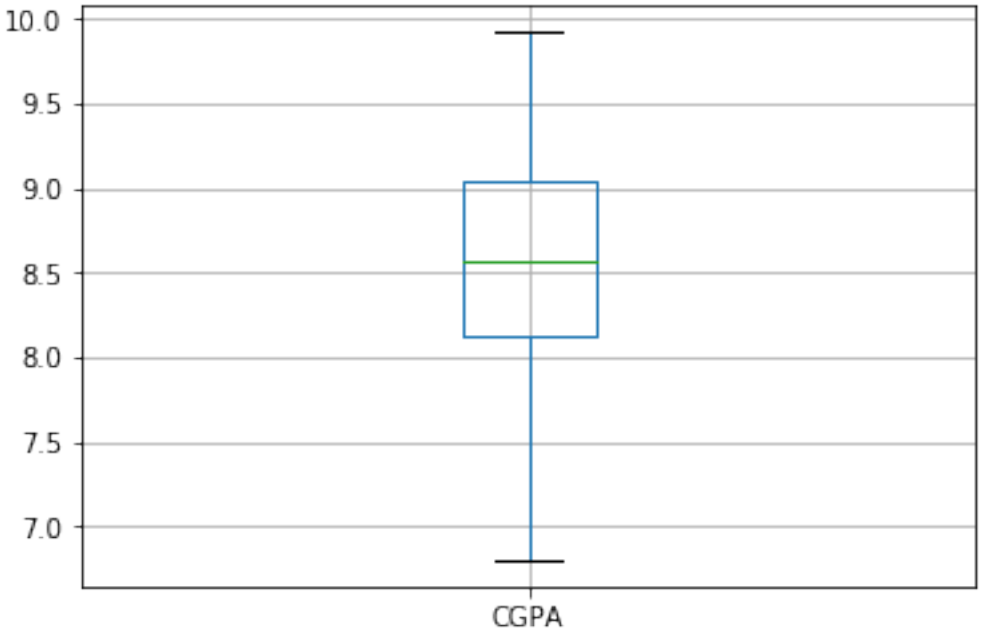

340
330
320
310
300
290
GRE Score

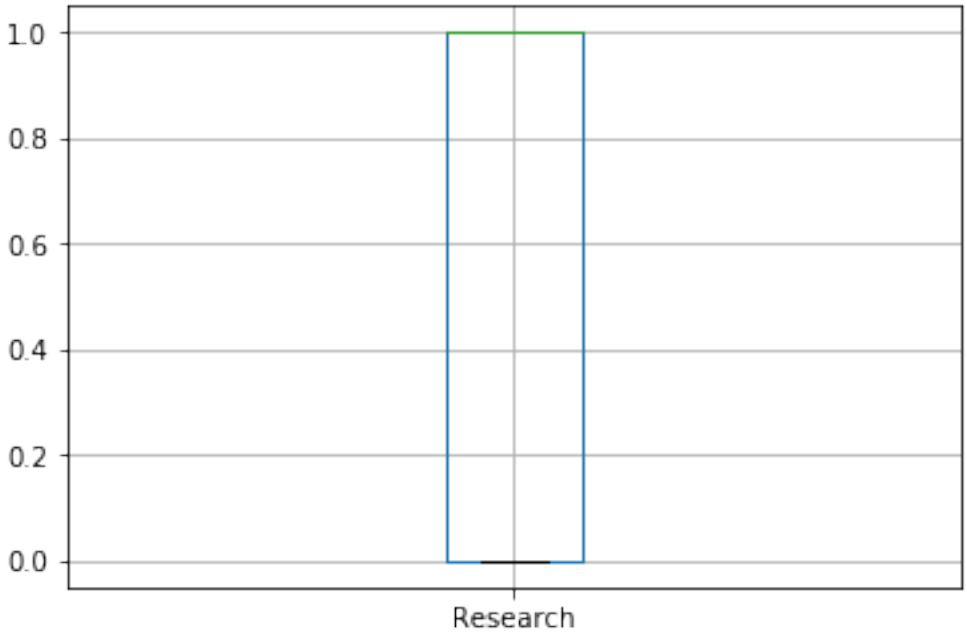

120
115
110
105
100
95
TOEFL Score

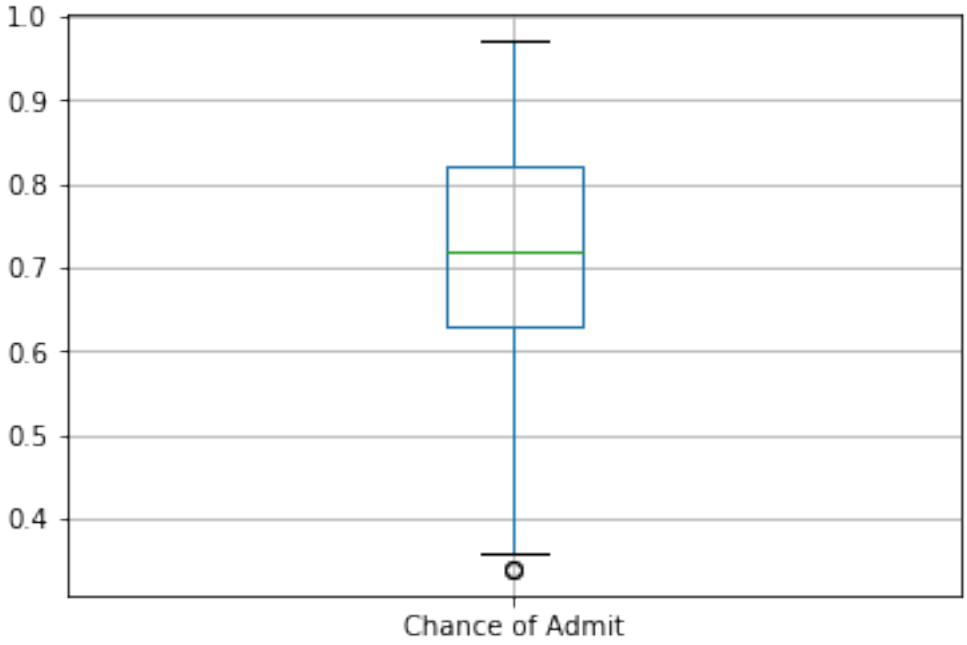

5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
University Rating


5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
SOP


5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
LOR


10.0
9.5
9.0
8.5
8.0
7.5
7.0
CGPA


1.0
0.8
0.6
0.4
0.2
0.0
Research


1.0
0.9
0.8
0.7
0.6
0.5
0.4
Chance of Admit

3 Data Exploration

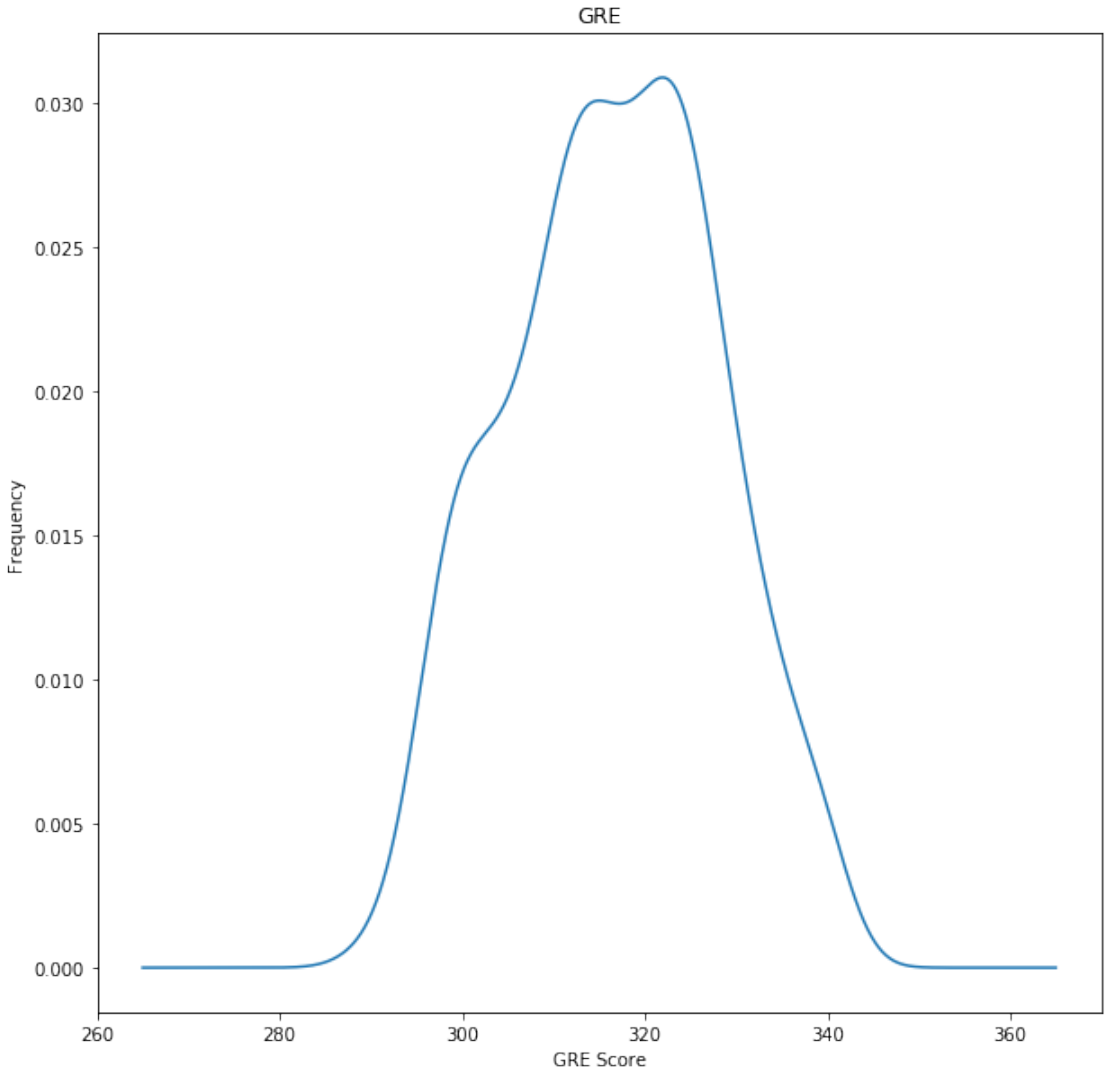
