PREDICTION OF GRADUATE ADMISSION

**Importing and reading data**

In [1]:

**import** **pandas** **as** **pd**

**import** **numpy** **as** **np**

**import** **warnings**

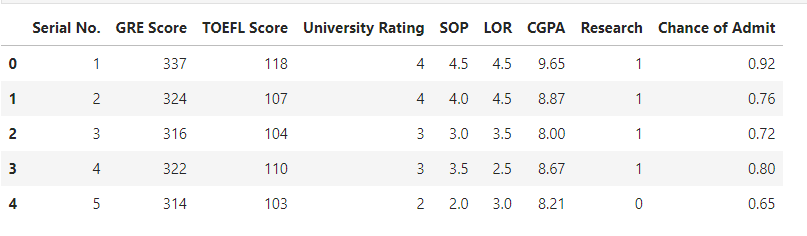
warnings**.**filterwarnings("ignore")

In [2]:

admission\_df **=** pd**.**read\_csv(r"C:\Users\Srividhya\Desktop\admission prediction dataset\Admission\_predict.csv")

In [3]:

admission\_df**.**head()

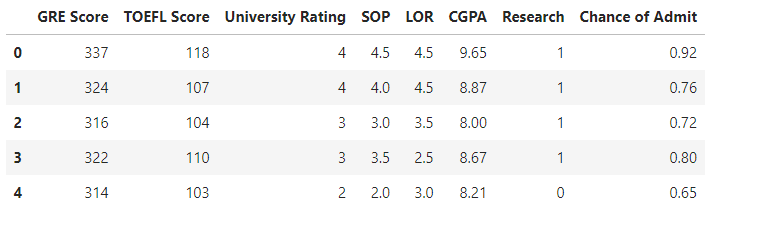


admission\_df**.**drop(columns **=** "Serial No.", inplace **=** **True**)

Dropping the column Serial No. since it is not necessary.

In [5]:

admission\_df**.**head()



**Checking the dataset for any null values**

In [6]:

admission\_df**.**isnull()**.**sum()

Out[6]:

GRE Score 0

TOEFL Score 0

University Rating 0

SOP 0

LOR 0

CGPA 0

Research 0

Chance of Admit 0

dtype: int64

There are no null values and hence it is good to proceed with analysis.

In [7]:

admission\_df**.**dtypes

Out[7]:

GRE Score int64

TOEFL Score int64

University Rating int64

SOP float64

LOR float64

CGPA float64

Research int64

Chance of Admit float64

dtype: object

There are no data types that are categorical. Hence, preprocessing is not necessary.

**Distribution of scores**

In [8]:

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

**%matplotlib** inline

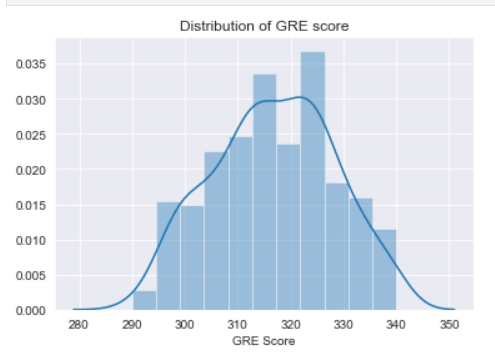
sns**.**set\_style("darkgrid")

In [9]:

plt1 **=** sns**.**distplot(admission\_df['GRE Score'])

plt**.**title("Distribution of GRE score")

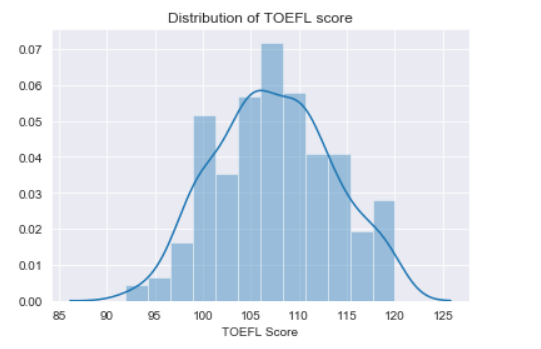
plt**.**show()



plt1 **=** sns**.**distplot(admission\_df['TOEFL Score'])

plt**.**title("Distribution of TOEFL score")

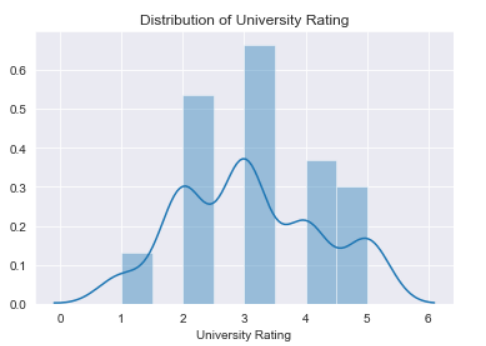
plt**.**show()



plt1 **=** sns**.**distplot(admission\_df['University Rating'])

plt**.**title("Distribution of University Rating")

