INNOMATİCS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

An Exploratory Data Analysis on the Aspiring Mind Employment Outcome (AMEO) 2015 study from AMCAT

By
Hannah Igboke

Who's the Data Analyst?

- I am Hannah Igboke, a graduate of Chemical Engineering and a budding data analyst. I am versed in the use of spreadsheets, SQL, Python libraries (Pandas, Numpy, Matplotlib, and many others), and Power BI to extract relevant and domain-specific insights from data. Some of my famous projects include building a scalable database for Olist stores and the Maven Toys sales performance analysis, among others.

- For this project, this is a concise report capturing my data analysis workflow, insights, and conclusion for the exploratory data analysis of the Aspiring Mind Employment Outcome Study conducted in 2015.

- Have a look at my portfolio below :
  Linkedin - Github

INNOMATICS RESEARCH LABS

About AMCAT

AMCAT, known as Aspiring Minds Computer Adaptive Test is an AI-based computer adaptive test which evaluates job applicants on critical areas like communication skills, logical reasoning, quantitative skills, and job-specific domain skills thereby helping recruiters identify the suitability of a candidate for different job roles.

INNOMATICS RESEARCH LABS

Analysis objectives

Following the study conducted in 2015 the AMCAT team were able to gather concrete data with which they hoped to understand what has become of candidates since they took part in the tests and find interesting patterns from the study.

This analysis breaks down the objectives into two:

- **Univariate and Bivariate analysis of variables**
- **Answers and conclusions to relevant hypothesis or questions.**

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

```
#import the pandas library and read data into a dataframe

import pandas as pd

amcat = pd.read_csv('AMCAT.csv')
amcat.head()
```

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	...	ComputerScience	MechanicalEngg	ElectricalEngg
0	train	203097	420000.0	6/1/12 0:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.3	...	-1	-1	-1
1	train	579905	500000.0	9/1/13 0:00	present	assistant manager	Indore	m	10/4/89 0:00	85.4	...	-1	-1	-1
2	train	810601	325000.0	6/1/14 0:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.0	...	-1	-1	-1
3	train	267447	1100000.0	7/1/11 0:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.6	...	-1	-1	-1
4	train	343523	200000.0	3/1/14 0:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.0	...	-1	-1	-1

5 rows × 39 columns

This shows a view of the first five elements of the data.

INNOMATICS RESEARCH LABS

Understanding the data

```
# summary statistics on the numerical columns

amcat.describe()
```

	ID	Salary	10percentage	12graduation	12percentage	CollegeID	CollegeTier
count	3.998000e+03	3.998000e+03	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
std	3.632182e+05	2.127375e+05	9.850162	1.653599	10.999933	4802.261482	0.262270
min	1.124400e+04	3.500000e+04	43.000000	1995.000000	40.000000	2.000000	1.000000
25%	3.342842e+05	1.800000e+05	71.680000	2007.000000	66.000000	494.000000	2.000000
50%	6.396000e+05	3.000000e+05	79.150000	2008.000000	74.400000	3879.000000	2.000000
75%	9.904800e+05	3.700000e+05	85.670000	2009.000000	82.600000	8818.000000	2.000000
max	1.298275e+06	4.000000e+06	97.760000	2013.000000	98.700000	18409.000000	2.000000

8 rows × 27 columns

Observations

- The DOJ and DOB columns needs to be converted from object to the date type
- The DOL column though it contains date values would be left in the object type since it contains 'present' string values which indicates that the candidate still works at a company.
- The 'Unnamed: 0' column appears to be irrelevant fo this exploratory data analysis, and hence would need to be removed or dropped
- College City Tier and College tier are categorical columns, the data type would therefore be converted from int to object.
- There appear to be no null values (na) 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.

This means that for such columns like 'ComputerScience' and others like it, that candidates can take only and not the other since one candidate in this case cannot belong to more than one domain or field.

