SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

Samsung Innovation Campus

Artificial Intelligence Course

HR Analysis

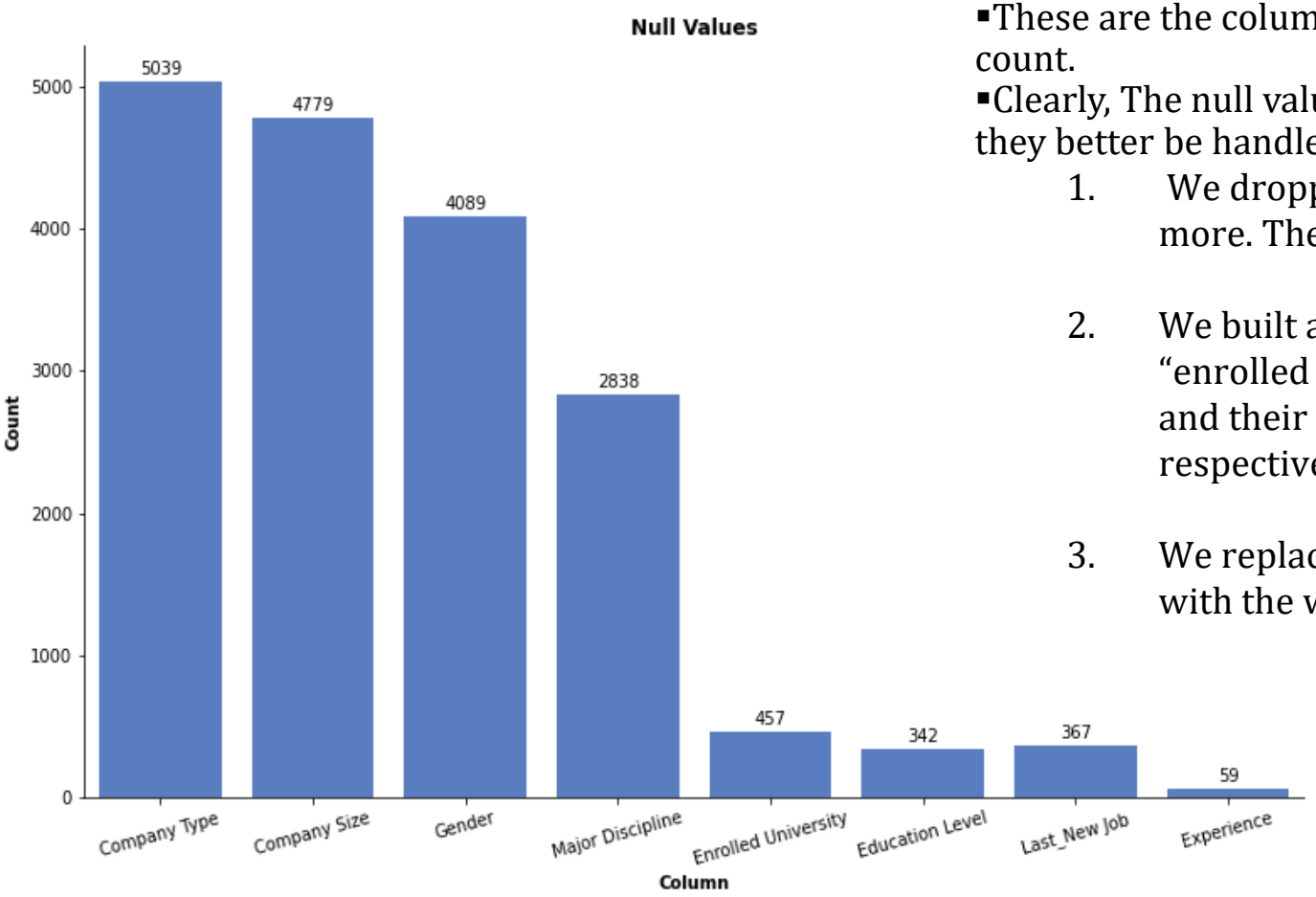
❑Data Description:

- A training institute that conducts training for analytics/ data science wants to expand their business to manpower recruitment (data science only) as well. The company gets a large number of signups for their training. Now, the company wants to connect these enrollees with their clients who are looking to hire employees working in the same domain. Before that, it is important to know which of these candidates are really looking for new employment. They have student information related to demographics, education, experience, and features related to training as well. They have student information related to demographics, education, experience, and features related to training as well.
- To understand the factors that lead a person to look for a job change the agency wants you to design a model that uses the current credentials/demographics/ experience to predict the probability of an enrollee to look for a new job.

Contents

- Clarification on NULL values
- Exploratory Data Analysis
    1. Univariate Charts
    2. Bivariate Charts
    3. Multivariate Chart
- Machine Learning Models
- Conclusion
- Problems we faced in our investigation

➢Dealing with NULL values



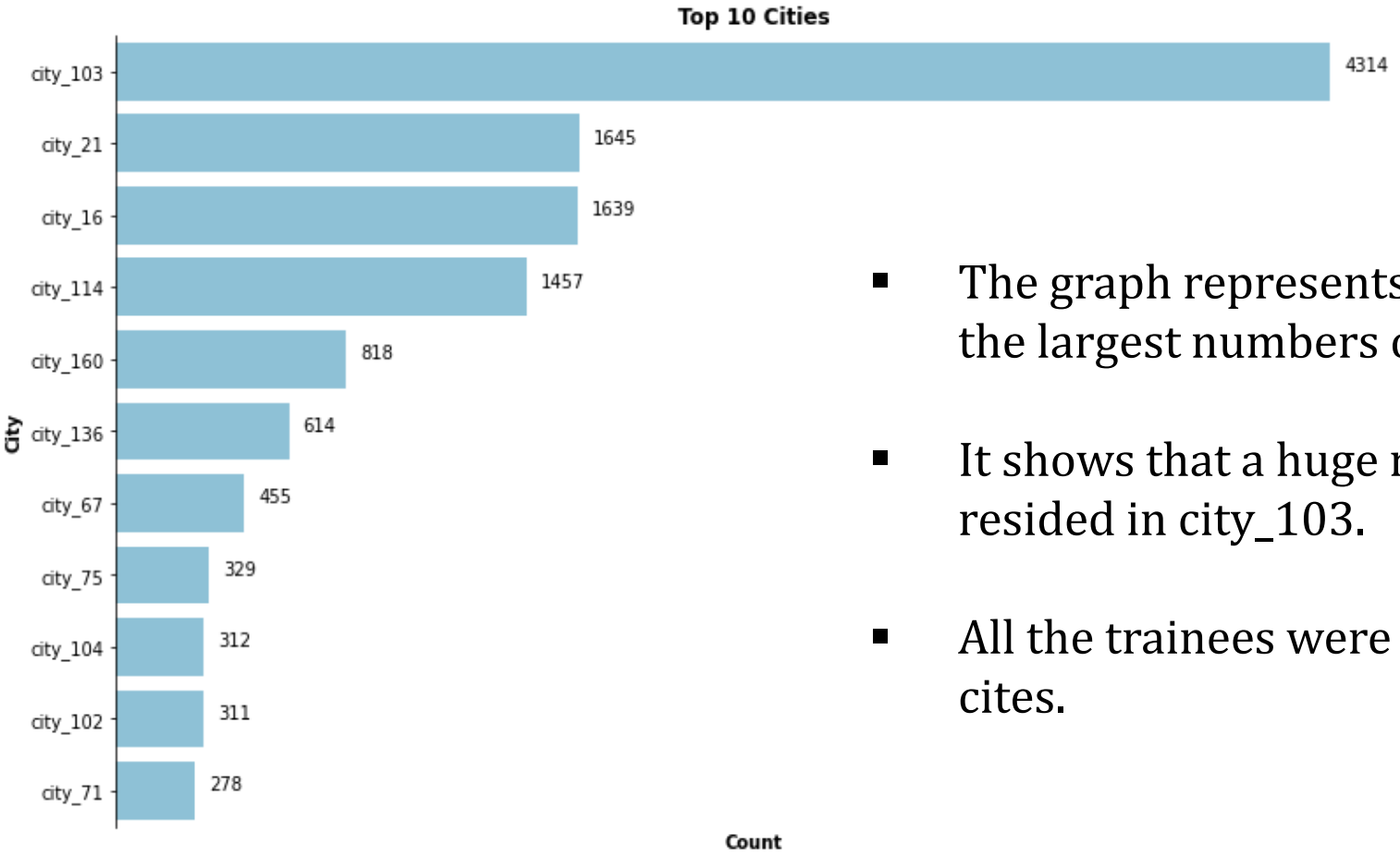
- These are the columns containing null values and their count.
- Clearly, The null values are too many to be dropped, so they better be handled.

1. We dropped the rows containing four nulls or more. They only formed 1.2% of our data.

2. We built a classification model on the columns “enrolled university” and “major discipline” and their accuracy were 85% and 88% respectively.

3. We replaced the remaining of the null values with the word “Unknown”.

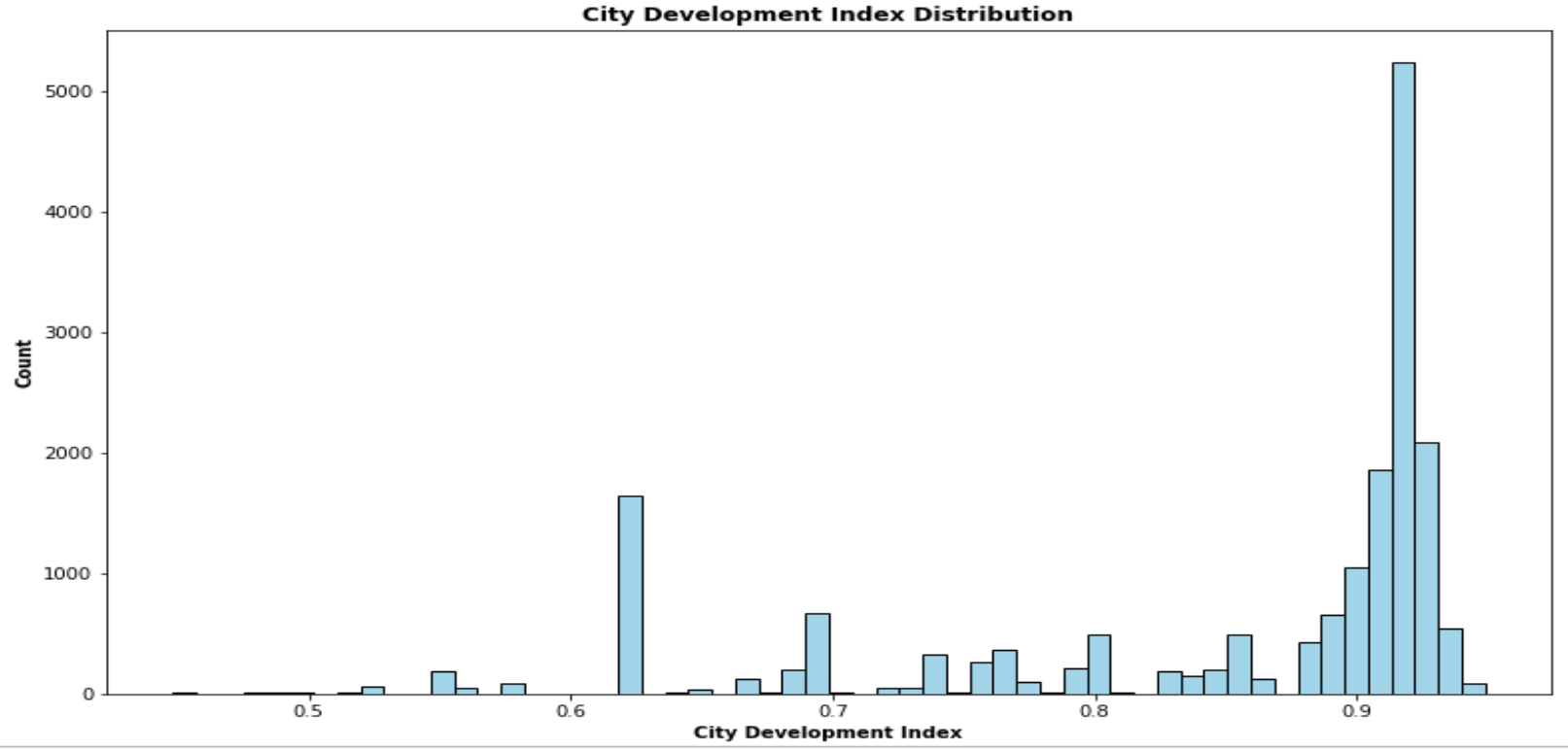
Exploratory Data Analysis
Univariate Charts



