INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Analysis of AMCAT Data

About me

• Kuppala Sri Sai Raviteja, Civil Engineering graduate transitioning to Data Science .Focused on applying statistical analysis, machine learning, and data visualization techniques.

• Gaining hands-on experience with Python, SQL, and analytical tools. Demonstrated strong interest in expanding skill set within data science field.

• Background in web development, now pivoting towards data-driven roles. Actively developing expertise in statistical methods and machine learning algorithms.

• Proficient in using Python for data analysis and manipulation. Skilled in SQL for database management and querying.

• LinkedIn profile: www.linkedin.com/in/raviteja-kuppala

• Github profile : https://github.com/raviteja4812023

INNOMATICS RESEARCH LABS

Business Problem

Overview

The employment landscape for engineering graduates is multifaceted, with salary outcomes influenced by various factors such as skills, job titles, locations, and demographics. Aspiring Minds' AMEO dataset offers valuable insights into these factors, providing a foundation for understanding and optimizing salary outcomes Specific Challenges.

Solution

Conduct thorough EDA and feature engineering, apply advanced ML algorithms for predictive modelling, and translate insights into actionable recommendations for improved employability and salary negotiations. Predictive Modelling: Accurately estimating salaries considering dynamic market factors is challenging.

Use Case Domain Understanding

Overview

The use case domain revolves around understanding employment outcomes and salary determinants specifically tailored for engineering graduates. It encompasses a range of factors that influence salary outcomes, including skills, job titles, locations, demographics, and industry trends.

Key Components

Salary Determinants: Factors impacting salaries likes kills, job roles, industries, and locations.

Employment Trends: Analysing industry-specific trend sand emerging roles.

Skill Enhancement: Recommending skill development and career advancement opportunities.

Salary Negotiation: Providing tactics for effective Salary negotiation.

Summary of the Data

Dataset Overview:

Total data points: 3,998
Variables: 39 columns, including ID, Salary, Date of Joining (DOJ, Date of Leaving (DOL), Designation, Job City, Gender, Date of Birth (DOB), academic performance (10th and 12th percentages), college details (college ID, tier, state), degree information, specialization, GPA, graduation year, and various skill scores (English, Logical, Quantitative, Domain, Computer Programming, etc.).

Data Cleaning and Preprocessing:

Handling Missing Values: Utilized imputation techniques for missing data in certain columns.
Removing Duplicates: Ensured data integrity by removing duplicate entries.
Standardization: Standardized data formats and encoded categorical variables for analysis.
Outlier Detection: Identified and addressed outliers in relevant columns for accurate analysis.

Analysis Workflow

- Understanding the data - initial exploratory data analysis
- Data cleaning and transformation
- Univariate analysis - Visual and non visual analysis
- Bivariate analysis
- Solutions to hypothesis or questions
- Conclusion

Understanding the Data

```
# summary statistics on the numerical columns
amcat.describe()
```

	ID	Salary	10percentage	12graduation	12percentage	CollegeID	CollegeTier	collegeGPA	CollegeCityID	CollegeCityTier	...	ComputerScience	MechanicalEngg
count	3.998000e+03	3.998000e+03	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	...	3998.000000	3998.000000
mean	6.637945e+05	3.076998e+05	77.925443	2008.087544	74.466366	5156.851426	1.925713	71.486171	5156.851426	0.300400	...	90.742371	22.974737
std	3.632182e+05	2.127375e+05	9.850162	1.653599	10.999933	4802.261482	0.262270	8.167338	4802.261482	0.458489	...	175.273083	98.123311
min	1.124400e+04	3.500000e+04	43.000000	1995.000000	40.000000	2.000000	1.000000	6.450000	2.000000	0.000000	...	-1.000000	-1.000000
25%	3.342842e+05	1.800000e+05	71.680000	2007.000000	66.000000	494.000000	2.000000	66.407500	494.000000	0.000000	...	-1.000000	-1.000000
50%	6.396000e+05	3.000000e+05	79.150000	2008.000000	74.400000	3879.000000	2.000000	71.720000	3879.000000	0.000000	...	-1.000000	-1.000000
75%	9.904800e+05	3.700000e+05	85.670000	2009.000000	82.600000	8818.000000	2.000000	76.327500	8818.000000	1.000000	...	-1.000000	-1.000000
max	1.298275e+06	4.000000e+06	97.760000	2013.000000	98.700000	18409.000000	2.000000	99.930000	18409.000000	1.000000	...	715.000000	623.000000

8 rows × 27 columns

- The DOJ and DOB columns needs to be converted from object to the date type.
- The 'Unnamed: 0' column appears to be irrelevant fo this exploratory data analysis, and hence would need to be removed or dropped.
- There appear to be no null values in any columns.
- However, some columns contain -1 and other negative values, which indicates that these values are not available and will be replaced with 0 instead.

INNOMATICS RESEARCH LABS

Data Cleaning and Transformation

```
cat_df = amcat.select_dtypes(include = ['object'])
num_df = amcat.select_dtypes(include = ['int64', 'float64'])
```

```
# to remove unnecessary
amcat.drop('Unnamed: 0', axis=1, inplace=True)
```

```
[11] # convert DOJ and DOB into datatime
     amcat['DOB'] = pd.to_datetime(amcat['DOB'])
     amcat['DOJ'] = pd.to_datetime(amcat['DOJ'])
```

```
# replace -ve values with 0
for col in columns_to_check:
    amcat.loc[amcat[col] < 0, col] = 0

# to do the count once more
negative_counts = {col: (amcat[col] < 0).sum() for col in columns_to_check}

for col, count in negative_counts.items():
    print("Num of -ve values in '{}': {}".format(col, count))
```

INNOMATICS RESEARCH LABS

Univariate analysis – Non visual analysis

```
Column name:  Degree
count                                                 3998
nunique                                                  4
unique     [B.Tech/B.E., MCA, M.Tech./M.E., M.Sc. (Tech.)]
Name: Degree, dtype: object
Value counts:
 Degree
B.Tech/B.E.      3700
MCA               243
M.Tech./M.E.       53
M.Sc. (Tech.)       2
Name: count, dtype: int64
```

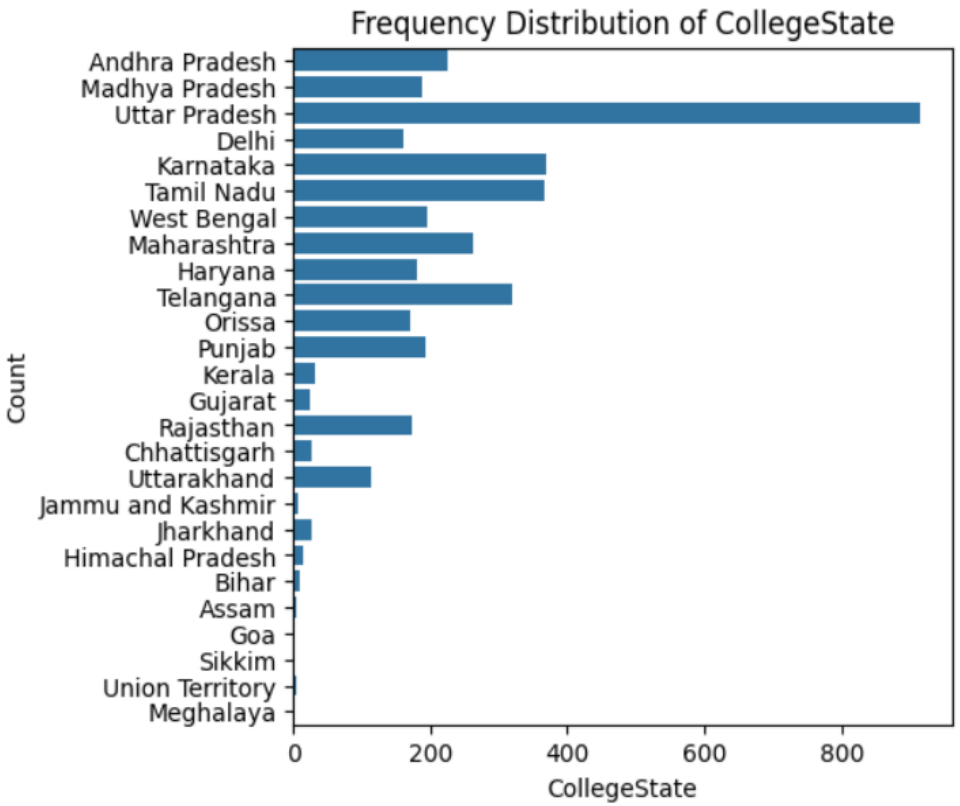
```
Column name:  Gender
count         3998
nunique          2
unique      [f, m]
Name: Gender, dtype: object
Value counts:
 Gender
m    3041
f     957
Name: count, dtype: int64
```

