INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Project on AMCAT

INNOMATICS
RESEARCH LABS

ABOUT ME:-

- I’m Subhash Veerla. I have completed my graduation in Btech
- I’m a passionate individual with an insatiable curiosity for data science. My enthusiasm drives me to explore the intricacies of data analysis, machine learning, and predictive modeling. I thrive on uncovering meaningful insights from complex datasets, constantly seeking to expand my knowledge and skills in this dynamic field through data-driven decision-making.

OBJECTIVES OF THE PROJECT:

- To analyze the given dataset.
- The analysis aims to gain insights and understanding from the provided dataset.
- Focusing on the relationship between various features and the target variable which is salary.
- To come up with a conclusion and insight.

SUMMARY OF THE DATA:

- The Aspiring Mind Employment Outcome 2015(AMEO) dataset , released by Aspiring Minds, focuses on employment outcomes for engineering graduates. It includes dependent variables such as Salary, Job Titles , and Job Locations, along with standardized scores in cognitive skills, technical skills and personality skills. With around 40 independent variables and 4000 data points, these variables encompass both continuous and categorical data. The dataset also includes demographic features and unique identifiers for each candidate.

STEPS FOR DATA CLEANING:

- Understanding the Data
- Handle missing values
- Dealt with duplicates and handled outliers
- Used statistical methods to clean the data when required
- Validated the data and tested the cleaned data
- Imported required libraries
- Read the data, selecting columns, filtering the data
- Creating and dropping columns/rows

GIVEN DATA:

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	...	ComputerScience	MechanicalEngg
0	train	203097	420000.0	6/1/12 0:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.30	...	-1	-1
1	train	579905	500000.0	9/1/13 0:00	present	assistant manager	Indore	m	10/4/89 0:00	85.40	...	-1	-1
2	train	810601	325000.0	6/1/14 0:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.00	...	-1	-1
3	train	267447	1100000.0	7/1/11 0:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.60	...	-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	78.00	...	-1	-1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3993	train	47916	280000.0	10/1/11 0:00	10/1/12 0:00	software engineer	New Delhi	m	4/15/87 0:00	52.09	...	-1	-1
3994	train	752781	100000.0	7/1/13 0:00	7/1/13 0:00	technical writer	Hyderabad	f	8/27/92 0:00	90.00	...	-1	-1
3995	train	355888	320000.0	7/1/13 0:00	present	associate software engineer	Bangalore	m	7/3/91 0:00	81.86	...	-1	-1
3996	train	947111	200000.0	7/1/14 0:00	1/1/15 0:00	software developer	Asifabadbanglore	f	3/20/92 0:00	78.72	...	438	-1
3997	train	324966	400000.0	2/1/13 0:00	present	senior systems engineer	Chennai	f	2/26/91 0:00	70.60	...	-1	-1

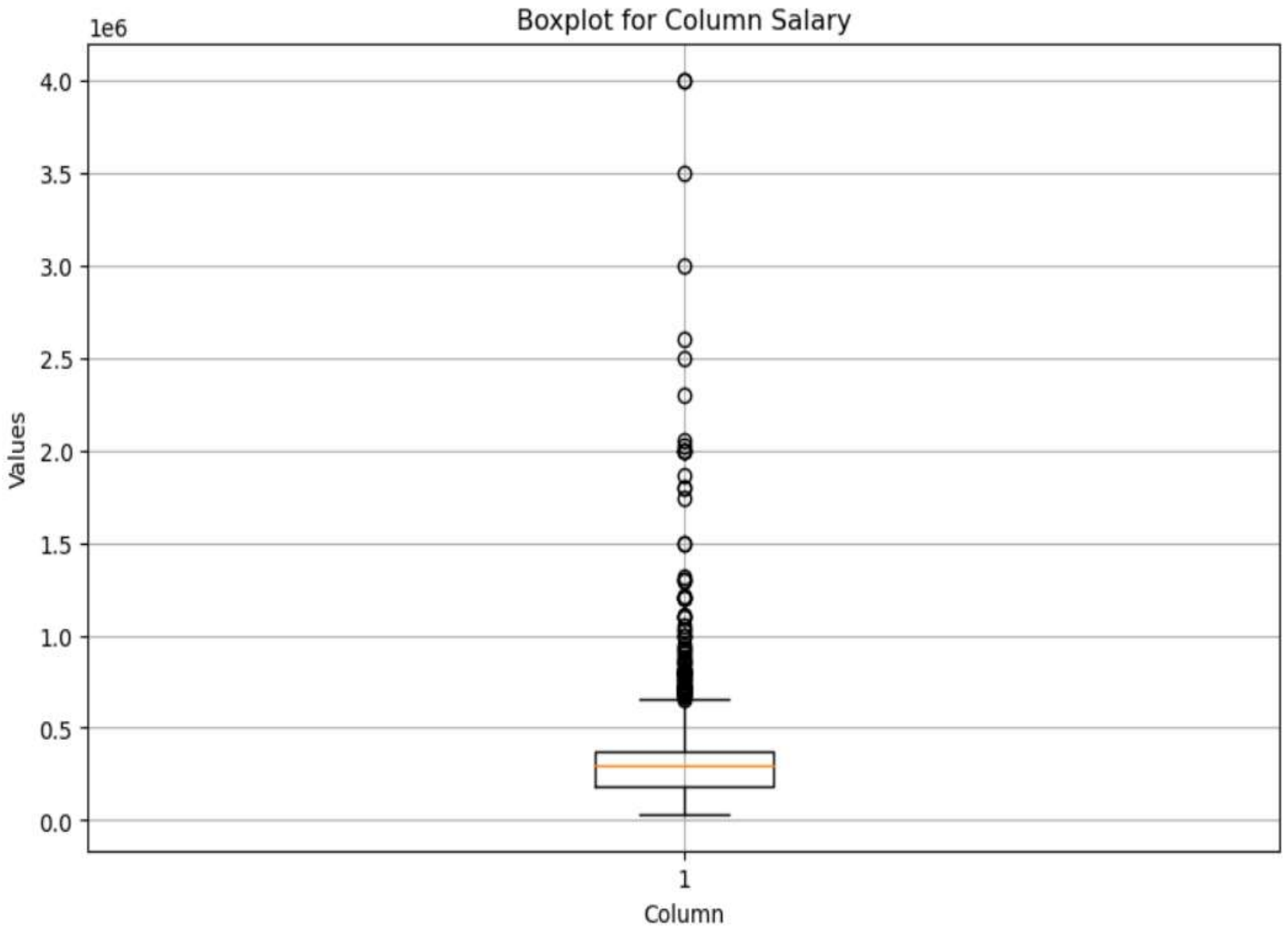
3998 rows × 39 columns

	ID	Salary	10percentage	12graduation	12percentage	CollegeID	CollegeTier	collegeGPA	CollegeCityID	CollegeCityTier	...	Comp
count	3.998000e+03	3.998000e+03	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	...	
std	3.632182e+05	2.127375e+05	9.850162	1.653599	10.999933	4802.261482	0.262270	8.167338	4802.261482	0.458489	...	
min	1.124400e+04	3.500000e+04	43.000000	1995.000000	40.000000	2.000000	1.000000	6.450000	2.000000	0.000000	...	
25%	3.342842e+05	1.800000e+05	71.680000	2007.000000	66.000000	494.000000	2.000000	66.407500	494.000000	0.000000	...	
50%	6.396000e+05	3.000000e+05	79.150000	2008.000000	74.400000	3879.000000	2.000000	71.720000	3879.000000	0.000000	...	
75%	9.904800e+05	3.700000e+05	85.670000	2009.000000	82.600000	8818.000000	2.000000	76.327500	8818.000000	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	...	

