INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

AMCAT Exploratory Data Analysis Project

About me

- **Background:** I hold a Bachelor's degree in Mechanical Engineering (B.Tech). However, my passion lies in the rapidly growing field of Data Science, which I find immensely promising in terms of its potential applications across various industries.

- **Motivation for Learning Data Science:** My interest in Data Science stems from my fascination with technology and data-driven decision-making. I believe that Data Science offers boundless opportunities for innovation and problem-solving, making it an ideal field for me to pursue my career ambitions.

- **Work Experience:** I have gained practical experience in the field of Data Science through my role as a Business Development Associate at Byju's, where I had the opportunity to work closely with data-driven strategies and analytics. Additionally, I have completed two internships focused on Data Science, where I worked on real-world projects, further honing my skills and understanding of the field.

- **LinkedIn Profile:** https://www.linkedin.com/in/sanchit-singla/

- **GitHub Profile:** https://github.com/sa-1-2/

Objective of the Project

- The objective of this project is to conduct an in-depth Exploratory Data Analysis (EDA) on the Aspiring Mind Employment Outcome 2015 (AMEO) dataset, focusing specifically on engineering graduates. The dataset comprises employment outcomes of engineering graduates, including dependent variables such as Salary, Job Titles, and Job Locations, as well as standardized scores from three different areas: cognitive skills, technical skills, and personality skills. Additionally, the dataset contains demographic features.
- The primary goals of the project are as follows:

1. Exploratory Data Analysis (EDA): Conduct comprehensive exploratory data analysis to gain insights into various aspects of the dataset.
2. Insight Generation: Extract meaningful insights and patterns from the data to understand the employment outcomes of engineering graduates.
3. Insightful Reporting: Summarize the findings and insights derived from the EDA process in a clear and concise manner

Summary of the data

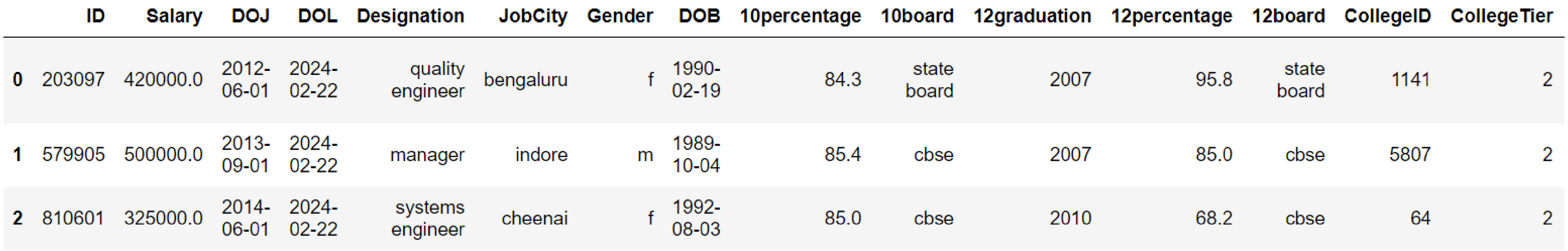
The Aspiring Mind Employment Outcome 2015 (AMEO) dataset contains comprehensive information on the employment outcomes of engineering graduates, with a focus on various factors influencing their career trajectories. Below is a summary of the dataset:

- The dataset is primarily limited to students with engineering disciplines.
- It comprises 3998 data points, each representing an individual engineering graduate.
- There are around 37 independent variables, and 1 dependent variable included in the dataset, encompassing both continuous and categorical features.
- Each data point is associated with a unique identifier for candidate identification purposes.
- There are 10 float, 17 int, and 12 object variables.
- Salary variable is a dependent variable lies in range 35K to 4000K.

Exploratory Data Analysis

❖ *Data Cleaning Steps*

1. Checked for null values and duplicated entries, and Unwanted columns like 'Unnamed: 0' were removed from the dataset.
2. Converted datetime columns such as 'DOJ' (Date of Joining), 'DOB' (Date of Birth), and 'DOL' (Date of Leaving) to the appropriate datetime format..
3. Rectified inconsistencies in columns like 'Designation', 'JobCity', '10board', '12board', and 'Specialization'..
4. Transformed columns such as 'collegeGPA' and 'DOL'to improve data consistency and accuracy..
5. Filled missing values for categorical features with the label "Missing".
6. After completing the data cleaning process, the dataset now appears as follows:



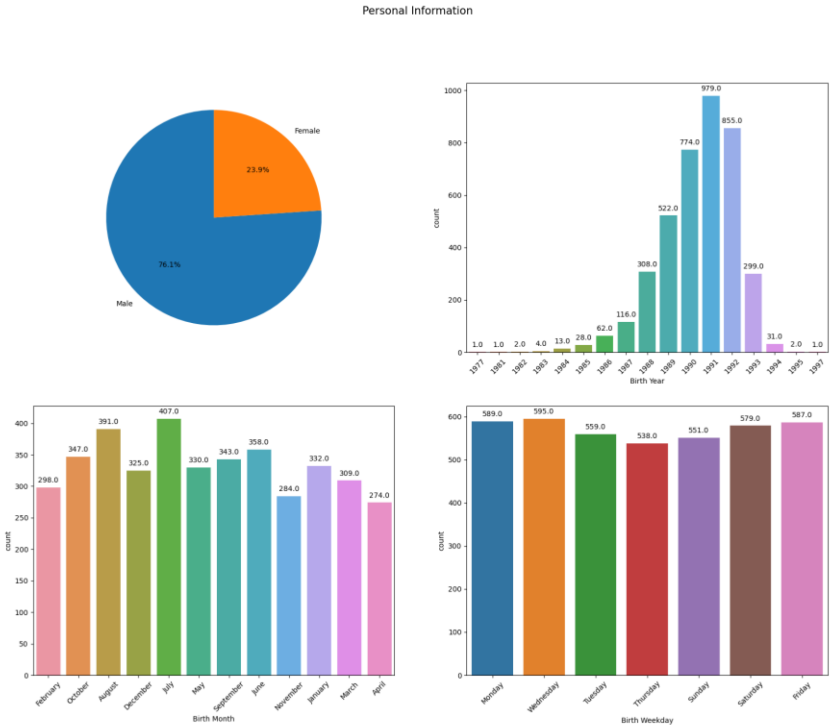
	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage	12board	CollegeID	CollegeTier
0	203097	420000.0	2012-06-01	2024-02-22	quality engineer	bengaluru	f	1990-02-19	84.3	state board	2007	95.8	state board	1141	2
1	579905	500000.0	2013-09-01	2024-02-22	manager	indore	m	1989-10-04	85.4	cbse	2007	85.0	cbse	5807	2
2	810601	325000.0	2014-06-01	2024-02-22	systems engineer	cheenai	f	1992-08-03	85.0	cbse	2010	68.2	cbse	64	2

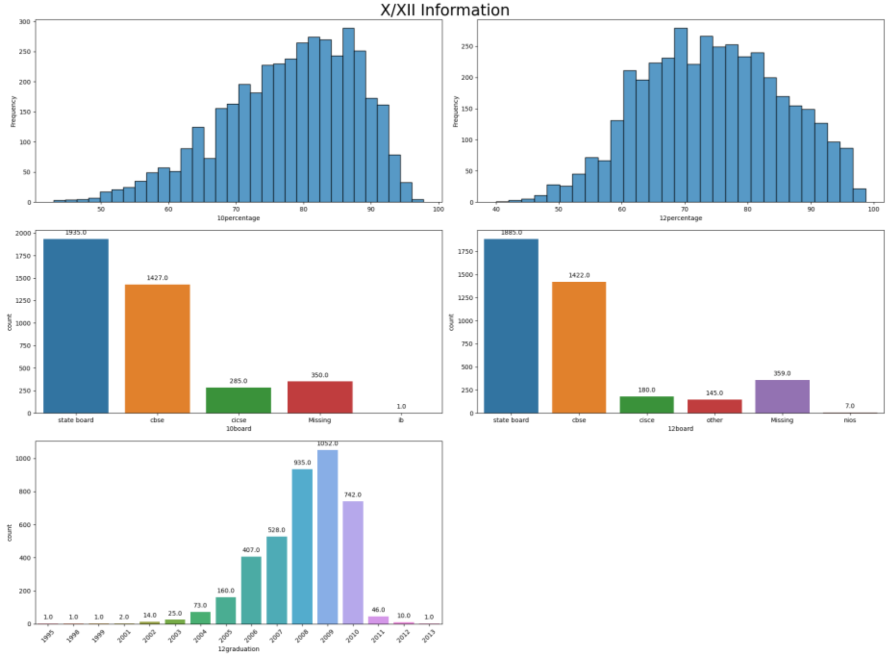
Exploratory Data Analysis

❖ *Univariate Analysis*

Performed Univariate Analysis to check the outliers, skewness in data, Range of values, different types and distribution of categorical values.

Some of the univariate Analysis Graphs are shown below:





INNOMATICS RESEARCH LABS

Insights from Univariate Analysis

❑ ***Personal Information***

- 23.9% of candidates are female, while 76.1% are male.
- Approximately 60% of candidates were born between the years 1990 and 1992.

❑ ***X/XII Information***

- The 10th percentile distribution is left-skewed, indicating that nearly 50% of students scored greater than or equal to 80.
- Similarly, the distribution of 12th percentile scores is also left-skewed, suggesting that only approximately 33.3% of students achieved scores of 80 or higher.

❑ ***College Information***

- 92.5% of candidates attended tier 2 colleges.
- 70% of candidates attended colleges located in tier 2 cities.
- Approximately 22% of candidates graduated from colleges in Uttar Pradesh.
- 50% of candidates achieved a GPA greater than 2.8.
- 92.5% of candidates hold a B.Tech/B.E degree.

Insights from Univariate Analysis

❑ ***Job Information***

- The majority of candidates are based in Bengaluru, Noida, Hyderabad, Pune, Chennai, and Delhi/NCR.
- Approximately 50% of candidates hold positions as Software Engineers, Developers, or in various other engineering fields.
- The median salary for candidates is 3 LPA, with a right-skewed distribution indicating a higher number of outliers.
- Nearly 23% of candidates commenced their jobs in the months of July and August.
- About 32% of candidates departed from their positions in the month of April.

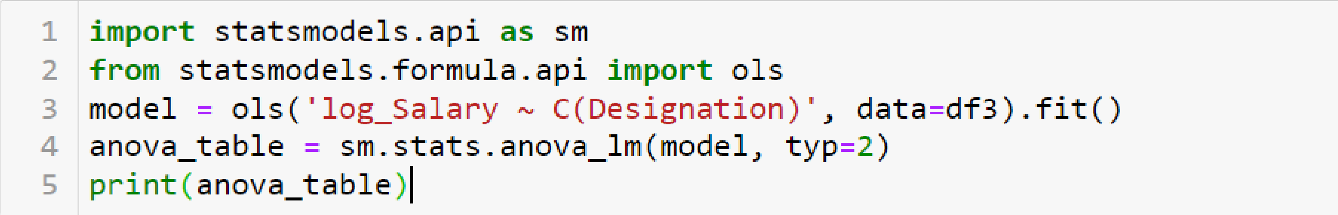
Bivariate Analysis

- Performed Bivariate Analysis to Examine Relationships Between Different Feature Types:

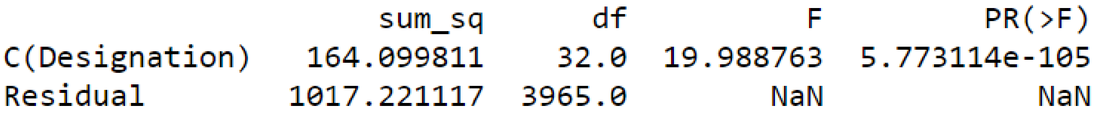
❖ ***Numeric Feature with Categorical Features:***

- Explored the relationship between numeric and categorical features using techniques such as:

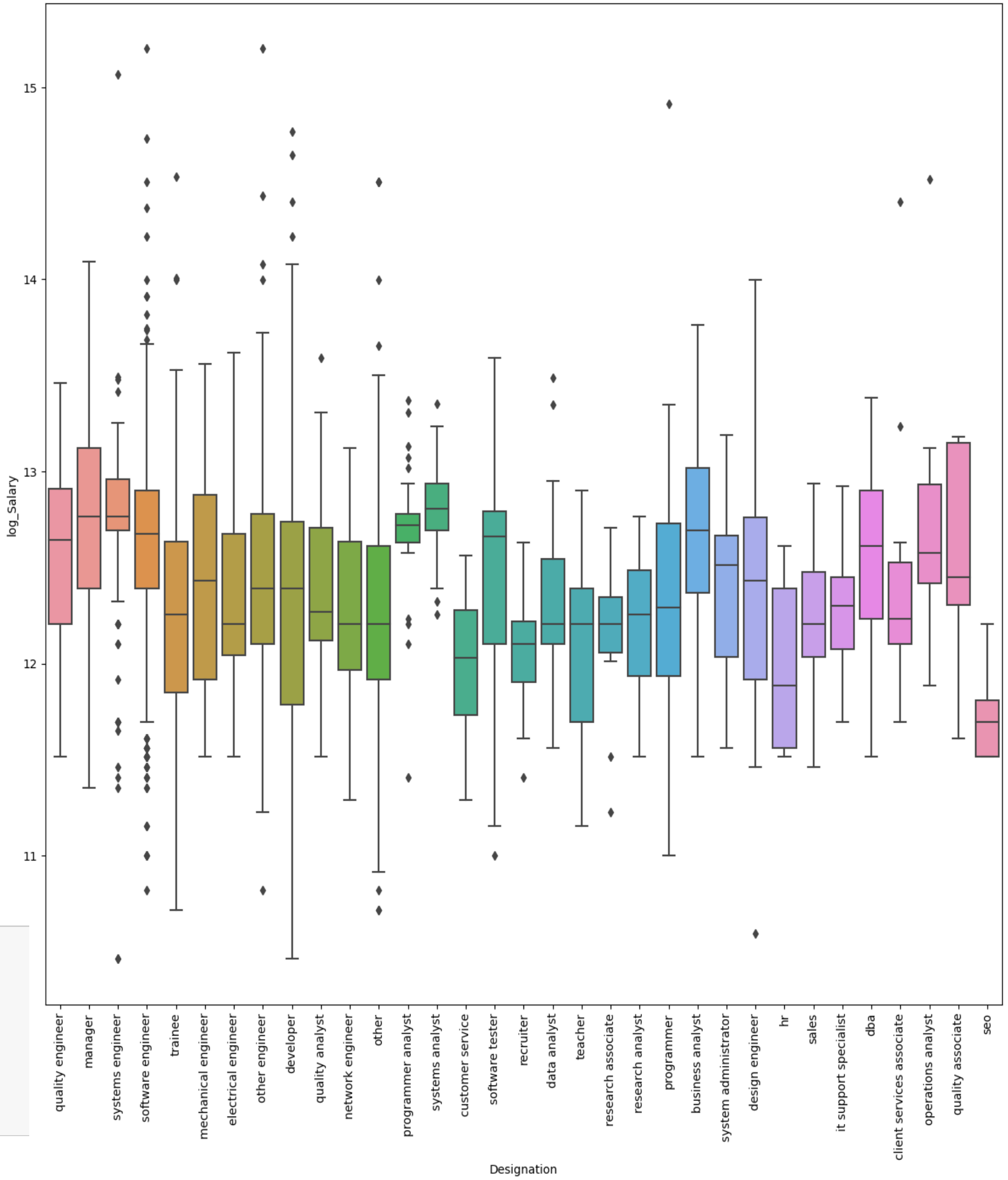
1. Box plots: Examined the distribution of numeric values across different categories of categorical variables
2. ANOVA test or independent ttest: Conducted analysis of variance to determine significant differences in numeric values across different categories of the categorical variable.



```python
import statsmodels.api as sm
from statsmodels.formula.api import ols
model = ols('log_Salary ~ C(Designation)', data=df3).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)
```



```
                     sum_sq      df          F         PR(>F)
C(Designation)   164.099811    32.0  19.988763  5.773114e-105
Residual        1017.221117  3965.0        NaN            NaN
```

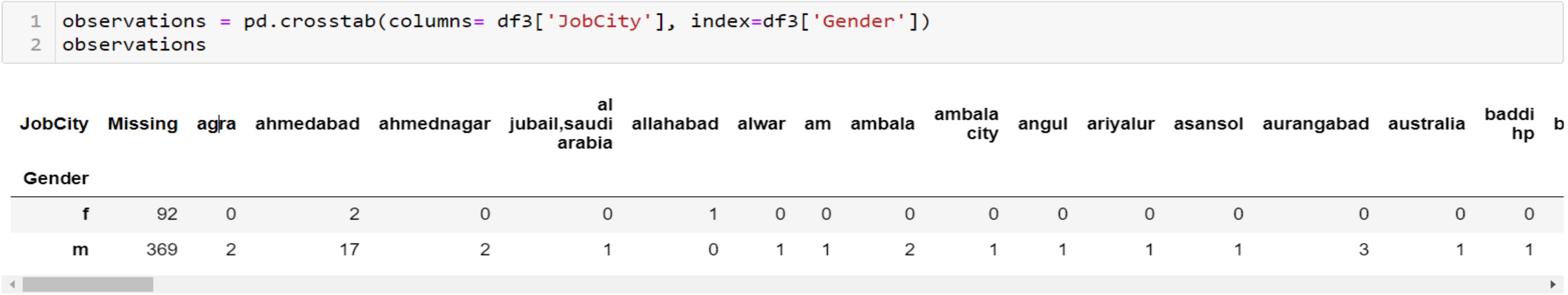


Bivariate Analysis

❖ ***Categorical and Categorical Features:***

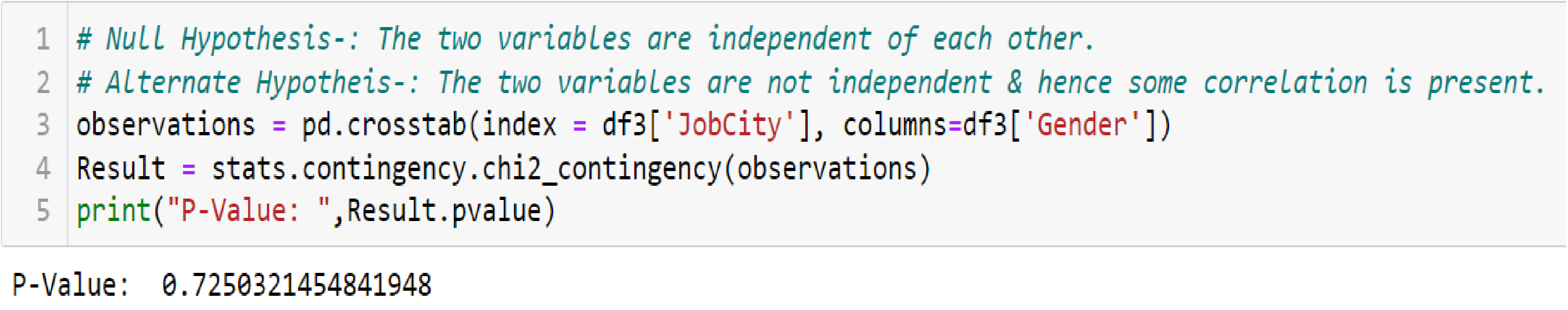
- Investigated the relationship between two categorical features through:

1. Crosstabulation: Generated contingency tables to display the frequency distribution of one categorical variable against another.
2. Chi-square test: Conducted a statistical test to assess the association between two categorical variables.



```python
observations = pd.crosstab(columns= df3['JobCity'], index=df3['Gender'])
observations
```

JobCity	Missing	agra	ahmedabad	ahmednagar	al jubail,saudi arabia	allahabad	alwar	am	ambala	ambala city	angul	ariyalur	asansol	aurangabad	australia	baddi hp	b
Gender																	
f	92	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	
m	369	2	17	2	1	0	1	1	2	1	1	1	1	3	1	1	



```python
# Null Hypothesis-: The two variables are independent of each other.
# Alternate Hypotheis-: The two variables are not independent & hence some correlation is present.
observations = pd.crosstab(index = df3['JobCity'], columns=df3['Gender'])
Result = stats.contingency.chi2_contingency(observations)
print("P-Value: ",Result.pvalue)
```

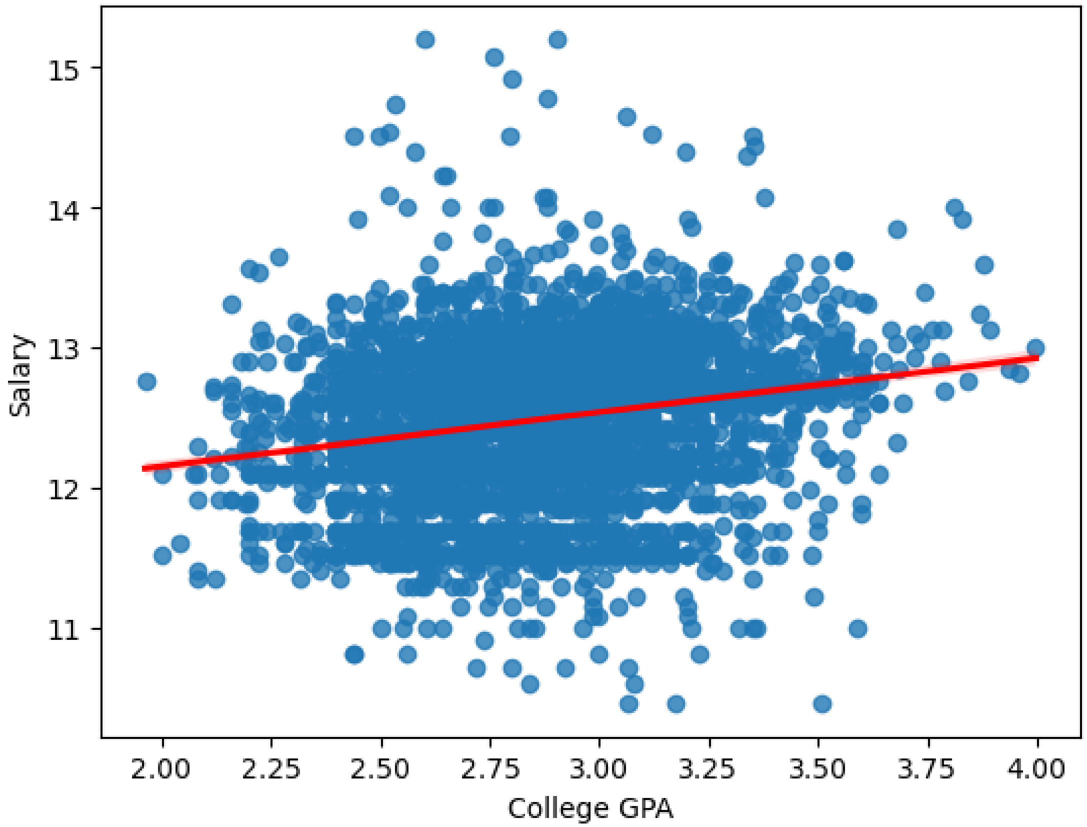
```
P-Value:  0.7250321454841948
```

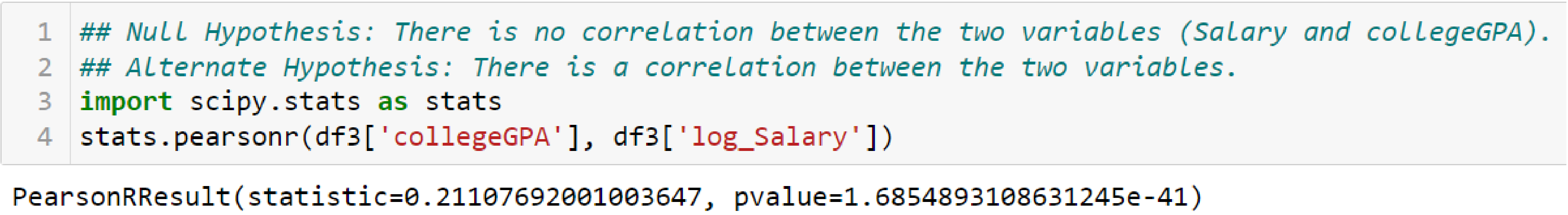
Bivariate Analysis

❖ ***Numerical and Numerical Features:***

- Examined the relationship between two numerical features using:

1. Scatter plots: Visualized the relationship between two numerical variables to identify patterns or correlations.
2. Correlation analysis: Calculated correlation coefficients (e.g., Pearson correlation) to quantify the strength and direction of the linear relationship between numerical variables.





```
## Null Hypothesis: There is no correlation between the two variables (Salary and collegeGPA).
## Alternate Hypothesis: There is a correlation between the two variables.
import scipy.stats as stats
stats.pearsonr(df3['collegeGPA'], df3['log_Salary'])
```

```
PearsonRResult(statistic=0.21107692001003647, pvalue=1.6854893108631245e-41)
```

Conclusions

- ***Job Designation and Salary:*** There is a clear correlation between job designation and salary levels, indicating that certain roles command higher compensation packages than others.
