INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis of AMCAT DATA

Done by –
ID-IN9240776
Name=B Koushik Aryan

Agenda

- Business Problem
- Objective of the Project
- Summary of the Data

- **Exploratory Data Analysis:**

a. Data Cleaning Steps
b. Data Manipulation Steps
c. Univariate Analysis Steps
d. Bivariate Analysis Steps

- Key Business Question
- Conclusion

INNOMATICS RESEARCH LABS

Motivation of the Project

- Predict engineering graduates' salaries based on academic and demographic factors.
- Improve job title and location matching for enhanced employee satisfaction and retention.
- Identify key factors influencing employee success, such as educational background and personality traits.
- Optimize recruitment strategies through data-driven insights.
- Foster employee development to improve organizational performance and long-term business outcomes.

Objective of the Project

- The objective of the problem is to leverage the dataset to develop predictive models,
- conduct exploratory analysis, and extract actionable insights that can inform strategic decisions aimed at optimizing the hiring process and improving employee outcomes for engineering graduates.
- By addressing these business problems, the objective is to enhance recruitment strategies, match candidates with suitable job roles and locations.
- Facilitate employee development to ultimately drive organizational performance and satisfaction.

Summary of Data set

- The dataset consists of employment outcomes for engineering graduates, focusing on key variables such as salary, job titles, and job locations. With approximately 39 independent variables and 3998 data points, it provides a comprehensive overview of candidates' backgrounds and skill sets.

Key features of the dataset include-

- **Demographic Information**: Details about the candidates' backgrounds, including age, gender, and location.
- **Educational Qualifications**: Information on grades, board affiliations, college tier, and GPA.
- **Standardized Assessment Scores**: Results from assessments evaluating cognitive, technical, and personality skills.
- **Diverse Engineering Disciplines**: Coverage of various engineering fields, including computer programming, electronics, and mechanical engineering.
- **AMCAT Personality Test Scores**: Insights into personality traits such as conscientiousness, agreeableness, and openness to experience.

Data Set

3]:

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage	12board	CollegeI
0	train	203097	420000	2012-06-01	present	senior quality engineer	Bangalore	f	1990-02-19	84.3	board ofsecondary education,ap	2007	95.8	board of intermediate education,ap	114
1	train	579905	500000	2013-09-01	present	assistant manager	Indore	m	1989-10-04	85.4	cbse	2007	85.0	cbse	580
2	train	810601	325000	2014-06-01	present	systems engineer	Chennai	f	1992-08-03	85.0	cbse	2010	68.2	cbse	6
3	train	267447	1100000	2011-07-01	present	senior software engineer	Gurgaon	m	1989-12-05	85.6	cbse	2007	83.6	cbse	692
4	train	343523	200000	2014-03-01	2015-03-01 00:00:00	get	Manesar	m	1991-02-27	78.0	cbse	2008	76.8	cbse	1136

Data set contains

- Shape=3998 rows & 39columns.
- Null values = 0
- Duplicated = 0

INNOMATICS RESEARCH LABS

Information of Dataset

```
Data columns (total 39 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Unnamed: 0               3998 non-null   object
 1   ID                       3998 non-null   int64
 2   Salary                   3998 non-null   int64
 3   DOJ                      3998 non-null   datetime64[ns
 4   DOL                      3998 non-null   object
 5   Designation              3998 non-null   object
 6   JobCity                  3998 non-null   object
 7   Gender                   3998 non-null   object
 8   DOB                      3998 non-null   datetime64[ns
 9   10percentage             3998 non-null   float64
 10  10board                  3998 non-null   object
 11  12graduation             3998 non-null   int64
 12  12percentage             3998 non-null   float64
 13  12board                  3998 non-null   object
 14  CollegeID                3998 non-null   int64
 15  CollegeTier              3998 non-null   int64
 16  Degree                   3998 non-null   object
 17  Specialization           3998 non-null   object
 18  collegeGPA               3998 non-null   float64
 19  CollegeCityID            3998 non-null   int64
 20  CollegeCityTier          3998 non-null   int64
 21  CollegeState             3998 non-null   object
 22  GraduationYear           3998 non-null   int64
 23  English                  3998 non-null   int64
 24  Logical                  3998 non-null   int64
 25  Quant                    3998 non-null   int64
 26  Domain                   3998 non-null   float64
 27  ComputerProgramming      3998 non-null   int64
 28  ElectronicsAndSemicon    3998 non-null   int64
 29  ComputerScience          3998 non-null   int64
 30  MechanicalEngg           3998 non-null   int64
 31  ElectricalEngg           3998 non-null   int64
 32  TelecomEngg              3998 non-null   int64
 33  CivilEngg                3998 non-null   int64
 34  conscientiousness        3998 non-null   float64
 35  agreeableness            3998 non-null   float64
 36  extraversion             3998 non-null   float64
 37  nueroticism              3998 non-null   float64
 38  openess to experience    3998 non-null   float64
```

The Dataset contains-

- 25 features are numerical
- 10 features are categorical
- 3 features are datetime