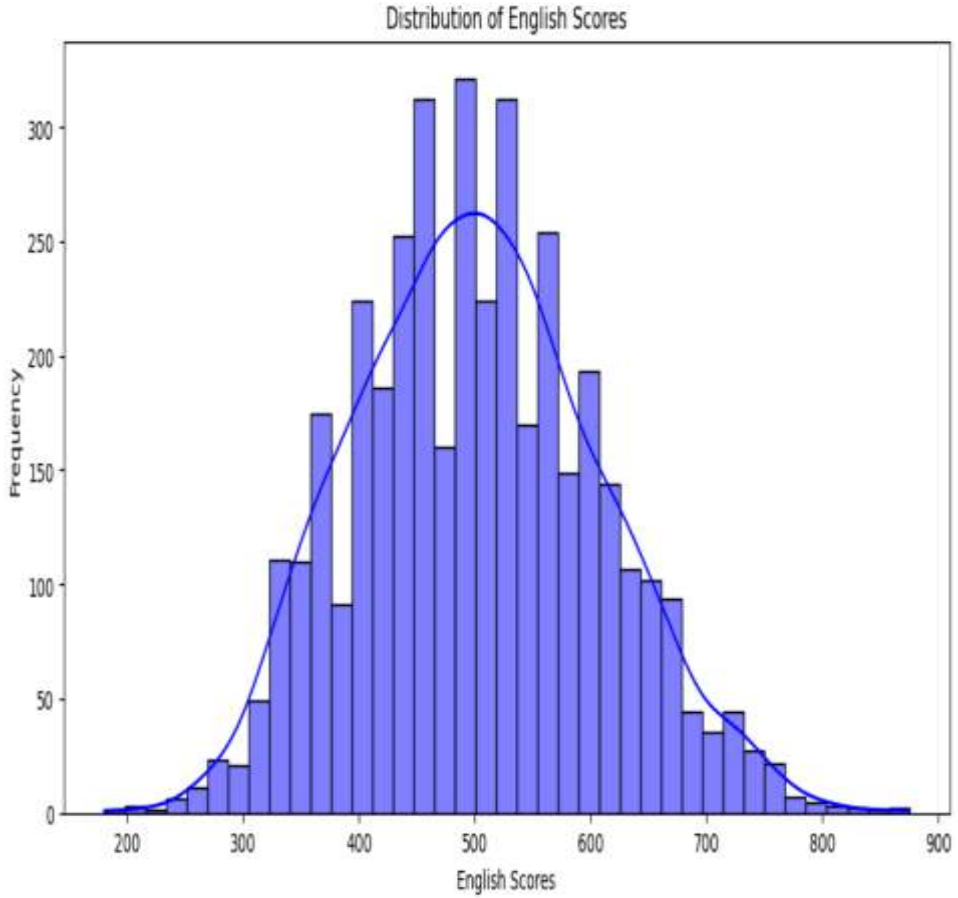
8 rows × 27 columns

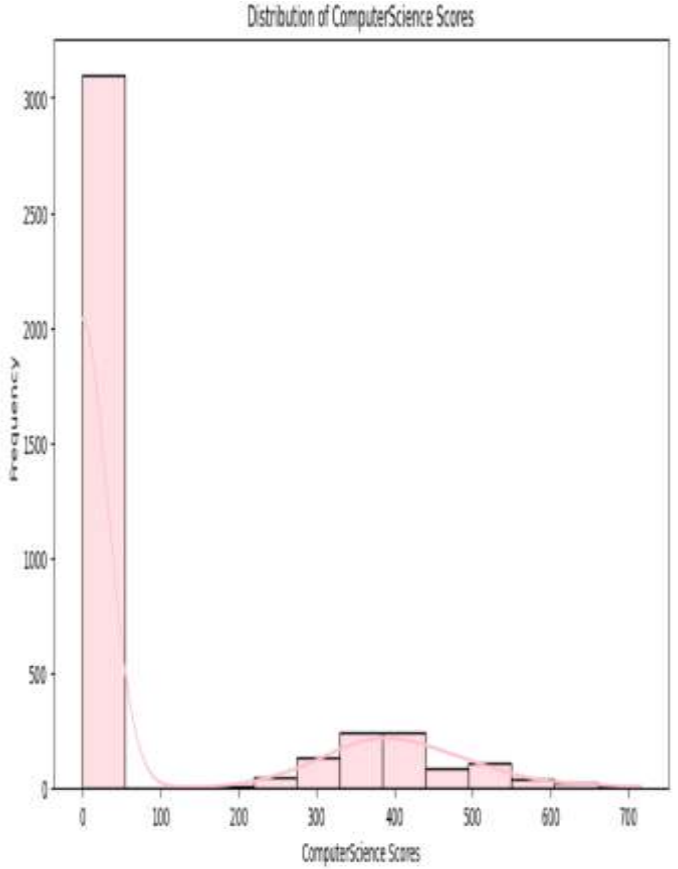
Data Visualisation:

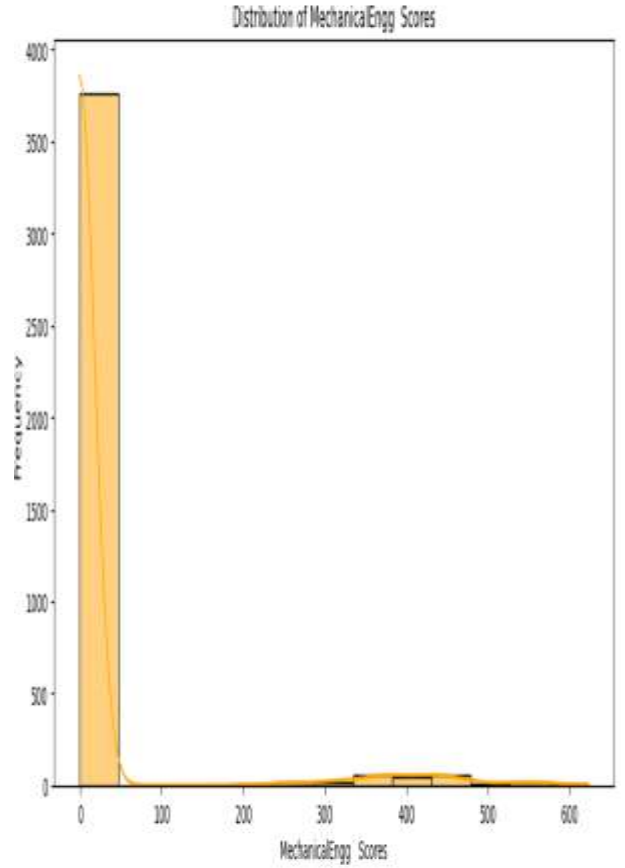
- Univariate Analysis Steps
- Bivariate Analysis Steps and
- Multivariate Analysis Steps



Observation : The boxplot visualizes the distribution of salary values, indicating a wide range of salaries with several outliers towards the higher end. The median salary lies within the lower quartile, suggesting a potential skewness towards lower incomes, while the upper whisker denotes considerable variability in higher salary ranges.