- ***Geographical Location and Salary Disparities:*** The geographical location of employment significantly influences salary differentials, with certain cities offering higher compensation compared to others.
- ***Gender and Earning Potential:*** Gender does not seem to play a significant role in determining earning potential, suggesting a relatively equitable distribution of salaries across genders.
- ***Educational Qualifications and Salary Impact:*** Educational qualifications, particularly the type of degree obtained, have a substantial impact on salary levels. For instance, MCA graduates tend to earn less than those with B.Tech/B.E. or M.Tech/M.E. degrees.
- ***Specializations and Salary Trends:*** Salary trends vary among different specializations, with candidates in computer science & engineering, information technology, and electronics engineering generally commanding higher salaries compared to those in civil engineering or computer application fields.

Conclusions

- ***College GPA and Its Marginal Impact:*** While College GPA may have some correlation with salary levels, its impact appears to be marginal compared to other factors such as educational qualifications and job designation.
- ***Job Designations and City Distribution:*** Various job designations are spread across different cities, each with its unique salary structures, highlighting the importance of geographical location in salary negotiations.
- ***Gender-specific Characteristics of Job Designations and Specializations:*** Certain job designations and specializations exhibit gender-specific characteristics, indicating potential disparities in occupational distribution across genders.

Overall, these observations suggest that while factors such as job designation, geographical location, and educational qualifications significantly influence salary levels, gender does not seem to be a determining factor in earning potential. Additionally, disparities in salary levels exist across different specializations, highlighting the importance of considering multiple factors in understanding salary trends and making informed career decisions.

THANK YOU

INNOMATICS RESEARCH LABS