

Columns Transformation in pandas Part 1

Import Library

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

Reading Csv files in Pandas

```
In [8]: df = pd.read_excel("C:/Users/prata/Downloads/ESD.xlsx")
df1 = pd.read_csv("DimCustomer.csv") #It will work if csv file in same folder
```

```
In [3]: df.head()
```

```
Out[3]:
```

	EEID	Full Name	Job Title	Department	Business Unit	Gender	Ethnicity	Age	Hire Date	Annual Salary	Bonus %	Country	City	Exit Date
0	E02387	Emily Davis	Sr. Manger	IT	Research & Development	Female	Black	55	2016-04-08	141604	0.15	United States	Seattle	2021-10-16
1	E04105	Theodore Dinh	Technical Architect	IT	Manufacturing	Male	Asian	59	1997-11-29	99975	0.00	China	Chongqing	NaT
2	E02572	Luna Sanders	Director	Finance	Speciality Products	Female	Caucasian	50	2006-10-26	163099	0.20	United States	Chicago	NaT
3	E02832	Penelope Jordan	Computer Systems Manager	IT	Manufacturing	Female	Caucasian	26	2019-09-27	84913	0.07	United States	Chicago	NaT
4	E01639	Austin Vo	Sr. Analyst	Finance	Manufacturing	Male	Asian	55	1995-11-20	95409	0.00	United States	Phoenix	NaT

```
In [4]: df.loc[(df["Bonus %"] == 0), "GetBonus"] = "No Bonus" #In dataSet I am creating New column(GetBonus) in basis of Bonus
print(df)
```





	EEID	Full Name	Job Title	Department	\
0	E02387	Emily Davis	Sr. Manger	IT	
1	E04105	Theodore Dinh	Technical Architect	IT	
2	E02572	Luna Sanders	Director	Finance	
3	E02832	Penelope Jordan	Computer Systems Manager	IT	
4	E01639	Austin Vo	Sr. Analyst	Finance	
..	
995	E03094	Wesley Young	Sr. Analyst	Marketing	
996	E01909	Lillian Khan	Analyst	Finance	
997	E04398	Oliver Yang	Director	Marketing	
998	E02521	Lily Nguyen	Sr. Analyst	Finance	
999	E03545	Sofia Cheng	Vice President	Accounting	

	Business Unit	Gender	Ethnicity	Age	Hire Date	Annual Salary	\
0	Research & Development	Female	Black	55	2016-04-08	141604	
1	Manufacturing	Male	Asian	59	1997-11-29	99975	
2	Speciality Products	Female	Caucasian	50	2006-10-26	163099	
3	Manufacturing	Female	Caucasian	26	2019-09-27	84913	
4	Manufacturing	Male	Asian	55	1995-11-20	95409	
..	
995	Speciality Products	Male	Caucasian	33	2016-09-18	98427	
996	Speciality Products	Female	Asian	44	2010-05-31	47387	
997	Speciality Products	Male	Asian	31	2019-06-10	176710	
998	Speciality Products	Female	Asian	33	2012-01-28	95960	
999	Corporate	Female	Asian	63	2020-07-26	216195	

	Bonus %	Country	City	Exit Date	GetBonus
0	0.15	United States	Seattle	2021-10-16	NaN
1	0.00	China	Chongqing	NaT	No Bonus
2	0.20	United States	Chicago	NaT	NaN
3	0.07	United States	Chicago	NaT	NaN
4	0.00	United States	Phoenix	NaT	No Bonus
..
995	0.00	United States	Columbus	NaT	No Bonus
996	0.00	China	Chengdu	2018-01-08	No Bonus
997	0.15	United States	Miami	NaT	NaN
998	0.00	China	Chengdu	NaT	No Bonus
999	0.31	United States	Miami	NaT	NaN

[1000 rows x 15 columns]

```
In [5]: df.loc[(df["Bonus %"] > 0), "GetBonus"] = "Bonus" #same as above line
print(df)
```



	EEID	Full Name	Job Title	Department	\
0	E02387	Emily Davis	Sr. Manger	IT	
1	E04105	Theodore Dinh	Technical Architect	IT	
2	E02572	Luna Sanders	Director	Finance	
3	E02832	Penelope Jordan	Computer Systems Manager	IT	
4	E01639	Austin Vo	Sr. Analyst	Finance	
..	
995	E03094	Wesley Young	Sr. Analyst	Marketing	
996	E01909	Lillian Khan	Analyst	Finance	
997	E04398	Oliver Yang	Director	Marketing	
998	E02521	Lily Nguyen	Sr. Analyst	Finance	
999	E03545	Sofia Cheng	Vice President	Accounting	

	Business Unit	Gender	Ethnicity	Age	Hire Date	Annual Salary	\
0	Research & Development	Female	Black	55	2016-04-08	141604	
1	Manufacturing	Male	Asian	59	1997-11-29	99975	
2	Speciality Products	Female	Caucasian	50	2006-10-26	163099	
3	Manufacturing	Female	Caucasian	26	2019-09-27	84913	
4	Manufacturing	Male	Asian	55	1995-11-20	95409	
..	
995	Speciality Products	Male	Caucasian	33	2016-09-18	98427	
996	Speciality Products	Female	Asian	44	2010-05-31	47387	
997	Speciality Products	Male	Asian	31	2019-06-10	176710	
998	Speciality Products	Female	Asian	33	2012-01-28	95960	
999	Corporate	Female	Asian	63	2020-07-26	216195	

	Bonus %	Country	City	Exit Date	GetBonus
0	0.15	United States	Seattle	2021-10-16	Bonus
1	0.00	China	Chongqing	NaT	No Bonus
2	0.20	United States	Chicago	NaT	Bonus
3	0.07	United States	Chicago	NaT	Bonus
4	0.00	United States	Phoenix	NaT	No Bonus
..
995	0.00	United States	Columbus	NaT	No Bonus
996	0.00	China	Chengdu	2018-01-08	No Bonus
997	0.15	United States	Miami	NaT	Bonus
998	0.00	China	Chengdu	NaT	No Bonus
999	0.31	United States	Miami	NaT	Bonus

[1000 rows x 15 columns]

In [6]: `df # calling dataset`



Out[6]:

	EEID	Full Name	Job Title	Department	Business Unit	Gender	Ethnicity	Age	Hire Date	Annual Salary	Bonus %	Country	City	Exit Date	GetI
0	E02387	Emily Davis	Sr. Manger	IT	Research & Development	Female	Black	55	2016-04-08	141604	0.15	United States	Seattle	2021-10-16	
1	E04105	Theodore Dinh	Technical Architect	IT	Manufacturing	Male	Asian	59	1997-11-29	99975	0.00	China	Chongqing	NaT	No
2	E02572	Luna Sanders	Director	Finance	Speciality Products	Female	Caucasian	50	2006-10-26	163099	0.20	United States	Chicago	NaT	
3	E02832	Penelope Jordan	Computer Systems Manager	IT	Manufacturing	Female	Caucasian	26	2019-09-27	84913	0.07	United States	Chicago	NaT	
4	E01639	Austin Vo	Sr. Analyst	Finance	Manufacturing	Male	Asian	55	1995-11-20	95409	0.00	United States	Phoenix	NaT	No
...
995	E03094	Wesley Young	Sr. Analyst	Marketing	Speciality Products	Male	Caucasian	33	2016-09-18	98427	0.00	United States	Columbus	NaT	No
996	E01909	Lillian Khan	Analyst	Finance	Speciality Products	Female	Asian	44	2010-05-31	47387	0.00	China	Chengdu	2018-01-08	No
997	E04398	Oliver Yang	Director	Marketing	Speciality Products	Male	Asian	31	2019-06-10	176710	0.15	United States	Miami	NaT	
998	E02521	Lily Nguyen	Sr. Analyst	Finance	Speciality Products	Female	Asian	33	2012-01-28	95960	0.00	China	Chengdu	NaT	No
999	E03545	Sofia Cheng	Vice President	Accounting	Corporate	Female	Asian	63	2020-07-26	216195	0.31	United States	Miami	NaT	

1000 rows × 15 columns



```
In [16]: df1.columns # cheeking columns in Dataset
```



```
Out[16]: Index(['CustomerKey', 'GeographyKey', 'CustomerAlternateKey', 'Title',  
          'FirstName', 'MiddleName', 'LastName', 'NameStyle', 'BirthDate',  
          'MaritalStatus', 'Suffix', 'Gender', 'EmailAddress', 'YearlyIncome',  
          'TotalChildren', 'NumberChildrenAtHome', 'EnglishEducation',  
          'SpanishEducation', 'FrenchEducation', 'EnglishOccupation',  
          'SpanishOccupation', 'FrenchOccupation', 'HouseOwnerFlag',  
          'NumberCarsOwned', 'AddressLine1', 'AddressLine2', 'Phone',  
          'DateFirstPurchase', 'CommuteDistance', 'fullname'],  
          dtype='object')
```

```
In [17]: df1["fullname"] = df1["FirstName"]+ " " + df1["LastName"] # Concat first name and last name and created one new column Fullname  
print(df1)
```