We are aiming to explore the data by visualising the variables under different circumstances. We first explore the data by visualising each variable by itself. We then proceed to pair descriptive features with the target feature and understand their relationships. Finally, we create a scatter matrix to understand the aggregate behaviour of data.

3.1 Univariate Visualisations

3.1.1 Density plot of GRE Score variable

Looking at this graph we can see that the data is dense in between the range of 310-340.

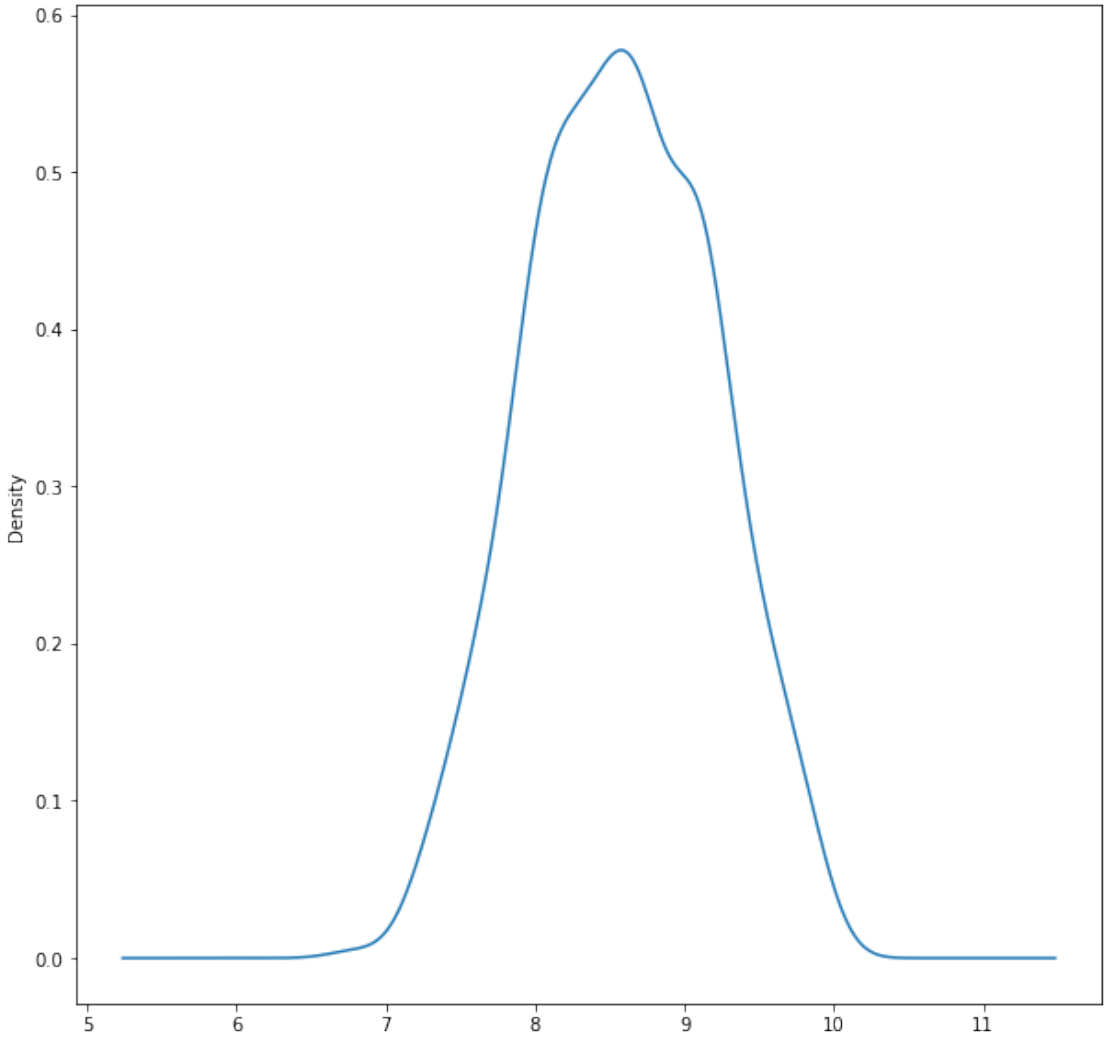
```
In [17]: Admissions['GRE Score'].plot(kind='density',figsize=(10,10))
         plt.xlabel('GRE Score')
         plt.ylabel('Frequency')
         plt.title('GRE')
         plt.show()
```



3.1.2 Density Plot for CGPA variable

Looking at this graph we can see that the data is dense in between the range of 8-9.

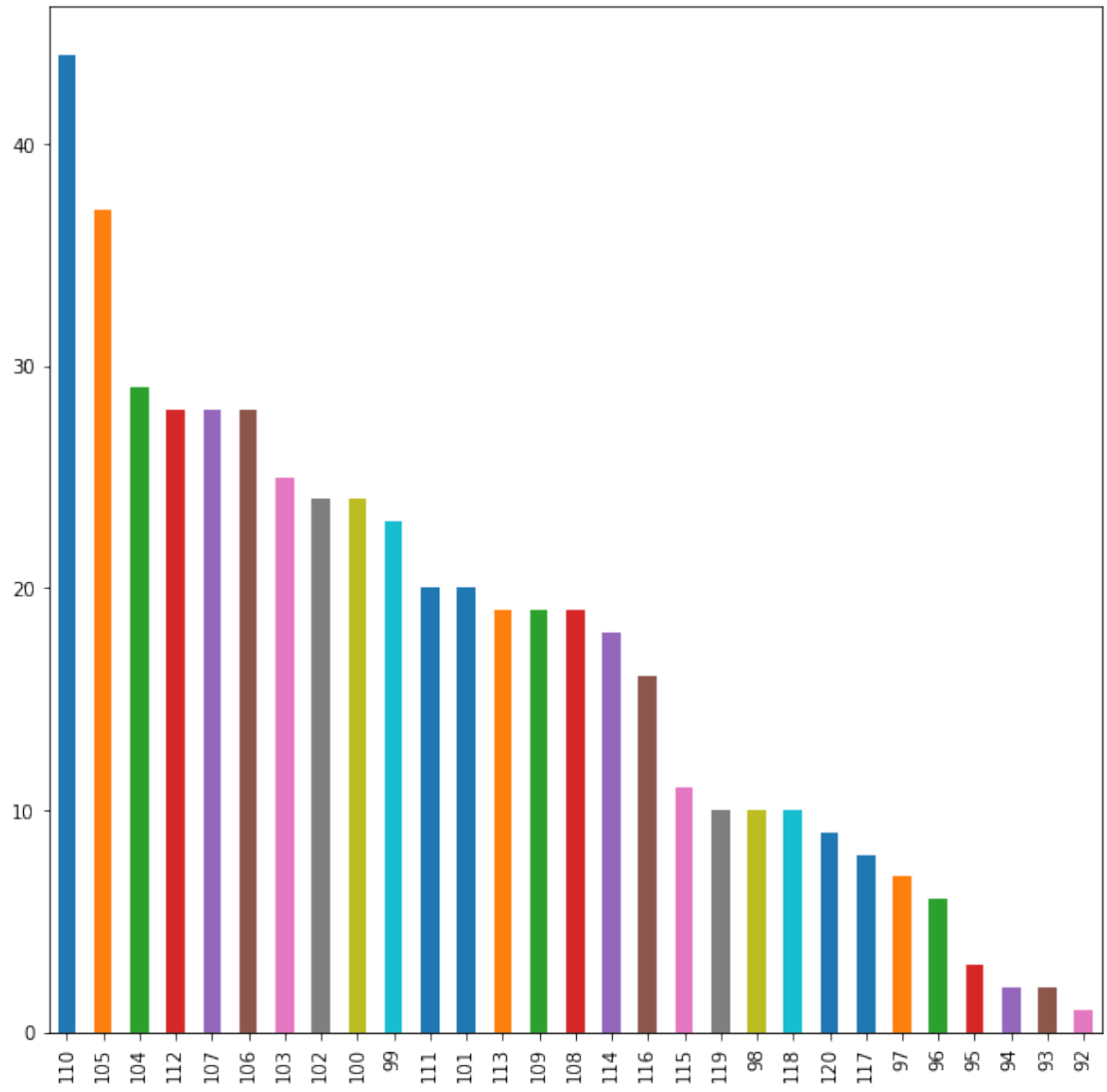
```
In [36]: Admissions['CGPA'].plot(kind='density')
         plt.gcf().set_size_inches((10, 10))
```



3.1.3 Bar Plot for TOEFL Score variable

Looking at the Bar plot we can see that the data is ranging from 110-92. Which means that most of the students have scored high in TOEFL exam. There are also students who have scored the highest mark i.e. 110.

