INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO dataset 2015

About me

- Myself Ibteda Azeem
- I have completed Btech in Electrical and Electronics Engineering
- Data science skills are in high demand across various industries. As organizations increasingly rely on data to drive decision-making, there's a growing need for professionals who can extract insights from data. I am a data science enthusiast who wants to apply my skills to solve real world problems

- Github link : https://github.com/Azeem-I?tab=repositories
- Linkedin link : https://www.linkedin.com/in/azeem-i-198359270/

Objective Of The Project

- The Project aims at performing Exploratory Data Analysis on the given dataset and finding the patterns existing between the features that contains the employment outcome of Engineering graduates
- The dataset was released by Aspiring Minds from the Aspiring Mind Employment Outcome 2015 (AMEO). The study is primarily limited only to students with engineering disciplines.

About The Dataset

- The dataset was released by Aspiring Minds from the Aspiring Mind Employment Outcome 2015 (AMEO).
- The study is primarily limited only to students with engineering disciplines. The dataset contains the employment outcomes of engineering graduates as dependent variables (Salary, Job Titles, and Job Locations) along with the standardized scores from three different areas – cognitive skills, technical skills and personality skills. The dataset also contains demographic features.
- ID : A unique ID to identify a candidate
- Salary: Annual CTC offered to the candidate (in INR)
- DOJ : Date of joining the company
- DOL :Date of leaving the company
- Designation : Designation offered in the job
- JobCity : Location of the job (city)
- Gender : Candidate’s gender

About The Dataset

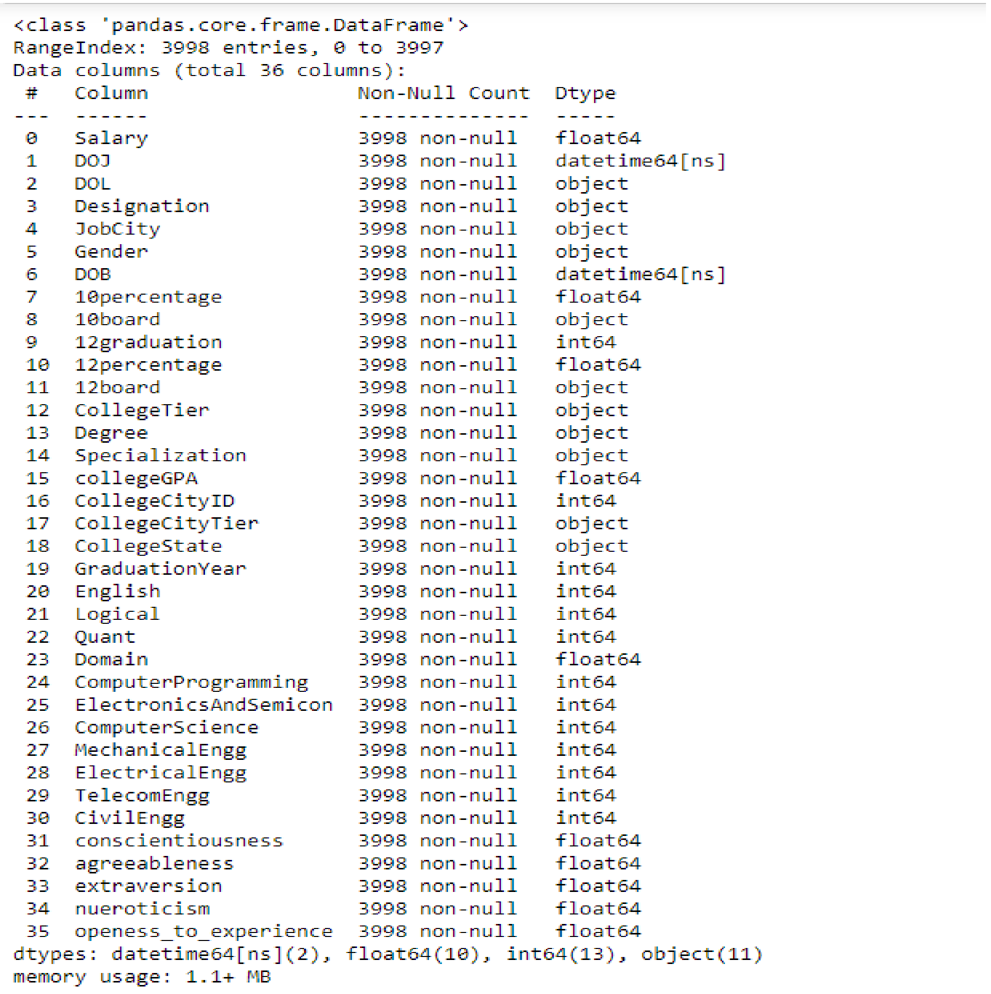
- DOB : Date of birth of candidate
- 10percentage : Overall marks obtained in grade 10 examinations
- 10board: The school board whose curriculum the candidate followed in grade 10
- 12graduation : Year of graduation - senior year high school
- 12percentage : Overall marks obtained in grade 12 examinations
- 12board:The school board whose curriculum the candidate followed in grade 12
- CollegeID :Unique ID identifying the college which the candidate attended
- CollegeTier :Tier of college
- Degree: Degree obtained/pursued by the candidate
- Specialization : Specialization pursued by the candidate
- CollegeGPA : Aggregate GPA at graduation
- CollegeCityID :A unique ID to identify the city in which the college is located in
- CollegeCityTier : The tier of the city in which the college is located

- CollegeState : Name of States
- GraduationYear : Year of graduation (Bachelor's degree)
- English : Scores in AMCAT English section
- Logical : Scores in AMCAT Logical section
- Quant : Scores in AMCAT Quantitative section
- Domain :Scores in AMCAT's domain module
- ComputerProgramming : Score in AMCAT's Computer programming section
- ElectronicsAndSemicon : Score in AMCAT's Electronics & Semiconductor Engineering section
- ComputerScience : Score in AMCAT's Computer Science section
- MechanicalEngg : Score in AMCAT's Mechanical Engineering section
- ElectricalEngg : Score in AMCAT's Electrical Engineering section
- TelecomEngg : Score in AMCAT's Telecommunication Engineering section
- CivilEngg : Score in AMCAT's Civil Engineering section
- conscientiousness : Scores in one of the sections of AMCAT's personality test
- agreeableness : Scores in one of the sections of AMCAT's
- neuroticism: Scores in one of the sections of AMCAT's personality test
- openess_to_experience : Scores in one of the sections of AMCAT's personality test

INNOMATICS RESEARCH LABS

General Information about Data

- - There are 3998 entries and 38 features of the given dataset.
- - 10 features are of float64 datatype,17 of int64 and 12 object datatype columns.
- - There are no missing values in the dataset.
- - There are no duplicate entries in the dataset

