INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EXPLORATORY DATA ANALYSIS ON AMCAT

About me

- **MANEKUNTA RAMESH KUMAR**
- **B.TECH**
- I like to solve problems using static methods, and I am also interested in programming languages. I want to learn data science, as it offers high-demand career opportunities.
- **linkedin :** www.linkedin.com/in/manekunta-ramesh-kumar
- **github :** https://github.com/Manekuntaramesh

Business Problem and Use case domain understanding:

Perform **Exploratory Data Analysis (EDA)** on the data-set given below. Consider **Salary** as a target variable.

Objects:

- Each row in the dataset represents an individual employee or job record.

Attributes (Columns):

- Unnamed: 0: Appears to be an index or placeholder, possibly indicating the source of the data (e.g., "train").
- ID: A unique identifier for each employee or record.
- Salary: The employee's salary.
- DOJ: Date of Joining.
- DOL: Date of Leaving (or "present" if still employed).
- Designation: The job title of the employee.
- JobCity: The city where the job is located.
- Gender: Gender of the employee (m/f).
- DOB: Date of Birth.
- 10percentage: Likely a score or metric, possibly related to performance.

- ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg: Attributes that might indicate qualifications or degrees in various engineering fields.
- Conscientiousness, Agreeableness, Extraversion, Neuroticism, Openness_to_Experience: Traits likely derived from personality assessments, indicating various psychological attributes.

- **Exploratory Data Analysis:**

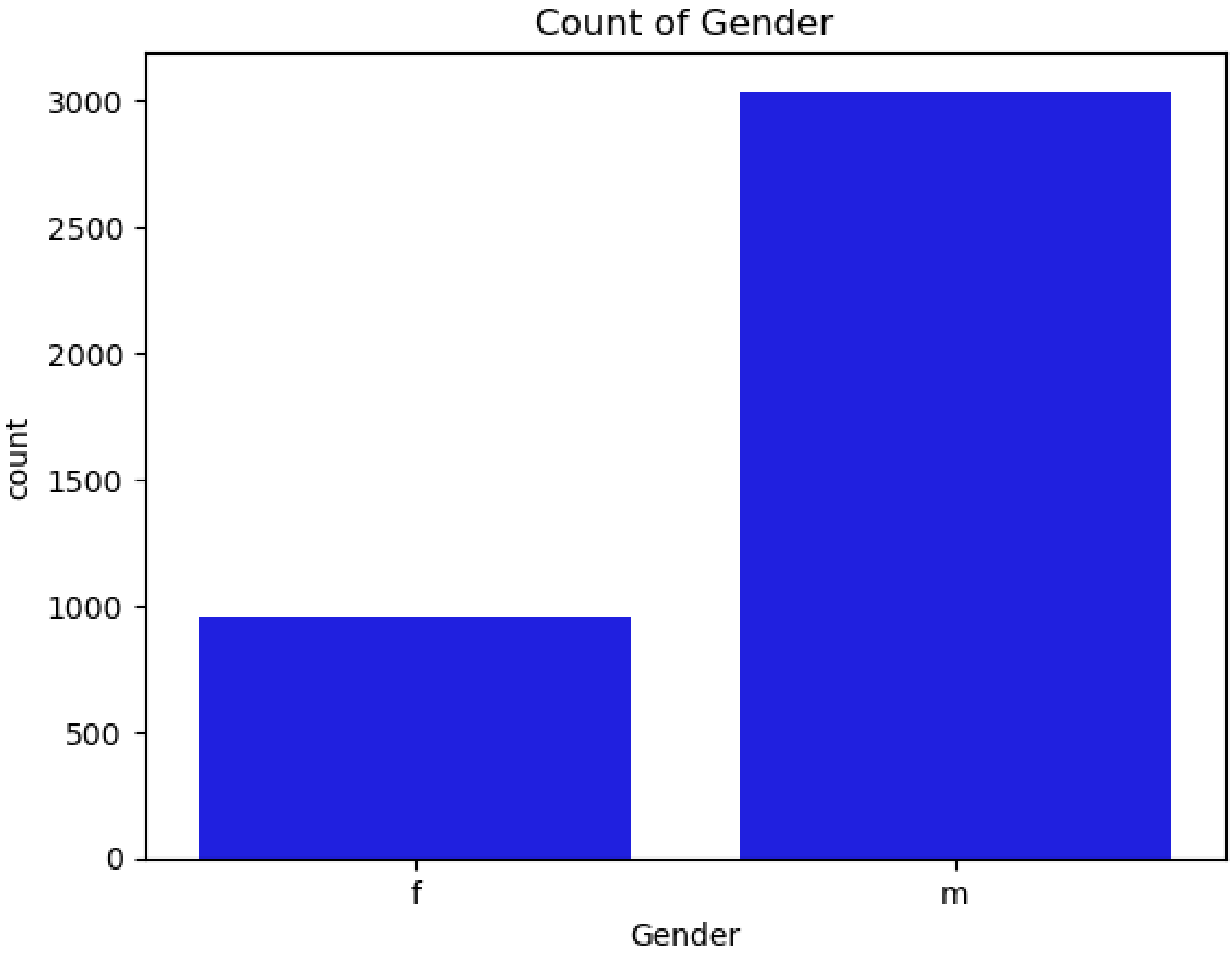
•**1. Data Cleaning Steps**

•Identify Missing Values: Check for NaNs using `df.isnull().sum()`.
•Handle Missing Values: Remove or impute (mean/median/mode).
•Remove Duplicates: Use `df.drop_duplicates()`.
•Correct Data Types: Ensure columns have appropriate types (e.g., `astype()`).
•Filter Outliers: Identify using Z-score or IQR methods.
•Standardize/Normalize Data: Scale features for consistency.

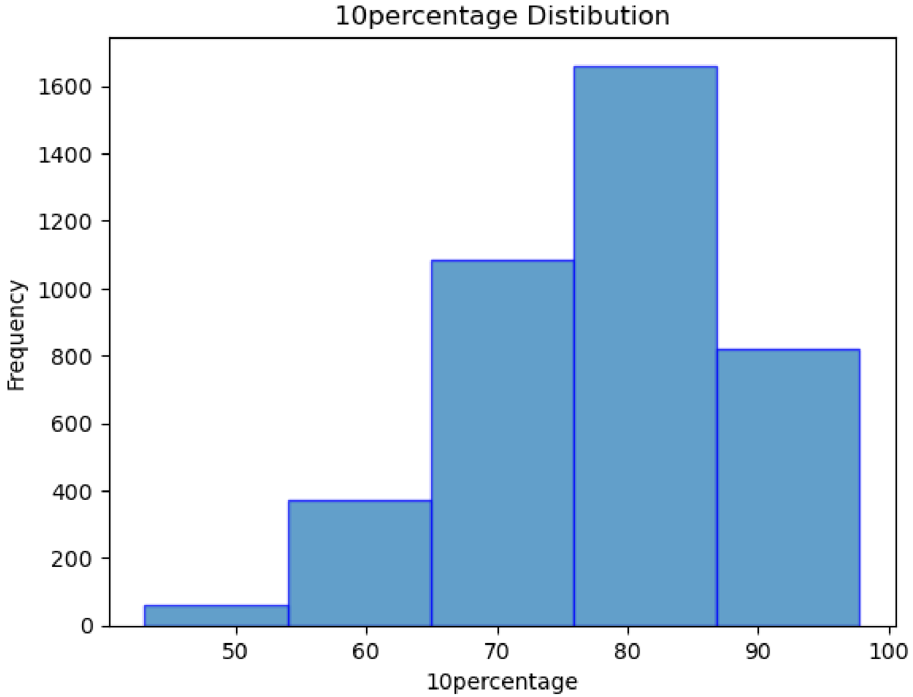
•**2. Data Manipulation Steps**
•Select Relevant Columns: Focus on key features.
•Filter Rows: Subset data based on conditions (e.g., `df[df['column'] > value]`).
•Create New Features: Derive variables (e.g., age from birthdate).
•Group By Operations: Summarize data using aggregation.
•Merging/Joining Datasets: Combine datasets with `merge()`.

- **3. Univariate Analysis Steps:**
  - Descriptive Statistics: Calculate mean, median, mode, etc. (df.describe()).
  - Visualizations: Use histograms, bar charts, and box plots.
  - Distribution Analysis: Assess distribution shape (normality, skewness).

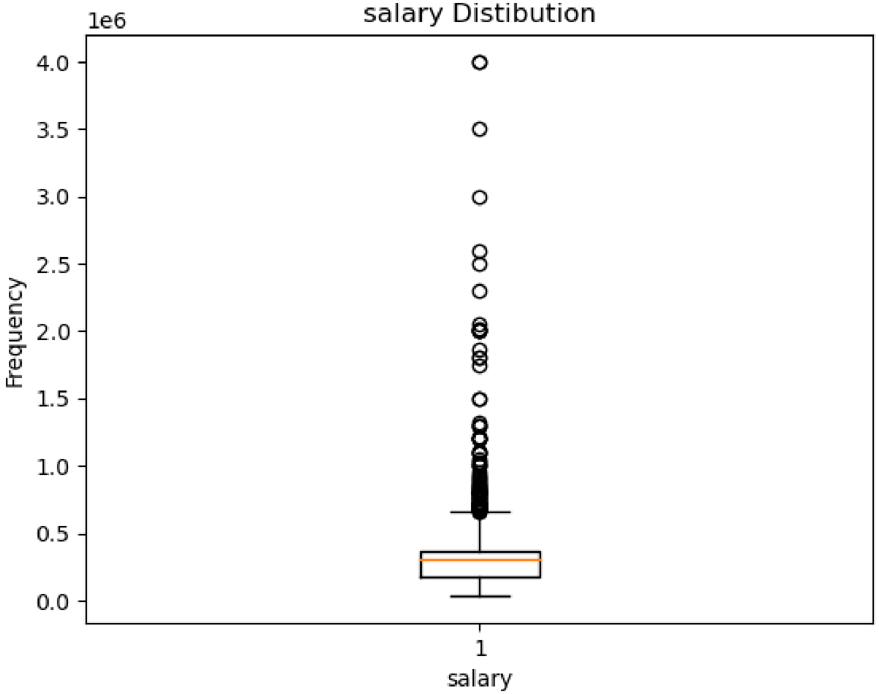
- **Count Plots :**
- A count plot is a type of visualization used to display the counts of observations in categorical data. It shows the frequency of each category in a dataset, making it easy to compare different groups.
- **Axes:**
- **X-axis: "Gender"**
- **Y-axis: "Count"**