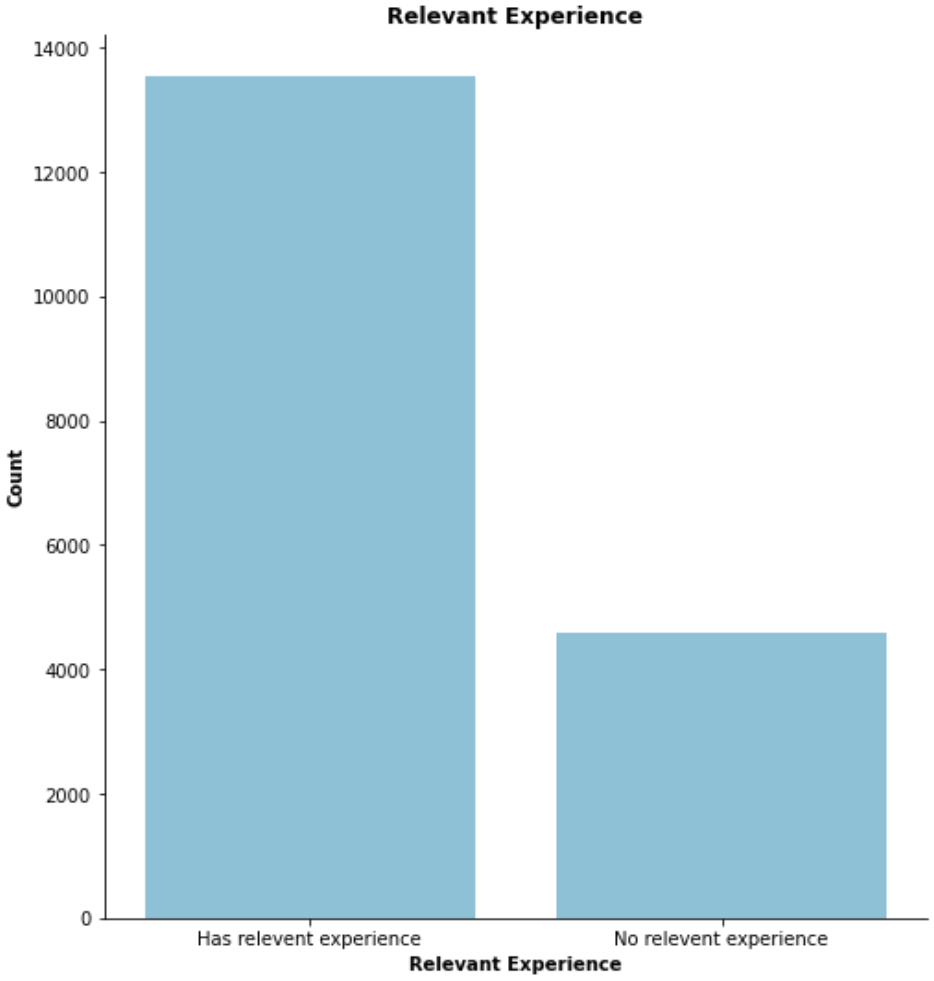
- The graph represents the top 10 cities with the largest numbers of trainees.
- It shows that a huge number of trainees resided in city_103.
- All the trainees were distributed among 123 cites.

Exploratory Data Analysis
Univariate Charts

- The graph represents the distribution of the city development index of the trainees across our data.
- It shows that most trainees cities had development index around 0.9.

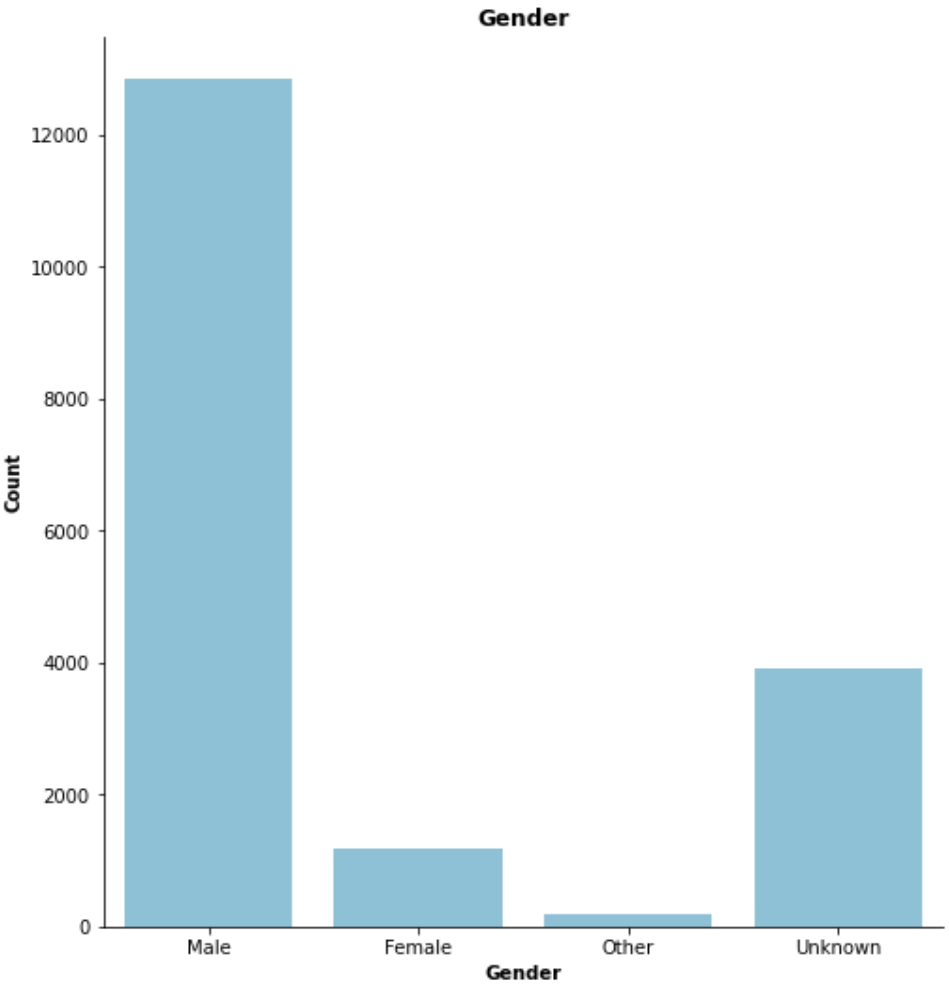


Exploratory Data Analysis
Univariate Charts



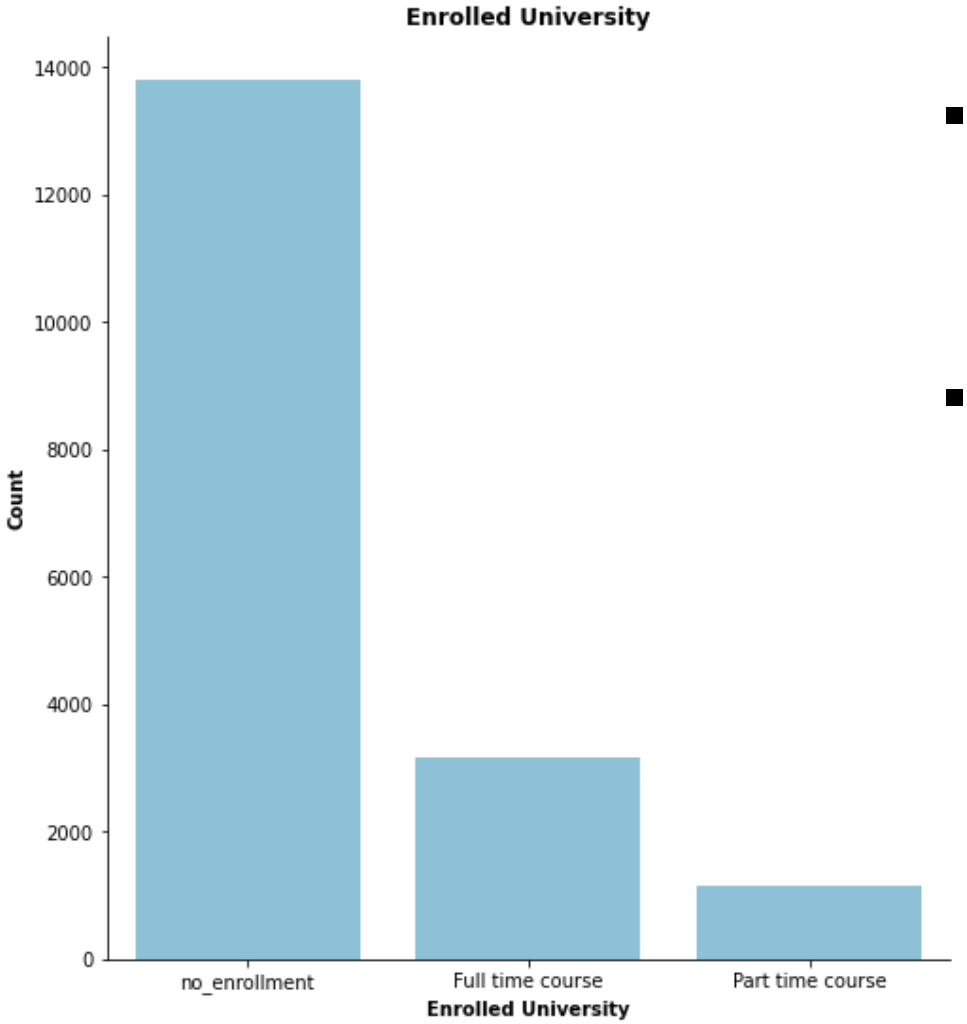
- The graph represents the number of trainees with relevant experience to the field of data science.
- It shows that most trainees accepted to the program has relevant experience to data science.