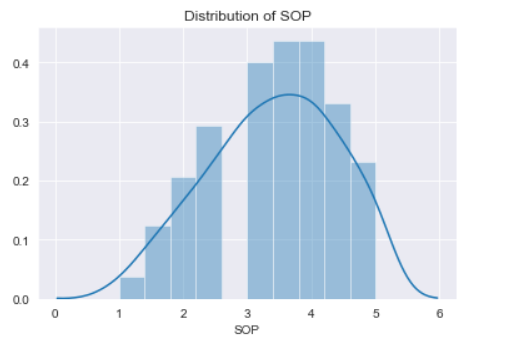
plt**.**show()



plt1 **=** sns**.**distplot(admission\_df['SOP'])

plt**.**title("Distribution of SOP")

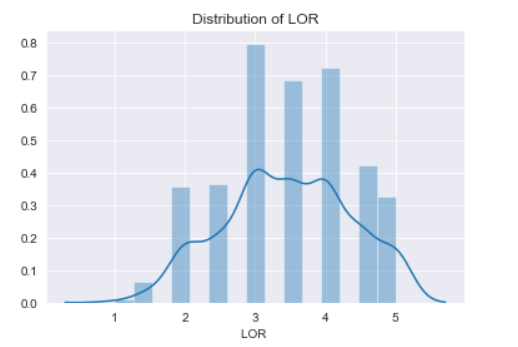
plt**.**show()



plt1 **=** sns**.**distplot(admission\_df['LOR '])

plt**.**title("Distribution of LOR")

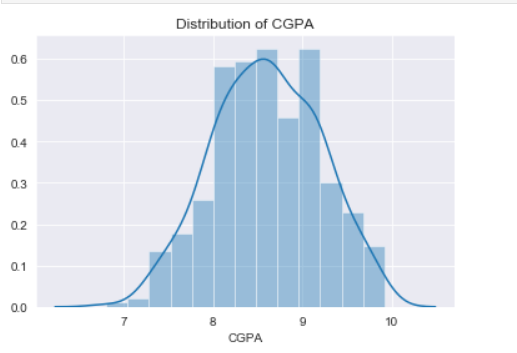
plt**.**show()



plt1 **=** sns**.**distplot(admission\_df['CGPA'])

plt**.**title("Distribution of CGPA")

plt**.**show()



With the help of these distribution plots, we can easily see the scores of GRE, TOEFL, and so on that are most commonly scored by the students. The most common scores in:

1. GRE --> between 320 to 330 and then 310 to 320
2. TOEFL --> between 105 and 110. Mostly above 100.
3. University Rating --> 3
4. SOP --> between 3 and 4, mostly near 4
5. LOR --> 3 and 4
6. CGPA --> above 8.5 and also, 9 is mostly scored cgpa.

**Distribution of Scores against its mean score**

In [15]:

**import** **plotly**

**import** **plotly.graph\_objs** **as** **go**

avg\_gre\_score **=** admission\_df['GRE Score']**.**mean()

data **=** [go**.**Histogram(

x **=** admission\_df['GRE Score']

)]

