The Dataset is from kaggle and this data is designed by a company who wants to hire data scientists and it makes trainings and it would like to know who from the candidates who successfully passed some of these trainings will work for it after completing the trainings as this will reduce cost and much time. dataset consists of 14 column and all these columns are information about those people by using this data we will be able to predict the possibility of those people to work for this company or they will look for new job in other company

Columns:

- Enrollee_id: unique number given to every candidate
- City: code of the city of the candidates
- City_development_index : the development index of the candidates cities
- Gender: male or female
- Relevent_experience: relevant experience of the candidates
- enrolled_university : type university courses the candidate enrolled it if exist
- education_level
- experience : years of experience
- company_size: the number of employess who in their current company in which they worked
- company_type : The type of the current company of the candidate
- last_new_job: the difference between the previous and the current job in years
- training_hours: number of trainings hours the candidate completed in the company
- target: consists of 2 states (0,1) where 0 means that the candidate is not looking for a job change and 1 means that the candidate is looking for a job change

Cleaning

Quality :

We drop null values in all table using dropna()

```
<'class 'pandas.core.frame.DataFrame>
Int64Index: 8877 entries, 1 to 19155
:Data columns (total 14 columns)
   Column                  Non-Null Count  Dtype   #
   -----  --------------                  ------  ---
  enrollee_id             8877 non-null   object   0
  city                    8877 non-null   object   1
 city_development_index  8877 non-null   float64   2
gender                  8877 non-null   category   3
  relevent_experience     8877 non-null   object   4
  enrolled_university     8877 non-null   object   5
  education_level         8877 non-null   object   6
  major_discipline        8877 non-null   object   7
  experience              8877 non-null   object   8
  company_size            8877 non-null   object   9
  company_type            8877 non-null   object  10
  last_new_job            8877 non-null   object  11
   training_hours          8877 non-null   int64  12
 target                  8877 non-null   float64  13
dtypes: category(1), float64(2), int64(1), object(10)
memory usage: 979.7+ KB
```

1- enrollee_id

the id must be a fixed number we will make it 5 numbers

we change the value of enrolle_id column to string using astype(str) and we make it fixed number using str.padstr(5,fillchar='0') to make it 5 numbers and complete the other by zeros

```
test_clean.enrollee_id=test_clean.enrollee_id.astype(str).str.pad(5,fillchar='0')
test_clean.enrollee_id.value_counts()
```

```
1    02061
1    21730
1    28134
1    09028
1    25693
..
1    02252
1    30530
1    28021
1    06595
1    31688
Name: enrollee_id, Length: 19158, dtype: int64
```

2- Gender (2 issues)

we will change the type to category using astype('category')

and we will remove the ‘other’ values to be only:

-male

-female

```
test_clean=test_clean[test_clean.gender!='Other']
test_clean.gender.value_counts()
```

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

```
test_clean.info()
```

```
<'class 'pandas.core.frame.DataFrame>
Int64Index: 8877 entries, 1 to 19155
:Data columns (total 14 columns)
   Column                  Non-Null Count  Dtype    #
   -----  --------------                   ------   ---
  enrollee_id             8877 non-null   object   0
  city                    8877 non-null   object   1
 city_development_index  8877 non-null   float64  2
gender                  8877 non-null   category  3
  relevent_experience     8877 non-null   object   4
  enrolled_university     8877 non-null   object   5
  education_level         8877 non-null   object   6
  major_discipline        8877 non-null   object   7
  experience              8877 non-null   object   8
  company_size            8877 non-null   object   9
  company_type            8877 non-null   object  10
  last_new_job            8877 non-null   object  11
   training_hours          8877 non-null   int64   12
 target                  8877 non-null   float64  13
dtypes: category(1), float64(2), int64(1), object(10)
memory usage: 979.7+ KB
```

3-relevant_experience (1 issue)

we will change the type to category using astype('category')

- Has relevent experience

- No relevent experience

```
test_clean.info()
```

```
<'class 'pandas.core.frame.DataFrame>
Int64Index: 8877 entries, 1 to 19155
:Data columns (total 14 columns)
   Column                  Non-Null Count  Dtype   #
   -----  --------------                   ------  ---
  enrollee_id              8877 non-null   object   0
  city                     8877 non-null   object   1
 city_development_index  8877 non-null   float64   2
gender                   8877 non-null   category   3
relevent_experience      8877 non-null   category   4
  enrolled_university      8877 non-null   object   5
  education_level          8877 non-null   object   6
  major_discipline         8877 non-null   object   7
  experience               8877 non-null   object   8
  company_size             8877 non-null   object   9
  company_type             8877 non-null   object  10
  last_new_job             8877 non-null   object  11
   training_hours           8877 non-null    int64  12
 target                   8877 non-null   float64  13
dtypes: category(2), float64(2), int64(1), object(9)
memory usage: 919.1+ KB
```