INNOMATICS RESEARCH LABS

Data Manipulation Steps

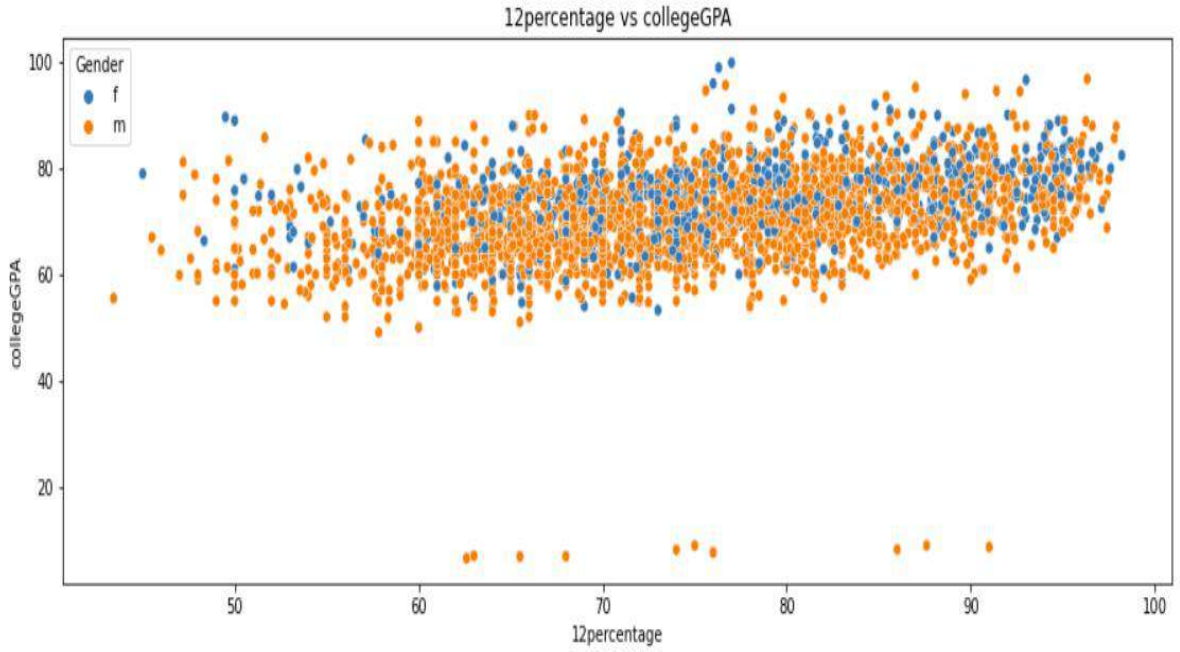
Data set contains

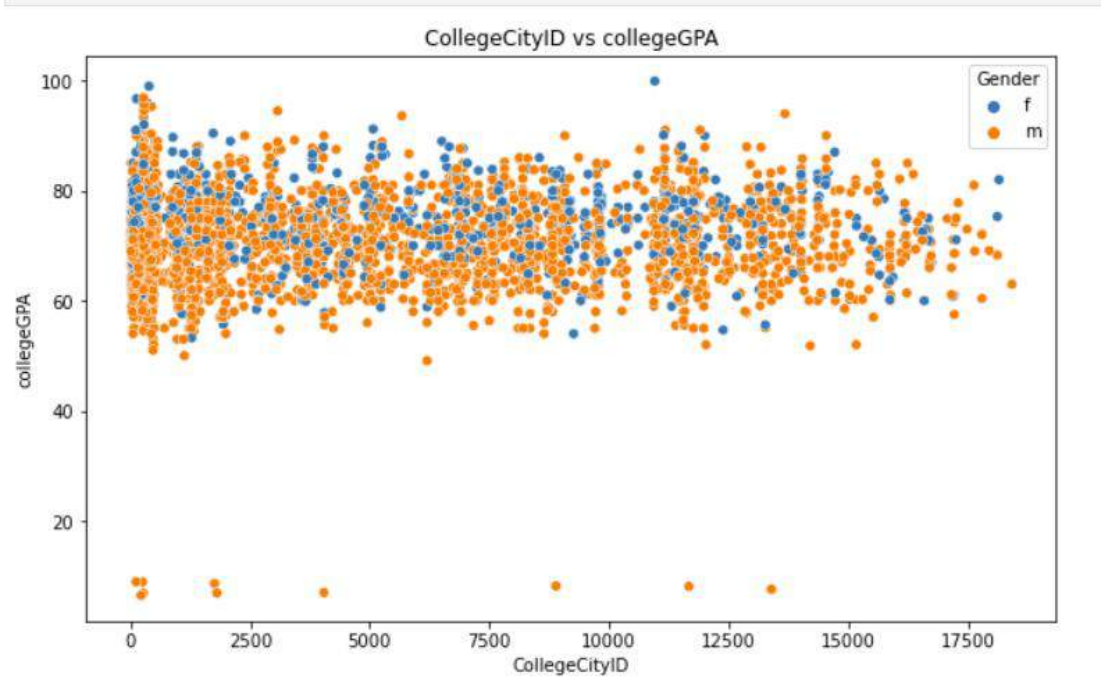
- Shape=3998 rows & 39columns.
- Null values = 0
- Duplicated = 0
- Outliers = 2398

Cleaning steps

- Removing irrelevant data points from specific columns
- Identifying inconsistence of Data
- Replacing the irrelevant values with suitable values
- 10 board and 12 board columns contain 0 value which is missing value. Filled with mean of that specific column.
- Detecting outliers using IQR- and using visual plots etc

Identifying Outliers



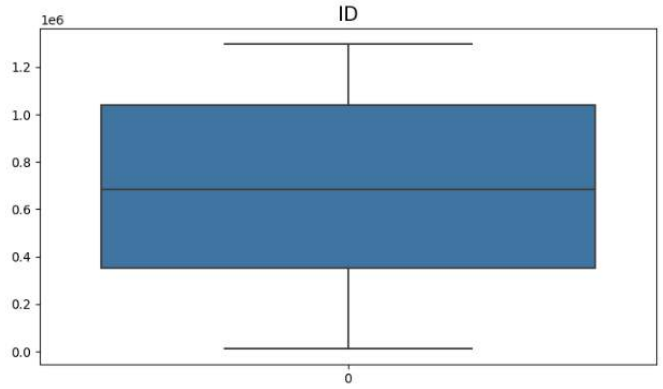


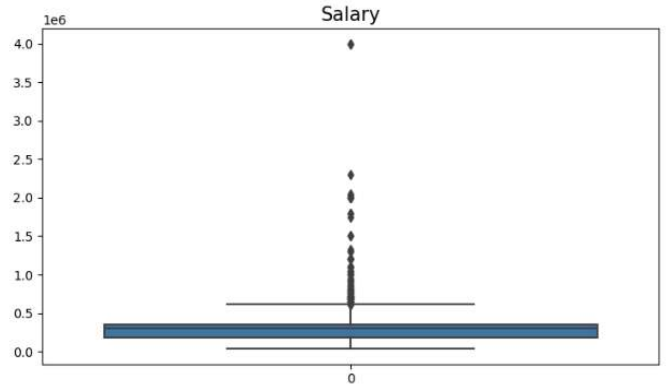
Observation from Scatter plot:

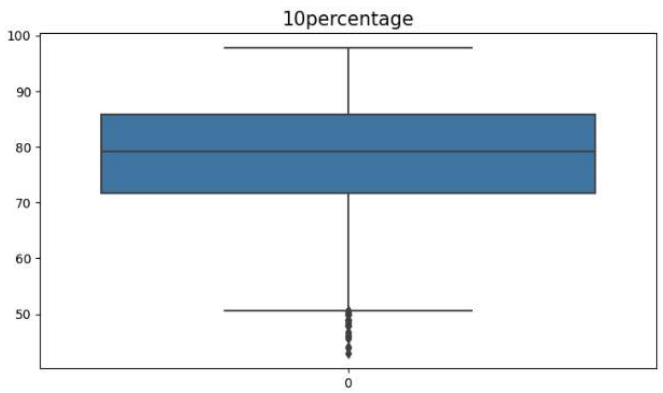
- We can some outliers in both Scatterplots at bottom of the plots

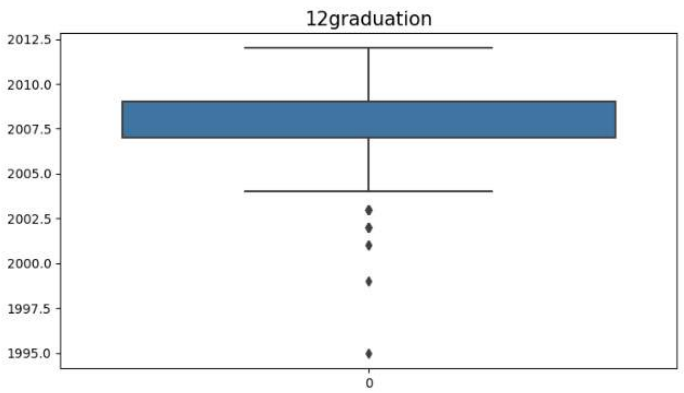
INNOMATICS RESEARCH LABS

Identifying Outliers







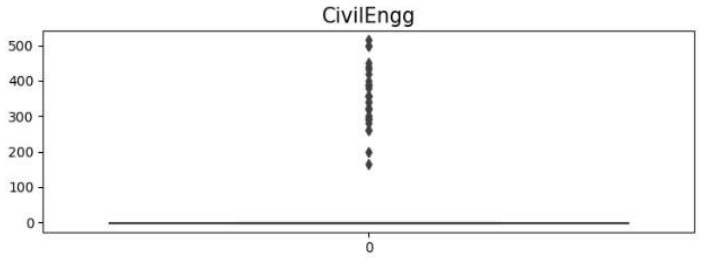


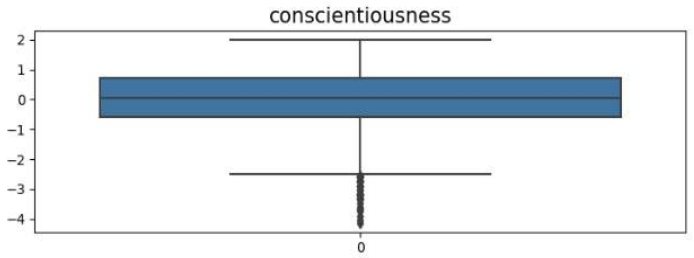
Observation from above charts:

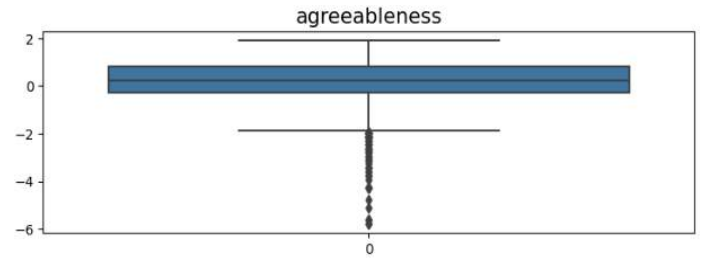
- We can more numbers of outliers in Salary,10oercentage ,12Graduation plots

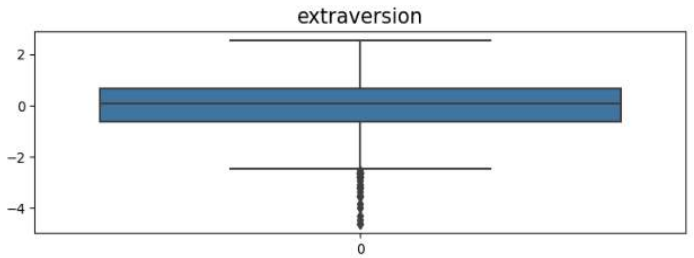
INNOMATICS RESEARCH LABS

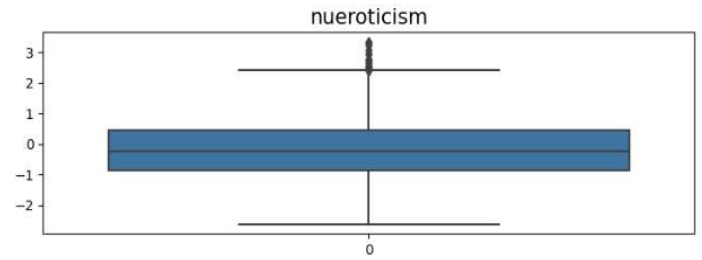
Identifying Outliers

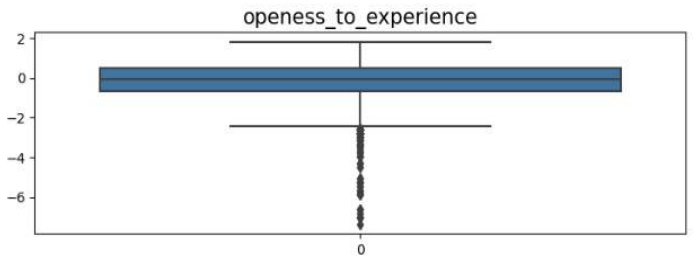












Observation from above charts:

- We can more numbers of outliers in all box plots

INNOMATICS RESEARCH LABS

Identifying Outliers

```
q1=dt.quantile(0.25)
q3=dt.quantile(0.75)
iqr=q3-q1
lb=q1-1.5*iqr
ub=q3+1.5*iqr
outilers=(dt<lb)|(dt>ub)
outilers.sum()
```

```
10board                    0
10percentage              25
12board                    0
12graduation              28
12percentage               0
CivilEngg                 36
CollegeCityID              0
CollegeCityTier            0
CollegeID                  0
CollegeState               0
CollegeTier              235
ComputerProgramming        1
ComputerScience          793
DOL                        0
Degree                     0
Designation                0
Domain                   198
ElectricalEngg           143
ElectronicsAndSemicon      0
English                    8
Gender                     0
GraduationYear             2
ID                         0
JobCity                    0
Logical                   13
MechanicalEngg           189
Quant                     18
Salary                   111
Specialization             0
TelecomEngg              308
Unnamed: 0                 0
agreeableness            102
collegeGPA                28
conscientiousness         51
extraversion              33
nueroticism               19
openess_to_experience     83
dtype: int64
```

Observation from Analysis:

- By using IQR we found there are huge no of outliers in the data set
- Above o/p gives shows the outlies in each respective column

INNOMATICS RESEARCH LABS

Data Visualization

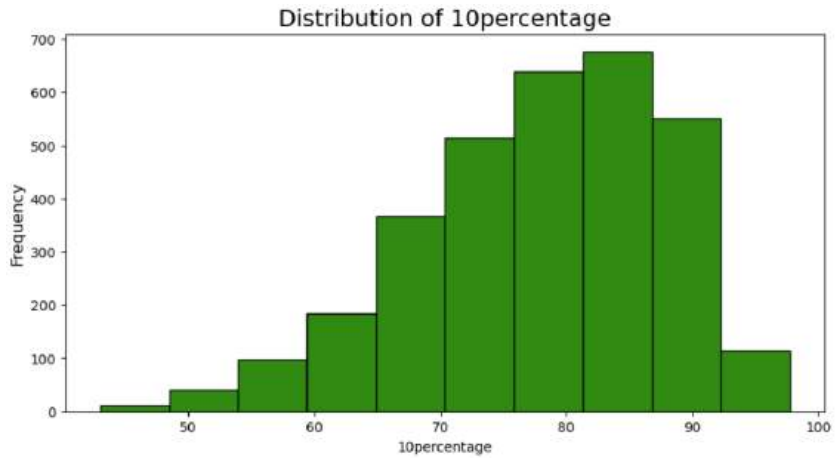
Univariate Plot:

- A univariate plot visualizes the distribution of a single variable. Common examples include histograms, box plots, and density plots, which help to show the central tendency, spread, and outliers in the data
- While univariate plots focus on understanding the distribution of one variable.

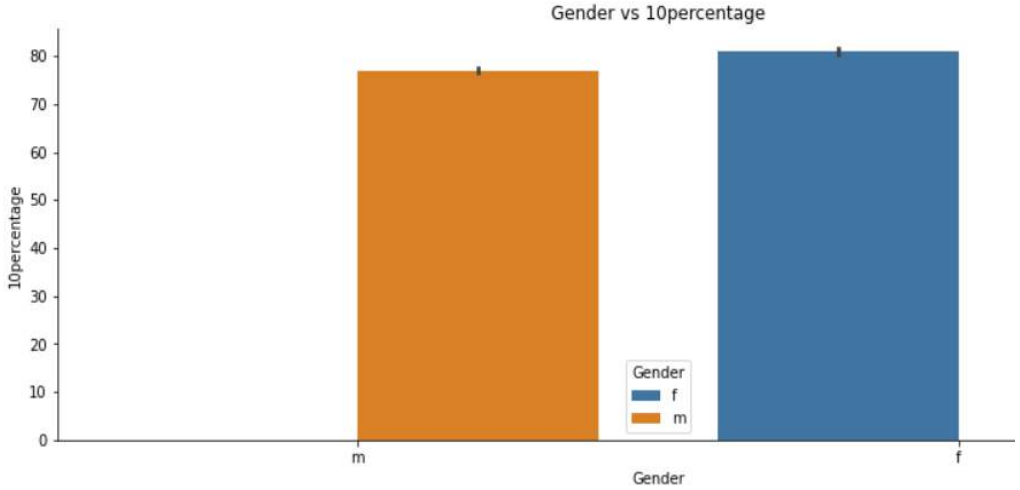
Bivariate Plot:

- A bivariate plot shows the relationship between two variables. Scatter plots, line plots, and heatmaps are typical examples that help illustrate correlations, patterns, or trends between the two variables
- bivariate plots help analyze relationships and interactions between two variables, providing deeper insights into how they may affect one another

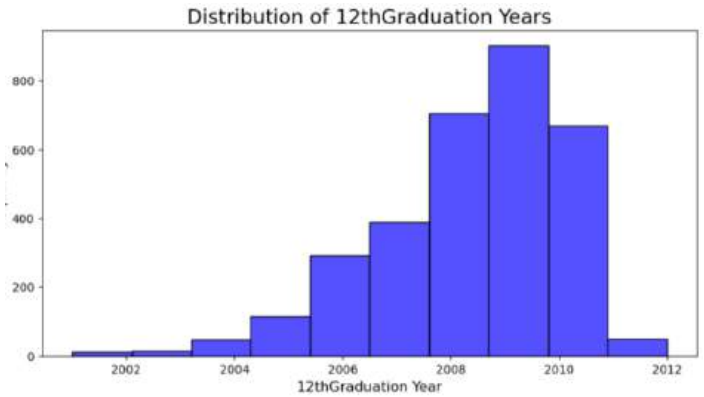
EDA of 10th and 12th Standard Students



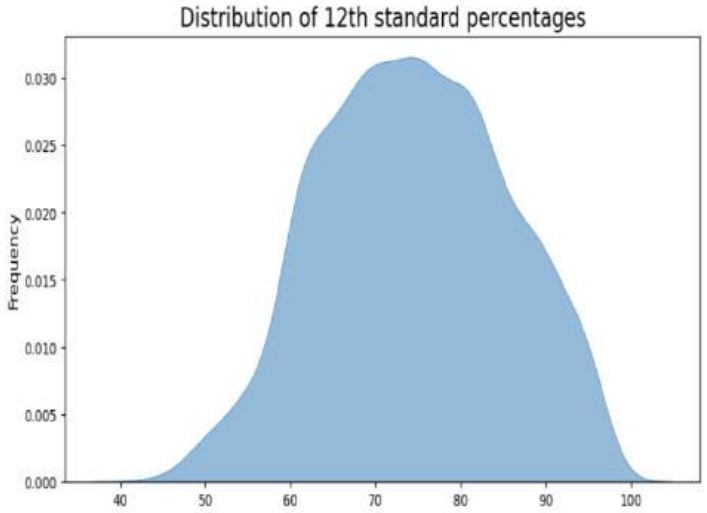
- Highest distribution of 10th% between 80% to 90%



- Girls achieved higher percentage scores than boys

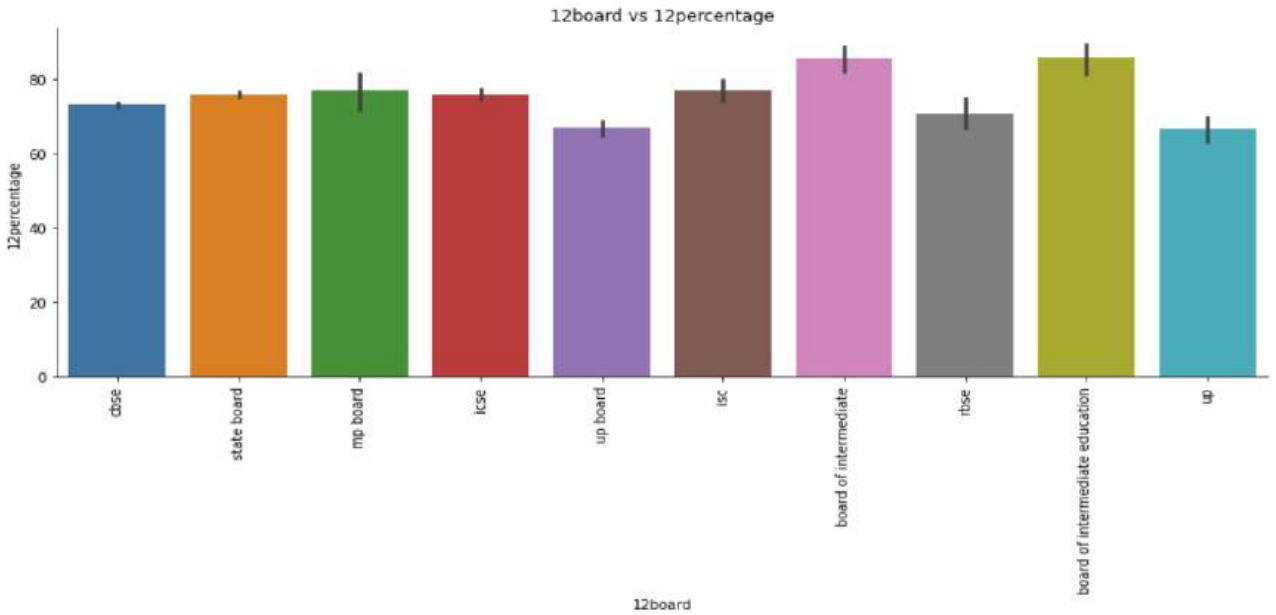


- There was a substantial increase in the number of 12thgraduates between 2008 to 2009



- This chart indicates that a majority of students scored between 70% and 80% in their 12th standard examinations
- We can observe there are some students who scored low marks at 40%

INNOMATICS RESEARCH LABS



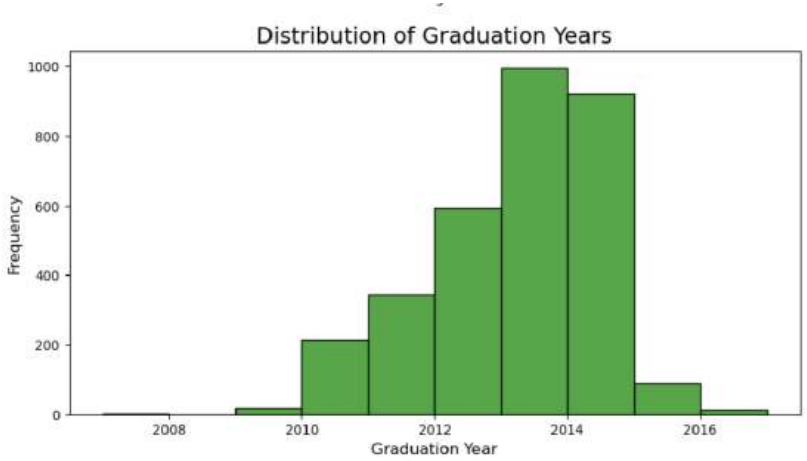

```
dt.groupby(['12board'])['12percentage'].min().sort_values(ascending=False).tail(10)
```

```
12board
puc                                    49.00
rajasthan board                        48.34
department of pre university education 47.83
cbse                                   47.60
up board                               47.20
gshseb                                 47.20
state board                            46.00
pue                                    45.50
chse                                   45.00
p u board, karnataka                   43.42
Name: 12percentage, dtype: float64
```

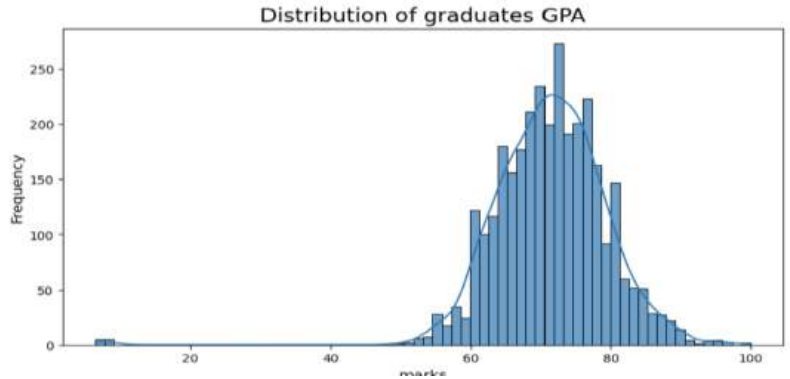
Observations from Above Chart-

- The bar chart shows that students from the IPE board achieved the highest percentages compared to other boards.
- The non-visualized data frame lists the lowest percentage of 12th standard
- There are no failure candidates in the intermediate exams, indicating a strong overall performance.
- The lowest percentage recorded among students is 43%.
- This data leads to the conclusion that there is a 100% pass rate in the interboard examinations.

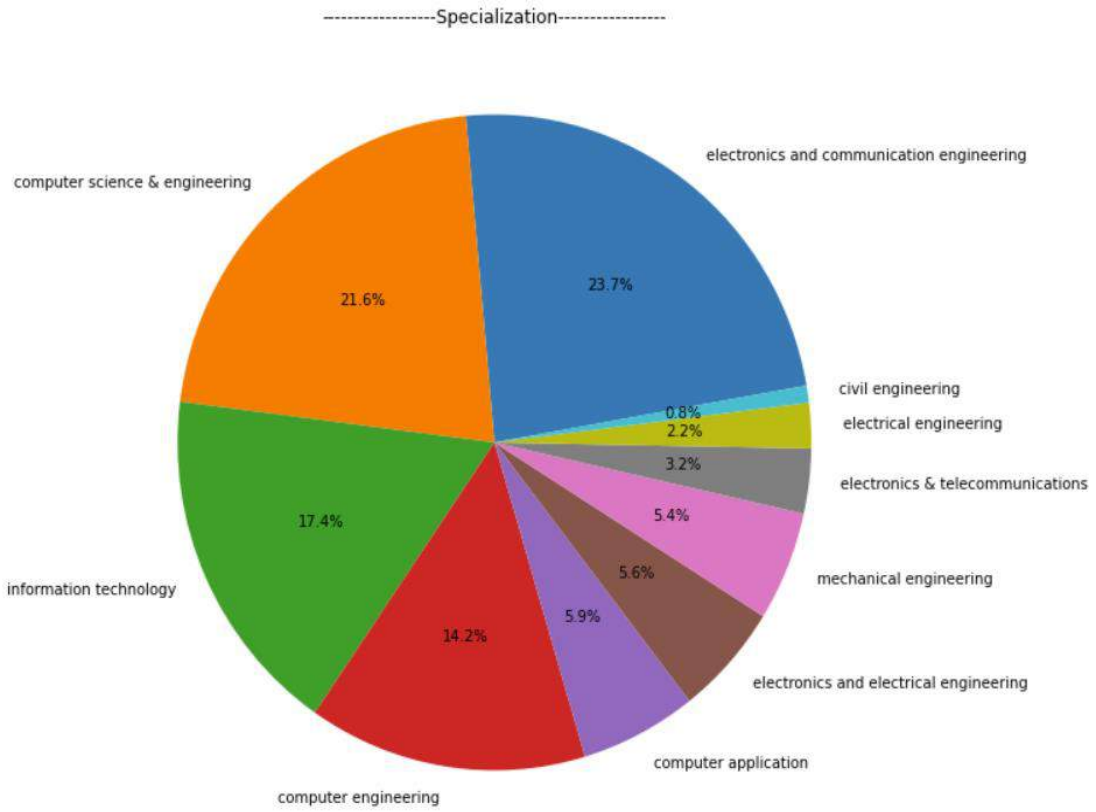
EDA of Graduates



- There was a substantial increase in the number of graduates between 2013 and 2015