For the persoanlity traits assessments, it means that there was no valid score or assessment provided for that particular trait.

```
# to check the column characteristics

amcat.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3998 entries, 0 to 3997
Data columns (total 39 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Unnamed: 0             3998 non-null   object
 1   ID                     3998 non-null   int64
 2   Salary                 3998 non-null   float64
 3   DOJ                    3998 non-null   object
 4   DOL                    3998 non-null   object
 5   Designation            3998 non-null   object
 6   JobCity                3998 non-null   object
 7   Gender                 3998 non-null   object
 8   DOB                    3998 non-null   object
 9   10percentage           3998 non-null   float64
 10  10board                3998 non-null   object
 11  12graduation           3998 non-null   int64
 12  12percentage           3998 non-null   float64
 13  12board                3998 non-null   object
 14  CollegeID              3998 non-null   int64
 15  CollegeTier            3998 non-null   int64
 16  Degree                 3998 non-null   object
 17  Specialization         3998 non-null   object
 18  collegeGPA             3998 non-null   float64
 19  CollegeCityID          3998 non-null   int64
```

Snapshots of initial exploratory analysis to understand the data

INNOMATICS RESEARCH LABS

Data Transformation

Data cleaning and formatting

```
# Converting to date time data types

amcat['DOJ'] = pd.to_datetime(amcat['DOJ'])
amcat['DOB'] = pd.to_datetime(amcat['DOB'])

amcat.dtypes
```

```
# converting from int to object

amcat['CollegeTier'] = amcat['CollegeTier'].astype(object)
amcat['CollegeCityTier'] = amcat['CollegeCityTier'].astype(object)

amcat.dtypes
```

```
# to remove the 'Unamed: 0' column

amcat.drop(columns = ['Unnamed: 0'], inplace = True)

amcat.columns
```

```
Index(['ID', 'Salary', 'DOJ', 'DOL', 'Designation', 'JobCity', 'Gender', 'DOB',
       '10percentage', '10board', '12graduation', '12percentage', '12board',
       'CollegeID', 'CollegeTier', 'Degree', 'Specialization', 'collegeGPA',
       'CollegeCityID', 'CollegeCityTier', 'CollegeState', 'GraduationYear',
       'English', 'Logical', 'Quant', 'Domain', 'ComputerProgramming',
       'ElectronicsAndSemicon', 'ComputerScience', 'MechanicalEngg',
       'ElectricalEngg', 'TelecomEngg', 'CivilEngg', 'conscientiousness',
       'agreeableness', 'extraversion', 'nueroticism',
       'openess_to_experience'],
      dtype='object')
```

```
# replacing negative values with 0

# recall the list - columns_to_check

#to replace neative values with 0 in these columns

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))
```

```
Num of -ve values in 'Domain': 0
Num of -ve values in 'ComputerProgramming': 0
Num of -ve values in 'ElectronicsAndSemicon': 0
Num of -ve values in 'ComputerScience': 0
Num of -ve values in 'MechanicalEngg': 0
Num of -ve values in 'ElectricalEngg': 0
Num of -ve values in 'TelecomEngg': 0
Num of -ve values in 'CivilEngg': 0
Num of -ve values in 'conscientiousness': 0
Num of -ve values in 'agreeableness': 0
Num of -ve values in 'extraversion': 0
Num of -ve values in 'nueroticism': 0
Num of -ve values in 'openess_to_experience': 0
```

Data cleaning and transformation steps

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:
 B.Tech/B.E.       3700
MCA                 243
M.Tech./M.E.         53
M.Sc. (Tech.)         2
Name: Degree, dtype: int64
```

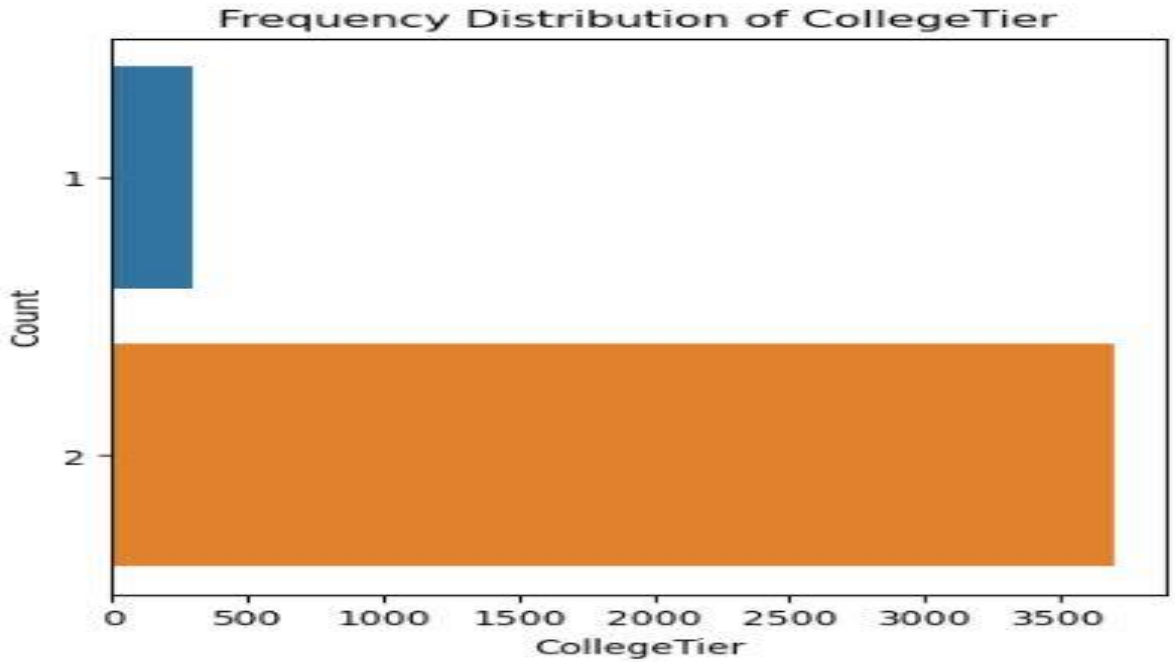
```
Column name:  English
min         180.000000
max         875.000000
mean        501.649075
median      500.000000
std         104.940021
Name: English, dtype: float64
Column name:  Logical
min         195.000000
max         795.000000
mean        501.598799
median      505.000000
std          86.783297
Name: Logical, dtype: float64
Column name:  Quant
min         120.000000
max         900.000000
mean        513.378189
median      515.000000
std         122.302332
```

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

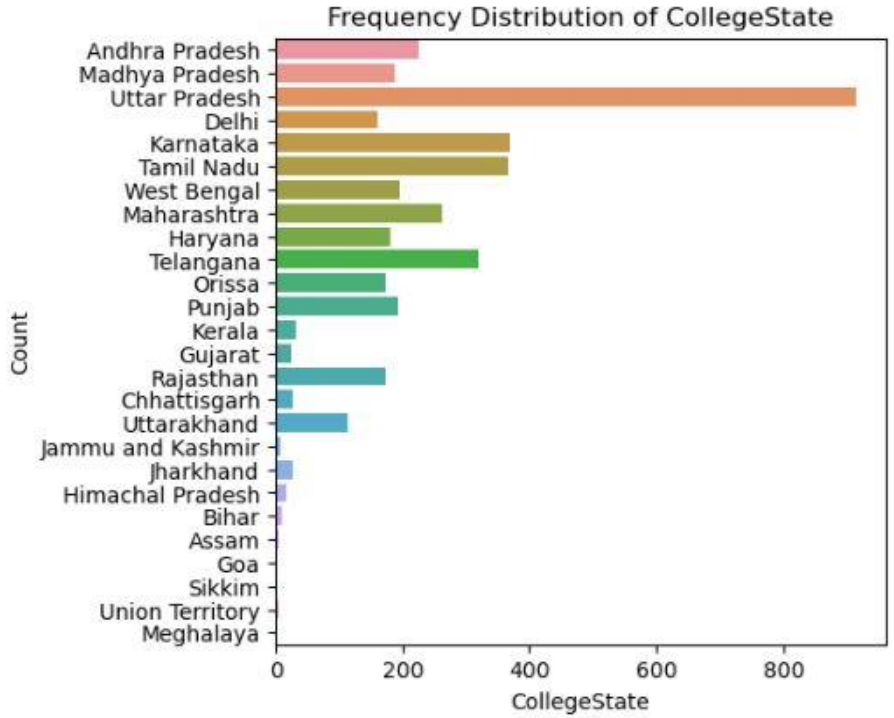
In the study conducted, there were more candidates with a B.Tech/B.E and very few with an M.Sc. Analysis showed that the males were in greater number that year. Also, it shows relevant statistics of the test scores for candidates in the English, Quant and Logical sections.

INNOMATICS RESEARCH LABS

Univariate analysis – visual analysis



The College City Tier refers to the tier of the city in which the college is located. I found that most candidates are from the College city Tier tagged O while a few of them are from the tier



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.

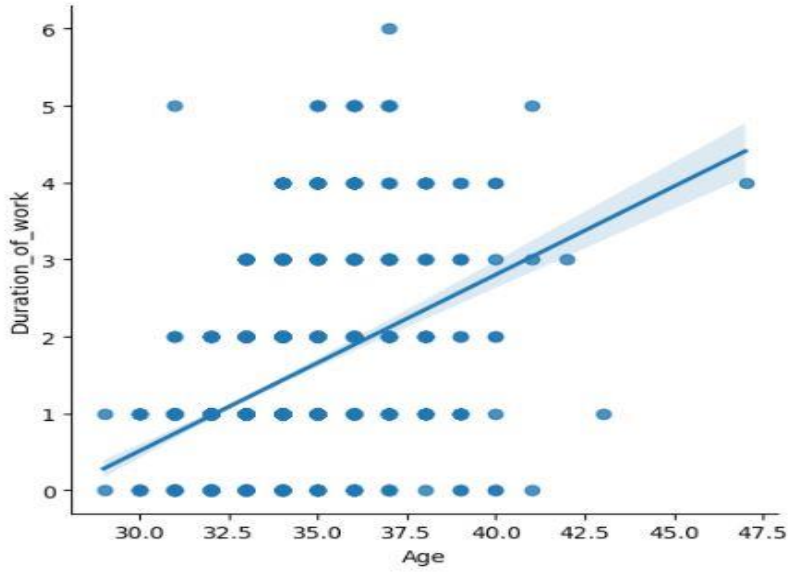
INNOMATICS RESEARCH LABS

Bivariate analysis

Age vs Experience

