Problem statement for the US college dataset analysis:

A complete csv file with columns like private , graduation rates , Student Faculty ratio,etc of more than 500 colleges/universities.
https://www.kaggle.com/datasets/yashgpt/us-college-data/discussion

Objective:
The objective is to perform exploratory data analysis (EDA) and regression analysis to:
1. Analyze the relationship between the number of applications and the acceptance rate.
2. Develop a predictive model to estimate acceptance rates based on the number of applications received.
3. Evaluate if the regression model is goodness of fit or not.

Data Description:
The dataset contains the following variables:
- Dependent Variable:
  - Accept: The number of applications accepted by each college in US.
- Independent Variable:
  - Apps : The total number of applications received by each college in US .

Analysis Steps:
1. Create scatter plots to visualize the relationship between the number of applications and the acceptance rate.
2. Fit a linear regression model to the data to quantify the relationship between the number of applications and acceptance rate.
3. Analyze the residuals by plotting a residual vs actual plot and an actual vs fitted plot to assess the model's performance.
4. Calculate the summary statistics and coefficient of determination (R-squared) to evaluate the goodness of fit of the regression model.

By completing these steps, the admissions office will gain insights into the factors influencing acceptance rates and develop a predictive model to assist in future admissions decision-making processes.

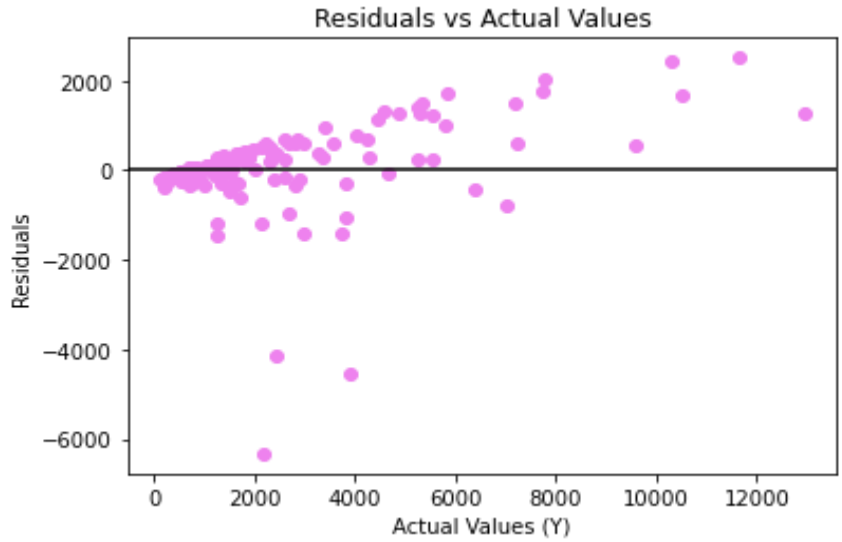
PYTHON LIBRARIES:

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

DATASET:
```
df = pd.read_csv('College_data.csv')
```

RESIDUAL VS ACTUAL PLOT:
CODE:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

data = pd.read_csv('College_Data.csv')

# Extract independent and dependent variables
X = data['Apps'].values.reshape(-1, 1)  # Reshape to 2D array for scikit-learn
Y = data['Accept'].values