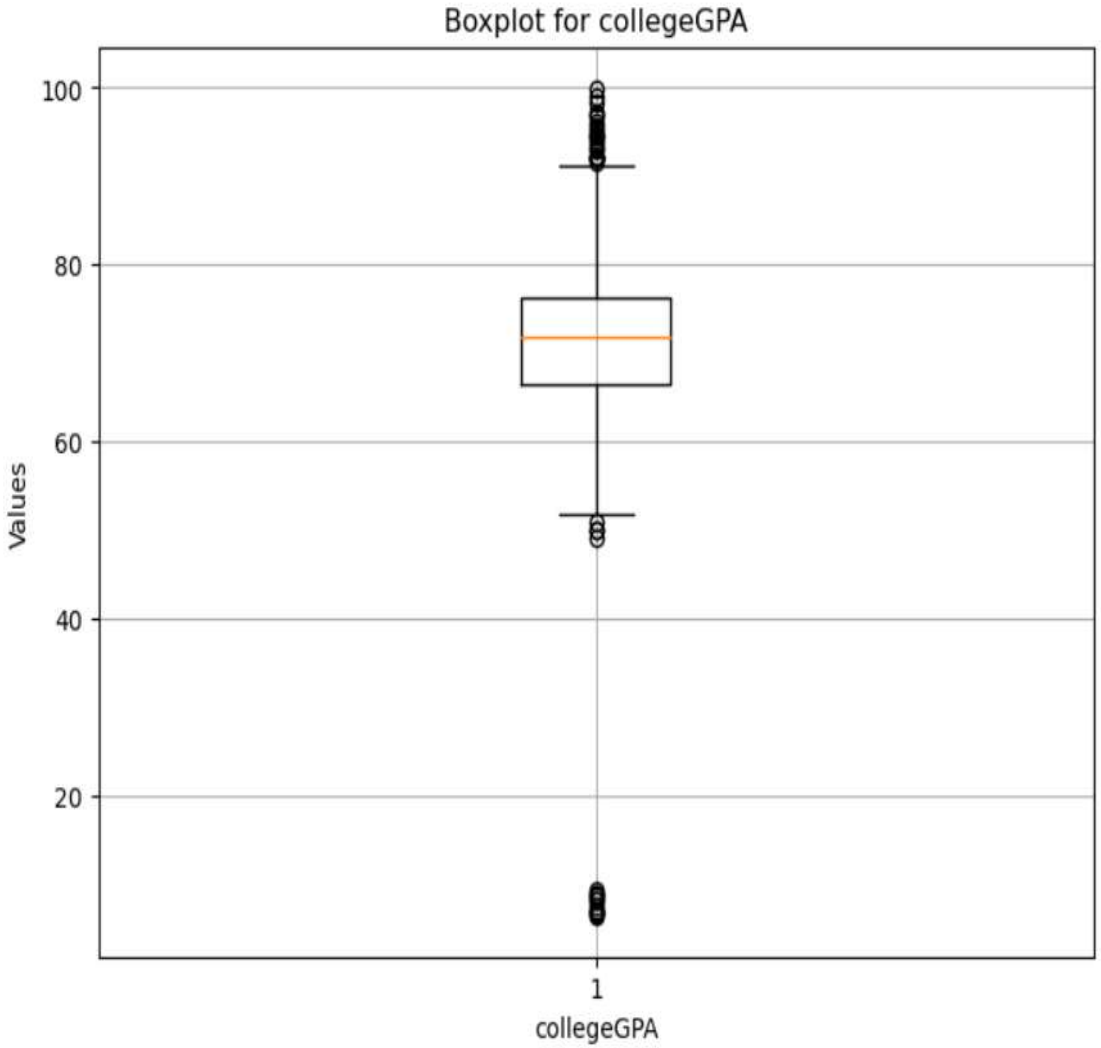




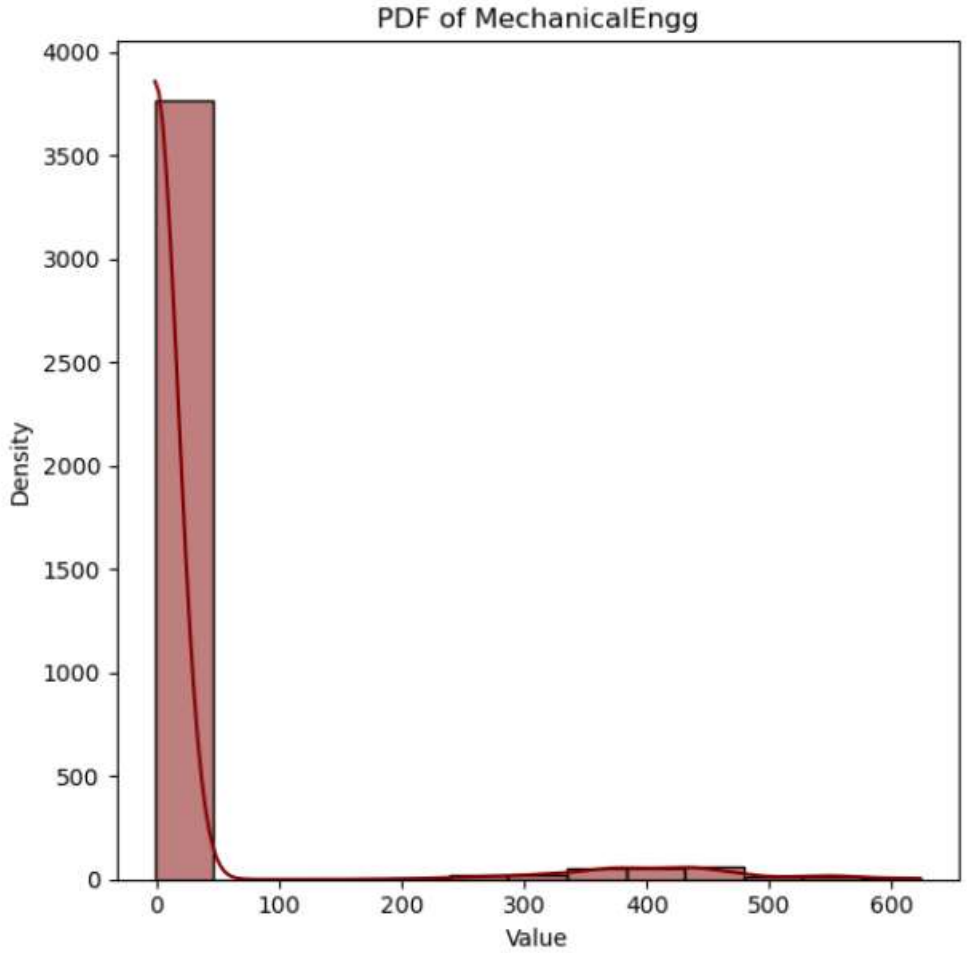


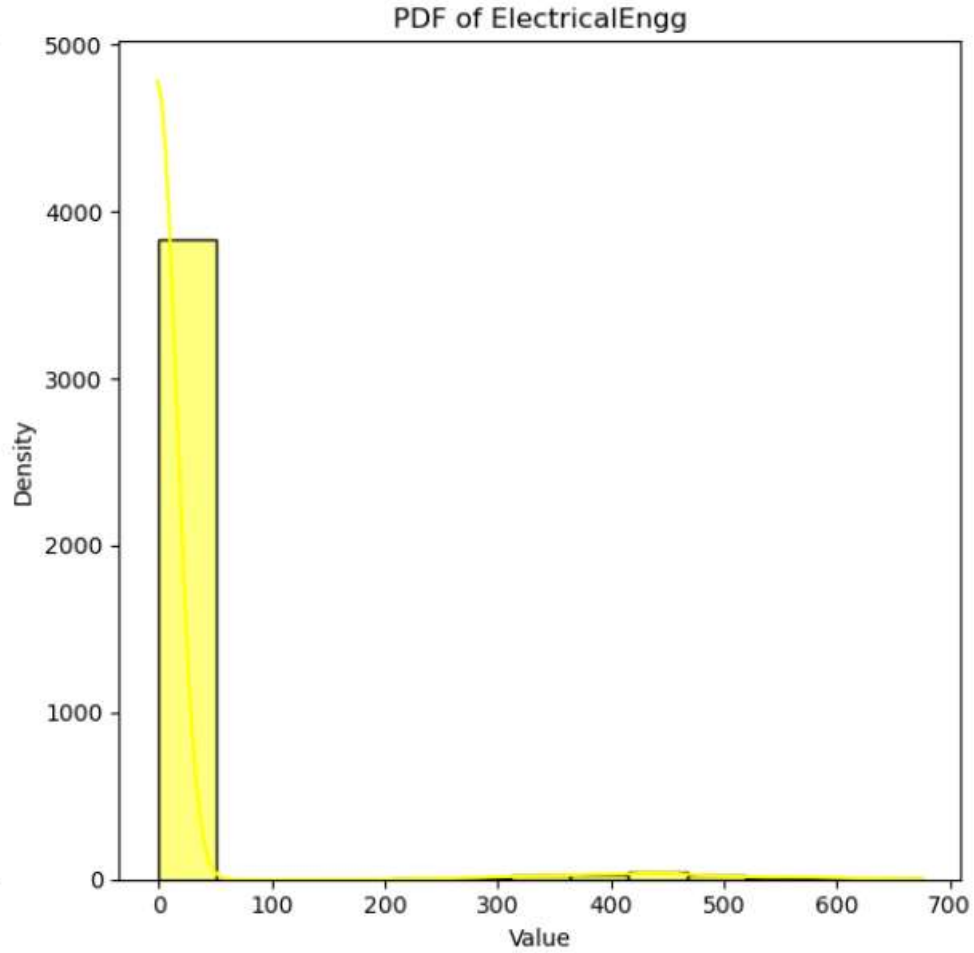
Observation : The histograms represents the 'English','ComputerScience', and 'MechanicalEngg' columns, showcasing the distribution of scores for each subject. Kernel density estimates (KDE) are included to provide smoother representations of the distributions. These visualizations offer insights into the spread and central tendencies of scores in each subject.

