INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on Aspiring Mind Employment Outcome 2015 (AMEO) Dataset

About me

Hi there! My name is Shubham Prajapati, and I'm passionate about the intersection of technology and data. Over the past year, I've delved deep into the world of Data Science and Machine Learning, exploring its intricacies and applications.

With a Bachelor's degree in Computer Science under my belt, I've always been drawn to the realm of technology. I'm a tech enthusiast at heart, constantly seeking out the latest developments and innovations in the field. From coding to exploring new software tools, I love immersing myself in all tech-related things.

My journey into Data Science was sparked by my innate curiosity about data. I find joy in uncovering patterns, extracting insights, and deriving meaning from datasets. It's this curiosity that led me to pursue a career in Data Science, where I can leverage my analytical skills to tackle real-world challenges.

During my academic and professional journey, I had the opportunity to work as a Data Analyst intern for a month, where I gained hands-on experience in data manipulation, visualization, and analysis. This experience further fueled my passion for data-driven decision-making and reinforced my desire to excel in the field of Data Science.

As I continue on this exciting path, I eagerly look forward to learning new and cutting-edge technologies in the ever-evolving landscape of tech. Whether it's mastering new algorithms, exploring emerging tools, or diving into advanced topics, I'm always ready to embrace the next challenge and expand my knowledge horizons.

Objective

The objective of this Exploratory Data Analysis (EDA) is to gain comprehensive insights into the dataset through rigorous univariate and bivariate analyses. By examining individual variables (univariate analysis) and exploring relationships between pairs of variables (bivariate analysis), we aim to uncover patterns, trends, and associations within the data. Through this process, we seek to identify key characteristics, understand distributions, detect outliers, and reveal potential correlations or dependencies that may inform subsequent modeling or decision-making processes.

Data Summary

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. The dataset contains around 37 independent variables and 3998 data points. The independent variables are both continuous and categorical in nature. The dataset contains a unique identifier for each candidate. Below mentioned table contains the details for the original dataset.

The dataset contains 20 Continuous Numerical Columns including the dependent variable (Salary). There are 10 Categorical Columns, 3 ID related columns and 5 Date related columns

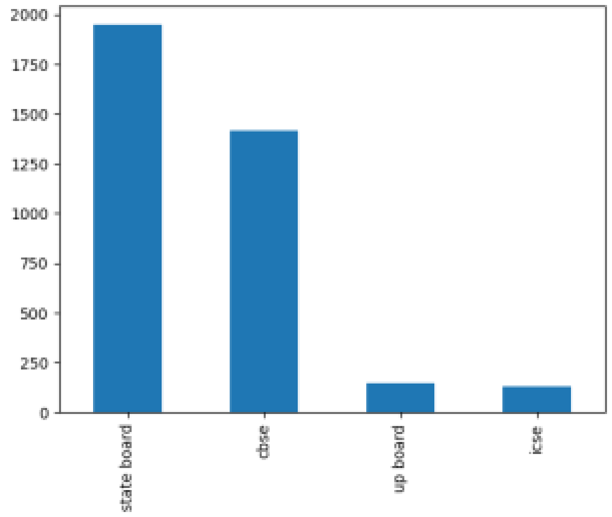
Data Cleaning

- The **10board** and **12board** column are very inconsistent, it contains vast amount of board names, some board names are written incorrectly.
- To clean this, I replaced all the inconsistent names with the correct one and replaced all the other board with state board. I replaced null values with ‘no board mentioned’
- Next is the JobCity column. In this column, I cleaned the top 10 city/state names (excluding the ones not mentioned), as they accounted for 85% of the data. I also filled the null values with 'not mentioned’.
- The **Designation** column exhibited some inconsistencies, such as 'software engineer' being written as 'software engg.' Therefore, I addressed these inconsistencies by removing them, along with any other inconsistencies present in the column.

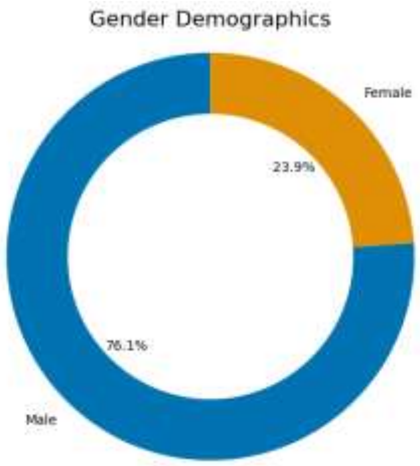
Feature Engineering

- I performed feature engineering to create a new column called **Age** from the **DOB** column. This allows us to explore potential correlations with other columns.

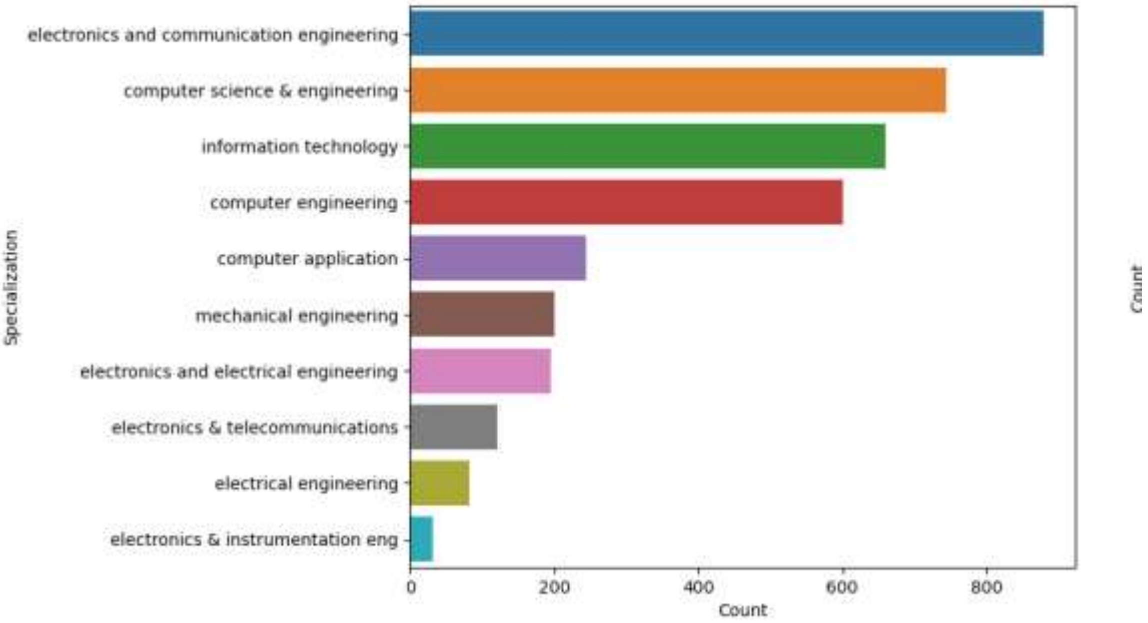
Dissecting Variables: Univariate Examination



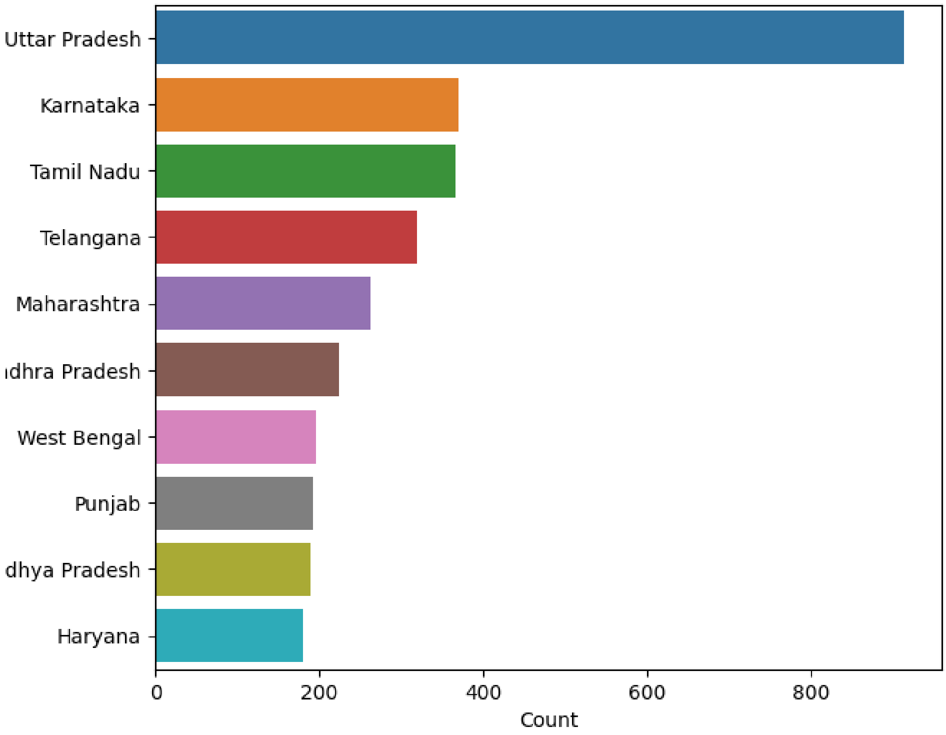
- The bar graph shows that the majority of people are from state board and CBSE board.
- If we exclude the state board, CBSE emerges as the primary contributor.



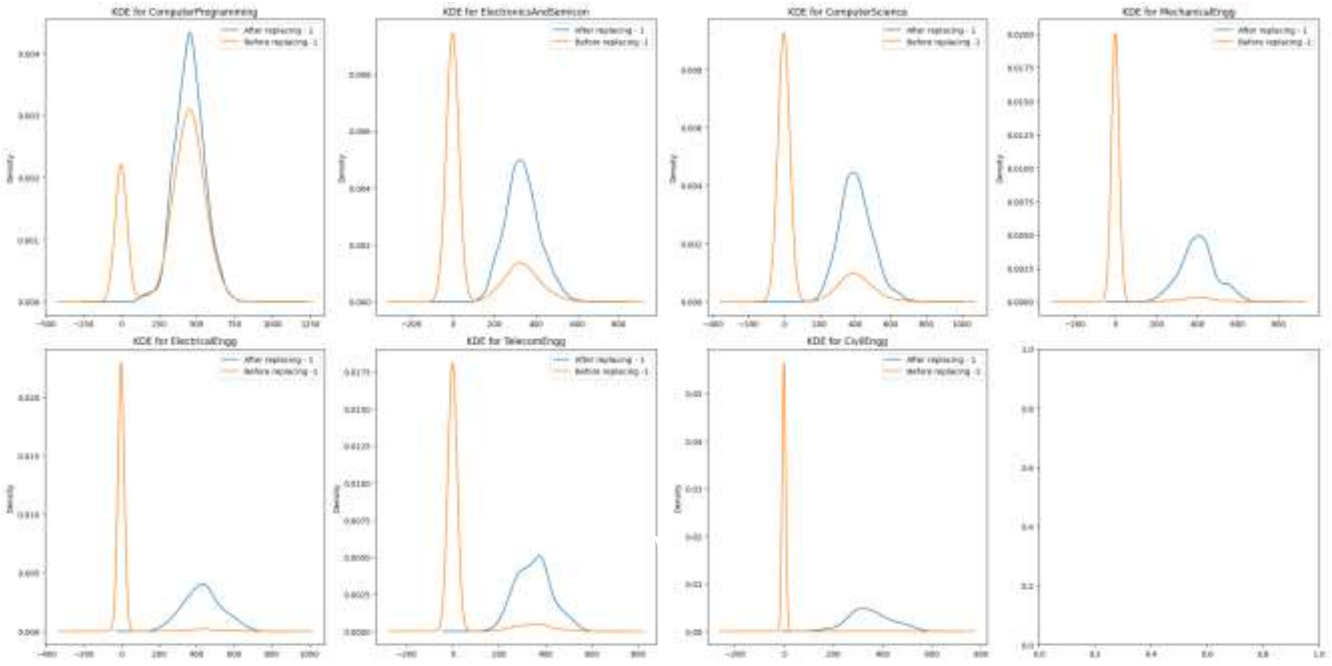
- The Donut Chart shows that the majority of people are Men.



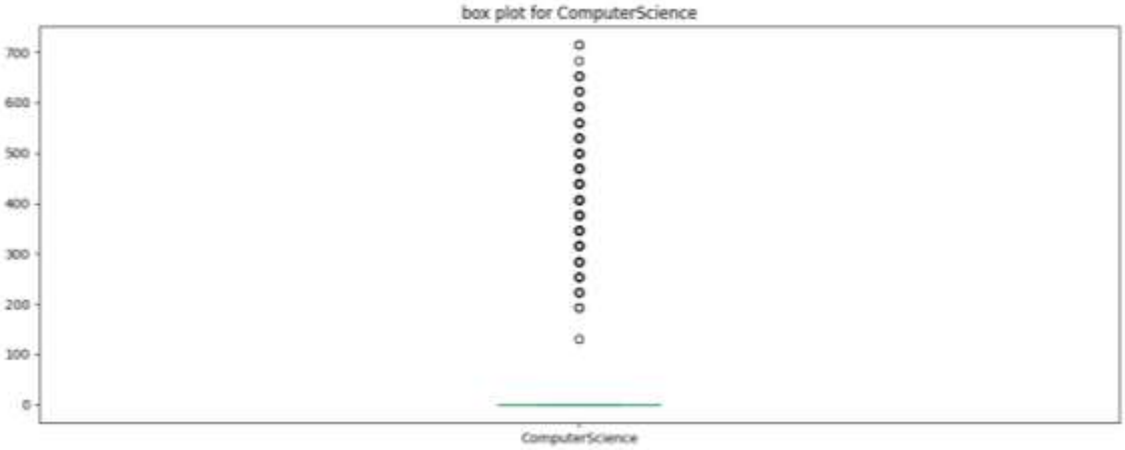
- The data reveals that the majority of individuals have backgrounds in electronics and computer sciences.
- The electronics market experienced a significant boom around 2015 following the discovery of blue LED technology, which catalyzed the transformation of screen technologies.



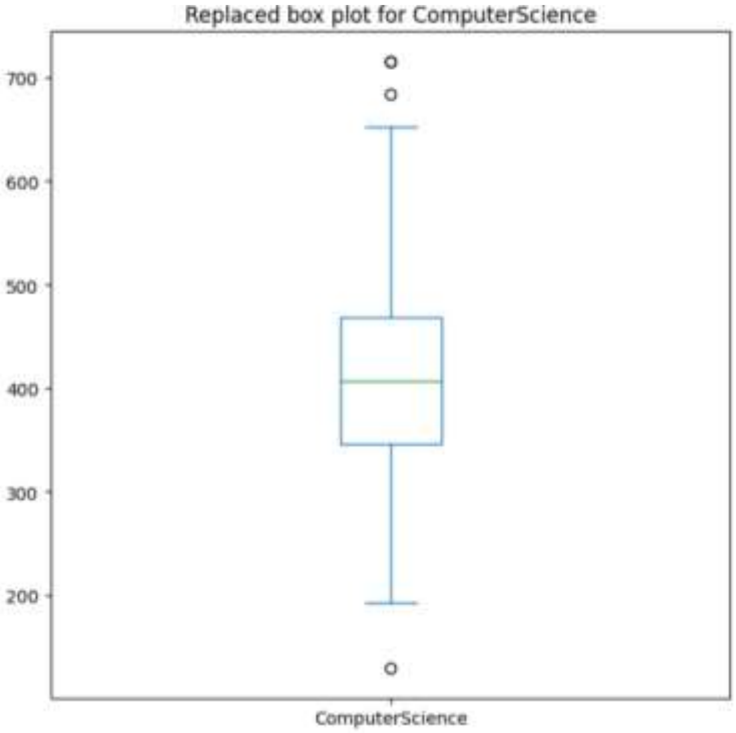
- Many individuals pursued their college education in Uttar Pradesh, possibly due to the concentration of technical colleges in areas like Noida and Greater Noida.



- The dataset includes -1 values in the AMCAT score-related columns. To visualize the impact of replacing these -1 values with NaN, two probability distributions were plotted. The orange curve represents the distribution before replacing -1 values by NaN, while the blue curve represents the distribution after replacing -1 values.
- We can observe that after replacing -1 with a NaN value, the distribution exhibits characteristics of a normal distribution.



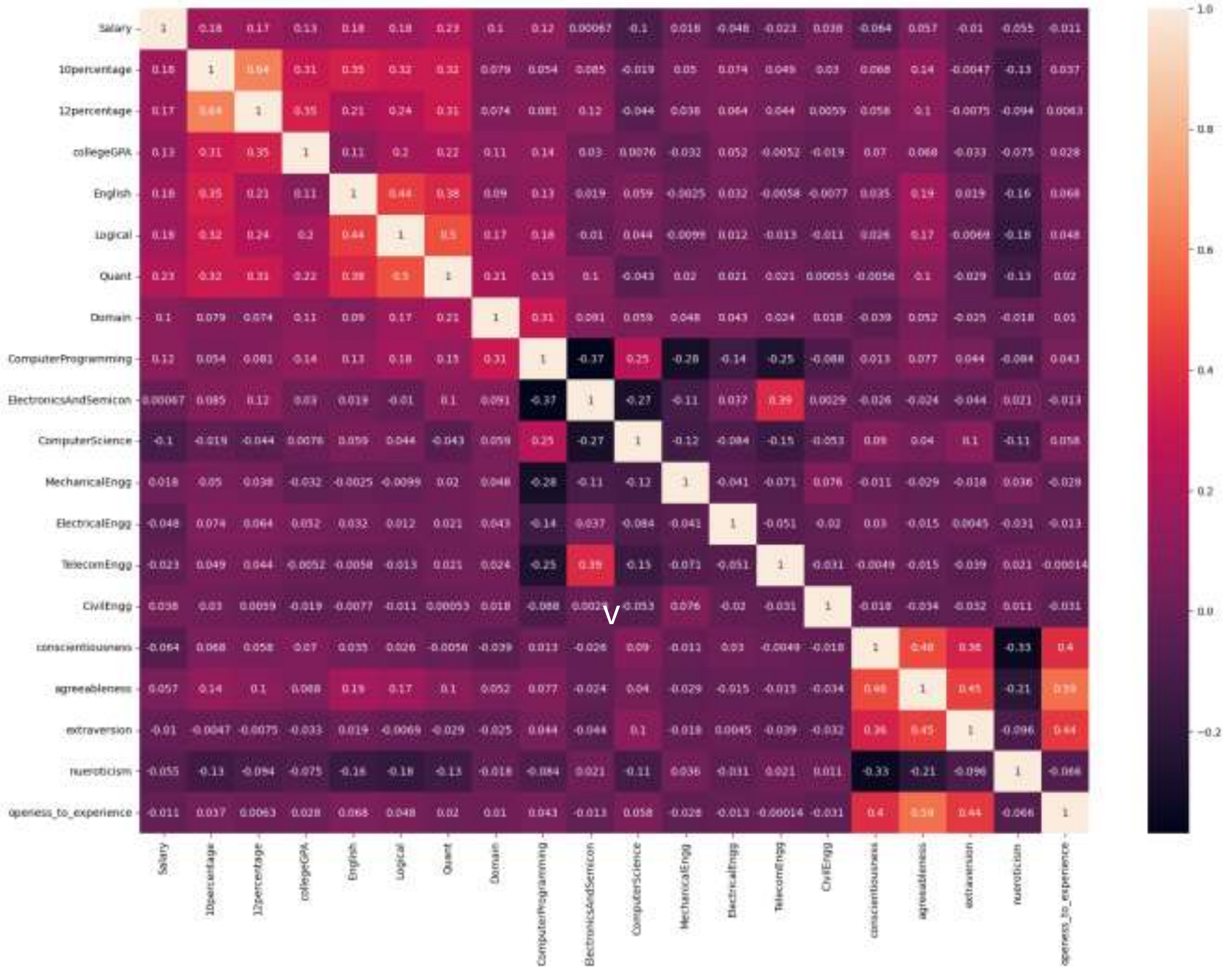
- The box plot on the upper side represents the data before replacing -1 values with NaN, while the box plot on the right side represents the data after replacing -1 values with NaN.
- It is evident that -1 values significantly contribute to the presence of outliers in the dataset. Therefore, removing them is recommended.



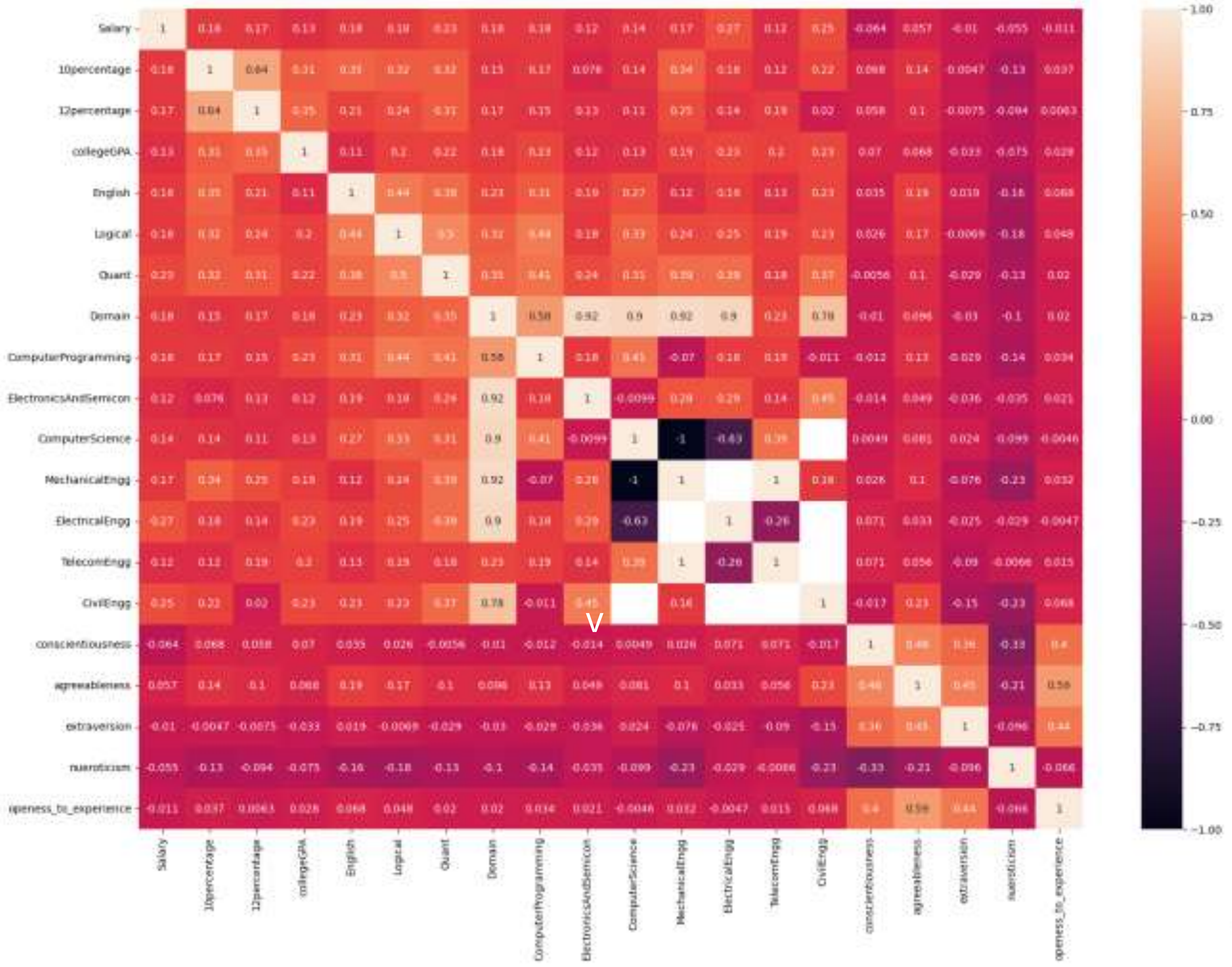
Exploring Relationships: Correlation Analysis

We determine correlation either by calculating correlation coefficients or by visually analyzing scatter plots. In this analysis, I created a heatmap of the correlation matrix and generated scatter plots to explore relationships between variables.

I created two heatmaps: one using the original data before replacing -1 values with NaN, and one using the modified data after replacing them. Let's examine the differences revealed by these plots.

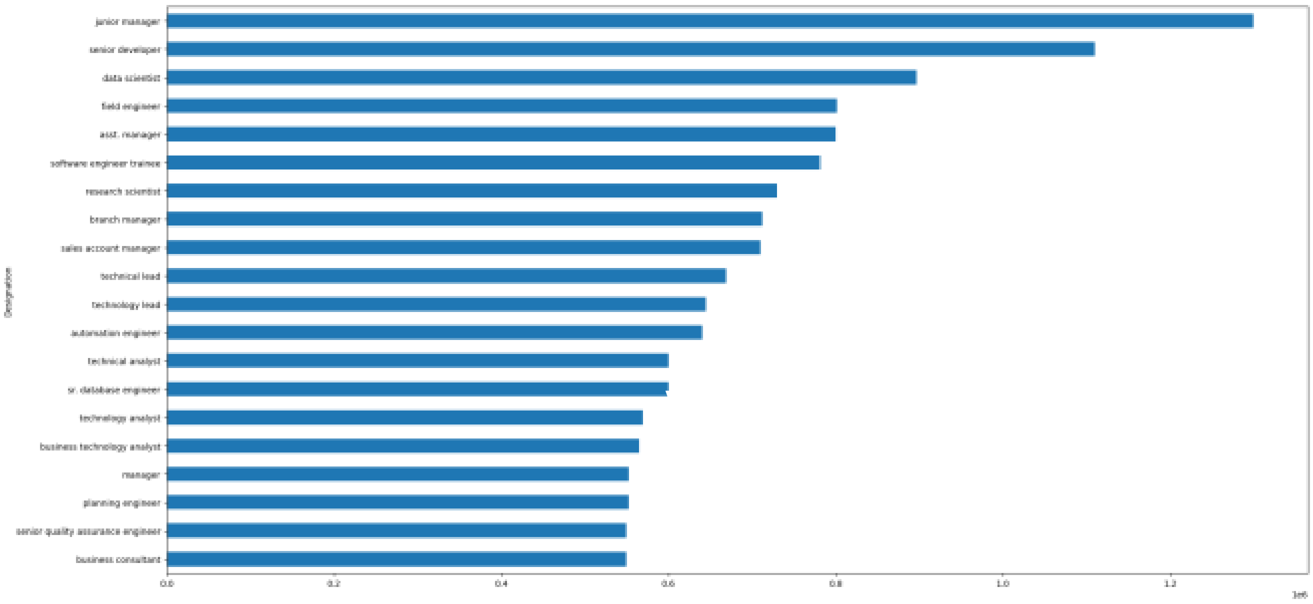


This heatmap represents the correlation matrix before replacing the -1 values with NaN. The correlation appears to be very low, as indicated by the dull colors across the entire heatmap. Unfortunately, the heatmap fails to provide meaningful insights about the correlation.