Exploratory Data Analysis
Univariate Charts



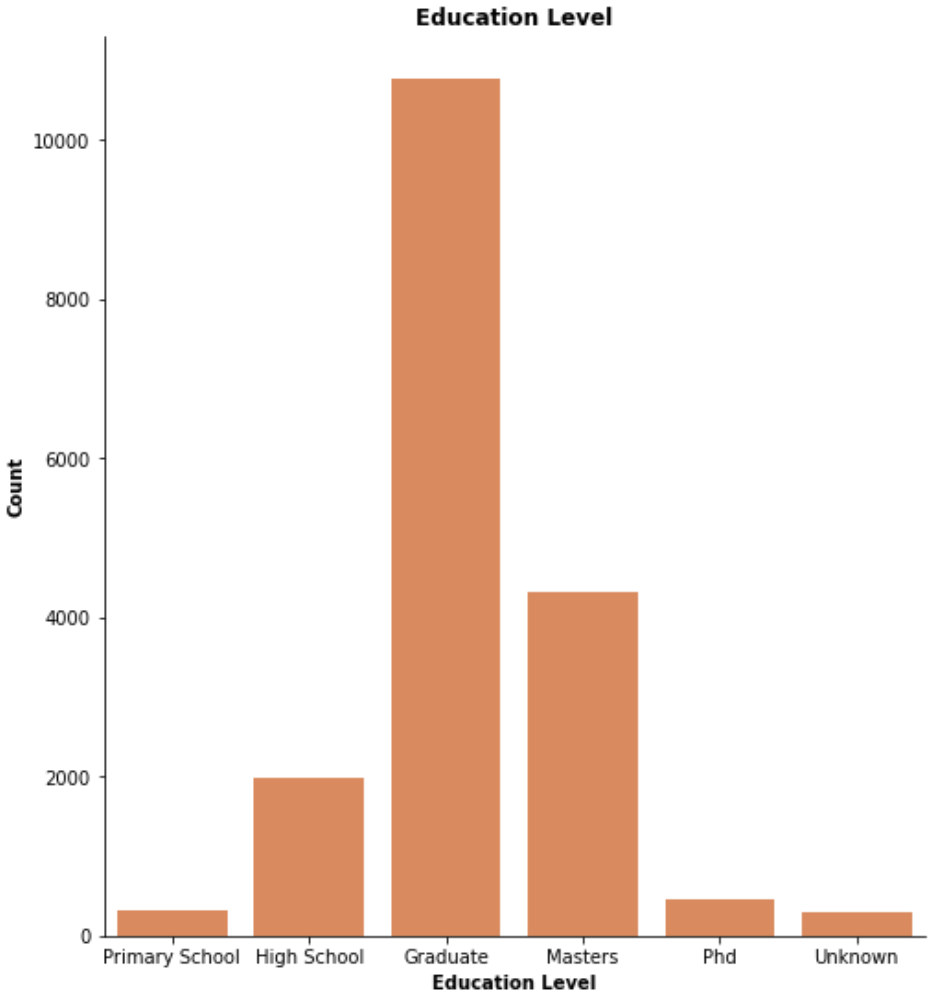
- The graph represents the number of trainees according to their gender.
- It shows that males are by far the largest species accepted to the program.

Exploratory Data Analysis
Univariate Charts



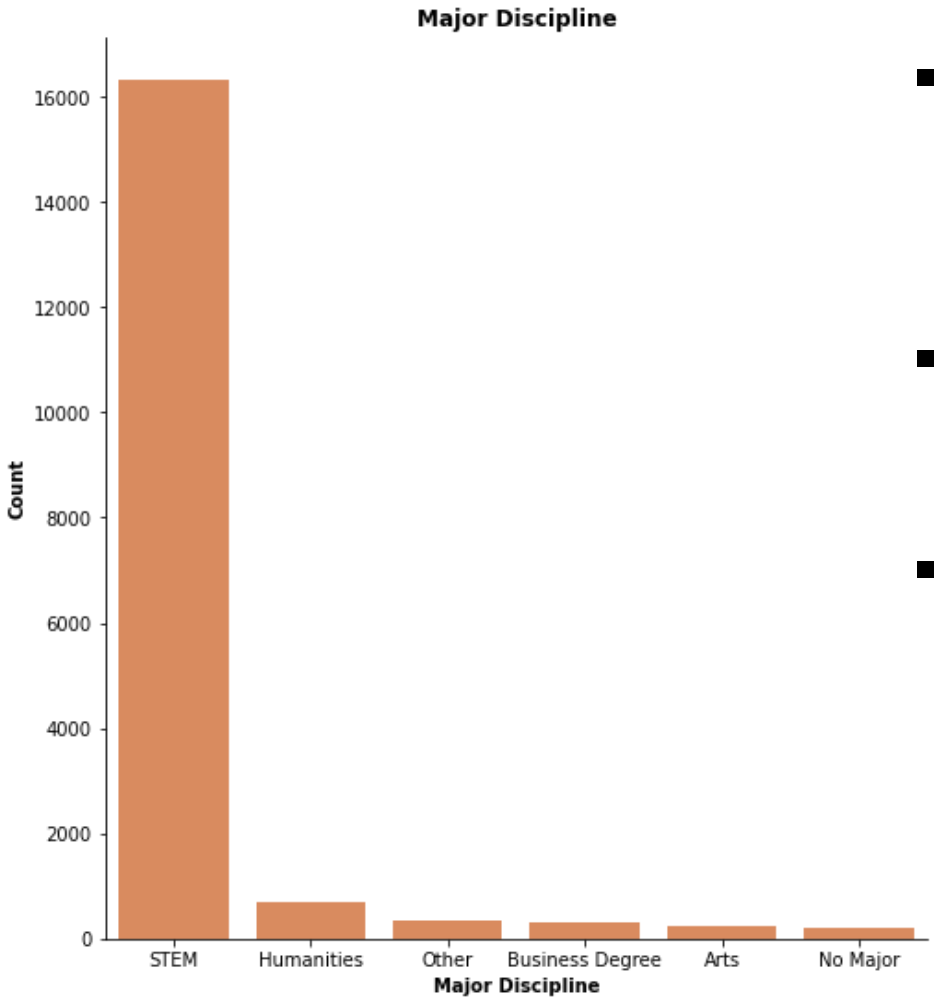
- The graph represents the number of trainees according to their university course enrollment status.
- Nearly 60% of the trainees have no university course enrollment.

Exploratory Data Analysis
Univariate Charts



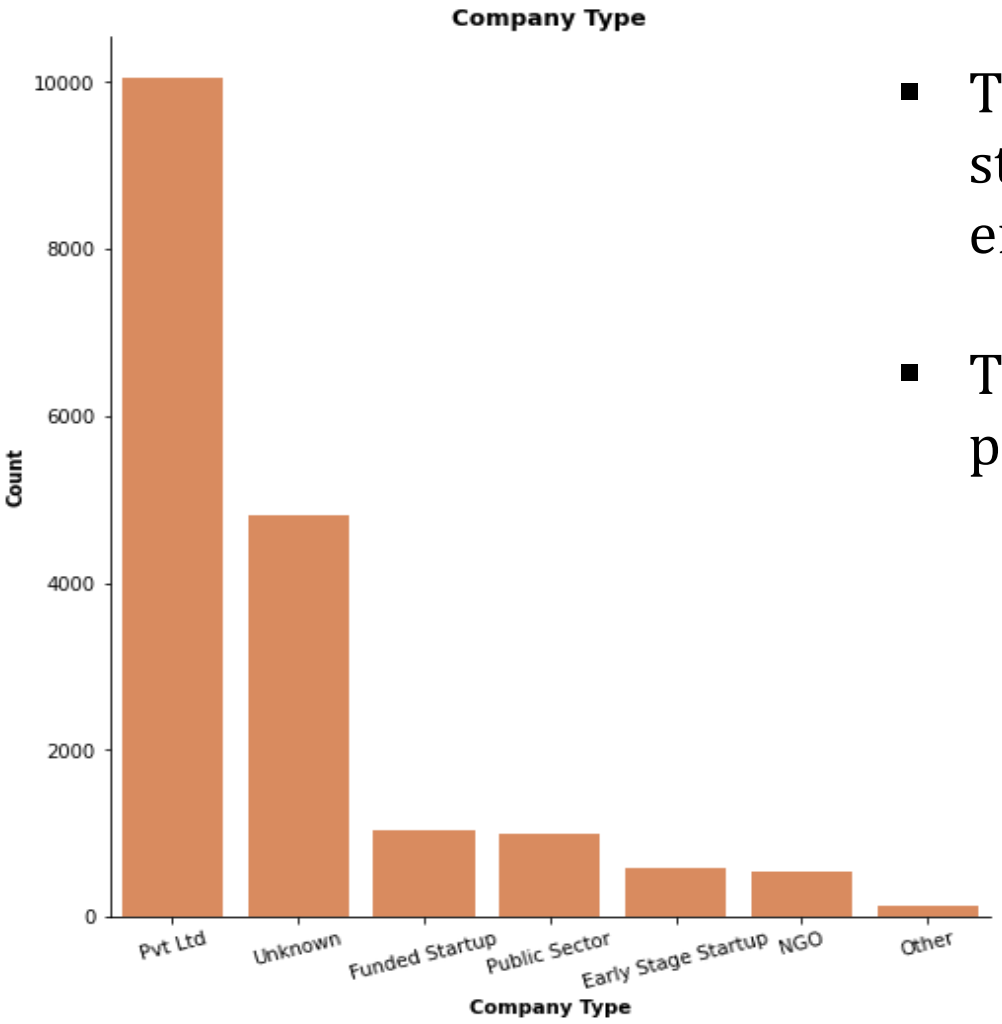
- The graph represents the counts of trainees with respond to their education level.
- Clearly, the graduate trainees were the majority in the training program and by a long shot.

Exploratory Data Analysis
Univariate Charts



- This graph represents the counts of trainees according to their major discipline.
- Clearly, 90% of trainees are STEM trainees.
- This feature shows the imbalance of the data.

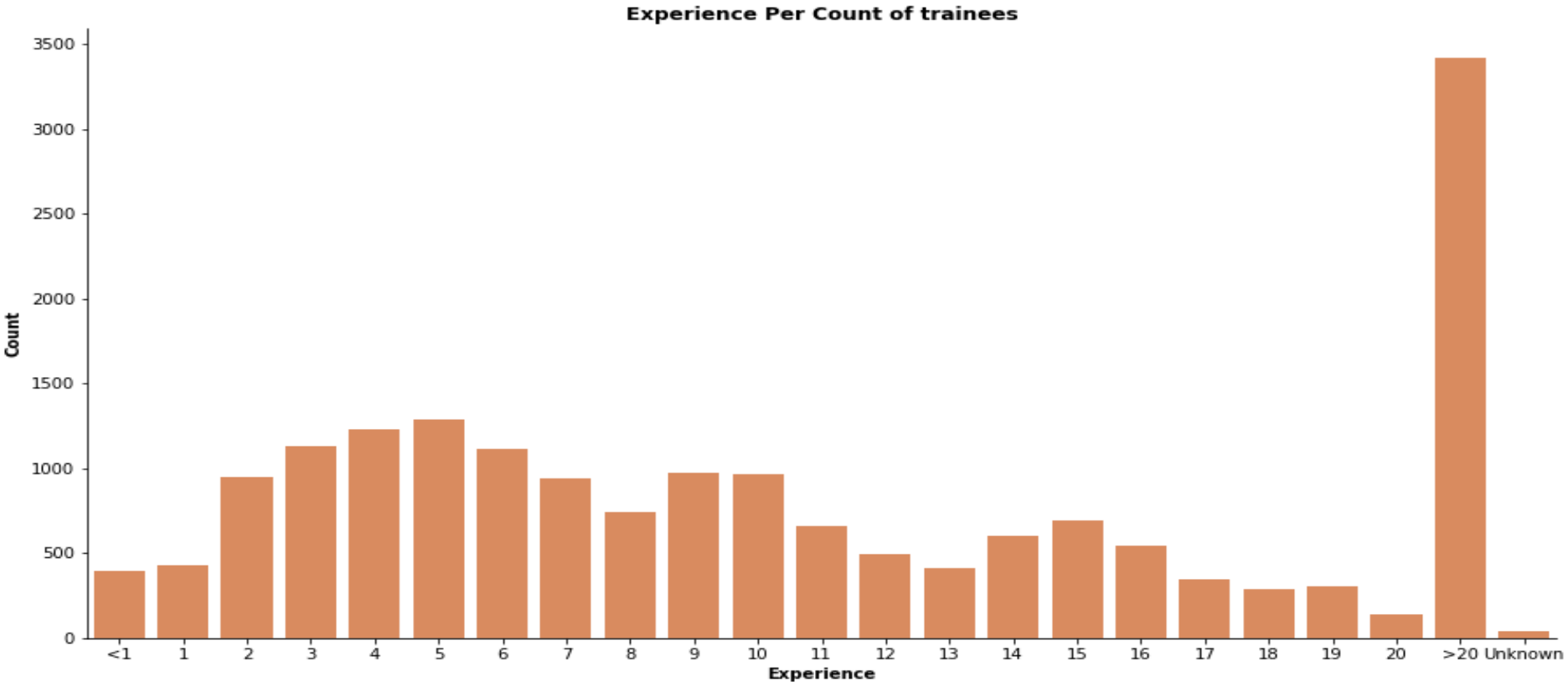
Exploratory Data Analysis
Univariate Charts



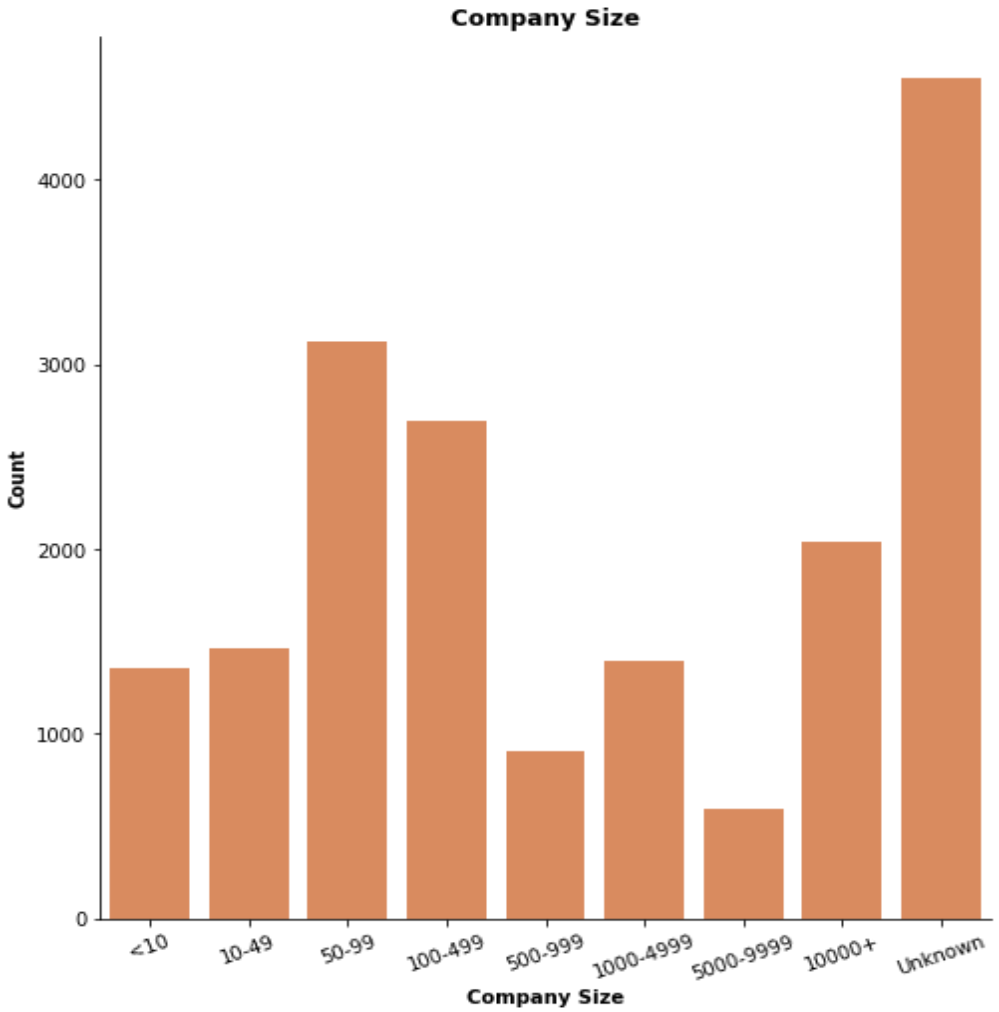
- This graph represents the number of students with respect to their current employer.
- The majority of trainees work for private limited companies.

➢ Exploratory Data Analysis
Univariate Charts

- The graph represents the counts of trainees by their years of experience.
- Clearly, The largest number of trainees have more than 20 years of experience.

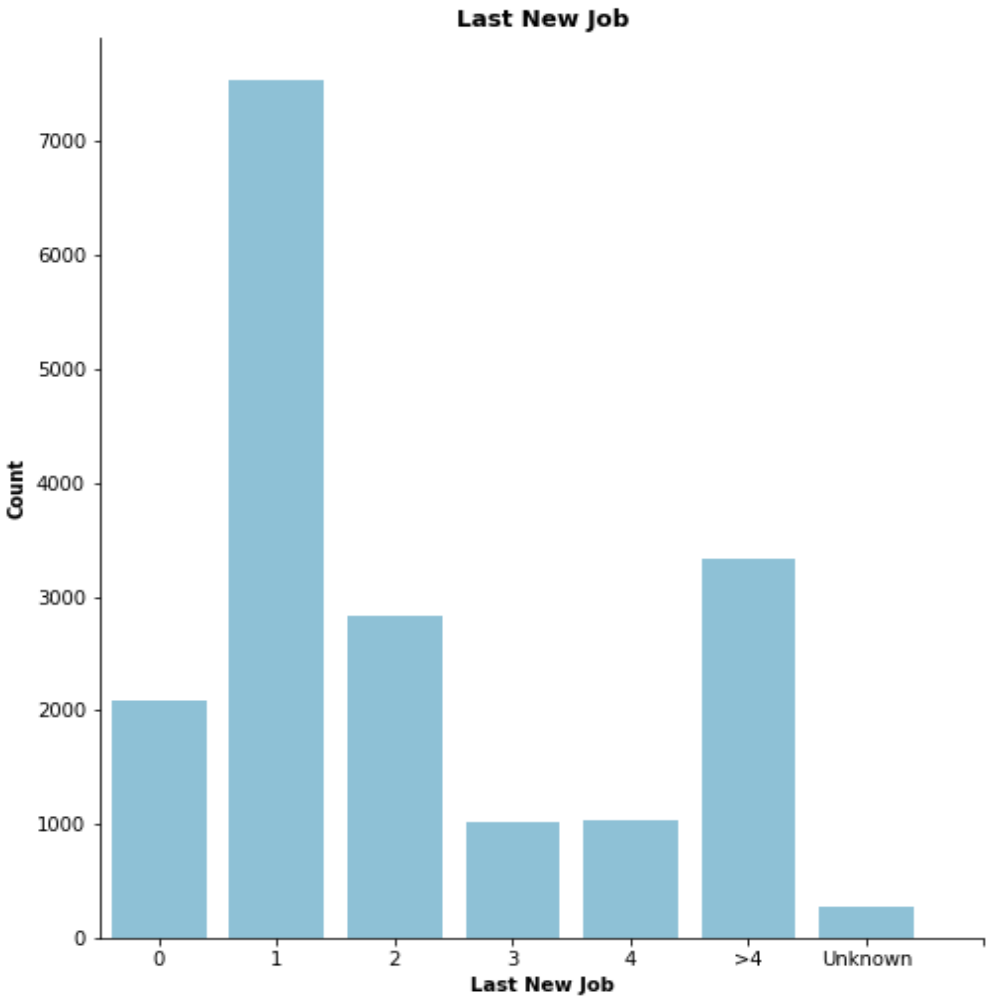


Exploratory Data Analysis
Univariate Charts



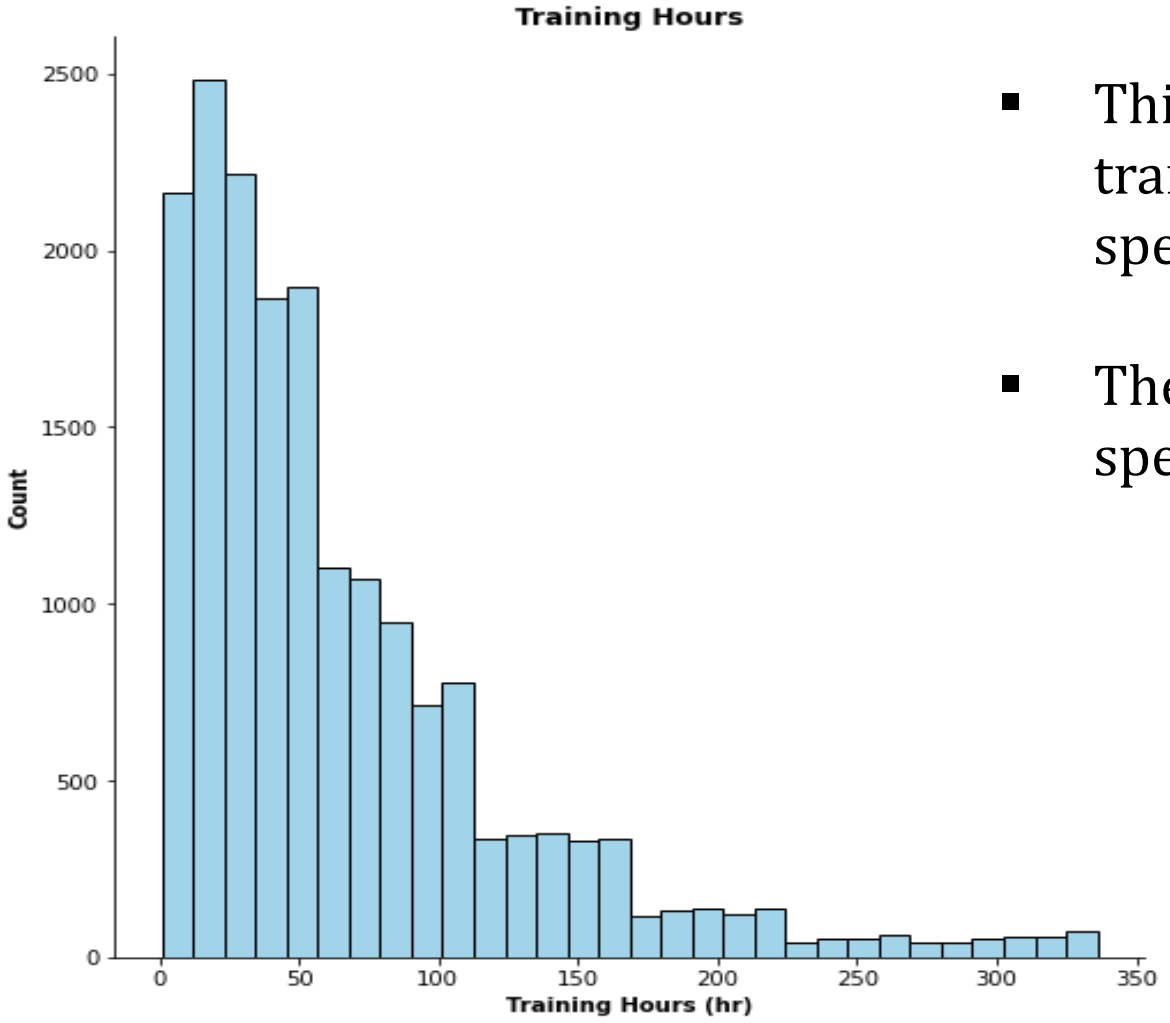
- This graph represents the number of employees of the current company the trainee works for.
- This feature contains a lot of missing data.
- Large number of trainees work for companies with employees ranging from 50 to 500.

Exploratory Data Analysis
Univariate Charts



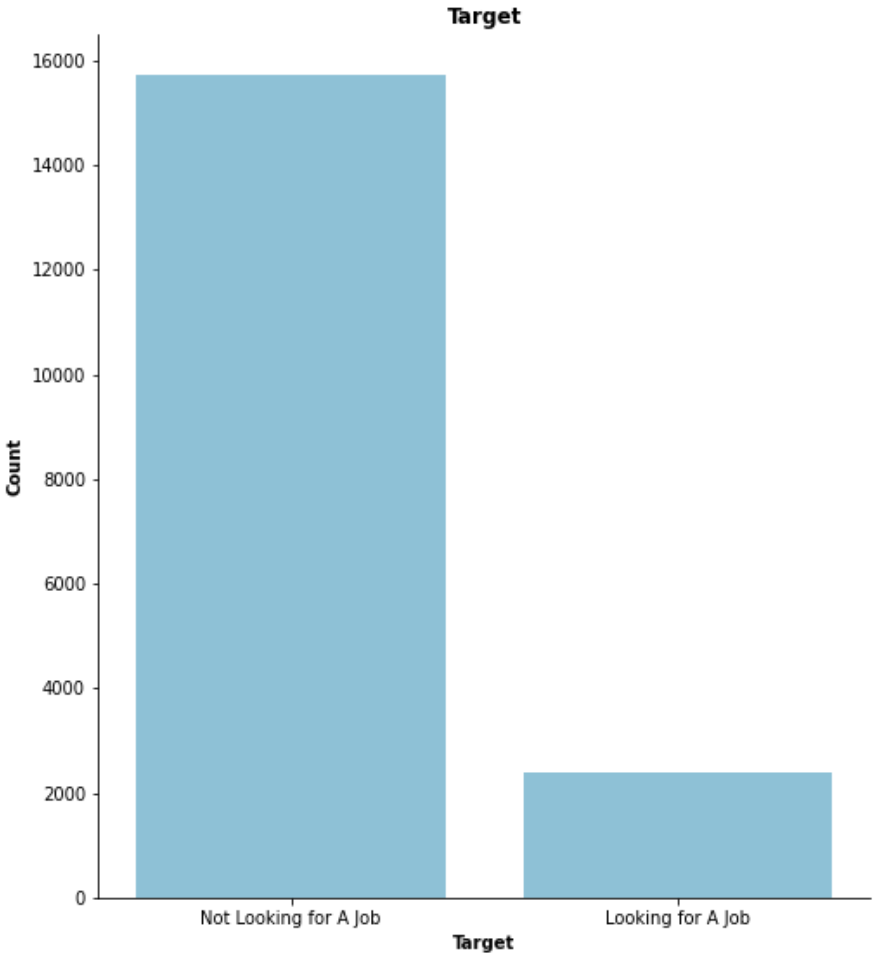
- This graph shows the difference in years between the trainee current job and previous job.
- Most of them didn't spend more than one year before getting their current job.

➢ Exploratory Data Analysis
Univariate Charts



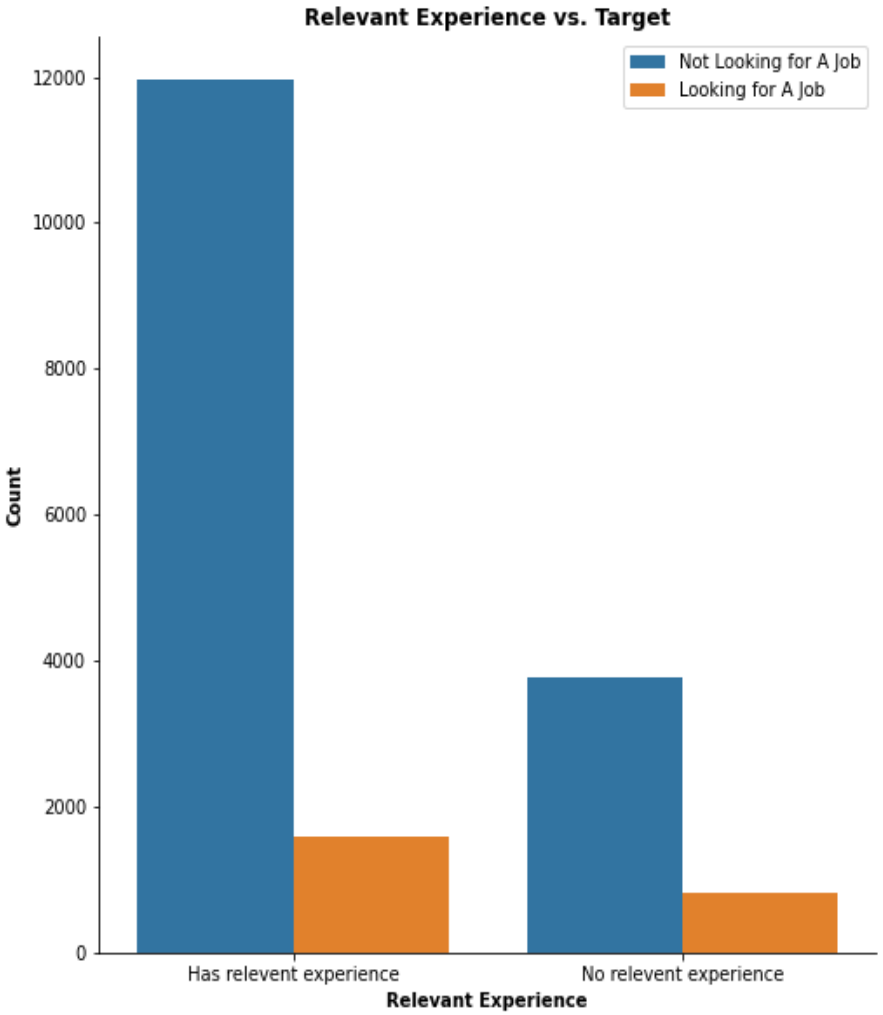
- This graph shows the count of trainees over the training hours spent.
- The majority of students did not spend more 100 hours of training.

Exploratory Data Analysis
Univariate Charts



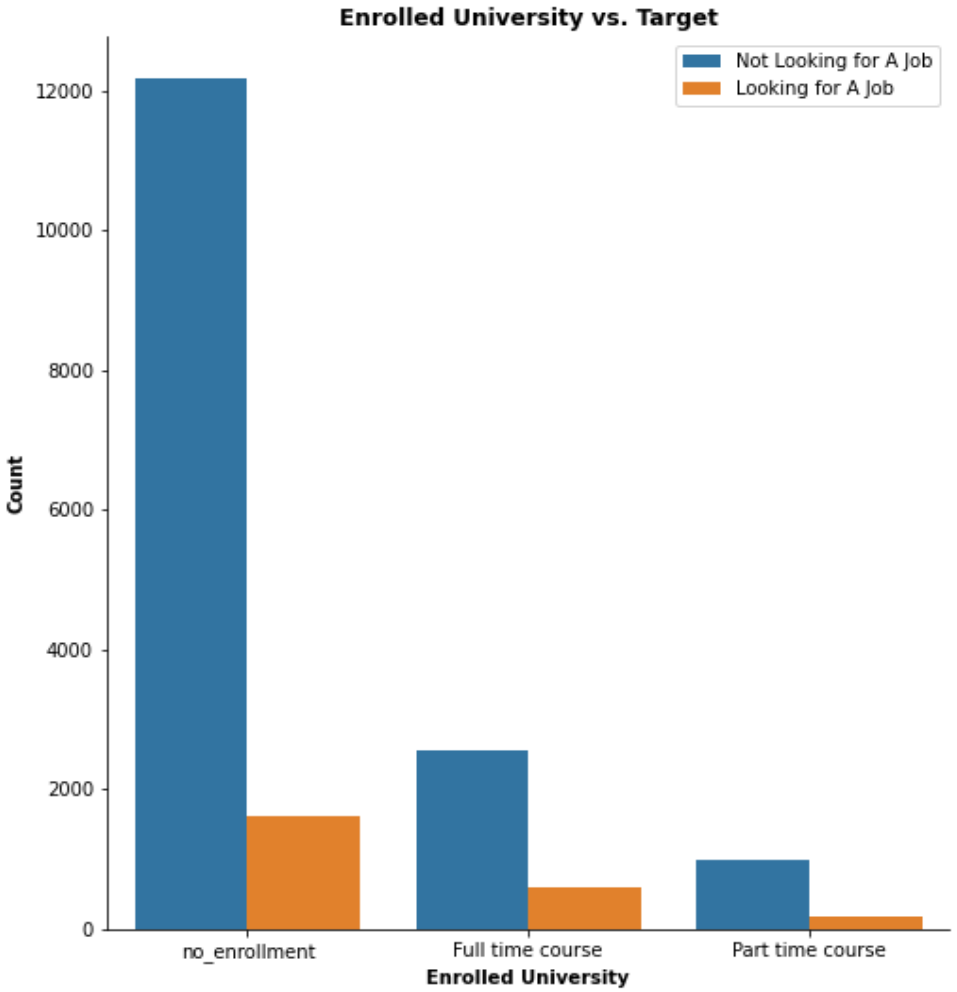
- This graph defines our main of interest (target) of whether the trainee looked for another job or not.
- Clearly, most of the trainees applied for the training program, were not looking for a job.
- The difference between them is quite large and makes us understand why the company asked us to build them a prediction model.

Exploratory Data Analysis
Bivariate Charts



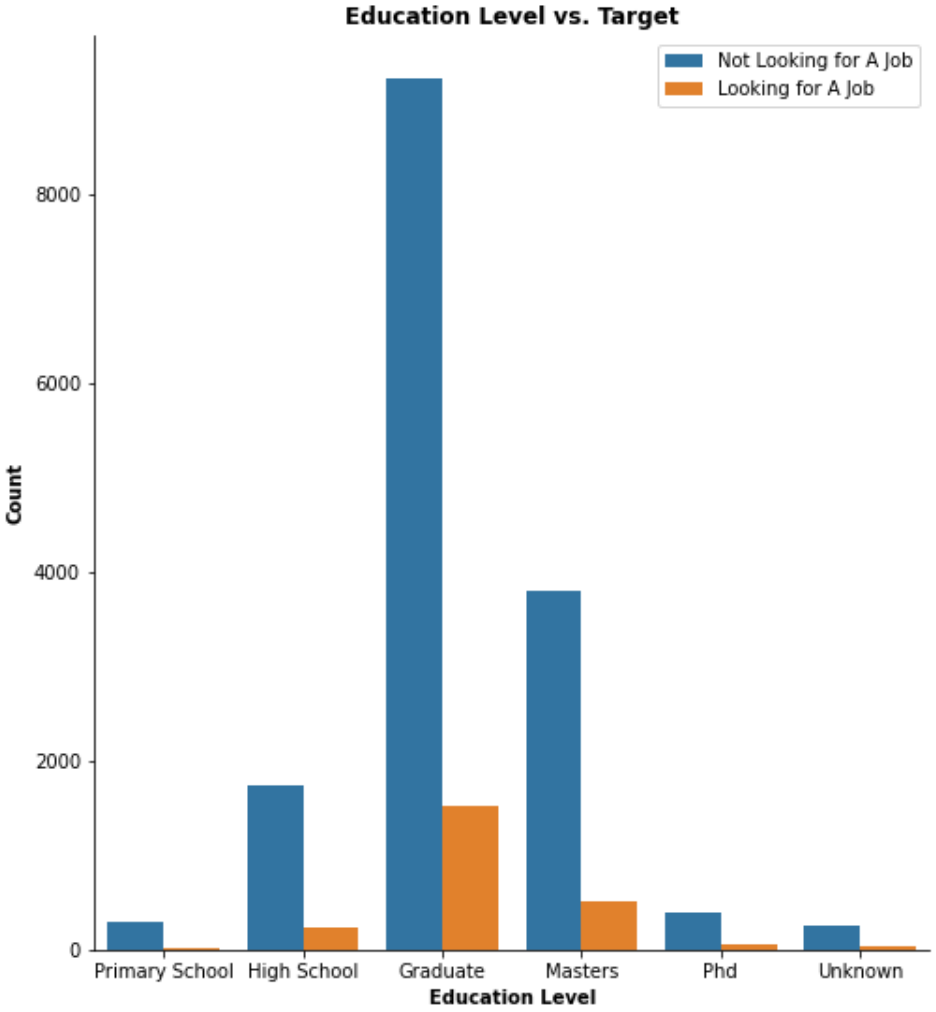
- Clearly, many of the trainees with relevant experience to data science applied to the training program not looking for a job.
- Maybe because they already had a job in data science .
- Maybe they just applied to improve their technical skills.
- Maybe they applied for a chance to get a better job, but not really seeking a new job.

Exploratory Data Analysis
Bivariate Charts



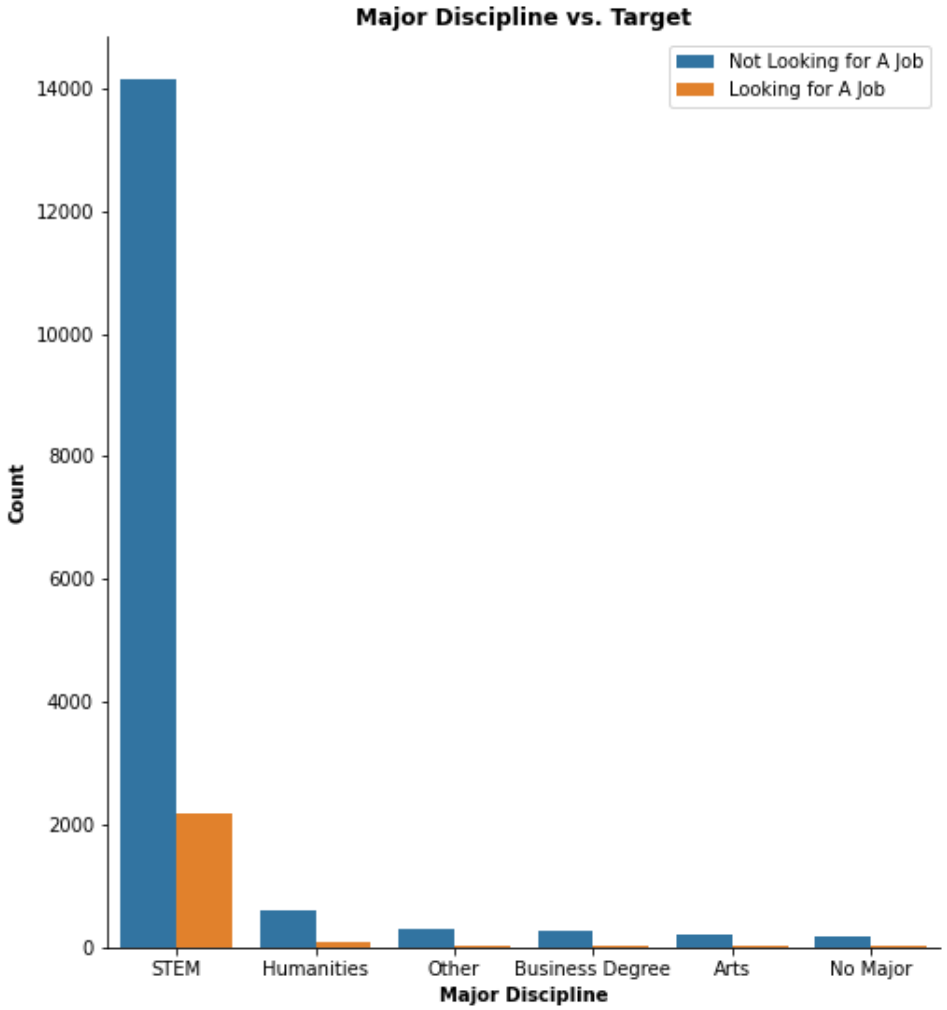
- This graph represents the number of trainees who looked for a job or not based on their university course enrollment status.
- It is obvious that trainees who were not enrolled into any university courses were not most likely to be looking for a new job.
- We can also see that trainees with part time university course tended to be looking for a new job.

Exploratory Data Analysis
Bivariate Charts



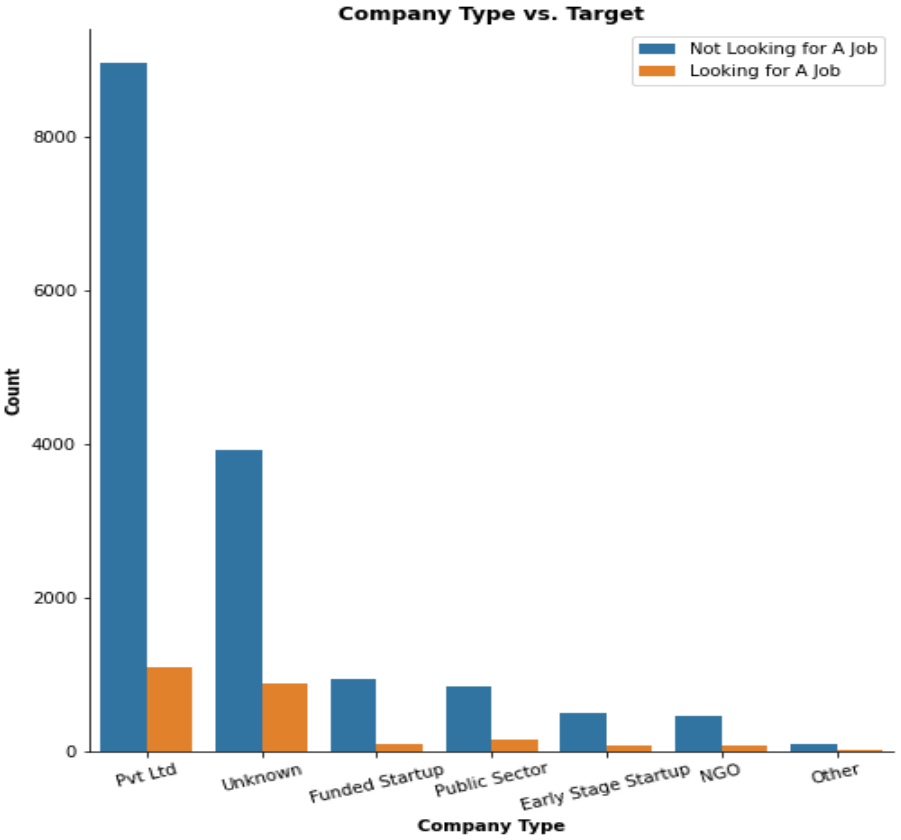
- This graph represents the number of trainees who applied looking for a new job or not based on their education level.
- It is obvious that nearly all of the trainees with primary school and PHD education levels were not looking for a new job.
- It also shows that the majority of graduates were not really looking for a new job.

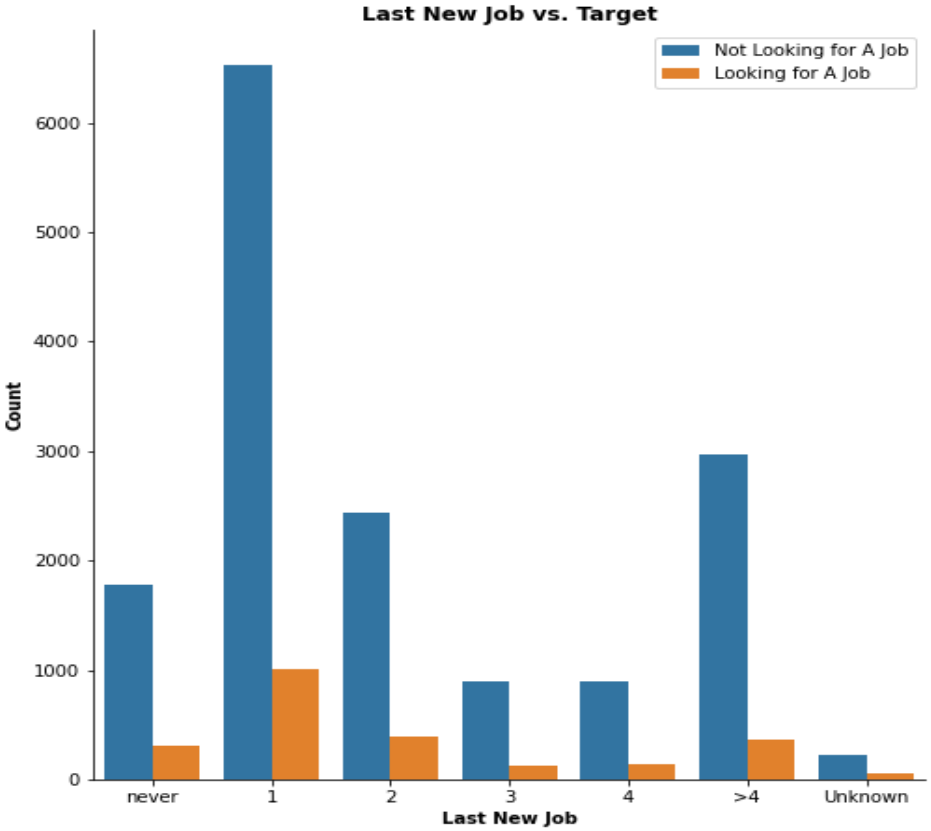
Exploratory Data Analysis
Bivariate Charts



- This graph represents the number of trainees who applied looking for a new job or not based on their major.
- Clearly, all the majors other than stem were not really looking for a new job.
- It also shows that a large number of trainees from STEM major were not really looking for a new job.
- So, we can conclude that you can't judge whether the trainee is looking for a new job or not based on their major.

Exploratory Data Analysis
Bivariate Charts





- We can see that most trainees who worked in a private limited company were not looking for a new job.

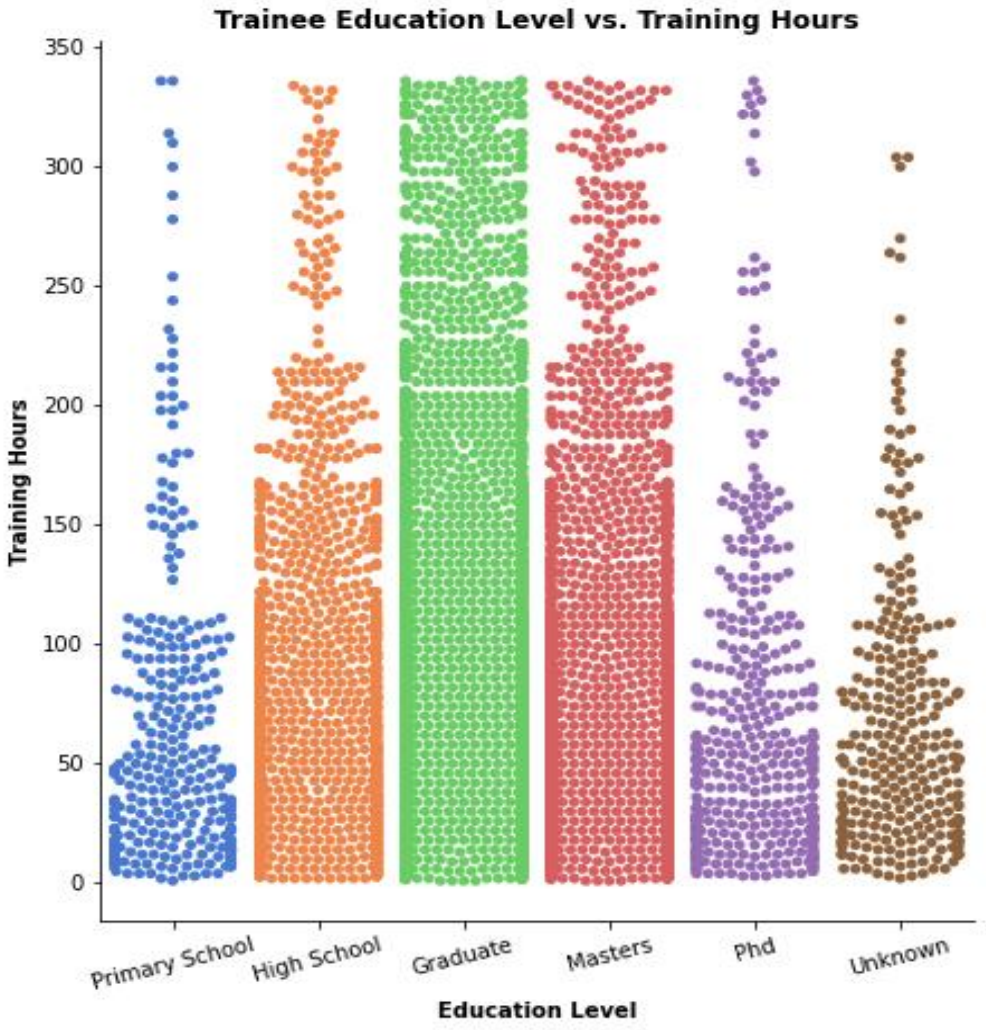
- We can see that most trainees who had one year of difference between their previous and current job were not really looking for a new job.

Exploratory Data Analysis
Bivariate Charts

- This graph represents the number of trainees who applied looking for a new job or not based on their experience.

- Clearly, most the trainees who had more that 20 years of experience were not looking for a new job.

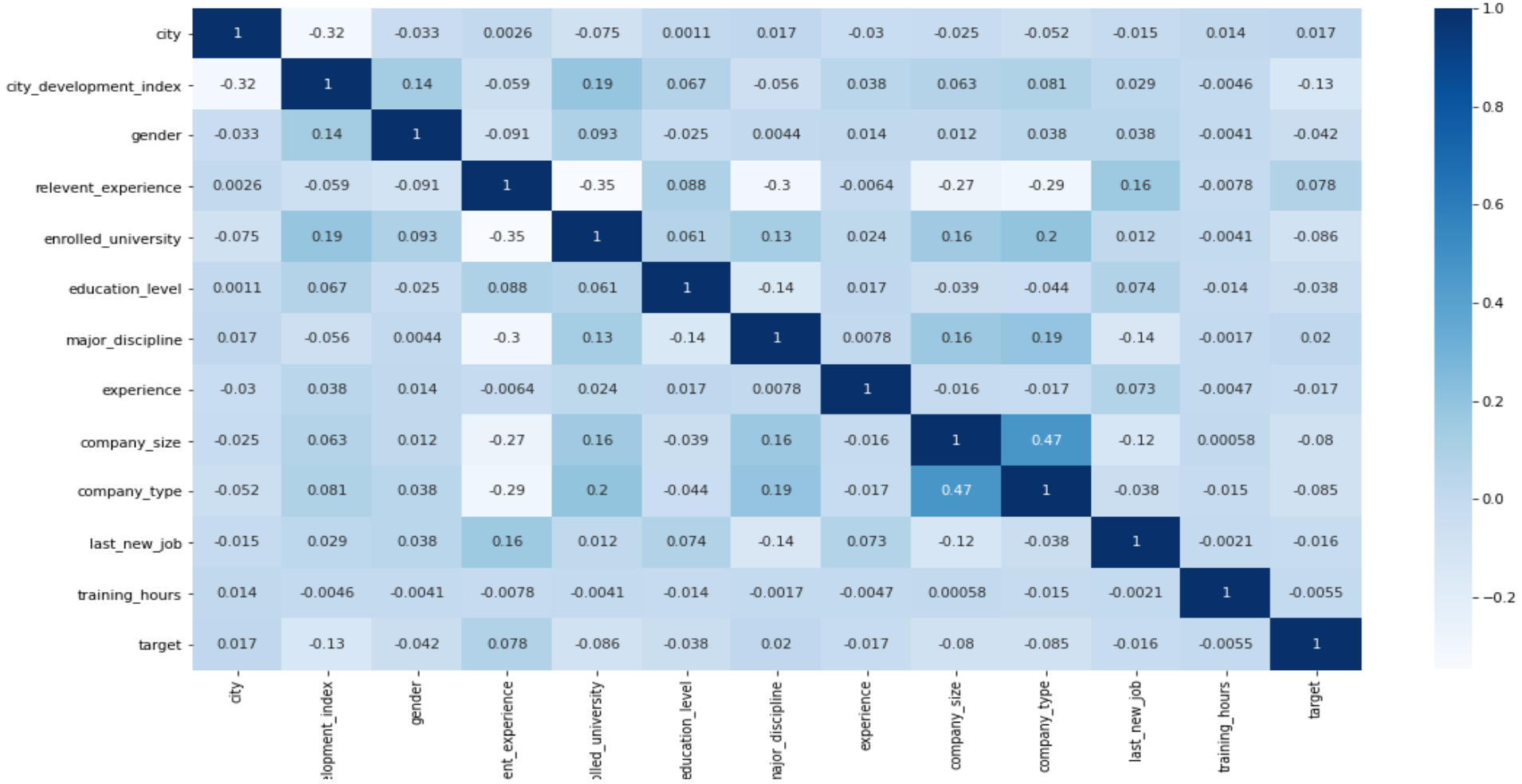
➢ Exploratory Data Analysis
Multivariate Chart



- This graph represents the number of trainees classified by their education level and the number of training hours they have completed.
- We can see that graduates are the majority and they have completed the most training hours.
- Then comes trainees with master's education level.

Machine Learning Models Correlation Matrix

- We have created three machine learning models to try to reach the best accuracy because of the so many null values in the data.
- The correlation matrix clarifies the correlations between the features.