4-enrolled_university (1 issue)

we will change the type to category using astype('category')

- no_enrollment
- Full time course
- Part time course

```
test_clean.info()
```

```
<'class 'pandas.core.frame.DataFrame>
Int64Index: 8877 entries, 1 to 19155
:Data columns (total 14 columns)
   Column                  Non-Null Count  Dtype   #
   -----  --------------                   ------  ---
  enrollee_id              8877 non-null   object   0
  city                     8877 non-null   object   1
 city_development_index  8877 non-null   float64   2
gender                   8877 non-null   category   3
relevent_experience      8877 non-null   category   4
enrolled_university      8877 non-null   category   5
  education_level          8877 non-null   object   6
  major_discipline         8877 non-null   object   7
  experience               8877 non-null   object   8
  company_size             8877 non-null   object   9
  company_type             8877 non-null   object  10
  last_new_job             8877 non-null   object  11
   training_hours           8877 non-null    int64  12
 target                   8877 non-null   float64  13
dtypes: category(3), float64(2), int64(1), object(8)
memory usage: 858.5+ KB
```

5-education_level (1 issue)

we will change the type to category using astype('category')

- Graduate
- masters
- Phd

```
test_clean.info()
```

```
<'class 'pandas.core.frame.DataFrame>
Int64Index: 8877 entries, 1 to 19155
:Data columns (total 15 columns)
   Column                  Non-Null Count  Dtype   #
   -----  --------------                  ------  ---
  enrollee_id             8877 non-null   object   0
  city                    8877 non-null   object   1
 city_development_index  8877 non-null   float64   2
gender                   8877 non-null   category   3
relevent_experience      8877 non-null   category   4
enrolled_university      8877 non-null   category   5
education_level          8877 non-null   category   6
major_discipline         8877 non-null   category   7
   experience               8877 non-null   int32   8
company_type             8877 non-null   category   9
   last_new_job             8877 non-null   int32  10
   training_hours           8877 non-null   int64  11
  target                   8877 non-null   object  12
   company_size_from        8877 non-null   int32  13
   company_size_to          8877 non-null   int32  14
dtypes: category(6), float64(1), int32(4), int64(1), object(3)
memory usage: 607.6+ KB
```

6- major_discipline (1 issue)

we change the type to category using astype('category')

- STEM
- humanities
- Business
- Degree Arts
- No Major

```
test_clean.info()
```

```
<'class 'pandas.core.frame.DataFrame>
Int64Index: 8877 entries, 1 to 19155
:Data columns (total 15 columns)
   Column                  Non-Null Count  Dtype   #
   -----  --------------                   ------  ---
  enrollee_id             8877 non-null   object   0
  city                    8877 non-null   object   1
 city_development_index  8877 non-null   float64   2
gender                   8877 non-null   category   3
relevent_experience      8877 non-null   category   4
enrolled_university      8877 non-null   category   5
education_level          8877 non-null   category   6
major_discipline         8877 non-null   category   7
   experience               8877 non-null   int32   8
company_type             8877 non-null   category   9
   last_new_job             8877 non-null   int32  10
   training_hours           8877 non-null   int64  11
  target                   8877 non-null   object  12
   company_size_from        8877 non-null   int32  13
   company_size_to          8877 non-null   int32  14
dtypes: category(6), float64(1), int32(4), int64(1), object(3)
memory usage: 607.6+ KB
```

7- company_type (2 issues)

- Full state names sometimes, abbreviations other times

  we will change using The function replace()

  Pvt Ltd to **private limited company**

  NGO to **Non-Governmental Organisation**

- we will change the type to category using astype('category')

- private limited company

- Funded Startup

- Public Sector

- Early Stage Startup

- Non-Governmental Organization

```
test_clean.company_type=test_clean.company_type.replace('Pvt Ltd','private limited company')
test_clean.company_type=test_clean.company_type.replace('NGO','Non-Governmental Organisation')
test_clean.company_type.value_counts()
```

```
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
```

```
test_clean.info()
```

```
<'class 'pandas.core.frame.DataFrame>
Int64Index: 8877 entries, 1 to 19155
:Data columns (total 14 columns)
   Column                  Non-Null Count  Dtype   #
   -----  --------------                  ------  ---
  enrollee_id             8877 non-null   object   0
  city                    8877 non-null   object   1
 city_development_index  8877 non-null   float64   2
gender                   8877 non-null   category   3
relevent_experience      8877 non-null   category   4
enrolled_university      8877 non-null   category   5
  education_level          8877 non-null   object   6
  major_discipline         8877 non-null   object   7
   experience               8877 non-null   int32   8
  company_size             8877 non-null   object   9
company_type             8877 non-null   category  10
   last_new_job             8877 non-null   int32  11
   training_hours           8877 non-null   int64  12
 target                  8877 non-null   float64  13
dtypes: category(4), float64(2), int32(2), int64(1), object(5)
memory usage: 728.7+ KB
```

8-Experience (2 issues)

-must be integer and handle the > and < signs and we will handle these signs by making >**20** To **21** and **<1** to **0** By using replace()

-change the type using the function astype(int)

```
test_clean.experience.value_counts()
```

```
1868    21
571      5
545     10
533      6
531      9
480      7
474      4
427      3
397      8
393     15
367     11
333     14
298     16
294      2
290     12
230     13
206     17
176     19
166     18
120      1
93       0
85      20
Name: experience, dtype: int64
```

9-last_new_job (2 issues)

-we will change **never** to **0** and **4<** to **5** using replace()

-we will change the type into integer using astype(int)

```
test_clean.last_new_job=test_clean.last_new_job.replace('>4','5')
test_clean.last_new_job=test_clean.last_new_job.replace('never','0')
test_clean.last_new_job.value_counts()
```

```
3798    1
1949    5
1561    2
604     3
594     4
371     0
Name: last_new_job, dtype: int64
```

```
test_clean.last_new_job=test_clean.last_new_job.astype(int)
```

10- target(1 issue)

-We will change the type of target from **float** into **integer** using astype(int)

```
test_clean.target=test_clean.target.astype(int)
test_clean.info()
```

```
<'class 'pandas.core.frame.DataFrame>
Int64Index: 8877 entries, 1 to 19155
:Data columns (total 14 columns)
   Column                   Non-Null Count  Dtype   #
   -----  --------------                   ------  ---
  enrollee_id              8877 non-null   object   0
  city                     8877 non-null   object   1
 city_development_index  8877 non-null   float64   2
gender                   8877 non-null   category   3
relevent_experience      8877 non-null   category   4
enrolled_university      8877 non-null   category   5
education_level          8877 non-null   category   6
major_discipline         8877 non-null   category   7
   experience               8877 non-null   int32   8
  company_size             8877 non-null   object   9
company_type             8877 non-null   category  10
   last_new_job             8877 non-null   int32  11
   training_hours           8877 non-null   int64  12
   target                   8877 non-null   int32  13
dtypes: category(6), float64(1), int32(3), int64(1), object(3)
memory usage: 572.9+ KB
```

Tidiness :

1-company_size

we will change the value **10000+** to **10000-99999** and **10/49** to **10-49** and **<10** to **1-10**

using replace()

we will make two columns company_size_from and company_size_to and split company_size by using str.split('-',1).str

and we change the two columns type to integer using astype(int)

and we will delete the company_size using drop('company_size',axis=1)

```
test_clean.company_size_from=test_clean.company_size_from.astype(int)
test_clean.company_size_to=test_clean.company_size_to.astype(int)
test_clean=test_clean.drop('company_size',axis=1)
test_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8877 entries, 0 to 8876
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype   
---  ------                  --------------  -----   
 0   enrollee_id             8877 non-null   object  
 1   city                    8877 non-null   object  
 2   city_development_index  8877 non-null   float64 
 3   gender                  8877 non-null   category
 4   relevent_experience     8877 non-null   category
 5   enrolled_university     8877 non-null   category
 6   education_level         8877 non-null   category
 7   major_discipline        8877 non-null   category
 8   experience              8877 non-null   int32   
 9   company_type            8877 non-null   category
 10  last_new_job            8877 non-null   int32   
 11  training_hours          8877 non-null   int64   
 12  target                  8877 non-null   int32   
 13  company_size_from       8877 non-null   int32   
 14  company_size_to         8877 non-null   int32   
dtypes: category(6), float64(1), int32(5), int64(1), object(2)
memory usage: 572.9+ KB
```