Annapoornima S

225229101

Lab 6: Pandas Data Cleaning

```
In [1]: import pandas as pd
    df = pd.read_csv("train_hr.csv")
          df.head(10)
 Out[1]:
                                                                                                                                             KPIs_met >80%
              employee_id department
                                        region education gender recruitment_channel no_of_trainings age previous_year_rating length_of_service
                                                                                                                                                       awards
                              Sales &
                                                Master's &
           0
                    65438
                                       region_7
                                                                            sourcing
                                                                                                 1 35
                                                                                                                        5.0
                                                                                                                                          8
                                                                                                                                                    1
                            Marketing
                    65141
                            Operations
                                     region_22 Bachelor's
                                                                               other
                                                                                                     30
                                                                                                                        5.0
                                                                                                                                          4
                                                                                                                                                    0
                              Sales &
                                      region_19 Bachelor's
                     7513
                                                                            sourcing
                                                                                                                        3.0
                                                                                                                                                    0
                                                              m
                                                                                                 1
                             Marketing
                              Sales &
           3
                     2542
                                      region_23 Bachelor's
                                                              m
                                                                               other
                                                                                                 2
                                                                                                     39
                                                                                                                        1.0
                                                                                                                                          10
                                                                                                                                                    0
                             Marketing
           4
                    48945
                           Technology
                                     region_26 Bachelor's
                                                                               other
                                                                                                     45
                                                                                                                        3.0
                                                                                                                                          2
                                                                                                                                                    0
           5
                    58896
                             Analytics
                                       region_2 Bachelor's
                                                                            sourcing
                                                                                                 2
                                                                                                     31
                                                                                                                        3.0
                                                                                                                                          7
                                                                                                                                                    0
                    20379
                            Operations
                                     region_20 Bachelor's
                                                                                                     31
                                                                                                                        3.0
                                                                                                                                          5
                                                                                                                                                    0
                                                Master's &
                    16290
                            Operations region_34
                                                              m
                                                                            sourcing
                                                                                                     33
                                                                                                                        3.0
                                                                                                                                          6
                                                                                                                                                    0
                                                   above
           8
                    73202
                             Analytics region 20 Bachelor's
                                                                               other
                                                                                                     28
                                                                                                                        4.0
                                                                                                                                          5
                                                                                                                                                    0
                                                              m
                                                                                                 1
                              Sales &
                                                Master's &
                                       region_1
                    28911
                                                              m
                                                                            sourcing
                                                                                                 1 32
                                                                                                                        5.0
                                                                                                                                          5
                                                                                                                                                    1
                            Marketing
In [21]: column_names = df.columns
          print(column_names)
          df.dtypes
          for i in column names:
               print("{} is unique : {}".format(i,df[i].is_unique))
          Index(['department', 'region', 'education', 'gender', 'recruitment_channel',
                   'no_of_trainings', 'age', 'awards_won?', 'avg_training_score',
                   'is_promoted'],
                 dtype='object')
          department is unique : False
          region is unique : False
          education is unique : False
          gender is unique : False
          recruitment_channel is unique : False
          no_of_trainings is unique : False
          age is unique : False
          awards_won? is unique : False
          avg_training_score is unique : False
          is_promoted is unique : False
 In [3]: df.index.values
 Out[3]: array([
                                       2, ..., 54805, 54806, 54807], dtype=int64)
 In [4]: 0 in df.index.values
 Out[4]: True
 In [5]: df.set_index("employee_id",inplace=True)
```

:	department	region	education	gender	recruitment_channel	no_of_training	gs age	previou	s_year_rating	length_of_service	KPIs_me >80%	t awar
employee_id												
65438	Sales & Marketing	region_7	Master's 8 above		sourcing		1 35		5.0	8		1
65141	Operations	region_22	Bachelor's	m m	other		1 30		5.0	4)
7513	Sales & Marketing	region_19	Bachelor's	m m	sourcing		1 34		3.0	7)
2542	Sales & Marketing	region_23	Bachelor's	m m	other		2 39		1.0	10)
48945	Technology	region_26	Bachelor's	m m	other		1 45		3.0	2)
3030	Technology	region_14	Bachelor's	m m	sourcing		1 48		3.0	17)
74592	Operations	region_27	Master's 8 above		other		1 37		2.0	6)
13918	Analytics	region_1	Bachelor's	m m	other		1 27		5.0	3		1
13614	Sales & Marketing	region_9	NaN	m	sourcing		1 29		1.0	2)
51526	HR	region 22	Bachelor's	m	other		1 27		1.0	5)
54808 rows >		Lumn_nam	es[i] for	' i in [8	,9,10]]							
4	_drop = [col				,9,10]]							
: columns_to_	_drop = [col				,9,10]]							
columns_to_ df.drop(col df	_drop = [col	op, inpla		axis=1)	,9,10]]	nt_channel no	o_of_train	nings a	ge awards_w	on? avg_training_	score is_	promot
<pre>columns_to_ df.drop(col df employee_id</pre>	_drop = [co. Lumns_to_dro departm	op, inpla	ace=True	axis=1)	gender recruitme		o_of_train					promot
columns_to_ df.drop(col df employee_id 65438	_drop = [co. Lumns_to_dro departn Sales & Marke	op, inpla	egion ion_7 Mas	education	gender recruitme	sourcing	o_of_trair	1 :	35	0	49	promot
: columns_to_ : df.drop(col : df : employee_id 65438 65141	_drop = [co. Lumns_to_dro departn Sales & Marko Operat	nent resting regions regions	egion ion_7 Mas	education ter's & above	gender recruitme	sourcing other	o_of_train	1 ;	35	0	49 60	promot