Histogram Plot:



Box Plot:



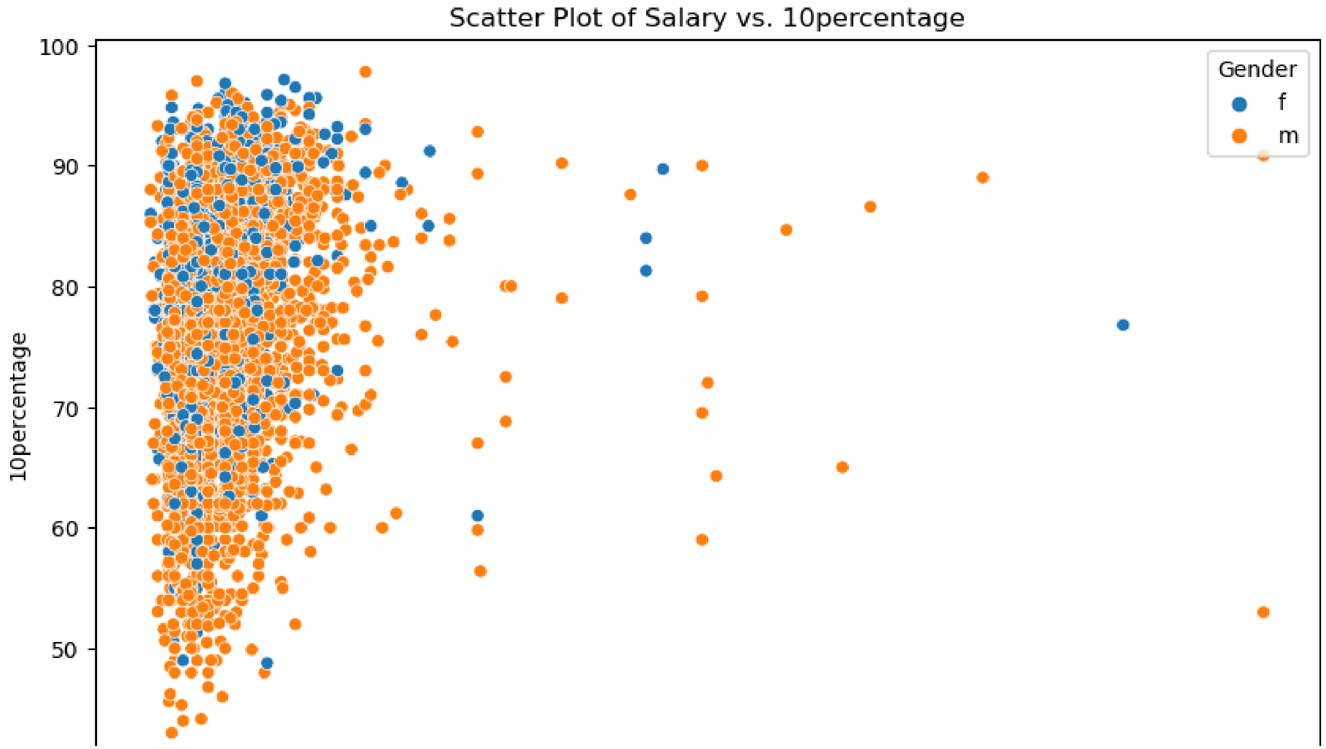
- **Histogram Plot:**
- A histogram plot (or hist plot) is a graphical representation used to visualize the distribution of a continuous variable. It divides the data into bins (intervals) and counts the number of observations that fall within each bin, allowing you to see the shape and spread of the data.
- **Axes:**
- **X-axis:** Represents the bins (intervals) of the variable.
- **Y-axis:** Represents the frequency (count) of observations within each bin.

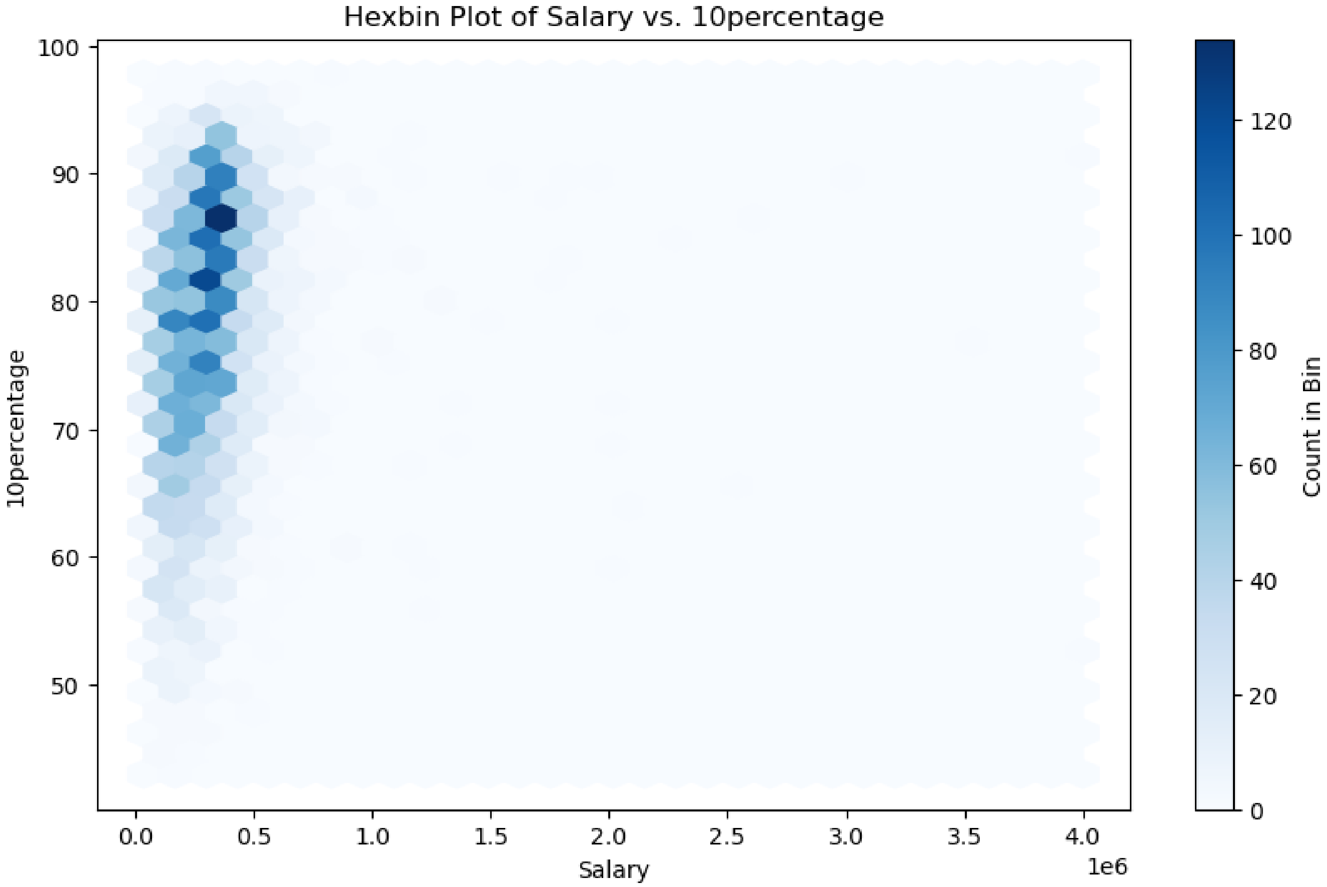
- **Box plot:**
- A boxplot (or box-and-whisker plot) is a graphical representation that summarizes the distribution of a dataset based on five summary statistics: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. It provides a visual way to identify the center, spread, and potential outliers in the data
- **Axes:**
- **X-axis:** Represents the bins (intervals) of the variable.
- **Y-axis:** Represents the frequency (count) of observations within each bin.

- ## 4. Bivariate Analysis Steps
- Correlation Analysis: Calculate correlation coefficients and visualize with heatmaps.
- Scatter Plots: Show relationships between two numerical variables.
- Grouped Box Plots: Visualize distributions across categories.
- Chi-Square Tests: Test relationships between categorical variables.
- Regression Analysis: Fit models to quantify relationships

- ## Scatter plot:

- A scatter plot is a type of data visualization that uses dots to represent the values obtained for two different variables—one plotted along the x-axis and the other along the y-axis. This type of plot is useful for identifying relationships, trends, and patterns between the two variables.
- Key Features of Scatter Plots
- **Axes:**
- **X-axis:** Salary
- **Y-axis:** 10percentage





- **<u>Herbin Plot:</u>**

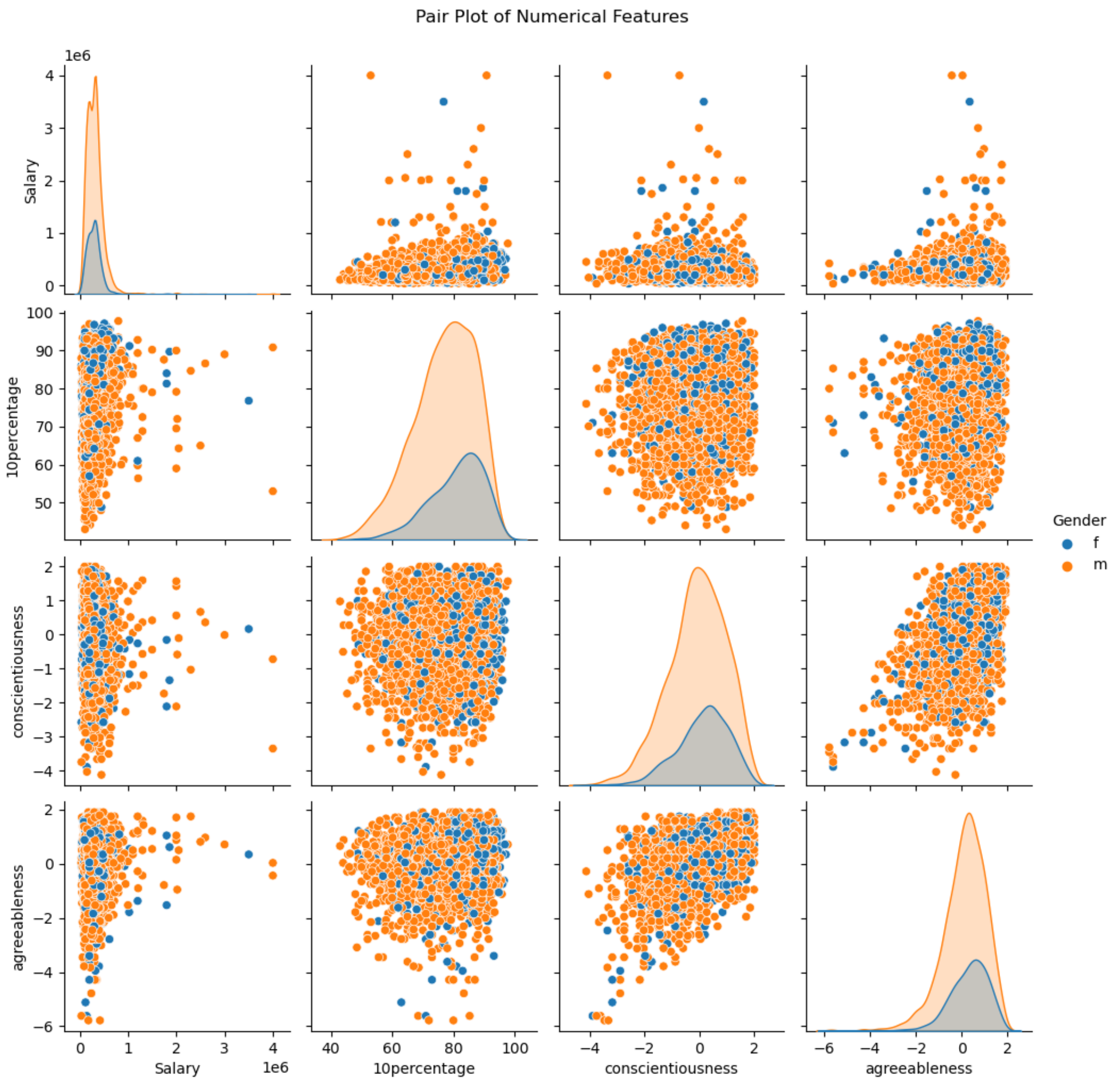
A herbin plot, commonly known as a hexbin plot, is a type of data visualization used to display the relationship between two continuous variables. It is particularly useful when dealing with large datasets, as it summarizes the number of points that fall into hexagonal bins, allowing for better visualization of density and clustering.