	CustomerKey	GeographyKey	CustomerAlternateKey	Title	FirstName	\
0	11000	26	AW00011000	NaN	Jon	
1	11001	37	AW00011001	NaN	Eugene	
2	11002	31	AW00011002	NaN	Ruben	
3	11003	11	AW00011003	NaN	Christy	
4	11004	19	AW00011004	NaN	Elizabeth	
...	
18479	29479	209	AW00029479	NaN	Tommy	
18480	29480	248	AW00029480	NaN	Nina	
18481	29481	120	AW00029481	NaN	Ivan	
18482	29482	179	AW00029482	NaN	Clayton	
18483	29483	217	AW00029483	NaN	Jésus	

	MiddleName	LastName	NameStyle	BirthDate	MaritalStatus	...	\
0	V	Yang	0	1971-10-06	M	...	
1	L	Huang	0	1976-05-10	S	...	
2	NaN	Torres	0	1971-02-09	M	...	
3	NaN	Zhu	0	1973-08-14	S	...	
4	NaN	Johnson	0	1979-08-05	S	...	
...	
18479	L	Tang	0	1969-06-30	M	...	
18480	W	Raji	0	1977-05-06	S	...	
18481	NaN	Suri	0	1965-07-04	S	...	
18482	NaN	Zhang	0	1964-09-01	M	...	
18483	L	Navarro	0	1965-06-06	M	...	

	SpanishOccupation	FrenchOccupation	HouseOwnerFlag	NumberCarsOwned	\
0	Profesional	Cadre	1	0	
1	Profesional	Cadre	0	1	
2	Profesional	Cadre	1	1	
3	Profesional	Cadre	0	1	
4	Profesional	Cadre	1	4	
...	
18479	Administrativo	Employé	1	0	
18480	Administrativo	Employé	1	0	
18481	Administrativo	Employé	0	0	
18482	Administrativo	Employé	1	0	
18483	Administrativo	Employé	1	0	

	AddressLine1	AddressLine2	Phone	\
0	3761 N. 14th St	NaN	1 (11) 500 555-0162	
1	2243 W St.	NaN	1 (11) 500 555-0110	
2	5844 Linden Land	NaN	1 (11) 500 555-0184	
3	1825 Village Pl.	NaN	1 (11) 500 555-0162	
4	7553 Harness Circle	NaN	1 (11) 500 555-0131	



```

...
18479      111, rue Maillard      ...      NaN  1 (11) 500 555-0136
18480      9 Katherine Drive      NaN  1 (11) 500 555-0146
18481      Knaackstr 4      NaN  1 (11) 500 555-0144
18482      1080, quai de Grenelle      NaN  1 (11) 500 555-0137
18483      244, rue de la Centenaire      NaN  1 (11) 500 555-0141

```

```

      DateFirstPurchase CommuteDistance      fullname
0      2011-01-19      1-2 Miles      Jon Yang
1      2011-01-15      0-1 Miles      Eugene Huang
2      2011-01-07      2-5 Miles      Ruben Torres
3      2010-12-29      5-10 Miles      Christy Zhu
4      2011-01-23      1-2 Miles      Elizabeth Johnson
...
18479      2012-09-04      0-1 Miles      Tommy Tang
18480      2013-07-17      0-1 Miles      Nina Raji
18481      2011-08-13      0-1 Miles      Ivan Suri
18482      2012-09-18      0-1 Miles      Clayton Zhang
18483      2012-09-09      0-1 Miles      Jésus Navarro

```

[18484 rows x 30 columns]

In [18]: `df1 # calling dataset`



Out[18]:

	CustomerKey	GeographyKey	CustomerAlternateKey	Title	FirstName	MiddleName	LastName	NameStyle	BirthDate	MaritalStatus	...
0	11000	26	AW00011000	NaN	Jon	V	Yang	0	1971-10-06	M	...
1	11001	37	AW00011001	NaN	Eugene	L	Huang	0	1976-05-10	S	...
2	11002	31	AW00011002	NaN	Ruben	NaN	Torres	0	1971-02-09	M	...
3	11003	11	AW00011003	NaN	Christy	NaN	Zhu	0	1973-08-14	S	...
4	11004	19	AW00011004	NaN	Elizabeth	NaN	Johnson	0	1979-08-05	S	...
...
18479	29479	209	AW00029479	NaN	Tommy	L	Tang	0	1969-06-30	M	...
18480	29480	248	AW00029480	NaN	Nina	W	Raji	0	1977-05-06	S	...
18481	29481	120	AW00029481	NaN	Ivan	NaN	Suri	0	1965-07-04	S	...
18482	29482	179	AW00029482	NaN	Clayton	NaN	Zhang	0	1964-09-01	M	...



CustomerKey	GeographyKey	CustomerAlternateKey	Title	FirstName	MiddleName	LastName	NameStyle	BirthDate	MaritalStatus	...
18483	29483	217	AW00029483	NaN	Jésus	L	Navarro	0	1965-06-06	M ...

In []:

