# Exploratory Data Analysis Project

# On AMCAT Data

Libraries Used:

* NumPy
* Pandas
* Seaborn
* Matplotlib
* SciPy

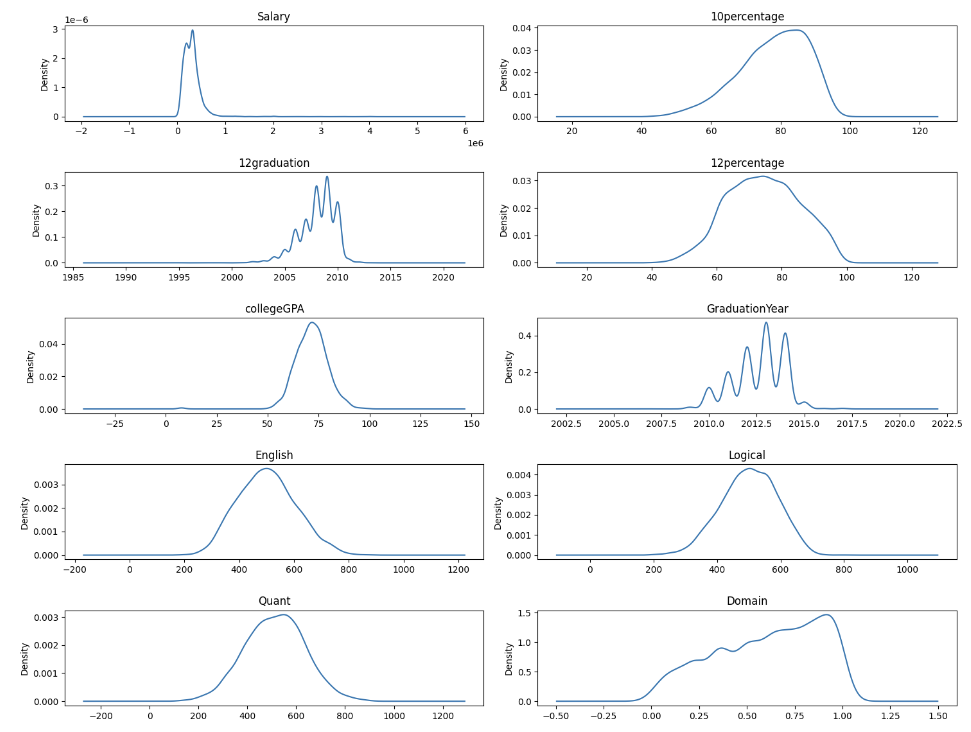
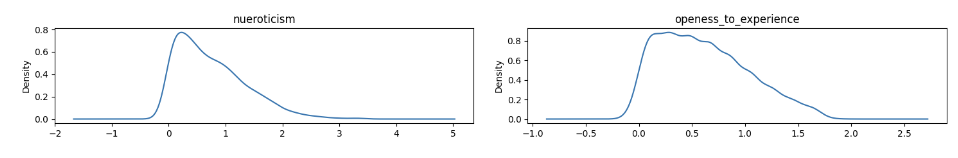
Introduction:

* The AMCAT test is an online employability test for both graduates and ug students.
* The AMCAT dataset consists of the details of the candidates such as salary, DOJ, DOL, city, gender, etc. along with a unique ID column to identify the candidates. In total there are 38 features to assess the salary of the candidates.
* The scores of each candidate in each of the section makes it easier for the employers to see their strong and weak areas.
* There are 3998 records and 38 features among which 27 are numerical columns, 3 are datetime and the rest are categorical
* Although there are no null values, there are values such as -1 and 0 in features such as JobCity and all the scores, except for English, Logical and Quant since these are mandatory sections in AMCAT. The rest of the sections are optional sections and candidates choose to attempt only those sections in which they have their specialization. E.g.: A candidate whose specialization is Computer Science & Engineering will choose to attempt only ‘ComputerProgramming’ or ‘ComputerScience’ in the exam. Hence the rest of the score features would be null values for this candidate which are represented as -1s. Thus, replace all -1s as NaN values

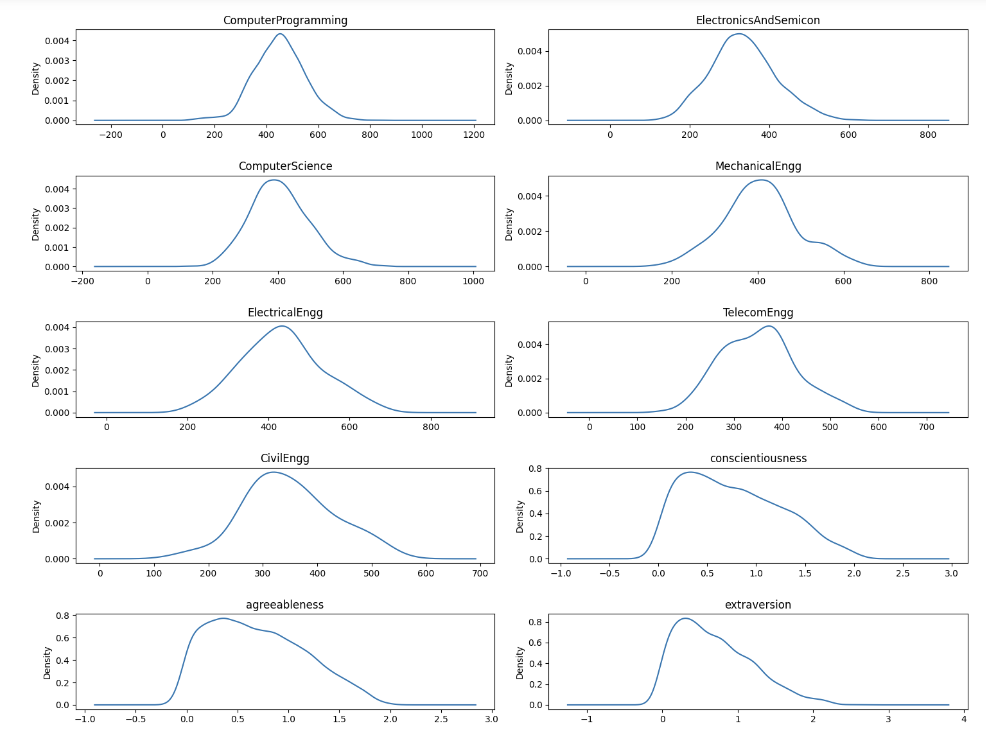
Objective:

Taking the Salary column as the target feature, performing EDA to analyze and see how salary varies with gender, 12th percentage, College GPA and Degree

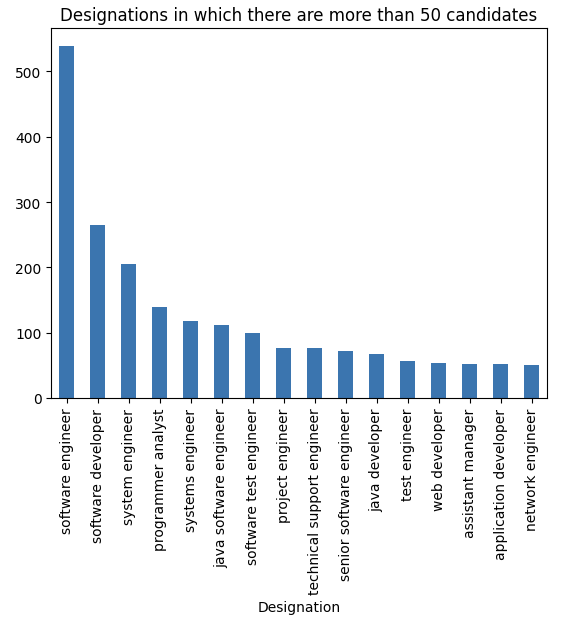
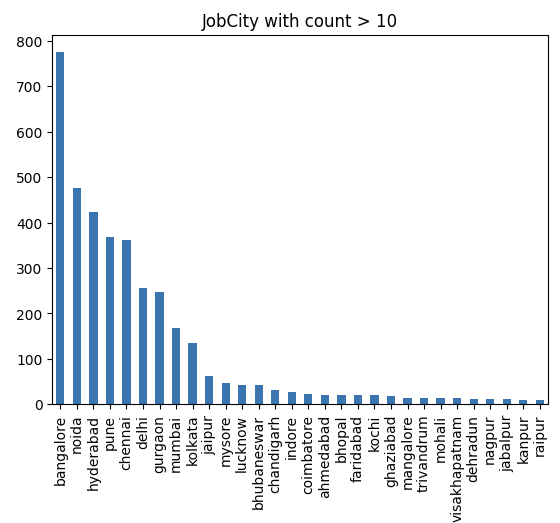
Univariate Analysis (Numerical):