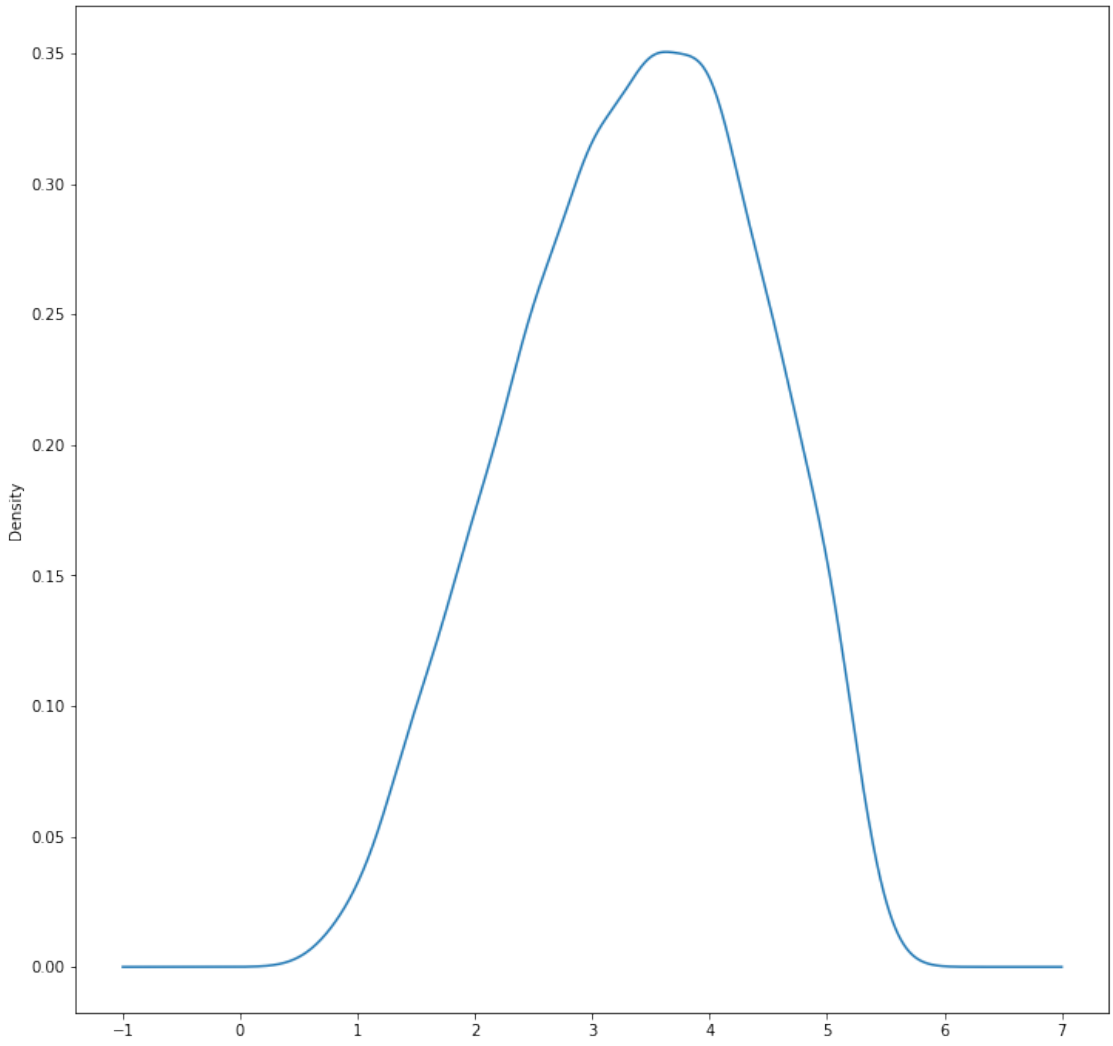
```
In [37]: Admissions['TOEFL Score'].value_counts().plot(kind='bar')
         plt.gcf().set_size_inches((10, 10))
```



3.1.4 Density Plot for SOP variable

Looking at this graph we can see that the data is dense in between the range of 2-4. For some profiles, SOP strength was asked to the applicants themselves. A few values were extrapolated using other parameters.

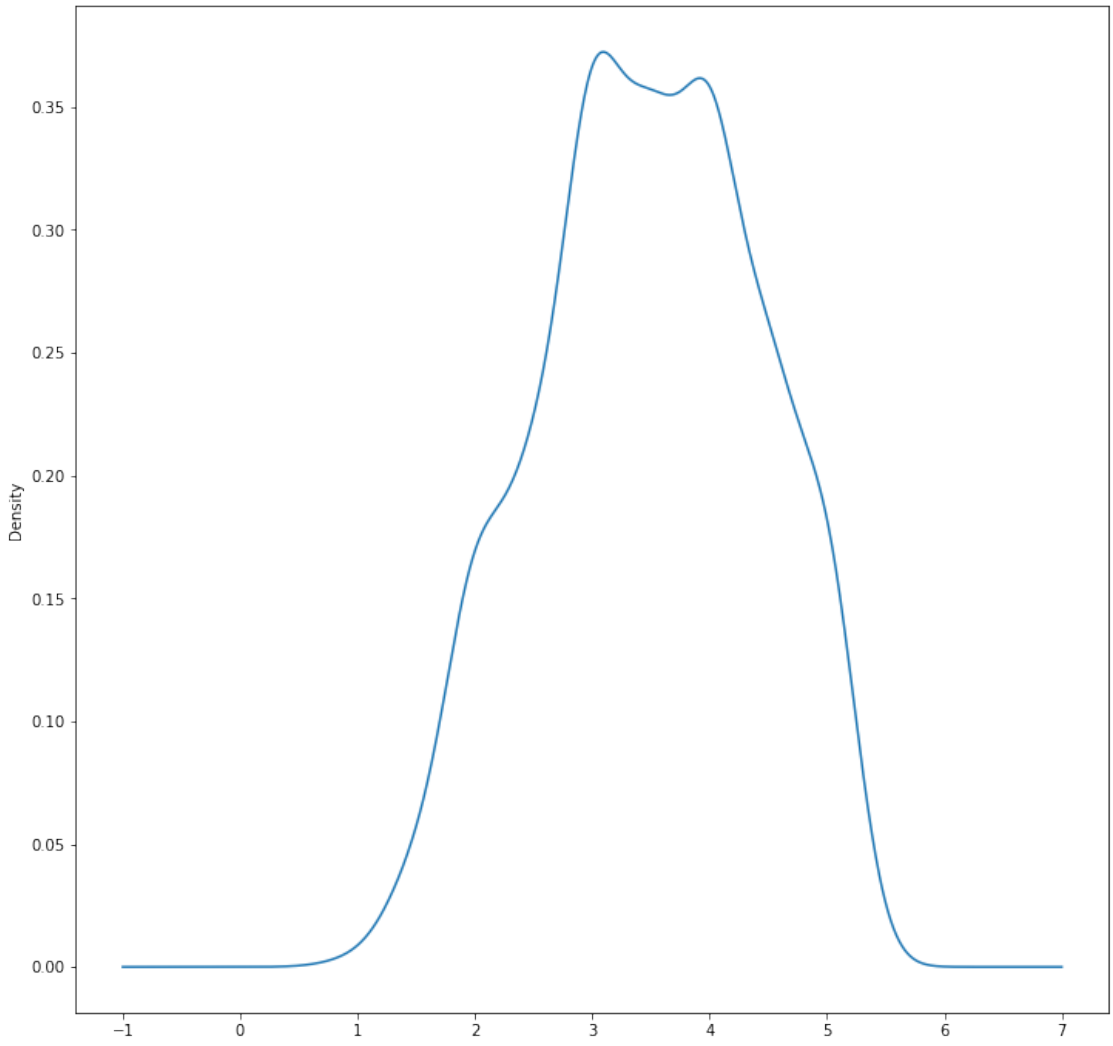
```
In [20]: Admissions['SOP'].plot(kind='density')
         plt.gcf().set_size_inches((12, 12))
```



3.1.5 Density Plot for LOR variable

Looking at this graph we can see that the data is dense in between the range of 3-5. For some profiles, LOR strength was asked to the applicants themselves. A few values were extrapolated using other parameters.

