Visualization

- we use two packages to visualize our data and these packages are matplotlib package and seaborn package , these packages can do a lot in visualization and each package have many plots on it , we use some of these plots like bar plot in matplotlib package and count plot ,box plot in seaborn package .
- first thing we import the packages we need to use it in code.

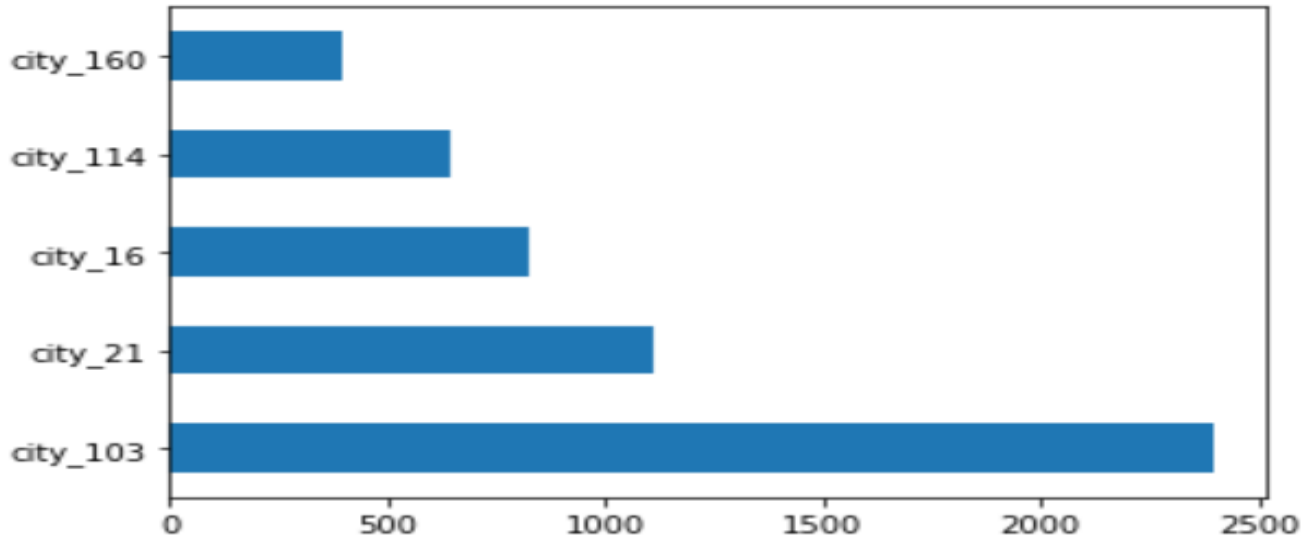
```
In [16]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns
         %matplotlib inline
         %reload_ext autoreload
         %autoreload 2
```

- Here we use matplotlib package and we use bar and barh from it .
- when we analyze city development index column We conclude that :-

1. The mean value of city development is 0.83
2. The median is 0.91
3. The standard deviation equals 0.11
4. The first half of the values are less than 0.91 and the second half is more than 0.91.
5. Looking at the standard deviation we can see, that the values do not differ from the average of values.It means that most of the candidates are from well-developed cities.
6. Most of the candidates come from the city_103. Next are city_21, city_16, city_114, city_160.

```
city_dev_Mean: 0.8441975892757027
city_dev_Median: 0.91
city_dev_Standard deviation: 0.11639971276054528
```

Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x20978125408>

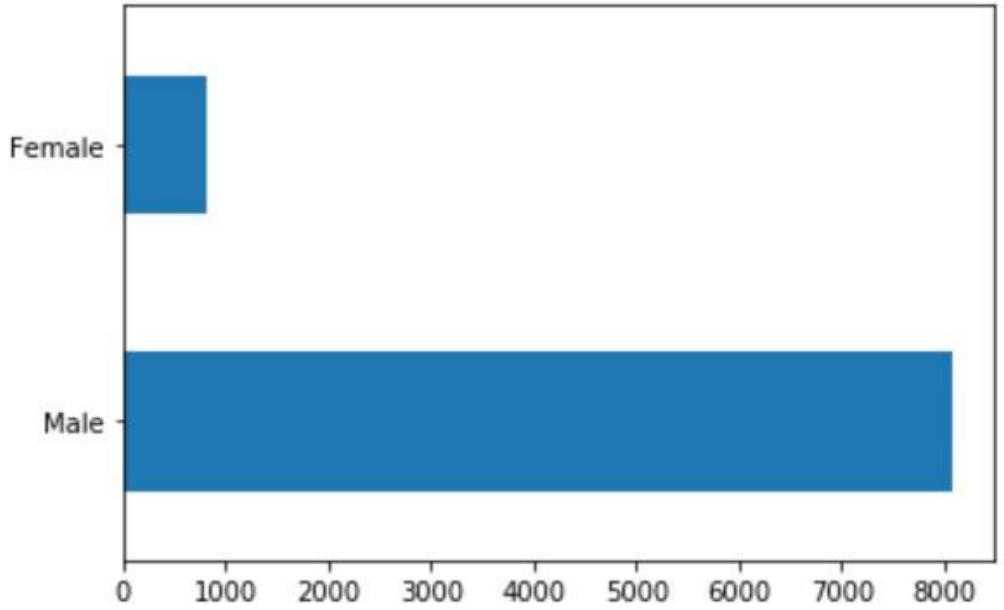


❖ When we analyze the gender column we conclude that :-

- We can see a big difference between the genders. In candidates, list prevails men.
- There are 8073 men and 804 women .

```
Male      8073
Female     804
Name: gender, dtype: int64
```

```
Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0x209781804c8>
```

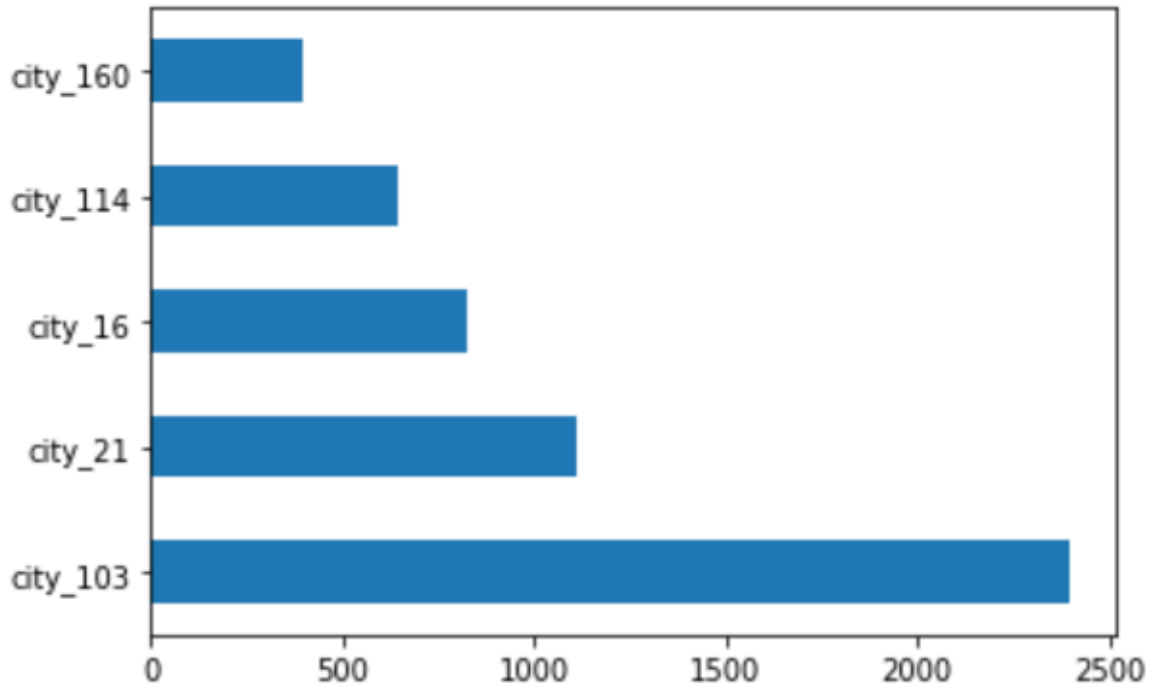


❖ When we analyze relevent experience column we conclude that :-

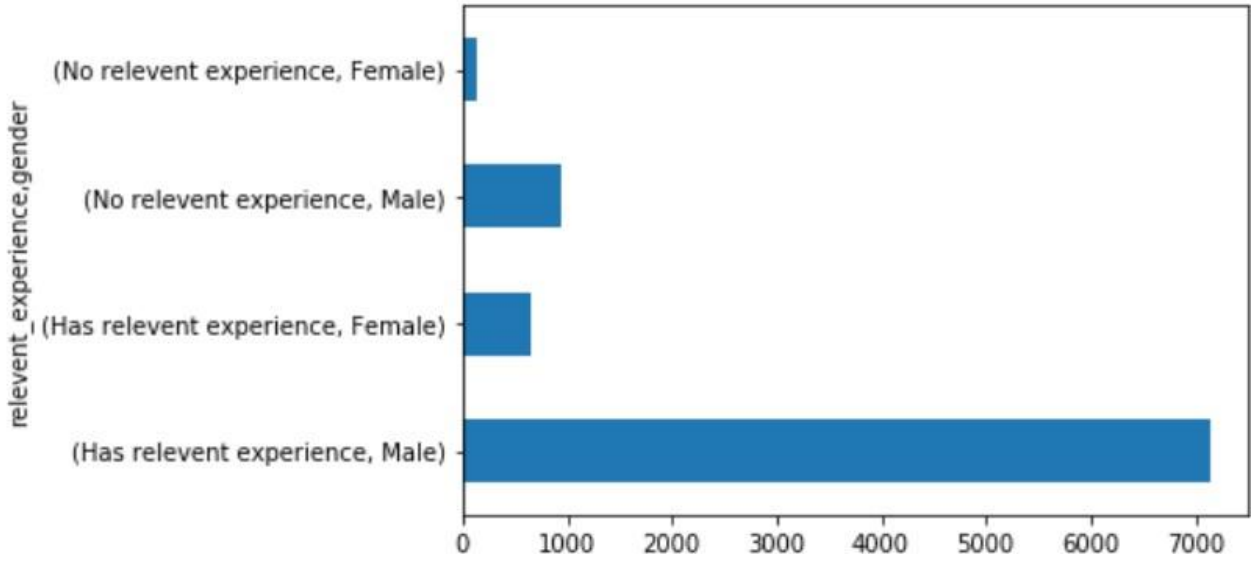
- Around 7798 people have relevant experience and 1079 hasn't it. Based on genders, people mostly have a relevant experience.
- In the male group, 11.5% of men haven't got relevant experience. The group of women fared worse. In this group, 17.9% of women haven't got relevant experience.

Out[22]:
```
Has relevent experience    7798
No relevent experience     1079
Name: relevent_experience, dtype: int64
```

Out[20]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x20978125408>
```



```
Out[24]: relevent_experience      gender
         Has relevent experience  Male      7138
                                  Female     660
         No relevent experience   Male       935
                                  Female     144
         Name: gender, dtype: int64
```

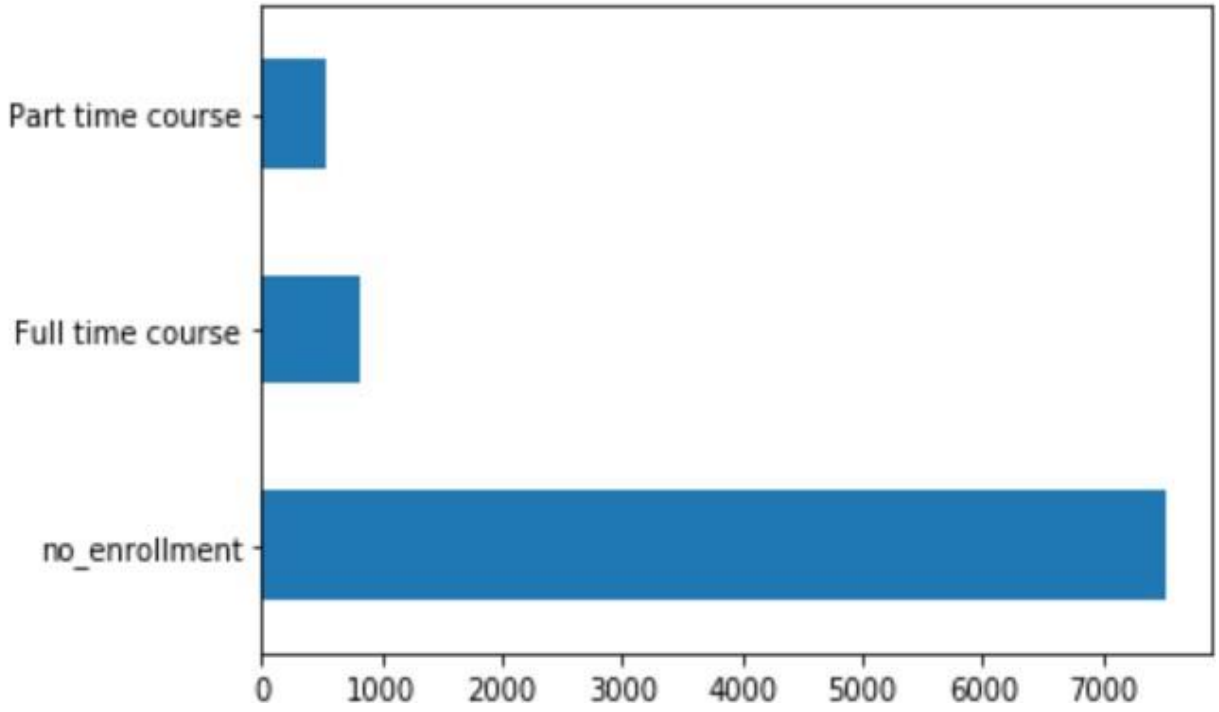


❖ When we analyze Enrolled university column we conclude that :-

- The most of people did not attend college.
- 822 people attended college in a full-time course. 522 people attended college on a part-time course.
- 15.14% of candidates attended college.

Out[27]:
```
no_enrollment       7533
Full time course     822
Part time course     522
Name: enrolled_university, dtype: int64
```

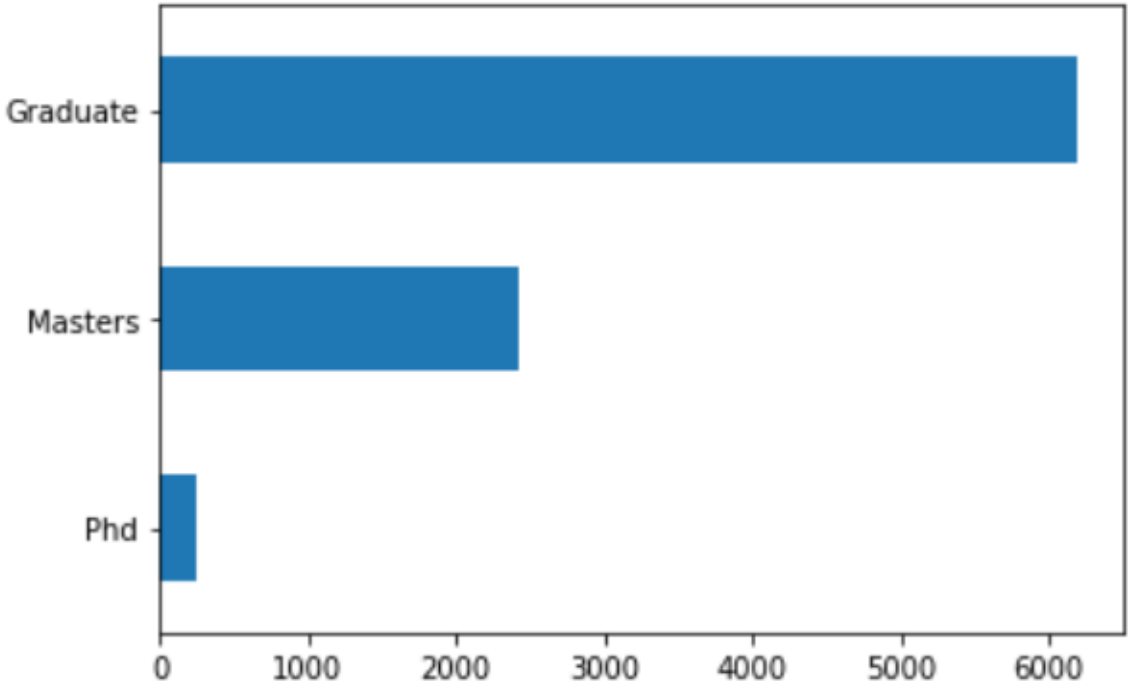
Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x20978268048>



❖ When we analyze education level column we conclude that :-

- The most candidates are in Graduate level

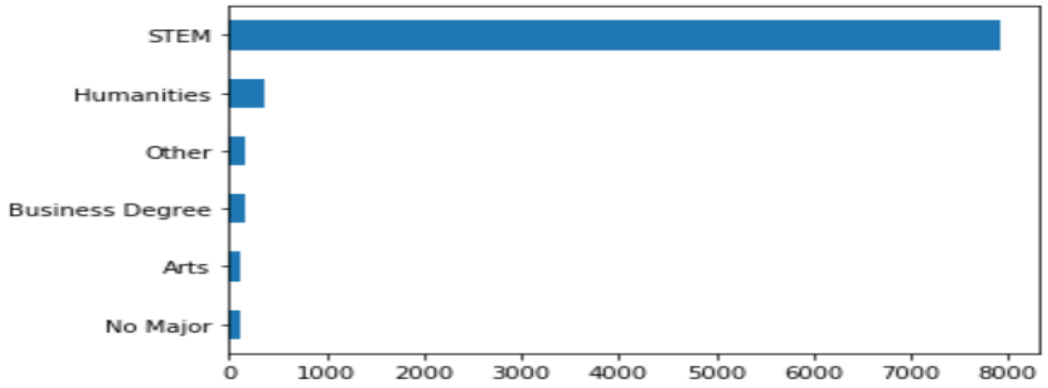
```
Out[115]: Phd          254
          Masters     2433
          Graduate    6190
          Name: education_level, dtype: int64
```



❖ When we analyze Major discipline column we conclude that :-

- The vast majority of candidates specialize in the STEM discipline (7924 candidates).
- Humanities (374)
- Other (175)
- Business Degree (168)
- Arts (126)
- No Major (110)
- The most candidates who are specialized in STEM, have a Graduate level. Data Science relies heavily on math and science. This explains why so many candidates specialize in STEM.

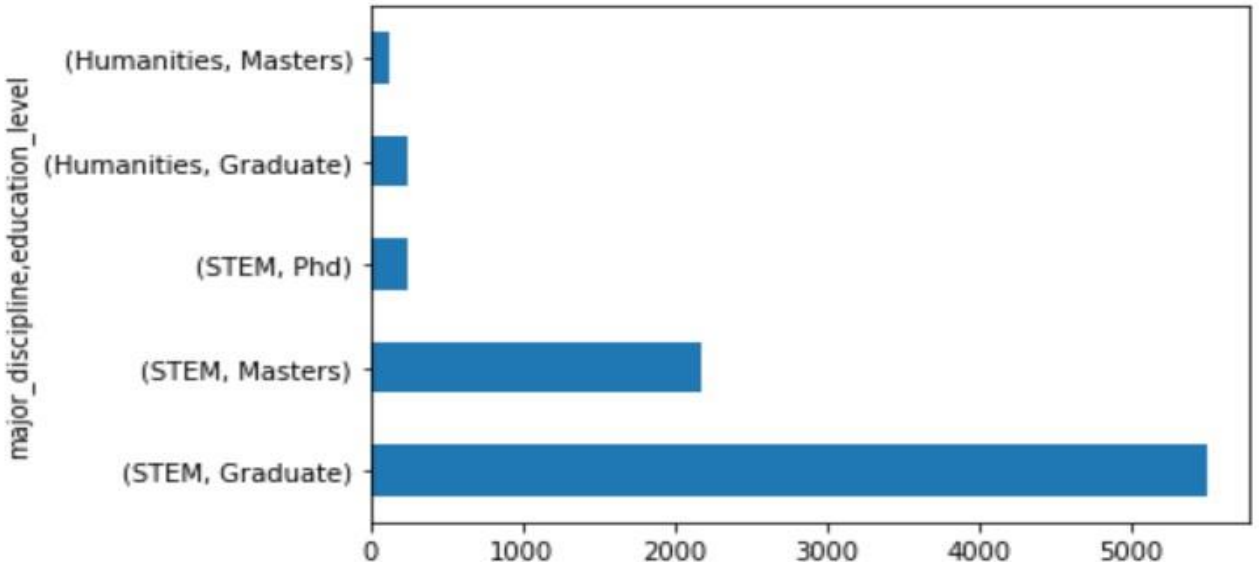
```
Out[31]: No Major              110
         Arts                  126
         Business Degree       168
         Other                 175
         Humanities            374
         STEM                 7924
         Name: major_discipline, dtype: int64
```



Out[32]:

major_discipline	education_level	
STEM	Graduate	5505
	Masters	2180
	Phd	239
Humanities	Graduate	236
	Masters	128

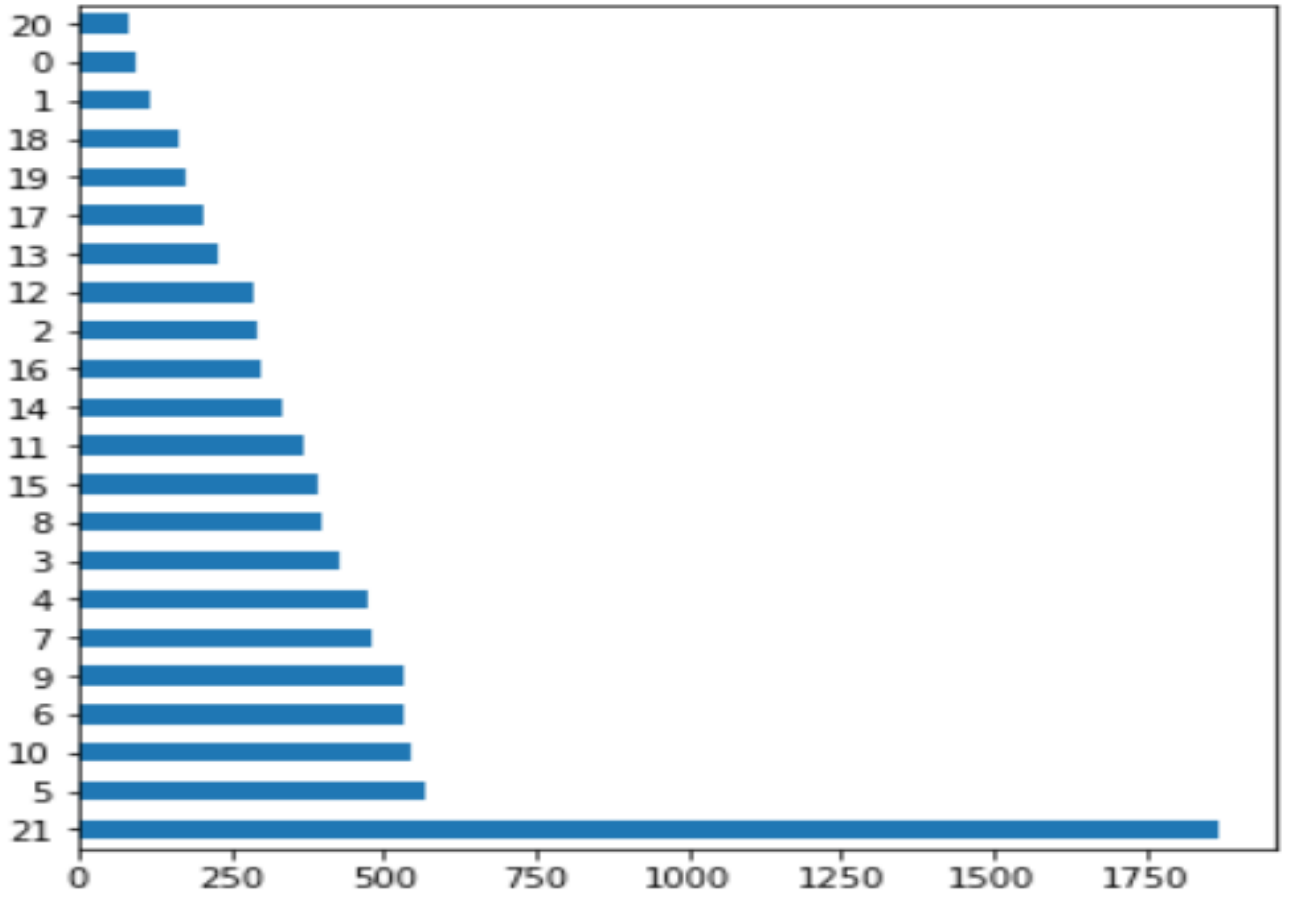
Name: education_level, dtype: int64



❖ When we analyze experience column we conclude that :-

- About 1868 candidates have over 20 years of experience in Data Science. There are also many people with 5 years of experience
- We can see the increasing popularity of Data Science. Most candidates have between 6 and 10 years of expeiance and between 11 and 15 years of experience.

Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x1e056

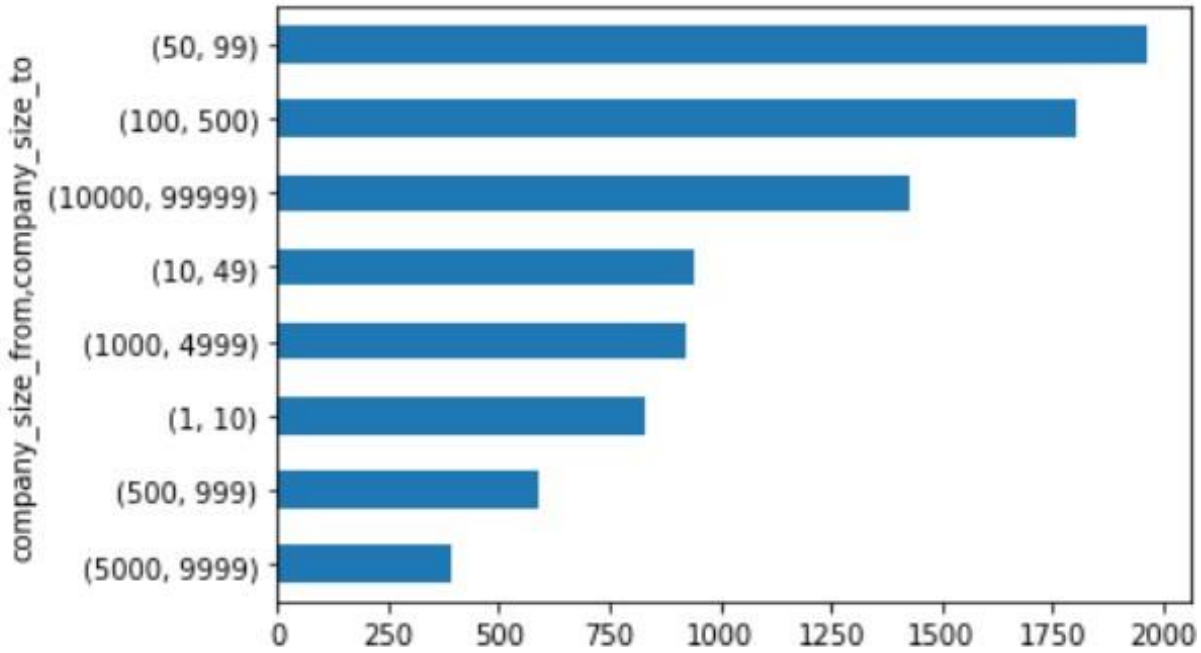



```
Candidates with experience between less than one and five years:  1979
Candidates with experience between six and ten years:  8877
Candidates with experience between eleven and fifteen years:  8877
Candidates with experience between sixteen and more than twenty years:  2799
```

❖ When we analyze company size column we conclude that :-

- The most candidates work in small companies (from 50 to 99 and from 100 to 500 workers).

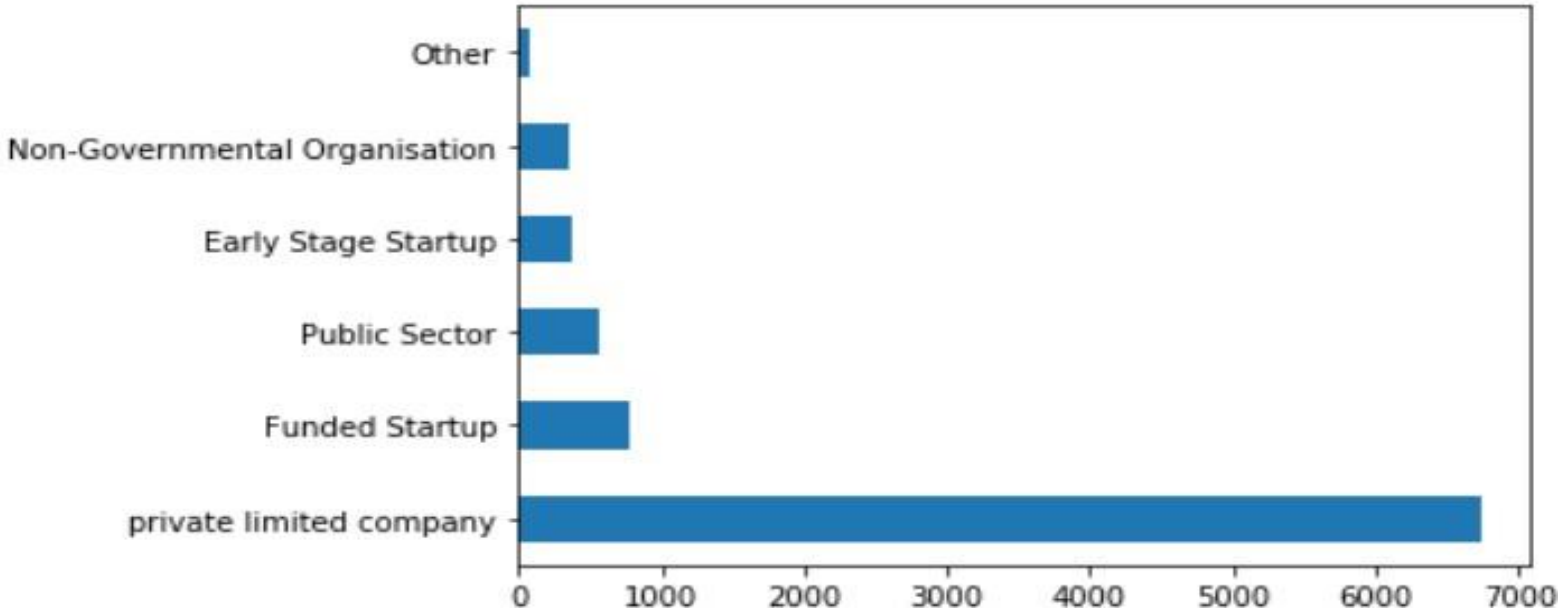
```
Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0x1e0566ca688>
```



❖ When we analyze company type column we conclude that :-

- Most of the companies which candidates work is a private limited company (6738). Next are Funded Startup (775) and Public sector (559).

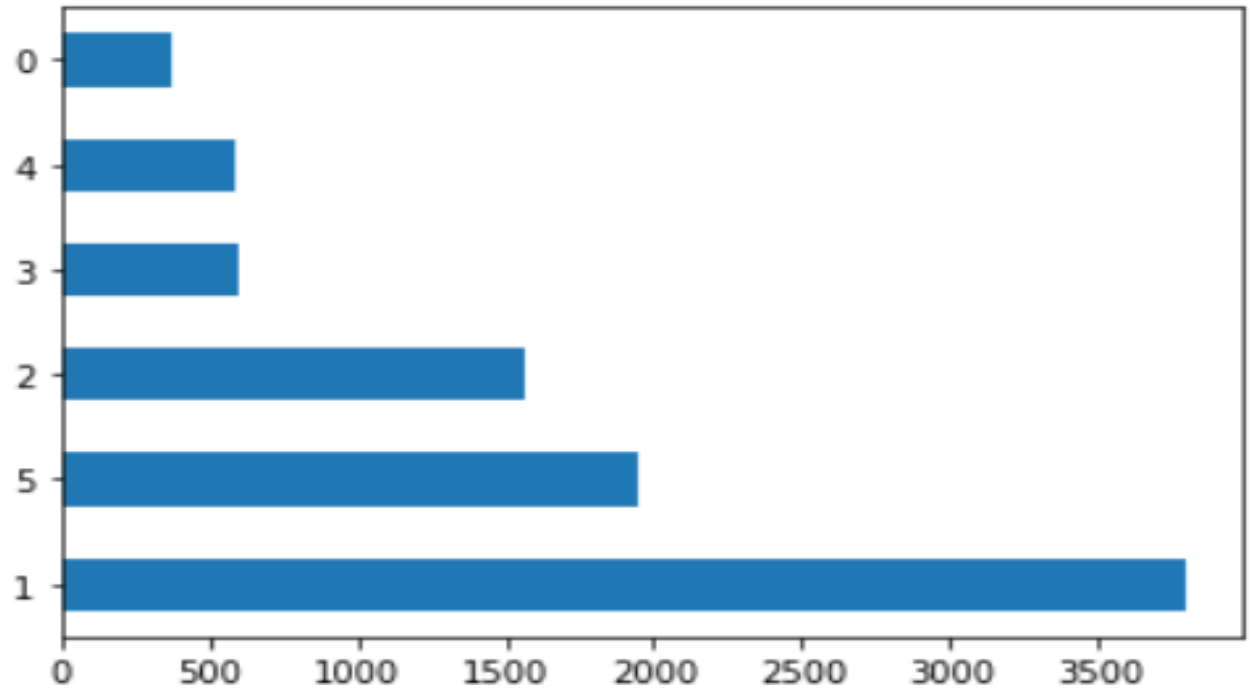
```
Out[22]: private limited company          6738
         Funded Startup                    775
         Public Sector                     559
         Early Stage Startup               382
         Non-Governmental Organisation     352
         Other                              71
         Name: company_type, dtype: int64
```



❖ When we analyze Last new job column we conclude that :-

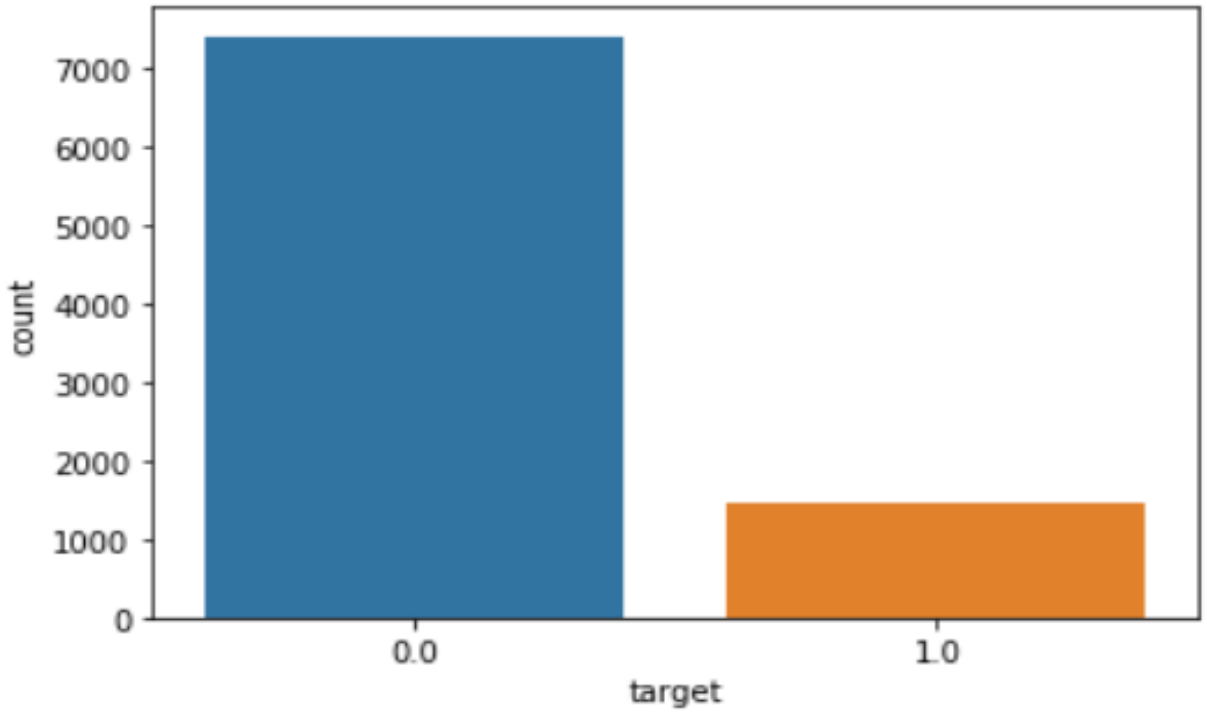
- The most common difference between the candidate's past and current job is 1 year.

```
Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x1e056
```



- ❖ Here we use seaborn package and we use count plot from it .
- ❖ Target
  - 0 – Not looking for job change,
  - 1 – Looking for a job change
  - ✓ As we can see, here we have imbalanced data, the number of 1 (Looking for a job change) < 0 (Not looking for job change)
  - ✓ the probability that the person will not change the job is higher

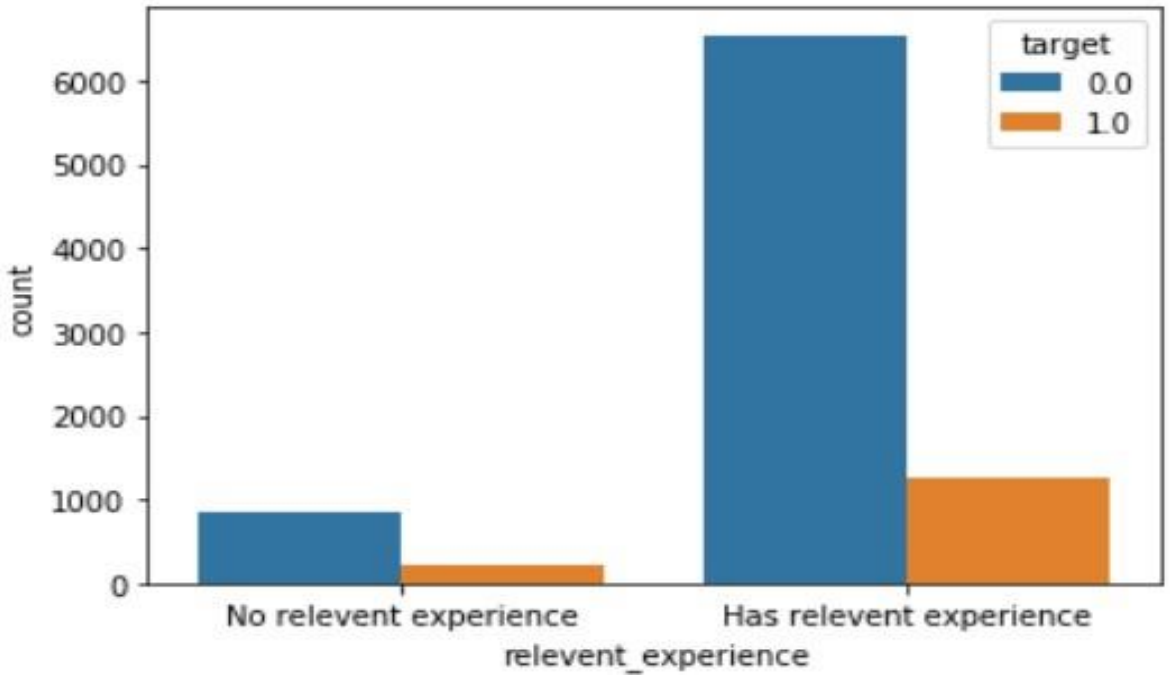
```
In [24]: sns.countplot(file['target'])
Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0x1e05689ea88>
```



❖ When we compare relevent exeprience column with target column we conclude that :-

- The most candidates with relevent experience and with not relevent experience are not looking for a job chance

```
In [25]: sns.countplot(file['relevent_experience'],hue=file['target'])
Out[25]: <matplotlib.axes._subplots.AxesSubplot at 0x1e056766ac8>
```

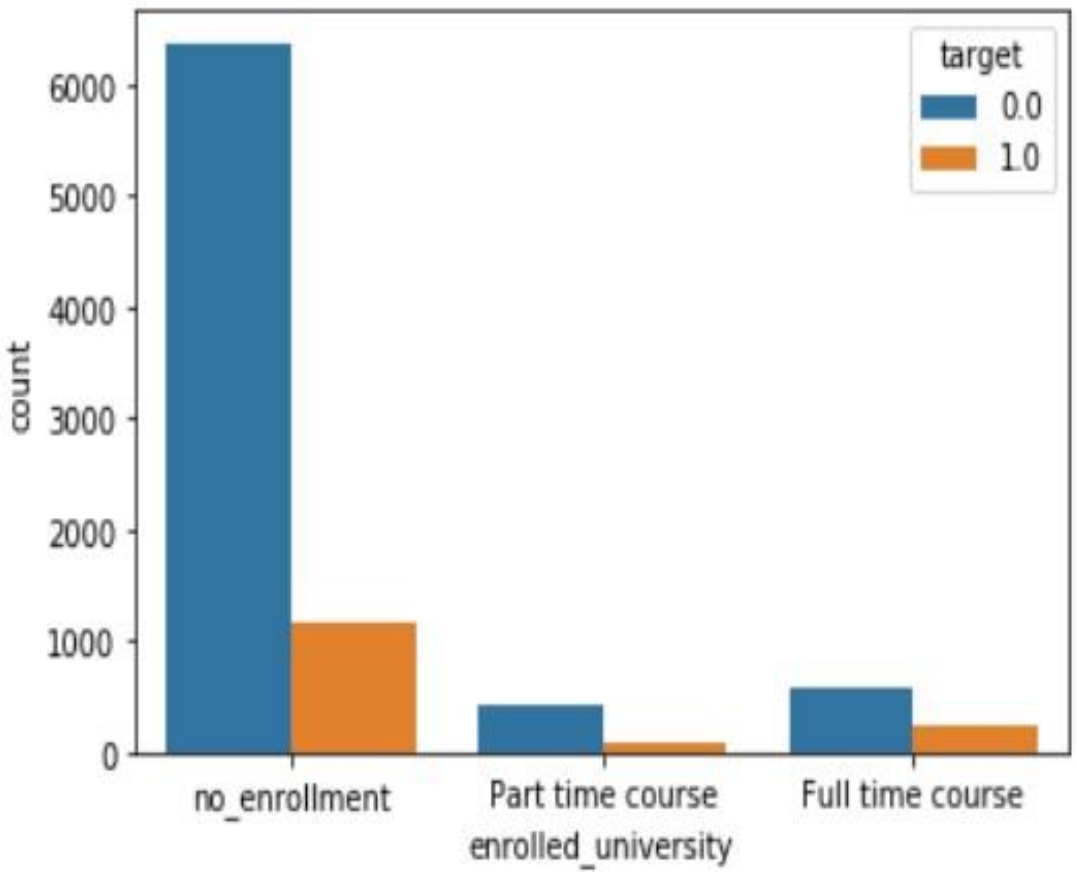


❖ When we compare enrolled university column with target column we conclude that :-

- Most of the person are having no enrollment in any university and those without any enrollment have no interest in changing their job
- Part Time course enrolled students are very less and they also dont want to change
- Full time enrolled are not very much also but in proportion they have higher chance of changing the job than others

```
In [26]: sns.countplot(file['enrolled_university'],hue=file['target'])
```

```
Out[26]: <matplotlib.axes._subplots.AxesSubplot at 0x1e056965088>
```

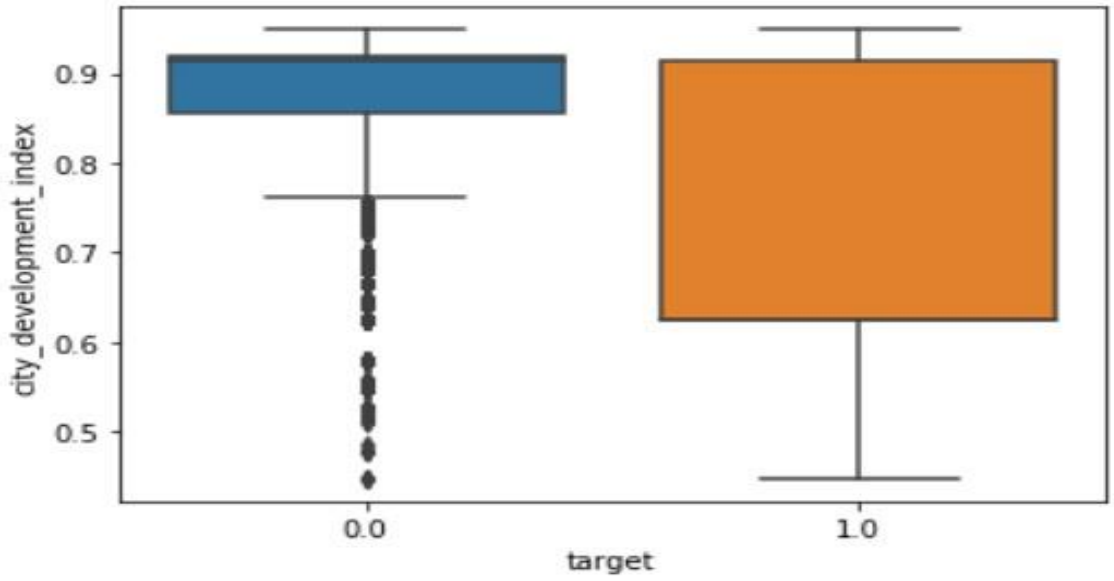


❖ Here we use box plot from seaborn package

❖ When we compare city development index column with target column we conclude that :-

- Most of the people who are not changing jobs are from city with high development index so basically they are having a comfortable life in respective city are not willing to change jobs
- People in cities with less development index tend to change their jobs for better life style maybe!

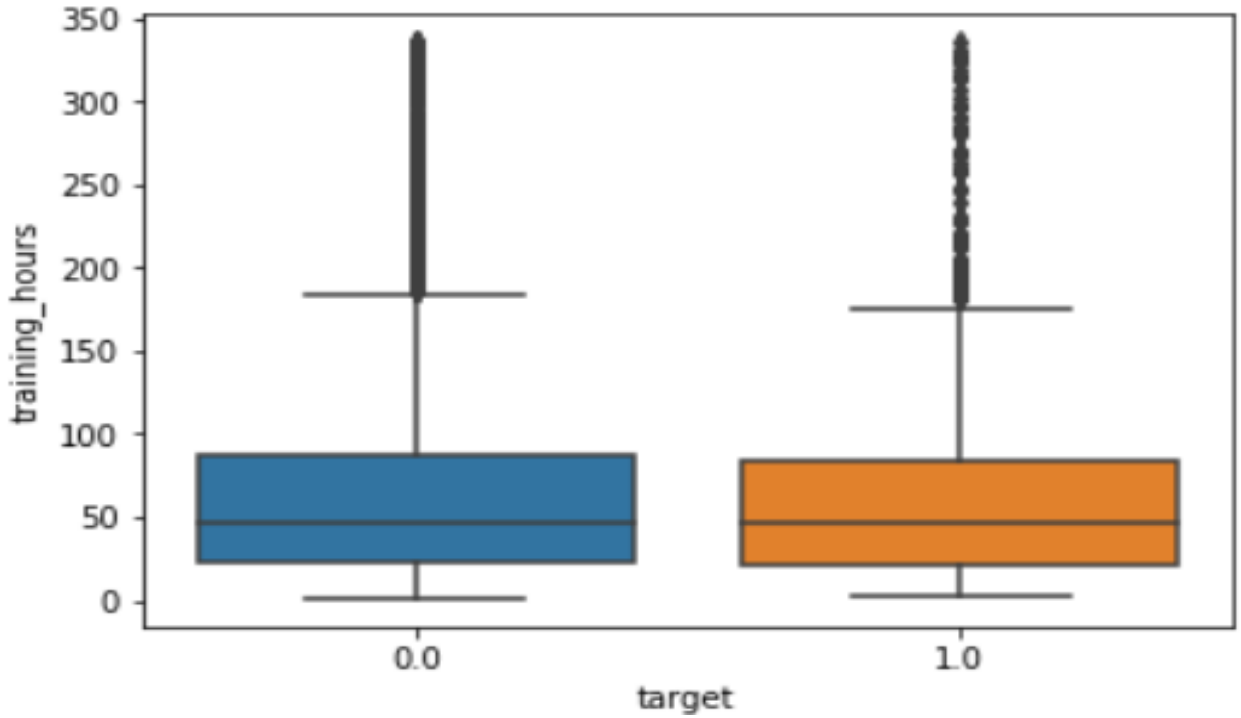
```
In [27]: sns.boxplot(x='target',y='city_development_index',data=file)
Out[27]: <matplotlib.axes._subplots.AxesSubplot at 0x1e0569e5b48>
```



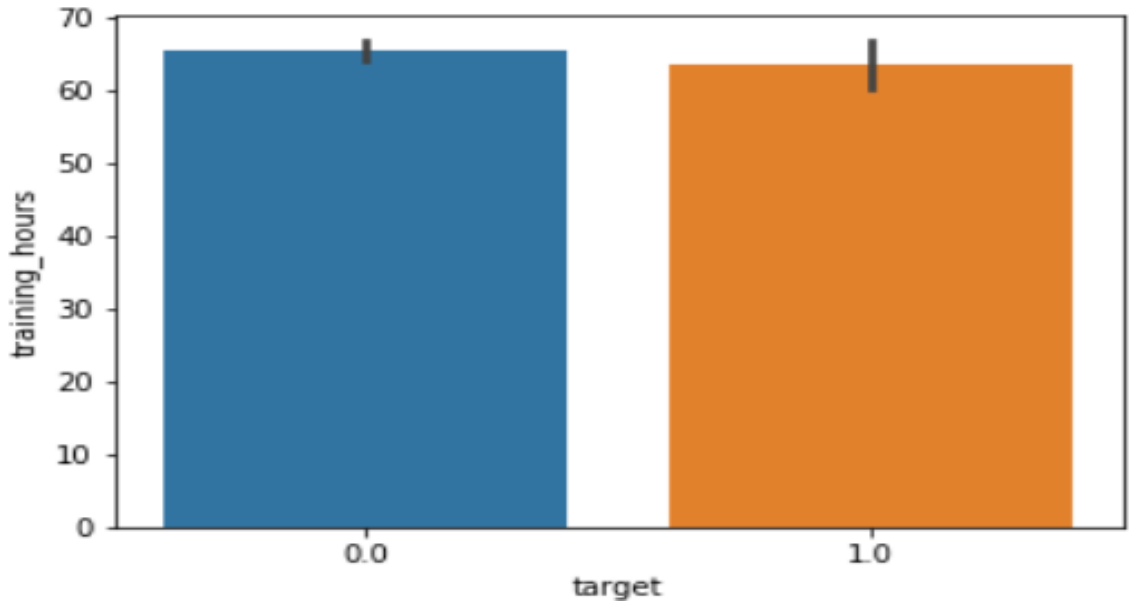
❖ When we compare training hours column with target column we conclude that :-

- There is no much difference between training hours of those who are wishing to change and those who are not , so not much can be deduced from this classification

```
In [28]: sns.boxplot(y='training_hours',x='target',data=file)
Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x1e056a5ad08>
```



```
In [29]: sns.barplot(y='training_hours',x='target',data=file)
Out[29]: <matplotlib.axes._subplots.AxesSubplot at 0x1e056ae0588>
```

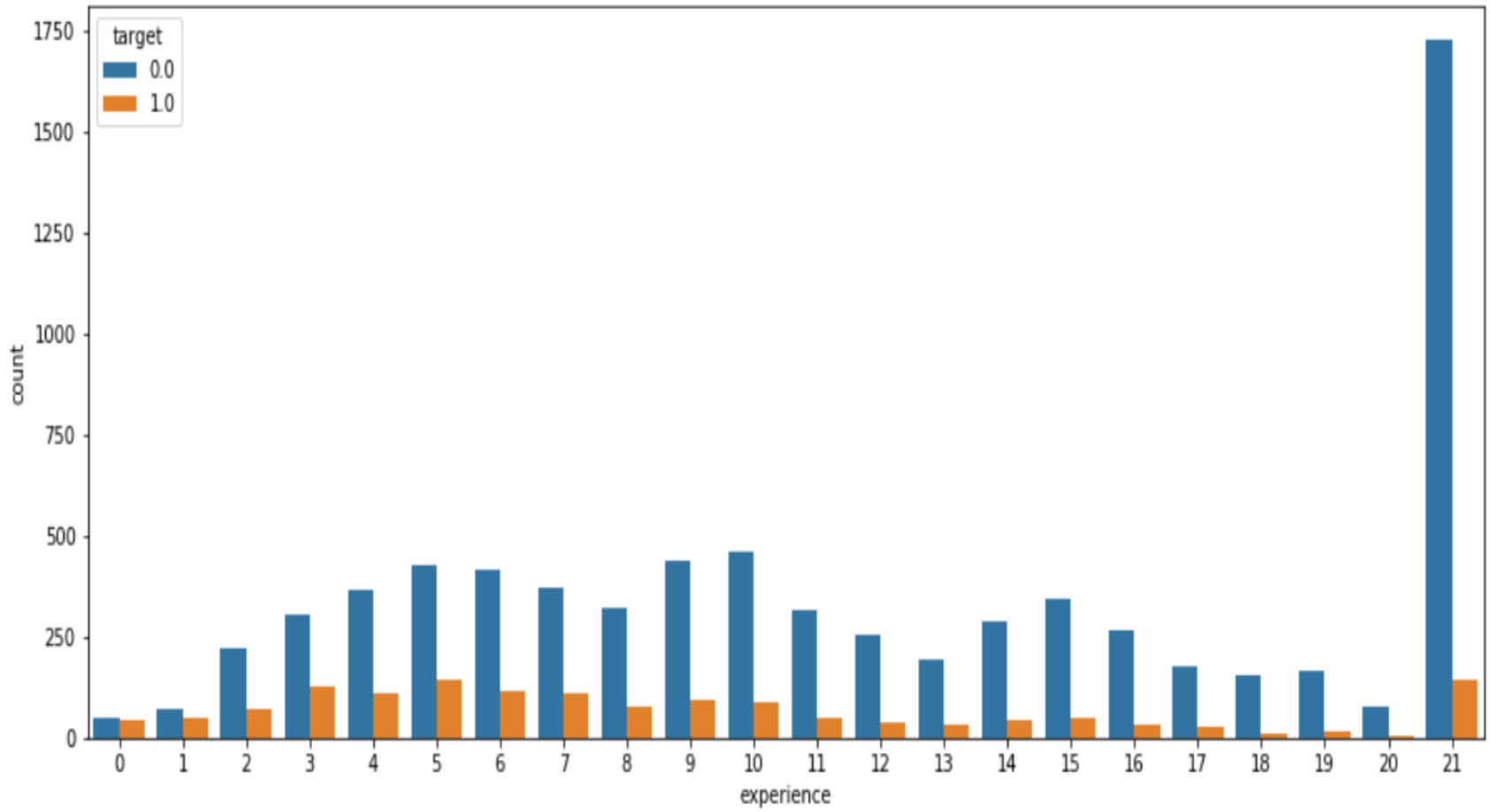


❖ When we compare experience column with target column we conclude that :-

- people who have experience less than 1 year have more tendency to change their job while those with more than 20 years of experience have very less tendency of changing jobz
- As experience increase , tendency to change the job becomes more and more less

In [33]:
```
plt.figure(figsize=(15,6))
sns.countplot(file['experience'],hue=file['target'])
```

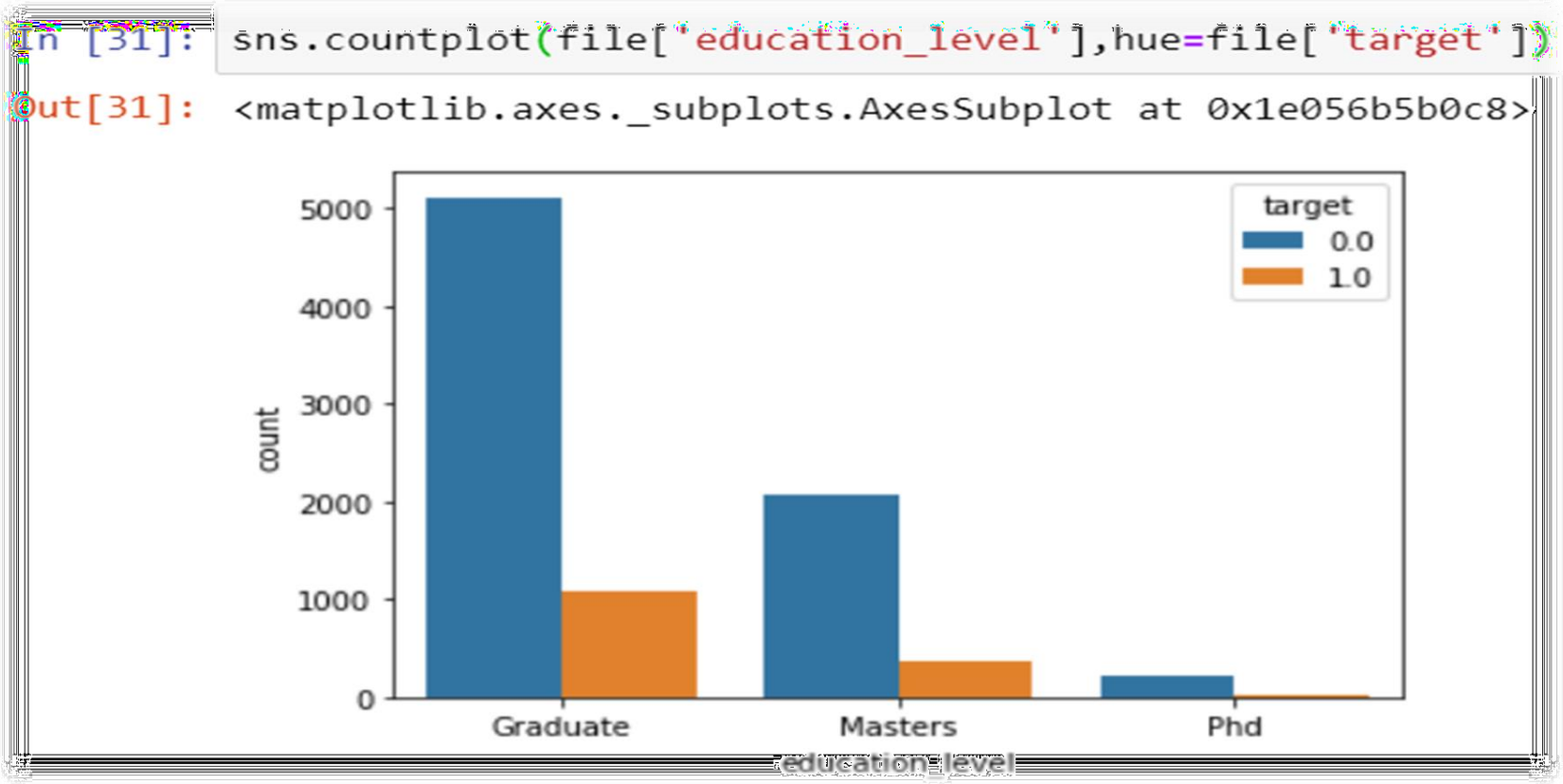
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x1e0566cd108>



❖ When we compare education level column with target column we conclude that :-

- 1.Graduates have very less chance of leaving
- Regardless of any education level , there is a very less chance of changing job

```
In [31]: sns.countplot(file['education_level'],hue=file['target'])
Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x1e056b5b0c8>
```

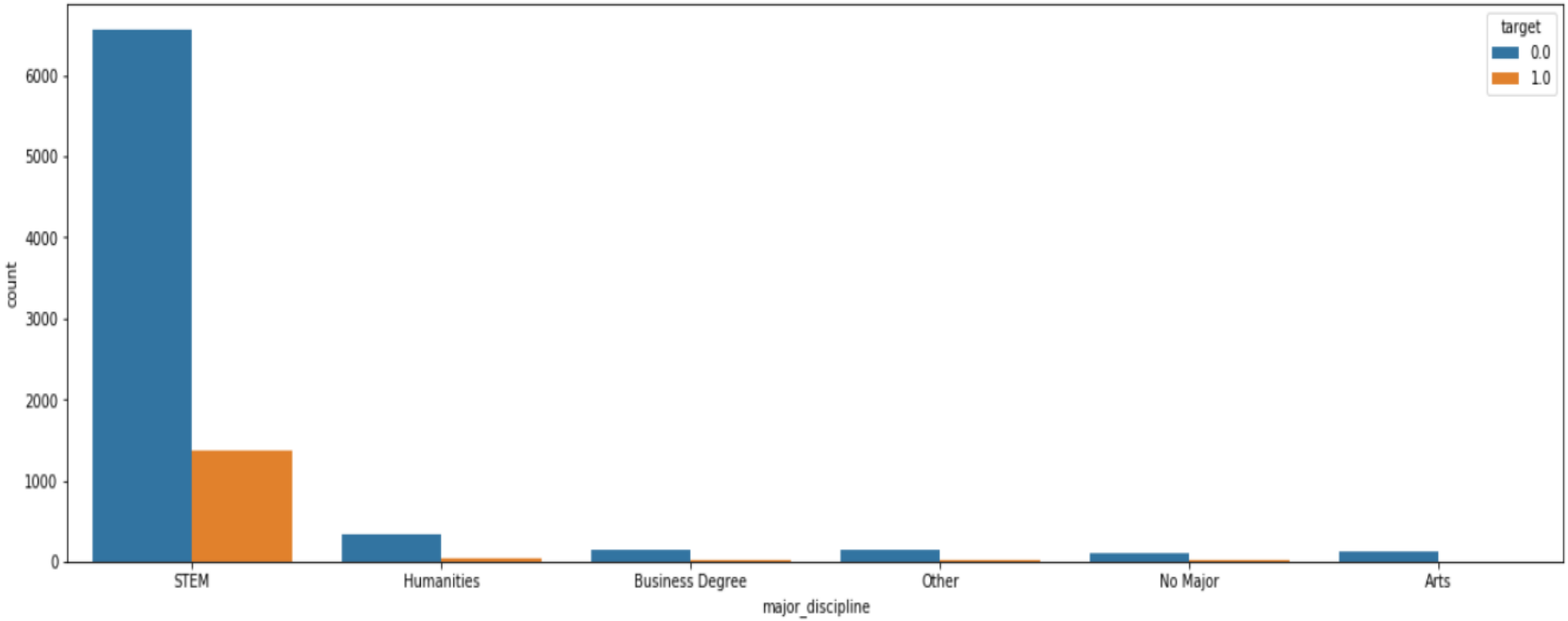


❖ When we compare major discipline column with target column we conclude that :-

- 1.Mostly people are STEM
- In STEM , people tend not to change the job
- People not tend to change job in any field

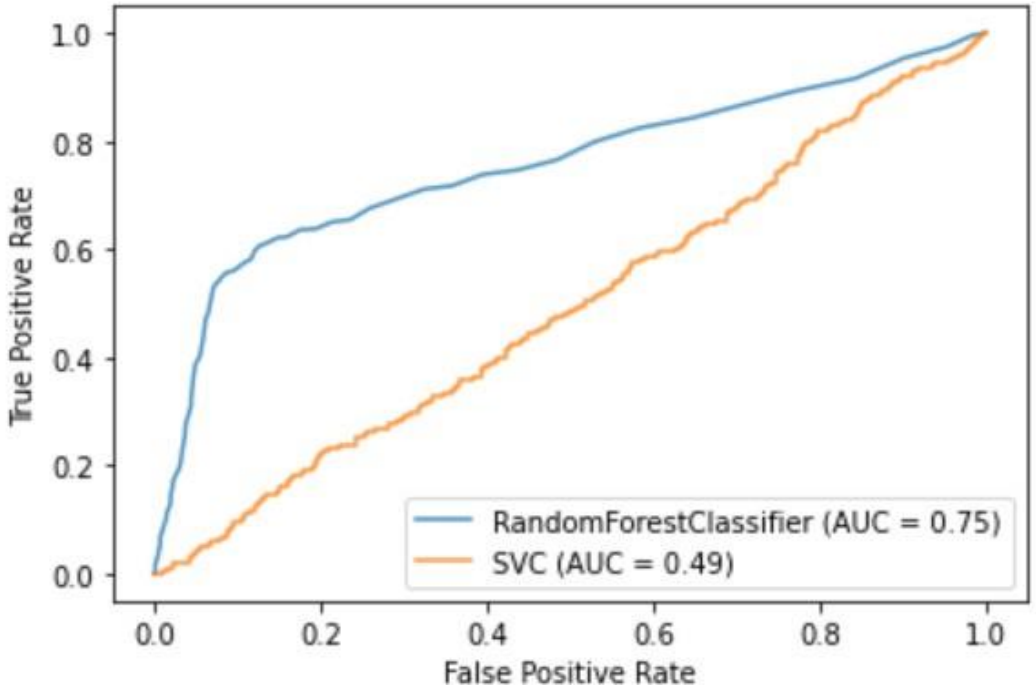
```
plt.figure(figsize=(20,6))
sns.countplot(file['major_discipline'],hue=file['target'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e056c07488>
```



❖ visualization of score(y_pred)_the relation between true positive rate and false positive rate

```
Out[305]: <sklearn.metrics._plot.roc_curve.RocCurveDisplay at 0x7efd4c2a5048>
```



Out[307]: <matplotlib.axes._subplots.AxesSubplot at 0x7efd4251d240>