*# Vertical dashed line to indicate the average app rating*

layout **=** {'shapes': [{

'type' :'line',

'x0': avg\_gre\_score,

'y0': 0,

'x1': avg\_gre\_score,

'y1': 100,

'line': { 'dash': 'dashdot'}

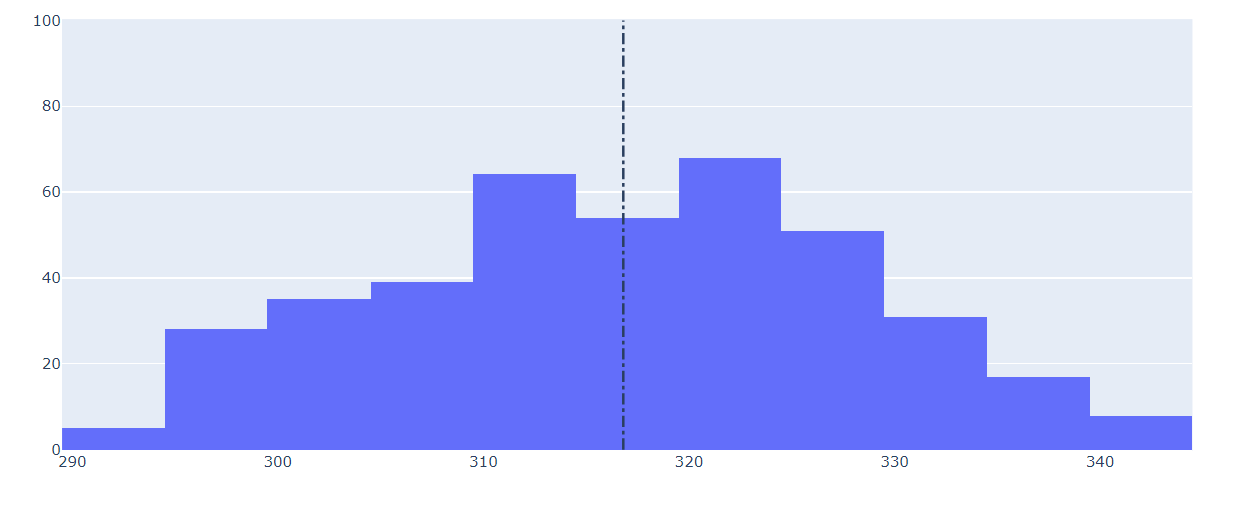
}]

}

print(avg\_gre\_score)

plotly**.**offline**.**iplot({'data': data, 'layout': layout})

316.8075



avg\_toefl\_score **=** admission\_df['TOEFL Score']**.**mean()

data **=** [go**.**Histogram(

x **=** admission\_df['TOEFL Score']

)]

*# Vertical dashed line to indicate the average app rating*

layout **=** {'shapes': [{

'type' :'line',

'x0': avg\_toefl\_score,

'y0': 0,

'x1': avg\_toefl\_score,

'y1': 100,

'line': { 'dash': 'dashdot'}

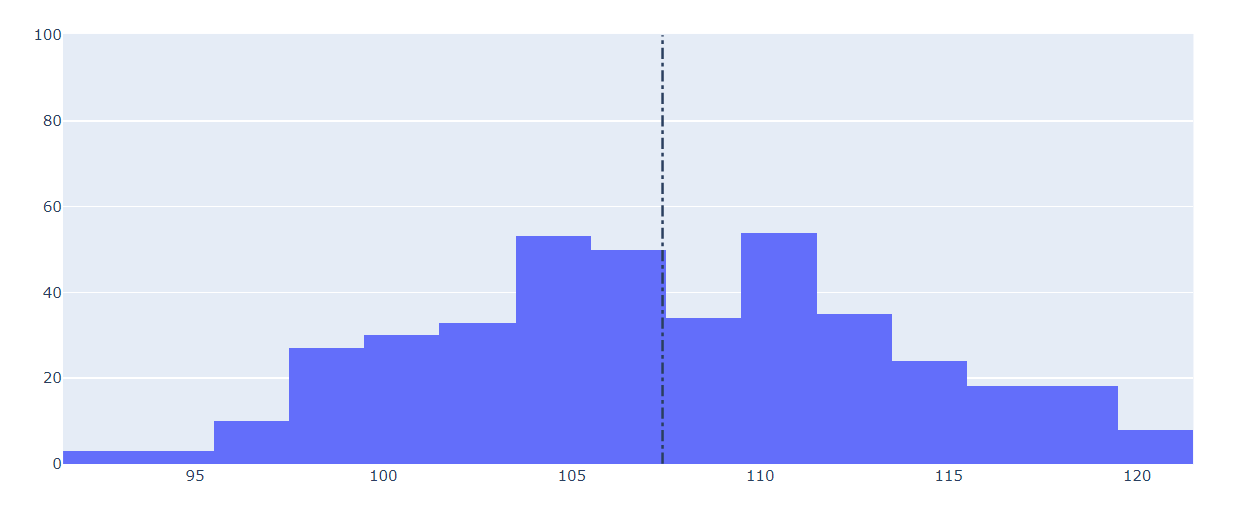
}]

}

print(avg\_toefl\_score)

plotly**.**offline**.**iplot({'data': data, 'layout': layout})

107.41



avg\_sop\_score **=** admission\_df['SOP']**.**mean()

data **=** [go**.**Histogram(

x **=** admission\_df['SOP']

)]

*# Vertical dashed line to indicate the average app rating*

layout **=** {'shapes': [{

'type' :'line',

'x0': avg\_sop\_score,

'y0': 0,

'x1': avg\_sop\_score,

'y1': 100,

'line': { 'dash': 'dashdot'}

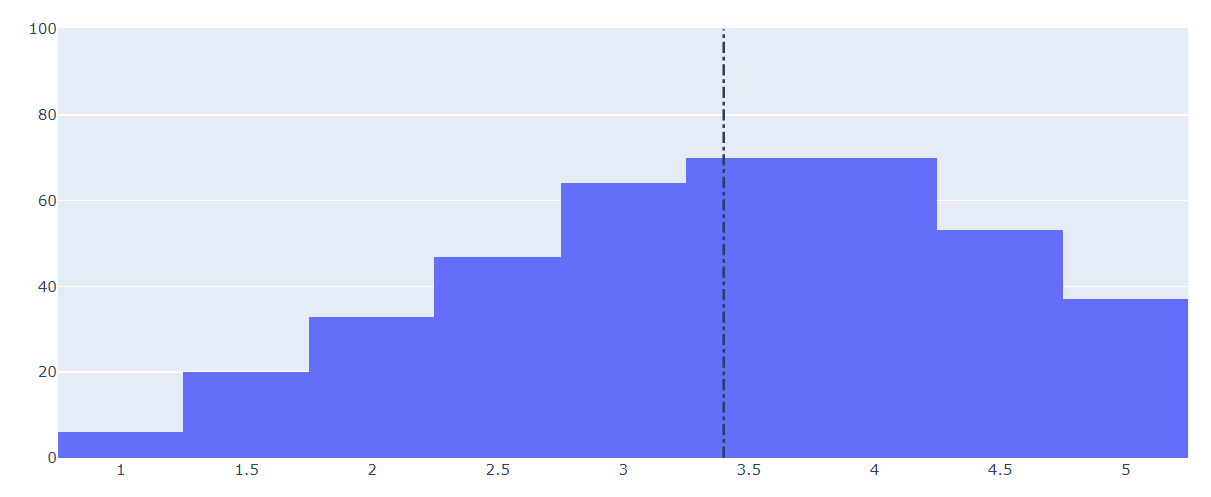
}]

}

print(avg\_sop\_score)

plotly**.**offline**.**iplot({'data': data, 'layout': layout})

3.4



avg\_lor\_score **=** admission\_df['LOR ']**.**mean()

data **=** [go**.**Histogram(

x **=** admission\_df['LOR ']

)]

*# Vertical dashed line to indicate the average app rating*

layout **=** {'shapes': [{

'type' :'line',

'x0': avg\_lor\_score,

'y0': 0,

'x1': avg\_lor\_score,

'y1': 100,

'line': { 'dash': 'dashdot'}

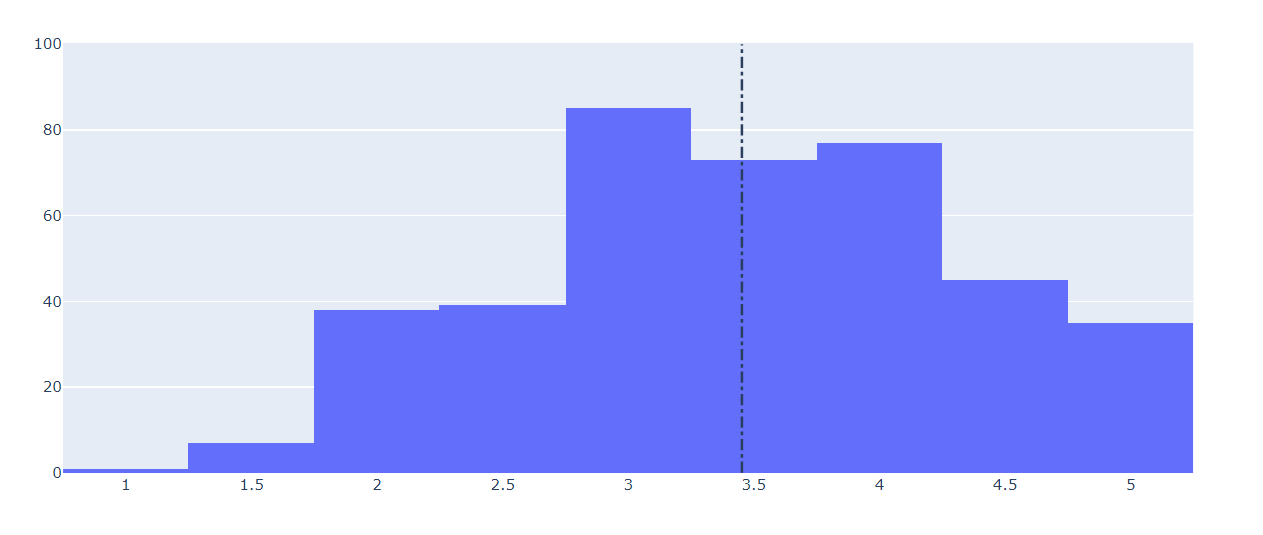
}]

}

print(avg\_lor\_score)

plotly**.**offline**.**iplot({'data': data, 'layout': layout})

3.4525



avg\_cgpa\_score **=** admission\_df['CGPA']**.**mean()

data **=** [go**.**Histogram(

x **=** admission\_df['CGPA']

)]

*# Vertical dashed line to indicate the average app rating*

layout **=** {'shapes': [{

'type' :'line',

'x0': avg\_cgpa\_score,

'y0': 0,

'x1': avg\_cgpa\_score,

'y1': 100,

'line': { 'dash': 'dashdot'}

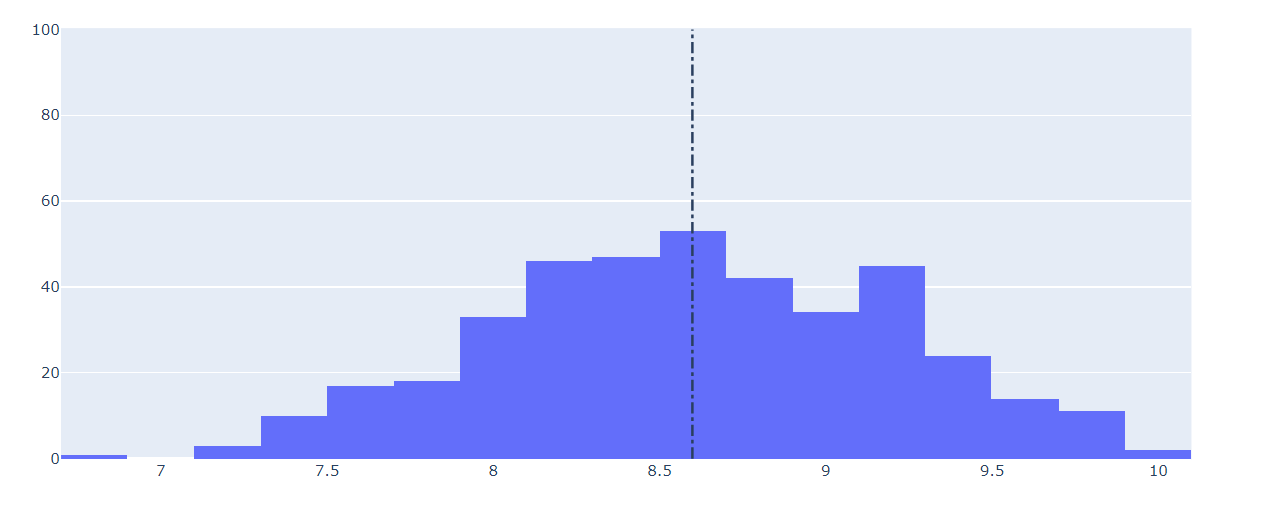
}]

}

print(avg\_cgpa\_score)

plotly**.**offline**.**iplot({'data': data, 'layout': layout})

8.598924999999998



From the analysis, it is clear that:

**The mean score of:**

1. GRE --> 315 - 319
2. TOEFL --> 106 - 107
3. SOP --> 3.5
4. LOR --> 3.5
5. CGPA --> 8.5 - 8.69

**Relation between different factors in the dataset**

In [20]:

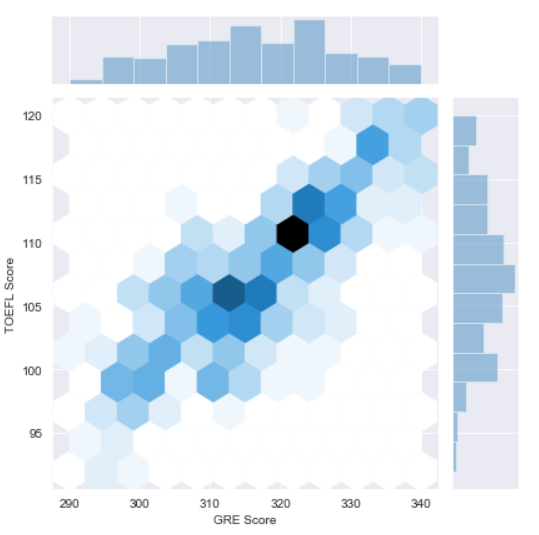
plt2 **=** sns**.**jointplot(x **=** admission\_df['GRE Score'], y **=** admission\_df['TOEFL Score'], kind **=** 'hex')