INNOMATICS RESEARCH LABS

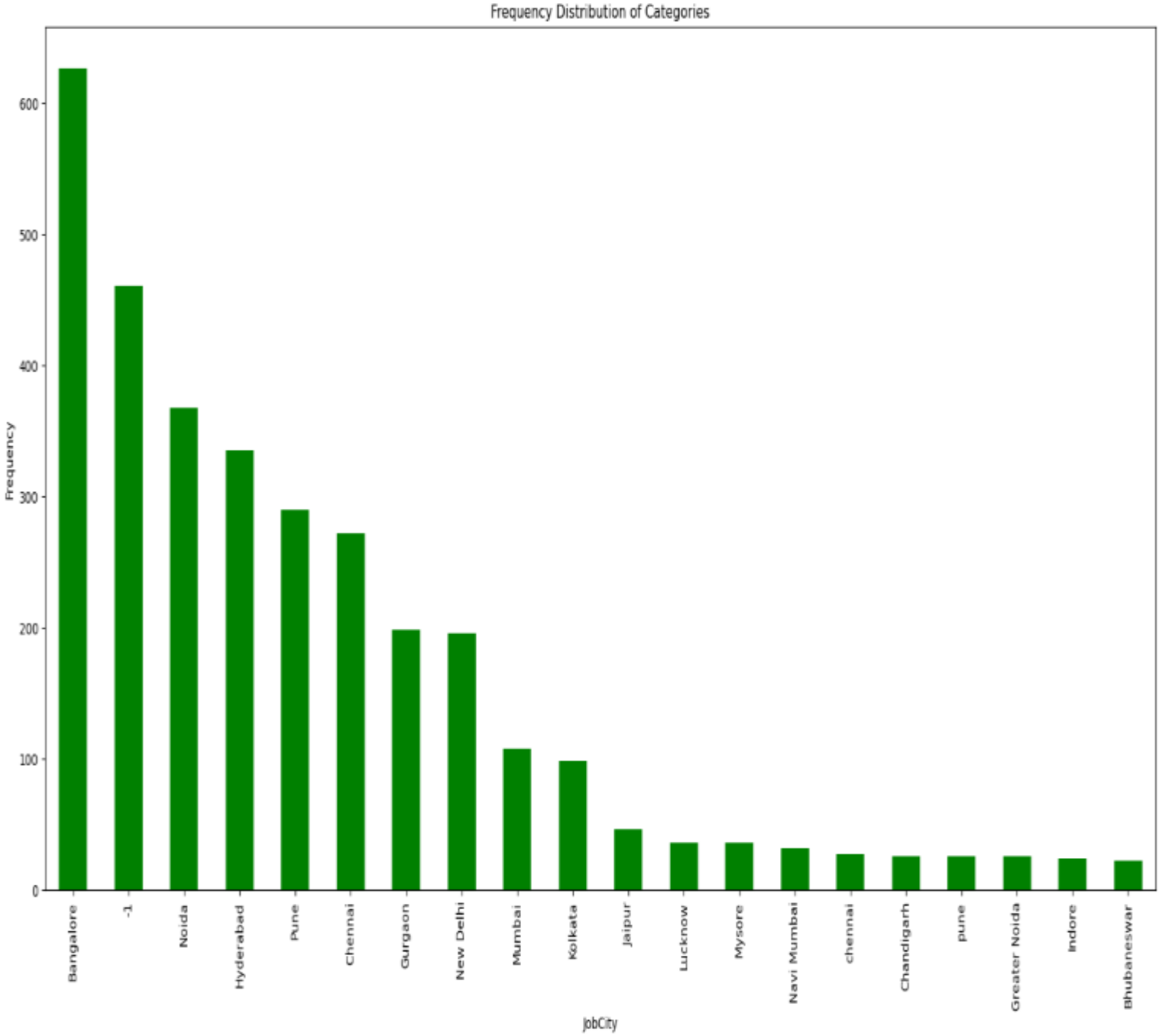


Observation: The boxplot illustrates the distribution of college GPAs, with a median value close to the center of the interquartile range. There are a few outliers on the lower end and some are in upper range, suggesting a relatively normal distribution with some students performing below the median GPA.