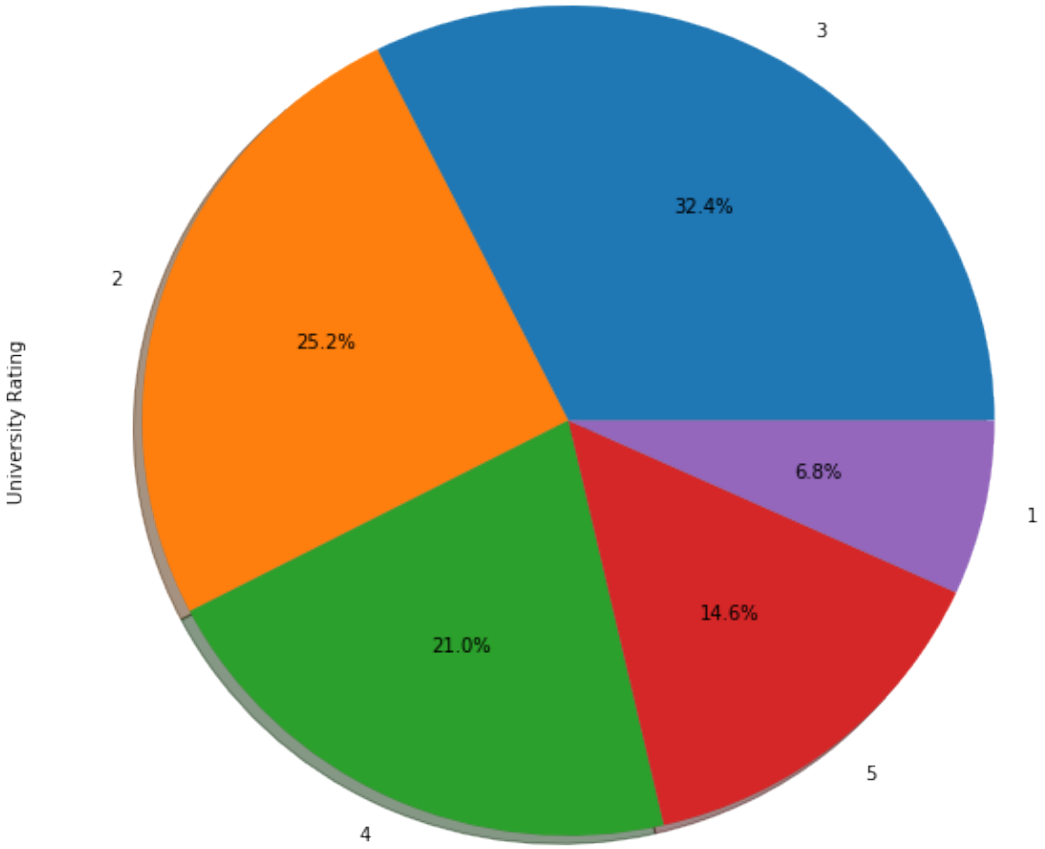
```
In [21]: Admissions['LOR'].plot(kind='density')
         plt.gcf().set_size_inches((12, 12))
```



3.1.6 Pie Chart of University Rating Variable

Looking at the pie chart below we can infer that University Rating 3 has the most number of data points followed by University Rating 2 coming in at second.

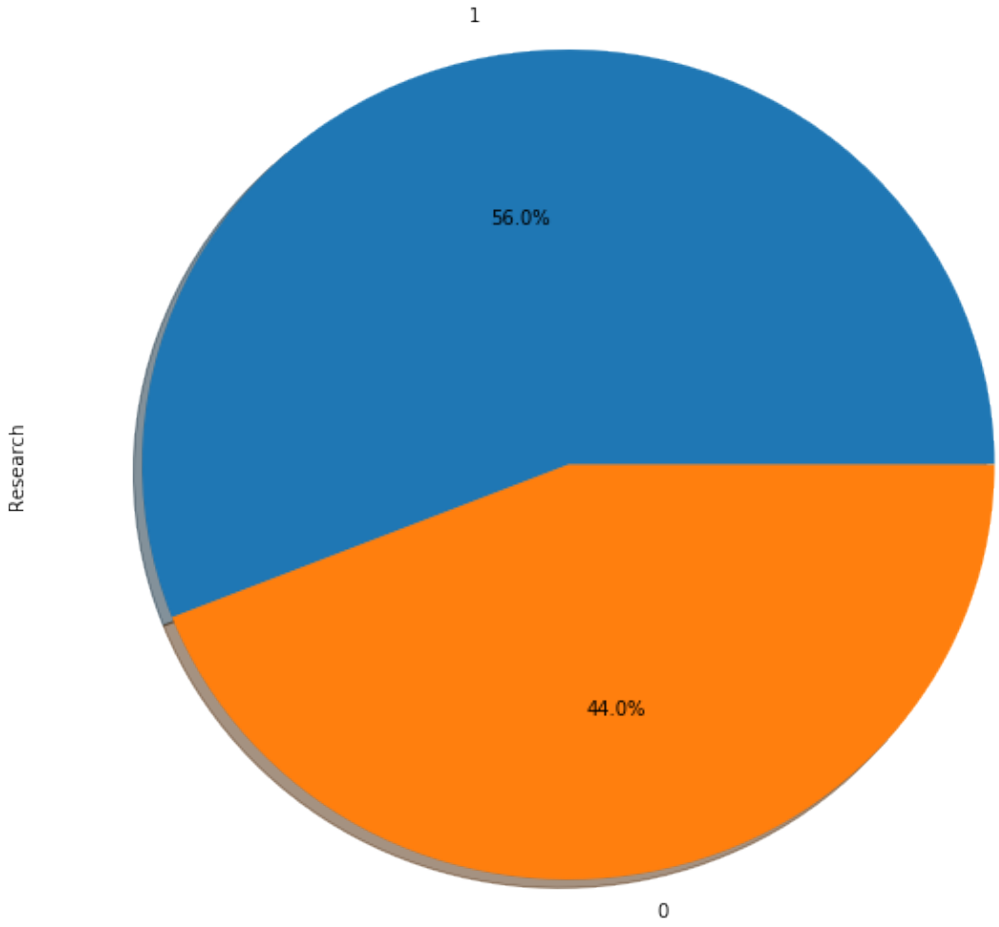
```
In [25]: Admissions['University Rating'].value_counts().plot(kind='pie',shadow=True,autopct='%
         plt.gcf().set_size_inches((10, 10))
```



3.1.7 Pie Chart of Research Variable

Looking at the pie chart below we can infer that in the data set 56% of the students have research experience whereas 44% do not have research experience.

```
In [24]: Admissions['Research'].value_counts().plot(kind='pie',shadow=True,autopct='%1.1f%%')
         plt.gcf().set_size_inches((10, 10))
```



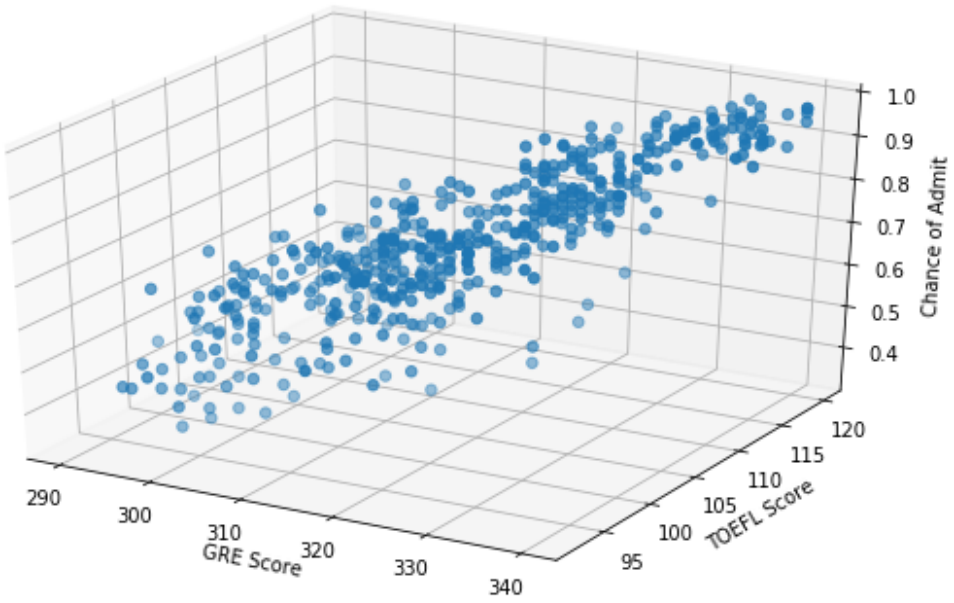
3.2 Multi-Variate Visualisation

3.2.1 3D Scatter Plot of GRE Score, TOEFL Score across Chance of Admit

We have created a 3d scatter plot comparing three variables i.e GRE Score, TOEFL Score and Chance of Admit. Looking at the plot we can clearly observe a positive linear relation between the three. This implies that the higher the student's GRE and TOEFL scores are; higher their chances of admission.

```
In [28]: from mpl_toolkits.mplot3d import Axes3D
         fig=plt.figure(figsize=(10,6))
         ax=fig.add_subplot(111,projection='3d')
         ax.scatter(Admissions['GRE Score'],Admissions['TOEFL Score'],Admissions['Chance of Adm
         ax.set_xlabel('GRE Score')
         ax.set_ylabel('TOEFL Score')
```

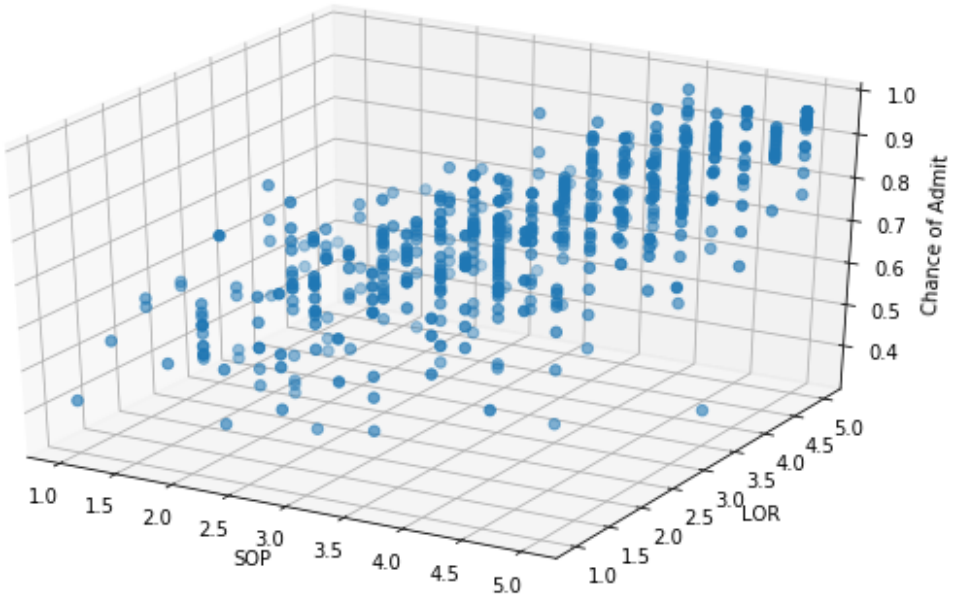
```
ax.set_zlabel('Chance of Admit')
plt.show()
```



3.2.2 3D Scatter Plot of SOP , LOR weights across Chance of Admit

We have created a 3D scatter plot comparing three variables i.e SOP, LOR and Chance of Admit. Looking at the plot we can clearly observe a positive linear relation between the three. This implies that the higher the student's SOP and LOR weights; higher their chances of admission.

```
In [29]: from mpl_toolkits.mplot3d import Axes3D
         fig=plt.figure(figsize=(10,6))
         ax=fig.add_subplot(111,projection='3d')
         ax.scatter(Admissions['SOP'],Admissions['LOR'],Admissions['Chance of Admit'],s=30)
         ax.set_xlabel('SOP')
         ax.set_ylabel('LOR')
         ax.set_zlabel('Chance of Admit')
         plt.show()
```

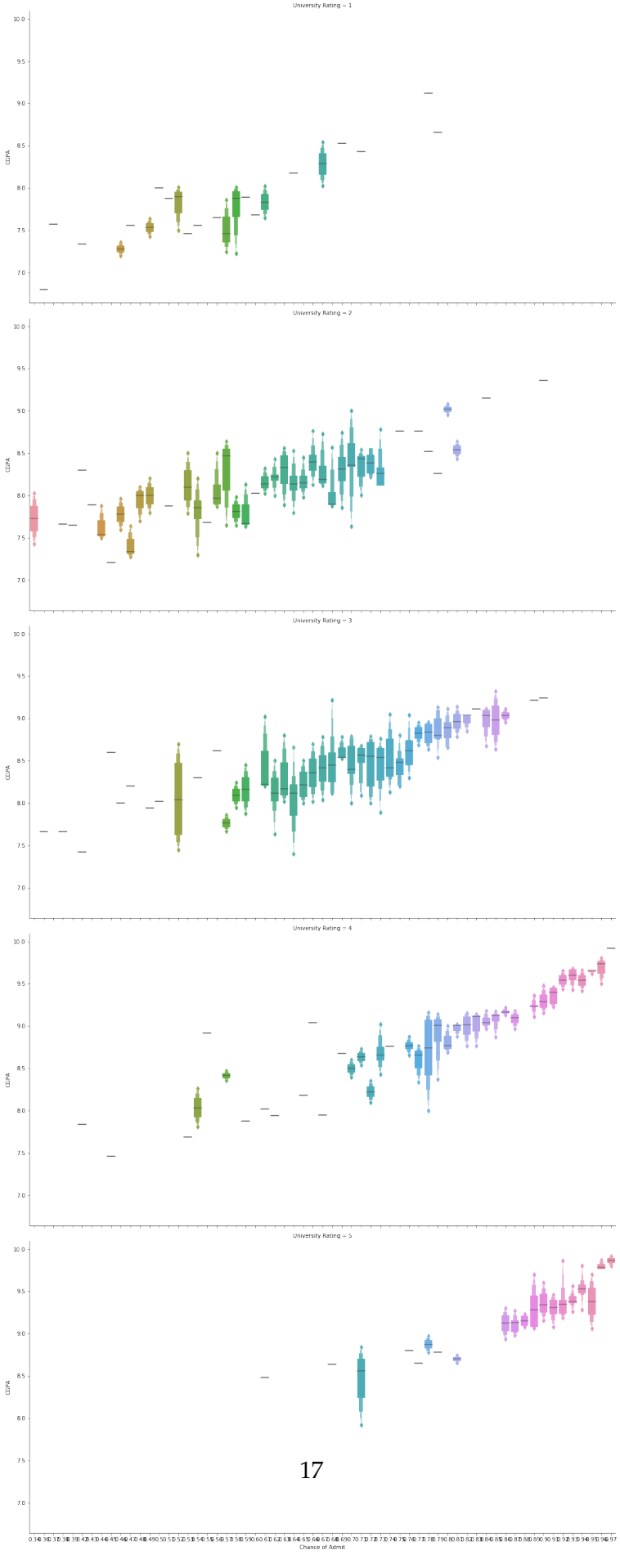


3.2.3 Boxen Plot of Chance of Admit and CGPA grouped by University Rating

The below plots helps us understand the impact of a CGPA on Chance of Admit. There is a positive linear relationship between the two. However, we take a step further and visualise this spanning across different university ratings. We can observe that higher the CGPA, the student has a better chance of admit at a university with higher rating.

```
In [30]: sns.catplot(x='Chance of Admit',y='CGPA',row='University Rating',kind='boxen',data=Adm
```

```
Out[30]: <seaborn.axisgrid.FacetGrid at 0xc8c26d8>
```

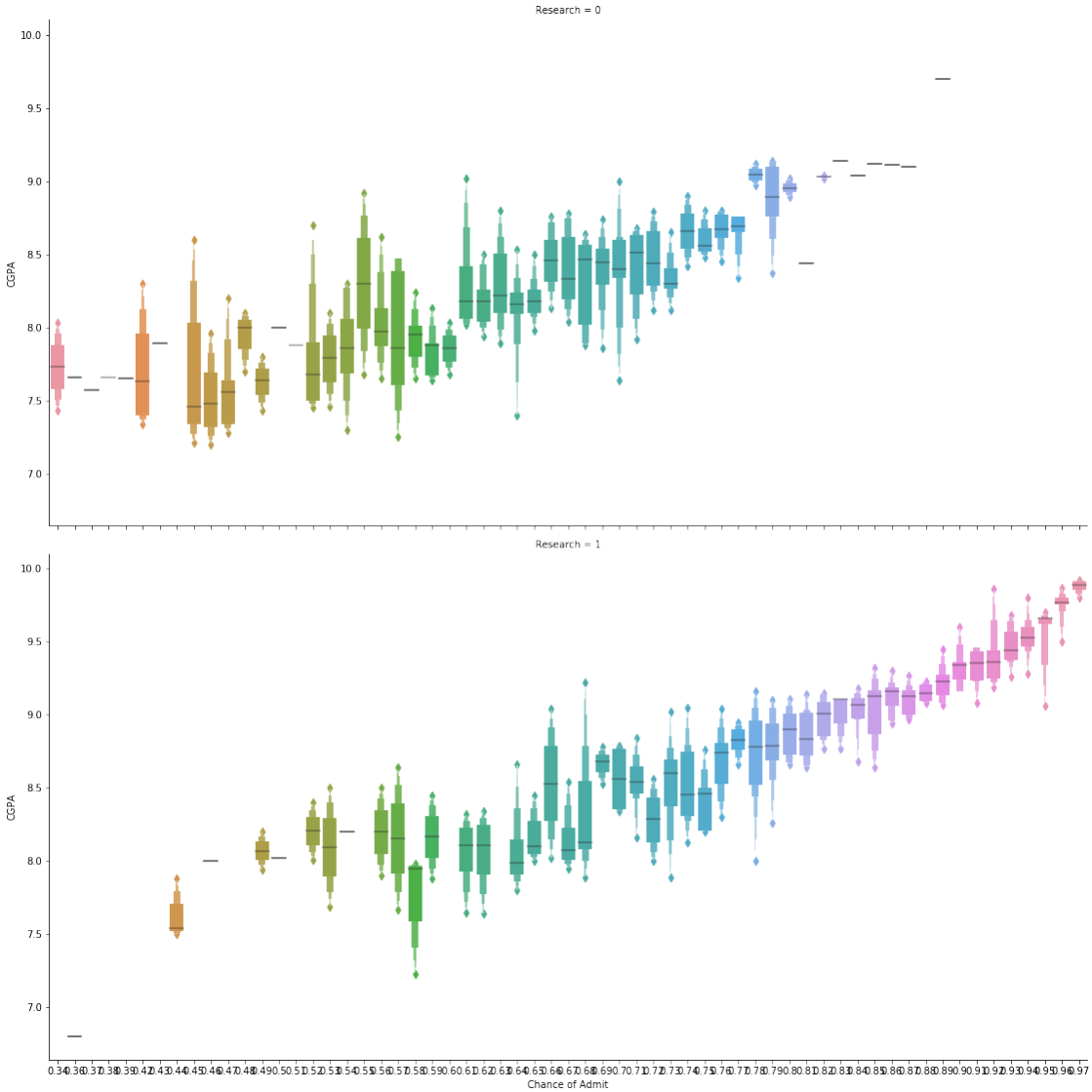

University Rating = 1
University Rating = 2
University Rating = 3
University Rating = 4
University Rating = 5
CGPA
Chance of Admit