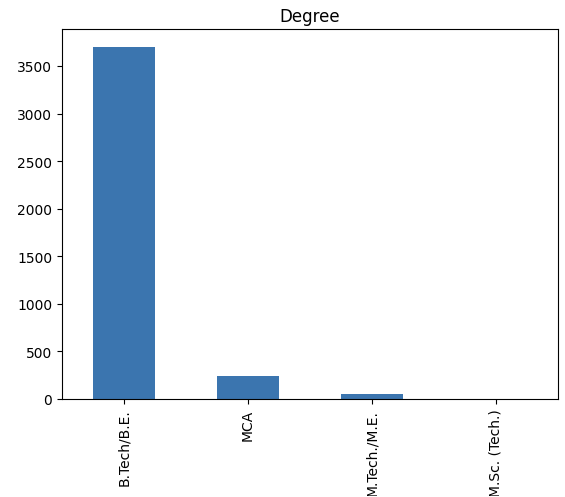
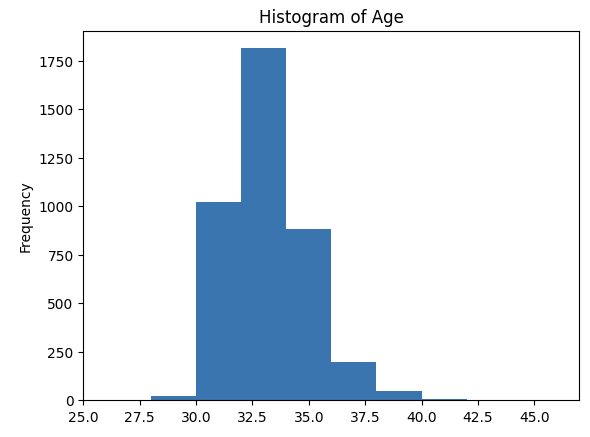
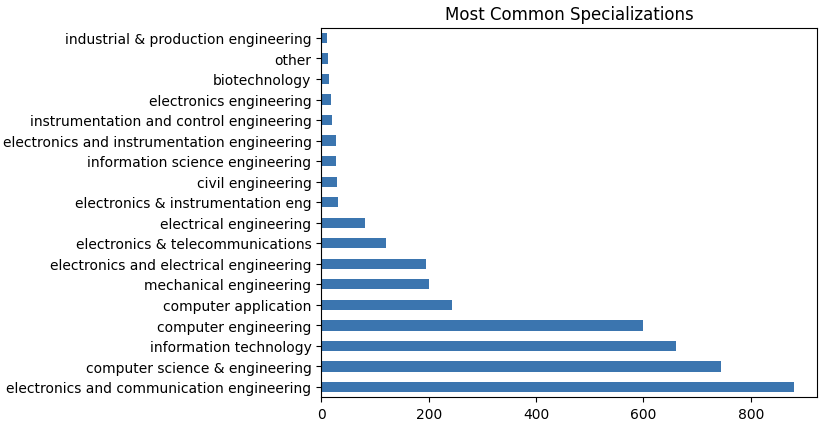
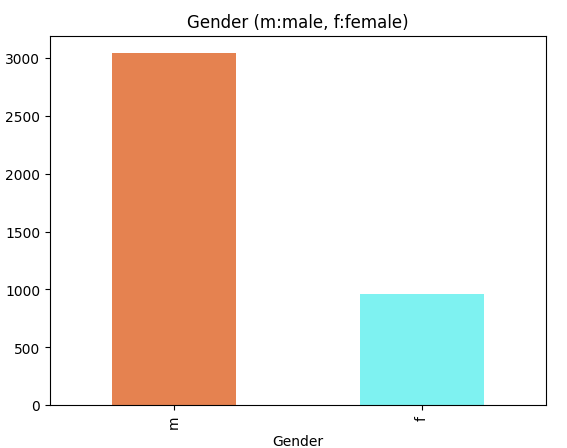


* Most of the features approximately follow normal distribution and some of them are skewed.
* Since Graduation years for 10th, 12th and Bachelor’s is a discrete numerical, the density plot is not a smooth curve
* Salary is shown as 106 and it does not follow normal distribution. The average salary among these 4000 candidates is Rs. 307699.85 and the median is Rs. 300000 with a huge standard deviation of 212737.5
* While the median of 10th, 12th and collegeGPA is all below 70 percent, it has gradually decreased from 79.15 in 10th to 71.72 in college GPA.