# Split the data into training and testing sets
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

# Fit the regression model
model = LinearRegression()
model.fit(X_train, Y_train)

# Make predictions on the testing set
Y_pred = model.predict(X_test)

# Calculate residuals
residuals = Y_test - Y_pred

# Plot residuals vs actual values
plt.scatter(Y_test, residuals, color='violet')
plt.axhline(y=0, color='black', linestyle='-')  # Add a horizontal line at y=0
plt.title('Residuals vs Actual Values')
plt.xlabel('Actual Values (Y)')
plt.ylabel('Residuals')
Plt.show()
```



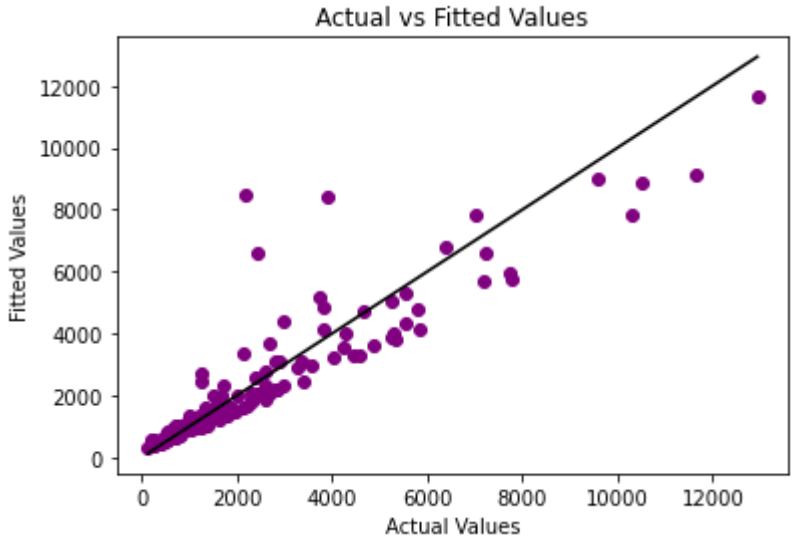
INTERPRETATION:

- A plot is such that the residuals can be contained in an **outward opening funnel** then such pattern indicates that the variance of errors is not constant, but it is an increasing function of y.
- Initially tight clustering: This indicates that for lower values of the independent variable(s), the model's predictions are relatively accurate, as evidenced by the residuals (the differences between observed and predicted values) being consistently close to zero. This suggests that the model is performing well in capturing the relationship between the variables in this range.
- Increasing scatter away from the horizontal line: As the values of the independent variable(s) increase, the spread or variability of the residuals also increases. This widening scatter implies that the model's predictive accuracy decreases as the independent variable(s) increase, indicating a violation of the assumption of constant variance, i.e., heteroscedasticity.

ACTUAL VS FITTED PLOT:
CODE:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

data = pd.read_csv('College_Data.csv')

# Extract independent and dependent variables
X = data['Apps'].values.reshape(-1, 1)  # Reshape to 2D array for scikit-learn
Y = data['Accept'].values

# Split the data into training and testing sets
```

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

# Fit the regression model
model = LinearRegression()
model.fit(X_train, Y_train)

# Make predictions on the testing set
Y_pred = model.predict(X_test)

# Plot actual vs fitted values
plt.scatter(Y_test, Y_pred, color='purple')
plt.plot([min(Y_test), max(Y_test)], [min(Y_test), max(Y_test)], color='black')  # Add a
diagonal line for reference
plt.title('Actual vs Fitted Values')
plt.xlabel('Actual Values')
plt.ylabel('Fitted Values')
plt.show()
```



INTERPRETATION:

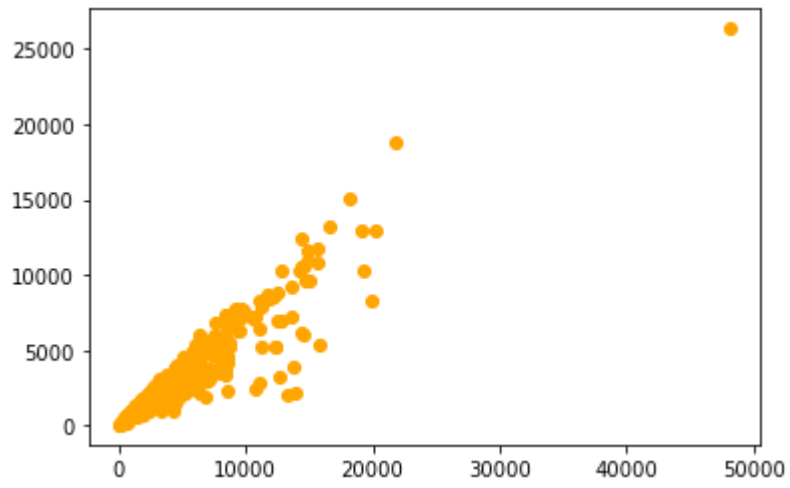
- Initial clustering along the diagonal line: This indicates that the model's predictions closely match the actual observed values for a certain range of the independent variable(s). Essentially, it suggests that the model is performing well within this range and that there is a linear relationship between the independent and dependent variables.

- Gradual scattering away from the diagonal line: As the values of the independent variable(s) increase or decrease beyond the range where the points are clustered, the scatter of the points from the diagonal line increases. This suggests that the model's predictive accuracy decreases as the independent variable(s) move away

from the range where it initially performed well. This could indicate issues like heteroscedasticity or nonlinearity.

SCATTER PLOT:
CODE:

```
# US College dataset
data = pd.read_csv('College_Data.csv')

X = data['Apps']  # Extracting the 'X' column
Y = data['Accept']  # Extracting the 'Y' column

#convert them in numpy arrays
X_values = X.values
Y_values = Y.values

#Plotting scatter diagram
plt.scatter(x, y, color='orange', label='Data')
```



INTERPRETATION:

- Tightly clustered initially: The tight clustering of points at the beginning suggests a strong relationship or correlation between the two variables being plotted. This indicates that as one variable increases or decreases, the other variable tends to exhibit a similar pattern. It suggests a certain level of predictability or consistency in the relationship between the variables within this range.

- Increasing scatter as it moves upward from left to right: As the values of the independent variable(s) increase (moving from left to right), the scatter of the points also increases. This widening scatter implies that the relationship between the variables becomes more variable or less predictable as the values of the independent

variable(s) increase. This could indicate potential issues such as heteroscedasticity, where the variability of the residuals changes across different levels of the independent variable(s).

Regression line plot:
CODE:

```
# US college data
df = pd.read_csv('College_Data.csv')
x = data['Apps']  # Extracting the 'X' column
y = data['Accept']  # Extracting the 'Y' column

# Performing linear regression
slope, intercept, r_value, p_value, std_err = stats.linregress(x, y)
line = slope * x + intercept

# Create scatter plot with regression line
plt.scatter(x, y, color='pink', label='Data')
plt.plot(x, line, color='red', linewidth=1, label='Regression Line')

#  labels and title
plt.xlabel('X')
plt.ylabel('Y')
plt.title('Regression Plot')

# Add legend
plt.legend()

# Show plot
Plt.show()
```

INTERPRETATION:

- Initial tight clustering around the regression line: This suggests that the model's predictions are relatively accurate for lower values of the independent variable(s). The tight clustering indicates that the observed values closely follow the trend predicted by the regression line within this range.

- Gradual divergence from the regression line as points move upward: As the values of the independent variable(s) increase, the points start to deviate from the regression line. This divergence could indicate potential issues such as heteroscedasticity, where the variability of the residuals increases as the values of the independent variable(s) increase.

- Potential nonlinearity or omitted variable bias: Another interpretation could be that the relationship between the independent and dependent variables is not strictly linear. The increasing deviation from the regression line as the values of the independent variable(s) increase might suggest that the relationship is more complex than initially assumed. Alternatively, it could indicate that there are important variables not included in the model that are influencing the relationship.

MODAL ADEQUACY:
REGRESSION EQUATION , REGRESSION COEFFICIENTS AND R^2 VALUE:

```
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

data = pd.read_csv('College_Data.csv')
```

```
# Extract independent and dependent variables
X = data['Apps'].values.reshape(-1, 1)
y = data['Accept'].values

# Create linear regression model
model = LinearRegression()

# Fit the model to the data
model.fit(X, y)

# Regression equation
beta_1 = model.coef_[0]
beta_0 = model.intercept_
print(f"Regression equation: y = {beta_0} + {beta_1}x")

# Regression coefficients
print(f"Regression coefficients:")
print(f"Beta knot (intercept): {beta_0}")
print(f"Beta 1 (slope): {beta_1}")

# R-squared value
y_pred = model.predict(X)
r_squared = r2_score(y, y_pred)
print(f"R-squared value: {r_squared}")

# Mean Squared Error (MSE)
mse = mean_squared_error(y, y_pred)
print(f"Mean Squared Error (MSE): {mse}")
```

OUTPUT:

```
Regression equation: y = 225.2793201678469 + 0.597515371583679
5x
Regression coefficients:
Beta knot (intercept): 225.2793201678469
Beta 1 (slope): 0.5975153715836795
R-squared value: 0.8900989818886287
Mean Squared Error (MSE): 659431.1053174739
```

INTERPRETATION:

- Heteroscedasticity: Since the variability of the residuals is not constant, it implies tha t the assumption of constant variance is violated. This could indicate that there are u naccounted-for factors influencing the relationship between the independent and de pendent variables.

- R-squared of 0.8: This indicates that the model explains a significant portion (80%) of the variability in the dependent variable. A high R-squared value suggests that the m odel fits the data well.

Conclusion:

Given the presence of heteroscedasticity, it's important to be cautious when interpreting th e results. Although the model explains a large portion of the variability in the dependent vari able, the heteroscedasticity may affect the precision of the estimates and the reliability of th e statistical inferences. Further diagnostic tests or model adjustments may be necessary to a ddress the issue of heteroscedasticity and improve the overall validity of the regression anal ysis.