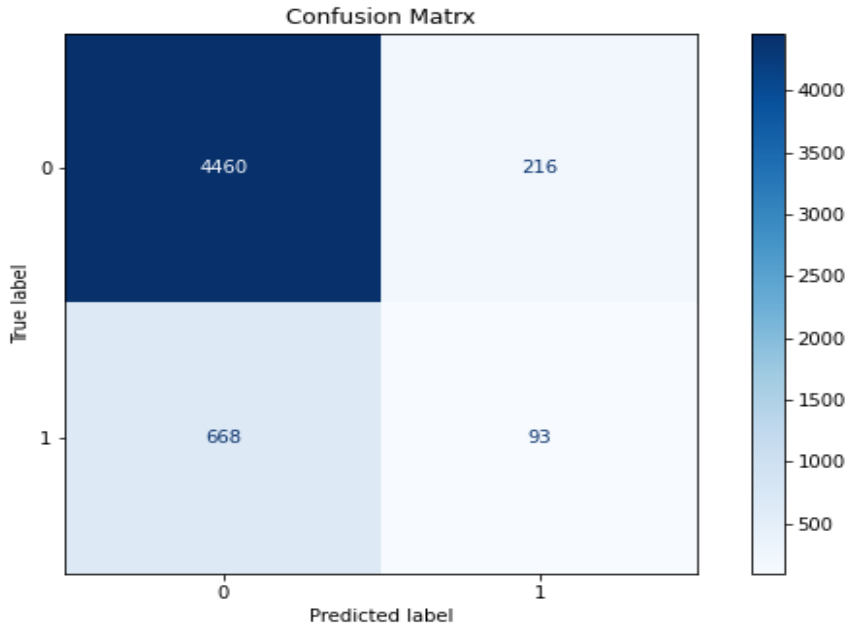


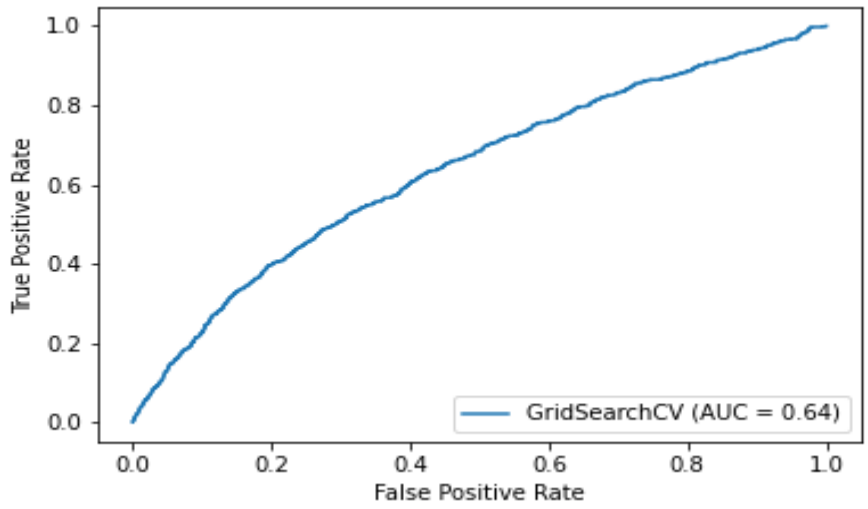
Machine Learning Models

Model Using SMOTE (Replacing NULL with 0)

- In this model we replaced the null values with as of conceding the null values a new class of the data.
- Classification Report:

	Precision	Recall	F1-Score	Support
0.0	0.87	0.95	0.91	4676
1.1	0.30	0.12	0.17	761
Accuracy			0.84	5437
Macro Avg	0.59	0.54	0.54	5437
Weighted Avg	0.79	0.84	0.81	5437



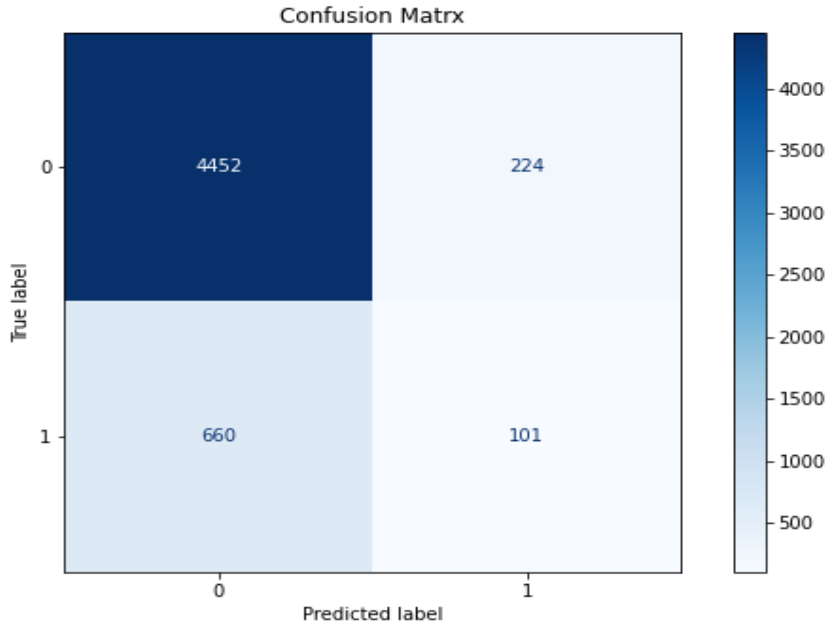


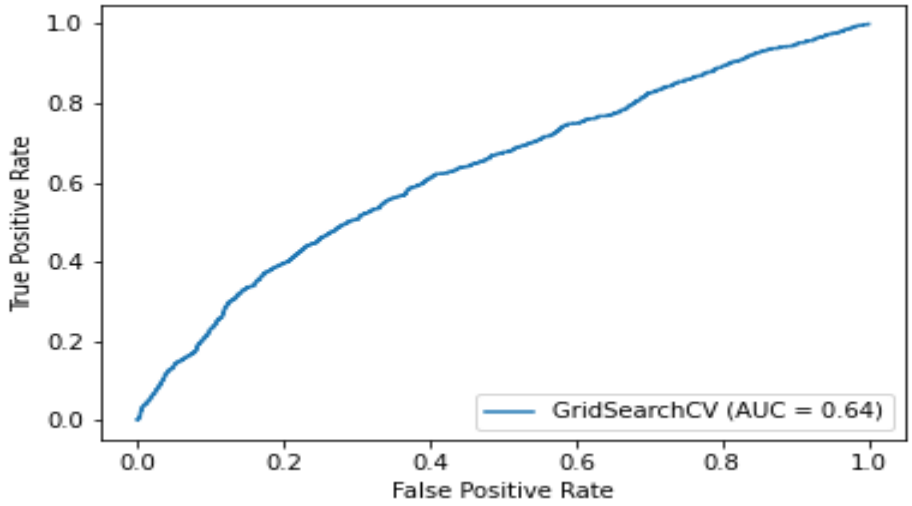
Machine Learning Models

Model Using SMOTE (Predicted NULL Values)

- In this model we used the model we created to predicted our null values on “enrolled_university” and “major_discipline” columns and replaced the remaining null values with 0.
- Classification Report:

	Precision	Recall	F1-Score	Support
0.0	0.87	0.95	0.91	4676
1.1	0.31	0.13	0.19	761
Accuracy			0.84	5437
Macro Avg	0.59	0.54	0.55	5437
Weighted Avg	0.79	0.84	0.81	5437



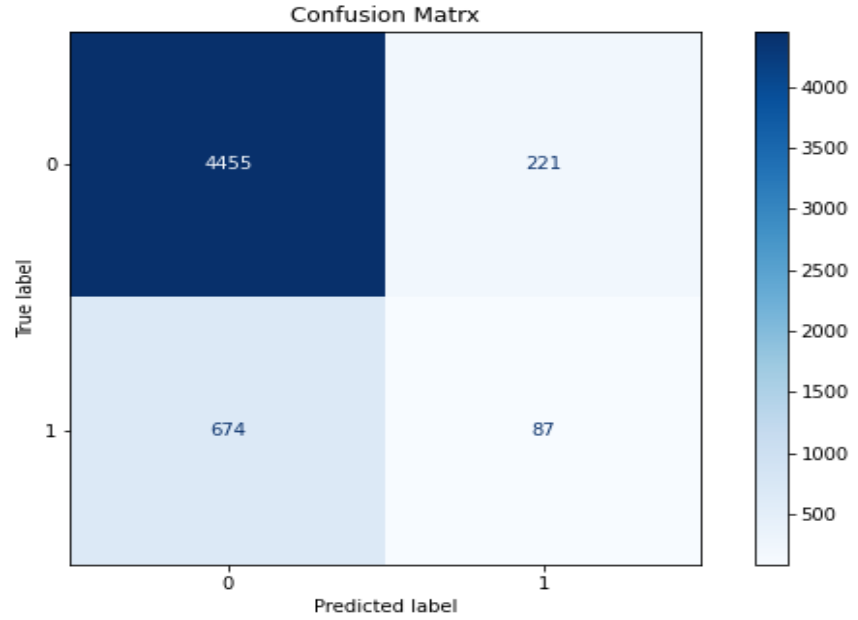


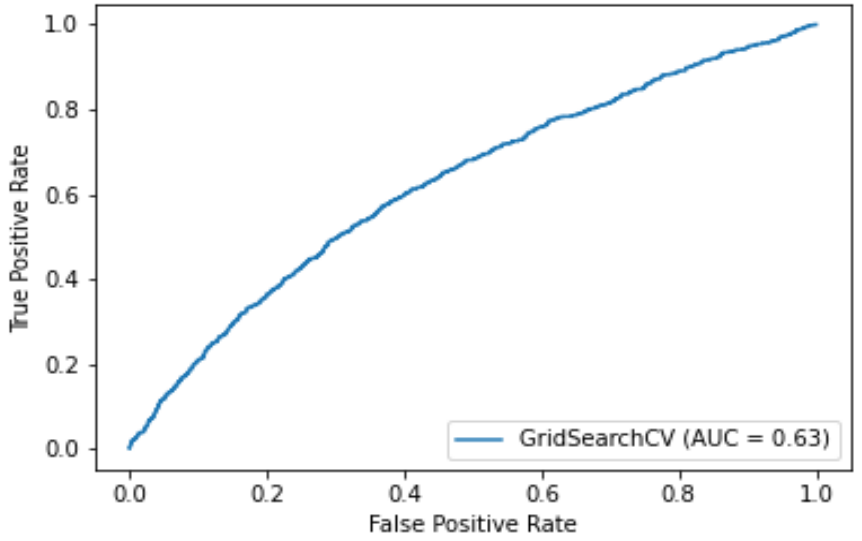
Machine Learning Models

Model Using SMOTE(Replacing NULL Values with Mode)

- In this model we replaced the null values in each column with the most frequent value in that column.
- Classification Report:

	Precision	Recall	F1-Score	Support
0.0	0.87	0.95	0.91	4676
1.1	0.28	0.11	0.16	761
Accuracy			0.84	5437
Macro Avg	0.58	0.53	0.54	5437
Weighted Avg	0.79	0.84	0.80	5437





➢Conclusion

1. I have used the univariate charts to explore the data in each column and get a clear understanding of the data.

2. Bivariate charts to explore the relation between two features and see how each feature could affect the target.

3. Multivariate swarm plot to explore the relationship between the number of trainees classified by their education level and the number of training hours they have completed in training.

4. I have created three models to try to reach the best accuracy of the predicting whether the future will apply to look for new employment or not.

5. The three models were created due to the ambiguity I faced of how to deal with the so many null values.

6. I used the SMOTE algorithm to overcome the imbalance of the data.

➢Problems we faced in our investigation

1. The data contains a lot of null values that affected the analysis, and the accuracy of the machine learning models.

2. There are a lot of features with imbalance data.

3. The target data is imbalanced which caused the machine learning model to be biased to the majority.

SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations