Business Problem

Context

Jamboree has helped thousands of students like you make it to top colleges abroad. Be it GMAT, GRE or SAT, their unique problem-solving methods ensure maximum scores with minimum effort. They recently launched a feature where students/learners can come to their website and check their probability of getting into the IVY league college. This feature estimates the chances of graduate admission from an Indian perspective.

How can you help here?

Your analysis will help Jamboree in understanding what factors are important in graduate admissions and how these factors are interrelated among themselves. It will also help predict one's chances of admission given the rest of the variables.

Column Profiling:

- Serial No. (Unique row ID)
- GRE Scores (out of 340)
- TOEFL Scores (out of 120)
- University Rating (out of 5)
- Statement of Purpose and Letter of Recommendation Strength (out of 5)
- Undergraduate GPA (out of 10)
- Research Experience (either 0 or 1)
- Chance of Admit (ranging from 0 to 1)

Concept Used:

- Exploratory Data Analysis
- Linear Regression

How to begin:

- Import the dataset and do usual exploratory data analysis steps like checking the structure & characteristics of the dataset.
- Drop the unique row Identifier if you see any. This step is important as you don't want your model to build some understanding based on row numbers.
- Use Non-graphical and graphical analysis for getting inferences about variables. This can be done by checking the distribution of variables of graduate applicants.
- Once you've ensured that students with varied merit apply for the university, you can start understanding the relationship between different factors responsible for graduate admissions.
- Check correlation among independent variables and how they interact with each other.
- Use Linear Regression from (Statsmodel library) and explain the results.

  Test the assumptions of linear regression:

```
* Multicollinearity check by VIF score
* Mean of residuals
* Linearity of variables (no pattern in residual plot)
* Test for Homoscedasticity
* Normality of residuals
* Do model evaluation- MAE, RMSE, R2 score, Adjusted R2.
* Provide actionable Insights & Recommendations
```

Import Libraries

In [1]:

```
from jamboree import BasicDataChecks, ExploratoryDataAnalysis, HypothesisTesting, LinearRegression, print_format, lr_models
from sklearn.preprocessing import StandardScaler, MinMaxScaler
%load_ext autoreload
%autoreload 2
import warnings
warnings.filterwarnings('ignore')
```

Basic Data Checks - Null, Duplicate, etc.

In [2]:
```
basic_data_checks = BasicDataChecks("Jamboree_Admission.csv")
basic_data_checks.check
data = basic_data_checks.data
```

Data head:

Serial No.	gre_score	toefl_score	university_rating	sop	lor_	cgpa	research	chance_of_admit_
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

Data tail:

Serial No.	gre_score	toefl_score	university_rating	sop	lor_	cgpa	research	chance_of_admit_
495	332	108	5	4.5	4.0	9.02	1	0.87
496	337	117	5	5.0	5.0	9.87	1	0.96
497	330	120	5	4.5	5.0	9.56	1	0.93
498	312	103	4	4.0	5.0	8.43	0	0.73
499	327	113	4	4.5	4.5	9.04	0	0.84

Null Values:

```
gre_score           0
toefl_score         0
university_rating   0
sop                 0
lor_                0
cgpa                0
research            0
chance_of_admit_    0
dtype: int64
```

Duplicate Values:

0

Data Info:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 500 entries, 0 to 499
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   gre_score          500 non-null    int64
 1   toefl_score        500 non-null    int64
 2   university_rating  500 non-null    int64
 3   sop                500 non-null    float64
 4   lor_               500 non-null    float64
 5   cgpa               500 non-null    float64
 6   research           500 non-null    int64
 7   chance_of_admit_   500 non-null    float64
dtypes: float64(4), int64(4)
memory usage: 35.2 KB
None
```

Data Description:

	gre_score	toefl_score	university_rating	sop	lor_	cgpa	research	chance_of_admit_
count	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.000000	500.00000
mean	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.560000	0.72174
std	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.496884	0.14114
min	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.34000
25%	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.000000	0.63000
50%	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.000000	0.72000

```
75%    325.000000   112.000000          4.000000    4.000000    4.00000    9.040000    1.000000          0.82000
max    340.000000   120.000000          5.000000    5.000000    5.00000    9.920000    1.000000          0.97000
```

Observations

- No null values in the data set.
- No duplicate values in the data set.
- Data is available for **500** students.
- GRE scores on average are 316 with $\pm$ 11 fluctuation (which is low and good). Students have indeed scored the highest possible 340 as well. The minimum between 500 students is 290, which as a simple google search reveals is good many universities.
- TOEFL score also display a good average of 107 with $\pm$ 6 fluctuation. Minimum score is 92. As per Google, a score is anything above 90.
- University ratings out 5, have an average of 3, implying that most students indeed make it to an average rate university (rating of 3). SOP and LOR strength rating support the same fact.
- Undergrad CGPA of more than 8.5 on average is required for admits. Possible exceptions exist, wherein a GPA as low as 6.8 also exists. It would be interesting to check the strength of LOR and SOPs for such students.
- Majority of the students (~56%) have research experience of atmost 1 year.
- The chances of admit, on average to a university are 72%.

Additional Views

- It would be interesting to see, how much the strenth of the LOR and SOP affect the chances of admission, given a range of GPAs.
- Following the above, possible outliers could also be identified for students with less than average GPA, but high strength of SOP and LOR, getting into highly-ranked universities.

Data Exploration

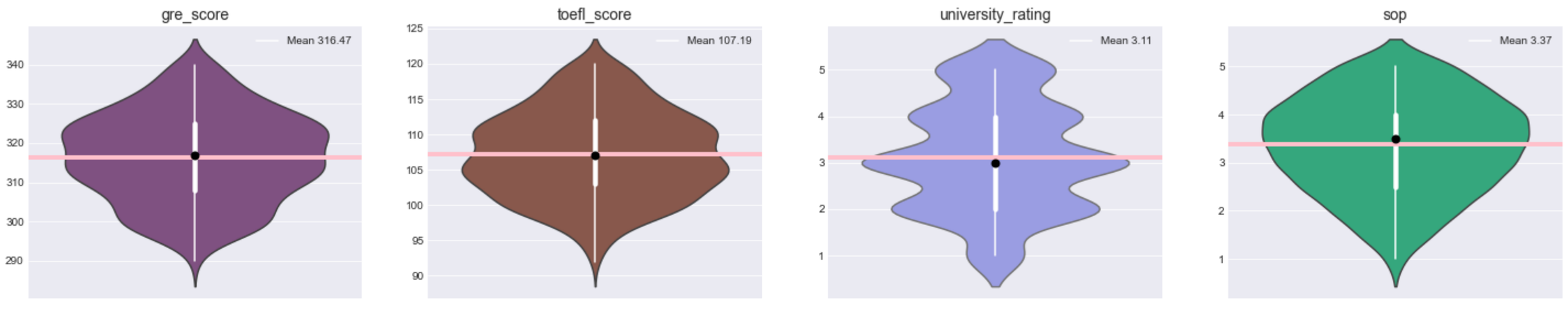
Performing univariate and bivariate exploratory data analysis with relevant comments.

Variable Distribution

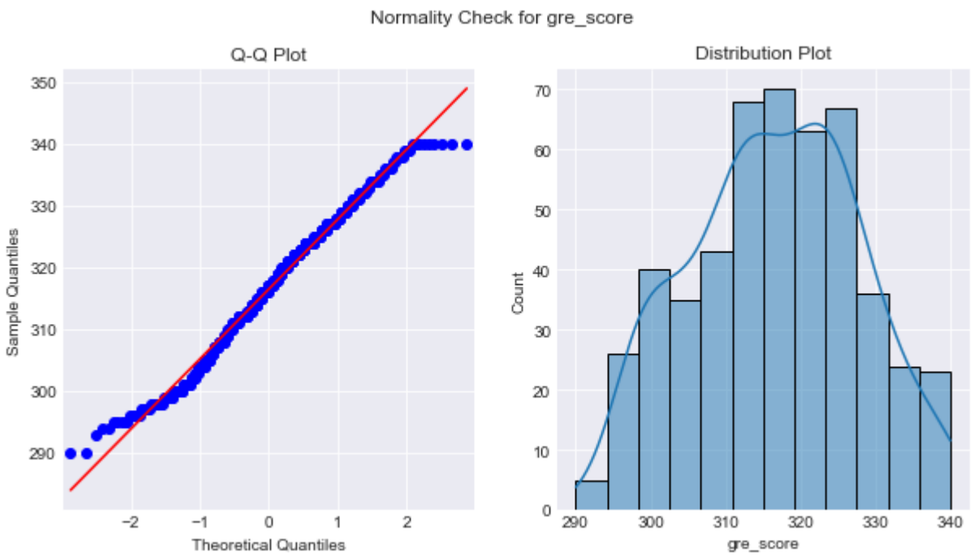
In [3]:

```
eda = ExploratoryDataAnalysis(data, [])
eda.univariate_analysis(num_lines = 160, width = 25, height = 10, ncols = 4, box=False, hist = False)
```

Violin Plot of Continuous Variables:

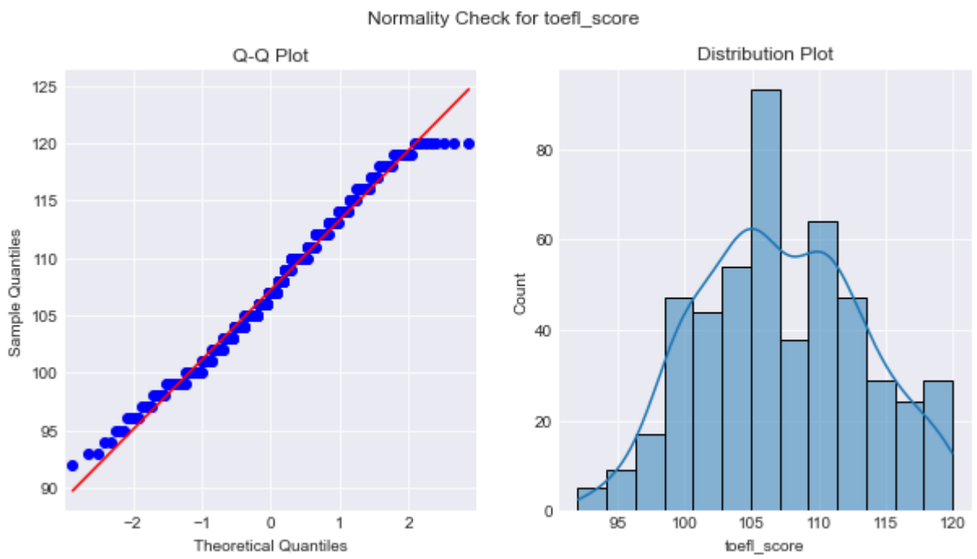


Normality Check for: gre_score



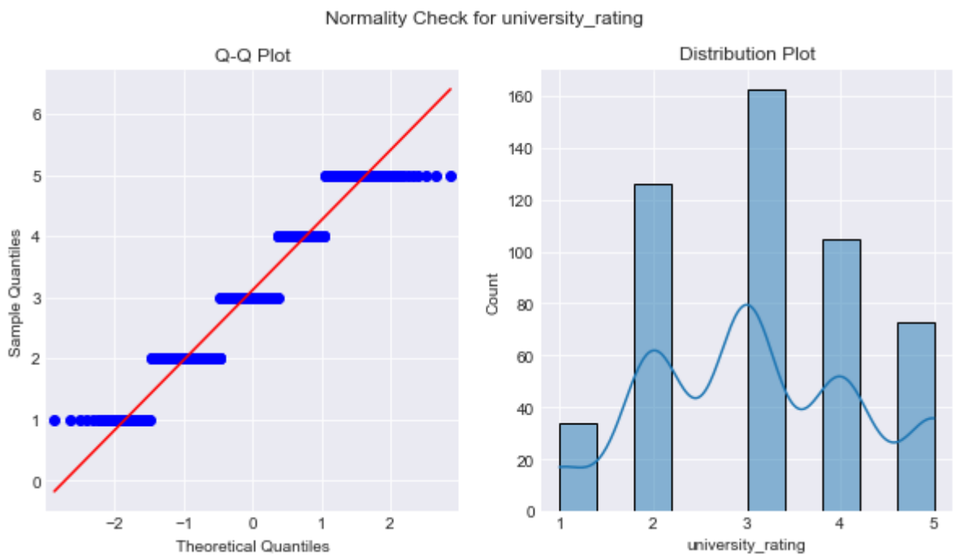
Shapiro-Wilk p-val: 0.0 | alpha: 0.05
We have sufficient evidence to say that gre_score doesn't come
from a normal distribution.

Normality Check for: toefl_score



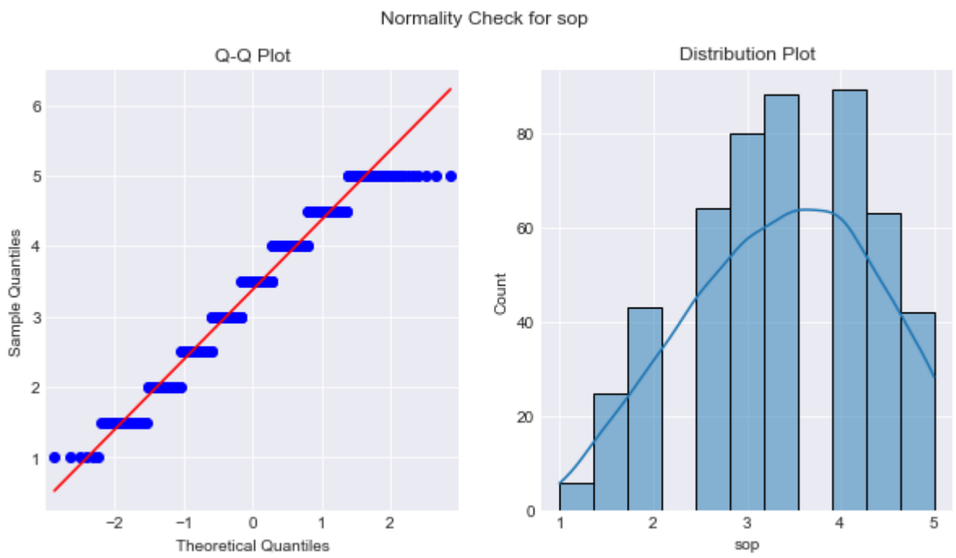
Shapiro-Wilk p-val: 0.0 | alpha: 0.05
We have sufficient evidence to say that toefl_score doesn't come
from a normal distribution.

Normality Check for: university_rating



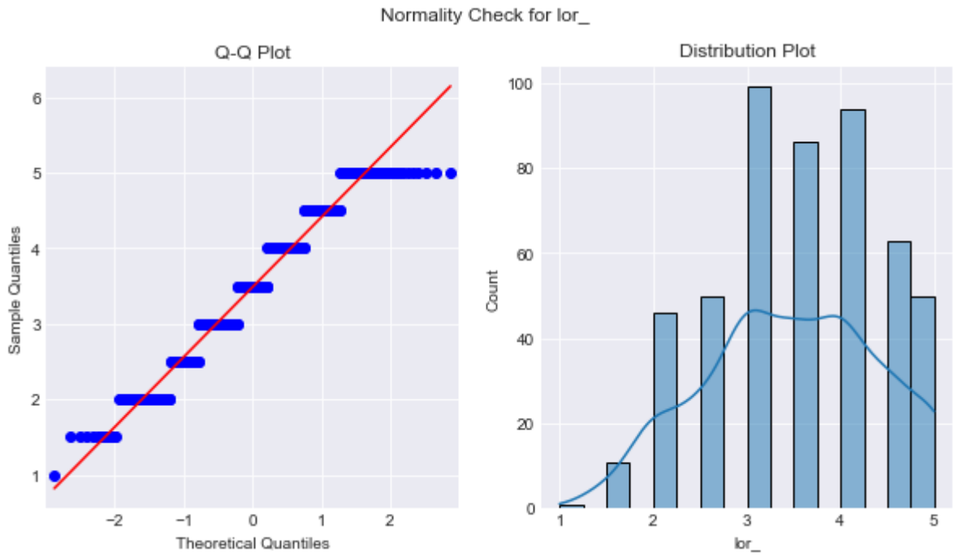
Shapiro-Wilk p-val: 0.0 | alpha: 0.05
We have sufficient evidence to say that university_rating doesn't come
from a normal distribution.

Normality Check for: sop



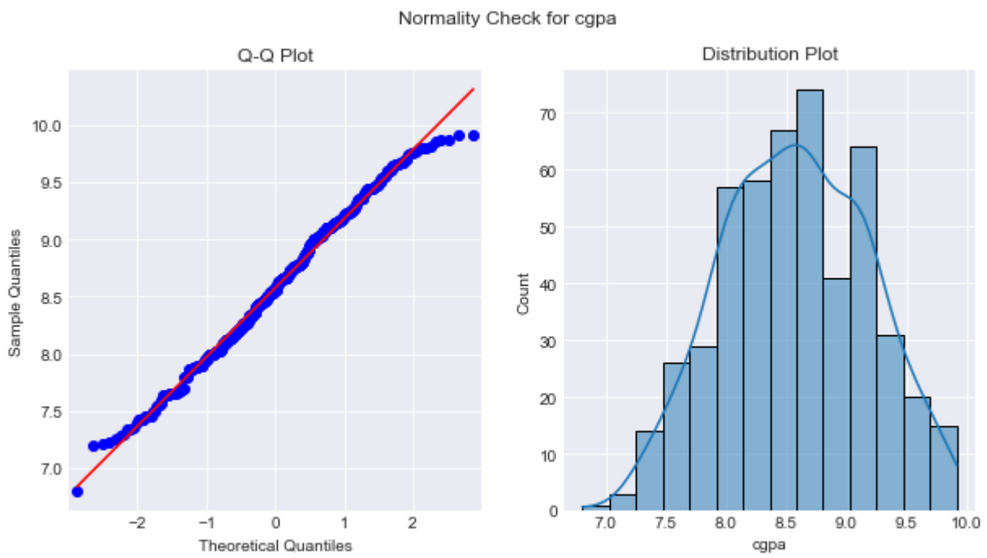
Shapiro-Wilk p-val: 0.0 | alpha: 0.05
We have sufficient evidence to say that sop doesn't come
from a normal distribution.

Normality Check for: lor_



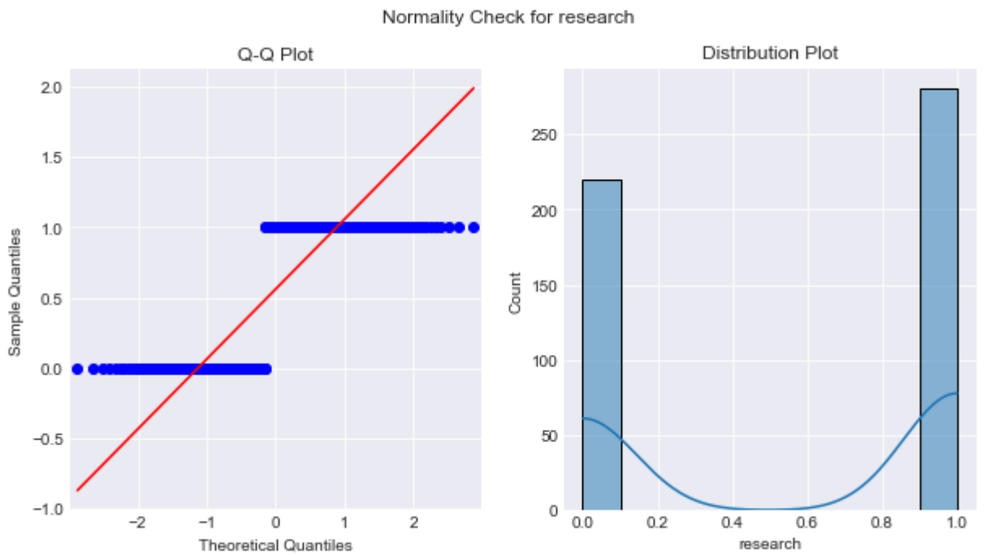
Shapiro-Wilk p-val: 0.0 | alpha: 0.05
We have sufficient evidence to say that lor_ doesn't come
from a normal distribution.

Normality Check for: cgpa



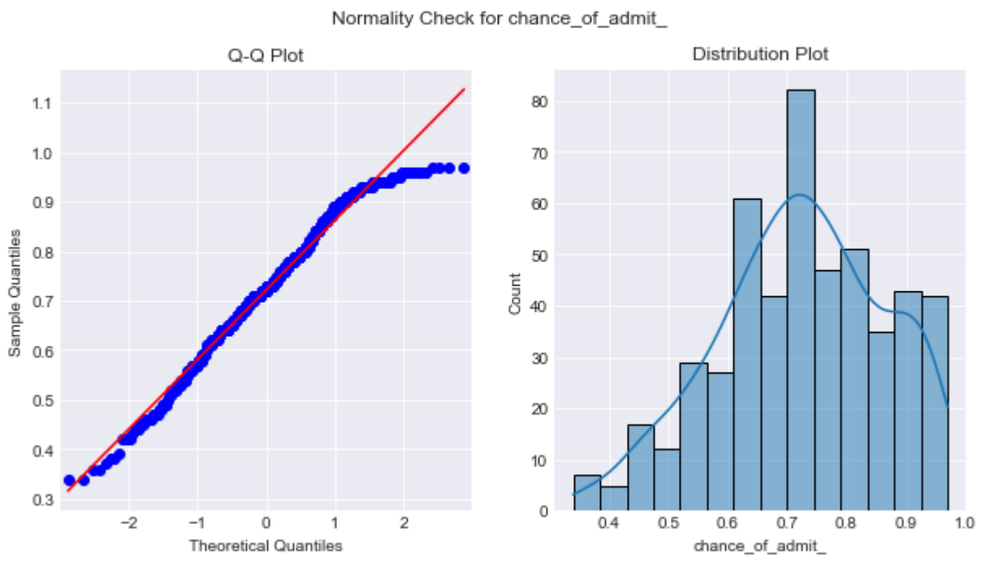
Shapiro-Wilk p-val: 0.01 | alpha: 0.05
We have sufficient evidence to say that cgpa doesn't come
from a normal distribution.

Normality Check for: research



Shapiro-Wilk p-val: 0.0 | alpha: 0.05
We have sufficient evidence to say that research doesn't come
from a normal distribution.

Normality Check for: chance_of_admit_




```
---------------------------------------------------------------------------
Shapiro-Wilk p-val: 0.0 | alpha: 0.05
We have sufficient evidence to say that chance_of_admit_ doesn't come
from a normal distribution.
---------------------------------------------------------------------------
```

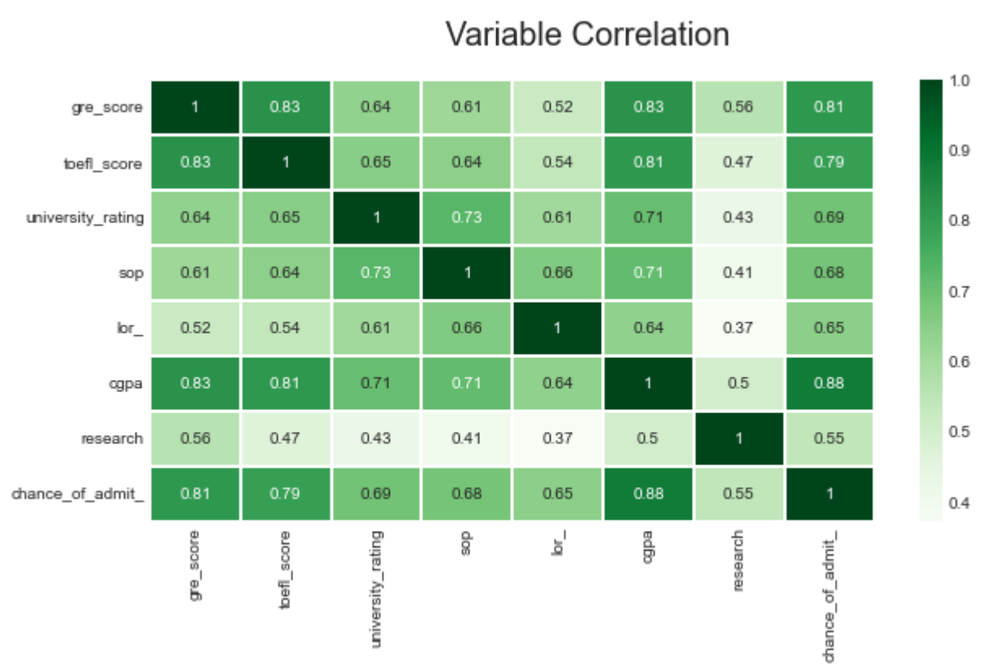
Observations

- None of the variables follow a normal distribution.
- None of the variables are affected by extreme outliers.

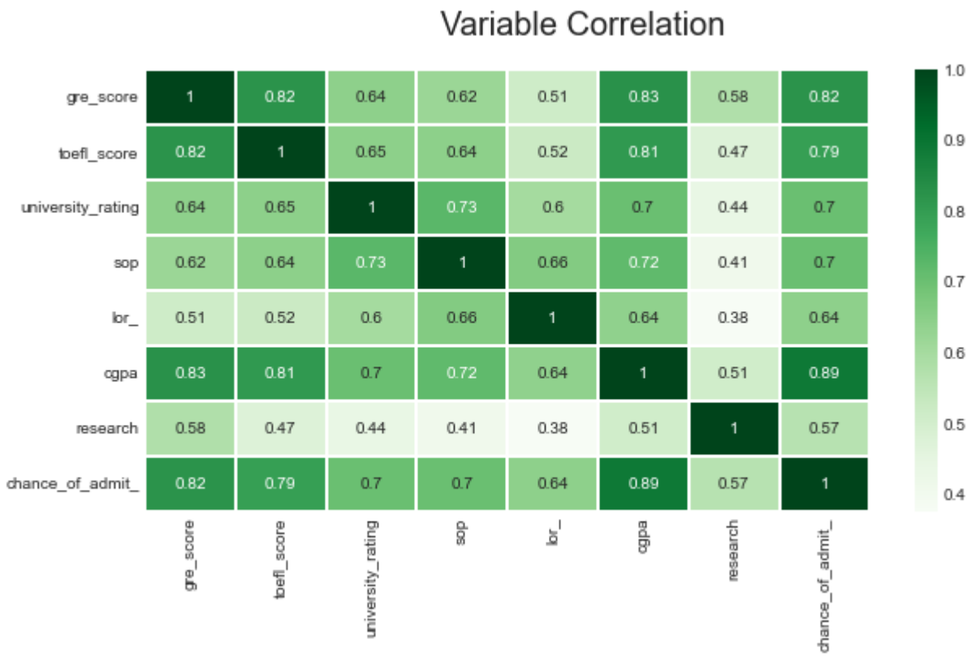
Correlation

In [4]:
```
eda.bivariate_analysis(corr=True)
```

Pearson's Correlation between Continuous Variables:



Spearman's Correlation between Continuous Variables:

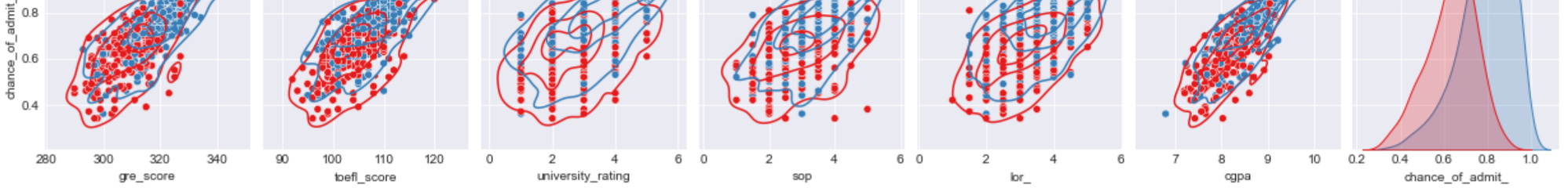


Observations

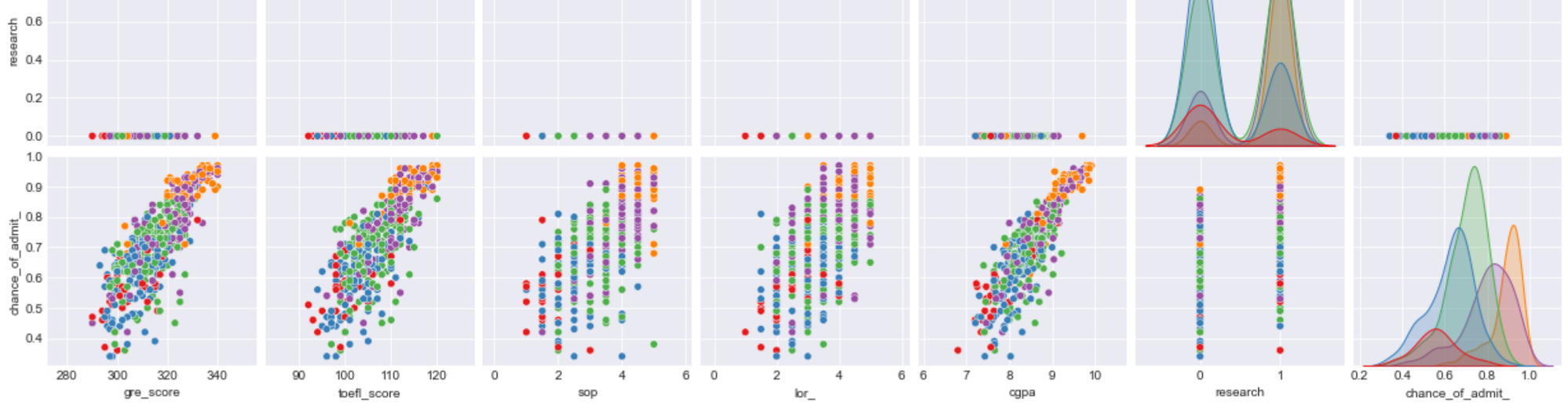
- GRE Score, TOEFL Score & CGPA have the maximum effect on admission.
- The three variables between themselves are also closely related.
- SOP is related to CGPA and and university rating.
- LOR is closely related to SOP.

In [5]:
```
eda.bivariate_analysis(pairplot=True)
```

Pairplot with Research Experience:



Pairplot with University Rating:



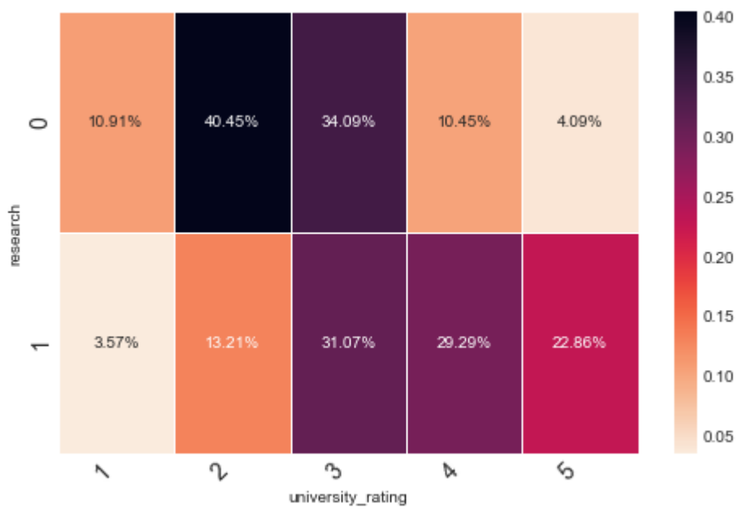
Observations

- People with higher GRE and TOEFL have done some research work. They also seem to have higher GPAs and consequently higher chances of admission.
- Between university rating and chances of admission, there seems to be a positive trend for all ratings, given research is done. Most students at university rating 5 have done research.

Contigency between University Rating and Research Experience

In [6]:
```
eda.bivariate_analysis(contigency_table=True)
```

University_Rating_and_Reasearch_Experience:



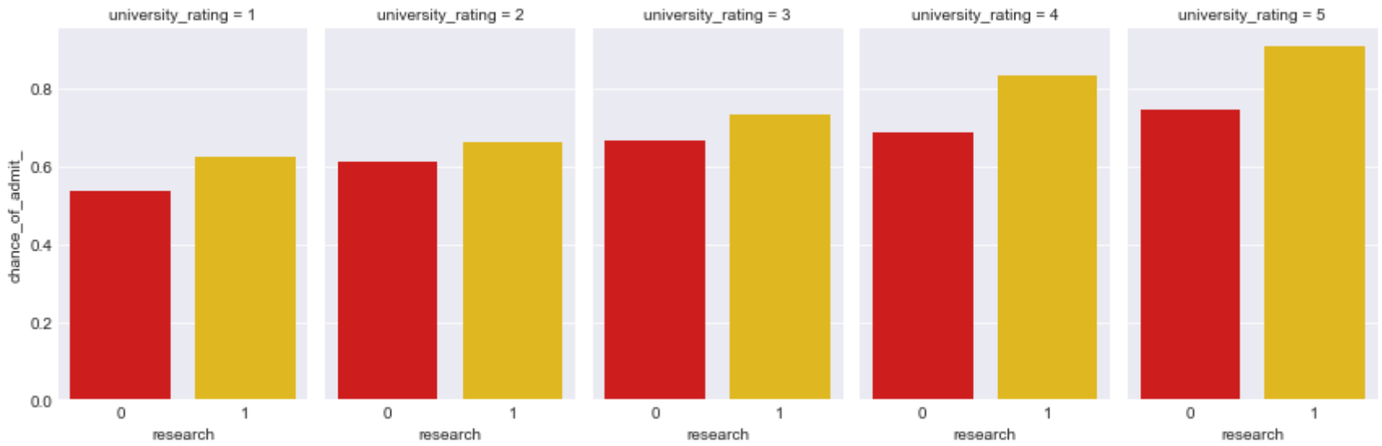
Observations

There is clear relationship positive relationship between better university rating and research done.

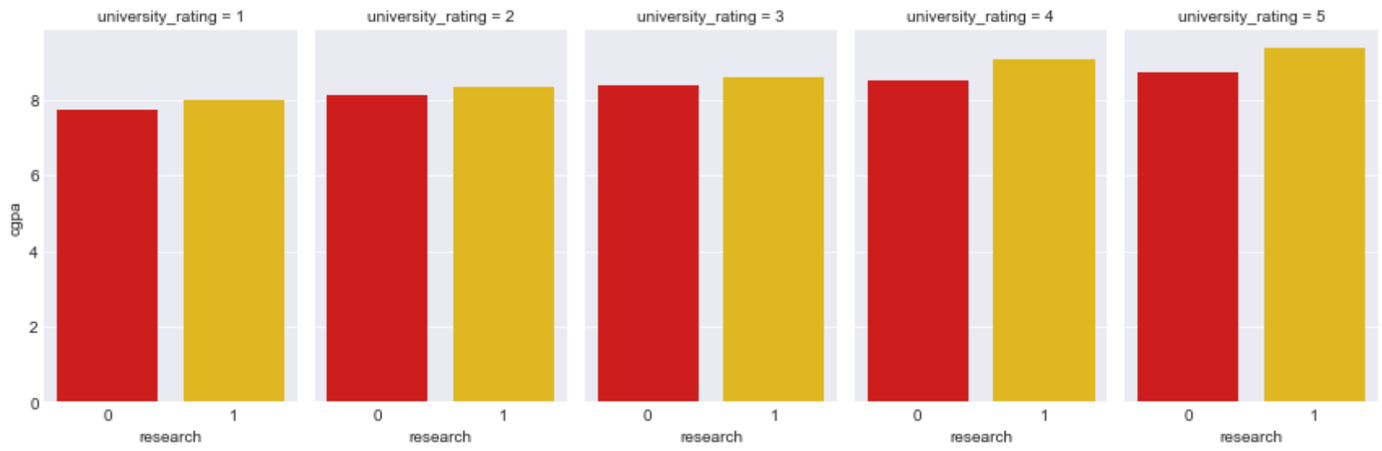
Relationship with University Rating & Research Experience between Variables

In [7]:
```
eda.bivariate_analysis(catplot=True)
```

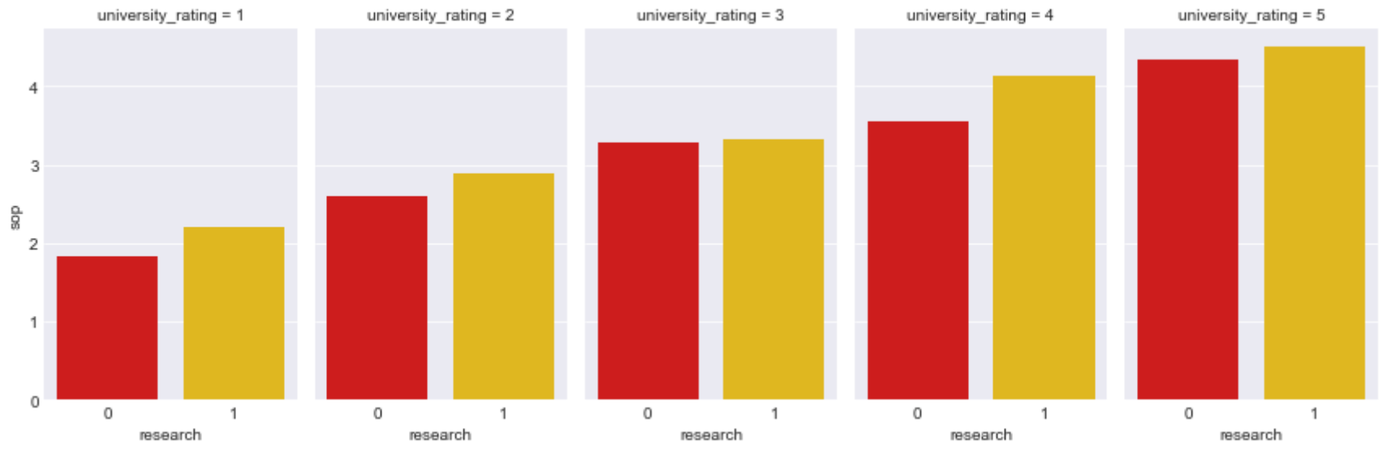
Chances of Admit Based on Research Experience for Different University Rating:



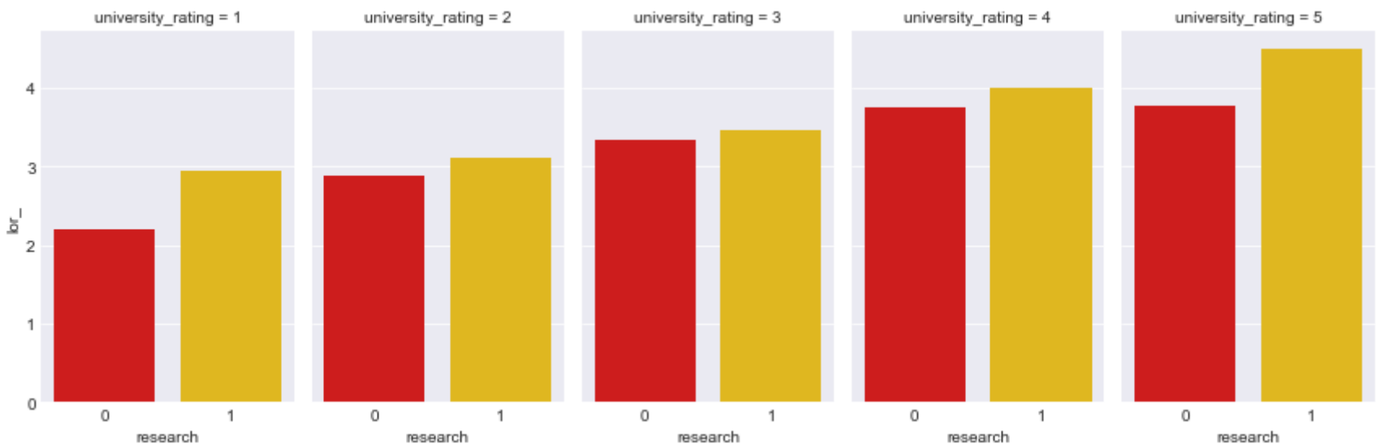
CGPA & the Relationship b/w Research Experience + Different University Rating:



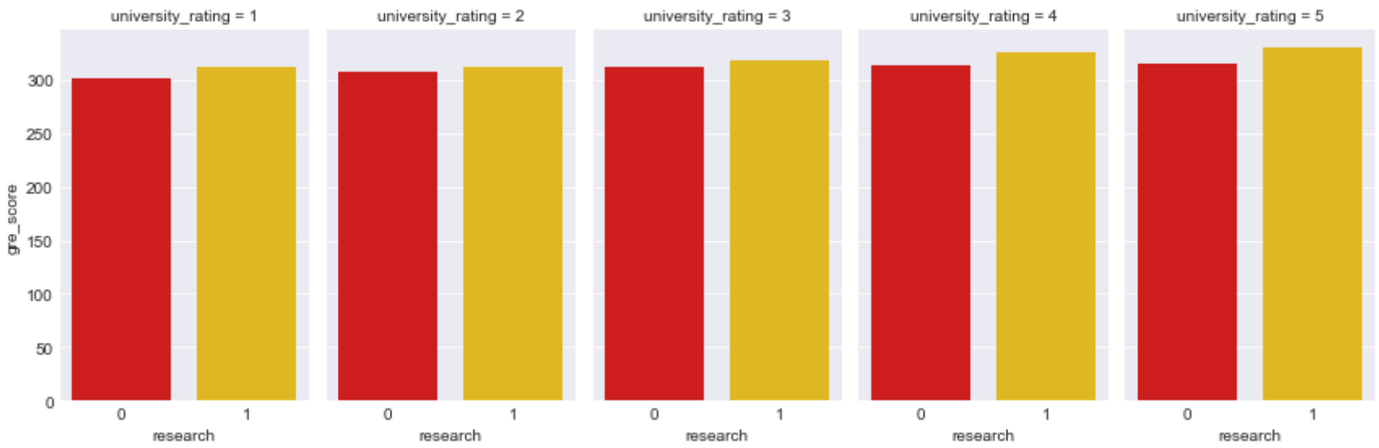
SOP Strength & the Relationship b/w Research Experience + Different University Rating:



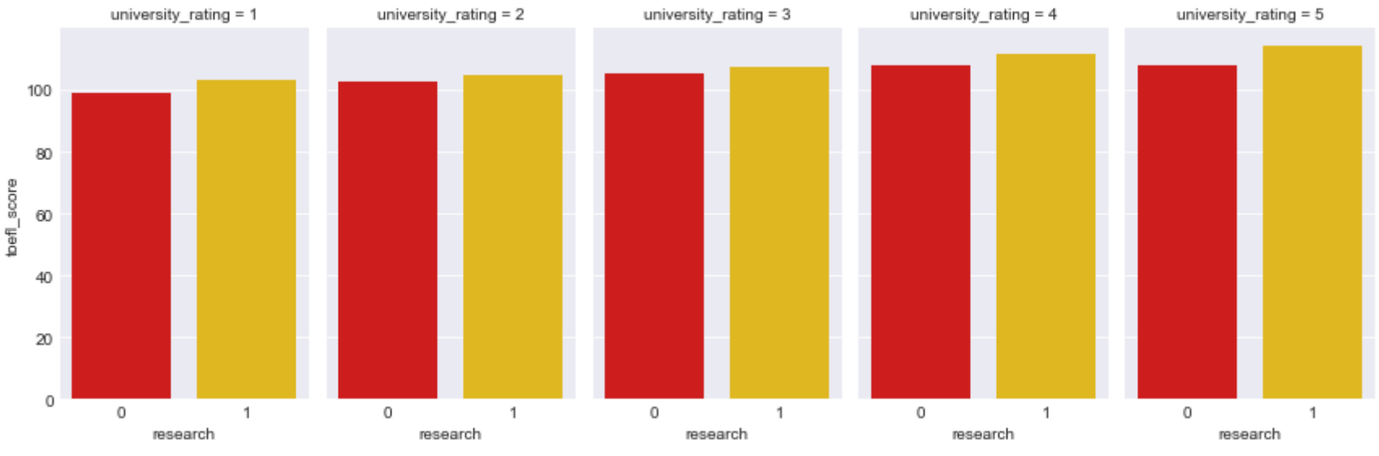
LOR Strength & the Relationship b/w Research Experience + Different University Rating:



GRE Score & the Relationship b/w Research Experience + Different University Rating:



TOEFL Score & the Relationship b/w Research Experience + Different University Rating:



Regression Modelling

Steps involved:

1. Data Pre-processing
2. Building the model
3. Testing the assumptions
4. Evaluating the performance

Data Pre-processing

1. Duplicate value check - No duplicates
2. Missing value treatment - No missing values
3. Outlier treatment - Not needed
4. Feature engineering - Scaled using Standard Scaler | Removal with VIF
5. Data preparation for modeling

In [8]:
```
reg_model = LinearRegression(data = data, endog_var = "chance_of_admit_", scale = True, scaling_method = StandardScaler)
reg_model.split_train_test(test_size=0.2, random_state = 42)
reg_model.multicollinearity_check(threshold = 4.5, remove_multi_col = True)
```

Columns Dropped: ['cgpa']

Out[8]:

	VIF	Features	status
0	4.489983	gre_score	Available
1	3.664298	toefl_score	Available
2	2.572110	university_rating	Available
3	2.785764	sop	Available
4	1.977698	lor	Available
5	4.654540	cgpa	Dropped
6	1.518065	research	Available

Building the Model

1. Build the Linear Regression model and comment on the model statistics
2. Display model coefficients with column names

In [9]:
```
trained_model = reg_model.fit_ols_regression(add_constant=True, col_to_drop=None)
reg_model.model_summary
```

Out[9]:

OLS Regression Results

Dep. Variable:	chance_of_admit_	**R-squared:**	0.771
Model:	OLS	**Adj. R-squared:**	0.768
Method:	Least Squares	**F-statistic:**	220.9
Date:	Sat, 05 Mar 2022	**Prob (F-statistic):**	1.48e-122
Time:	10:42:09	**Log-Likelihood:**	512.81
No. Observations:	400	**AIC:**	-1012.
Df Residuals:	393	**BIC:**	-983.7
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7242	0.003	213.825	0.000	0.718	0.731
gre_score	0.0543	0.007	8.317	0.000	0.041	0.067
toefl_score	0.0336	0.006	5.366	0.000	0.021	0.046
university_rating	0.0094	0.005	1.737	0.083	-0.001	0.020
sop	0.0120	0.006	2.157	0.032	0.001	0.023
lor_	0.0257	0.005	5.537	0.000	0.017	0.035
research	0.0121	0.004	2.906	0.004	0.004	0.020

Omnibus:	64.130	**Durbin-Watson:**	2.042
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	108.515
Skew:	-0.947	**Prob(JB):**	2.73e-24
Kurtosis:	4.711	**Cond. No.**	4.80

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Observations

- The Adjusted R-square indicates there that we can explain 76.8% of the chances of admissions with our current variables.
- F-statistic is simple stating that our model performs better with the independent variables than with just a constant.
- A lower AIC/BIC value indicates the model is a good fit. We should compare it with other model iterations.
- University rating as a predictor variable is statistically insignificant. We should remove it.
- The second part of the summary deals more with normality. The data clearly is not normal as proved by the results and EDA earlier.

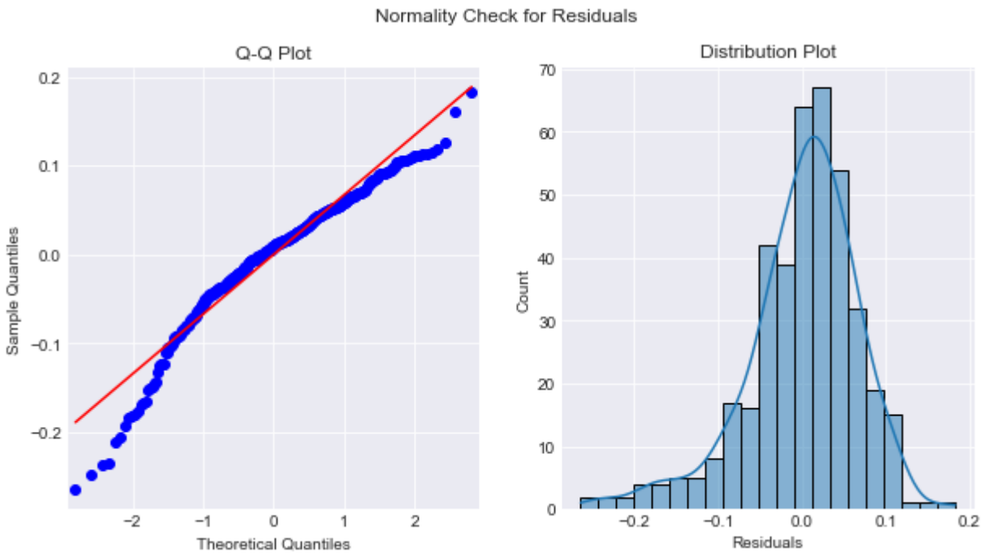
Testing the Assumptions

1. The **mean of residuals** is nearly zero
2. **Linearity of variables** (no pattern in the residual plot)
3. Test for **Homoscedasticity**
4. **Normality of residuals** (almost bell-shaped curve in residuals distribution, points in QQ plot are almost all on the line)

In [10]:
```
reg_model.residual_analysis
```

```
--------------------------------------------------------------------------------
Residuals Check
```

Normality Check for: Residuals



Shapiro-Wilk p-val: 0.0 | alpha: 0.05
We have sufficient evidence to say that Residuals doesn't come
from a normal distribution.

Goldfeld-Quandt Homoskedasticity Test
Alpha: 0.05 | pvalue: 0.49
Null: Error terms are homoscedastic | Alternative: Error terms are heteroscedastic.
We do not have sufficient evidence to reject the null hypothesis.

Observations

- Mean of the residuals is zero and there doesn't seem to be any linearity between them.
- The residuals however are not normally distributed, but they are homoscedastic.

Evaluating the Performance

1. Metrics to be checked - MAE, RMSE, R2, Adj R2
2. Train and test performances are checked
3. Comments on the performance measures and if there is any need to improve the model or not

In [11]:
```
reg_model.performance_analysis
```

Out[11]:

	Train	Test
MAE	0.05	0.054
MSE	0.005	0.005
RMSE	0.067	0.072
R^2 scikit-learn	0.7713	0.7478
Adj. R^2 scikit-learn	0.7672	0.7287

Observations

- RMSE indicates that on average we are off in our estimates by 6.7% on train data and 7.2% on test data.
- R^2 and Adj. R^2 on test data decrease in performance.

Iterations of the Model

All Variables Exist

In [12]:
```
reg_model1, trained_model1, performance1 = lr_models(data = data, endog_var="chance_of_admit_",
                                                     remove_multi_col = False, add_constant=True,
                                                     col_to_drop = None, residual_analysis = False)
print()
print(reg_model1.model_summary)
print()
print(performance1)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       chance_of_admit_   R-squared:                       0.821
Model:                           OLS   Adj. R-squared:                  0.818
Method:                Least Squares   F-statistic:                     257.0
Date:               Sat, 05 Mar 2022   Prob (F-statistic):          3.41e-142
Time:                       10:42:10   Log-Likelihood:                561.91
No. Observations:                400   AIC:                            -1108.
Df Residuals:                    392   BIC:                            -1076.
Df Model:                          7
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const                0.7242      0.003    241.441      0.000       0.718       0.730
gre_score            0.0267      0.006      4.196      0.000       0.014       0.039
toefl_score          0.0182      0.006      3.174      0.002       0.007       0.030
university_rating    0.0029      0.005      0.611      0.541      -0.007       0.012
sop                  0.0018      0.005      0.357      0.721      -0.008       0.012
lor_                 0.0159      0.004      3.761      0.000       0.008       0.024
cgpa                 0.0676      0.006     10.444      0.000       0.055       0.080
research             0.0119      0.004      3.231      0.001       0.005       0.019
==============================================================================
Omnibus:                       86.232   Durbin-Watson:                   2.050
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              190.099
```

```
Skew:                          -1.107   Prob(JB):                     5.25e-42
Kurtosis:                       5.551   Cond. No.                         5.65
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

                        Train    Test
MAE                     0.043   0.043
MSE                     0.004   0.004
RMSE                    0.059   0.061
R^2 scikit-learn       0.8211  0.8188
Adj. R^2 scikit-learn  0.8174  0.8029
```

Observations

- The performance does get a bump by including GPA, but we can still remove university rating and SOP strength.

Without University Rating and SOP

In [13]:

```
reg_model3, trained_model3, performance3 = lr_models(data = data, endog_var="chance_of_admit_",
                                                     remove_multi_col = False, add_constant=True,
                                                     col_to_drop = ["university_rating", "sop"], residual_analysis = False)

print()
print(reg_model3.model_summary)
print()
print(performance3)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       chance_of_admit_   R-squared:                       0.821
Model:                           OLS   Adj. R-squared:                  0.818
Method:                Least Squares   F-statistic:                     360.8
Date:               Sat, 05 Mar 2022   Prob (F-statistic):          1.36e-144
Time:                       10:42:10   Log-Likelihood:                 561.54
No. Observations:                400   AIC:                            -1111.
Df Residuals:                    394   BIC:                            -1087.
Df Model:                          5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.7242      0.003    241.830      0.000       0.718       0.730
gre_score      0.0269      0.006      4.245      0.000       0.014       0.039
toefl_score    0.0191      0.006      3.391      0.001       0.008       0.030
lor_           0.0172      0.004      4.465      0.000       0.010       0.025
cgpa           0.0691      0.006     11.147      0.000       0.057       0.081
research       0.0122      0.004      3.328      0.001       0.005       0.019
==============================================================================
Omnibus:                       84.831   Durbin-Watson:                   2.053
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              185.096
Skew:                          -1.094   Prob(JB):                     6.41e-41
Kurtosis:                       5.514   Cond. No.                         4.76
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

                        Train    Test
MAE                     0.043   0.043
MSE                     0.004   0.004
RMSE                    0.059   0.061
R^2 scikit-learn       0.8207  0.8155
Adj. R^2 scikit-learn   0.818  0.8036
```

Observations

- We only see a slight increase in the R^2 and Adj. R^2.

Business Actionables and Insights

Good Predictors: GRE Scores | TOEFL Scores | LOR Strength | CGPA | Research Experience

1. Students aiming for higher education need to have a good college GPA, ideally above 8.5.
2. Having some research experience boosts the chances of getting into a college but the change is not that high.
3. LORs boost the chances of admission. SOPs are not a very good predictor of college admissions. However, the quality of LORs is directly related to the quality of SOPs.
4. GRE and TOEFL scores are important to increase chances of admission.

Business Focus:

1. They should target the sophomores more specifically and motivate them to pursue some sort of research.
2. For freshman's, they should advise to keep a consistently high GPA to ensure a minimum of 8.5 or above throughout college life. If targeted, it would be easy to transition freshers to pursue some research.
3. While a good GRE & TOEFL score is needed across all universities, having high scores, increase the chance of admission to a highly ranked university. Therefore, based on the specific university aim for a student, there can be slight variation in such scores and the focus on them can be appropriately decided.
4. LOR and SOP strength closely relate to the university strength. A strong LOR has a direct relationship with a strong SOP. Since SOP is more specific to the student, Jamboree should ensure that LORs are of high quality. The stronger the LOR, the better the chances of getting into a highly ranked university.

Summary:

1. Consistent GRE and TOEFL Scores on average should be the aim and student preparation can be targeted accordingly.
2. Better GPA and good LORs give a significant boost to the type of university. LOR formats can be created for the students. Freshers to be targeted to ensure good GPA consistently.

Outcome:

If done, Jamboree with a linear regression model can predict the chances of admission with approximately 80% accuracy and 5% to 6% variation.

Improvements:

1. The relationship is not linear and non-linear models have the possibility of giving better results.
2. Variables like:
    - Internship/Work experience
    - Rating of undegrad college
    - Relationship between the degree pursued during graduation and degree to pursue for higher ed.
    - Financial background and strength

    Can improve the model to explain chances of admission.
3. Type of scores and students can actually be segregated based on the aim for a specific university/course/country by clustering techniques.