```
amcat['Age'] = 2024 - amcat['DOB'].dt.year
sns.lmplot(x="Age", y="Duration_of_work", data=amcat)
```

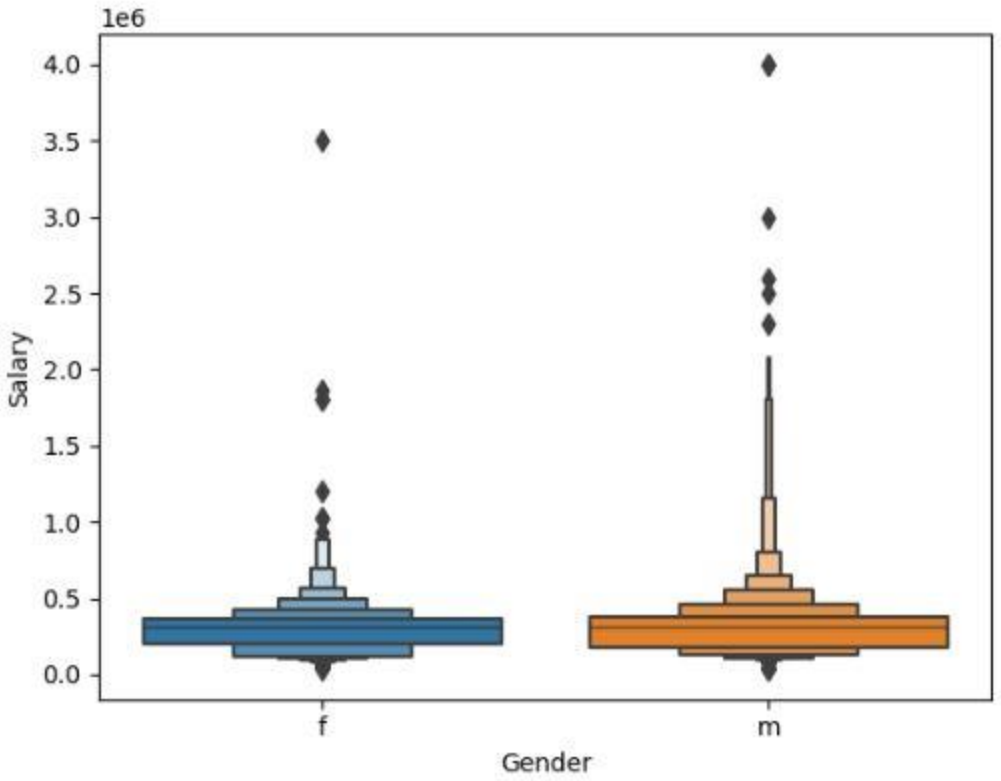
```
<seaborn.axisgrid.FacetGrid at 0x2a3dcf69190>
```



Salary vs Gender

```
sns.boxenplot(data=amcat, x="Gender", y="Salary")
```

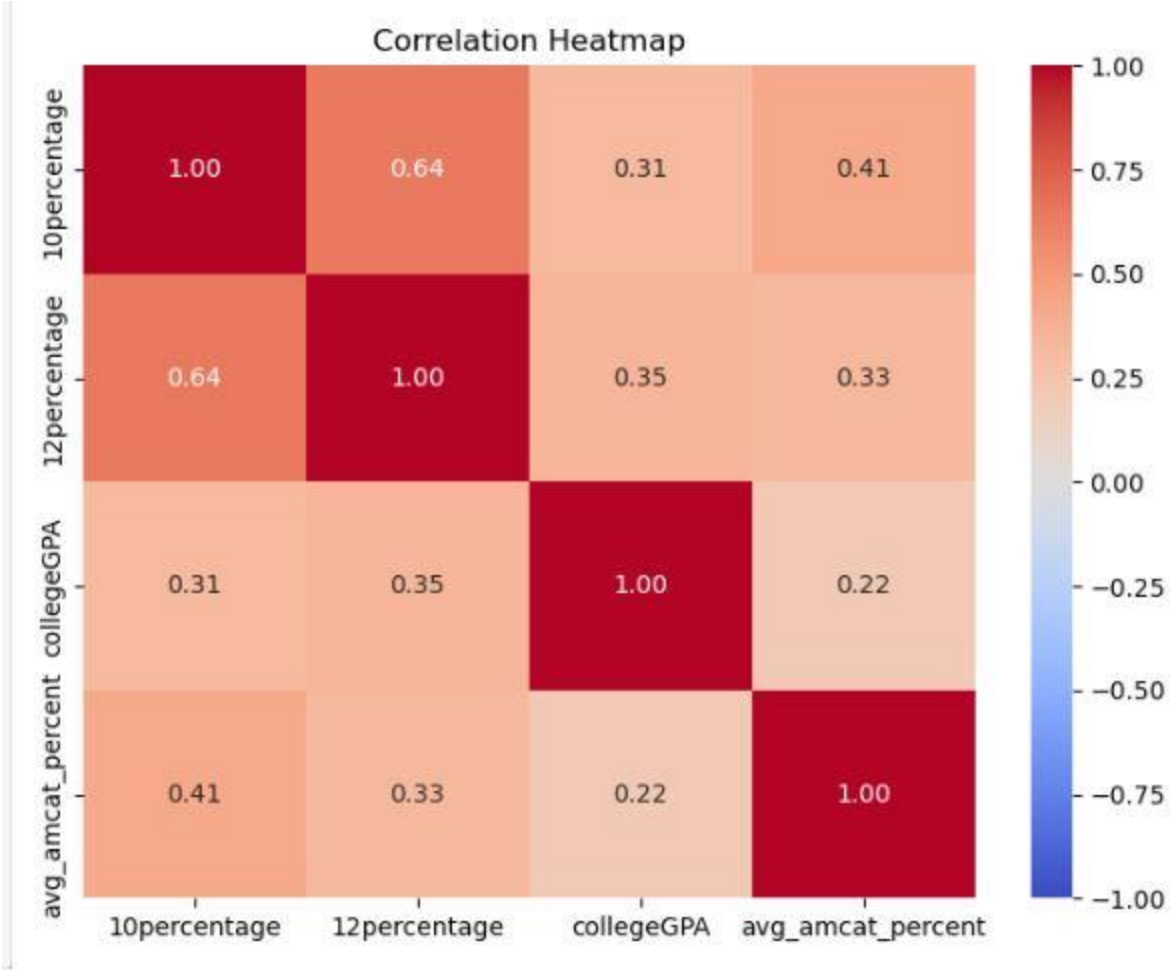
```
<AxesSubplot:xlabel='Gender', ylabel='Salary'>
```



From the plot, it is seen that the distribution of salary for males is higher, which also means they earn more. But as we saw in earlier analysis, the number of males to females in this study is unequal. Also, the plot shows a relationship between the age of duration(length) of work for participants.

INNOMATICS RESEARCH LABS

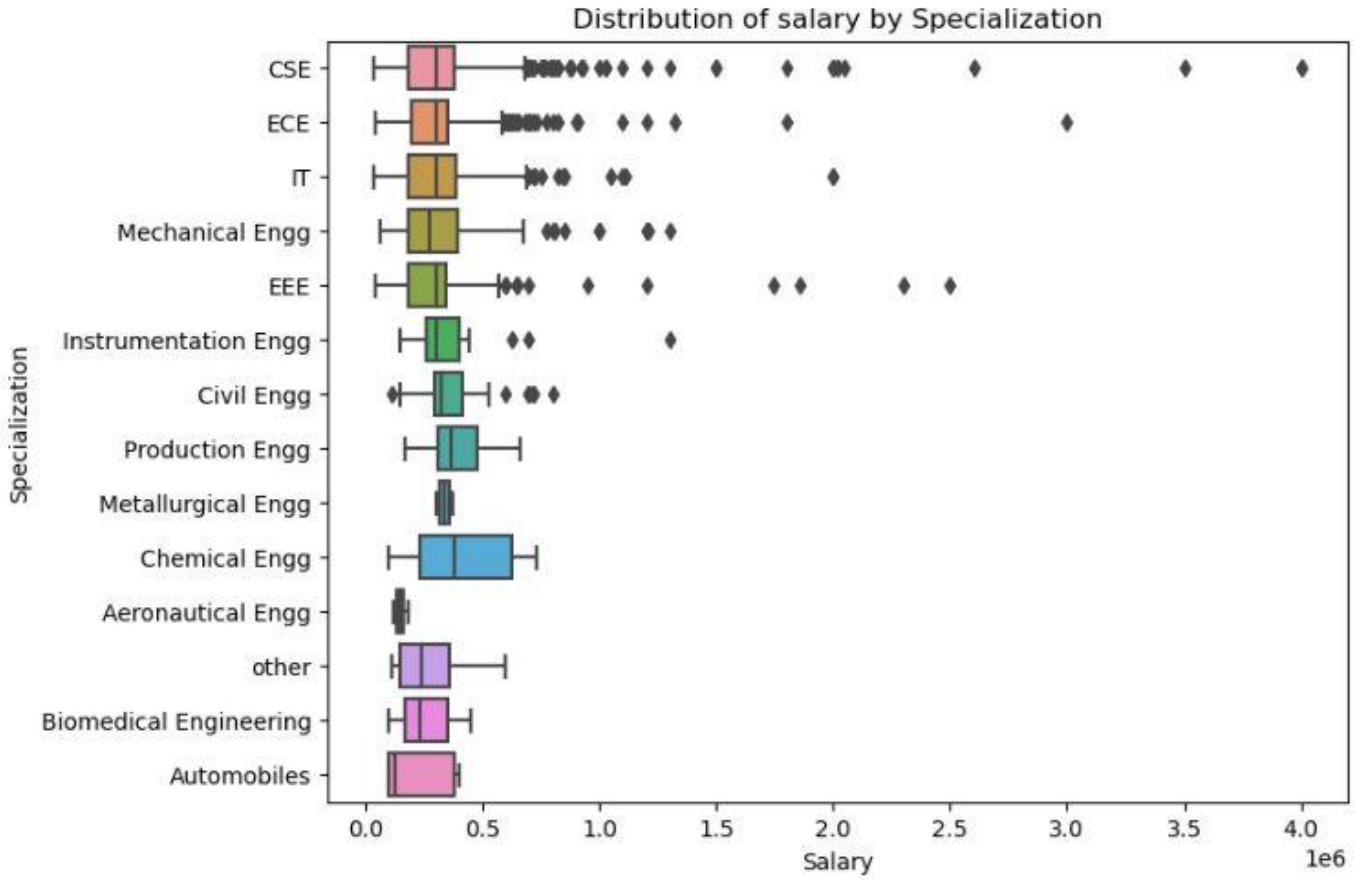
Is there a correlation between college GPA and AMCAT scores?



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

INNOMATICS RESEARCH LABS

What specialization earns more salary?



- ✓ **Chemical engineering has a wider spread of salary range.**
- ✓ **Computer Science (CSE) has the most outlier cases - larger salary cases compared to other fields.**
- ✓ **Aeronautical engineering and Metallurgical Engineering had the least spread of salary ranges**
- ✓ **Production and Chemical Engineers have the highest median salary amongst the different specializations**

INNOMATICS RESEARCH LABS

Is there a relationship between gender and specialization?

```
from scipy.stats import chi2_contingency

# Creating a contingency table
contingency_table = pd.crosstab(amcat['Specialization'], amcat['Gender'])

# Perform chi-square test for independence
chi2, p, dof, expected = chi2_contingency(contingency_table)
print("Chi-square statistic:", chi2)
print("p-value:", p)
```

```
Chi-square statistic: 76.89814311812007
p-value: 4.21134346506098l4e-11
```

Since the p-value (4.21e-11) is much smaller than the typical significance level of 0.05, I reject the null hypothesis - H0: Gender and specialization are independent.

Therefore, I conclude that there is a significant relationship between gender and specialization in the data provided

INNOMATICS RESEARCH LABS

Conclusion

Following the insights generated from my analysis, I can make the following conclusions:

- **In the 2015 study, there were more male candidates compared to the female candidates who took part in the tests.**
- **The maximum scores for each AMCAT test section was 900. A perfect score was achieved in the Quant section alone with the highest scores in the English and Logical section being 875 and 795 respectively.**
- **There is little correlation or association between a candidates college GPA and AMCAT scores.**
- **Production and Chemical Engineers have the highest median salary amongst the different specializations.**
- **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.**

For the full blown analysis check my Github repository.

INNOMATICS RESEARCH LABS

THANK YOU

INNOMATICS RESEARCH LABS