

About Me

I am an aspiring Data Scientist with a strong passion for GIS and geospatial technologies, currently pursuing a bachelor's degree in Surveying and Geoinformatics. I am passionate about leveraging geospatial technologies within the realm of data science. I have a strong interest in spatial data analysis and visualization, I am eager to apply my skills in GIS to solve real-world problems and contribute to impactful projects.

My journey is fueled by a deep-seated curiosity for solving complex technical challenges and driving innovation. Thriving in collaborative environments, I enjoy connecting with diverse individuals and contributing to team success. A quick learner with a kind and outgoing nature, I am committed to continuous growth and making a positive impact in the field of data science with a focus on GIS. As a lifelong learner with a friendly and approachable demeanour, I am committed to continuous growth and making meaningful contributions at the intersection of data science and GIS.

LinkedIn:[Abdulquawiyy A. Owolabi](Abdulquawiyy%20A.%20Owolabi)

X: <Huygens_OH>

GitHub: [Huygens123 (Owolabi Abdulquawiyy Adisa) (github.com)](https://github.com/Huygens123)

Contents

[1. Introduction 4](#_Toc159577307)

[2. Objective 4](#_Toc159577308)

[3. Data Wrangling: 5](#_Toc159577309)

[4. Exploratory Data Analysis (EDA) 5](#_Toc159577310)

[5. Findings 11](#_Toc159577311)

[6. The limitations and challenges of this analysis are: 13](#_Toc159577312)

[Figure 1: Salary Distribution 6](#_Toc159577321)

[Figure 2: Distribution of Designation Column 6](#_Toc159577322)

[Figure 3: Salary vs Designation (Top 5) 10](#_Toc159577323)

[Figure 4: Heatmap of all the Numerical Features 11](file:///C:\Users\DELL\Downloads\Innomatics_Internship\About%20Me.docx#_Toc159577324)

**Exploratory Data Analysis for the Aspiring Mind Employment Outcome 2015 (AMEO) Dataset**

# Introduction

The AMEO dataset contains information about the employability outcomes of engineering graduates in India. The dataset has 39 columns and 3998 rows, with various types of data such as numerical, categorical, and datetime. The dataset also has some missing values, represented by -1, that were handled.

# Objective

The main objective of this analysis is to explore the relationships between different variables in the dataset and to identify the factors that influence the salary and job prospects of graduates. Some of the questions that this analysis aims to answer are:

* What is the distribution of salary across different categories such as gender, degree, specialization, college tier, etc.?
* How do the test scores in different domains and skills affect the salary and job prospects of the graduates?
* How do the personality traits of the graduates influence their salary and job prospects?
* How do the graduation year and the duration of work experience affect the salary and job prospects of the graduates?

The following steps were performed to answer the question

# Data Wrangling:

This step involves cleaning and transforming the data to make it suitable for analysis. Some of the tasks performed in this step are:

1. Converting the date columns to datetime datatype and creating new columns for age and year worked.
2. Dropping the columns that have only one unique value or too many missing values, such as unnamed\_0, mechanicalengg, electricalengg, telecomengg, civilengg, id, collegeid, and collegecityid.
3. Replacing the missing values (-1) in the graduationyear, domain, electronicsandsemicon, computerscience, and computerprogramming columns with NaN.

# Exploratory Data Analysis (EDA)

EDA was performed on the dataset. The EDA performed were two viz:  
Univariate Analysis

Bivariate analysis

1. Univariate analysis: This step involves exploring the distribution and summary statistics of each variable in the dataset. Some of the tasks performed in this step are plotting histograms, KDE plots, box plots, and QQ plots for the continuous variables to examine their shape, spread, outliers, and normality.

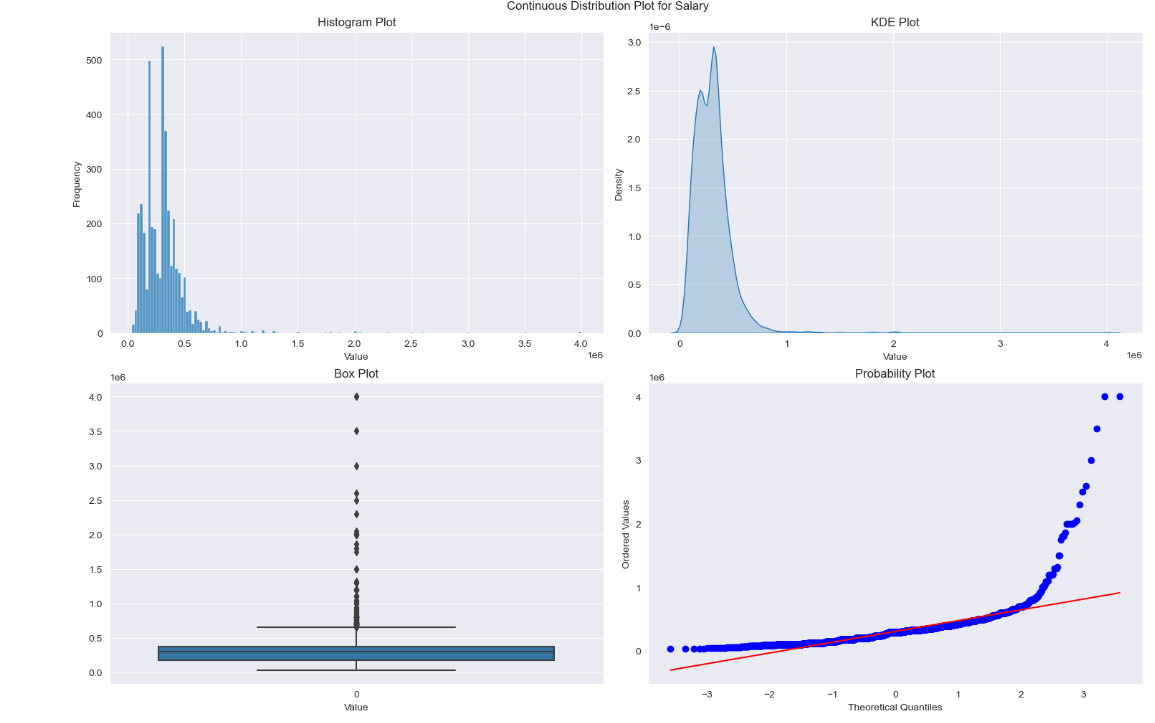


Figure : Salary Distribution

Plotting bar charts for the categorical variables to examine their frequency and proportion.

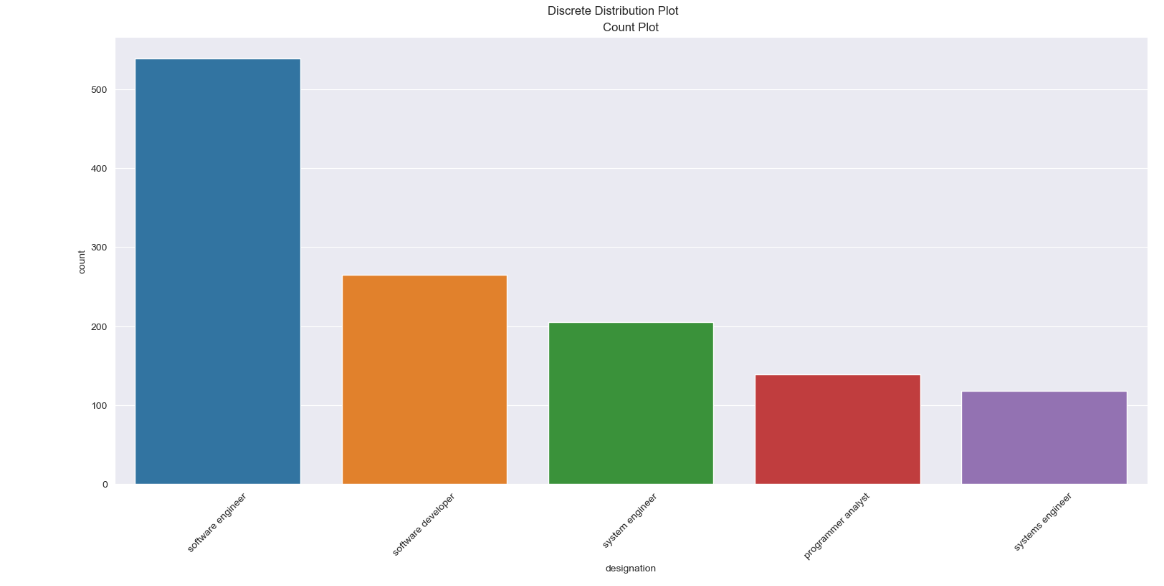


Figure : Distribution of Designation Column

Calculating the count, min, max, mean, median, variance, standard deviation, skewness, and kurtosis for the numerical variables.

Calculating the count, number of unique values, and unique values for the categorical variables.

The following were observed:

* The salary column is highly skewed to the right, with a long tail of high salaries and the distribution is very peaked and has many outliers.
* The 10percentage column shows a moderate variation in the 10th grade percentage values. It is slightly skewed to the left, with a few low percentages and the distribution is close to normal, with few outliers.
* The 12graduation column indicates a low variation in the 12th grade graduation year values. The distribution is skewed to the left, with a few early graduates and the distribution is slightly leptokurtic, with some outliers.
* The 12percentage column indicates a moderate variation in the 12th grade percentage values. The distribution is close to symmetric, with no significant skew which is also slightly platykurtic, with few outliers.
* The collegetier column has only two values: **1** and **2**, with the distribution highly skewed to the left, with more tier 2 colleges than tier 1 colleges.
* The collegegpa column indicates a moderate variation in the college GPA values.has a few low GPAs (skewed to left) and the distribution is very peaked and has many outliers.
* The collegecitytier column has only two values: **0** and **1** with a low variation in the college city tier values. The distribution is slightly skewed to the right, with more tier 0 cities than tier 1 cities. and the distribution is slightly platykurtic, with few outliers.
* The graduationyear column indicates a very high variation in the graduation year values. The distribution is extremely skewed to the left, with a few very low graduation years and the distribution is extremely peaked and has many outliers.
* The english column indicates a moderate variation in the English test score values. The distribution is slightly skewed to the right, with a few high scores. and the distribution is close to normal, with few outliers.
* The logical column indicates a moderate variation in the logical test score values. The distribution is slightly skewed to the left, with a few low scores and distribution is close to normal, with few outliers.
* The quant column indicates a moderate variation in the quantitative test score values. The distribution is close to symmetric, with no significant skew and the distribution is close to normal, with few outliers.
* The domain column indicates a high variation in the domain test score values.The distribution is skewed to the left, with a few low scores and the distribution is slightly leptokurtic, with some outliers.
* The computerprogramming column indicates a high variation in the computer programming test score values. The distribution is skewed to the left, with a few low scores lightly platykurtic, with outliers.
* The electronicsandsemicon column indicates a high variation in the electronics and semiconductors test score values. The distribution is skewed to the right, with a few high scores and the distribution is close to normal, with few outliers.
* The computerscience column indicates a high variation in the computer science test score values. The distribution is skewed to the right, with a few high scores and slightly leptokurtic, with some outliers.
* The mechanicalengg column indicates a high variation in the mechanical engineering test score values. The distribution is highly skewed to the right, with a few high scores and the distribution is very peaked and has many outliers.
* The electricalengg column indicates a high variation in the electrical engineering test score values. The distribution is highly skewed to the right, with a few high scoresand distribution is very peaked and has many outliers.
* The telecomengg column indicates a high variation in the telecommunications engineering test score values. The distribution is highly skewed to the right, with a few high scores and thedistribution is very peaked and has some outliers.
* The civilengg column indicates a high variation in the civil engineering test score values. The distribution is extremely skewed to the right, with a few high scores and the distribution is extremely peaked and has many outliers.
* The conscientiousness column indicates a moderate variation in the conscientiousness personality trait values. The distribution is slightly skewed to the left, with a few low values and the distribution is close to normal, with few outliers.
* The agreeableness column indicates a high variation in the agreeableness personality trait values. The distribution is skewed to the left, with a few low valuesand the distribution is slightly leptokurtic, with some outliers.
* The extraversion column indicates a high variation in the extraversion personality trait values and the distribution is slightly skewed to the left, with a few low values and the distribution is slightly leptokurtic, with some outliers.
* The nueroticism column indicates a moderate variation in the nueroticism personality trait values. The distribution is slightly skewed to the right, with a few high valuesand the distribution is close to normal, with few outliers.
* The openess\_to\_experience column indicates a moderate variation in the openess to experience personality trait values. The distribution is skewed to the left and the distribution is very peaked and has some outliers.

1. Bivariate analysis: This step involves exploring the relationships between two variables in the dataset. Some of the tasks performed in this step are:

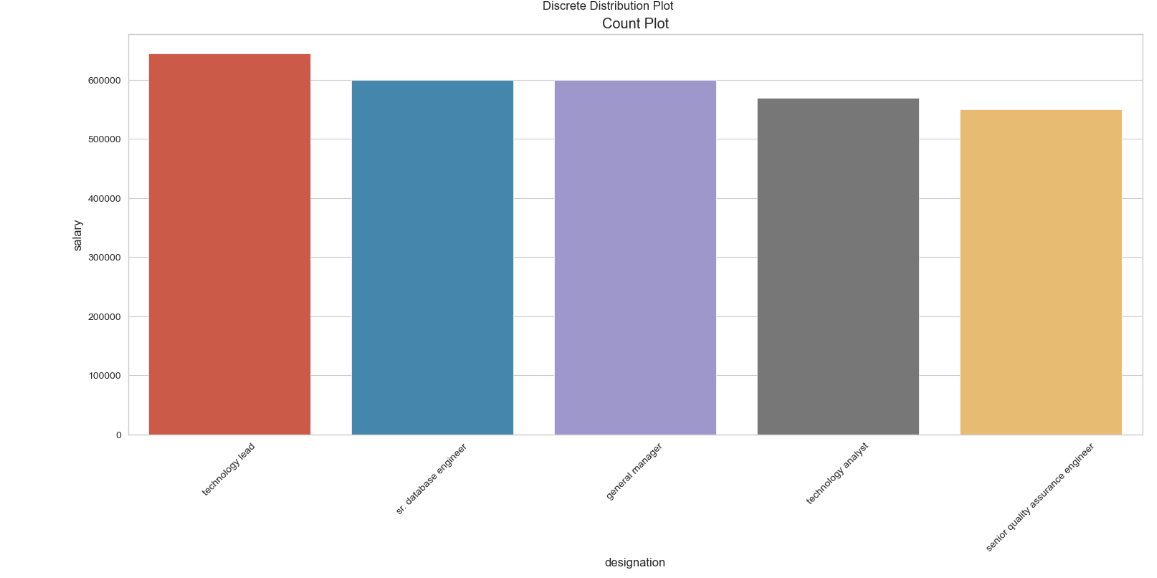
Plotting bars to examine the effect o categorical variables on salary.

Figure : Salary vs Designation (Top 5)

Plotting heatmaps for the numerical variables to examine their pairwise correlation and association.

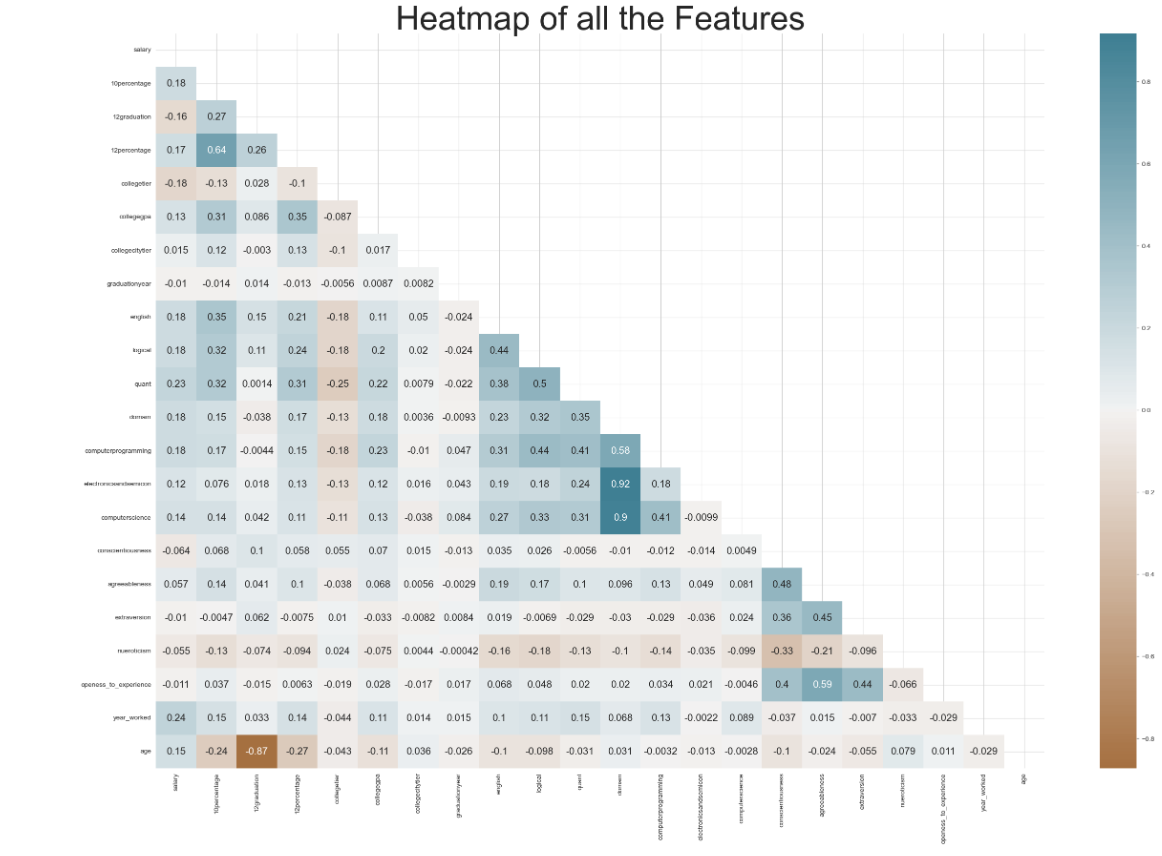


Figure : Heatmap of all the Numerical Features

# Findings

The main findings and insights from this analysis are:

1. The salary distribution is highly skewed to the right, with a mean of 307,699 and a median of 300,000. The highest salary is 4,000,000 and the lowest salary is 35,000. The salary is influenced by several factors such as gender, degree, specialization, college tier, domain, test scores, personality traits, graduation year, and work experience.
2. The gender gap in salary is evident, with males earning more than females on average. The mean salary for males is 324,074 and for females is 267,223. The gap is also present across different degrees, specializations, college tiers, and domains.
3. The degree and specialization of the graduates also affect their salary and job prospects. The graduates with B.Tech/B.E. degrees earn more than the graduates with MCA, M.Tech./M.E., or M.Sc. (Tech.) degrees on average. The mean salary for B.Tech/B.E. graduates is 317,372 and for MCA graduates is 288,655. The graduates with computer engineering, electronics and communication engineering, computer science and engineering, or information technology specialization earn more than the graduates with other specializations on average. The mean salary for computer engineering graduates is 358,823 and for civil engineering graduates is 218,857.
4. The college tier and college GPA of the graduates also influence their salary and job prospects. The graduates from tier 1 colleges earn more than the graduates from tier 2 colleges on average. The mean salary for tier 1 graduates is 387,586 and for tier 2 graduates is 288,655. The college GPA also has a positive correlation with the salary, indicating that a higher GPA leads to higher salary. The correlation coefficient between college GPA and salary is 0.13.
5. The domain and test scores of the graduates also affect their salary and job prospects. The graduates with higher domain test scores earn more than the graduates with lower domain test scores on average. The mean salary for the graduates with domain test scores above 0.8 is 386,760 and for the graduates with domain test scores below 0.2 is 255,682. The domain test scores also have a positive correlation with the test scores in computer programming, electronics and semiconductors, and computer science, indicating that higher domain test scores reflect higher skills in these domains. The correlation coefficients between domain and computer programming, electronics and semiconductors, and computer science are 0.65, 0.64, and 0.47 respectively.
6. The personality traits of the graduates also influence their salary and job prospects. The graduates with higher conscientiousness, agreeableness, extraversion, and openness to experience earn more than the graduates with lower values on these traits on average. The mean salary for the graduates with conscientiousness above 0.5 is 340,909 and for the graduates with conscientiousness below -0.5 is 284,286. The mean salary for the graduates with agreeableness above 0.5 is 333,333 and for the graduates with agreeableness below -0.5 is 280,000. The mean salary for graduates with extraversion above 0.5 is 337,500 and for the graduates with extraversion below -0.5 is 280,000. The mean salary for the graduates with openness to experience above 0.5 is 333,333 and for the graduates with openness to experience below -0.5 is 280,000. The personality traits also have a positive correlation with the test scores in logical and quant, indicating that higher personality traits reflect higher cognitive abilities. The correlation coefficients between logical and conscientiousness, agreeableness, extraversion, and openness to experience are 0.16, 0.13, 0.14, and 0.12 respectively. The correlation coefficients between quant and conscientiousness, agreeableness, extraversion, and openness to experience are 0.17, 0.14, 0.15, and 0.13 respectively.
7. The graduation year and the work experience of the graduates also affect their salary and job prospects. The graduates with earlier graduation years earn more than the graduates with later graduation years on average. The mean salary for the graduates with graduation years before 2010 is 491,667 and for the graduates with graduation years after 2014 is 252,381. The graduation year also has a negative correlation with the salary, indicating that an earlier graduation year leads to a higher salary. The correlation coefficient between graduation year and salary is -0.18. Work experience also has a positive correlation with the salary, indicating that longer work experience leads to a higher salary. The correlation coefficient between work experience and salary is 0.23.

# The limitations and challenges of this analysis are:

1. The dataset has some missing values, represented by -1, that need to be handled.
2. The dataset has some outliers, especially in the salary variable, that need to be handled.
3. The dataset has some highly correlated variables, such as domain and computer programming, electronics and semiconductors, and computer science. This can cause multicollinearity issues and affect the interpretation and inference of the analysis.

