INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO Dataset

Thabasum Shaikh
22nd Feb 2024

About me

Greetings! I'm a proactive, responsible, and results-oriented professional currently pursuing a bachelor's degree in computer engineering. My passion lies in tackling technical challenges, conducting research, and pioneering new technologies. Thriving in collaborative environments, I relish connecting with diverse individuals. With my friendly demeanor and aptitude for quick learning, I excel in high-pressure situations and possess strong stress management skills.

My journey into Data Science is fueled by its transformative potential across industries. I'm captivated by the ability to glean invaluable insights from vast data sets, driving informed decision-making and fostering innovation. Data Science provides a robust toolkit for uncovering patterns, trends, and correlations that have profound impacts on businesses and society.

While my professional experience in Data Science is nascent, I'm enthusiastic about immersing myself in this field and applying my knowledge in real-world scenarios. I'm actively seeking entry-level roles where I can further develop my skills while making tangible contributions. If you're interested in delving deeper into my background and interests, I invite you to explore my LinkedIn profile here and my GitHub repository here. These platforms showcase my educational journey, projects, and more.

PROJECT OBJECTIVE

The primary objective of this project is to conduct an in-depth analysis of the provided dataset, with a specific focus on understanding the relationship between different features and the target variable, which is Salary. The main goals of this analysis are as follows:

1. **Pattern and Trend Identification**
2. **Relationship Exploration**
3. **Outlier Detection**

Summary of Data

1. Rows: 3998
2. Columns: 38
3. Data Types: Numeric (27 columns)
4. Categorical (9 columns)
5. Date time (2 columns)
6. Target Variable: Salary (numeric)
7. Features: ID: Unique identifier for each record
8. DOJ: Date of joining (date time)
9. DOL: Date of leaving (object)
10. Designation: Job title or position (object)
11. Job City: City where the job is located (object)
12. Gender: Gender of the employee (object)
13. DOB: Date of birth (date time)
14. 10percentage: Percentage obtained in 10th grade (float)
15. 10board: Board of education for 10th grade (object)
16. 12graduation: Year of graduation from 12th grade (int)
17. 12percentage: Percentage obtained in 12th grade (float)
18. 12board: Board of education for 12th grade (object)
19. College ID: Unique identifier for college (int)
20. College Tier: Tier of college (int)
21. Degree: Degree obtained (object)
22. Specialization: Field of specialization (object)
23. collegeGPA: Grade Point Average in college (float)
24. CollegeCityID: Unique identifier for college city (int)
25. CollegeCityTier: Tier of college city (int)
26. College State: State where the college is located (object)
27. Graduation Year: Year of graduation (int) English, Logical,
28. Quant: Scores in respective subjects (int)
29. Domain: Domain knowledge score (float) ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg: Scores in respective subjects (int) conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience: Personality traits scores (float) Memory Usage: 1.2+ MB

Data Cleaning and Processing Observations:

1. Timestamp format for DOJ, DOL, and DOB columns.
2. Job city column contains "-1" values, indicating missing values.
3. "10 board" and "12 board" columns contain "0" values, indicating missing values.
4. "College state" column contains "union territory," not a specific state.
5. "Graduation year" column contains "0," indicating missing values.
6. "Domain" column contains "-1," indicating missing values.

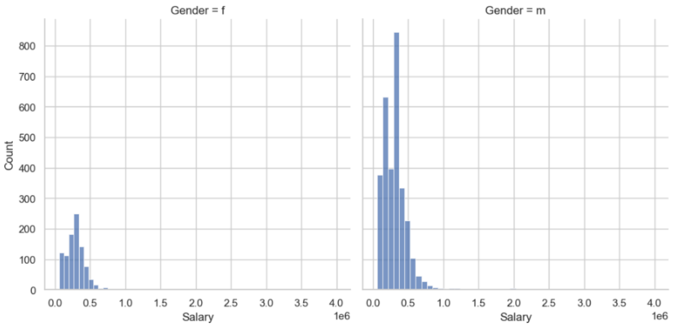
Actions Taken:

- Converted timestamp format to date using the datetime module for DOJ and DOL columns.
- Replaced "present" value in the DOL column with the present date.
- Dropped 12th and 10th graduation timestamp columns and created new columns:
    a) 12Gradage: Indicates the age of a person during 12th
    b) Graduation Gradage: Indicates the age of a person during their higher education graduation.
- Imputed missing values in the "Graduation year" column with the mode.
- Replaced "-1" values in the "Job city" column with NaN equivalents.
- Replaced "0" values in the "10 board" and "12 board" columns with appropriate missing value indicators.
- Replaced "union territory" in the "college state" column with the correct state name.
- Replaced "-1" values in the "Domain" column with appropriate missing value indicators.

UNI-VARIATE ANALYSIS

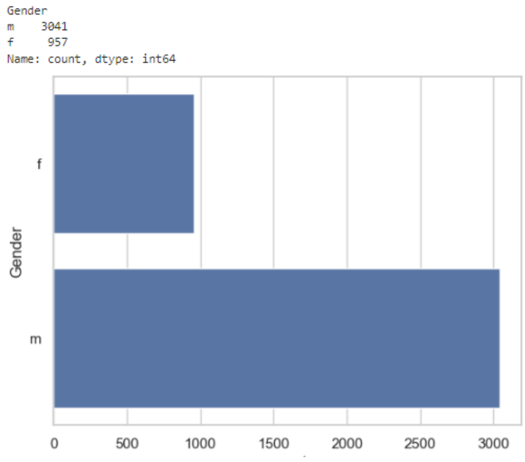
Salary Distribution:

- The salary data is right-skewed, indicating that there are outliers with very high salaries.
- The median salary for both genders is similar.
- Males have more outliers in their salary distribution compared to females.



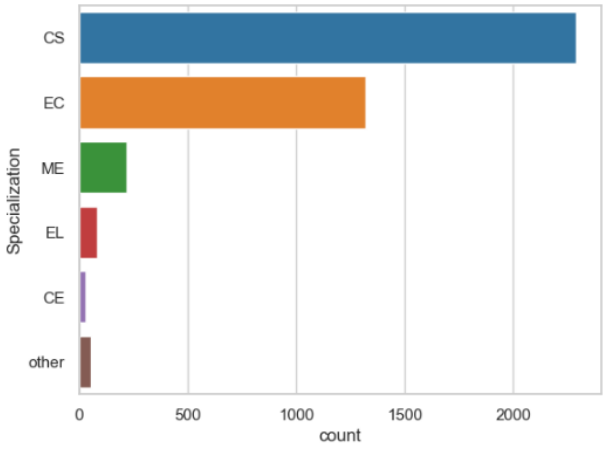
Gender Distribution:

- The ratio of males to females is approximately 3:1, indicating that there are significantly more men employed in the dataset.


```
Gender
m    3041
f     957
Name: count, dtype: int64
```

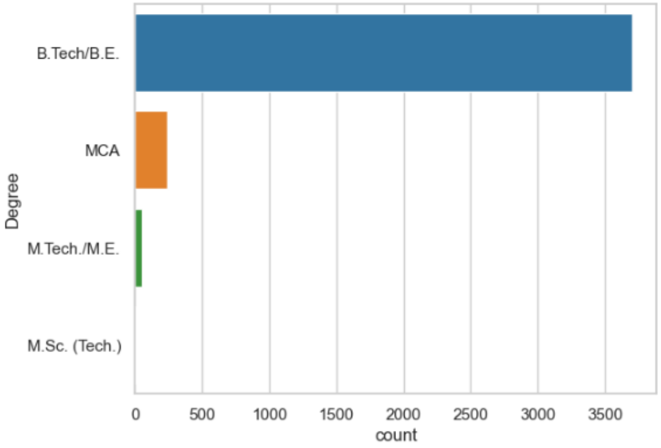
Salary Distribution by Specialization:

- The median salary for people from all specializations is nearly similar.
- People specializing in Computer Science (CS) and Electronics and Communication (EC) tend to have higher salaries compared to other specializations.



Degree Distribution of Amcat Aspirant

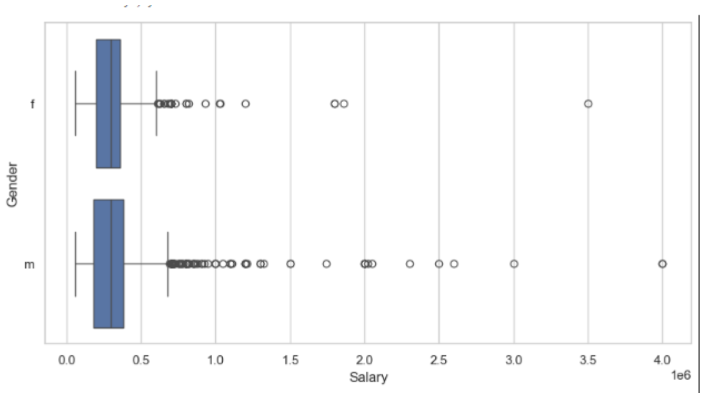
- The majority of Amcat aspirants hold a Bachelor's degree (Btech).



BIVARIATE ANALYSIS

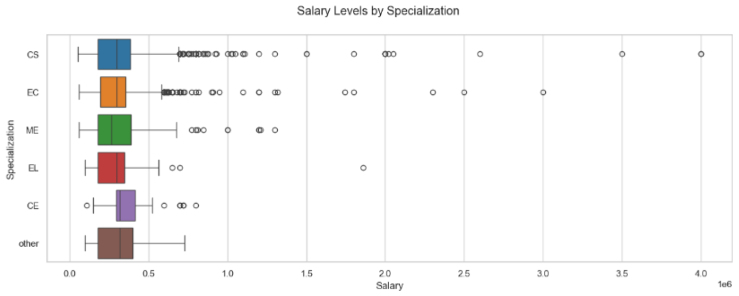
Salary Distribution by Gender

- The median salary for both genders is similar, but males have more outliers, indicating a higher number of males earning higher pay.
- There is not a significant difference in the median salary between males and females.



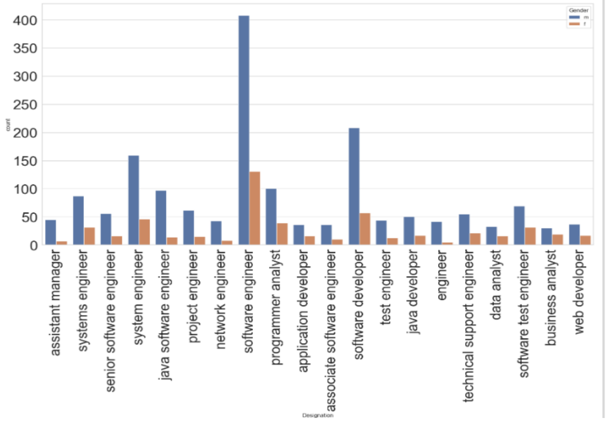
Salary Distribution by Specialization:

- People specializing in Computer Science (CS) and Electronics and Communication (EC) tend to have higher salaries compared to other specializations.
- The median salary for all specializations is nearly similar, but some specializations have higher numbers of individuals with higher salaries.



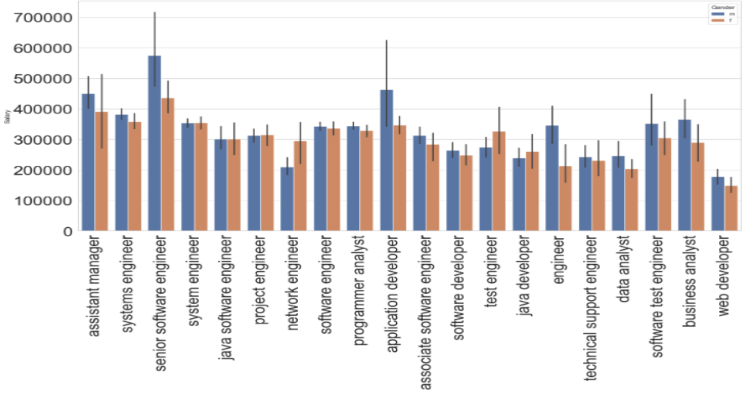
Gender Distribution in Top Designations:

- All general professions are more dominated by males, indicating a gender disparity in various job roles.
- The frequency of different job roles varies, with some roles being more common among males than females.