: columns_to_ : df.drop(col : df :	_drop = [co. Lumns_to_dro departn Sales & Marke Operat	op, inpla	egion ion_7 Mas ion_22 ion_19	education ter's & above Bachelor's	gender recruitme	sourcing other sourcing	o_of_trair	1 ; 1 ; 1 ;	35 30 34	0 0 0	49 60 50	promot
<pre>columns_to_ df.drop(col df employee_id 65438 65141 7513 2542</pre>	_drop = [co. Lumns_to_dro departm Sales & Marke Operat Sales & Marke Sales & Marke	pp, inpla	ace=True	education ter's & above Bachelor's Bachelor's	gender recruitme f m m m	sourcing other sourcing other	o_of_train	1 : 1 : 1 : 2 : 1	35 30 34	0 0 0 0	49 60 50 50	promot
df.drop(col df: employee_id 65438 65141 7513 2542 48945	_drop = [co. Lumns_to_dro departn Sales & Marke Operal Sales & Marke Sales & Marke Techno	pp, inplained regions	egion ion_7 Mas ion_22 in_19 in_23 in_26	education ter's & above Bachelor's Bachelor's Bachelor's	gender recruitme	sourcing other sourcing other	o_of_traiı	1 : 1 : 1 : 2 : 1 : 4	35 30 34 39	0 0 0 0	49 60 50 50 73	promot
: columns_to_ : df.drop(col : df : employee_id 65438 65141 7513 2542 48945	_drop = [co. Lumns_to_dro departn Sales & Marko Operat Sales & Marko Sales & Marko Techno	pp, inplainment regions regions regions regions regions regions regions regions regions	egion ion_7 Mas ion_22 in_19 in_23 in_26	education ter's & above Bachelor's Bachelor's Bachelor's	gender recruitme	sourcing other sourcing other other	o_of_trair	1 : 1 : 1 : 2 : 1 :	35 30 34 39 45	0 0 0 0 0	49 60 50 50 73	promot
: columns_to_ : df.drop(col : df : employee_id 65438 65141 7513 2542 48945 3030	_drop = [co. Lumns_to_dro departm Sales & Marke Operat Sales & Marke Sales & Marke Techno	pp, inplacement regions region	ace=True, agion agion agion agion_7 Mas an_22 an_19 an_23 an_26 an_14	education ter's & above Bachelor's Bachelor's Bachelor's Bachelor's	gender recruitme	sourcing other sourcing other other sourcing	o_of_train	1 : 1 : 2 : 1 : 4 :	35 30 34 39 45 	0 0 0 0 0	49 60 50 50 73 	promot
: columns_to_ : df.drop(col : df : employee_id 65438 65141 7513 2542 48945 3030 74592	_drop = [co. Lumns_to_dro departn Sales & Marke Operal Sales & Marke Techno Techno Operal	nent regions	ace=True	education ter's & above Bachelor's Bachelor's Bachelor's Bachelor's education	gender recruitme	sourcing other sourcing other other sourcing other	o_of_train	1 : 1 : 1 : 2 : 1 : · · · · · · · · · · · · · · · · ·	35 30 34 39 45 	0 0 0 0 0 	49 60 50 50 73 78 56	promot
df.drop(col df: employee_id 65438 65141 7513 2542 48945 3030 74592 13918	_drop = [co. Lumns_to_dro departm Sales & Marke Operat Sales & Marke Sales & Marke Techno	pp, inplainment regions region	ace=True, agion agion agion agion_7 Mas an_22 an_19 an_23 an_26 an_14	education ter's & above Bachelor's Bachelor's Bachelor's Bachelor's	gender recruitme	sourcing other sourcing other other sourcing	o_of_train	1 : 1 : 2 : 1 :	35 30 34 39 45 	0 0 0 0 0	49 60 50 50 73 	promot

54808 rows × 10 columns

```
In [10]: df['department'] = df['department'].fillna(' ')
           df
Out[10]:
                                department
                                               region
                                                             education gender recruitment_channel no_of_trainings age awards_won? avg_training_score is_promoted
            employee id
                   65438
                          Sales & Marketing
                                             region_7
                                                       Master's & above
                                                                             f
                                                                                            sourcing
                                                                                                                   1
                                                                                                                       35
                                                                                                                                       0
                                                                                                                                                          49
                                                                                                                                                                         0
                   65141
                                                                                                                       30
                                                                                                                                       0
                                                                                                                                                          60
                                                                                                                                                                         Ω
                                 Operations
                                            region 22
                                                             Bachelor's
                                                                             m
                                                                                               other
                                                                                                                   1
                                                                                                                                       0
                    7513 Sales & Marketing
                                            region_19
                                                             Bachelor's
                                                                             m
                                                                                            sourcing
                                                                                                                       34
                                                                                                                                                          50
                                                                                                                                                                         0
                    2542
                          Sales & Marketing
                                                             Bachelor's
                                                                                               other
                                                                                                                   2
                                                                                                                       39
                                                                                                                                       0
                                                                                                                                                          50
                                                                                                                                                                         0
                                            region 23
                                                                             m
                                                                                                                                       0
                   48945
                                Technology
                                                                                                                       45
                                                                                                                                                                         0
                                            region 26
                                                             Bachelor's
                                                                                                                                                          73
                                                                             m
                                                                                               other
                    3030
                                Technology
                                            region_14
                                                             Bachelor's
                                                                                                                       48
                                                                                                                                       0
                                                                                                                                                          78
                                                                                                                                                                         0
                                                                             m
                                                                                            sourcing
                   74592
                                                                                                                       37
                                                                                                                                       0
                                 Operations
                                            region_27 Master's & above
                                                                                               other
                                                                                                                                                          56
                                                                                                                                                                         0
                   13918
                                  Analytics
                                             region_1
                                                             Bachelor's
                                                                             m
                                                                                               other
                                                                                                                                       0
                                                                                                                                                          79
                                                                                                                                                                         0
                   13614 Sales & Marketing
                                                                  NaN
                                                                                            sourcing
                                                                                                                       29
                                                                                                                                       0
                                                                                                                                                          45
                                                                                                                                                                         0
                                             region_9
                   51526
                                       HR region_22
                                                             Bachelor's
                                                                                               othe
                                                                                                                       27
                                                                                                                                       0
                                                                                                                                                          49
                                                                                                                                                                         0
           54808 rows × 10 columns
In [11]: df['education'] = df['education'].fillna(99)
Out[11]:
                                department
                                               region
                                                             education gender recruitment_channel no_of_trainings age awards_won? avg_training_score is_promoted
            employee_id
                   65438
                          Sales & Marketing
                                             region_7
                                                       Master's & above
                                                                                            sourcing
                                                                                                                       35
                                                                                                                                       0
                                                                                                                                                          49
                                                                                                                                                                         0
                                                                                                                                       0
                   65141
                                 Operations
                                            region_22
                                                             Bachelor's
                                                                             m
                                                                                               other
                                                                                                                   1
                                                                                                                       30
                                                                                                                                                          60
                                                                                                                                                                         0
                    7513 Sales & Marketing
                                            region_19
                                                             Bachelor's
                                                                             m
                                                                                            sourcing
                                                                                                                   1
                                                                                                                       34
                                                                                                                                       0
                                                                                                                                                          50
                                                                                                                                                                         0
                    2542
                          Sales & Marketing
                                            region 23
                                                             Bachelor's
                                                                             m
                                                                                               other
                                                                                                                   2
                                                                                                                       39
                                                                                                                                       0
                                                                                                                                                          50
                                                                                                                                                                         0
                                                                                                                                       0
                   48945
                                Technology
                                            region 26
                                                             Bachelor's
                                                                             m
                                                                                               other
                                                                                                                       45
                                                                                                                                                          73
                                                                                                                                                                         0
                                Technology
                                                                                                                                       0
                    3030
                                            region 14
                                                             Bachelor's
                                                                             m
                                                                                            sourcing
                                                                                                                   1
                                                                                                                       48
                                                                                                                                                          78
                                                                                                                                                                         0
                   74592
                                 Operations
                                                                                                                       37
                                                                                                                                       0
                                                                                                                                                                         0
                                            region 27
                                                       Master's & above
                                                                                               other
                                                                                                                                                          56
                   13918
                                                                                               other
                                                                                                                       27
                                                                                                                                                          79
                                                                                                                                                                         0
                                  Analytics
                                             region 1
                                                             Bachelor's
                                                                             m
                   13614 Sales & Marketing
                                                                    99
                                                                                            sourcing
                                                                                                                       29
                                                                                                                                       0
                                                                                                                                                          45
                                                                                                                                                                         0
                                             region 9
                                                                             m
                   51526
                                       HR region_22
                                                                                                                       27
                                                                                                                                       0
                                                                                                                                                          49
                                                                                                                                                                         0
                                                             Bachelor's
                                                                                               othe
           54808 rows × 10 columns
           df['age'] = df['age'].fillna(df['age'].mean())
In [12]:
Out[12]:
                                                             education gender recruitment_channel no_of_trainings age awards_won? avg_training_score is_promoted
                                department
                                               region
            employee_id
                   65438
                                                                                                                       35
                                                                                                                                       0
                                                                                                                                                          49
                                                                                                                                                                         0
                          Sales & Marketing
                                             region_7
                                                       Master's & above
                                                                             f
                                                                                            sourcing
                                                                                                                   1
                   65141
                                 Operations
                                            region_22
                                                             Bachelor's
                                                                             m
                                                                                               othe
                                                                                                                       30
                                                                                                                                       0
                                                                                                                                                          60
                                                                                                                                                                         0
                    7513 Sales & Marketing
                                                             Bachelor's
                                                                                                                       34
                                                                                                                                       0
                                                                                                                                                          50
                                                                                                                                                                         0
                                            region_19
                                                                                            sourcing
                    2542
                          Sales & Marketing
                                            region_23
                                                             Bachelor's
                                                                                               other
                                                                                                                   2
                                                                                                                       39
                                                                                                                                       0
                                                                                                                                                          50
                                                                                                                                                                         0
                   48945
                                Technology
                                            region_26
                                                             Bachelor's
                                                                             m
                                                                                               other
                                                                                                                   1
                                                                                                                       45
                                                                                                                                       0
                                                                                                                                                          73
                                                                                                                                                                         n
                       ...
                                                                                                                                       0
                    3030
                                Technology
                                            region 14
                                                             Bachelor's
                                                                             m
                                                                                            sourcing
                                                                                                                   1
                                                                                                                       48
                                                                                                                                                          78
                                                                                                                                                                         0
                                                                                                                                       n
                   74592
                                                                                                                       37
                                                                                                                                                                         n
                                 Operations
                                            region_27
                                                       Master's & above
                                                                             f
                                                                                               other
                                                                                                                                                          56
                                                                                                                       27
                                                                                                                                       0
                   13918
                                  Analytics
                                             region_1
                                                             Bachelor's
                                                                             m
                                                                                               other
                                                                                                                                                          79
                                                                                                                                                                         0
                   13614 Sales & Marketing
                                                                    99
                                                                                                                       29
                                                                                                                                       0
                                                                                                                                                          45
                                                                                                                                                                         0
                                             region 9
                                                                             m
                                                                                            sourcing
                   51526
                                       HR region_22
                                                                                                                       27
                                                                                                                                       0
                                                                                                                                                          49
                                                                                                                                                                         0
                                                             Bachelor's
                                                                                               other
                                                                             m
           54808 rows × 10 columns
In [13]: import numpy as np
```

```
In [14]: df1 = pd.DataFrame(data={'col1':[np.nan,np.nan,2,3,4,np.nan,np.nan]})
In [15]: df1.fillna(method='pad', limit=1)
Out[15]: col1
         0 NaN
         1 NaN
         2 2.0
         3 3.0
         4 4.0
         5 4.0
         6 NaN
In [16]: df1.fillna(method='pad', limit=1)
Out[16]:
          col1
         0 NaN
         1 NaN
         2 2.0
         3 3.0
         4 4.0
         5 4.0
         6 NaN
In [17]: df1.fillna(method = 'bfill')
Out[17]: col1
         0 2.0
         1 2.0
         2 2.0
         3 3.0
         4 4.0
         5 NaN
         6 NaN
In [18]: df1.dropna()
Out[18]: col1
         2 2.0
         3 3.0
         4 4.0
In [19]: df1.dropna(axis=1)
Out[19]: _____
         1
         2
         3
         5
         6
```