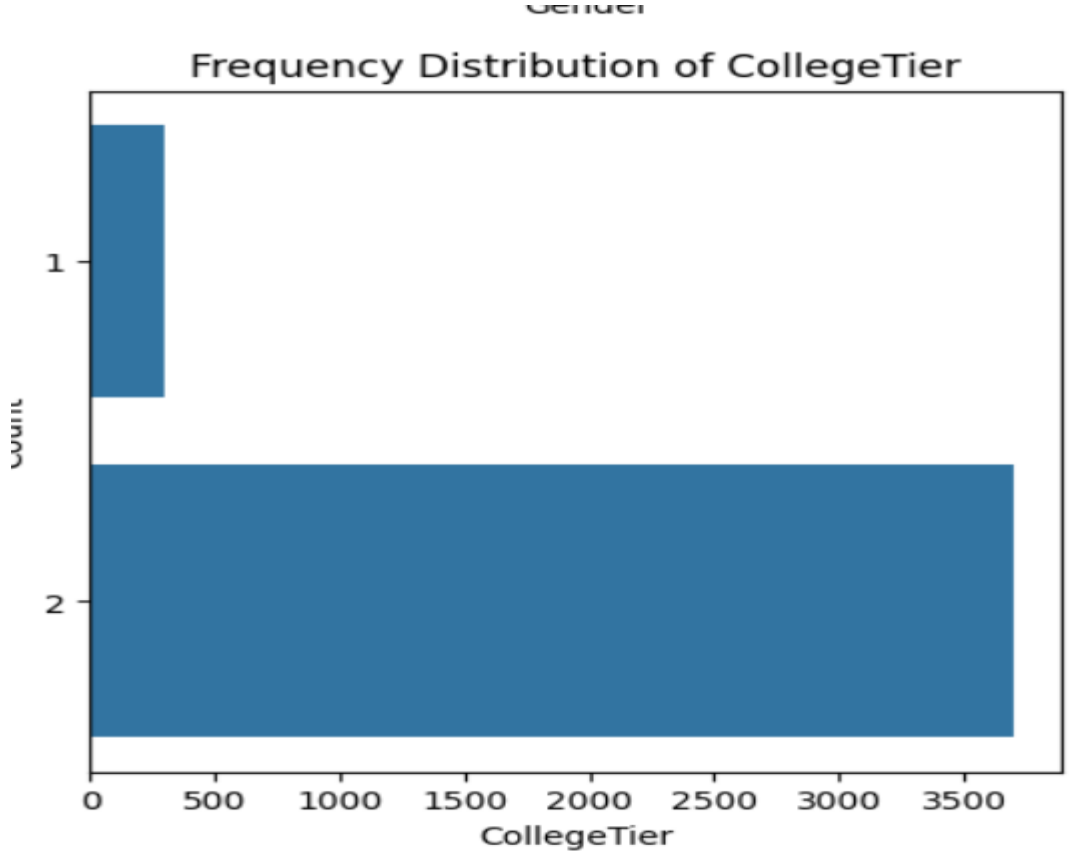
```
Name: English, dtype: float64
Column name:  Logical
min       195.000000
max       795.000000
mean      501.598799
median    505.000000
std        86.783297
```

```
Name: Logical, dtype: float64
Column name:  Quant
min       120.000000
max       900.000000
mean      513.378189
median    515.000000
std       122.302332
```