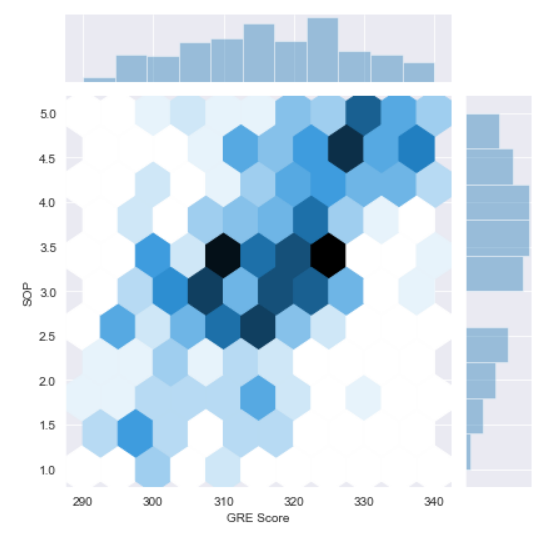
plt3 **=** sns**.**jointplot(x **=** admission\_df['GRE Score'], y **=** admission\_df['SOP'], kind **=** 'hex')

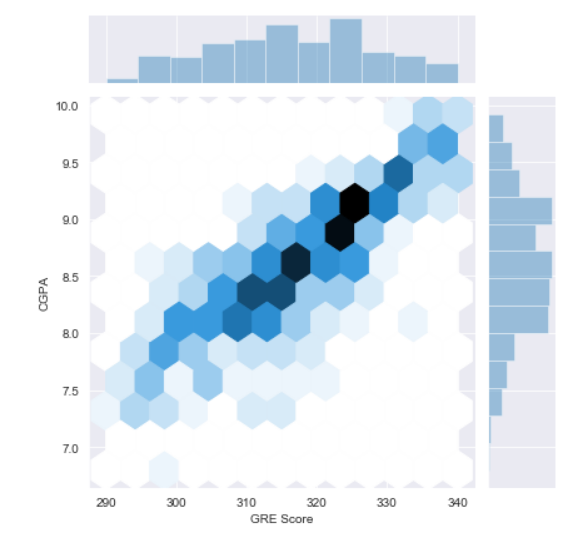
plt4 **=** sns**.**jointplot(x **=** admission\_df['GRE Score'], y **=** admission\_df['CGPA'], kind **=** 'hex')

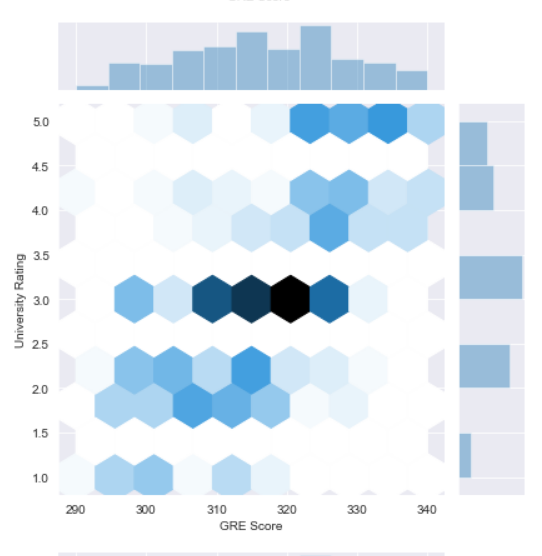
plt5 **=** sns**.**jointplot(x **=** admission\_df['GRE Score'], y **=** admission\_df['University Rating'], kind **=** 'hex')

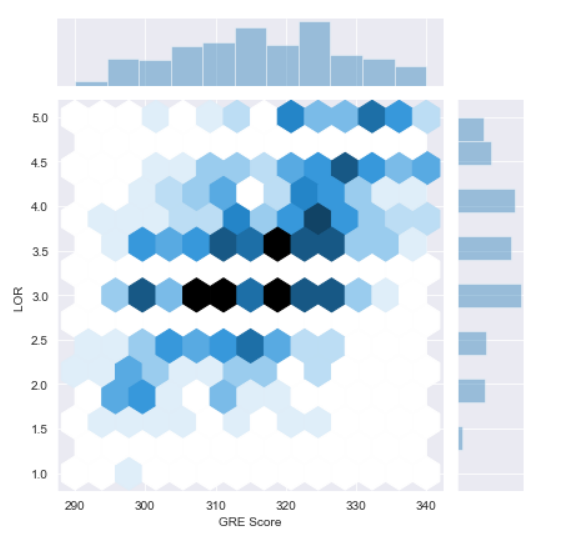
plt5 **=** sns**.**jointplot(x **=** admission\_df['GRE Score'], y **=** admission\_df['LOR '], kind **=** 'hex')