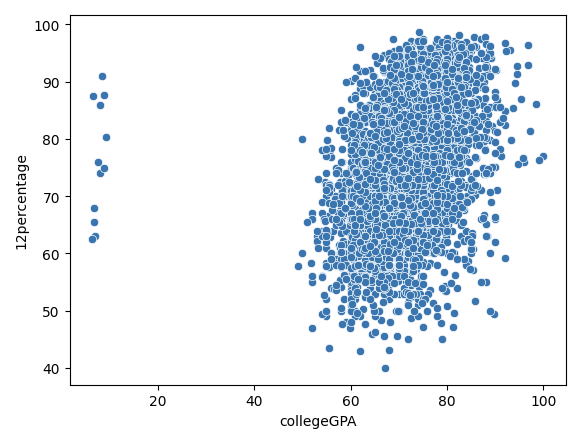
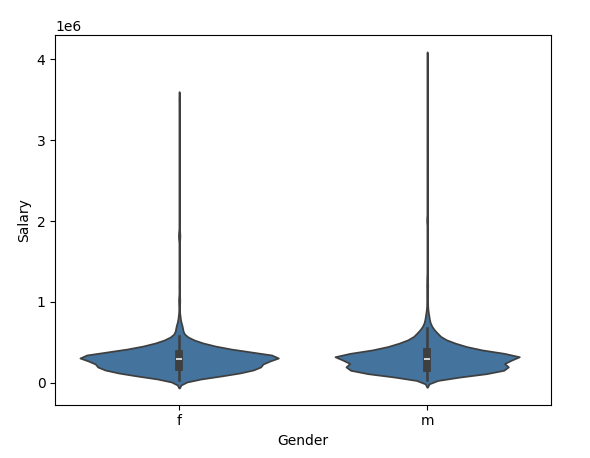
**Univariate Analysis (Categorical):**

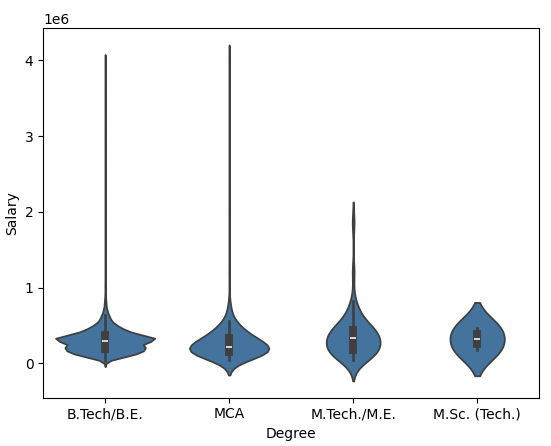
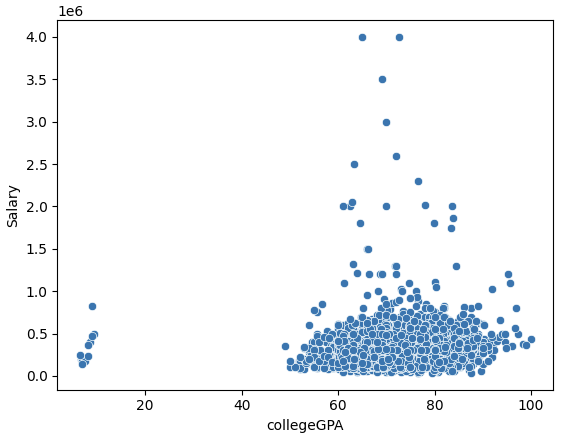
 

* Most of the candidates come from B. Tech/ B. E background and ECE candidates dominate the dataset.
* The meidan age of the candidates is 33 years and most of the candidates are men
* Most common city that they work in is Bangalore followed by Noida, Hyderabad and Pune

**Bivariate:**





* From the first plot, both males and females approximately earn the same amount except for some men who earn more than others (outliers).
* The next plot shows how candidates scored equally well in college although their 12th percentage varies in the range of 50-99%
* Irrespective of your college GPA, people earn the same amount. Scoring higher in college does not imply a higher paying job
* Some of the Bachelor’s and MCA students seem to be doing better than PG candidates since they earn higher salaries. Again, higher qualification does not ensure a higher paying job. But a large proportion of them most of them earn equally well.

**Research Questions:**

1. Using Confidence Interval approach, it shows that the average salary is not between 250k and 300k. The average salary is above 300k i.e., Rs. 338135.47
2. Using Chi-square test, we can conclude that there is a significant difference in specialization depending on gender.