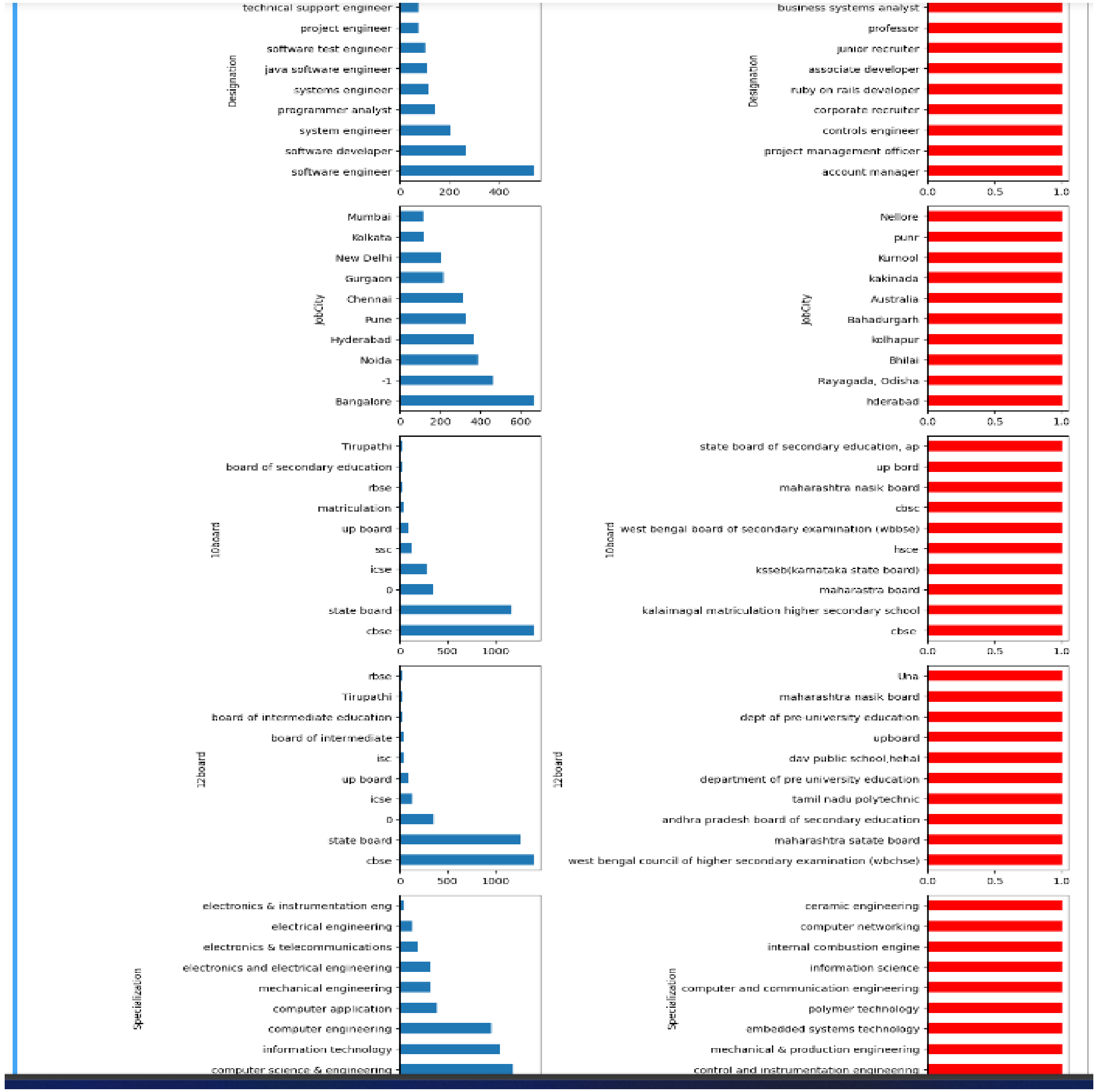


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 36 columns):
 #   Column                 Non-Null Count  Dtype         
---  ------                 --------------  -----         
 0   Salary                 3998 non-null   float64       
 1   DOJ                    3998 non-null   datetime64[ns]
 2   DOL                    3998 non-null   object        
 3   Designation            3998 non-null   object        
 4   JobCity                3998 non-null   object        
 5   Gender                 3998 non-null   object        
 6   DOB                    3998 non-null   datetime64[ns]
 7   10percentage           3998 non-null   float64       
 8   10board                3998 non-null   object        
 9   12graduation           3998 non-null   int64         
 10  12percentage           3998 non-null   float64       
 11  12board                3998 non-null   object        
 12  CollegeTier            3998 non-null   object        
 13  Degree                 3998 non-null   object        
 14  Specialization         3998 non-null   object        
 15  collegeGPA             3998 non-null   float64       
 16  CollegeCityID          3998 non-null   int64         
 17  CollegeCityTier        3998 non-null   object        
 18  CollegeState           3998 non-null   object        
 19  GraduationYear         3998 non-null   int64         
 20  English                3998 non-null   int64         
 21  Logical                3998 non-null   int64         
 22  Quant                  3998 non-null   int64         
 23  Domain                 3998 non-null   float64       
 24  ComputerProgramming    3998 non-null   int64         
 25  ElectronicsAndSemicon  3998 non-null   int64         
 26  ComputerScience        3998 non-null   int64         
 27  MechanicalEngg         3998 non-null   int64         
 28  ElectricalEngg         3998 non-null   int64         
 29  TelecomEngg            3998 non-null   int64         
 30  CivilEngg              3998 non-null   int64         
 31  conscientiousness      3998 non-null   float64       
 32  agreeableness          3998 non-null   float64       
 33  extraversion           3998 non-null   float64       
 34  nueroticism            3998 non-null   float64       
 35  openess_to_experience  3998 non-null   float64       
dtypes: datetime64[ns](2), float64(10), int64(13), object(11)
memory usage: 1.1+ MB
```

Data Cleaning and Manipulation

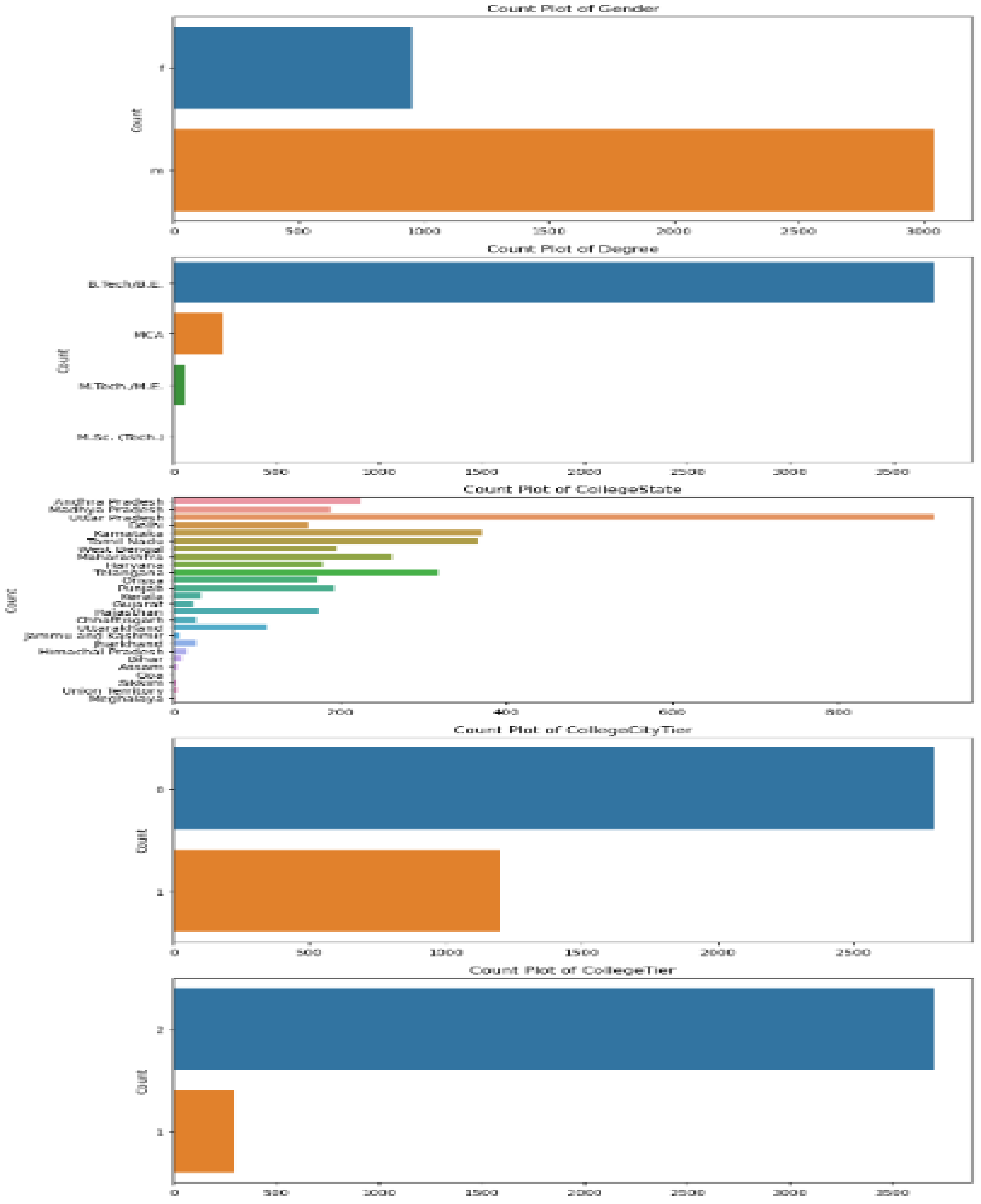
- Converted the columns conatining date of birth,date of joining to date time format.
- Created a new column age fro date of birth column.
- converted the CollegeTier and CollegeCityTier column to object datatype as they conatin discrete value.
- Applied Levenstein function to calculate to rearrange the JobCity,Designationand Specialization feature as they have high cardinality and many misspelled names by which we reduced the unique categories .
- Created New columns to calculate year of employment
- Separated the dataframe to analyze both working as well as those who have left the job.

Job City,10,12board,Specialization



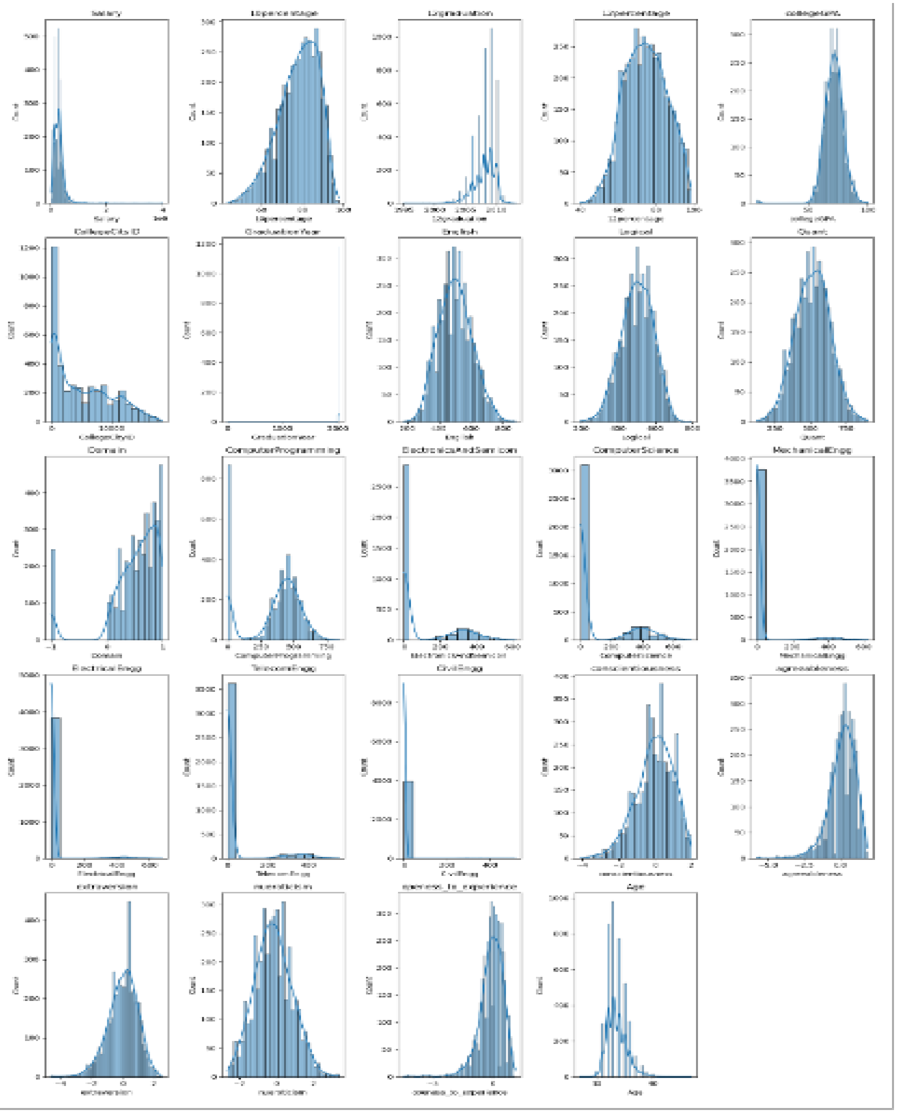
- Designation features has highest number in the category of Software Engineer which is 539,,265 software developers,205 system Engineers,130 Program analyst while Junior Software developer,human resource intern,senior quality asurance engineer and many more are one in number.
- In Job City we can see the mainly the metropolitan city if India like Bangalore,Noida,Hyderabad,Pune,Chennai,Gurgaon,New Delhi Kolkata are the places where most of the people are working while very few are working in Rayagardh,Bhilai,Kolhapur,Nellore.
- In the 10th and 12th board we can see that most of the employess have passed their high school and Intermediate from cbse,0 category board,icse while minimum numbers have passed from maharashtra state board,maharshtra nsik board,karnataka state board,west bengal seconadary education board.
- Maximum number of employees have done their specialization in Electronics and Communication Engineering,Computer Sciencee Engineering,Information Technology while very few have specialized in mechanical production,power system and automation,control and Instrumentaion.

Gender,Degree,CollegeState,CollegeTier,CollegeCityTier



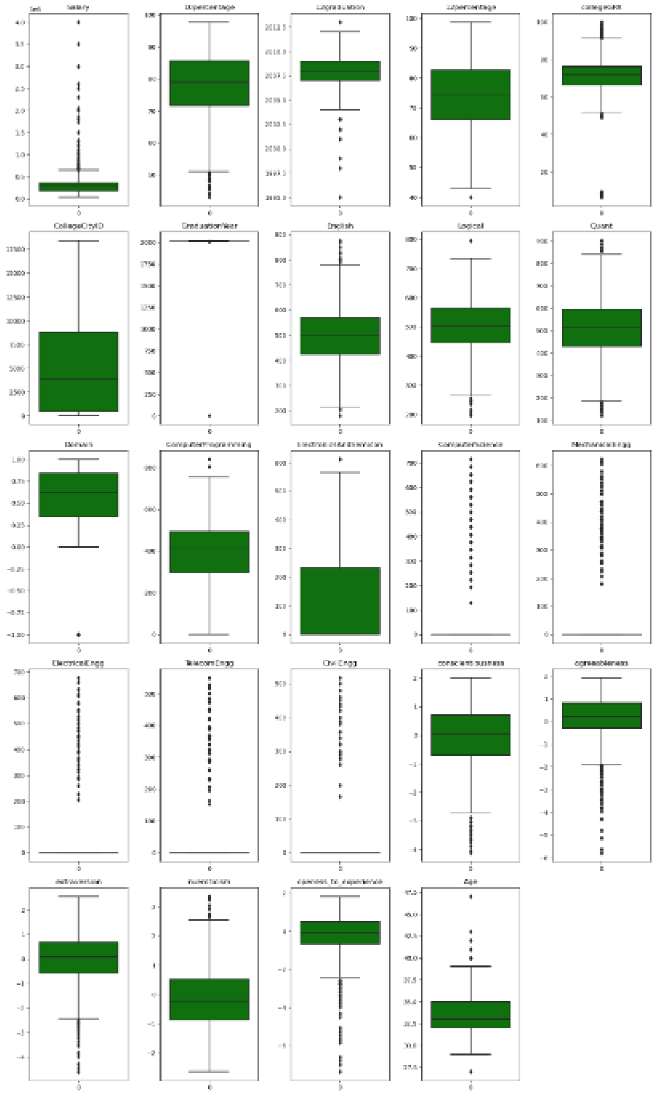
- 3,700 employes have done B.Tech / B.E ,243 have done MCA,53 have done MTech/M.E and two have done MSc .
- 3701 candidates are from college tier 2 while 297 are from College Tier 1
- 2797 are from college city tier while 1201 are from 1 College City Tier.
- Maximum number of College state is from Uttar Pradesh,Tamil nadu,Telangana ,Andhra Prdesh,West Bengal Punjab,MP,Haryana while least are from Goa,Meghalaya,Sikkim.
- 3041 are males and only 957 are females.

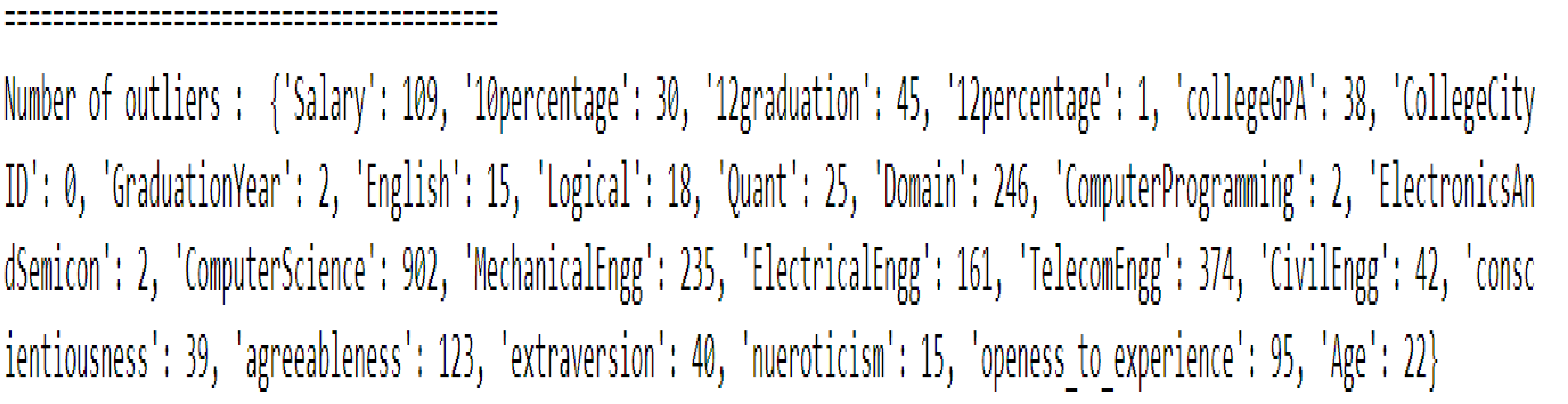
Distribution in Numerical Features



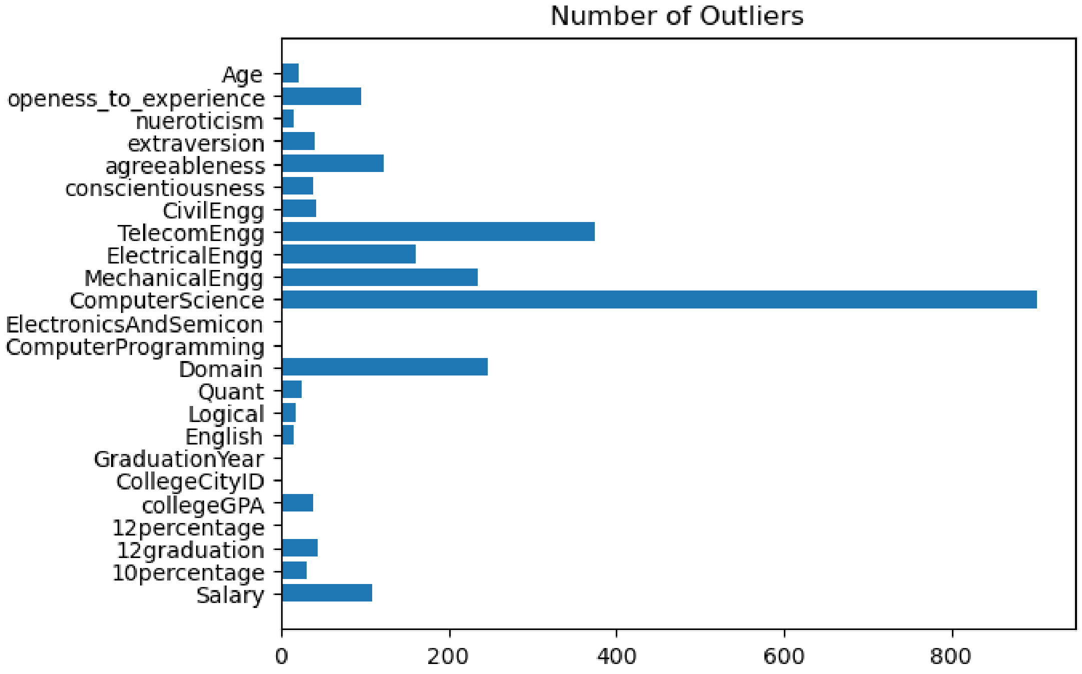
- Most of the continous features have outliers and that is the reason for the skewed distribution.
- ComputerScience Feature has more than 800 outliers while Telecomm,Mechanical,electrical<domain have around 200 outliers
- ElectronicsSemicon,ComputerProgramming,12percentage,CollegeCityId features does not have any outliers.
- High avearge value is recorded in English,Logical,Quant,Computer Programming and Domain while Electrical,Mechanical,Telecomm,Civil have low average marks.
- The English,Logical and quant section follows normal distribution.
- In ElectronicsAndSemicon,ComputerScience,Mechanical Engineering,Electrical Engieering,Telecom Engineering,Civil Engineering more than 2500 have 0 scores.
- In ComputerProgramming around 800 have zero score and around 250 have zero score score in the domain.
- The Salary ranges from 4000000 to35000.The average salary is 30,76,999 while median is 300000 .
- The average 10th percentage is 77% while highest is 97% and lowest is 43%.
- The average 12th percentage is 74% with 98.7 being the highest and 40 the lowest.
- The average college GPA is 71% ,minimum is 6.45 and maximum is 99.93
- Most of the employees have graduated in the year 2012 and 2013
- Average value of conscientiousness is -0.03,average value in agreeableness is 0.14,avaerage value in extraversion is 0.002,average value in neuroticism is -0.16,average value in openness to experience is -0.138.
- Most of the employees have age of 33 while 47 is the highest age and 27 is the lowest

Outliers

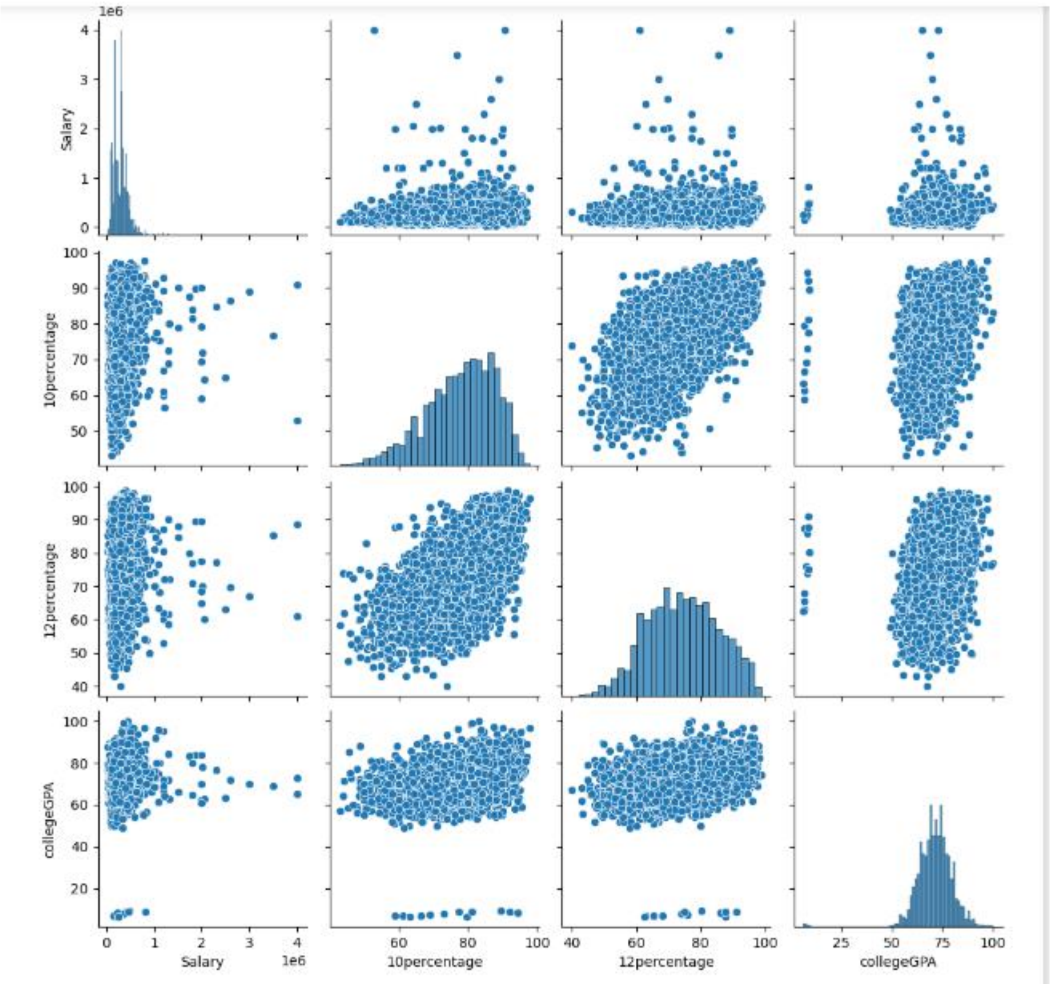


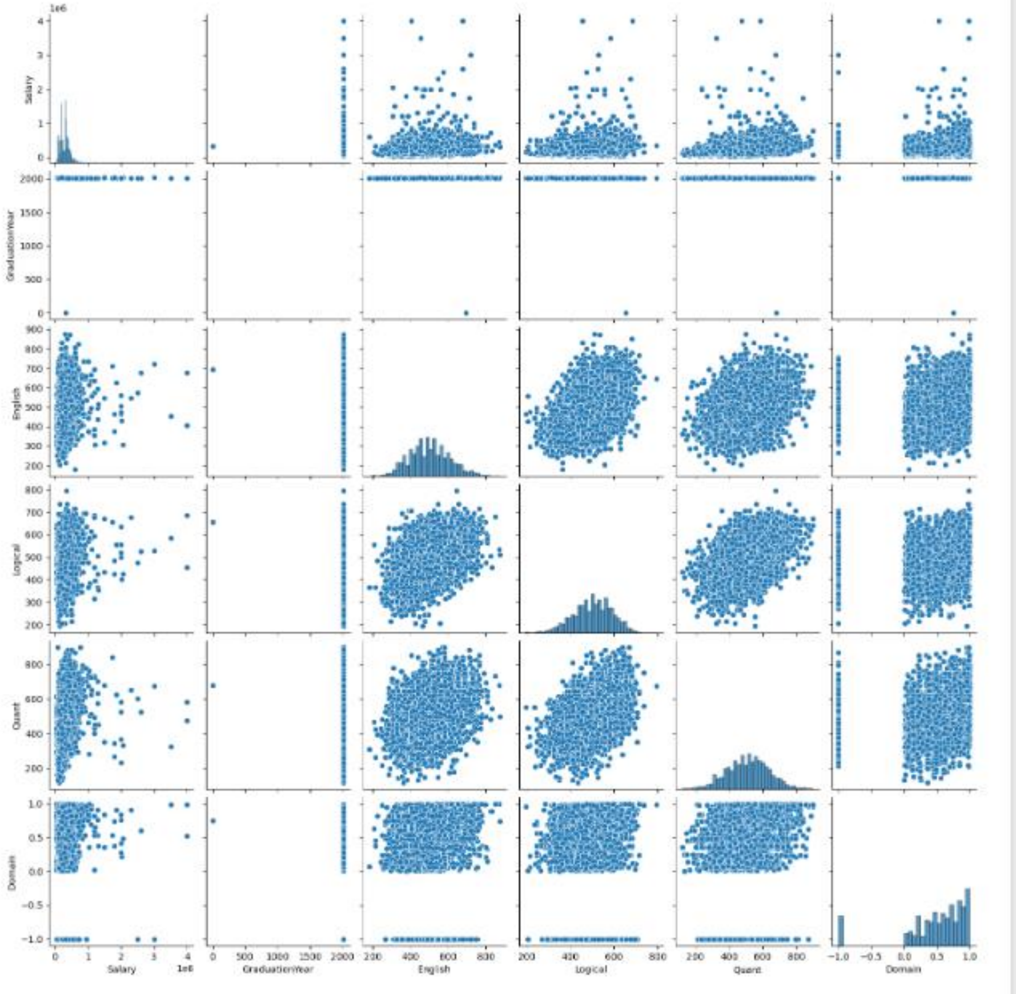