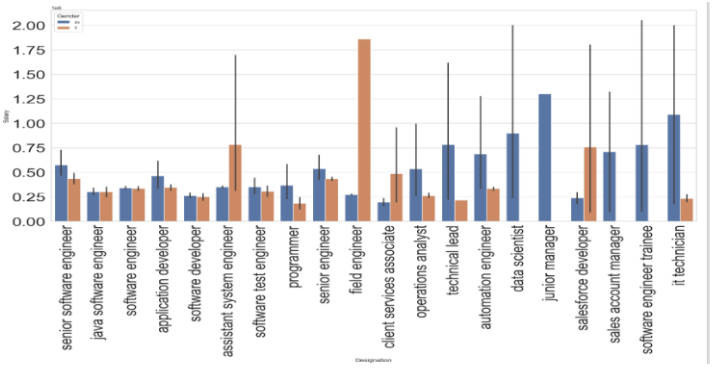
Relationship between Gender and Specialization:

- Men from CS, EC, and CE specializations earn slightly more than women from the same specializations.
- Women from the Electrical (EL) specialization earn more than men from the same specialization, suggesting a gender-based disparity in salary within specific specializations.



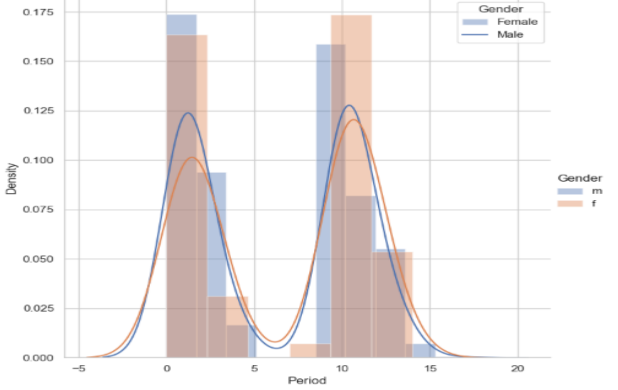
Salary Distribution by Gender and Designation

- Most of the high-paying jobs are in the IT domain.
- In 45% of top-paying roles, men are generally paid higher than women.
- In 20% of top-paying roles, women are paid higher than men
- In roles like junior manager, sales account manager, and software engineer trainee no women are working in these fields.
- Junior manager is highest paying for men and field engineer is the highest paying role for women.
- The discrepancy between pay based on gender might be because of other features like experience, specialization, etc.
- Software Engineer and Software developer are the most frequent and highest paying jobs



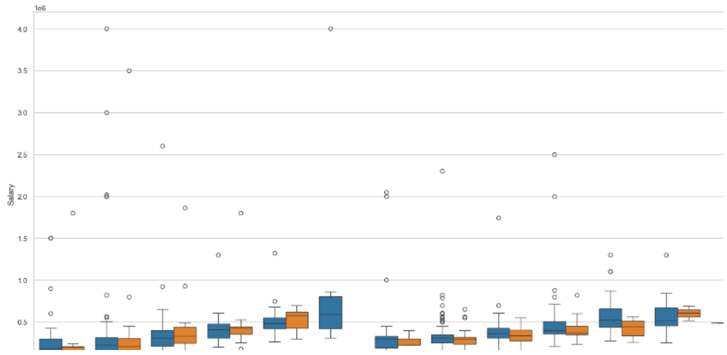
Distribution of Work Experience (Period) by Gender:

- The distribution plot shows a bi-modal distribution for both genders, indicating two distinct clusters of work experience.
- Females have an average work experience of approximately 6.63 years, while males have slightly less at 6.24 years for the entire dataset.
- In high-paying jobs, the average work experience decreases slightly for both genders, with females at 5.81 years and males at 5.74 years.
- Males generally exhibit higher average work experience compared to females, with males averaging around 5 years and females around 4.5 years.



