EXPLORATORY DATA ANALYSIS REPORT

PALAMANICKAM

NATIONAL INSTITUTE OF TECHNOLOGY , PUDUCHERRY

INTODUCTION

- **Dataset:** Engineering Graduates Salary Data in AMCAT Job Portal from Kaggle
- **Tools:** Google colab , Jupyter notebook
- **Description :** 8 rows x 28 columns
- Exploratory data analysis on the given dataset to gain insights
- This Dataset is also available at link

OBJECTIVE

The objective of this analysis is to perform Exploratory Data Analysis (EDA) on the engineering graduates' salary dataset to identify trends and insights regarding the relationship between different factors such as degree, specialization, and various engineering aspects with salary. By exploring these relationships, the analysis aims to uncover patterns in salary distributions across different degrees and specializations, identify high-demand engineering fields, and assess the market value of various courses. Ultimately, the analysis seeks to provide valuable insights for students, educators, and industry stakeholders to make informed decisions regarding education, career paths, and workforce planning in the engineering domain.

PYTHON LIBRARIES

CODE:

```
import pandas as pd

import numpy as ap

import matplotlib.pyplot as plt

import seaborn as sns

from scipy import stats
```

Import the required libraries from Python:

Pandas is functions for handling structured data like tables and time series . Numpy for scientific computing, providing support for large arrays and matrices.matplot to project the data in animated visualizations. To create a statistical graphics on data use Seaborn lib.Scipy will be hypothesis testing, probability calculations, and descriptive statistics on the data.

PROCESS

1. DATA CLEANING
2. UNIVARIATE ANALYSIS
3. BIVARIATE ANALYSIS

DATA CLEANING

- df = pd.read_csv('/content/drive/MyDrive/marbleAI_project/Engineering_graduate_salary.csv')

  Read the csv file data using pandas as pd

- df.columns = df.columns.str.strip()

  Remove the empty spaces in the column name

- df['DOB'] = pd.to_datetime(df['DOB'])

  Change the Datatype of date of birth column to datetime dtype

- df.isnull().sum().sort_values()

  Check whether null value is present in the columns

Null value Treatment :-

Missing completely random then drop the null value rows because it is unrelated to any variables or fill the value by mean, median , mode depends upon the feature.

Missing at random or Missing Not at random use KNN imputer(machine learning algorithm) or Predictive imputation on the features of the columns.

UNIVARIATE ANALYSIS

Univariate analysis are divided into two types Categorical and Numerical analysis:

First of all, select the numerical continuous features from the columns and store in a list. Summarize the numerical features - min, max, sum, mean, median, var, std, range, iqr to identity non visual analysis. Visual Analysis - Purpose: Helps us understand the Distribution of data and Outliers - Histogram Plot, KDE Plot and Box Plot

NON VISUAL ANALYSIS

```
for column in numeric_column:
    print('stat of ',column)
    print('mean :' , df[column].mean())
    print('minimum :' , df[column].min())
    print('maximum :' , df[column].max())
    print('standard deviation :' , df[column].std())
    print('skewness :' , df[column].skew())
    print('*'*10,'end','*'*20)
    print()
```

- Extracting the mean , minimum , maximum , spread of datapoints , skewness of all the Numerical columns in the features

VISUAL ANALYSIS

```
for i in numeric_column:
  fig,axes = plt.subplots(1,3,figsize=(20,8))
  plt.tight_layout(pad=2)
  sns.boxplot(df[i],ax=axes[0],fill=True)
  axes[0].set_title(f"Box plot - '{i}'")
  axes[0].set_xlabel(i)
  sns.kdeplot(df[i],fill=True,ax=axes[1])
  axes[1].set_title(f"Probability density function plot - '{i}'")
  axes[1].set_xlabel(i)
  axes[1].set_ylabel('density')
  sns.histplot(df[i],fill=True,ax=axes[2])
  axes[2].set_title(f"Histogram plot - '{i}'")
  axes[2].set_xlabel(i)
  axes[2].set_ylabel('frequency')
  plt.show
```

We are creating a subplot with three plots for all the columns in features: a box plot, a probability density function (PDF) plot, and a histogram. The box plot displays the distribution of the data's quartiles and outliers.

