EDWARD ODHIAMBO

GITHUB: https://github.com/EddOdhiambo (https://github.com/EddOdhiambo) Linkedin: https://www.linkedin.com/in/edward-odhiambo-656225276/ (https://www.linkedin.com/in/edward-odhiambo-656225276/)

This is a simple data cleaning project using a sample kaggle dataset enrollments.The aim of the project is to get to understand the type of data we are handling, import the data into the notebook, identify data summary to get the general understanding of what the data is all about.

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

In [2]:
```
import numpy as np
```

In [5]:
```
data = pd.read_csv("C:/dev/pythondataset/enrollments.csv")
```

In [8]:
```
data
```

Out[8]:

	account_key	status	join_date	cancel_date	days_to_cancel	is_udacity	is_canceled
0	448	canceled	2014-11-10	2015-01-14	65.0	True	True
1	448	canceled	2014-11-05	2014-11-10	5.0	True	True
2	448	canceled	2015-01-27	2015-01-27	0.0	True	True
3	448	canceled	2014-11-10	2014-11-10	0.0	True	True
4	448	current	2015-03-10	NaN	NaN	True	False
...	...	...	...	...	...	...	...
1635	1176	current	2015-08-12	NaN	NaN	False	False
1636	1110	current	2015-08-13	NaN	NaN	False	False
1637	1116	canceled	2015-08-15	2015-08-18	3.0	False	True
1638	874	current	2015-08-22	NaN	NaN	False	False
1639	686	current	2015-08-23	NaN	NaN	False	False

1640 rows × 7 columns

In [10]:
```
print("SUMMARY\n",data)
```

```
SUMMARY
       account_key    status   join_date cancel_date  days_to_cancel  \
0              448  canceled  2014-11-10  2015-01-14            65.0
1              448  canceled  2014-11-05  2014-11-10             5.0
2              448  canceled  2015-01-27  2015-01-27             0.0
3              448  canceled  2014-11-10  2014-11-10             0.0
4              448   current  2015-03-10         NaN             NaN
...            ...       ...         ...         ...             ...
1635          1176   current  2015-08-12         NaN             NaN
1636          1110   current  2015-08-13         NaN             NaN
1637          1116  canceled  2015-08-15  2015-08-18             3.0
1638           874   current  2015-08-22         NaN             NaN
1639           686   current  2015-08-23         NaN             NaN

      is_udacity  is_canceled
0           True         True
1           True         True
2           True         True
3           True         True
4           True        False
...          ...          ...
1635       False        False
1636       False        False
1637       False         True
1638       False        False
1639       False        False

[1640 rows x 7 columns]
```

In [11]:
```
data.head
```

Out[11]:
```
<bound method NDFrame.head of       account_key    status   join_date cancel_date  days_to_cancel  \
0             448  canceled  2014-11-10  2015-01-14            65.0
1             448  canceled  2014-11-05  2014-11-10             5.0
2             448  canceled  2015-01-27  2015-01-27             0.0
3             448  canceled  2014-11-10  2014-11-10             0.0
4             448   current  2015-03-10         NaN             NaN
...           ...       ...         ...         ...             ...
1635         1176   current  2015-08-12         NaN             NaN
1636         1110   current  2015-08-13         NaN             NaN
1637         1116  canceled  2015-08-15  2015-08-18             3.0
1638          874   current  2015-08-22         NaN             NaN
1639          686   current  2015-08-23         NaN             NaN

      is_udacity  is_canceled
0           True         True
1           True         True
2           True         True
3           True         True
4           True        False
...          ...          ...
1635       False        False
1636       False        False
1637       False         True
1638       False        False
1639       False        False

[1640 rows x 7 columns]>
```

In [12]:
```
data.tail
```

Out[12]:
```
<bound method NDFrame.tail of       account_key    status   join_date cancel_date  days_to_cancel  \
0             448  canceled  2014-11-10  2015-01-14            65.0
1             448  canceled  2014-11-05  2014-11-10             5.0
2             448  canceled  2015-01-27  2015-01-27             0.0
3             448  canceled  2014-11-10  2014-11-10             0.0
4             448   current  2015-03-10         NaN             NaN
...           ...       ...         ...         ...             ...
1635         1176   current  2015-08-12         NaN             NaN
1636         1110   current  2015-08-13         NaN             NaN
1637         1116  canceled  2015-08-15  2015-08-18             3.0
1638          874   current  2015-08-22         NaN             NaN
1639          686   current  2015-08-23         NaN             NaN

      is_udacity  is_canceled
0           True         True
1           True         True
2           True         True
3           True         True
4           True        False
...          ...          ...
1635       False        False
1636       False        False
1637       False         True
1638       False        False
1639       False        False

[1640 rows x 7 columns]>
```

In [13]:
```
data.dropna(inplace = True)
```

In [17]:
```
data
```

Out[17]:

	account_key	status	join_date	cancel_date	days_to_cancel	is_udacity	is_canceled
0	448	canceled	2014-11-10	2015-01-14	65.0	True	True
1	448	canceled	2014-11-05	2014-11-10	5.0	True	True
2	448	canceled	2015-01-27	2015-01-27	0.0	True	True
3	448	canceled	2014-11-10	2014-11-10	0.0	True	True
5	448	canceled	2015-01-14	2015-01-27	13.0	True	True
...	...	...	...	...	...	...	...
1624	1232	canceled	2015-07-16	2015-07-17	1.0	False	True
1625	1057	canceled	2015-07-17	2015-07-21	4.0	False	True
1628	1272	canceled	2015-07-17	2015-07-23	6.0	False	True
1632	807	canceled	2015-07-20	2015-07-22	2.0	False	True
1637	1116	canceled	2015-08-15	2015-08-18	3.0	False	True

988 rows × 7 columns

In [18]:
```
data.duplicated(). sum()
```

Out[18]:
```
4
```

In [14]:
```
data.duplicated()
```

Out[14]:
```
0       False
1       False
2       False
3       False
5       False
        ...
1624    False
1625    False
1628    False
1632    False
1637    False
Length: 988, dtype: bool
```

In [19]:
```
data.drop_duplicates
```

Out[19]:
```
<bound method DataFrame.drop_duplicates of      account_key    status   join_date cancel_date  days_to
_cancel  \
0             448  canceled  2014-11-10  2015-01-14           65.0
1             448  canceled  2014-11-05  2014-11-10            5.0
2             448  canceled  2015-01-27  2015-01-27            0.0
3             448  canceled  2014-11-10  2014-11-10            0.0
5             448  canceled  2015-01-14  2015-01-27           13.0
...           ...       ...         ...         ...            ...
1624         1232  canceled  2015-07-16  2015-07-17            1.0
1625         1057  canceled  2015-07-17  2015-07-21            4.0
1628         1272  canceled  2015-07-17  2015-07-23            6.0
1632          807  canceled  2015-07-20  2015-07-22            2.0
1637         1116  canceled  2015-08-15  2015-08-18            3.0

      is_udacity  is_canceled
0           True         True
1           True         True
2           True         True
3           True         True
5           True         True
...          ...          ...
1624       False         True
1625       False         True
1628       False         True
1632       False         True
1637       False         True

[988 rows x 7 columns]>
```

In [15]:
```
data
```

Out[15]:

	account_key	status	join_date	cancel_date	days_to_cancel	is_udacity	is_canceled
0	448	canceled	2014-11-10	2015-01-14	65.0	True	True
1	448	canceled	2014-11-05	2014-11-10	5.0	True	True
2	448	canceled	2015-01-27	2015-01-27	0.0	True	True
3	448	canceled	2014-11-10	2014-11-10	0.0	True	True
5	448	canceled	2015-01-14	2015-01-27	13.0	True	True
...	...	...	...	...	...	...	...
1624	1232	canceled	2015-07-16	2015-07-17	1.0	False	True
1625	1057	canceled	2015-07-17	2015-07-21	4.0	False	True
1628	1272	canceled	2015-07-17	2015-07-23	6.0	False	True
1632	807	canceled	2015-07-20	2015-07-22	2.0	False	True
1637	1116	canceled	2015-08-15	2015-08-18	3.0	False	True

988 rows × 7 columns

In [16]:
```
data.isnull(). sum()
```

Out[16]:
```
account_key       0
status            0
join_date         0
cancel_date       0
days_to_cancel    0
is_udacity        0
is_canceled       0
dtype: int64
```

In [20]:
```
data
```

Out[20]:

	account_key	status	join_date	cancel_date	days_to_cancel	is_udacity	is_canceled
0	448	canceled	2014-11-10	2015-01-14	65.0	True	True
1	448	canceled	2014-11-05	2014-11-10	5.0	True	True
2	448	canceled	2015-01-27	2015-01-27	0.0	True	True
3	448	canceled	2014-11-10	2014-11-10	0.0	True	True
5	448	canceled	2015-01-14	2015-01-27	13.0	True	True
...	...	...	...	...	...	...	...
1624	1232	canceled	2015-07-16	2015-07-17	1.0	False	True
1625	1057	canceled	2015-07-17	2015-07-21	4.0	False	True
1628	1272	canceled	2015-07-17	2015-07-23	6.0	False	True
1632	807	canceled	2015-07-20	2015-07-22	2.0	False	True
1637	1116	canceled	2015-08-15	2015-08-18	3.0	False	True

988 rows × 7 columns

In [ ]:
```

```