**Datacorp**

**Overview of project and business goals**  
'''The project we are working on is DataCorp, a leading company in Big Data and Data Science company.

**The objective:** Build a machine learning model that predicts the probability of a candidate seeking a new job or staying with DataCorp after completing the training  
  
**#Data validation**  
'''  
Before I did data validation and cleaning, the dataset contained 19158 1rows and 14 columns.  
After validating each column against the criteria provided, these are my observations:

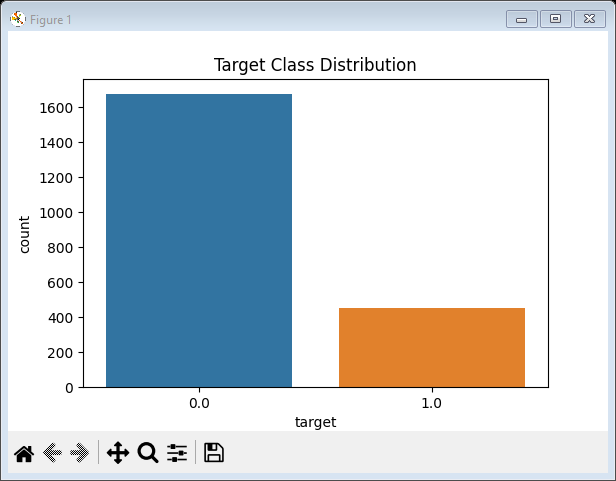
Gender  
'''On the gender column we had 4508 missing values. Since the data is anonymous I decided to fill the missing values with "other'''

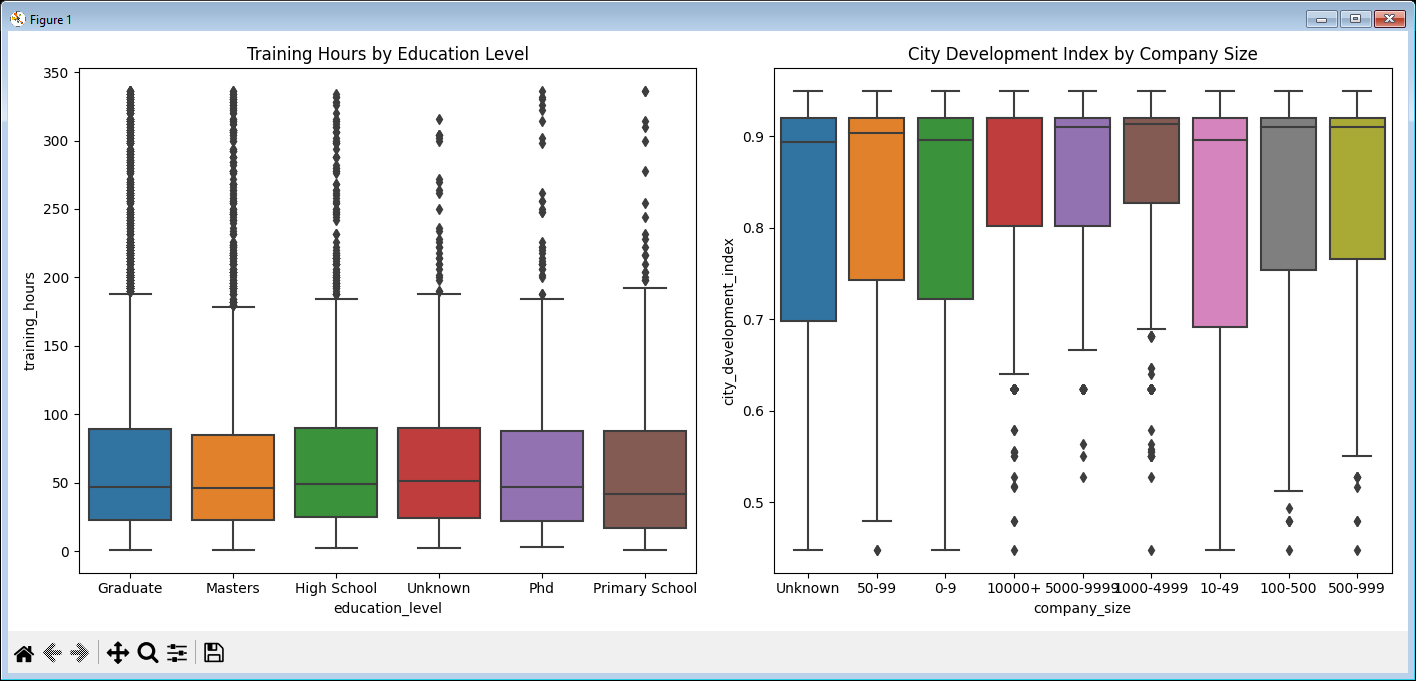
# enrolled university  
'''This column had 386 missing values. These were replaced by "unknown"'''

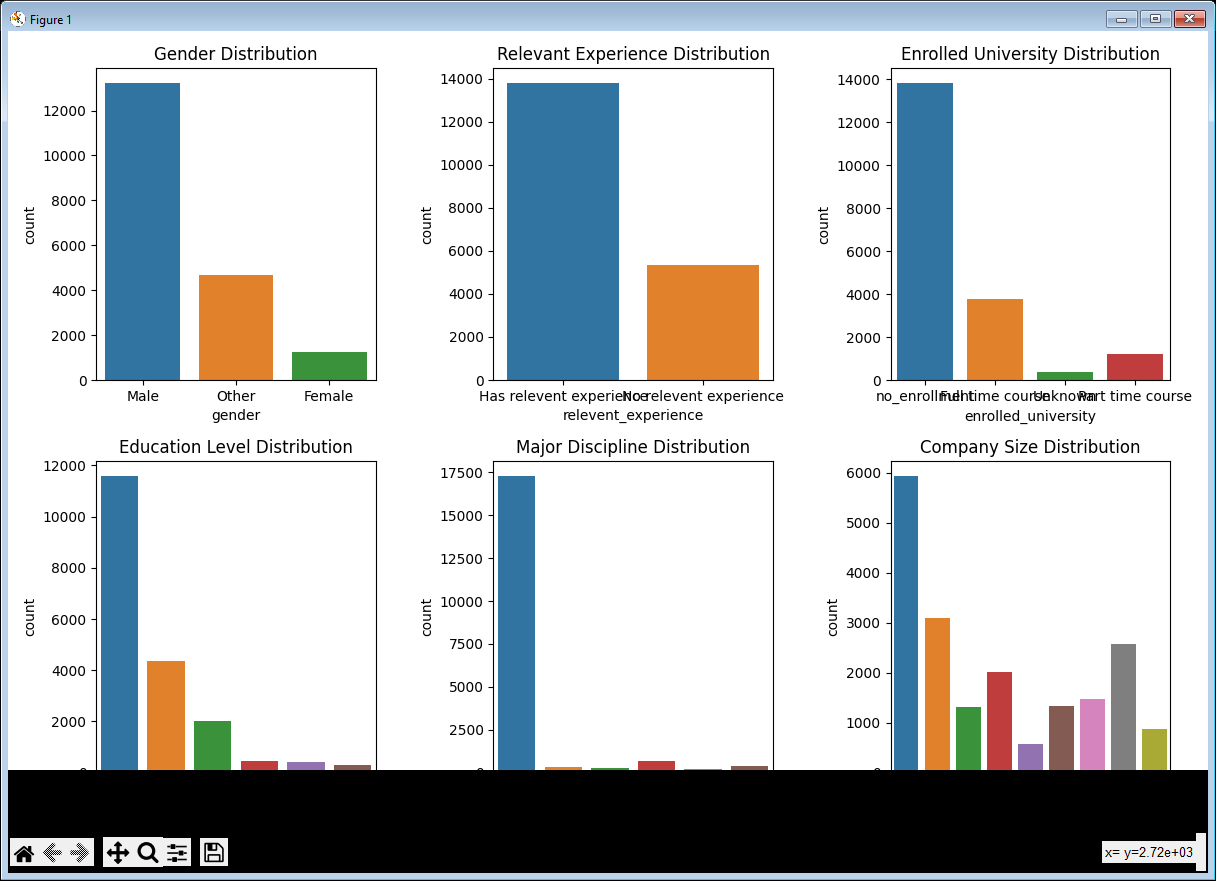
#education\_level  
'''This column had 460 missing values replaced by "unknown"'''  
  
#major\_discipline  
"This had 2813 missing values which I filled with the mode "  
  
#experience  
'''Experience column had 65 missing values. Since the data is categorical I opted to add another category "unknown".'''  
  
#company size  
'''This column has 5938 missing values that were filled with " unknown" being categorical'''  
  
#company type  
'''This column had 6140 missing values filled with "unknown"'''  
  
#last\_new\_job  
'''This column has 423 missing values filled with " unknonwn'''  
  
#Company size  
''' the company size has categories outside scope a 10/49 changed to 10-49 and <10 to 0-9. I decided to group similar categories together to reduce the number of unique values'''

**Key findings**

Given that our objective is to build a machine learning model that predicts the probability of a candidate seeking a new job or staying with DataCorp after completing the training, my observations are 450 out of 2129 people are likely to look for a job change.



The above box plot show the distribution in the training hours by the education level and city development by the company size.



The above visualizations show the various distributions that are there from Gender, Experience, enrollment of university, education level, major discipline and company size.

**RECOMMENDATIONs**

Based on the models insights I would suggest.

1.Identify trends over time and make data-driven decisions on resource allocation and strategy adjustments.

2. Enhance data collection process for more in depth analysis.

3.Improve quality of data collected due to a lot of missing values.

4.Continuous monitoring of key metrics to see the number of people looking for job change

5.Regular data analysis to stay informed on methods that are most beneficial.