INNOMATICS RESEARCH LABS

Univariate analysis – visual analysis

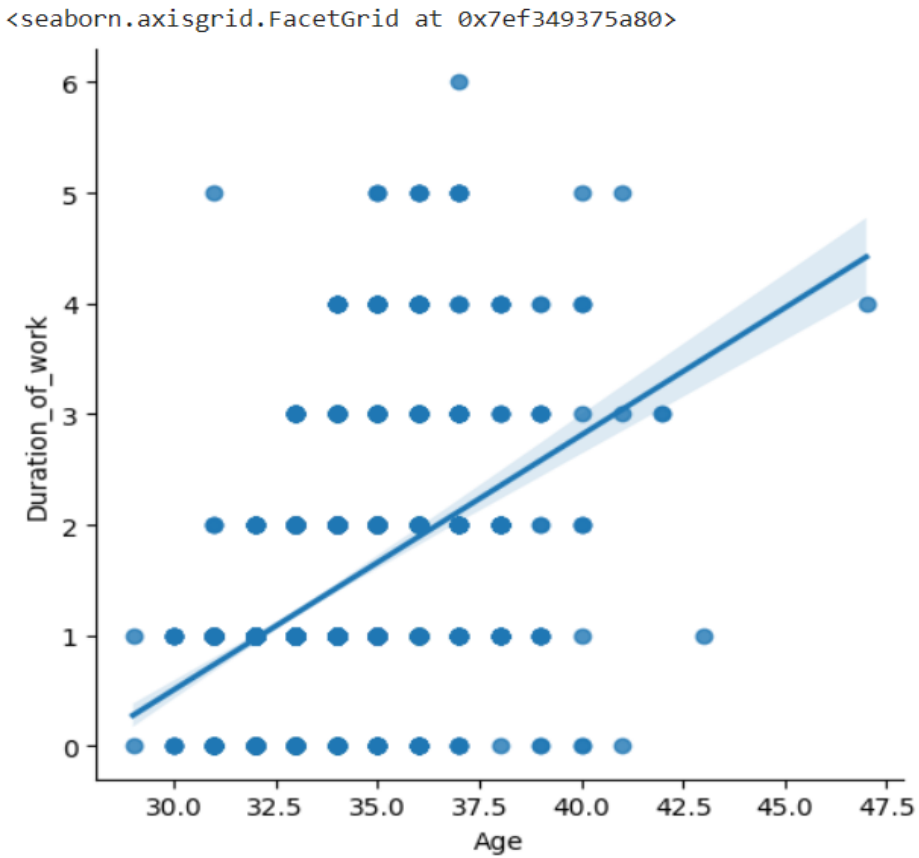


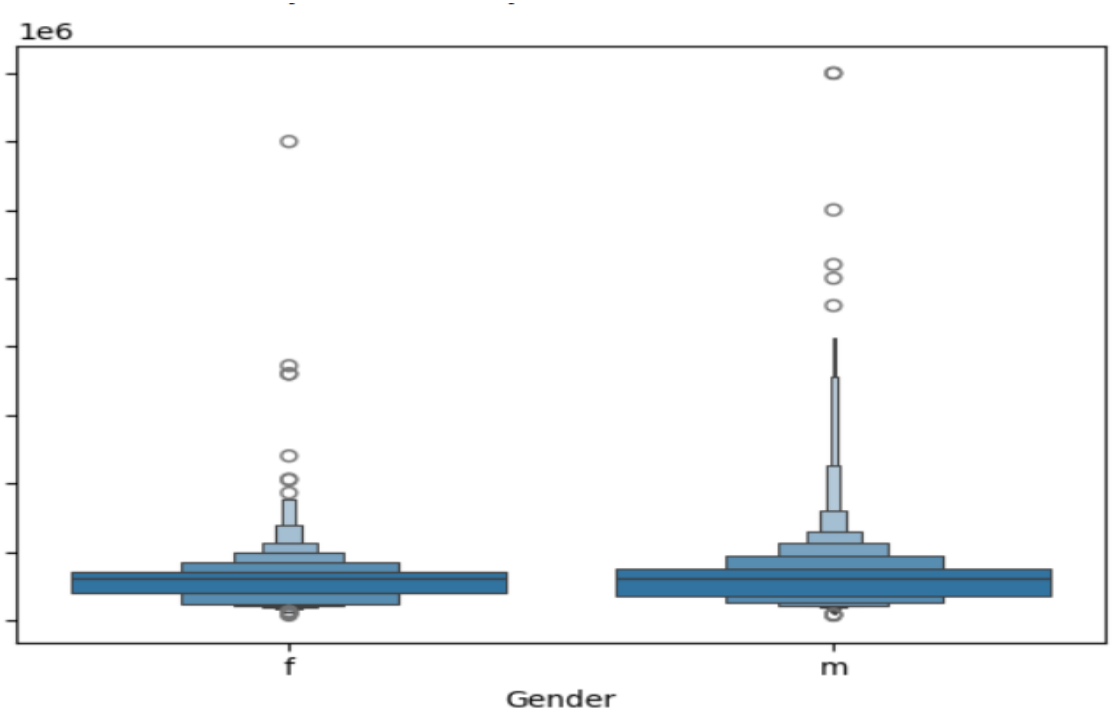


Very few of the candididates have an Msc. (Tech) degree, most of them rather, have a B.Tech\B.E degree.
Most candidates attended the Colleges in Uttar Pradesh state. Following that are Karnataka and Tamil Nadu states as the next state where most candidates attended college.
The number of candidates from College Tier 2 exceeds that of Tier 1.

INNOMATICS RESEARCH LABS

Bivariate analysis



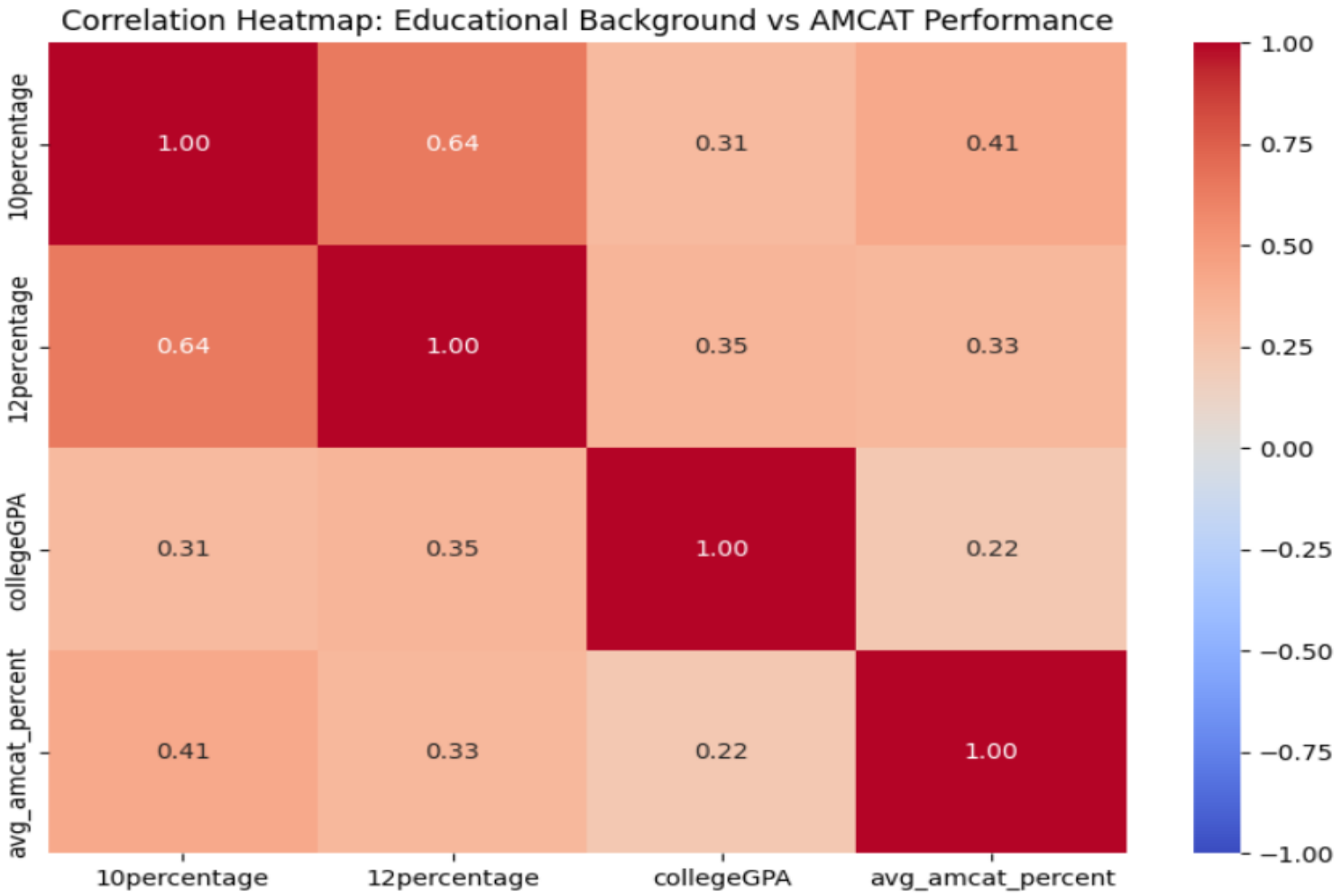


•The salary distribution analysis reveals a notable disparity between male and female earnings, with males generally commanding higher salaries.

•This imbalance in gender representation must be taken into account when interpreting the salary distribution results, as it may influence the overall picture of gender-based income differences in the study.

INNOMATICS RESEARCH LABS

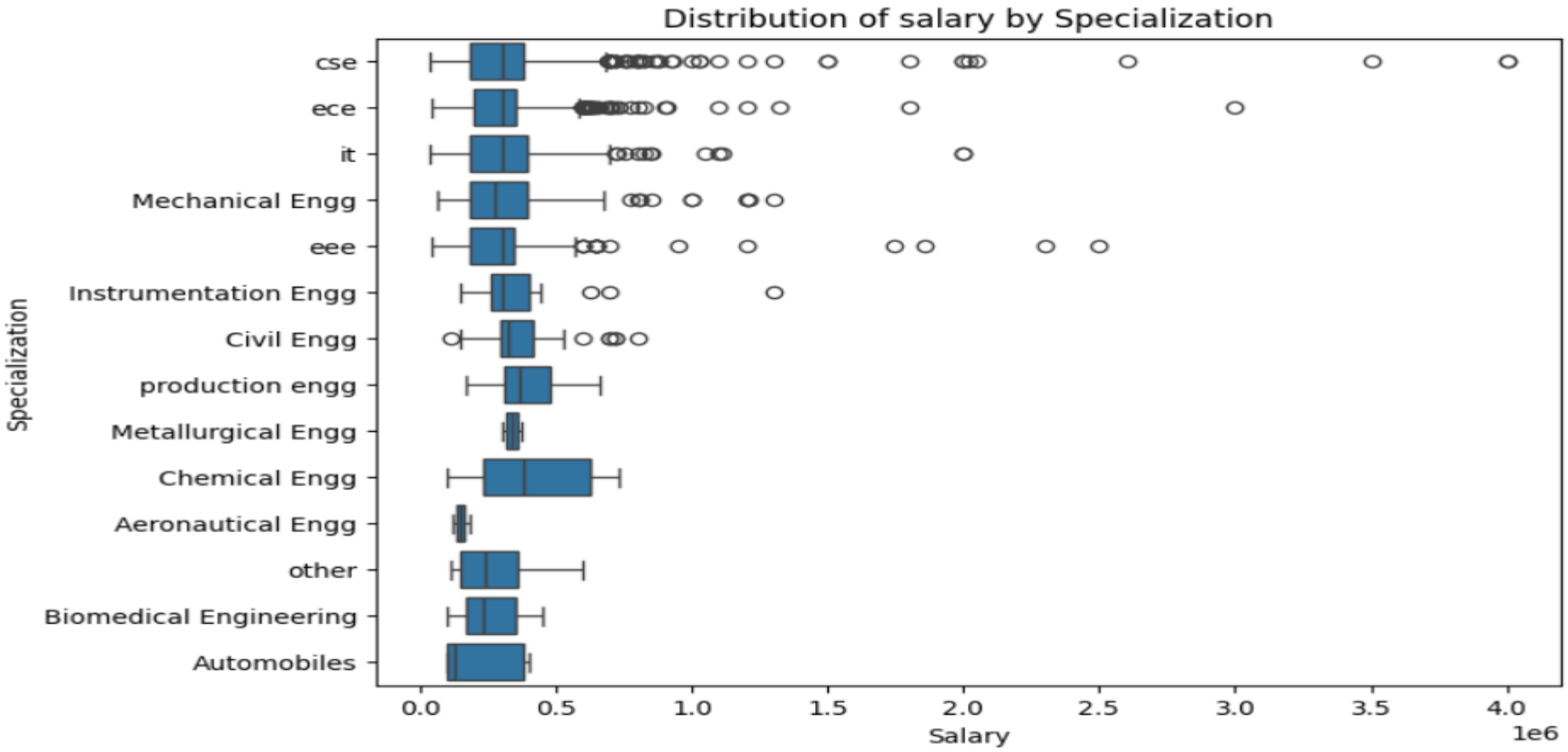
correlation between college GPA and AMCAT scores



The correlation between collegeGPA andAMCAT scores (avg_amcat_percent) is 0.22 which is relatively small and shows very little association between the college GPAof the candidate and his/her AMCATscores.

INNOMATICS RESEARCH LABS

Which specialization earns more salary



- Chemical engineering has a wider spread of salary range.
- Production and Chemical Engineers have the highest median salary amongst the different specializations.
- Aeronautical engineering and Metallurgical Engineering had the least spread of salary ranges.

INNOMATICS RESEARCH LABS

Conclusion

Based on the analysis conducted, the following conclusions can be drawn:

• The 2015 study revealed a higher participation of male candidates compared to female candidates in the tests.

• While the maximum attainable score for each AMCAT test section was 900, only the Quant section saw a perfect score. The English and Logical sections had peak scores of 875 and 795 respectively.

• Analysis indicates minimal correlation between candidates' college GPA and their AMCAT scores.

• Among various specializations, Production and Chemical Engineers command the highest median salaries.

• Regarding the geographical distribution of candidates' educational institutions:
Uttar Pradesh emerged as the state with the highest number of candidates attending college.
Karnataka and Tamil Nadu followed as the second and third most common states for candidates' college attendance.

THANK YOU

INNOMATICS RESEARCH LABS