Data Preprocessing assignment

December 31, 2021

##

Data Preprocessing assignment

[1]:
```
# Import libraries
import pandas as pd
```

[2]:
```
# Data
raw_data = {'first_name' : ['Jason', 'Molly', 'Tina', 'Jake', 'Amy'],
'last_name' : ['Miller', 'Jacobson', '.', 'Milner','Cooze'],
'age' : [42, 52, 36, 24, 73],
'preTestScore' : [4, 24, 31, '.', '.'],
'postTestScore' : ['25,000', '94,000', 57, 62, 70]}
```

[3]:
```
# Creating dataframe
df = pd.DataFrame(raw_data,
                  columns=['first_name', 'last_name', 'age', 'preTestScore',␣
 ↪'postTestScore'])

df
```

[3]:

	first_name	last_name	age	preTestScore	postTestScore
0	Jason	Miller	42	4	25,000
1	Molly	Jacobson	52	24	94,000
2	Tina	.	36	31	57
3	Jake	Milner	24	.	62
4	Amy	Cooze	73	.	70

Save the data frame into a .csv file as project.csv

[4]:
```
df.to_csv('project.csv', index = False)
```

Read the project.csv file and print the data frame.

[5]:
```
student_df = pd.read_csv('project.csv')
student_df
```

[5]:

	first_name	last_name	age	preTestScore	postTestScore
0	Jason	Miller	42	4	25,000
1	Molly	Jacobson	52	24	94,000

```
2       Tina          .   36           31            57
3       Jake     Milner   24            .            62
4        Amy      Cooze   73            .            70
```

Read the project.csv file without column heading

[6]:
```
student_df_no_heading = pd.read_csv('project.csv', header=None, skiprows=1)
student_df_no_heading
```

[6]:
```
       0         1   2   3       4
0  Jason    Miller  42   4  25,000
1  Molly  Jacobson  52  24  94,000
2   Tina         .  36  31      57
3   Jake    Milner  24   .      62
4    Amy     Cooze  73   .      70
```

Read the project.csv file and make two index columns, namely, 'First Name' and 'Last Name'.

[7]:
```
student_df = pd.read_csv('project.csv', index_col=['first_name','last_name'])
student_df
```

[7]:
```
                     age preTestScore postTestScore
first_name last_name
Jason      Miller     42            4        25,000
Molly      Jacobson   52           24        94,000
Tina       .          36           31            57
Jake       Milner     24            .            62
Amy        Cooze      73            .            70
```

Print the data frame in a Boolean form as True or False. True for Null/ NaN values and false for non-null values.

[8]:
```
student_df = pd.read_csv('project.csv', na_values=['.'])
student_df.isna()
```

[8]:
```
   first_name  last_name    age  preTestScore  postTestScore
0       False      False  False         False          False
1       False      False  False         False          False
2       False       True  False         False          False
3       False      False  False          True          False
4       False      False  False          True          False
```

Read the data frame by skipping the first 3 rows and print the data frame.

[9]:
```
student_df = pd.read_csv('project.csv', skiprows=3, header=None)
student_df
```

[9]:
```
      0       1   2   3   4
0  Tina       .  36  31  57
1  Jake  Milner  24   .  62
```

```
2   Amy   Cooze  73   .  70
```

[ ]: