In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
data = pd.read_csv(r'C:\Users\hp\Downloads\Chrome\lab2.csv')
data
```

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_di
0	32403	city_41	0.827	Male	Has relevent experience	Full time course	Graduate	STEM
1	9858	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM
2	31806	city_21	0.624	Male	No relevent experience	no_enrollment	High School	NaN
3	27385	city_13	0.827	Male	Has relevent experience	no_enrollment	Masters	STEM
4	27724	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
...	...	...	...	...	...	...	...	...
2124	1289	city_103	0.920	Male	No relevent experience	no_enrollment	Graduate	Humanities
2125	195	city_136	0.897	Male	Has relevent experience	no_enrollment	Masters	STEM
2126	31762	city_100	0.887	Male	No relevent experience	no_enrollment	Primary School	NaN
2127	7873	city_102	0.804	Male	Has relevent experience	Full time course	High School	NaN
2128	12215	city_102	0.804	Male	Has relevent experience	no_enrollment	Masters	STEM

2129 rows × 13 columns

In [3]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2129 entries, 0 to 2128
Data columns (total 13 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   enrollee_id             2129 non-null   int64  
 1   city                    2129 non-null   object 
 2   city_development_index  2126 non-null   float64
 3   gender                  1621 non-null   object 
 4   relevent_experience     2109 non-null   object 
 5   enrolled_university     2088 non-null   object 
 6   education_level         2066 non-null   object 
 7   major_discipline        1817 non-null   object 
 8   experience              2124 non-null   object 
 9   company_size            1507 non-null   object 
 10  company_type            1495 non-null   object 
 11  last_new_job            2089 non-null   object 
 12  training_hours          2109 non-null   float64
dtypes: float64(2), int64(1), object(10)
memory usage: 216.4+ KB
```

In [4]:
```
data.columns
```

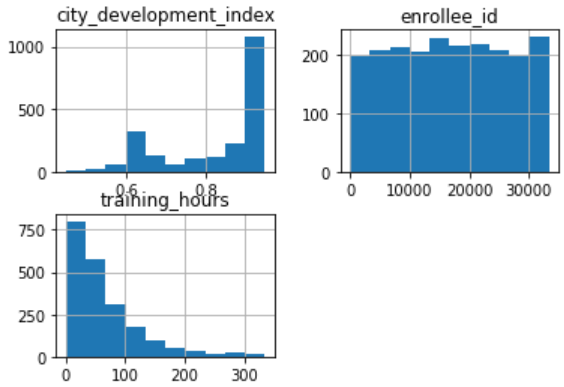
```
Index(['enrollee_id', 'city', 'city_development_index', 'gender',
       'relevent_experience', 'enrolled_university', 'education_level',
       'major_discipline', 'experience', 'company_size', 'company_type',
       'last_new_job', 'training_hours'],
      dtype='object')
```

In [5]:
```
data.isnull().sum()  #There are lots of null values in columns
```

```
enrollee_id               0
city                      0
city_development_index    3
gender                  508
relevent_experience      20
enrolled_university      41
education_level          63
major_discipline        312
experience                5
company_size            622
company_type            634
last_new_job             40
training_hours           20
dtype: int64
```

In [6]:
```
data.hist()
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000002340C1F0588>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000002340C69C408>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000002340C6D6208>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000002340C70F308>]],
      dtype=object)
```



In [7]:
```
data.values
```

```
array([[32403, 'city_41', 0.8270000000000001, ..., nan, '1', 21.0],
       [9858, 'city_103', 0.92, ..., 'Pvt Ltd', '1', 98.0],
       [31806, 'city_21', 0.624, ..., 'Pvt Ltd', 'never', 15.0],
       ...,
       [31762, 'city_100', 0.887, ..., 'Pvt Ltd', 'never', 18.0],
       [7873, 'city_102', 0.804, ..., 'Public Sector', '1', 84.0],
       [12215, 'city_102', 0.804, ..., 'Pvt Ltd', '2', 11.0]],
      dtype=object)
```

In [8]:
```
data['city'].head(20)
```

```
0      city_41
1     city_103
2      city_21
3      city_13
4     city_103
5      city_23
6      city_21
7     city_160
8     city_173
9      city_21
10    city_103
11     city_90
12     city_46
13     city_98
14    city_103
15     city_21
16     city_21
17     city_13
18     city_21
19     city_21
Name: city, dtype: object
```

In [12]:
```
data['gender'].count()
```

```
1621
```

In [14]:
```
data.shape
```

```
(2129, 13)
```

In [19]:
```
column = list(data.columns) # for operation ,we obtain of column in list
column
```

```
['enrollee_id',
 'city',
 'city_development_index',
 'gender',
 'relevent_experience',
 'enrolled_university',
 'education_level',
 'major_discipline',
 'experience',
 'company_size',
 'company_type',
 'last_new_job',
 'training_hours']
```

In [83]:
```
dict_data = {}  # fill all null values with backward fill  and store in new dictionary
for i in column :
    fill_col = data[i].fillna(method = 'bfill')
    a = fill_col.isnull().sum()
    print(i,a)
    dict_data[i]= fill_col
print(dict_data)
```

```
enrollee_id 0
city 0
city_development_index 0
gender 0
relevent_experience 0
enrolled_university 0
education_level 0
major_discipline 0
experience 0
company_size 0
company_type 0
last_new_job 0
training_hours 0
{'enrollee_id': 0       32403
1        9858
2       31806
3       27385
4       27724
        ...
2124     1289
2125      195
2126    31762
2127     7873
2128    12215
Name: enrollee_id, Length: 2129, dtype: int64, 'city': 0        city_41
1       city_103
2        city_21
3        city_13
4       city_103
          ...
2124    city_103
2125    city_136
2126    city_100
2127    city_102
2128    city_102
Name: city, Length: 2129, dtype: object, 'city_development_index': 0       0.827
1       0.920
2       0.624
3       0.827
4       0.920
        ...
2124    0.920
2125    0.897
2126    0.887
2127    0.804
2128    0.804
Name: city_development_index, Length: 2129, dtype: float64, 'gender': 0         Male
1       Female
2         Male
3         Male
4         Male
         ...
2124      Male
2125      Male
2126      Male
2127      Male
2128      Male
Name: gender, Length: 2129, dtype: object, 'relevent_experience': 0       Has relevent experience
1       Has relevent experience
2        No relevent experience
3       Has relevent experience
4       Has relevent experience
                ...
2124     No relevent experience
2125    Has relevent experience
2126     No relevent experience
2127    Has relevent experience
2128    Has relevent experience
Name: relevent_experience, Length: 2129, dtype: object, 'enrolled_university': 0       Full time course
1          no_enrollment
2          no_enrollment
3          no_enrollment
4          no_enrollment
              ...
2124       no_enrollment
2125       no_enrollment
```

```
2126       no_enrollment
2127    Full time course
2128       no_enrollment
Name: enrolled_university, Length: 2129, dtype: object, 'education_level': 0            Graduate
1              Graduate
2           High School
3                Masters
4              Graduate
              ...
2124           Graduate
2125            Masters
2126    Primary School
2127       High School
2128            Masters
Name: education_level, Length: 2129, dtype: object, 'major_discipline': 0            STEM
1              STEM
2              STEM
3              STEM
4              STEM
            ...
2124    Humanities
2125          STEM
2126          STEM
2127          STEM
2128          STEM
Name: major_discipline, Length: 2129, dtype: object, 'experience': 0        9
1         5
2        <1
3        11
4       >20
       ...
2124     16
2125     18
2126      3
2127      7
2128     15
Name: experience, Length: 2129, dtype: object, 'company_size': 0          <10
1        Oct-49
2        Oct-49
3        Oct-49
4        10000+
         ...
2124    100-500
2125    100-500
2126    100-500
2127    100-500
2128     10000+
Name: company_size, Length: 2129, dtype: object, 'company_type': 0            Pvt Ltd
1              Pvt Ltd
2              Pvt Ltd
3              Pvt Ltd
4              Pvt Ltd
             ...
2124    Public Sector
2125    Public Sector
2126          Pvt Ltd
2127    Public Sector
2128          Pvt Ltd
Name: company_type, Length: 2129, dtype: object, 'last_new_job': 0          1
1           1
2       never
3           1
4          >4
        ...
2124        4
2125        2
2126    never
2127        1
2128        2
Name: last_new_job, Length: 2129, dtype: object, 'training_hours': 0       21.0
1       98.0
2       15.0
3       39.0
4       72.0
        ...
2124    15.0
```

```
2125    30.0
2126    18.0
2127    84.0
2128    11.0
Name: training_hours, Length: 2129, dtype: float64}
```

In [84]:
```
new_data =pd.DataFrame(dict_data) # This is new DataFrame with filled all null values.
new_data
```

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_di
0	32403	city_41	0.827	Male	Has relevent experience	Full time course	Graduate	STEM
1	9858	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM
2	31806	city_21	0.624	Male	No relevent experience	no_enrollment	High School	STEM
3	27385	city_13	0.827	Male	Has relevent experience	no_enrollment	Masters	STEM
4	27724	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
...	...	...	...	...	...	...	...	...
2124	1289	city_103	0.920	Male	No relevent experience	no_enrollment	Graduate	Humanities
2125	195	city_136	0.897	Male	Has relevent experience	no_enrollment	Masters	STEM
2126	31762	city_100	0.887	Male	No relevent experience	no_enrollment	Primary School	STEM
2127	7873	city_102	0.804	Male	Has relevent experience	Full time course	High School	STEM
2128	12215	city_102	0.804	Male	Has relevent experience	no_enrollment	Masters	STEM

2129 rows × 13 columns

In [48]:
```
data['company_type'].values
```

```
array([nan, 'Pvt Ltd', 'Pvt Ltd', ..., 'Pvt Ltd', 'Public Sector',
       'Pvt Ltd'], dtype=object)
```

In [23]:
```
data.isnull().sum()
```

```
enrollee_id               0
city                      0
city_development_index    3
gender                  508
relevent_experience      20
enrolled_university      41
education_level          63
major_discipline        312
experience                5
company_size            622
company_type            634
last_new_job             40
training_hours           20
dtype: int64
```

In [40]:
```
data['gender'].isnull().sum()
```

```
508
```

In [42]:
```
gender = data['gender'].fillna(method= 'ffill')
gender
```

```
0         Male
1       Female
2         Male
3         Male
4         Male
         ...
2124      Male
2125      Male
2126      Male
2127      Male
2128      Male
Name: gender, Length: 2129, dtype: object
```

In [85]:
```
new_data.isnull().any() # this is proof of Not null values in new Dataframe
```

```
enrollee_id             False
city                    False
city_development_index  False
gender                  False
relevent_experience     False
enrolled_university     False
education_level         False
major_discipline        False
experience              False
company_size            False
company_type            False
last_new_job            False
training_hours          False
dtype: bool
```

In [89]:
```
new_data.isnull().sum() # sum of Null values 's columns
```

```
enrollee_id             0
city                    0
city_development_index  0
gender                  0
relevent_experience     0
enrolled_university     0
education_level         0
major_discipline        0
experience              0
company_size            0
company_type            0
last_new_job            0
training_hours          0
dtype: int64
```

filled all null values . below new DataFrame with filled

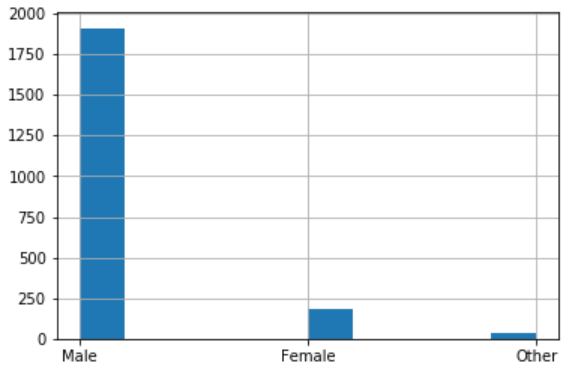
In [91]:
```
new_data.head(10)
```

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_disci
0	32403	city_41	0.827	Male	Has relevent experience	Full time course	Graduate	STEM
1	9858	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM
2	31806	city_21	0.624	Male	No relevent experience	no_enrollment	High School	STEM
3	27385	city_13	0.827	Male	Has relevent experience	no_enrollment	Masters	STEM
4	27724	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
5	217	city_23	0.899	Male	No relevent experience	Part time course	Masters	STEM
6	21465	city_21	0.624	Female	Has relevent experience	no_enrollment	Graduate	STEM
7	27302	city_160	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM
8	12994	city_173	0.878	Male	Has relevent experience	no_enrollment	Graduate	STEM
9	16287	city_21	0.624	Male	Has relevent experience	Full time course	Graduate	Other

EDA

In [93]:
```
new_data['gender'].hist() # Male population is high
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2340cc5fb88>
```



In [103]:
```
new_data[new_data['gender']=='Other'].count() # there are 34 other in gender category , not identi
fy
```

```
enrollee_id             34
city                    34
city_development_index  34
gender                  34
relevent_experience     34
enrolled_university     34
education_level         34
major_discipline        34
experience              34
company_size            34
company_type            34
last_new_job            34
training_hours          34
dtype: int64
```

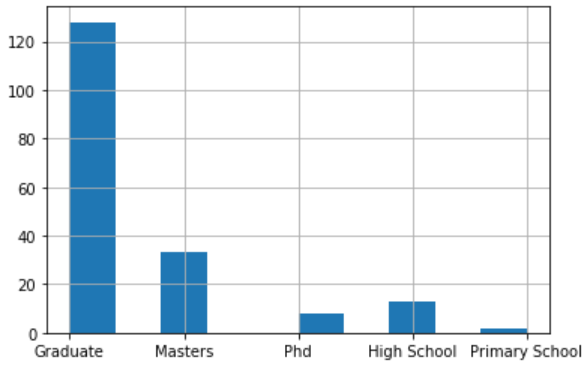
In [105]:
```
Female_data = new_data[new_data['gender']=='Female']
Female_data[Female_data['education_level']=='Graduate']
```

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_di
1	9858	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM
6	21465	city_21	0.624	Female	Has relevent experience	no_enrollment	Graduate	STEM
7	27302	city_160	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM
25	19139	city_103	0.920	Female	Has relevent experience	Part time course	Graduate	STEM
41	25855	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	Arts
...	...	...	...	...	...	...	...	...
2039	26950	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM
2047	13292	city_103	0.920	Female	Has relevent experience	Full time course	Graduate	STEM
2051	22037	city_61	0.913	Female	Has relevent experience	no_enrollment	Graduate	STEM
2052	29304	city_19	0.682	Female	Has relevent experience	no_enrollment	Graduate	STEM
2122	24507	city_90	0.698	Female	No relevent experience	no_enrollment	Graduate	STEM

128 rows × 13 columns

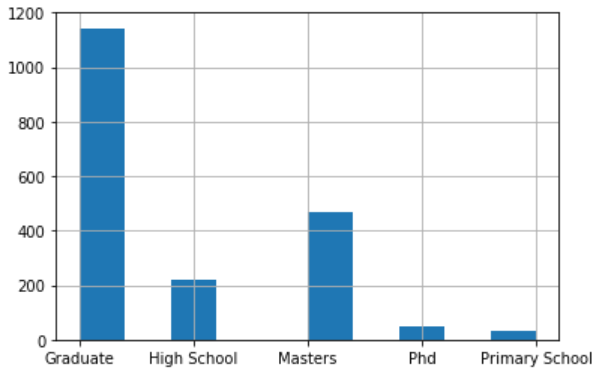
In [108]:
```
Female_data['education_level'].hist() # female data , lots of Female has Graduate degree,next Master
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2340dd4a148>
```



In [111]:
```
Male_data = new_data[new_data['gender']=='Male']
Male_data['education_level'].hist() # Graduate Male candidate is high
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2340deb8c48>
```

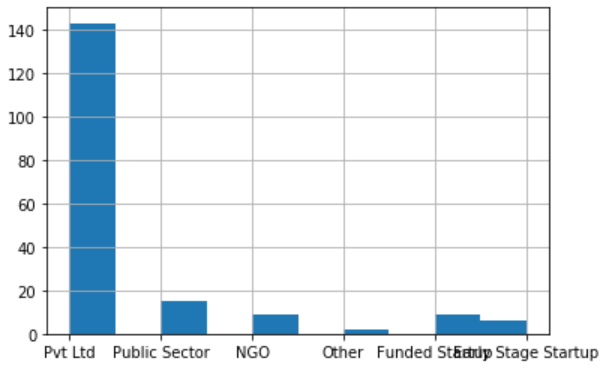


In [115]:
```
Female_data.groupby('company_type').count() # high population involved in PVT sector
```

company_type	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	n
Early Stage Startup	6	6	6	6	6	6	6	6
Funded Startup	9	9	9	9	9	9	9	9
NGO	9	9	9	9	9	9	9	9
Other	2	2	2	2	2	2	2	2
Public Sector	15	15	15	15	15	15	15	1
Pvt Ltd	143	143	143	143	143	143	143	1

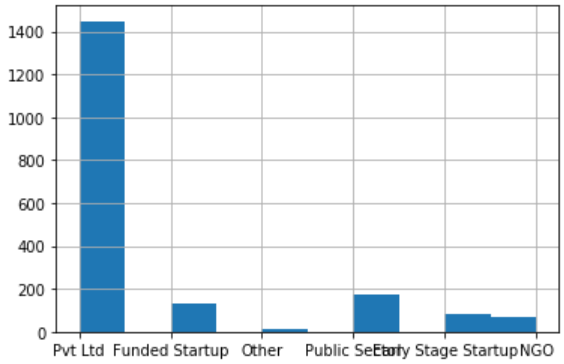
In [116]:
```
Female_data['company_type'].hist() # Female candidate also involved in pvt.sector.
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2340e01e448>
```



In [117]:
```
Male_data['company_type'].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2340e098a08>
```

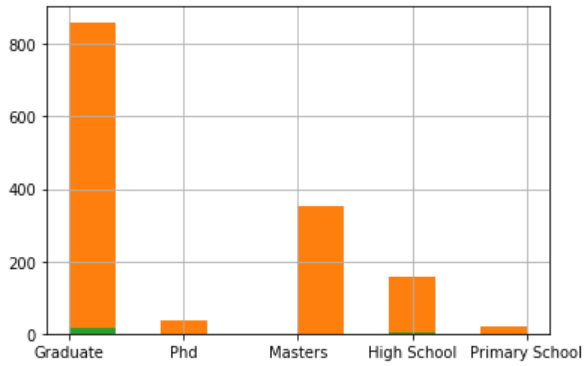


In [120]:
```
new_data.groupby('gender').count()
```

gender	enrollee_id	city	city_development_index	relevent_experience	enrolled_university	education_level	major_discipline
Female	184	184	184	184	184	184	184
Male	1911	1911	1911	1911	1911	1911	1911
Other	34	34	34	34	34	34	34

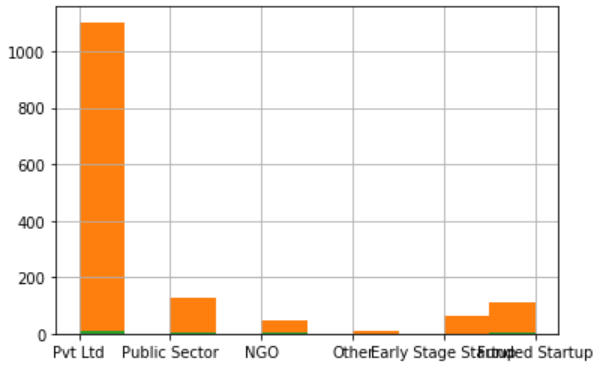
In [130]:
```
data.groupby('gender')['education_level'].hist() # this is group by divided Education level with gender
```

```
gender
Female    AxesSubplot(0.125,0.125;0.775x0.755)
Male      AxesSubplot(0.125,0.125;0.775x0.755)
Other     AxesSubplot(0.125,0.125;0.775x0.755)
Name: education_level, dtype: object
```



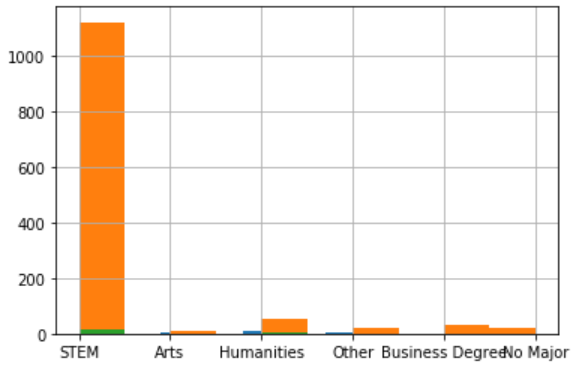
In [131]:
```
data.groupby('gender')['company_type'].hist() # group by comany type with gender
```

```
gender
Female    AxesSubplot(0.125,0.125;0.775x0.755)
Male      AxesSubplot(0.125,0.125;0.775x0.755)
Other     AxesSubplot(0.125,0.125;0.775x0.755)
Name: company_type, dtype: object
```



In [132]:
```
data.groupby('gender')['major_discipline'].hist()
```

```
gender
Female    AxesSubplot(0.125,0.125;0.775x0.755)
Male      AxesSubplot(0.125,0.125;0.775x0.755)
Other     AxesSubplot(0.125,0.125;0.775x0.755)
Name: major_discipline, dtype: object
```



mean ,standard deviation of training hours data

In [138]:
```
new_data['training_hours'].describe()  # till 75% of data  under 86. # standard deviation is 60.
```

```
count    2129.000000
mean       64.904180
std        60.032629
min         1.000000
25%        24.000000
50%        47.000000
75%        86.000000
max       334.000000
Name: training_hours, dtype: float64
```

In [143]:
```
new_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2129 entries, 0 to 2128
Data columns (total 13 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   enrollee_id             2129 non-null   int64  
 1   city                    2129 non-null   object 
 2   city_development_index  2129 non-null   float64
 3   gender                  2129 non-null   object 
 4   relevent_experience     2129 non-null   object 
 5   enrolled_university     2129 non-null   object 
 6   education_level         2129 non-null   object 
 7   major_discipline        2129 non-null   object 
 8   experience              2129 non-null   object 
 9   company_size            2129 non-null   object 
 10  company_type            2129 non-null   object 
 11  last_new_job            2129 non-null   object 
 12  training_hours          2129 non-null   float64
dtypes: float64(2), int64(1), object(10)
memory usage: 216.4+ KB
```

In [144]:
```
new_data
```

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_di
0	32403	city_41	0.827	Male	Has relevent experience	Full time course	Graduate	STEM
1	9858	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM
2	31806	city_21	0.624	Male	No relevent experience	no_enrollment	High School	STEM
3	27385	city_13	0.827	Male	Has relevent experience	no_enrollment	Masters	STEM
4	27724	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM
...	...	...	...	...	...	...	...	...
2124	1289	city_103	0.920	Male	No relevent experience	no_enrollment	Graduate	Humanities
2125	195	city_136	0.897	Male	Has relevent experience	no_enrollment	Masters	STEM
2126	31762	city_100	0.887	Male	No relevent experience	no_enrollment	Primary School	STEM
2127	7873	city_102	0.804	Male	Has relevent experience	Full time course	High School	STEM
2128	12215	city_102	0.804	Male	Has relevent experience	no_enrollment	Masters	STEM

2129 rows × 13 columns

note - Small observation during Analysis

In [ ]:
```
# Most of the Males and Female are involed in Pvt.sector .
#  They have graduation degree more in Both Male and Female.
# Mean values is arond 60 in training hours
# Many graduate people has less than 1 experience.
## till 75% of data  under 86. # standard deviation is 60.
# Stem is hightest taking discpline.
# In many case , high school candidate has obtain more than 84 traning hours completed.
```