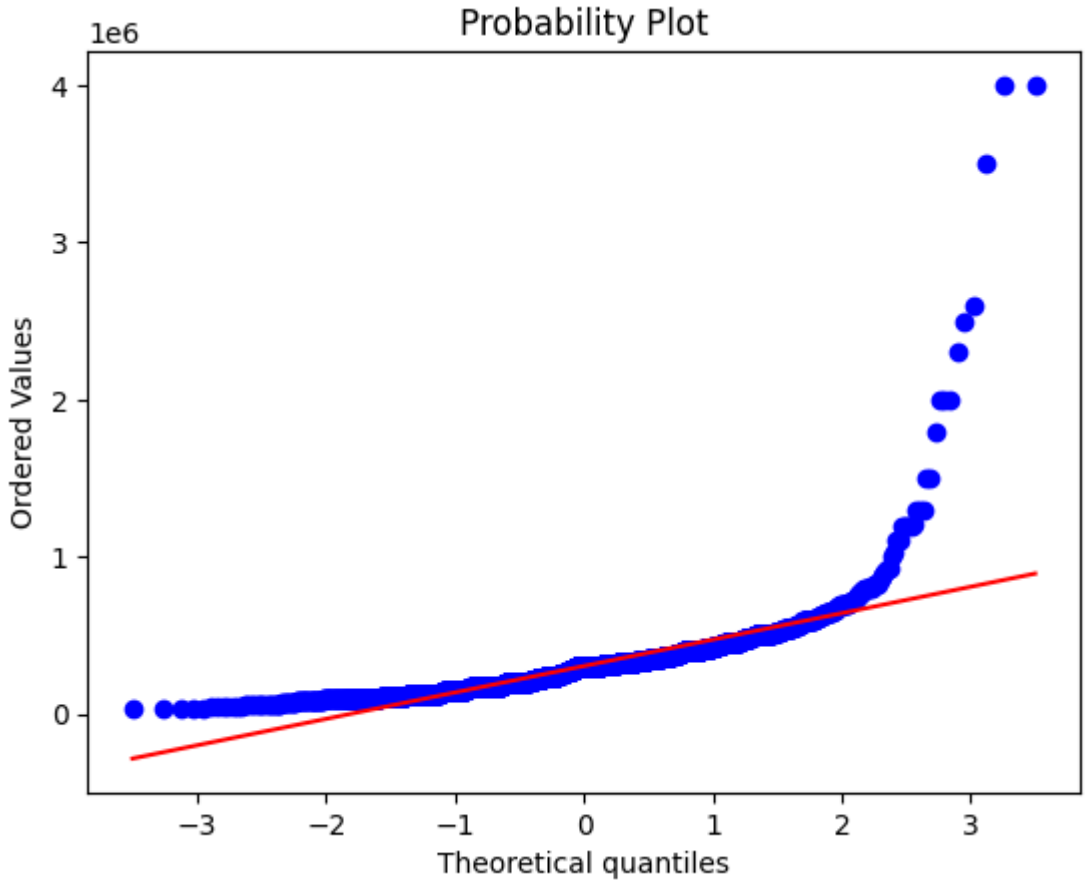
OUTLIERS TREATMENT

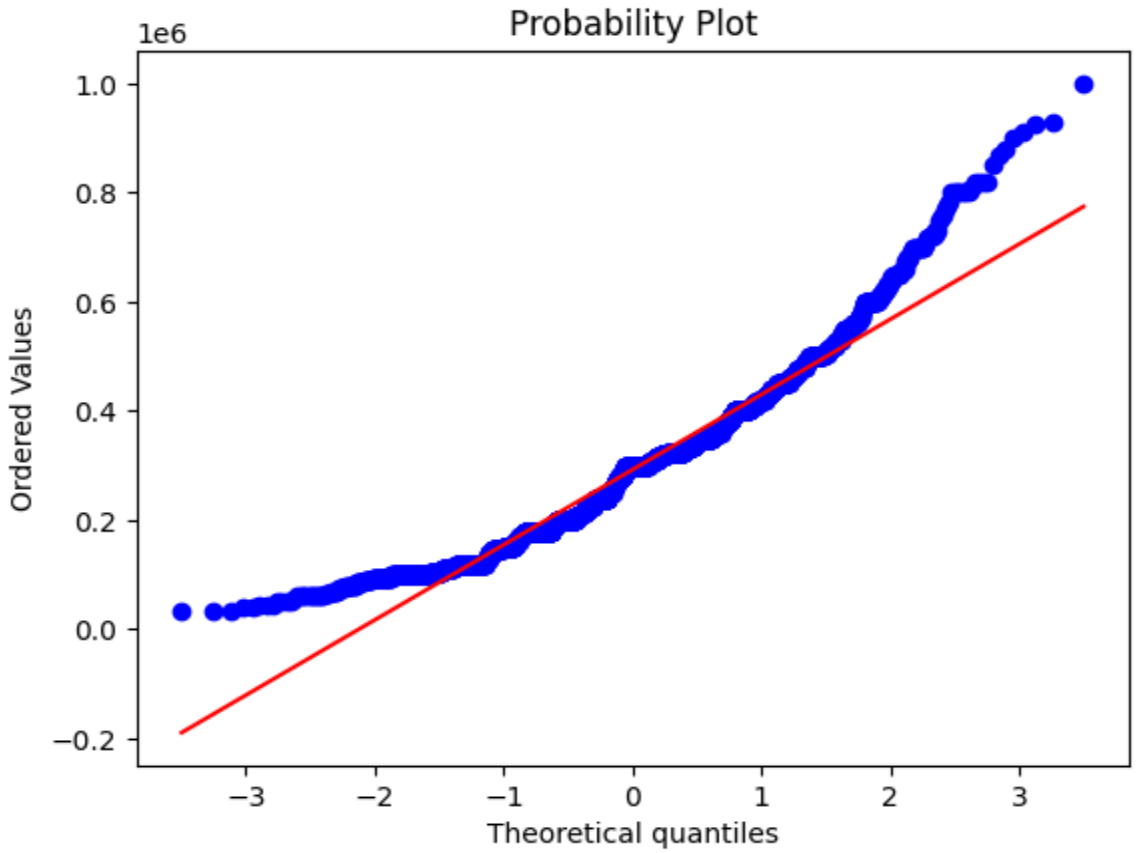
Probability plot of salary column which the values are less than 1 lakh rupees

```
df = df[df['Salary'] <= 1e6]
stats.probplot(df['Salary'],plot=plt,dist='norm')
```