```
========================================
Number of outliers :  {'Salary': 109, '10percentage': 30, '12graduation': 45, '12percentage': 1, 'collegeGPA': 38, 'CollegeCity
ID': 0, 'GraduationYear': 2, 'English': 15, 'Logical': 18, 'Quant': 25, 'Domain': 246, 'ComputerProgramming': 2, 'ElectronicsAn
dSemicon': 2, 'ComputerScience': 902, 'MechanicalEngg': 235, 'ElectricalEngg': 161, 'TelecomEngg': 374, 'CivilEngg': 42, 'consc
ientiousness': 39, 'agreeableness': 123, 'extraversion': 40, 'nueroticism': 15, 'openess_to_experience': 95, 'Age': 22}
```

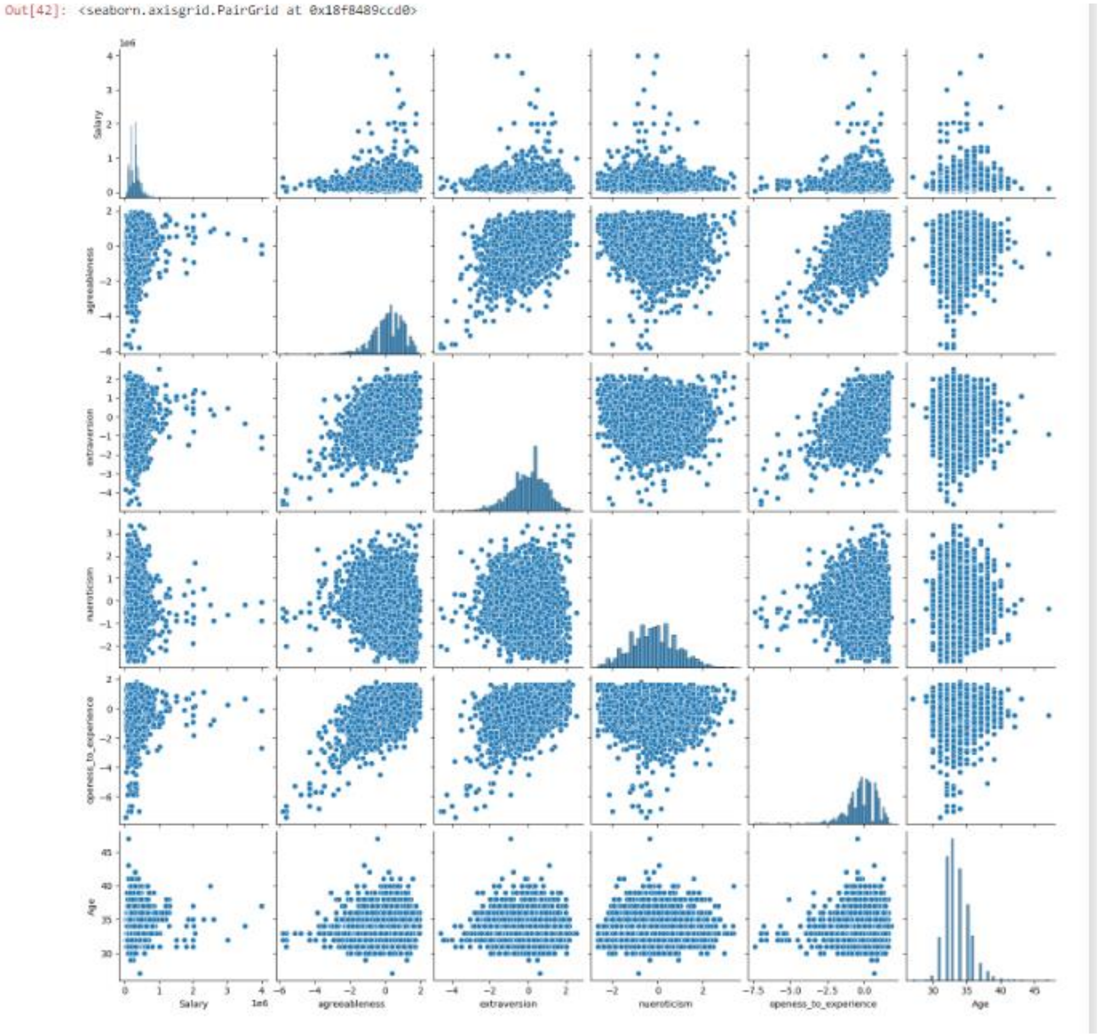


Bivariate Analysis between Numerical Columns





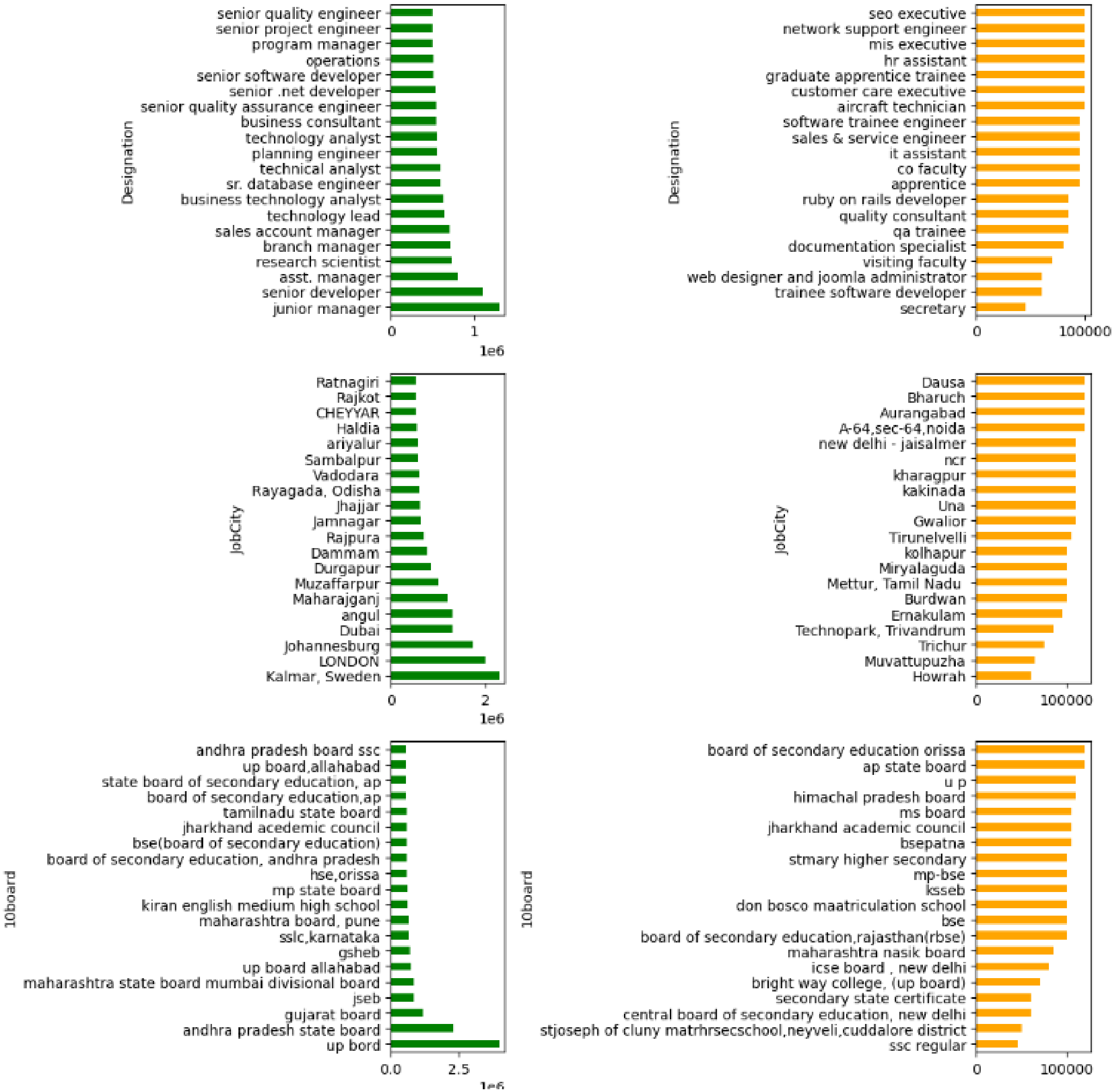
Bivariate Analysis between Numerical Columns

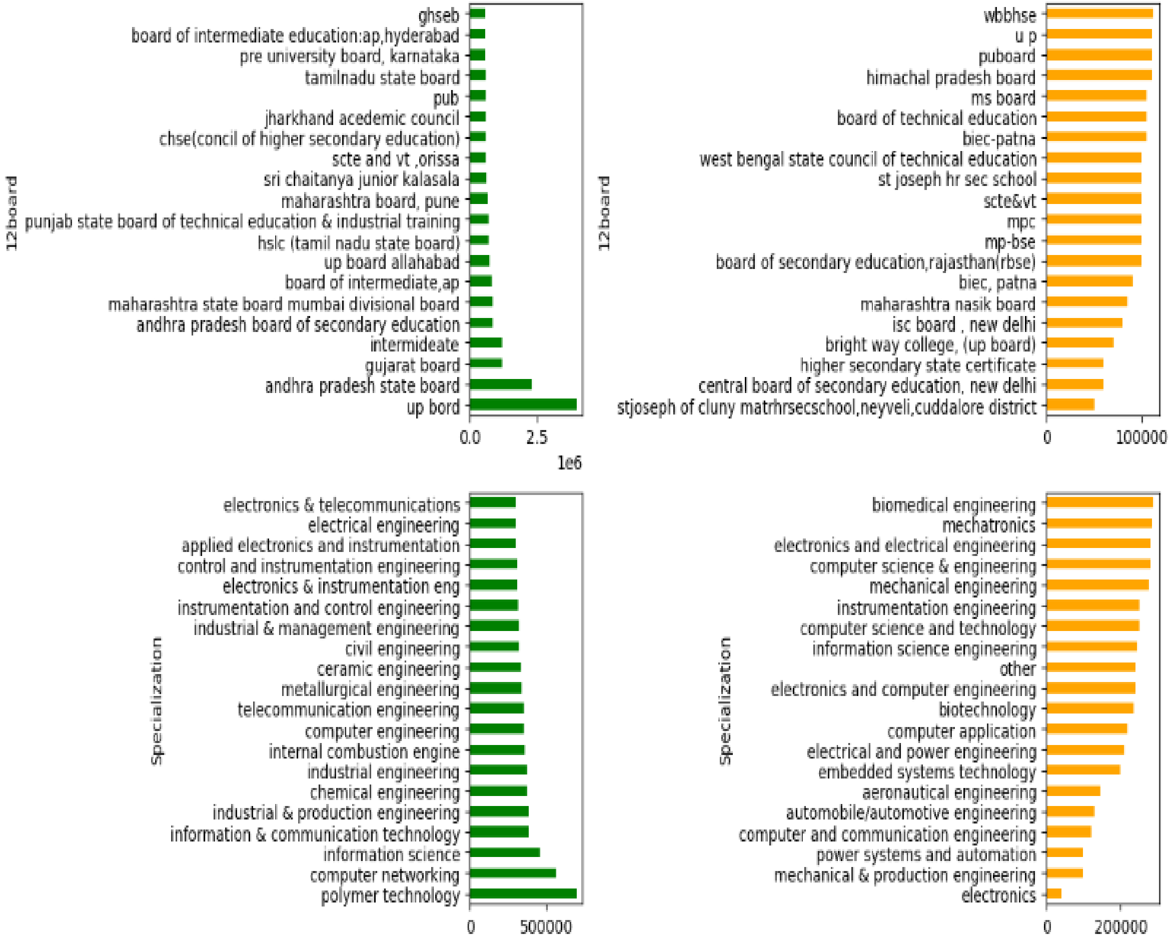


```
Out[42]: <seaborn.axisgrid.PairGrid at 0x18f8489ccd0>
```

- From the above scatter plots we can see that there is no relationship between salary with any of the features .
- There seems to positive linear relationship between 10th and 12th percentage which means those who have scored high in 10th are most likely to score high in 12th as well.
- English,Logical,Quant all have positive linear relationship with each other .Candidates who have scored high in English will have high marks in Logical and Quants as well and vice-versa.
- Agreeableness has linear poistive relation with extraversion and openness to experience .
- Agreeableness have linear negative relation with neuroticism which means that people having high agreeableness are most likely to have high openesst o experience and extraversion but they are less likely to have high neuroticism.
- Age Column is having no relationship with any other feature

Bivariate Analysis Between Numerical and Categorical Columns

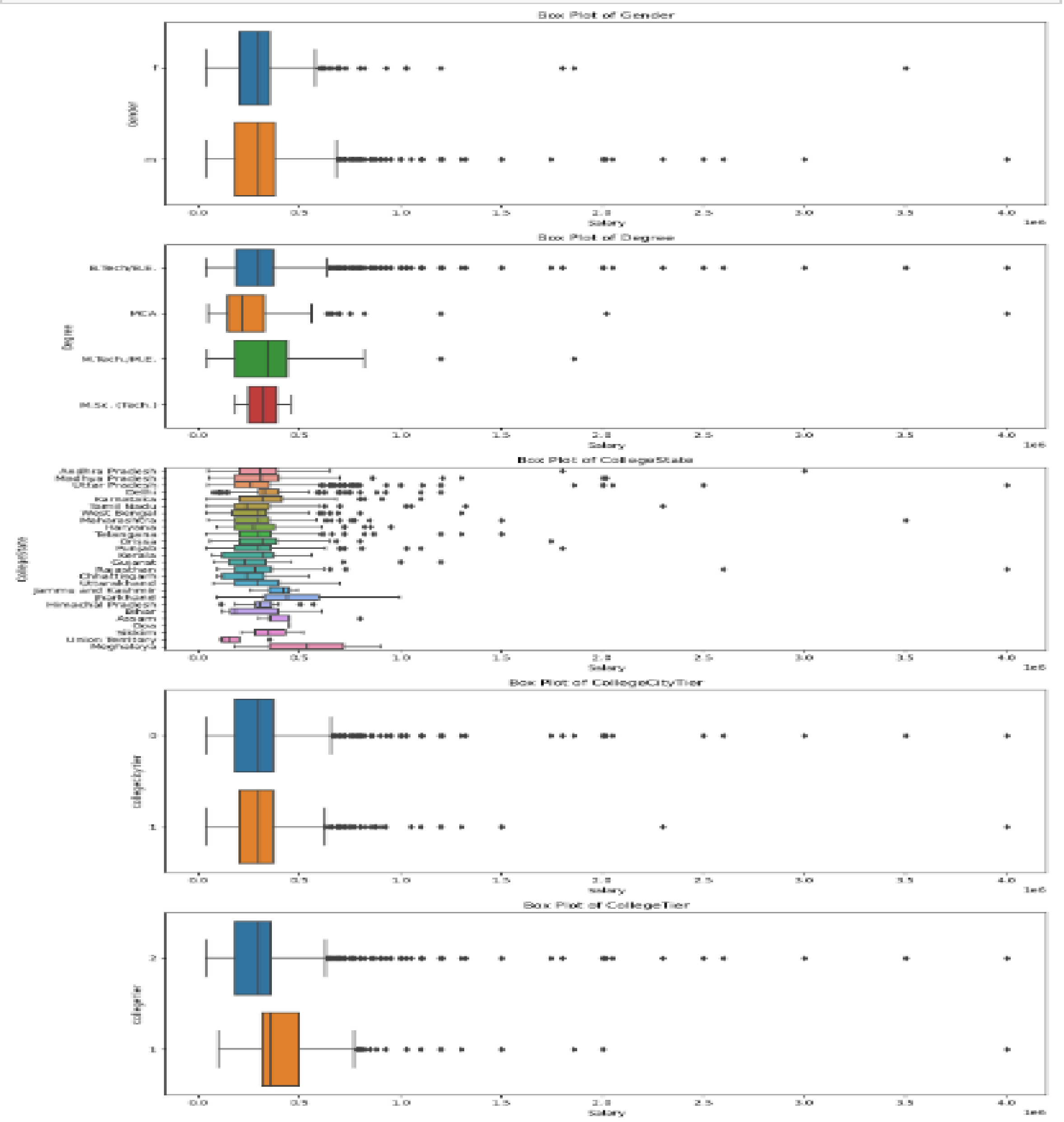




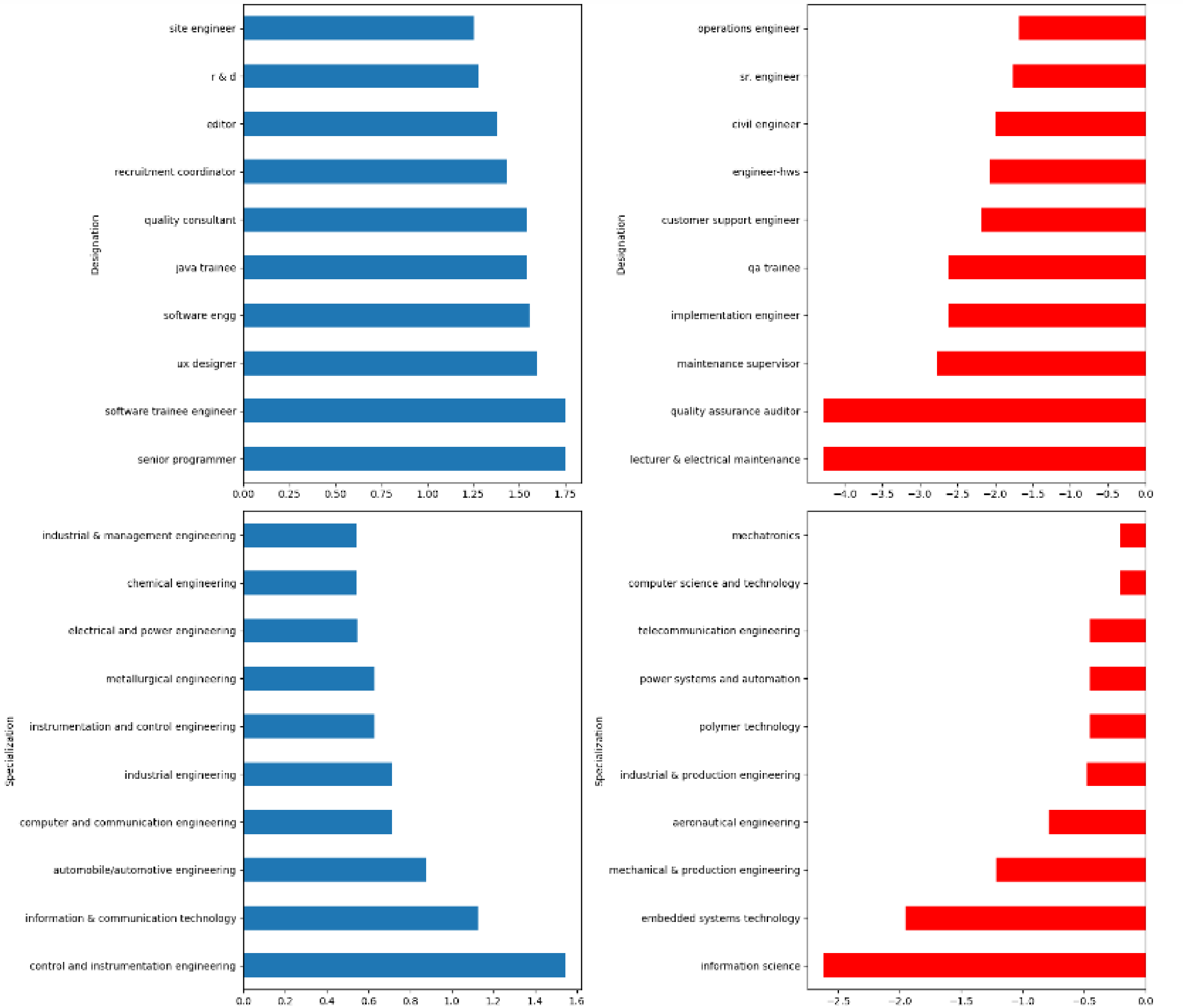
INNOMATICS
RESEARCH LABS

Bivariate Analysis Between Numerical and Categorical Columns

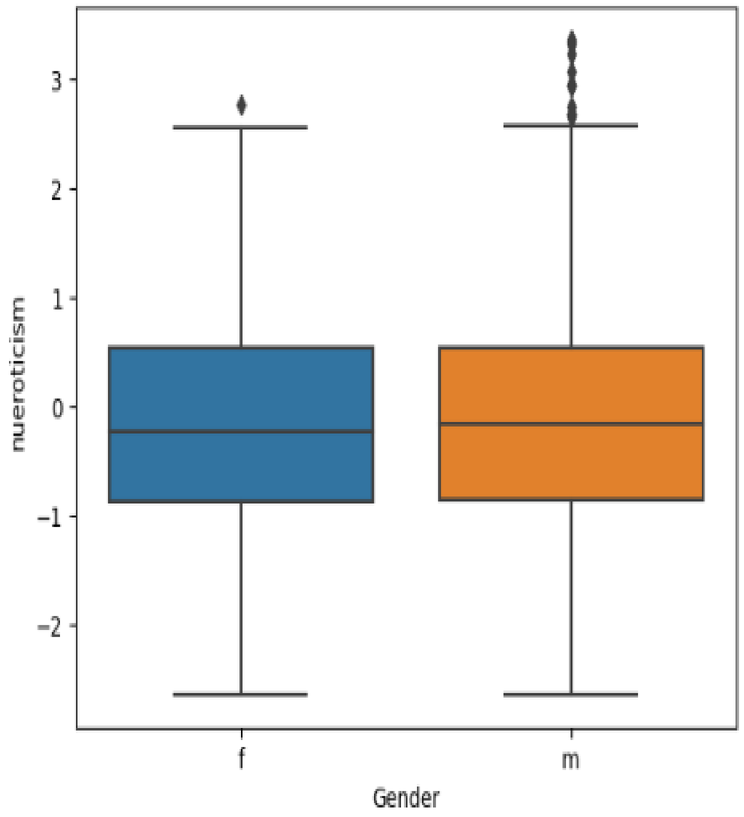
- Managerial and senior positon holders like Junior Manger,Senior Developer,Research scientist,Assistant manager,Branch Manager,senior Software developer have high income or salary while trainee position holders like secretary,trainee,visiting faculty,co faculty technician,qa trainee,hr assistant are having low salary.
- Candidates doing job in cities like Kalmar,Sweden,London,Johannesburg,Dubai,Dammam are having high income while those having jobs in cities like Howrah,Tichur,Budwan,Kolhapur are having low income.
- Candidates who have done their 10th and 12th board from UP board,Gujarat Board,Andhra Pradesh bard,jseb are having high income while those who have completed their 10th and 12th from cbse New Delhi,maharashtra nasik board,icse and isc New Delhi are having low income.
- Those who have done their specialization in Polymer Technology,Computer Networking,Information Science,Information and communication technology are earning the highest salary however those in electronics,mechanical and production engineering,power system and automation,automobile engineering are having lowest income.
- The median salary in both the genders are equal.
- Those who have done MTech / MSc(Tech) are having high median salary while MCA degree holders have lowest median salary.
- Those who have graduated from college Tier 1 are having high median salary.



NeuroticismVs Specialization,Designation and Gender

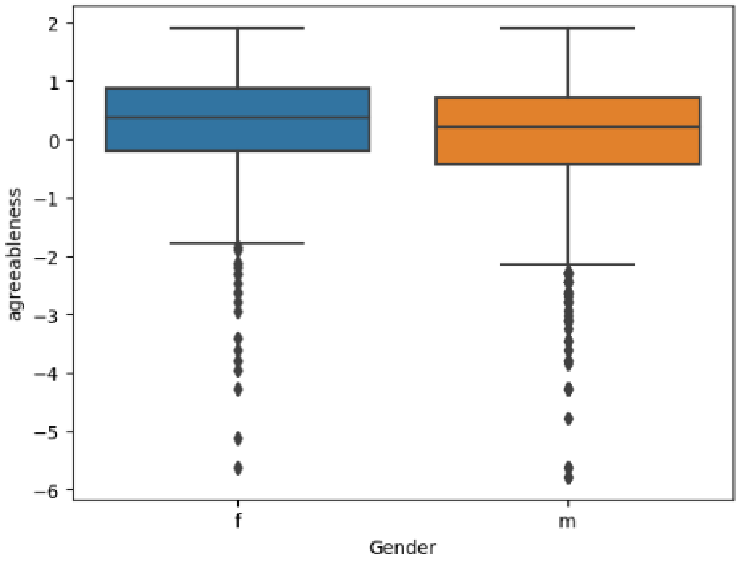


```
]: <AxesSubplot:xlabel='Gender', ylabel='nueroticism'>
```

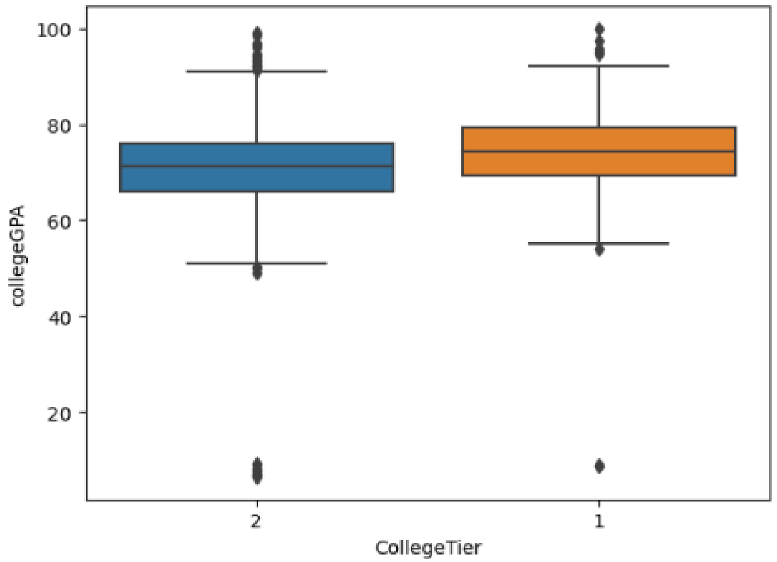


Insights

```
Out[55]: <Axes: xlabel='Gender', ylabel='agreeableness'>
```



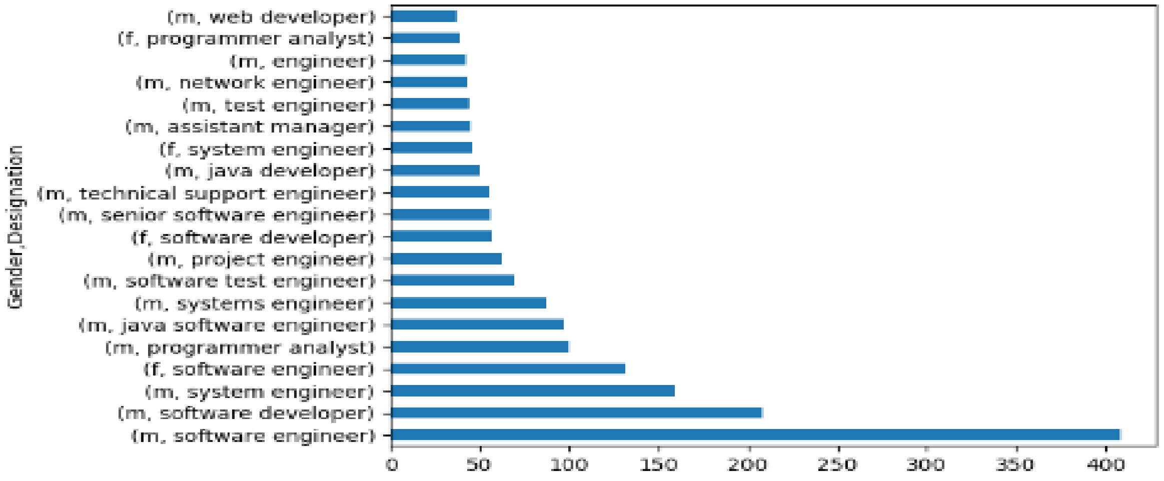
```
: <Axes: xlabel='CollegeTier', ylabel='collegeGPA'>
```



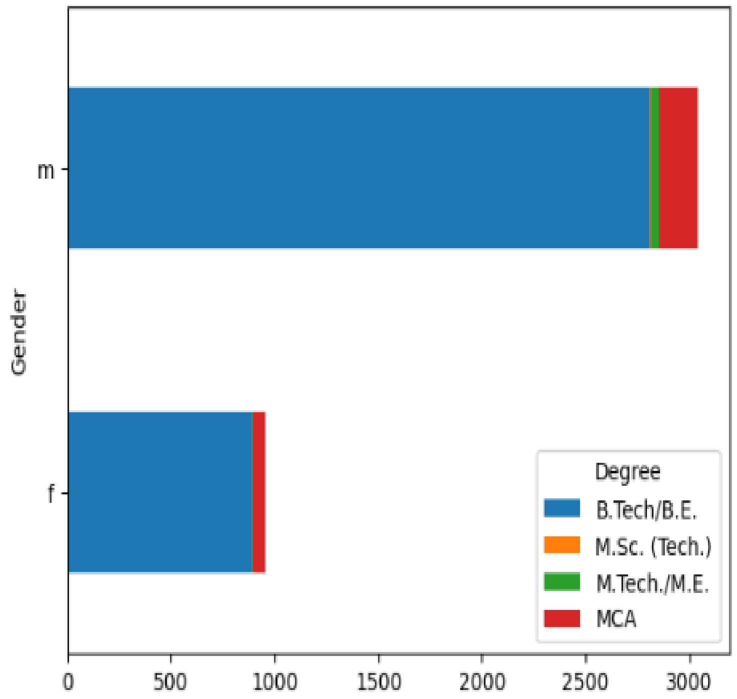
- Project Management Officer,full time loss prevention associate,telecom support engineer,digital marketing specialist,IT specialist are having high neuroticism whereas quality assurance editor,editor,document specialist,web intern,web designer are having low neuroticism.
- Candidates having specilization in Mechanical and automation,Mechanical production,Industrial Production are having high neuroticism.Those having specialization in Computer,communication,electronics,IT,controland instrumentation are having low neuroticism.
- Females and males are both having same median value of neuroticism.
- Those having designation as Senior Programmer,Software trainee engineer,UX designer have high agreeableness.
- Those having specialization in Control and Instrumentation have high agreeableness.
- Candiates belonging to College Tier 1 have high College GPA as compared to those belonging to college tier 2 .

Bivariate Analysis On Categorical vs Categorical Columns

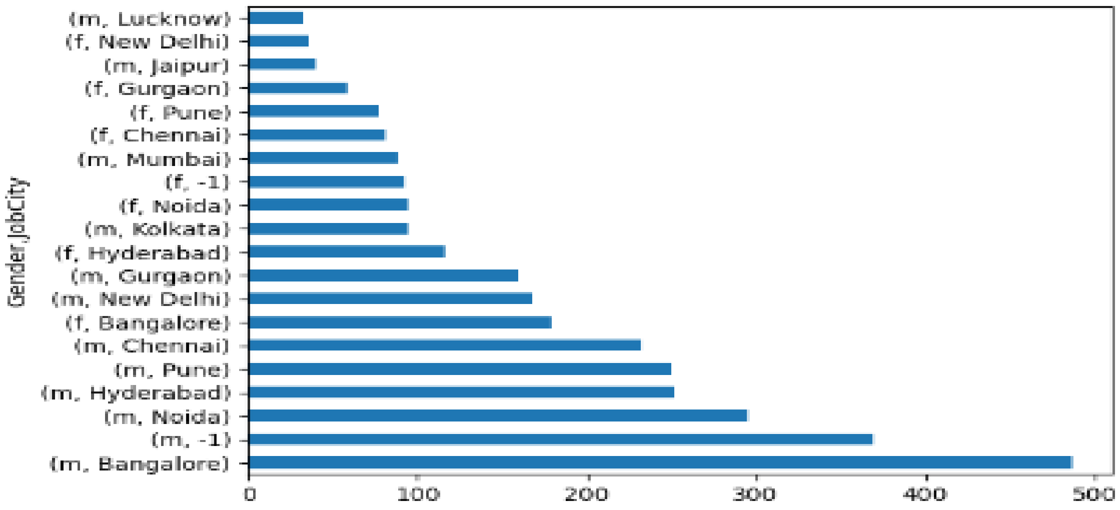
```
]: <Axes: ylabel='Gender,Designation'>
```



```
<Axes: ylabel='Gender'>
```

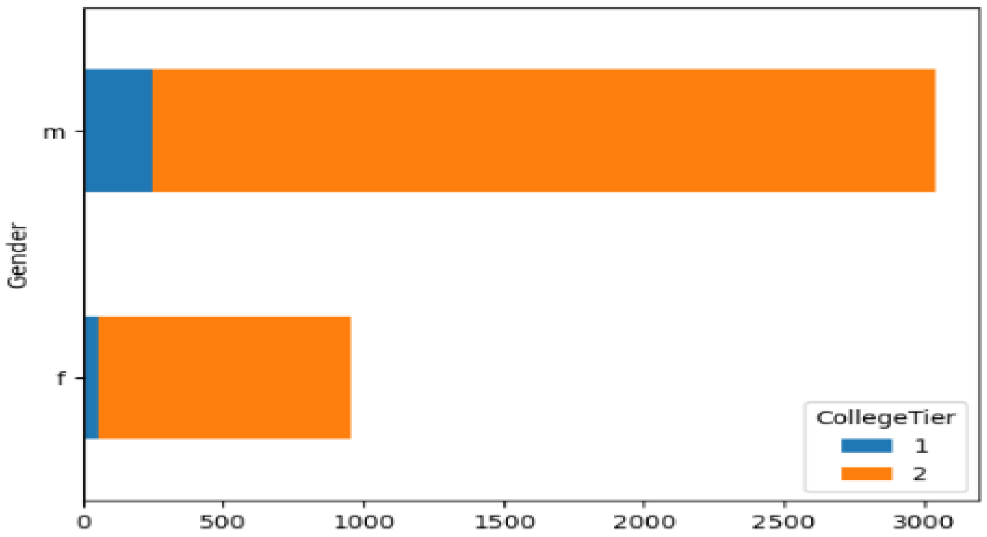


```
]: <Axes: ylabel='Gender,JobCity'>
```

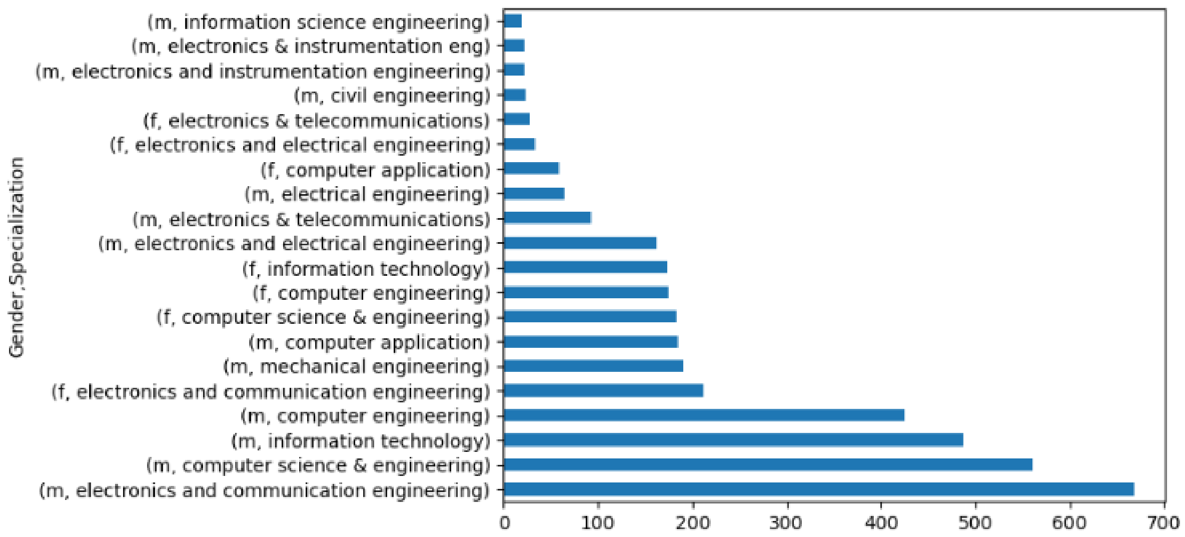


Bivariate Analysis On Categorical vs Categorical Columns

```
<Axes: ylabel='Gender'>
```



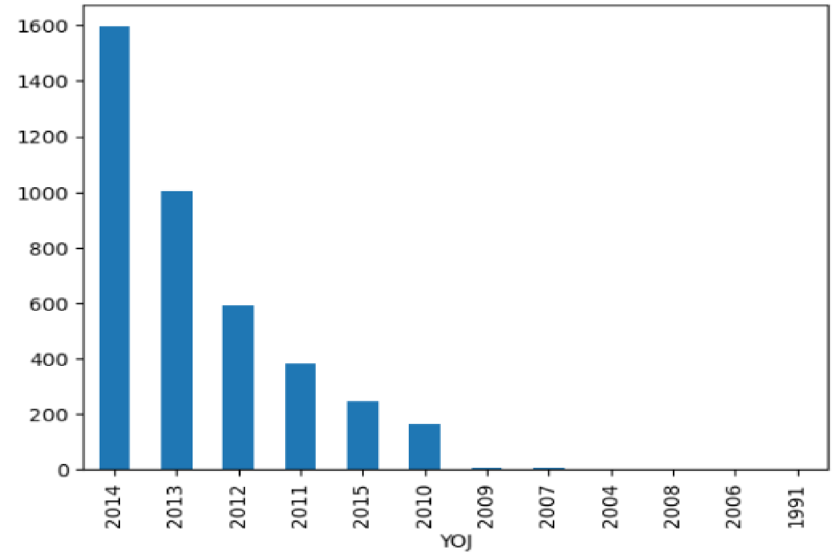
```
<Axes: ylabel='Gender,Specialization'>
```



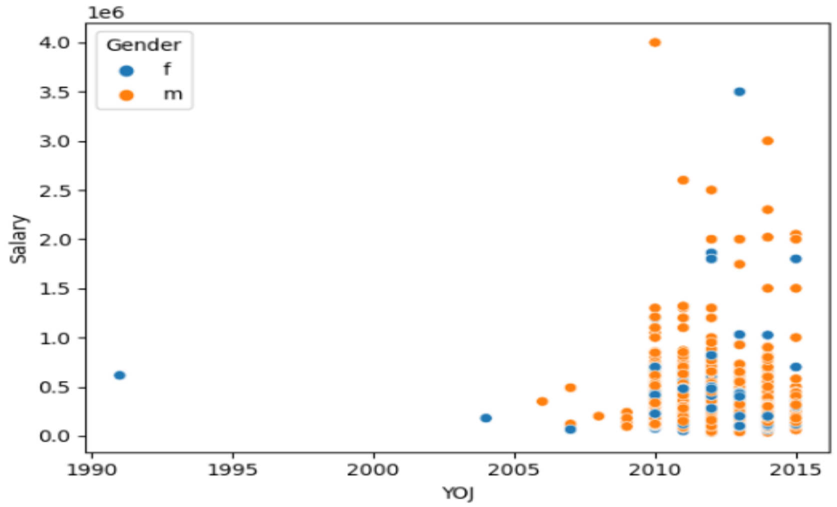
- Most of the males are Software engineer,Software Developer,Programme Analyst whereas Most of the Females are also Software Engineer,Software developer,System engineer,Programme Analyst
- Maximum Number of males and females have done Btech , after that maximum number in both gender are of MCA graduates.
- Most of the males are in Bangalore,Noida,Hyderabad and most of the females are in Bangalore,Hyderabad,Chennai.
- Distribution of Specialization on the higher side is also similar in both males and females .
- Most of the males as well as females belong to college tier 2
- Top 50 specialization vs designation belongs to electronics and computer related fields

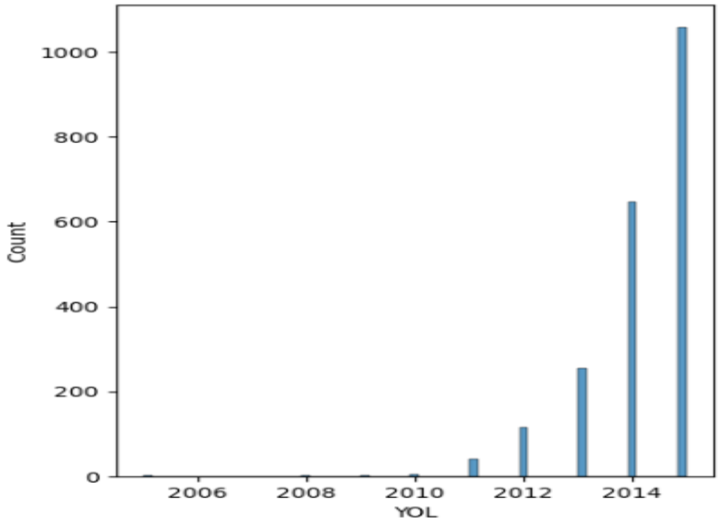
DateTime Columns Analysis

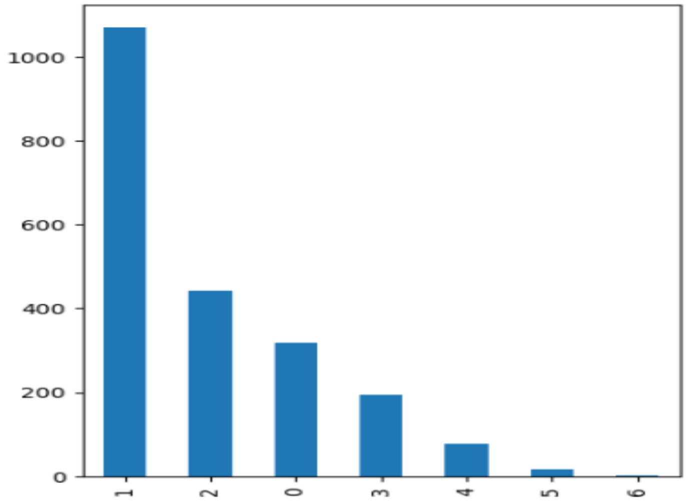
```
<Axes: xlabel='YOJ'>
```



```
<Axes: xlabel='YOJ', ylabel='Salary'>
```







- If we look at the year of leaving then maximum candidates have left in the year 2014 and 2013.
- More than 1000 candidates have worked only for a year and left the company followed by those who have worked for two years who are less than 500 in number.
- Around three hundred candidates have left without even completing one year

Conclusion

- The data is highly imbalanced if we consider gender into account.
- The most popular field of Specialization seems to be electronics and computer engineering,Information Technology.
- The most popular designation also belong in the field of Information Technology and Software development.
- Most of the Jobs are in the metropolotan cities of India like Bangalore,Hyderabad,chennai,pune,NewDelhi,Mumbai
- Those working outside India have very high income as compared to rest of candidates.
- Most of the candidates have started working after 2010 as it is pretty clear because of them have completed their graduation in 2012,2013 .
- Most of the candidates who have left the job have 1 year of experience in that job.
- There is relationship between gender and Specialization columns

THANK YOU