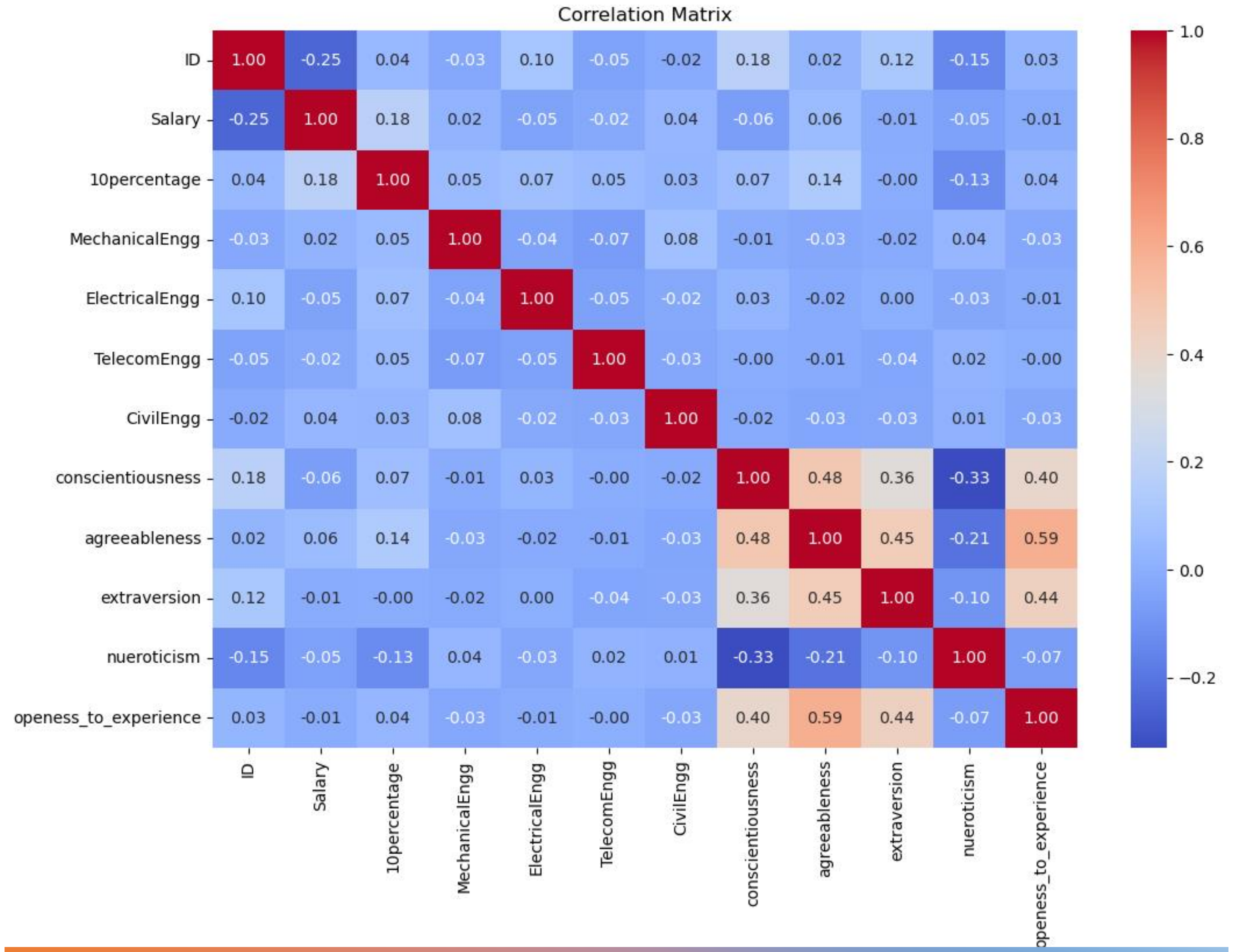
- **Axes:**
- **X-axis** : Salary
- **Y-axis:** 10percentage

Pair Plot:

A pair plot is a type of data visualization that displays pairwise relationships in a dataset. It creates a matrix of scatter plots (or other plots) for each pair of features, allowing you to quickly visualize correlations and distributions between multiple variables. Pair plots are particularly useful in exploratory data analysis (EDA) when working with datasets containing multiple continuous variables.

- **Matrix of Plots:** Each cell in the matrix corresponds to a scatter plot (or other visualizations) for a pair of features.
- **Diagonal Histograms or Density Plots:** The diagonal typically shows the distribution of each variable, using histograms or kernel density estimates (KDE).
- **Color Coding:** Points can be color-coded based on categorical variables to identify groups within the data.