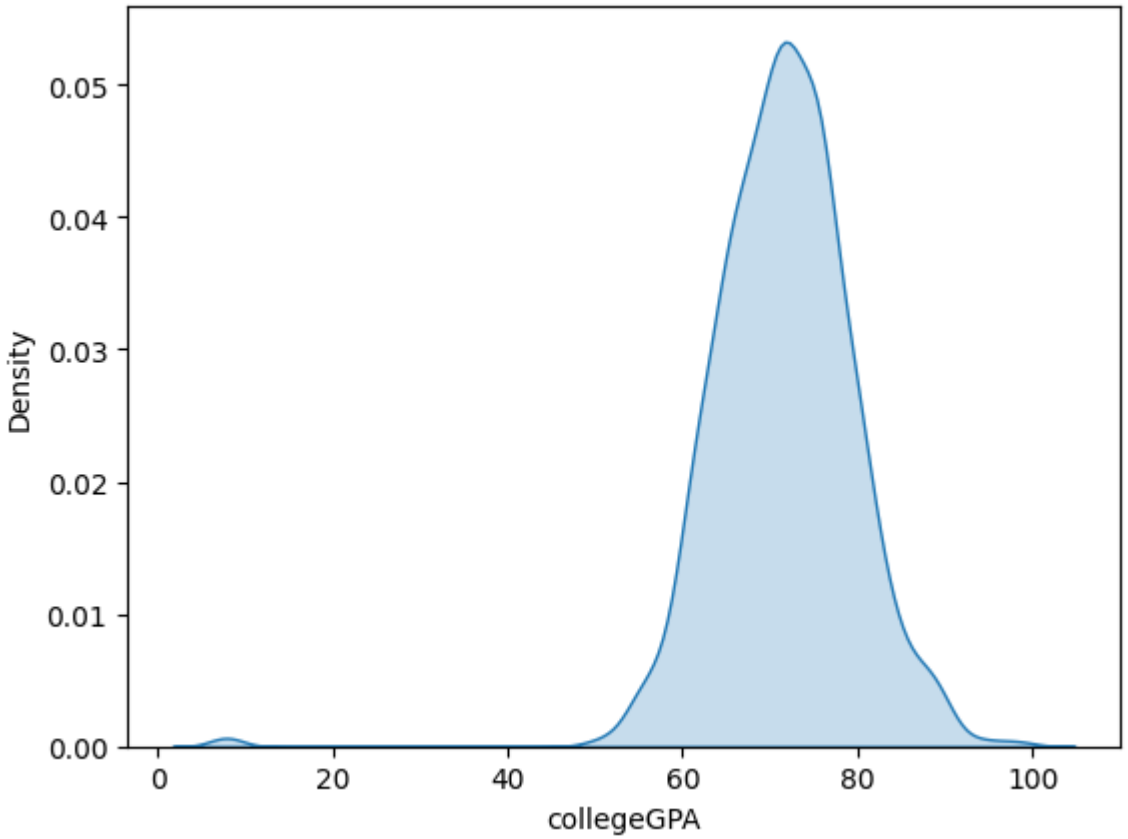
Frequency distribution plot of collage gpa column which is more than 20

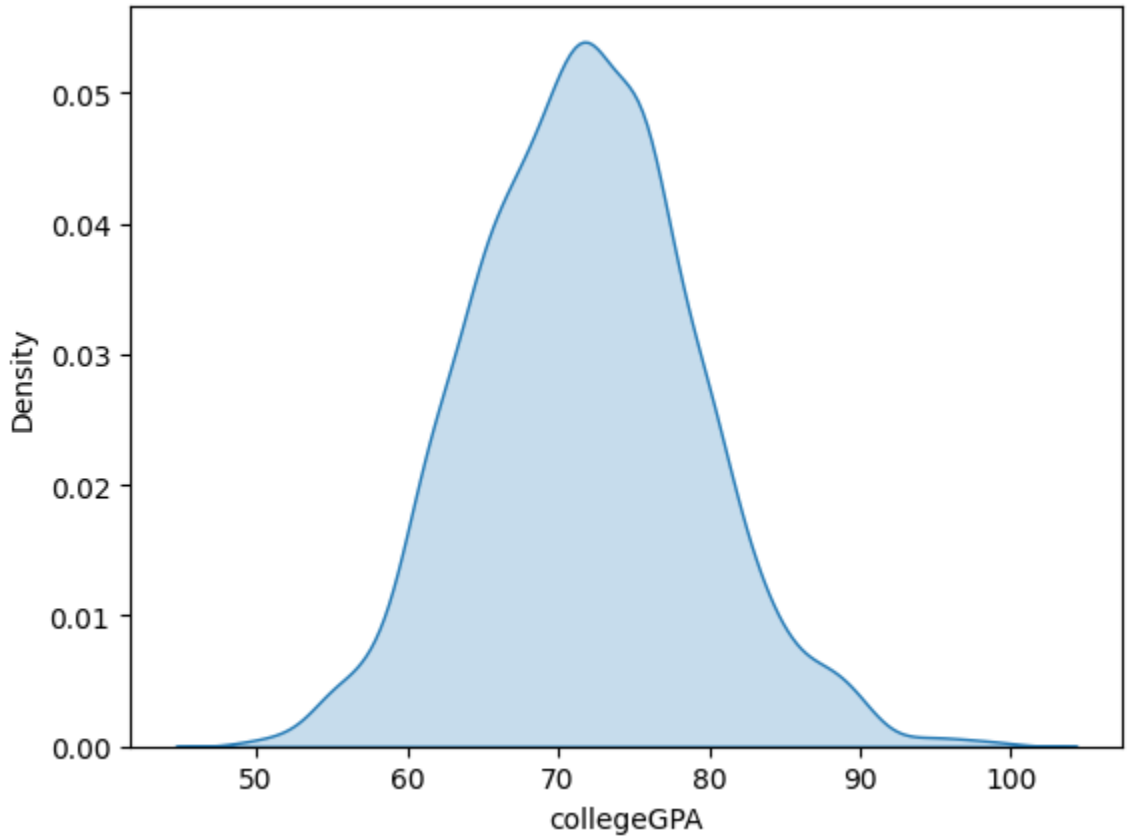
```
cleandf = df[df['collegeGPA'] > 20]
sns.kdeplot(cleandf['collegeGPA'],fill=True)
plt.show()
```





Above we can clearly see that the QQplot of Salary column before and after treatment of the outliers ,





Above we can clearly see that the KDE plot of Salary column before and after treatment of the outliers