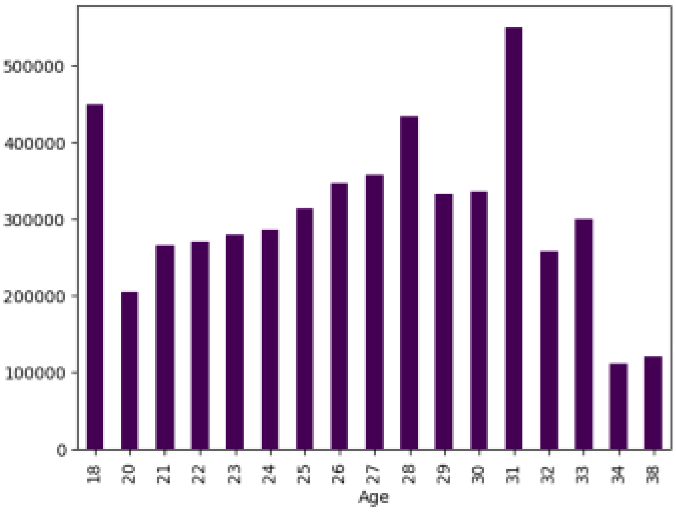


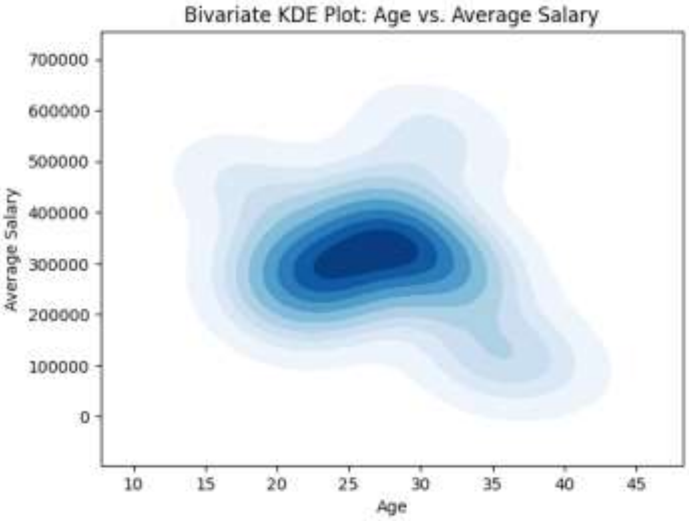
- This heatmap represents the correlation matrix after replacing the -1 values with NaN. There is a notable improvement in the correlation of the columns, as evidenced by the brighter colors compared to the previous heatmap. The enhanced visibility in this heatmap allows for a clearer understanding of the relationships between variables.
- There is a significant multicollinearity observed between the **Domain** column and the other AMCAT score columns.

Exploring Relationships: Bivariate Analysis



- It is evident that individuals holding the designations of **Manager**, **Developer**, and **Data Scientist** are earning more than **8 Lakhs per annum**.





- As observed, there is a higher density of individuals earning salaries between 2.5 LPA and 4 LPA within the age group of 20-33.

Exploring Relationships: Gender And Specialization

Chi-Square Test Results	
Chi-square statistic	104.46891913608455
P - Value	1.2453868176976918e-06
Degrees of Freedom	45

- Our null hypothesis was that specialization depends on gender, while our alternative hypothesis was that specialization does not depend on gender.
- As the dataset primarily consists of engineering students and exhibits low participation of females, it is possible that we do not have enough evidence to draw definitive conclusions.
- The p-value is much smaller than the significance level of 0.05. Therefore, we fail to reject the null hypothesis.

Hypothesis Testing

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.

One-Sample t Test Results	
t - statistic	7.1
P - Value	0.000
Degrees of Freedom	303

- Our null hypothesis was that the average salary of fresh graduates in Computer Science Engineering who take up jobs as a Programming Analyst, Software Engineer, Hardware Engineer, and Associate Engineer is 2.5-3 lakhs..
- The p-value is much smaller than the significance level of 0.05. Therefore, we fail to reject the null hypothesis.
- Since the Times of India claim is from January 18, 2019, and our available data is from 2015, there is a high likelihood that we do not have enough recent evidence to verify the claim.

INNOMATICS® RESEARCH LABS

THANK YOU

INNOMATICS®
RESEARCH LABS