3.2.4 Boxen Plot of Chance of Admit and CGPA grouped by Research

The below plots helps us understand the impact of a CGPA on Chance of Admit. There is a positive linear relationship between the two. However, we take a step further and visualise this spanning across Research Experience. We can observe that the students with Research experience have a higher chance of Admit in comparision with students who don't have research experience.

```
In [31]: sns.catplot(x='Chance of Admit',y='CGPA',row='Research',kind='boxen',data=Admissions,
```

```
Out[31]: <seaborn.axisgrid.FacetGrid at 0x10b6aac8>
```



4 Data Aggregation by Scatter Matrix

We have populated a scatter matrix on the overall data set. We will discuss the aggregate data behaviour as two components of Data Distribution and overall relationships among variables.

1. Data Distribution:

- GRE Score, TOEFL Score, University Rating and CGPA variables are normally distributed and symmetric.
- SOP, LOR and Chance of Admit are Normally distributed and are skewed to the left.

2. Relationships:

- GRE Score has a strong positive linear relationship with Chance of Admit, CGPA and TOEFL Score
- TOEFL Score has a strong positive linear relationship with Chance of Admit, CGPA and GRE Score
- SOP Score has a positive linear relationship with Chance of Admit, CGPA, TOEFL and GRE Score
- LOR Score has a positive linear relationship with Chance of Admit, CGPA, TOEFL and GRE Score
- Chance of Admit has a positive linear relationship with CGPA, TOEFL, LOR, SOP and GRE Score

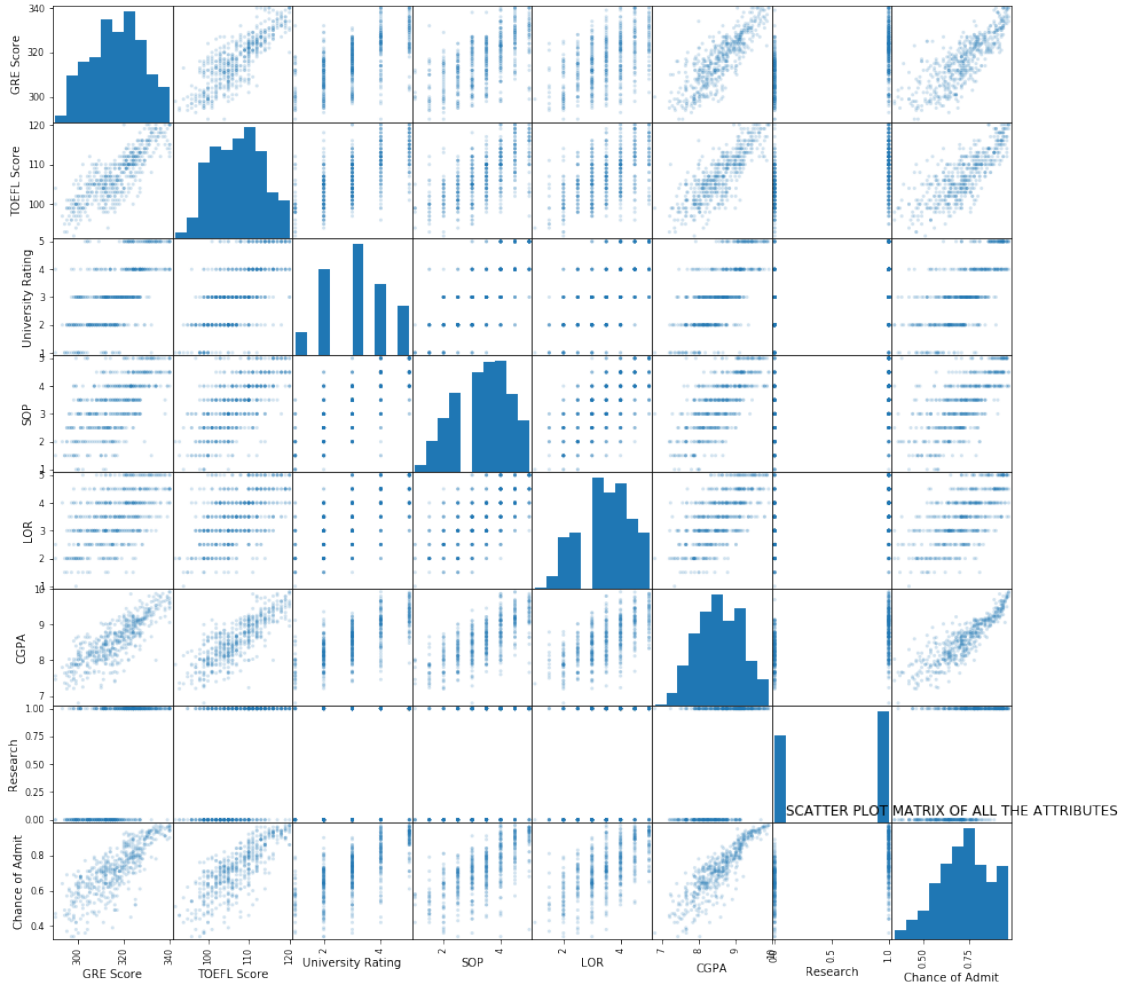
```
In [38]: from pandas.plotting import scatter_matrix
         numerics = ['int16', 'int32', 'int64', 'float16', 'float32', 'float64']
         Admissions_num = Admissions.select_dtypes(include=numerics)
         scatter_matrix(Admissions_num,alpha=0.2,figsize=(15,15),diagonal='hist')
         plt.title('SCATTER PLOT MATRIX OF ALL THE ATTRIBUTES')
         plots.show()


        ---------------------------------------------------------------------------

        NameError                                 Traceback (most recent call last)

        <ipython-input-38-4070132b8762> in <module>()
          4 scatter_matrix(Admissions_num,alpha=0.2,figsize=(15,15),diagonal='hist')
          5 plt.title('SCATTER PLOT MATRIX OF ALL THE ATTRIBUTES')
    ----> 6 plots.show()


        NameError: name 'plots' is not defined
```



5 Summary

In Phase 1, we did not find any null values, outliers. We removed all the whitespaces from the dataset. We dropped the attribute serial number as it does not provide any useful information. From the data exploration, we found that GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research are potentially useful features in predicting the chance of admit.

6 Citation

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