CATEGORICAL COLUMN ANALYSIS

Statistical Non Visual Analysis

Purpose: Helps us describe and summarize the data of the discrete column

count, nunique, unique, value_counts

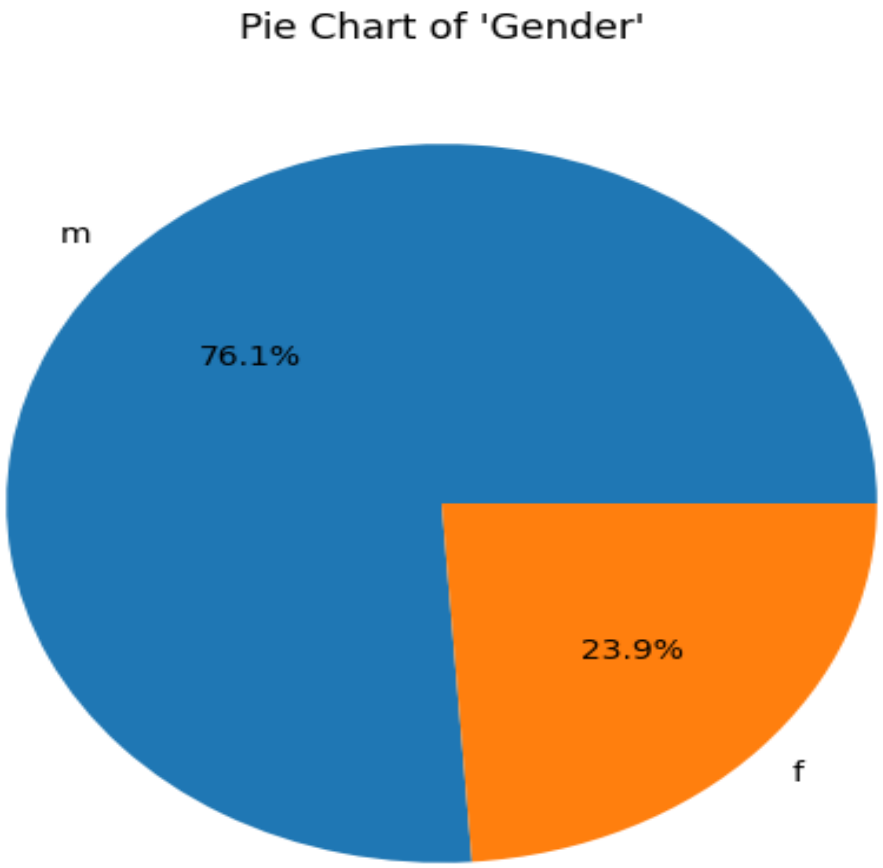
Visual Analysis

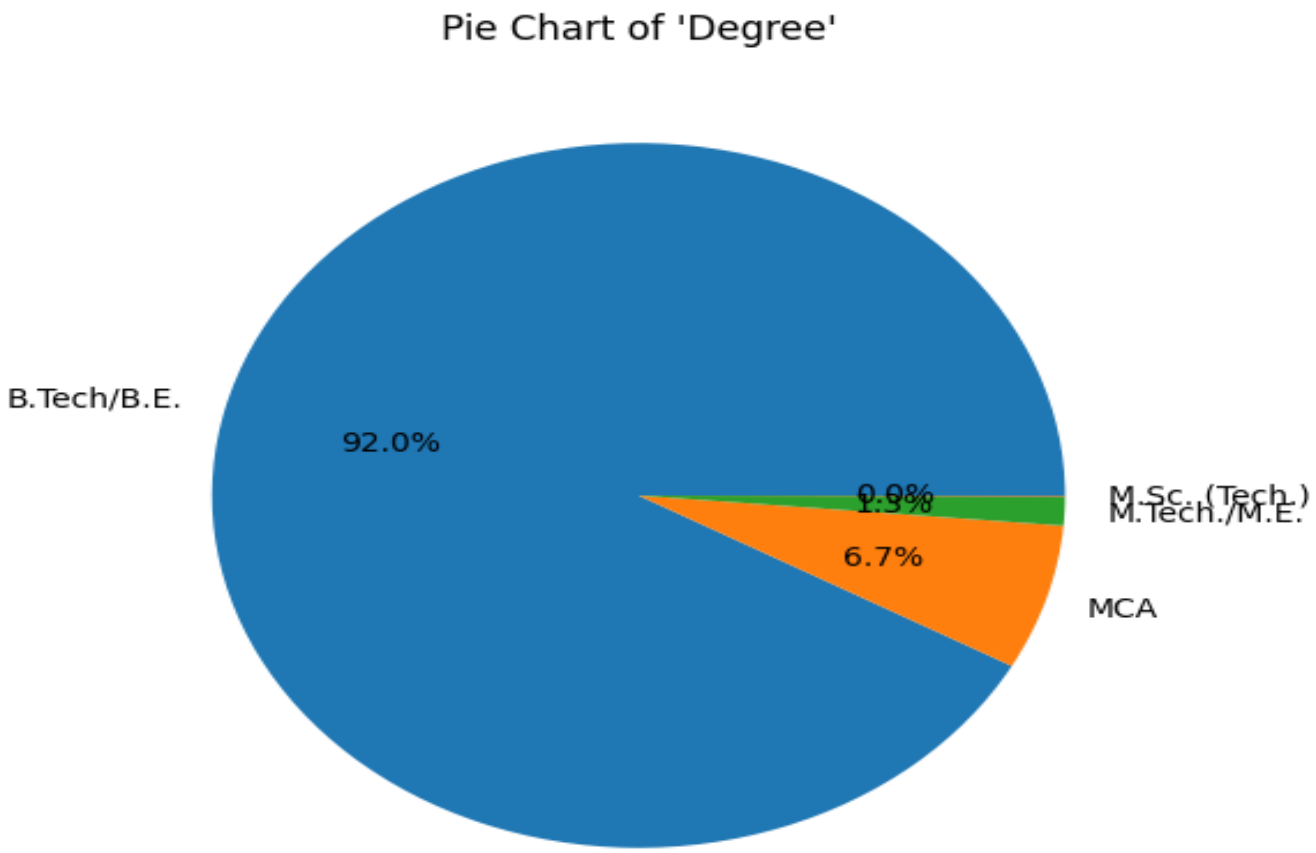
Purpose: Helps us understand how the data is distributed and Outliers

