Jamboree Education - Linear Regression

1. Import the dataset and do usual exploratory data analysis steps like checking the structure & characteristics of the dataset.

- The dataset of graduate applicants was imported, containing GRE, TOEFL, CGPA, SOP, LOR, Research, and *Chance of Admit*.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

url = "https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/839/original/Jamboree_Admission.csv"
df = pd.read_csv(url)
df.head()
```

	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

- Structure checks (shape, data types, missing values, duplicates) confirmed the dataset was clean and consistent

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

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

-

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

```
np.int64(0)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Serial No.         500 non-null    int64  
 1   GRE Score          500 non-null    int64  
 2   TOEFL Score        500 non-null    int64  
 3   University Rating  500 non-null    int64  
 4   SOP                500 non-null    float64
 5   LOR                500 non-null    float64
 6   CGPA               500 non-null    float64
 7   Research           500 non-null    int64  
 8   Chance of Admit    500 non-null    float64
dtypes: float64(4), int64(5)
memory usage: 35.3 KB
```

-
- Descriptive statistics highlighted variation in applicant scores, showing diversity in merit among students.

```
df.describe()
```

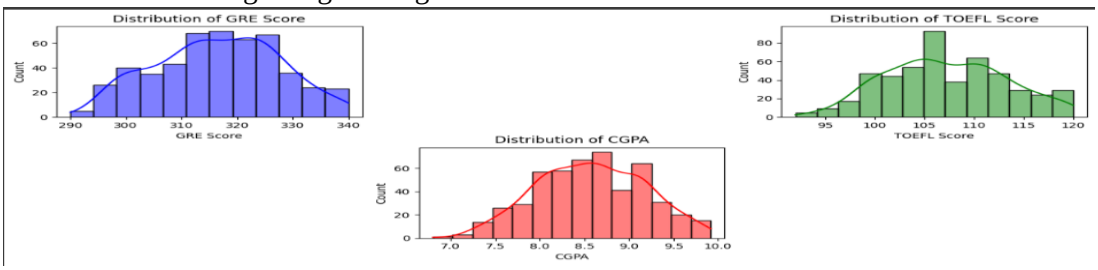
	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.000000	500.00000
mean	250.500000	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.560000	0.72174
std	144.481833	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.496884	0.14114
min	1.000000	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.34000
25%	125.750000	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.000000	0.63000
50%	250.500000	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.000000	0.72000
75%	375.250000	325.000000	112.000000	4.000000	4.000000	4.00000	9.040000	1.000000	0.82000
max	500.000000	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.000000	0.97000

2. 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
   - The column *Serial No.* was dropped as it was only a row identifier
   - Keeping this column would wrongly influence the model by treating row numbers as predictors.
   - Removing it ensured the dataset only contained meaningful features for modeling.

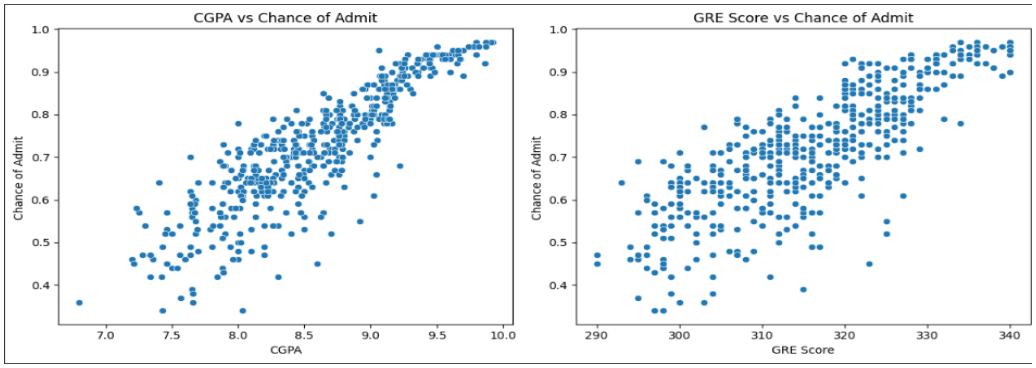
```
df.drop(columns = 'Serial No.', inplace= True)
df.head()
```

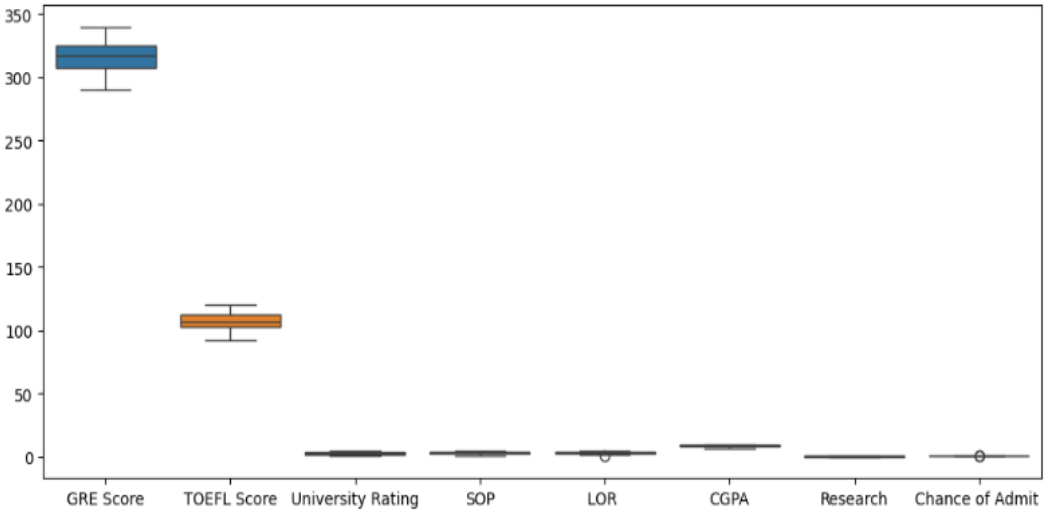
	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	337	118	4	4.5	4.5	9.65	1	0.92
1	324	107	4	4.0	4.5	8.87	1	0.76
2	316	104	3	3.0	3.5	8.00	1	0.72
3	322	110	3	3.5	2.5	8.67	1	0.80
4	314	103	2	2.0	3.0	8.21	0	0.65

3. Use Non-graphical and graphical analysis for getting inferences about variables
   - Histograms showed that GRE, TOEFL, and CGPA followed near-normal distributions, with most students scoring in higher ranges.

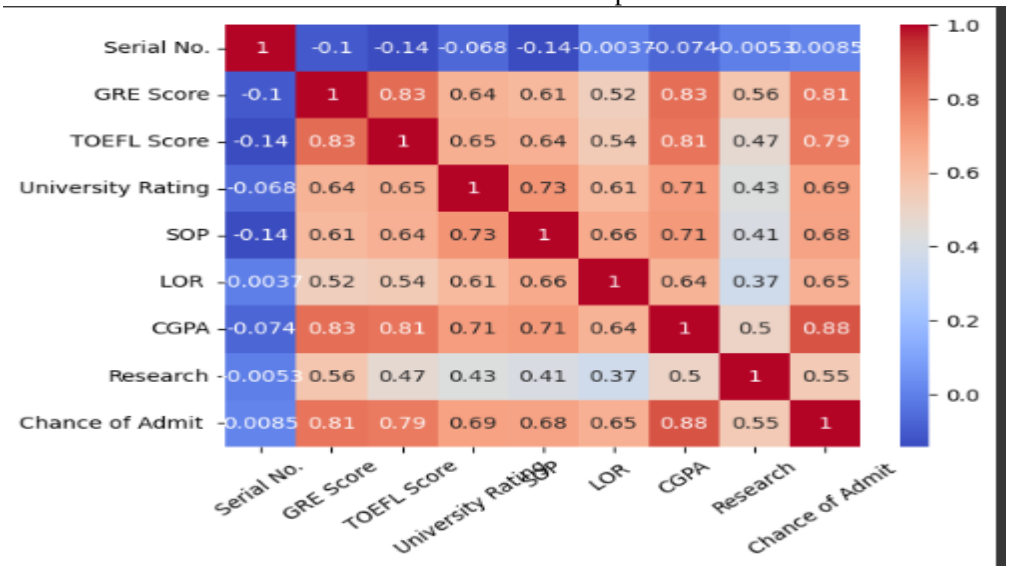


   - Scatterplots revealed positive relationships between CGPA, GRE, TOEFL, and the chance of admission.

- 

- Boxplots highlighted variation in scores, confirming applicants of varied academic merit applied.
- 

4. Check correlation among independent variables and how they interact with each other.

- A heatmap showed high correlations among GRE, TOEFL, and CGPA.
- Moderate correlations existed among LOR, SOP, and University Rating.
- Research showed weaker correlation but still added predictive value for admissions.
- 

5. Use Linear Regression from (Statsmodel library) and explain the results.

- The regression model confirmed that GRE, TOEFL, CGPA, and Research significantly influence admission chances.
- Coefficients were positive, showing higher scores and research increase the probability of admission.
- The model explained a large portion of variance in admission chances, validating its predictive strength.
-

```
df.columns = df.columns.str.strip()
X = df.drop(columns = 'Chance of Admit')
y = df[['Chance of Admit']]

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

6. Test the assumptions of linear regression:

a) Multicollinearity check by VIF score

- Initial VIF showed high multicollinearity among GRE, TOEFL, and CGPA
- Iterative dropping left *TOEFL Score* and *Research* with low VIF (<5)
- The reduced feature set improved stability of regression coefficients.
-

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

vif_data = pd.DataFrame()
vif_data['Features'] = X.columns
vif_data['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif_data
```

	Features	VIF
0	GRE Score	1308.061089
1	TOEFL Score	1215.951898
2	University Rating	20.933361
3	SOP	35.265006
4	LOR	30.911476
5	CGPA	950.817985
6	Research	2.869493

b) Mean of residuals

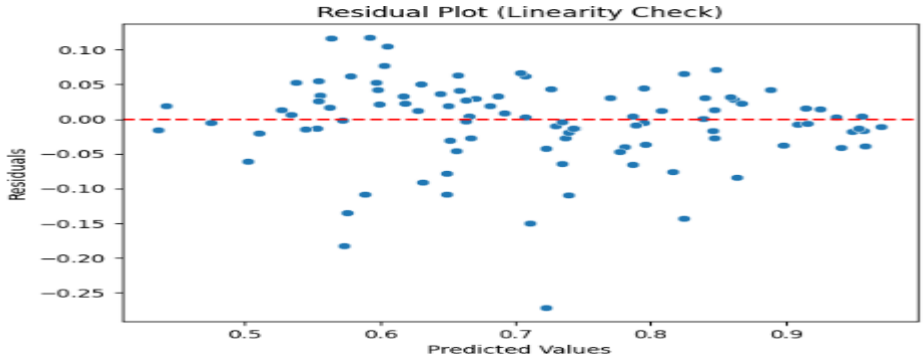
- Residuals from both train and test predictions had means close to 0.
- This satisfied the assumption of unbiased errors.
- No systematic overestimation or underestimation was found.
-

```
residuals = y_test - y_pred
print("Mean of residuals:", np.mean(residuals))
```

```
Mean of residuals: -0.005453623717661124
```

c) Linearity of variables (no pattern in residual plot)

- Residual vs predicted plots showed random scatter around zero.
- No obvious patterns (curves or funnels) were visible.
- This confirmed linearity between predictors and the dependent variable.



d) Test for Homoscedasticity

- Goldfeld-Quandt test gave p-value > 0.05.
- This indicated residuals had constant variance.
- Thus, homoscedasticity assumption was satisfied.

```
from statsmodels.stats.diagnostic import het_goldfeldquandt
import statsmodels.api as sm
from statsmodels.compat import lzip
import statsmodels.stats.api as sms

X_sm = sm.add_constant(X_train)
name = ['F statistic', 'p-value']
test = sms.het_goldfeldquandt(y_train, X_sm)
lzip(name, test)
```

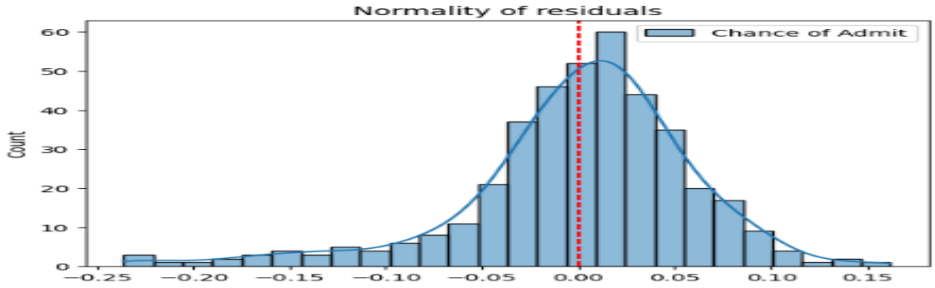
```
[('F statistic', np.float64(0.9506884043302326)),
 ('p-value', np.float64(0.6367845922443005))]
```

```
alpha = 0.05
if 0.6367845922443005 > alpha:
  print('Homoscedasticity satisfied')
else:
  print('Heteroscedasticity exists')
```

```
Homoscedasticity satisfied
```

e) Normality of residuals

- Histograms showed residuals roughly bell-shaped.
- Assumption of normal residuals was reasonably satisfied.



7. Model Evaluation (MAE, RMSE, R^2, Adjusted R^2)

- R^2 was high (>0.8) for full features, meaning ~82% of variance in admission chances was explained
- Adjusted R^2 was slightly lower but consistent, confirming model reliability.
- MAE and RMSE values were low, indicating accurate predictions.

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)

print(f' MAE is {mae}')
print(f' MSE is {mse}')
print(f' RMSE is {rmse}')
print(f' R2 is {r2}')
```

```
 MAE is 0.042722654277053636
 MSE is 0.003704655398788405
 RMSE is 0.06086588041578307
 R2 is 0.8188432567829631
```

```
n = len(y)
p = X.shape[1]

adj_r2 = 1 - (1 - r2) * (n - 1) / (n - p - 1)
print("Adjusted R²:", adj_r2)
```

```
Adjusted R²: 0.8162658234445094
```

Insights

1. **Strong Predictors** – CGPA, GRE, and TOEFL are highly correlated with admission chances, making them the most important predictors.
2. **Supporting Factors** – Research experience and quality of SOP/LOR also positively contribute, though their impact is smaller compared to academic scores.
3. **Applicant Diversity** – The dataset shows students with varied merit profiles apply, but those with strong academic and research records consistently have higher chances of admission.

Recommendations

1. **Prioritize Core Metrics** – Universities should give higher weightage to CGPA, GRE, and TOEFL while evaluating candidates, as they explain the majority of admission outcomes.
2. **Encourage Research & Projects** – Applicants with research experience or strong academic projects should be rewarded, even if their test scores are slightly lower.
3. **Adopt Data-Driven Admission Models** – Institutions can use regression-based models to set transparent cutoffs, balance academic and non-academic factors, and improve fairness in selection.