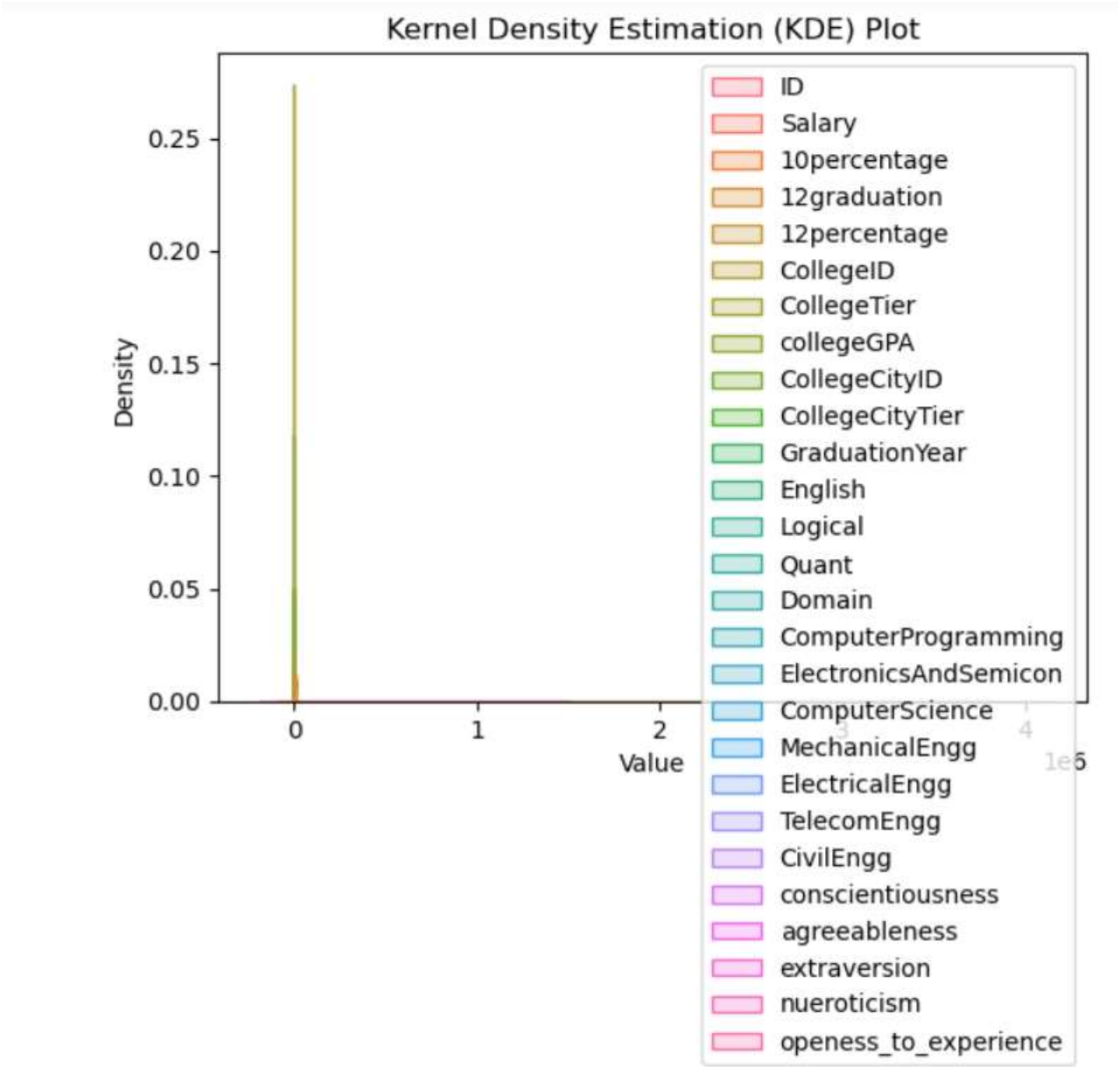




Observation:This visualization depicts the probability density function (PDF) of the MechanicalEngg column and ElectricalEngg, showcasing a distribution with a peak density around a certain value. The kernel density estimate (KDE) overlay provides a smoothed representation of the distribution's shape, aiding in identifying its central tendency and spread.