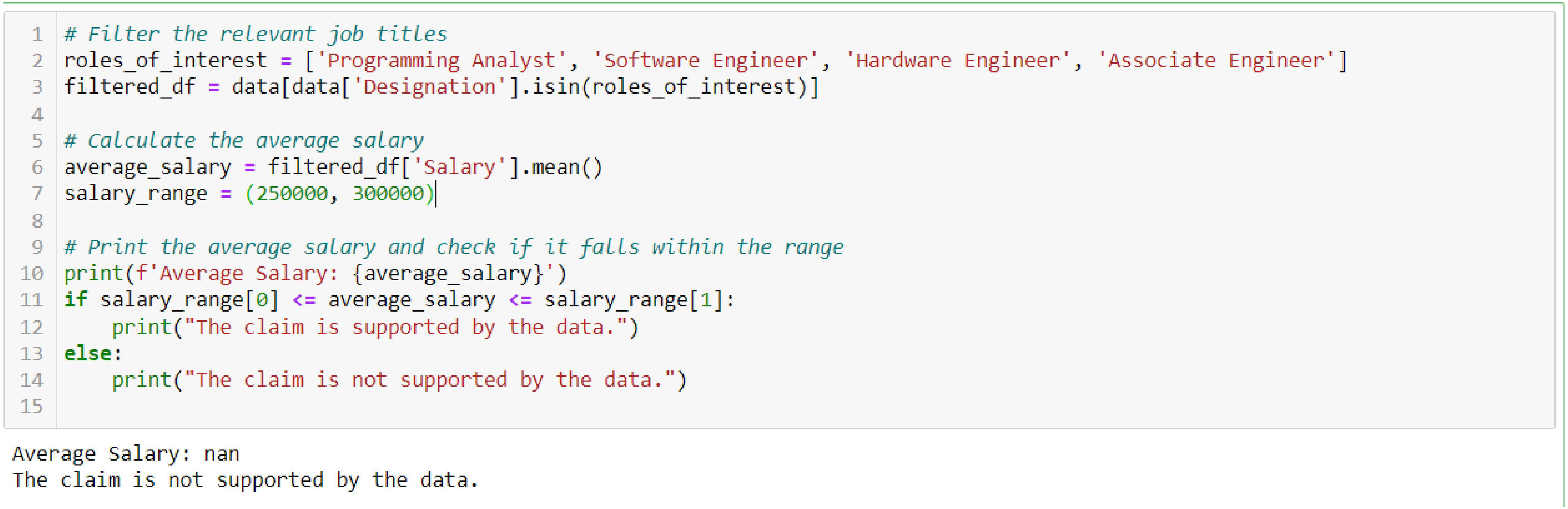
Correlation Matrix

	ID	Salary	10percentage	MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg	conscientiousness	agreeableness	extraversion	nueroticism	openess_to_experience
ID	1.00	-0.25	0.04	-0.03	0.10	-0.05	-0.02	0.18	0.02	0.12	-0.15	0.03
Salary	-0.25	1.00	0.18	0.02	-0.05	-0.02	0.04	-0.06	0.06	-0.01	-0.05	-0.01
10percentage	0.04	0.18	1.00	0.05	0.07	0.05	0.03	0.07	0.14	-0.00	-0.13	0.04
MechanicalEngg	-0.03	0.02	0.05	1.00	-0.04	-0.07	0.08	-0.01	-0.03	-0.02	0.04	-0.03
ElectricalEngg	0.10	-0.05	0.07	-0.04	1.00	-0.05	-0.02	0.03	-0.02	0.00	-0.03	-0.01
TelecomEngg	-0.05	-0.02	0.05	-0.07	-0.05	1.00	-0.03	-0.00	-0.01	-0.04	0.02	-0.00
CivilEngg	-0.02	0.04	0.03	0.08	-0.02	-0.03	1.00	-0.02	-0.03	-0.03	0.01	-0.03
conscientiousness	0.18	-0.06	0.07	-0.01	0.03	-0.00	-0.02	1.00	0.48	0.36	-0.33	0.40
agreeableness	0.02	0.06	0.14	-0.03	-0.02	-0.01	-0.03	0.48	1.00	0.45	-0.21	0.59
extraversion	0.12	-0.01	-0.00	-0.02	0.00	-0.04	-0.03	0.36	0.45	1.00	-0.10	0.44
nueroticism	-0.15	-0.05	-0.13	0.04	-0.03	0.02	0.01	-0.33	-0.21	-0.10	1.00	-0.07
openess_to_experience	0.03	-0.01	0.04	-0.03	-0.01	-0.00	-0.03	0.40	0.59	0.44	-0.07	1.00

- **Correlation Matrix :**

- A **correlation matrix** is a table that displays the correlation coefficients between multiple variables. It provides a quick way to assess the strength and direction of relationships among variables in a dataset. Each cell in the matrix shows the correlation between two variables, with values ranging from -1 to 1.
- Key Features of a Correlation Matrix
- **Values:**
- **+1:** Perfect positive correlation (as one variable increases, the other also increases).
- **-1:** Perfect negative correlation (as one variable increases, the other decreases).
- **0:** No correlation (no linear relationship between the variables).

- Research Question :
- 1.Testing the Salary Claim for Computer Science Graduates