**Insights:**

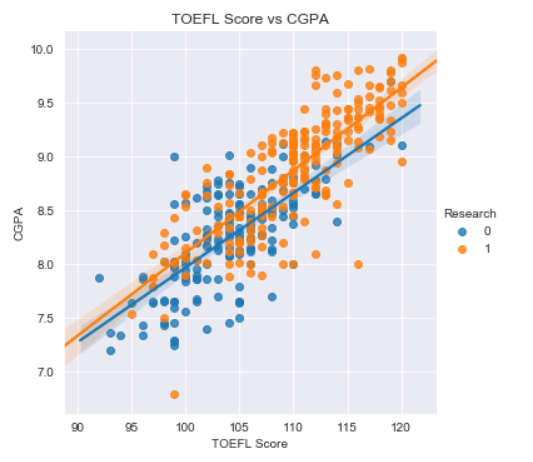
1. GRE score and TOEFL score are directly related. People with higher GRE scores tends to get high scores in TOEFL.
2. People with higher CGPA tends to get higher GRE score.
3. Various LOR scores are distributed among various GRE scores. So, GRE and LOR scores are not much related.
4. The relation between GRE score and SOP & GRE Score and University Rating are similar to the case of LOR and GRE. They are slightly related.

In [21]:

fig **=** sns**.**lmplot(x**=**"TOEFL Score", y**=**"CGPA", data**=**admission\_df, hue**=**"Research")

plt**.**title("TOEFL Score vs CGPA")

plt**.**show()



TOEFL Score and CGPA are highly related. People with high CGPA tend to get high TOEFL scores.

In [22]:

fig **=** sns**.**lmplot(x**=**"TOEFL Score", y**=**"SOP", data**=**admission\_df, hue**=**"Research")

plt**.**title("TOEFL Score vs SOP")

plt**.**show()



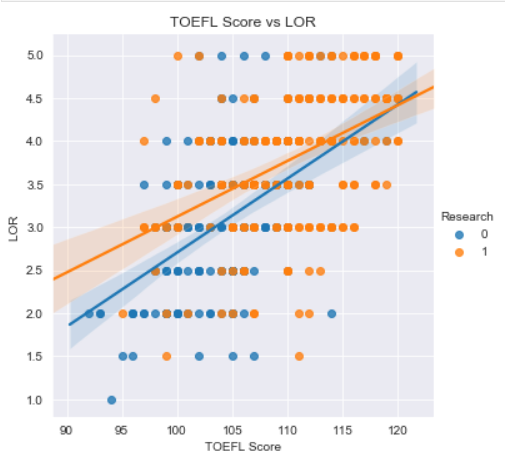
SOP and TOEFL score are not much related. Various SOP scores are distributed among various TOEFL scores.

In [23]:

fig **=** sns**.**lmplot(x**=**"TOEFL Score", y**=**"LOR ", data**=**admission\_df, hue**=**"Research")

plt**.**title("TOEFL Score vs LOR")

plt**.**show()



The case of relationship between TOEFL and LOR are the same as the case of relationship among TOEFL and SOP

In [24]:

fig **=** sns**.**lmplot(x**=**"CGPA", y**=**"SOP", data**=**admission\_df, hue**=**"Research")

plt**.**title("CGPA vs SOP")

plt**.**show()



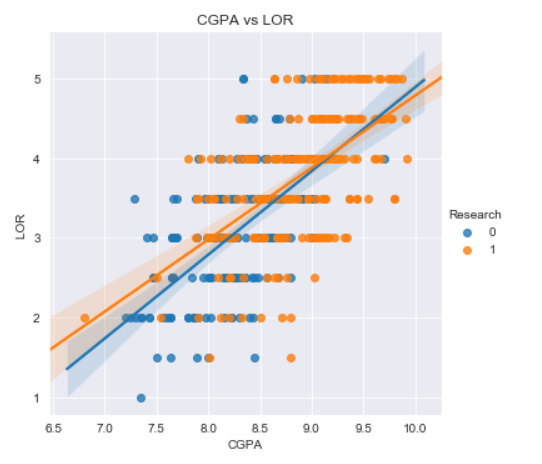
CGPA and SOP are not that much related. Since Statement of purpose(SOP) includes the accomplishments, the people with high CGPA tends to have good score of SOP.

In [25]:

fig **=** sns**.**lmplot(x**=**"CGPA", y**=**"LOR ", data**=**admission\_df, hue**=**"Research")

plt**.**title("CGPA vs LOR")

plt**.**show()



CGPA and LOR are not much related. Various CGPA scores have different kinds of LOR scores.