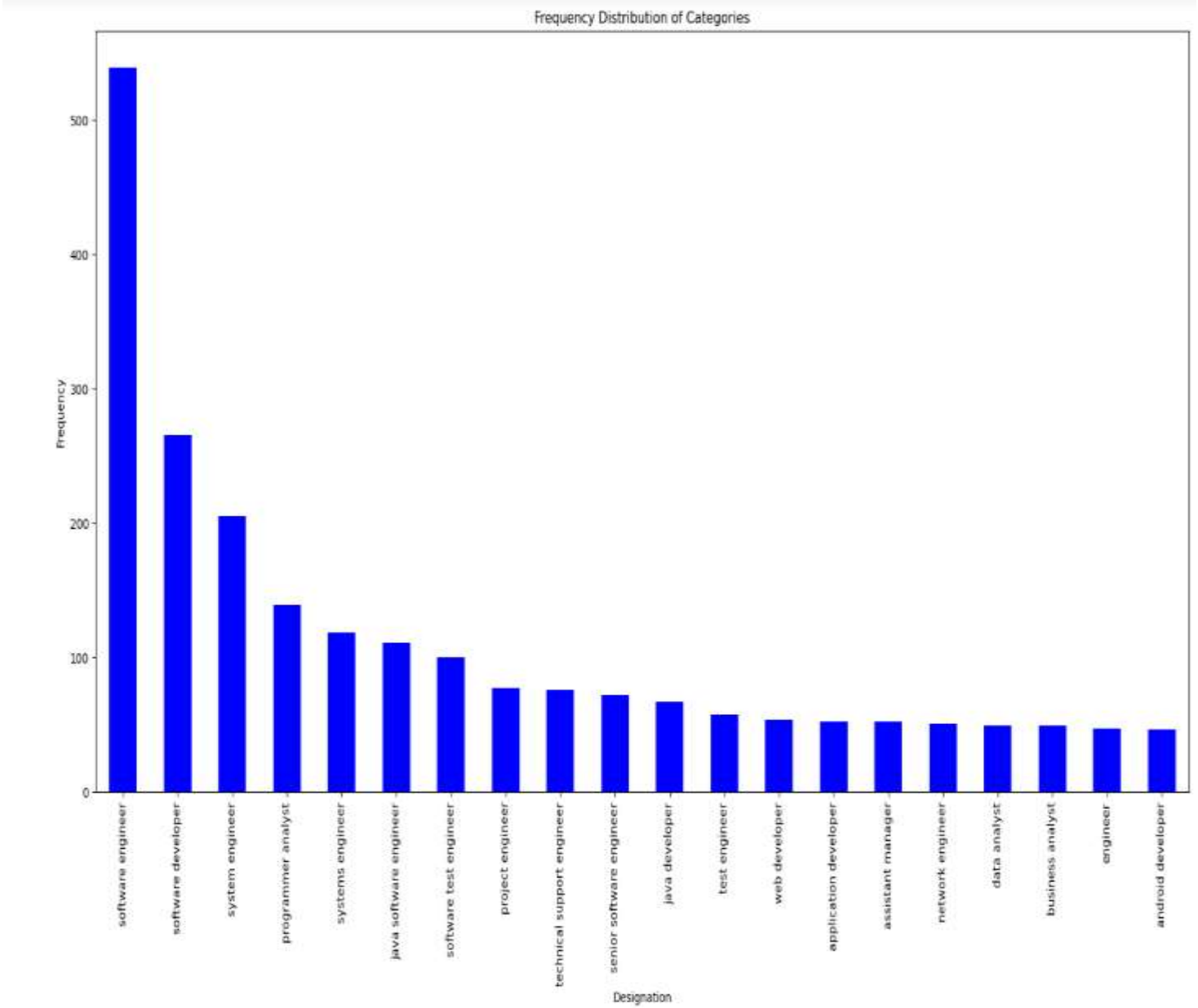


Observation : This bar plot illustrates the frequency distribution of the top 20 designations present in the dataset, showcasing the most common job titles. The ‘ JobCity' category 'Frequency' demonstrates a varied distribution, with certain designations appearing more frequently than others, while the horizontal orientation of the x-axis labels aids readability for a larger number of categories. from above bar we can say that the software engineer has the highest frequency followed by software developer.



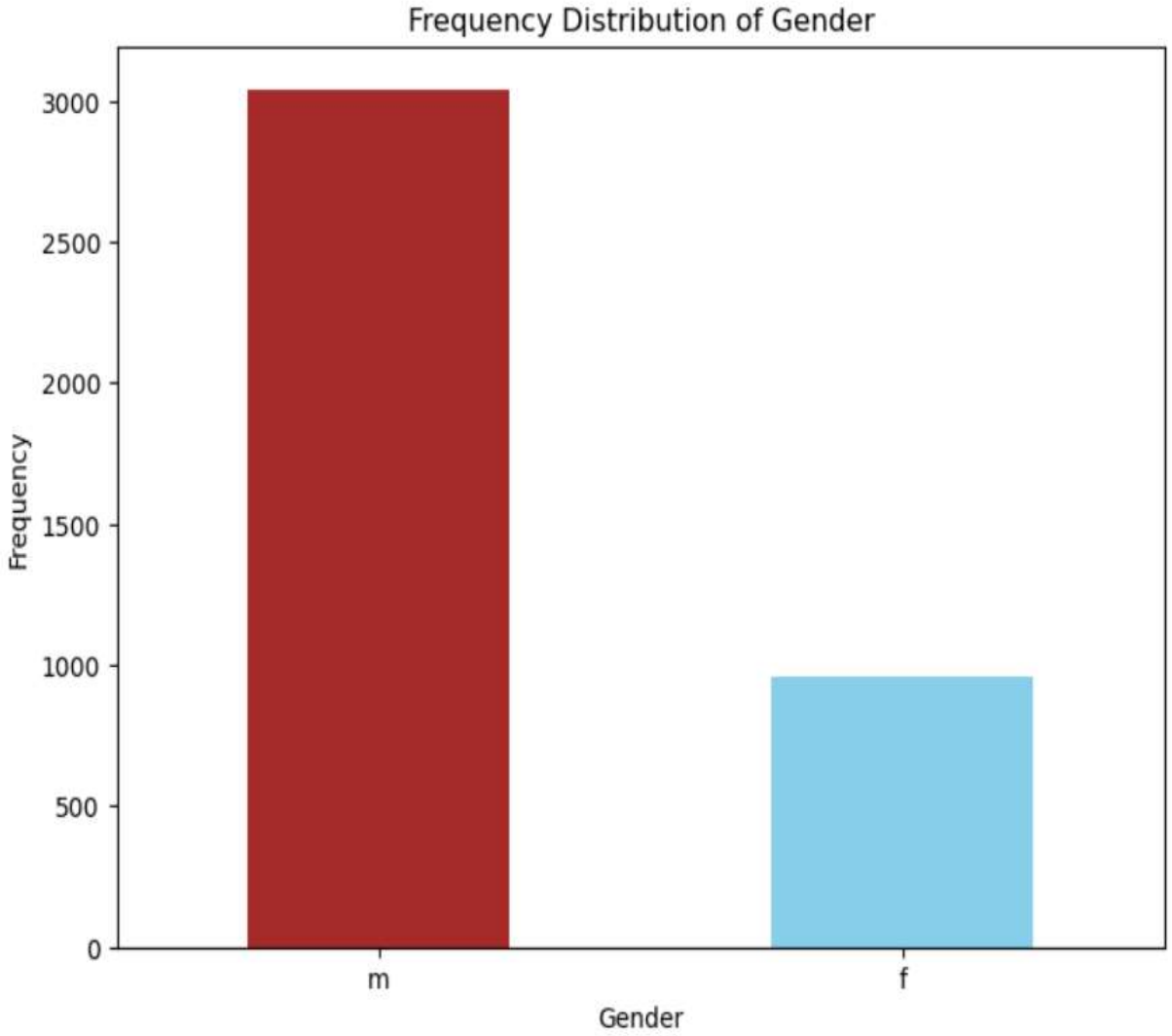
Observation : The Kernel Density Estimation (KDE) plot visually represents the distribution of multiple variables in the DataFrame df, highlighting the density of data points across their respective ranges. The shaded areas emphasize regions with higher data density, providing insights into the overall distribution pattern and concentration of values within the dataset.

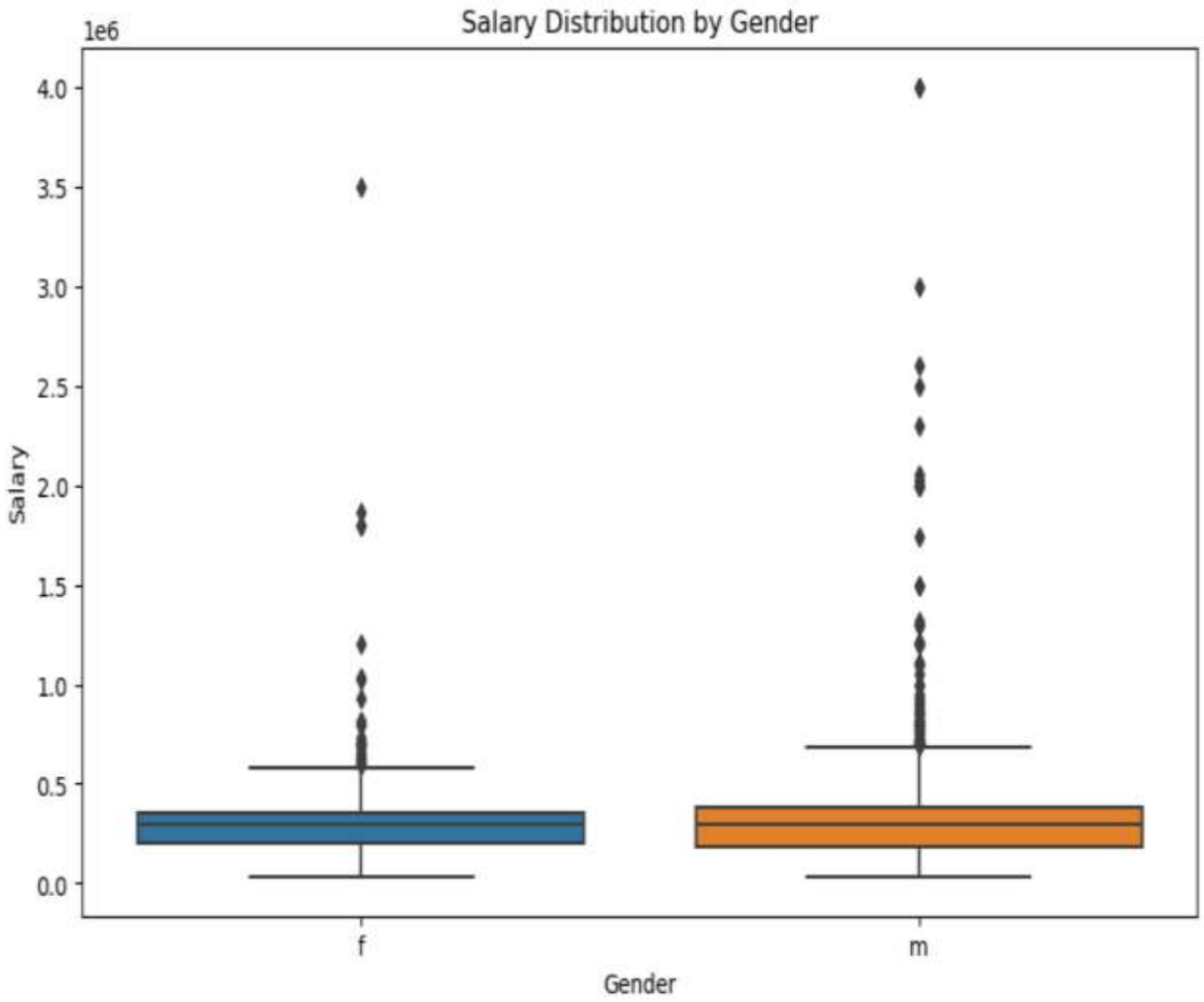
INNOMATICS RESEARCH LABS



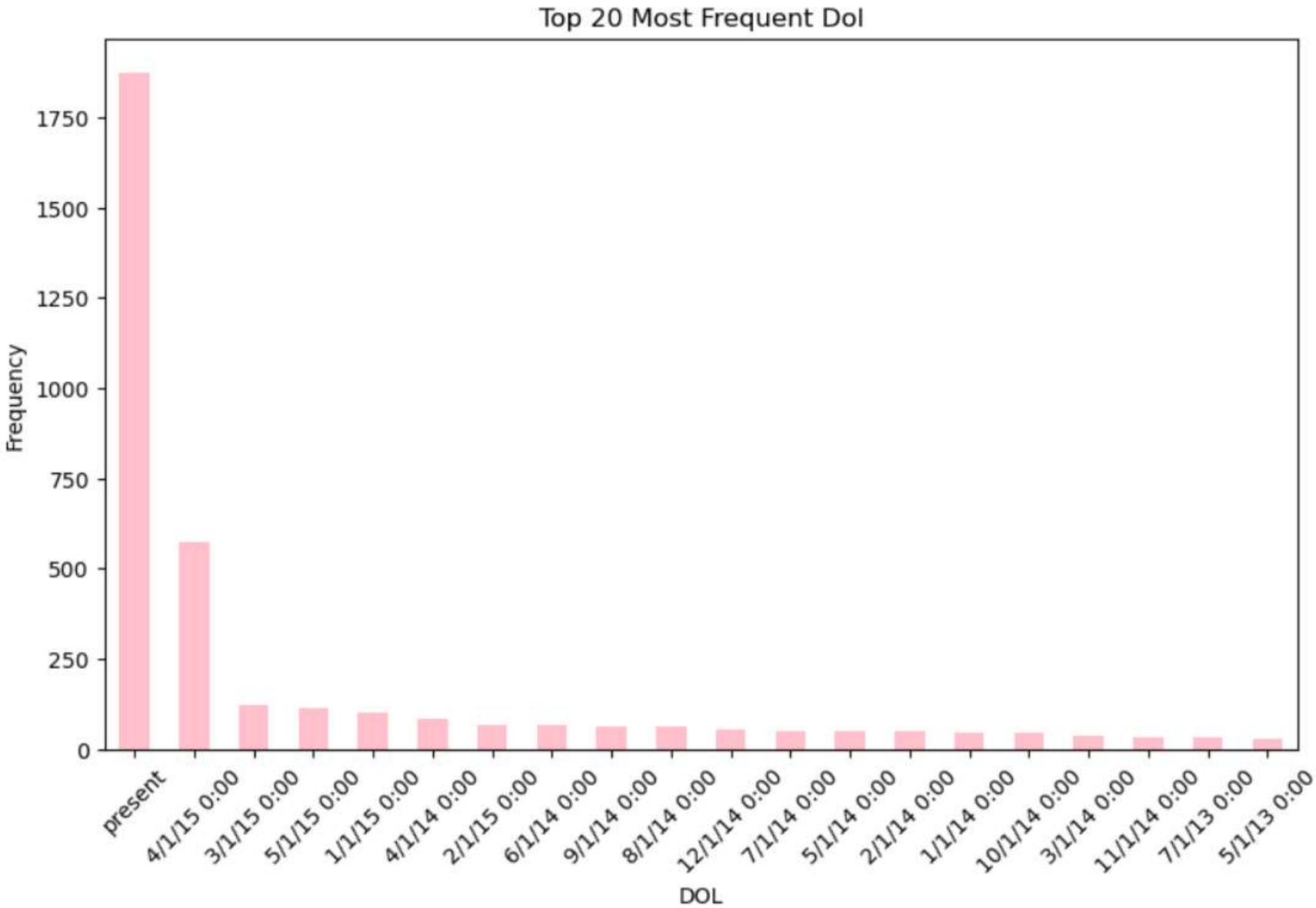
Observation:This bar plot displays the frequency distribution of the top 20 job cities recorded in the dataset, with 'Designation' on the x-axis and the frequency of occurrences on the y-axis. it represents the city Bangalore has the highest frequency

INNOMATICS RESEARCH LABS



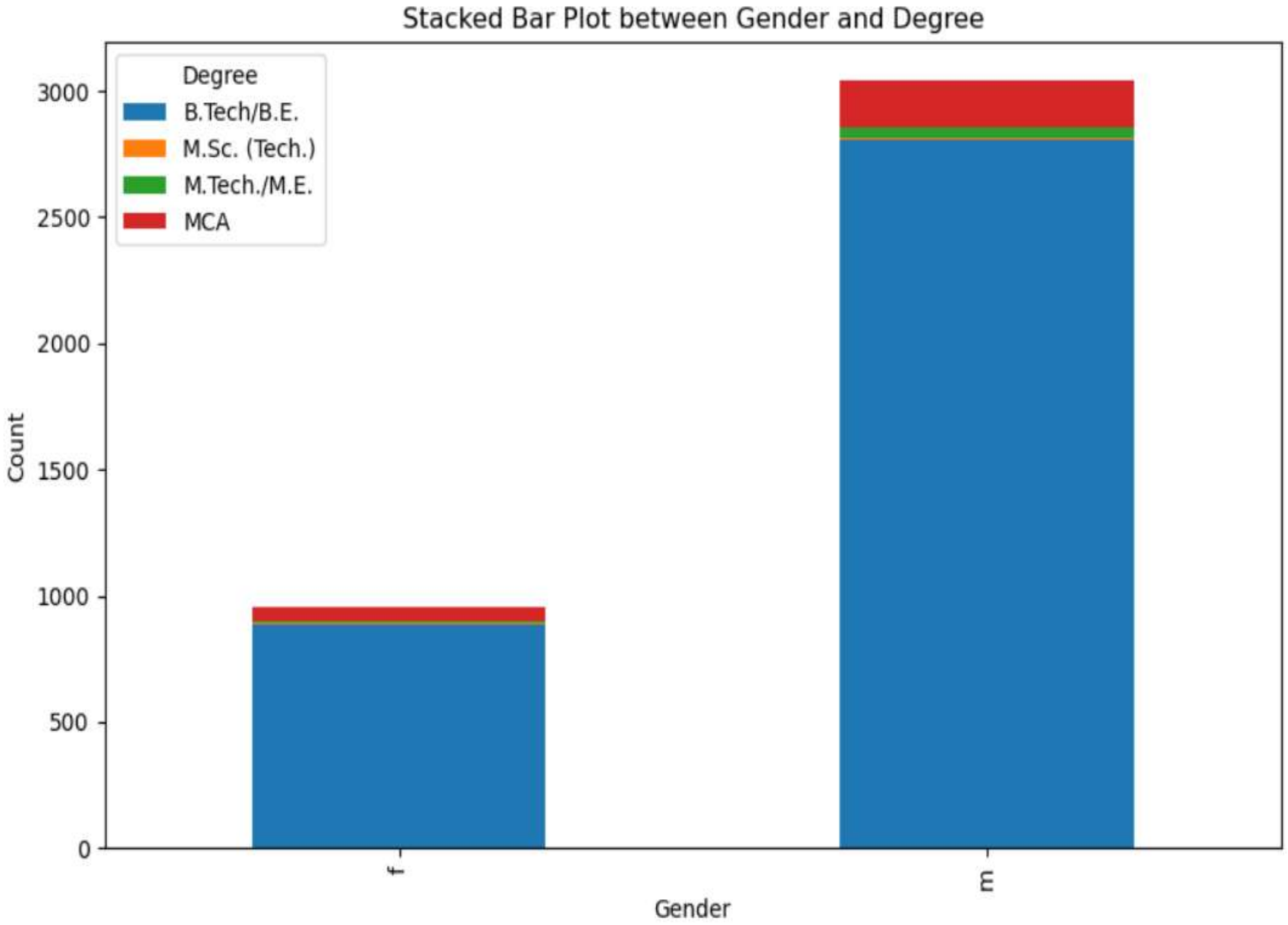


Observation:Look for significant differences in the median salary between genders. A higher or lower median might indicate gender disparities. as we can observes that the gender male as the highest salary outliers as compare to the female

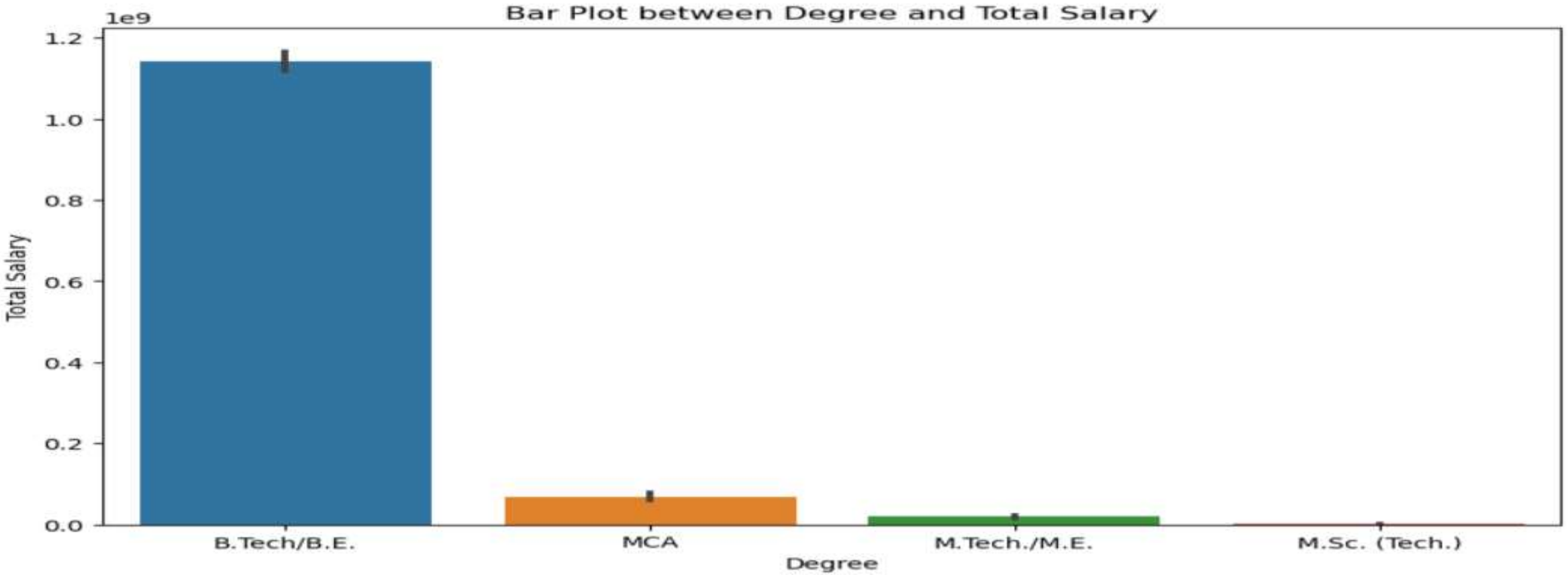


Observation:the above bar plot represents the the top 20 Dates of lossing. as we can see on present on this date has the highest number of lossing similarly 4/1/15 is the second highest date of lossing

INNOMATICS RESEARCH LABS



Observation: The stacked bar plot illustrates the distribution of degree types across genders. It highlights the gender-wise breakdown of degree categories, showcasing the relative frequencies of each degree type among males and females. The plot reveals the gender disparities in educational backgrounds, with certain degrees being more prevalent among one gender compared to the other. here we can observe that number of males has the has degree in B.Tech/BE and the male who are completed MCA has second highest number, in female also most of the female completed degree in B.Tech/B.E and few has done MCA andd very few has completed M.SC.(Tech.)



Observation: The strip plot between gender and specialization displays the distribution of different specializations across genders. It reveals the clustering of various specializations based on gender, highlighting potential gender-based preferences or trends in educational pursuits.

The bar plot illustrates the total salary aggregated by degree types, providing insights into the overall earning potential associated with different degrees. It enables comparison of the total salary earned across various degree categories, aiding in identifying degrees that contribute more significantly to overall earnings.

CONCLUSION:

- In conclusion, the extensive data analysis yields several notable discoveries about the factors impacting pay levels in dataset. While certain criteria such as tenure and college level, have a strong link with compensation, others such as gender and academic performance have no relationship. After removing outliers. Age does not appear to be a determining factor in compensation.

THANK YOU

INNOMATICS
RESEARCH LABS