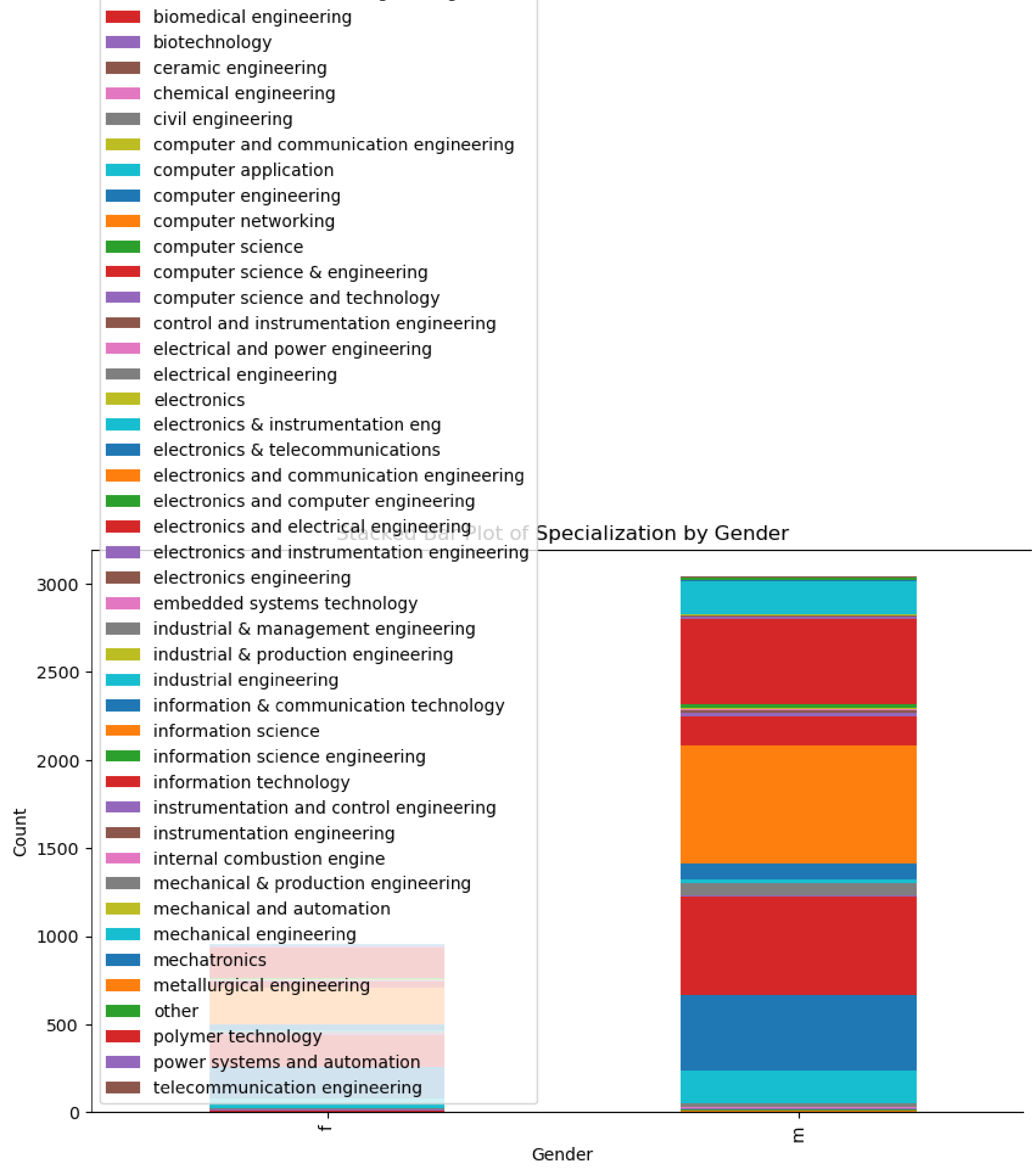
```python
# Filter the relevant job titles
roles_of_interest = ['Programming Analyst', 'Software Engineer', 'Hardware Engineer', 'Associate Engineer']
filtered_df = data[data['Designation'].isin(roles_of_interest)]

# Calculate the average salary
average_salary = filtered_df['Salary'].mean()
salary_range = (250000, 300000)

# Print the average salary and check if it falls within the range
print(f'Average Salary: {average_salary}')
if salary_range[0] <= average_salary <= salary_range[1]:
    print("The claim is supported by the data.")
else:
    print("The claim is not supported by the data.")

```

```
Average Salary: nan
The claim is not supported by the data.
```



2. Analyzing the Relationship Between Gender and Specializatio:

Steps to Analyze:

- **Cross-tabulation:**
- Create a cross-tabulation of gender and specialization to see the distribution.

Statistical Test:

- Use a Chi-square test to determine if there is a significant association between gender and specialization.

Chi – square:

The Chi-square test is a statistical method used to determine whether there is a significant association between two categorical variables. In the context of your research question regarding the relationship between gender and specialization, the Chi-square test can help determine if the distribution of specializations is independent of gender

- **Axes:**
- **X-axis** : Gender
- **Y-axis:** Count

.

INNOMATICS RESEARCH LABS

Conclusion:

- **For the Salary Claim:**

- Analyze the average salary for the specified roles and check if it lies within the claimed range. The results will either support or refute the claim made by the Times of India article.
- If the claim is supported, it can be beneficial for educational institutions and career advisors to highlight these expected salary ranges when counseling prospective students in the Computer Science field.
- Conversely, if the claim is refuted (for instance, if the average salary is significantly lower than 2.5 lakhs), this could prompt discussions about market conditions, the need for curriculum updates, or industry alignment with educational outcomes.

- **For Gender and Specialization:**

- The results from the Chi-square test will indicate whether there is a significant relationship between gender and specialization preferences.
- The findings may also reflect broader societal trends regarding gender roles and expectations in education and careers. Insights gained can help inform discussions about gender equity in education and the workforce.

Overall Summary:

The analyses derived from the dataset can offer valuable insights not just for academic or organizational strategies but also for societal understanding of labor market trends. By testing the salary claims and exploring the relationship between gender and specialization, we can better understand the dynamics at play in the field of Computer Science Engineering and related job markets.

THANK YOU