**Correlation between features**

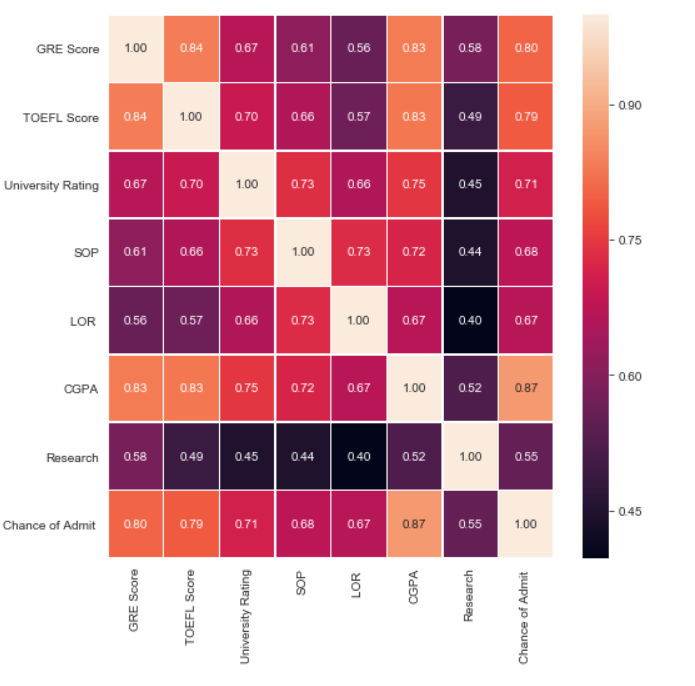
In [26]:

corr\_1 **=** admission\_df**.**corr()

fig, ax **=** plt**.**subplots(figsize**=**(8, 8))

sns**.**heatmap(corr\_1, linewidths**=**.5, annot**=True**, fmt**=**".2f")

plt**.**show()



With the correlation heatmap, the relationship between various features are easily identified.

**From this, the key takeaway is the relation between different features and their effect on chance of admit.**

The CGPA score has highest importance in the chance of admit with the correlation of 0.87, followed by GRE Score (correlation of 0.80) and then TOEFL(correlation of 0.79).

**Model fitting and prediction**

**Train-test-splitting**

In [27]:

**from** **sklearn.model\_selection** **import** train\_test\_split

X **=** admission\_df**.**drop(['Chance of Admit '], axis**=**1)

y **=** admission\_df['Chance of Admit ']

X\_train, X\_test, y\_train, y\_test **=** train\_test\_split(X,y,test\_size **=** 0.30)

**Decision Tree Model**

In [28]:

**from** **sklearn.metrics** **import** accuracy\_score

**from** **sklearn.tree** **import** DecisionTreeRegressor

**from** **sklearn.metrics** **import** mean\_squared\_error

model1 **=** DecisionTreeRegressor()

model1**.**fit(X\_train, y\_train)

predictions1 **=** model1**.**predict(X\_test)

print("Decision Tree: ",np**.**sqrt(mean\_squared\_error(y\_test, predictions1)))

Decision Tree: 0.10171856598805681

The error rate in the predictions by decision tree model is quite high.

**Other kind of Models and predictions along with error score**

In [29]:

**from** **sklearn.linear\_model** **import** LinearRegression

**from** **sklearn.ensemble** **import** RandomForestRegressor

**from** **sklearn.neighbors** **import** KNeighborsRegressor

**from** **sklearn.svm** **import** SVR

**from** **sklearn.ensemble** **import** GradientBoostingRegressor

models **=** {"Linear Regression": LinearRegression(), 'Random Forest':RandomForestRegressor(),

'KNN':KNeighborsRegressor(),'SVM':SVR(), 'GradientBoost':GradientBoostingRegressor()}

**for** model\_name, model **in** models**.**items():

predictor\_model **=** model

predictor\_model**.**fit(X\_train, y\_train)

predictions **=** predictor\_model**.**predict(X\_test)

print(str(model\_name) **+** ": "**+** str(np**.**sqrt(mean\_squared\_error(y\_test, predictions))))

Linear Regression: 0.07003277520238489

Random Forest: 0.07951608851882323

KNN: 0.08302047940116945

SVM: 0.09298628436257315

GradientBoost: 0.07599553659078155

**Insights:**

With the shown outputs of error rates in different models, it is clearly seen that the Linear Regression model has lesser error rate comparatively and hence is considered as the best model in this case. A simple model like Linear Regression performs the best with these kind of data.

**Conclusion:**

Thus, from the analysis, it is clear that the CGPA, GRE score, and TOEFL score are of high importance in getting a admit for graduate schools. These three features performs the best in predicting the chance of admit of the student.