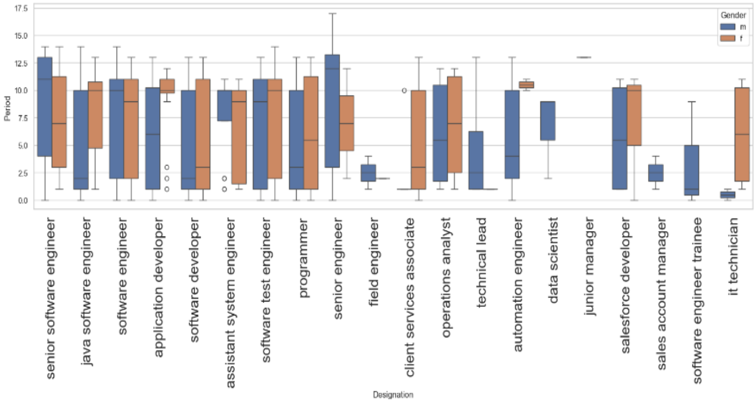
Relationship Between Work Experience and Salary by Gender:

- The median salary for both males and females shows a slight increase with work experience for the first five years.
- However, there is a sudden decrease in median salary in the sixth year, followed by a similar pattern in the subsequent years.
- Men and women with the same level of experience are paid nearly equally, with salaries ranging from 3.5 to 5 lakhs.



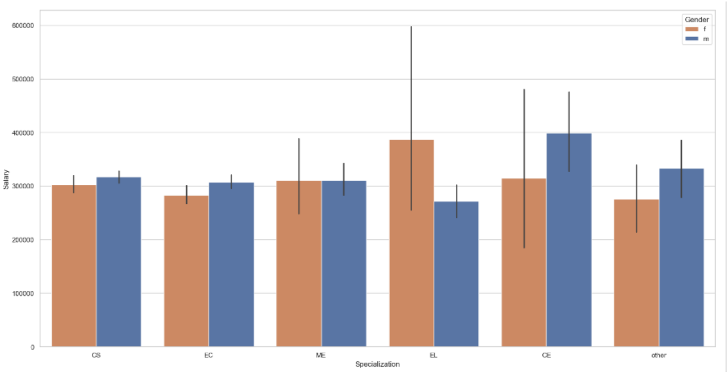
Experience Distribution by Designation and Gender

- The period distribution across different designations is skewed, indicating varying experience levels.
- Median experience differs between genders across all designations.
- However, experience alone doesn't explain salary disparities; some women with higher experience are paid less, and vice versa for men.
- Experience doesn't strongly correlate with salary. The designation with the highest experience is Senior Engineer.



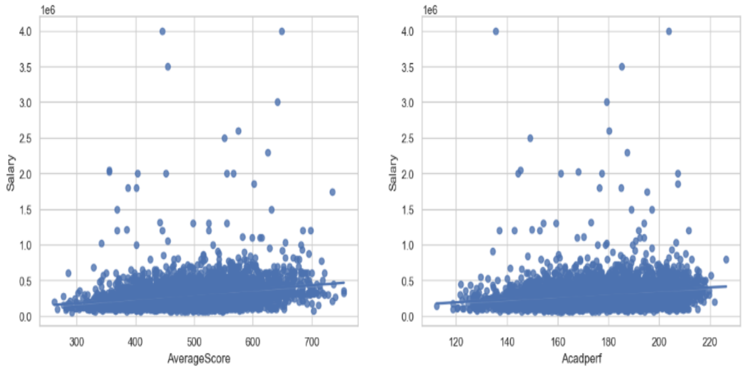
Salary Variation Across Specializations by Gender

- Men from CS, EC, and CE specializations earn slightly more than women from the same specializations.
- Women from the EL (Electrical Engineering) specialization earn significantly more than men from the same specialization.



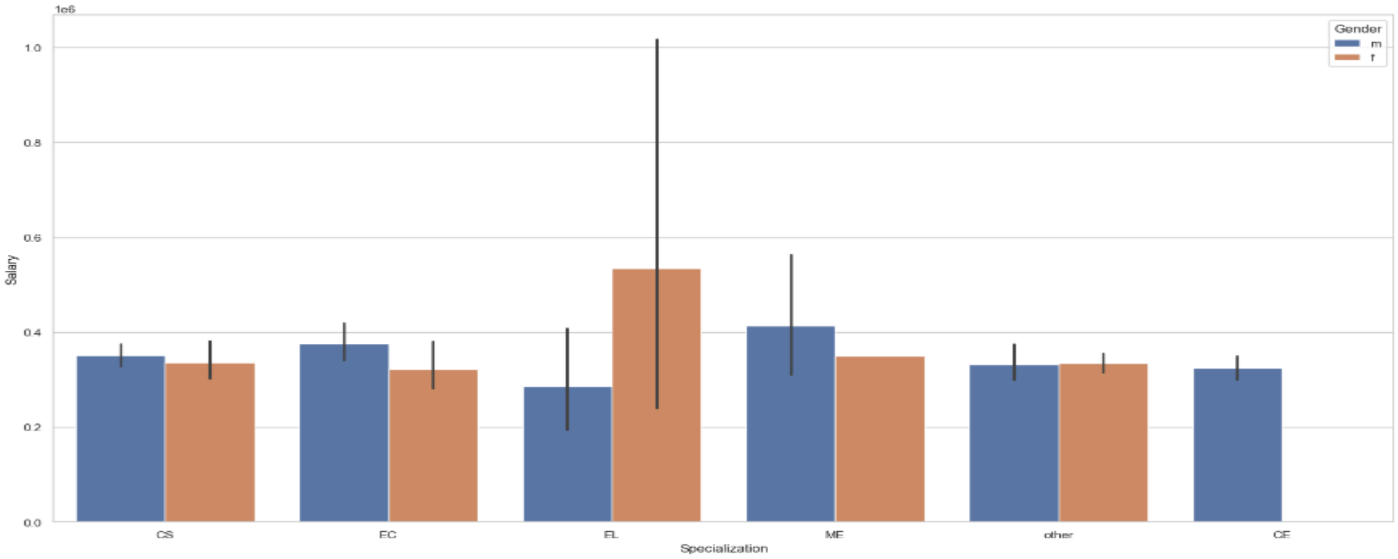
Relationship Between Academic Performance and Salary

- There is a positive correlation between the average score (combining logical, quantitative, and English test scores) and salary.
- Similarly, there is a positive correlation between academic performance (combining 10th, 12th, and college GPA) and salary.



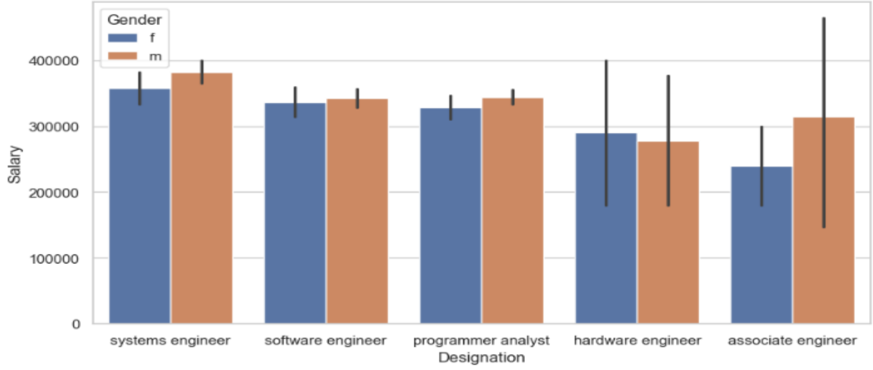
Gender-based Salary Distribution in High-Paying Jobs Across Specializations

- From the CE (Computer Engineering) specialization, only men are occupying higher-paying jobs.
- Specialization is not the primary reason for women being paid less, as the majority of people are from the CS (Computer Science) field, where men and women earn similar salaries in high-paying roles.

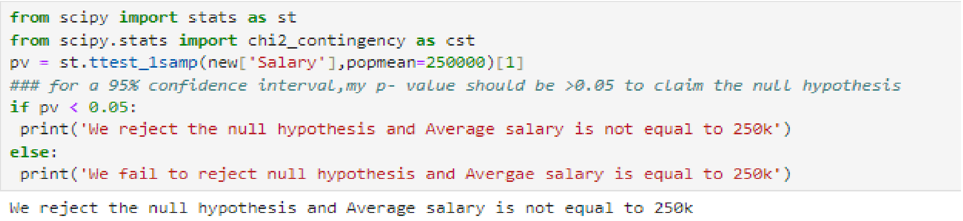


Times of India article dated Jan 18, 2019 states that "*After doing your Computer Science Engineering if you take up jobs as a Programming Analyst, Software Engineer, Hardware Engineer and Associate Engineer you can earn up to 2.5-3 lakhs as a fresh graduate.*"

- The salary for freshers, particularly in roles such as programmer analyst, software engineer, hardware engineer, associate engineer, and systems engineer, typically starts from 200k. Additionally, male individuals tend to earn more than female individuals in these entry-level positions.



- The 1-sample T-test was conducted to verify the claim that the average salary is not equal to 250k.
- The null hypothesis (H0: µ = 250k) was tested against the alternative hypothesis (H1: µ ≠ 250k).
- The resulting p-value was less than 0.05, indicating statistical significance.
- Therefore, we reject the null hypothesis and conclude that the average salary is not equal to 250k.



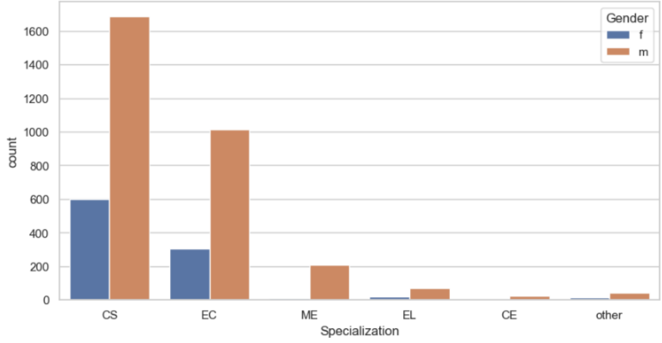
```
from scipy import stats as st
from scipy.stats import chi2_contingency as cst
pv = st.ttest_1samp(new['Salary'],popmean=250000)[1]
### for a 95% confidence interval,my p- value should be >0.05 to claim the null hypothesis
if pv < 0.05:
 print('We reject the null hypothesis and Average salary is not equal to 250k')
else:
 print('We fail to reject null hypothesis and Avergae salary is equal to 250k')
```

```
We reject the null hypothesis and Average salary is not equal to 250k
```

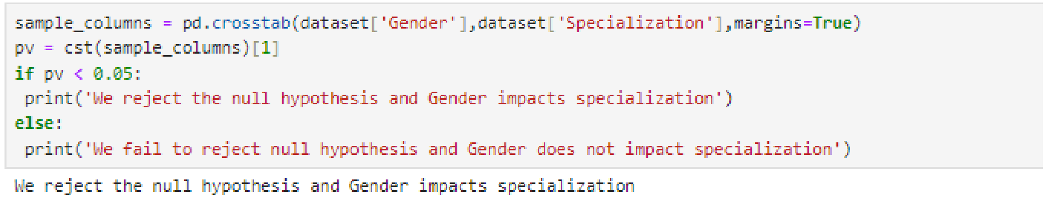
In summary, the analysis and hypothesis testing provide evidence to support the claim that the average salary for fresher's in the specified roles is not equal to 250k.

Is there a relationship between gender and specialization? (i.e. Does the preference of Specialization depend on the Gender?)

- Most of the Amcat aspirants in the dataset are from Computer Science (CS) and Electronics and Communication (EC) specializations.



Chi-Square Test to check the relation between Specialization and Gender:-



```
sample_columns = pd.crosstab(dataset['Gender'],dataset['Specialization'],margins=True)
pv = cst(sample_columns)[1]
if pv < 0.05:
 print('We reject the null hypothesis and Gender impacts specialization')
else:
 print('We fail to reject null hypothesis and Gender does not impact specialization')
```

```
We reject the null hypothesis and Gender impacts specialization
```

Chi-square test conducted, the following conclusions can be drawn:

- The chi-square test was performed to assess the relationship between specialization and gender.
- The null hypothesis (H0: Gender does not impact specialization) was tested against the alternative hypothesis (H1: Gender impacts specialization).
- The resulting p-value was less than 0.05, indicating statistical significance.

Therefore, we reject the null hypothesis and conclude that gender impacts specialization.

THANK YOU!!