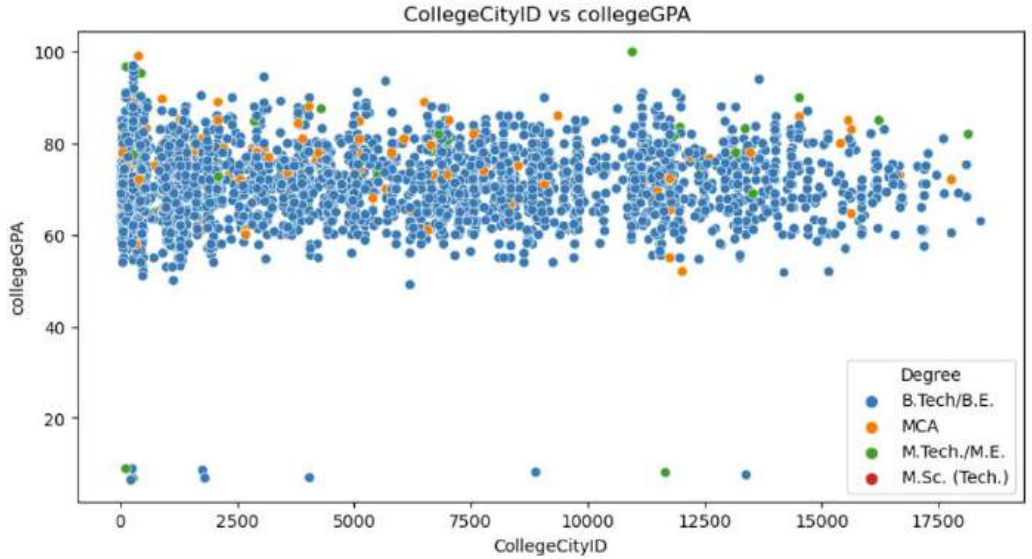
- Most of the students has the college GPA in the range of 60-80
- We can see some students are failed in graduation

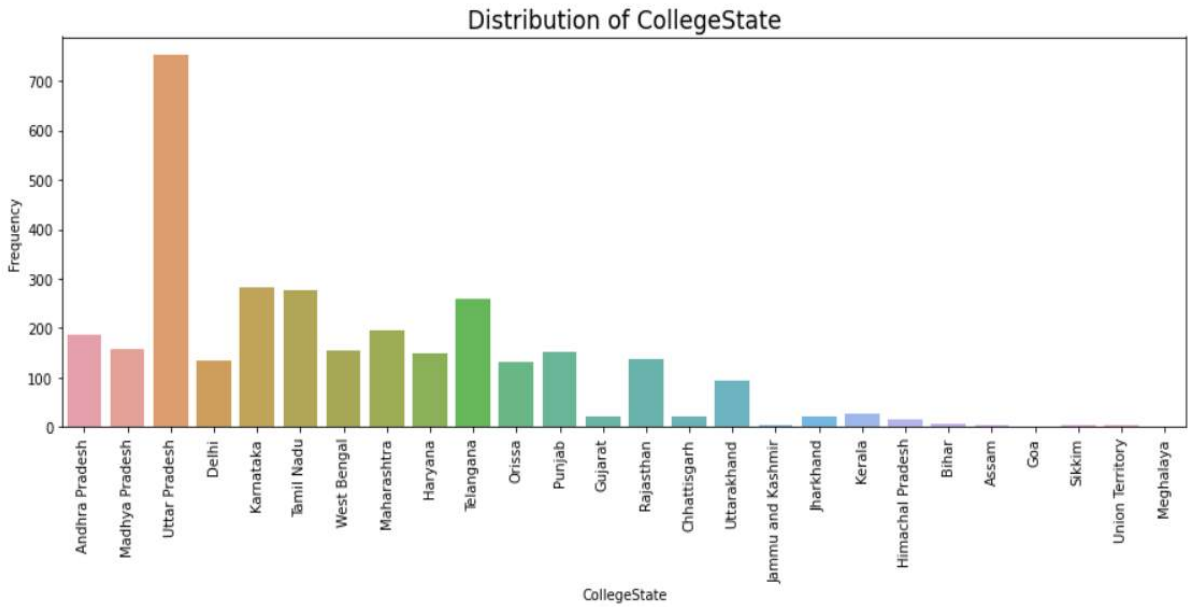


- Most of the graduates pursued ECE,CSE,IT
- ECE as the highest percentage
- Civil graduates as low percentage

INNOMATICS RESEARCH LABS

Graduates vs GPA





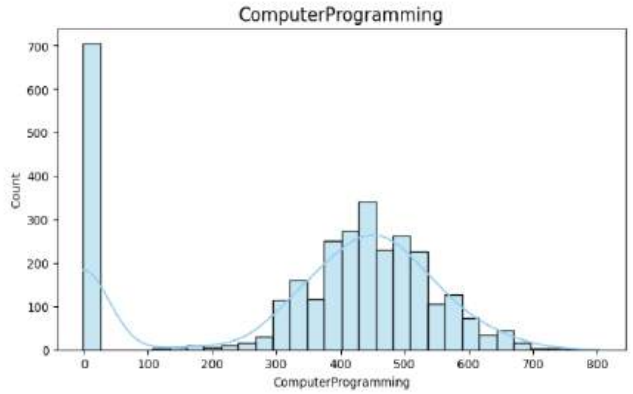
CollegeID	Degree	
10950	M.Tech./M.E.	99.93
388	MCA	99.00
275	B.Tech/B.E.	96.90
128	M.Tech./M.E.	96.70
347	B.Tech/B.E.	96.00
285	B.Tech/B.E.	95.70
443	M.Tech./M.E.	95.30
3076	B.Tech/B.E.	94.50
13668	B.Tech/B.E.	94.00
5671	B.Tech/B.E.	93.60

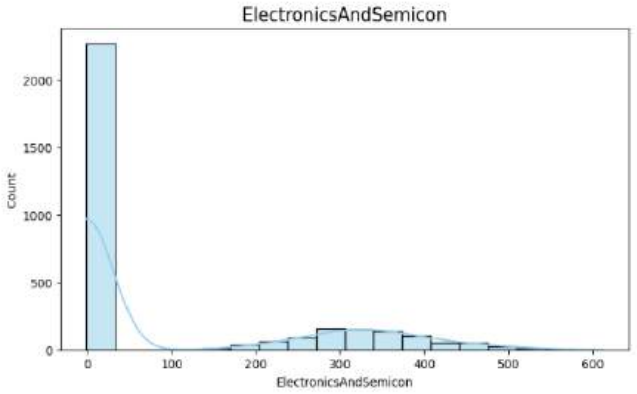
Name: collegeGPA, dtype: float64

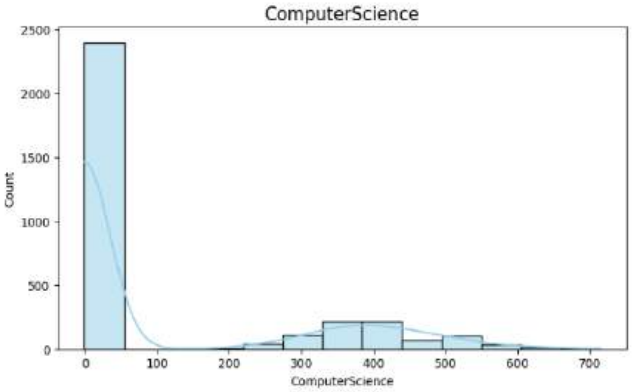
Observation from the Above Charts:

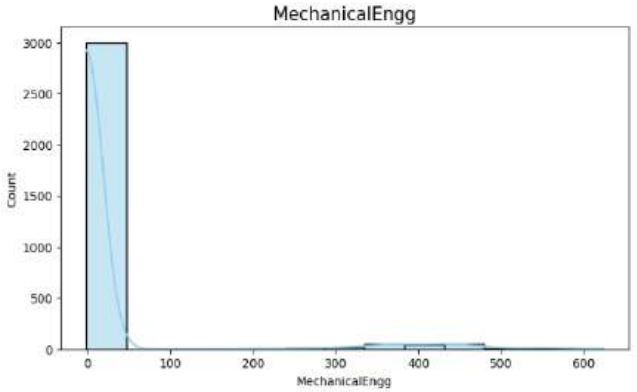
- It appears that a significantly higher number of students pursued B.Tech compared to other degrees.
- In scatter plot we can there are some candidates who got failed in graduation
- Compare to M.tech, M.sc graduates are less
- According to AMCAD Data most of the graduates are for uttarpradesh state
- In Groupby data frame shows top 10 percentage of graduated students including UG and P
- ID-10950 Mtech graduated secured with 99.93%

INNOMATICS RESEARCH LABS





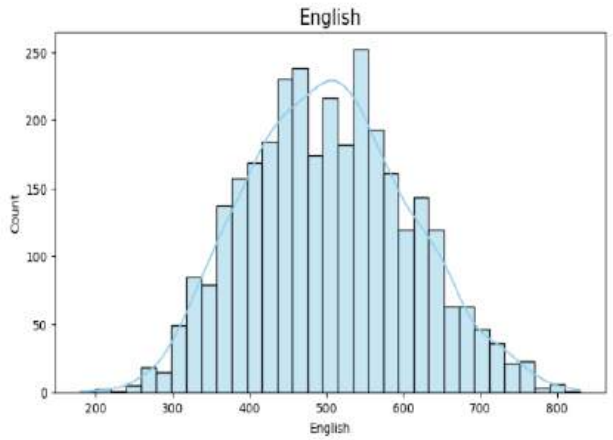


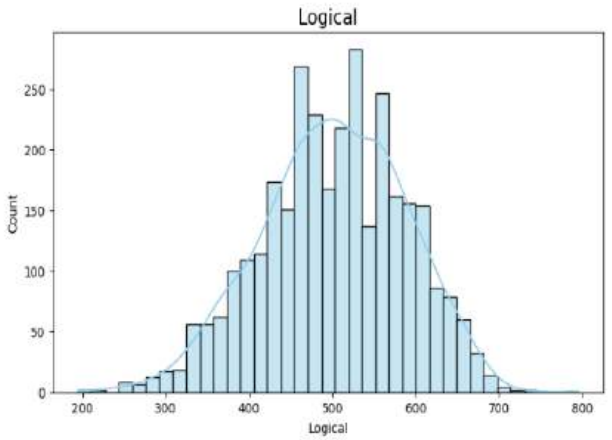


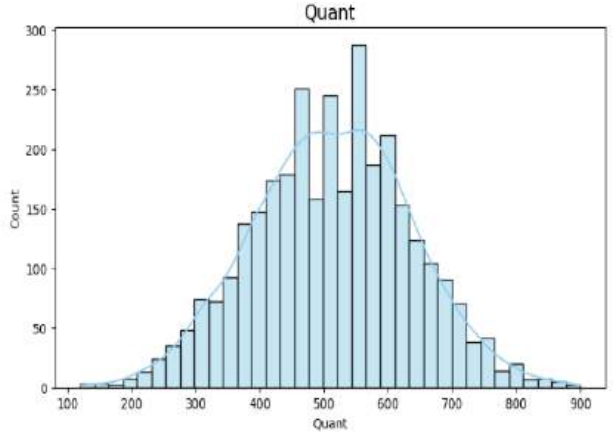
Observations from above Plot-

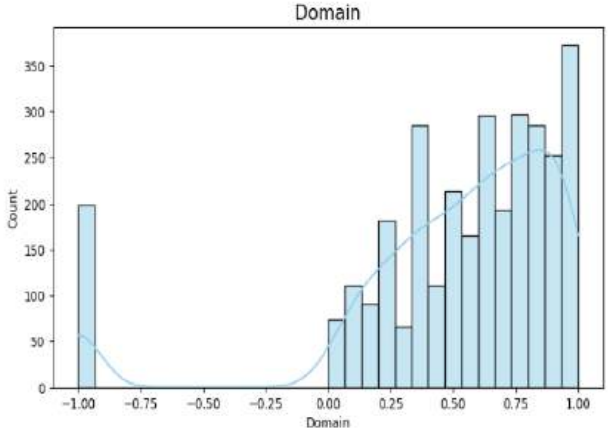
- In Computer ProgrammingThe majority of scores ranged between 416 and 459. The peak occurred at 455, with an average score of 452
- In Electronics and semiconductor Most scores fell between 0 and 79. The highest number of students scored 0, with an average score of 96
- In Computer Science and Mechanicaleng there are low score .

INNOMATICS RESEARCH LABS







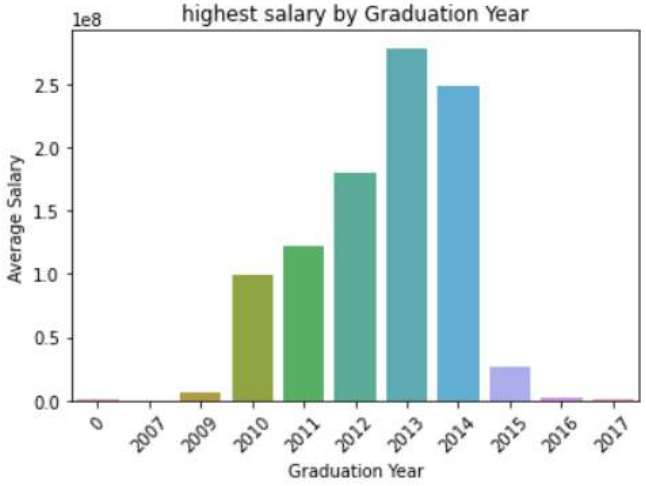


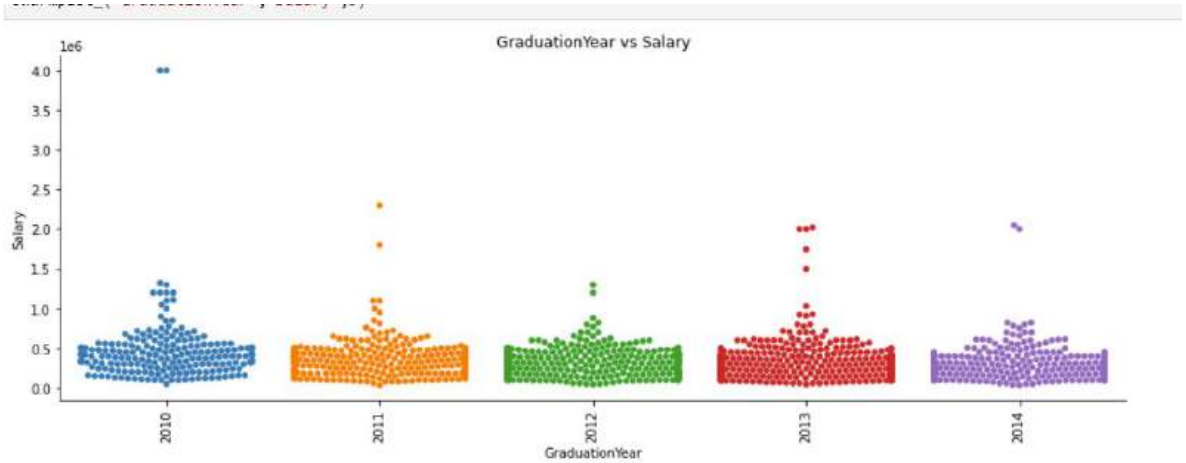
Observations from above Plot-

- In English Majority of the scores fell within the range of 389 to 545.
- In Quant :Majority of the scores were in between 425-608. The maximum number of students scored 605 with an average of 513
- In logical: Most scores fell within the range of 454 to 584, peaking at 495, with an average of 502.

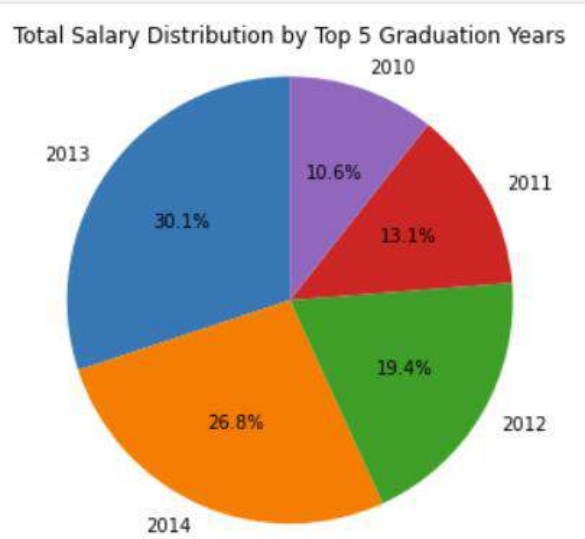
INNOMATICS RESEARCH LABS

Graduates vs Salaries

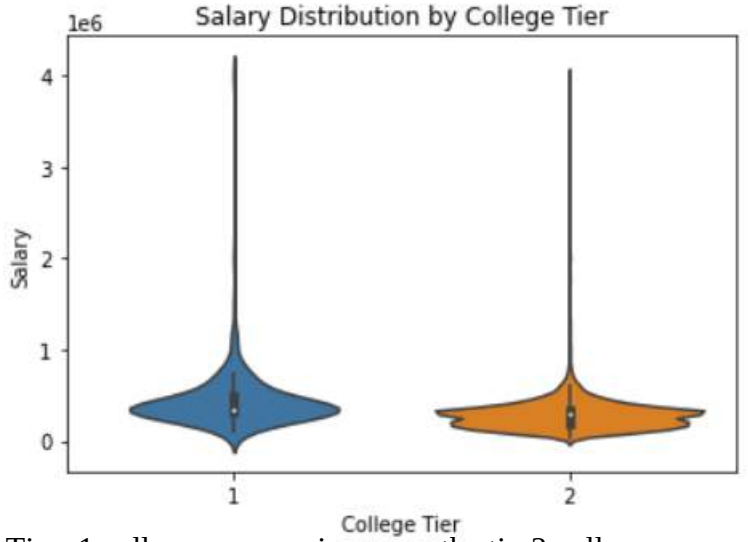




- In Above charts shows 2013 & 2014 graduates are earning more salary than other graduate batches



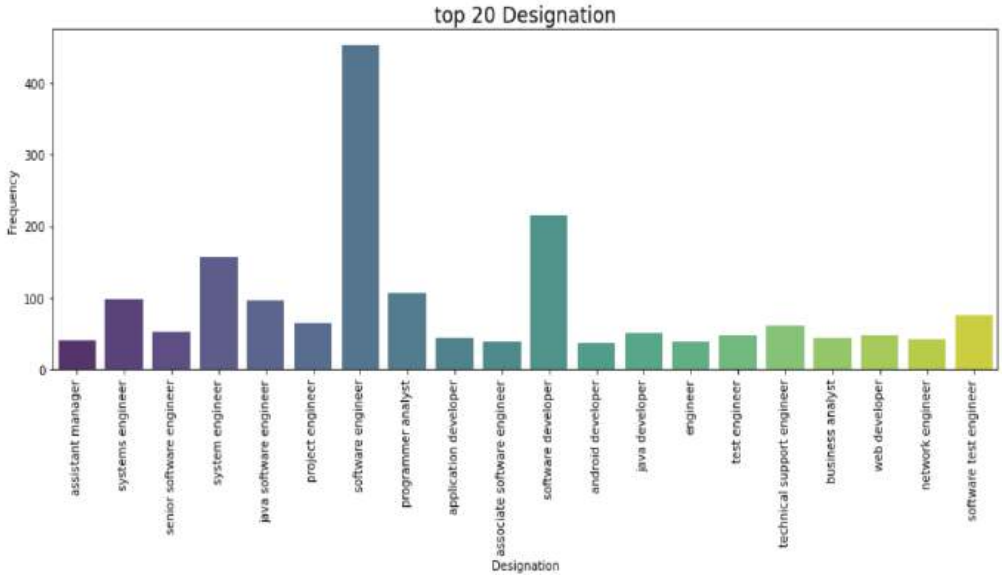
- Pie charts shows the percentages of the top 5 Gradutaion years



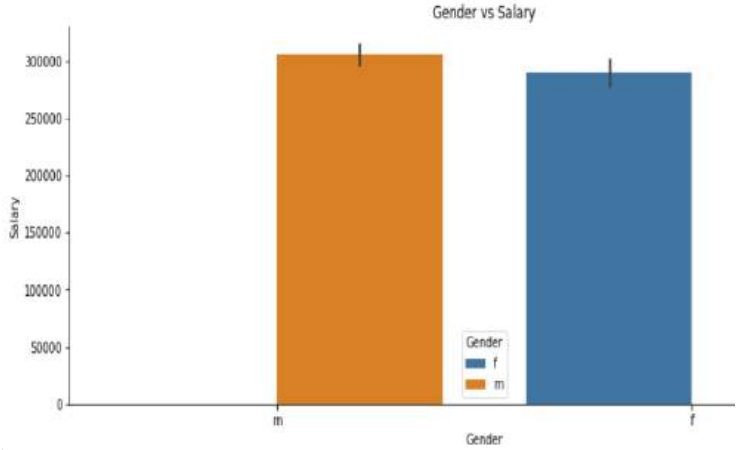
Tier -1 colleges are earning more the tie-2 colleges

INNOMATICS RESEARCH LABS

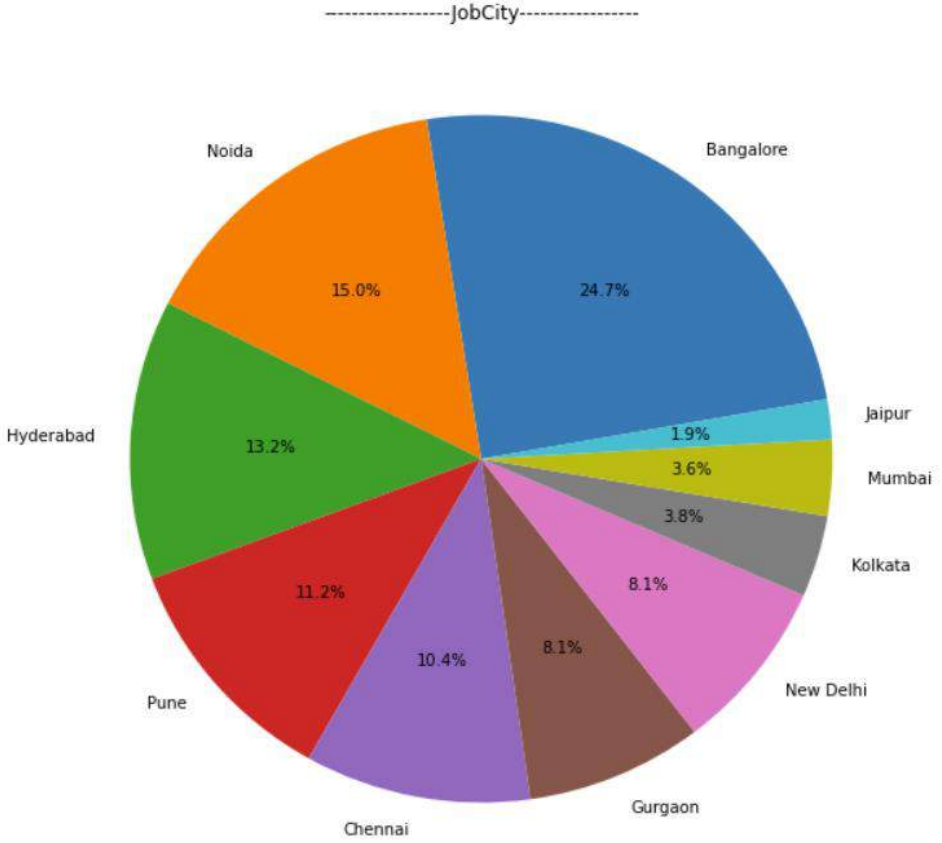
EDA of Employes



- The bar chart displays top 20 designations out them software engineers are more



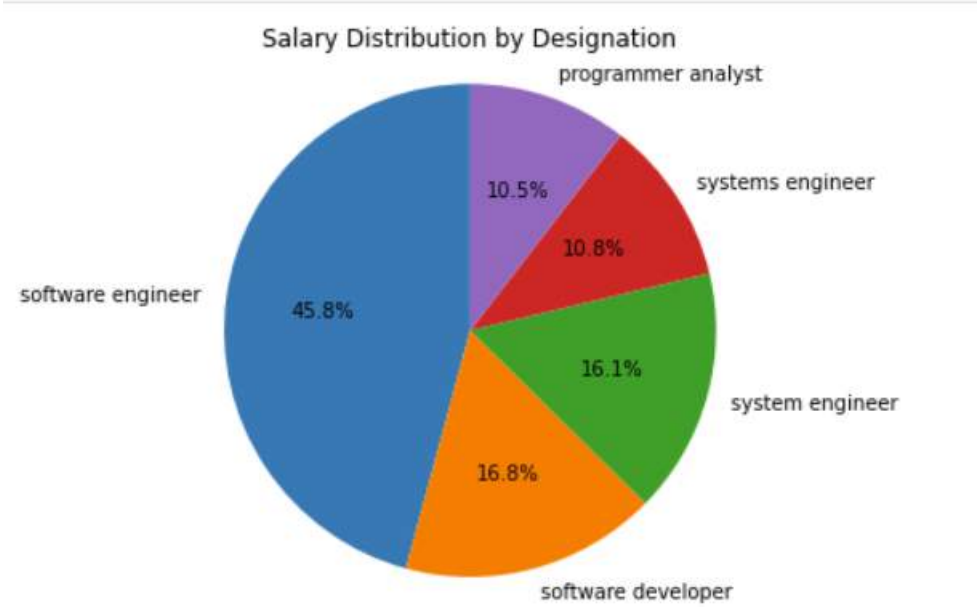
- Male employes earn more income than Female employes



- The majority of the working employees are from Bangalore
- Jaipur has less percentage of employes

INNOMATICS RESEARCH LABS

Designation vs Salary



Observation from above charts:

- Pie chart shows the proportion between top5 designations
- Software engineeries employes more salary than other employes
- In Swarm plot The salary distribution for software engineers is wider compared to other designations, with some outliers indicating much higher salaries, reaching above 2 million.
- The salary for programmer analysts is tightly clustered, indicating less variation in salary compared to other roles
- The majority of salaries for software developers cluster around a mid-level range, a few high outliers exist, indicating a broader potential for higher earnings in this role

INNOMATICS RESEARCH LABS

Designation vs Salary

```
dt.groupby(['ID','Designation'])['Salary'].sum().sort_values(ascending=False)
```

```
ID       Designation
41147    automation engineer          4000000
48107    senior software engineer     4000000
281074   software developer           2300000
342930   software engineer trainee    2050000
641560   operations analyst           2020000
1045685  data scientist               2000000
615010   it technician                2000000
803778   technical lead               2000000
1254777  salesforce developer         1800000
202950   client services associate    1800000
Name: Salary, dtype: int64
```

```
dt.groupby(['ID','Designation'])['Salary'].min().sort_values(ascending=False)
```

```
ID       Designation
211934   entry level management trainee    50000
920413   software developer                 50000
812145   jr. software engineer              50000
109571   marketing analyst                  50000
482761   maintenance engineer               50000
1088385  graduate apprentice trainee        45000
637137   training specialist                45000
578157   application developer              40000
242100   systems engineer                   35000
1272092  .net developer                     35000
Name: Salary, dtype: int64
```

Observation from above charts:

- The above Group by charts describes about maximum and minimum salaries of the employes from different roles
- Automation engineer and senior software engineer earns more salary
- System engineer and .net developer earns minimum salary
- Max_salary=40M
- Min_salary=35K

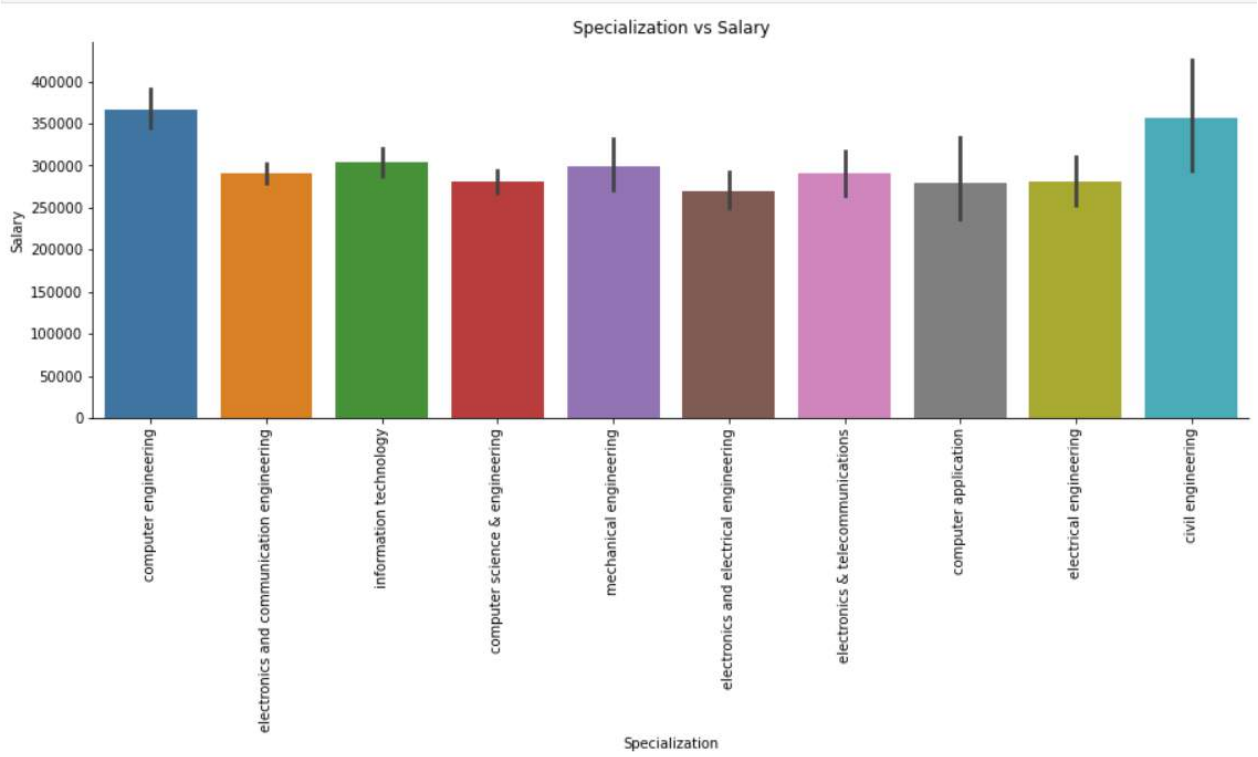
INNOMATICS RESEARCH LABS

Designation vs Salary

```
filt = df[df['Graduated'] == 'ungraduted']
filt.groupby(['ID', 'Designation'])['Salary'].max().reset_index()
```

	ID	Designation	Salary
0	240465	systems engineer	470000
1	249853	electrical project engineer	180000
2	262814	web developer	145000
3	287976	engineer	250000
4	299447	assistant professor	360000
5	385442	software engineer	820000
6	813008	it technician	180000
7	868740	product development engineer	240000
8	912934	mechanical engineer	400000
9	1262900	java software engineer	180000

- Those 9 students are ungraduated students who are working with different designations and there salaries
- Minimum salary-180k
- Maximum salary-470k



- In bar chart shows the Specialization vs Salary,computer engineering students earns more salary than other branches

INNOMATICS RESEARCH LABS

Conclusion

- The dataset analyzed factors influencing salary levels among engineering graduates, including tenure, college tier, and job designation, identifying Senior Software Engineers as the highest earners.
- Gender had minimal impact on average income, but female graduates earned lower salaries than the overall average.
- Academic performance indicators, such as GPA, did not consistently correlate with salary levels, indicating a complex relationship between education and earnings.
- Regional salary trends revealed significant variations across major cities and highlighted lucrative job roles.
- The study addressed gender pay disparities and explored the education-salary relationship to promote equitable employment practices.
- Recommendations for further analysis using machine learning were suggested to deepen insights into salary determinants.
- Overall, the project provides valuable insights into the employment dynamics of engineering graduates, aiding organizations and policymakers in refining employment strategies.

THANK YOU

INNOMATICS RESEARCH LABS