Bar/Count Plot

Pie chart

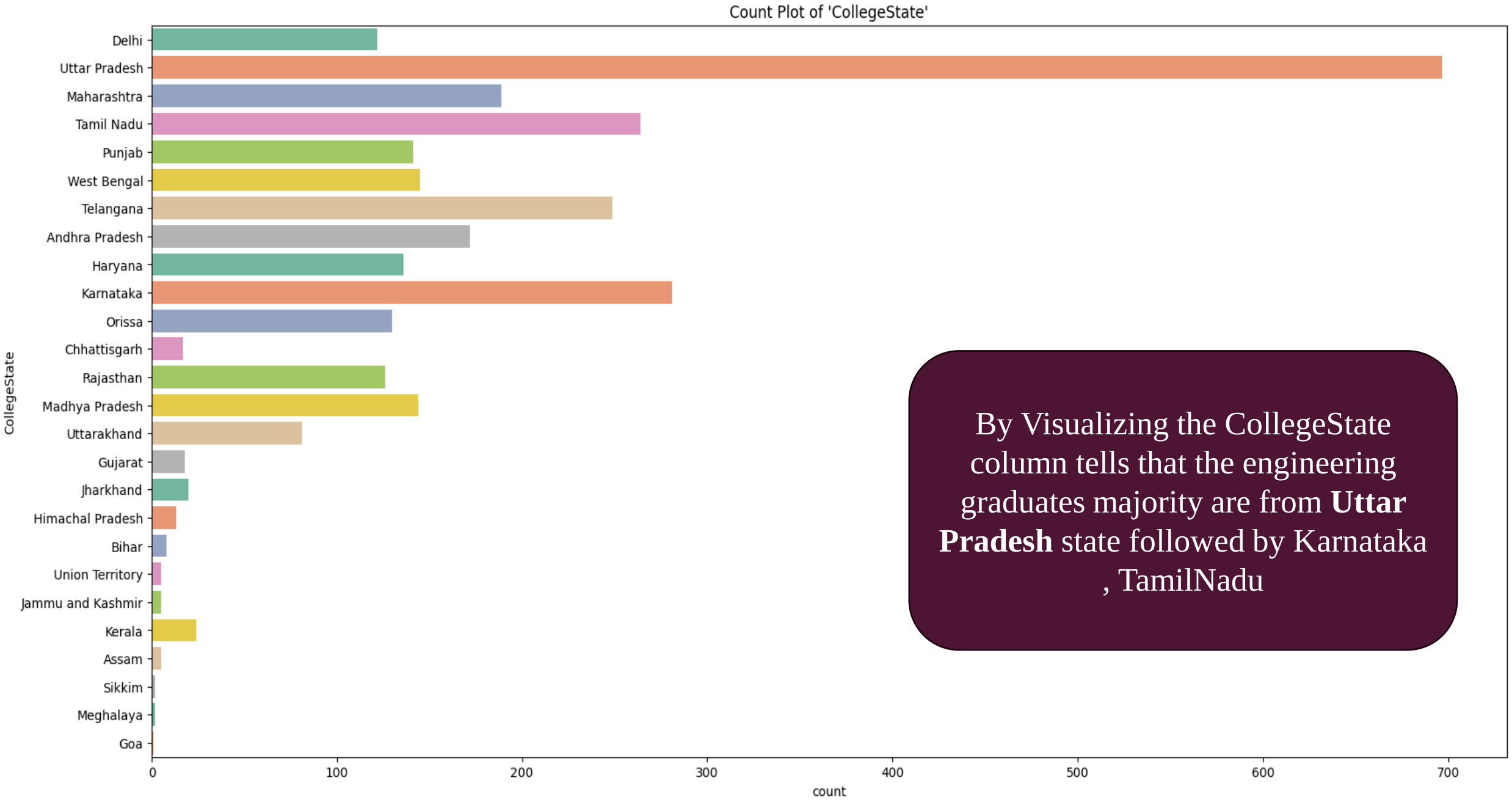
```
categorical_column = ['Gender','10board','12board','Degree','Specialization','College State']
for i in categorical_column:
  print('*'*20,i,20*'*')
  print(f'Value count of {i}:::')
  print(df[i].value_counts())
  print()
```





From the above pie chart we can identify that male are dominating in engineering field

From the above pie chart we can identify that majority of the students prefer only ug course and mostly not interested in higher studies

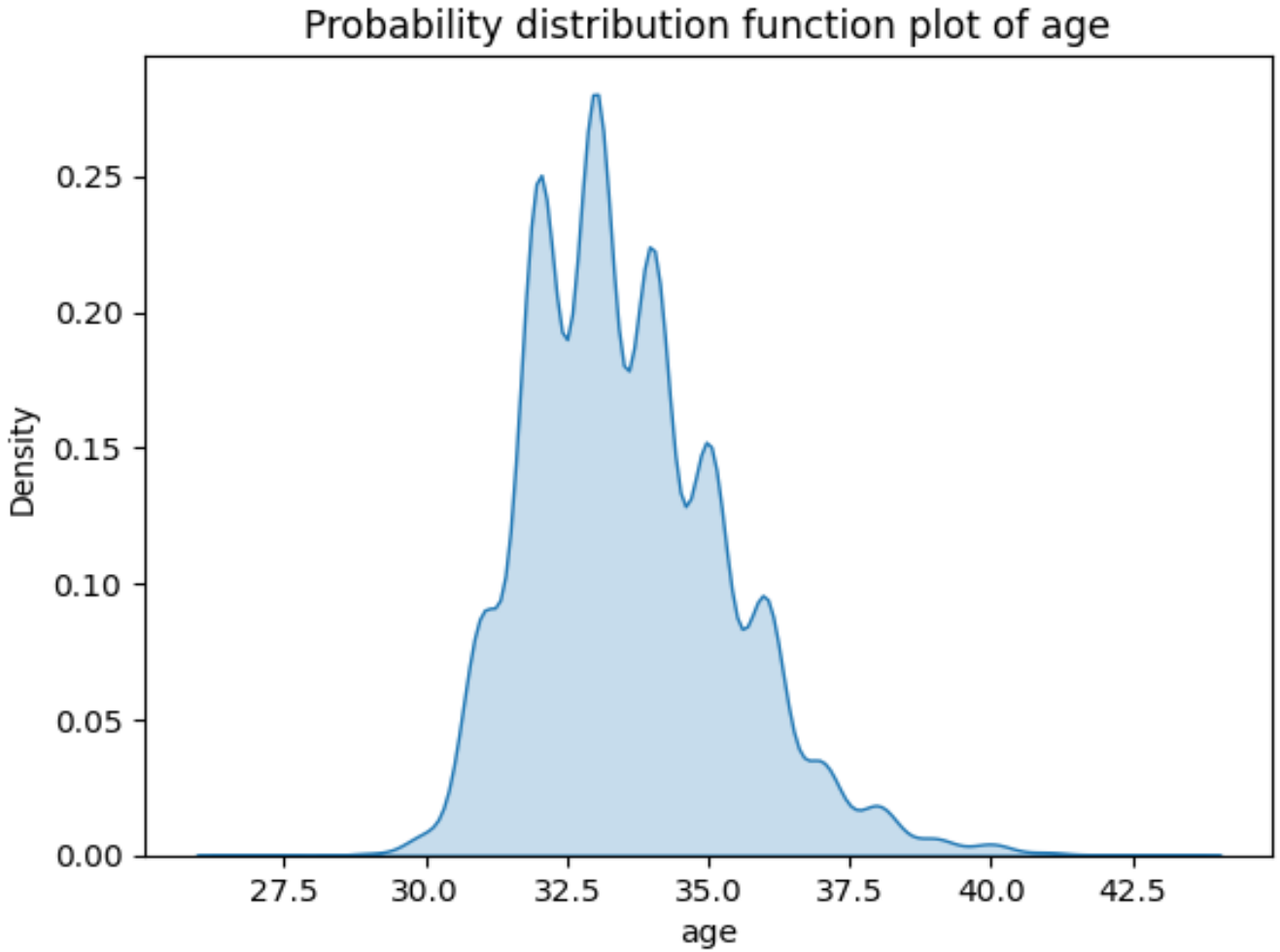


By Visualizing the CollegeState column tells that the engineering graduates majority are from **Uttar Pradesh** state followed by Karnataka , TamilNadu

EXTRACTING AGE FEATURE FROM ‘DOB’ COLUMN :

```
today = pd.to_datetime('today')
df['age'] = today.year - df['DOB'].dt.year
plt.title(f"Probability distribution function plot of age")
sns.kdeplot(df['age'],fill=True)
plt.show()
plt.pie(df['age'].value_counts(),
labels=df['age'].value_counts().index, autopct='%1.1f%%')
plt.title(f"Pie Chart of age")
```

The Age of the graduates are falling in the range between 30 to 40 years

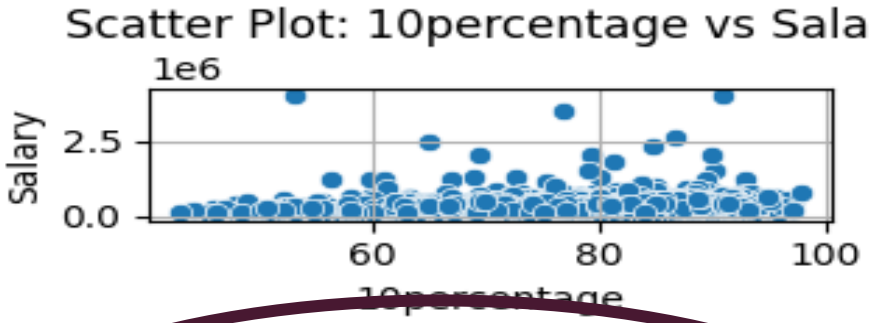


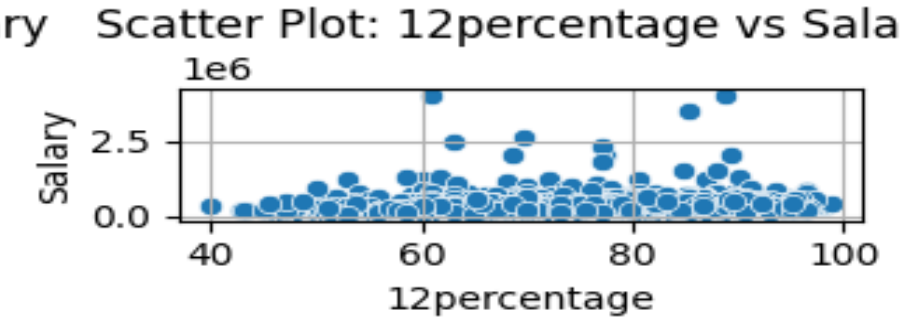
BIVARIATE ANALYSIS

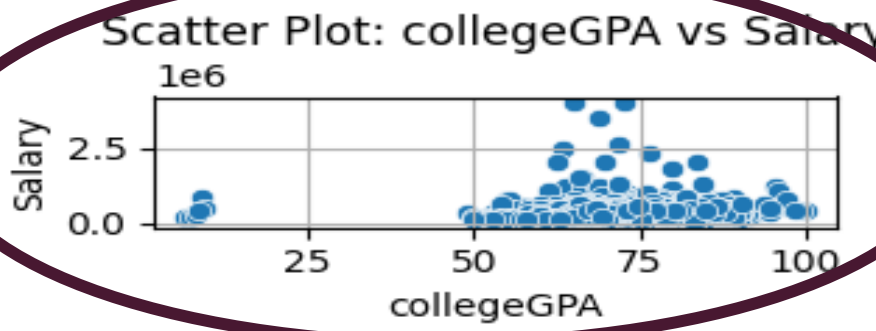
- Here , we going to fix the salary as the target column and perform visualization on salary and other numerical and categorical column in the data.
- Box plot – Numerical vs categorical data
- Scatter plot – Numerical vs Numerical data
- Bar plot – Categorical vs Categorical data

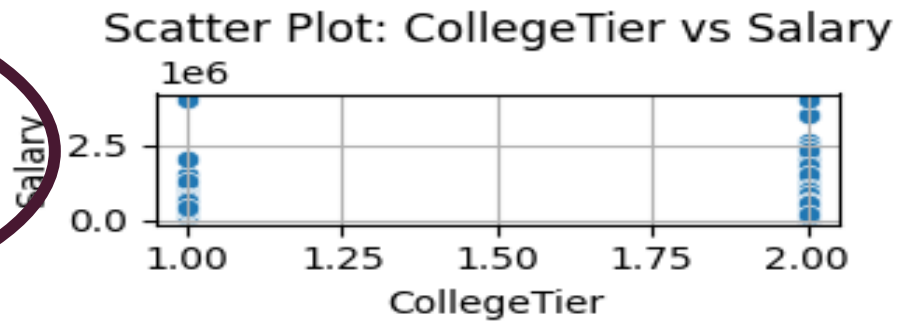
```
numeric =
['10percentage','12percentage','collegeGPA','CollegeTier','GraduationYear','age']
for j, i in enumerate(numeric, 1):
        plt.subplot(3,2, j)
        sns.scatterplot(x=i,y='Salary',data=df)
        plt.title(f"Scatter Plot: {i} vs Salary")
        plt.xlabel(i)
        plt.ylabel("Salary")
        plt.grid(True)
plt.tight_layout()
plt.show
```

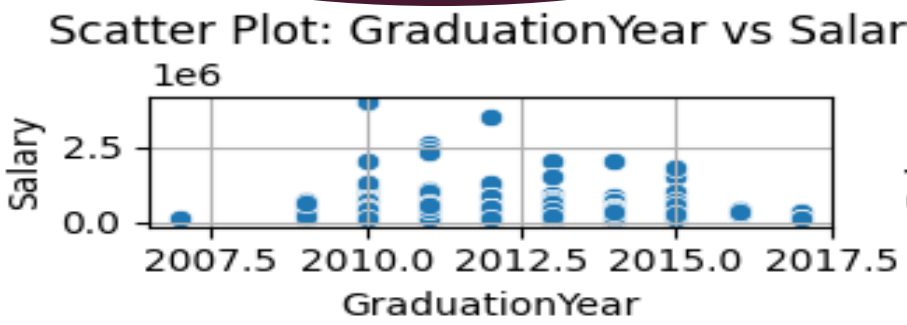
Scatter plot of salary vs numerical columns

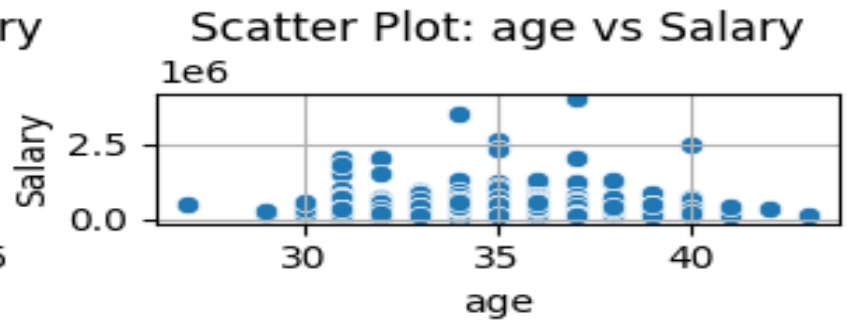






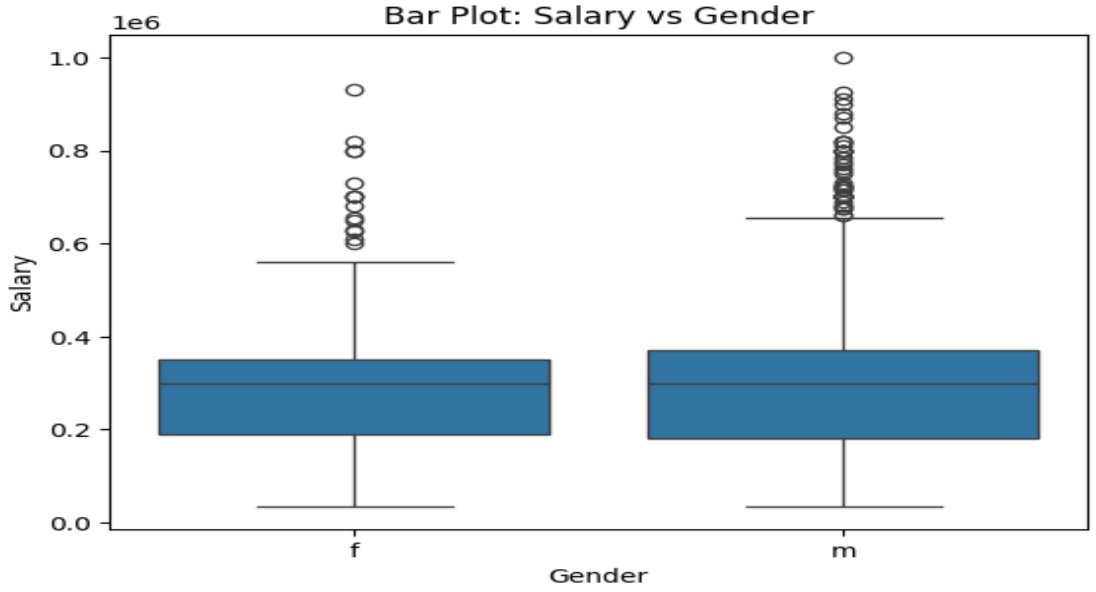


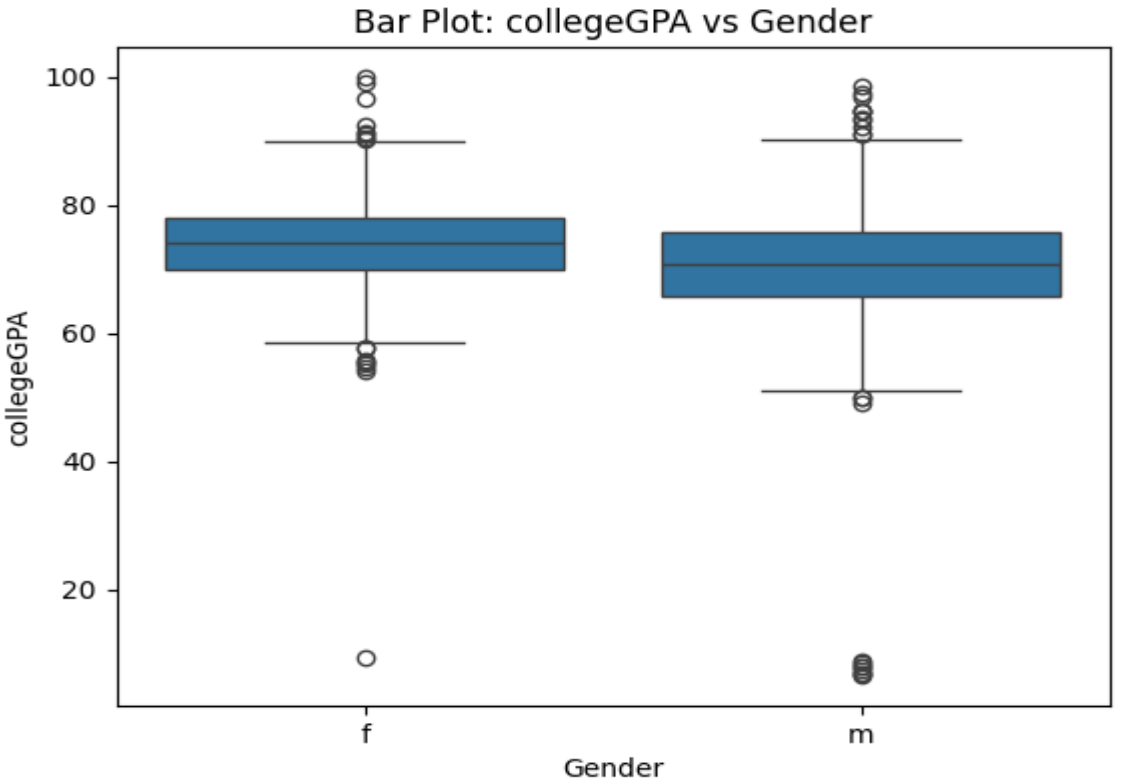




Here, we can see that Salary is higher for those are average studying student in the college(60% - 75%)

BOX PLOT FOR CATEGORICAL VS NUMERICAL COLUMN





Whatever the percentage of male is more tham female , the Average of CollegeGPA , Salary is same for both male and female.

So both are earning same in this generation

STACKED BAR PLOT FOR CATEGORICAL VS CATEGORICAL

```
column1 =
column2 = '


crosstab = pd.crosstab(df[column1], df[column2], normalize='index')
plt.title(f"{column1} vs {column2}")
plt.xlabel("Proportion")
plt.ylabel(column1)
# plt.legend(title=column2, bbox_to_anchor=(1, 1), loc='upper left')
plt.tight_layout()
plt.show()
```

CONCLUSION

In conclusion, the analysis of the engineering graduates' salary dataset reveals several key insights. Firstly, the average salary of a Indian engineering graduates is approximately **Rs.292,545.08**. Secondly, the pie chart illustrates male dominance in the engineering field. Additionally, it is evident that a majority of students prefer undergraduate courses over higher studies. Furthermore, analysis of the State where college located indicates that Uttar Pradesh has the highest number of engineering graduates, followed by Karnataka and Tamil Nadu. Moreover, there is a notable correlation between salary and academic performance, with higher salaries observed among students with average performance in college (60% - 75%). Interestingly, despite a higher proportion of males in the dataset, both male and female graduates earn similar salaries, suggesting gender parity in earnings within this generation of engineers. Overall, these insights provide valuable information for students, educators, and industry stakeholders, aiding in decision-making regarding education choices, career paths, and workforce planning in the engineering sector.

For more Insights and analysis visit my Github repository

Github Repository link : https://github.com/Palamanickam0806/Exploratory_Data_Analysis_on_Graduates_Salary

Dataset: link

LinkedIn account: https://www.linkedin.com/in/palamanickam-s-2ab81925b/recent-activity/all/

Email: Valliaravind@gmail.com

THANK

YOU