Working on Real Project with Python

(A part of Big Data Analysis)

Police Dataset

Here,

The data from a Police Check Post is given.

This data is available as a CSV file. We are going to analyze this data set using the Pandas DataFrame.

```
import pandas as pd
data = pd.read csv(r"C:\Users\ROHIT GREWAL\Desktop\DSL\Videos\10. Real
Project 3 - Police\Police Data.csv")
data.head()
   stop date stop time
                         country name driver gender
                                                      driver age raw \
0
   1/2/2005
                   1:55
                                  NaN
                                                              1985.0
                                  NaN
                                                   М
  1/18/2005
                  8:15
                                                              1965.0
  1/23/2005
                 23:15
                                  NaN
                                                   М
                                                              1972.0
                 17:15
3 2/20/2005
                                  NaN
                                                   М
                                                              1986.0
4 3/14/2005
                 10:00
                                  NaN
                                                              1984.0
                               violation_raw violation
   driver_age driver_race
search conducted
         20.0
                    White
                                    Speeding
                                              Speeding
0
False
1
         40.0
                    White
                                    Speeding
                                              Speeding
False
         33.0
                    White
                                    Speeding
                                              Speeding
False
                    White Call for Service
         19.0
                                                  0ther
False
         21.0
                    White
                                    Speeding
                                              Speeding
False
                stop outcome is arrested stop duration
  search type
drugs related stop
          NaN
                    Citation
                                    False
                                                0-15 Min
False
```

1	NaN	Citation	False	0-15 Min
False				
2	NaN	Citation	False	0-15 Min
False				
3	NaN	Arrest Driver	True	16-30 Min
False				
4	NaN	Citation	False	0-15 Min
False				

Instruction (For Data Cleaning)

1. Remove the column that only contains missing values

```
# df.isnull().sum()
# df.drop( columns = 'Column name' , inplace = True )
data.isnull().sum()
stop_date
                           0
stop time
country_name
                       65535
driver_gender
                        4061
driver_age_raw
                        4054
driver_age
                        4307
driver race
                        4060
violation raw
                        4060
                        4060
violation
search conducted
                           0
search type
                       63056
stop outcome
                        4060
is arrested
                        4060
stop duration
                        4060
drugs_related_stop
dtype: int64
data.drop( columns = 'country name', inplace = True)
data
       stop_date stop_time driver_gender
                                           driver age raw
driver_age \
        1/2/2005
                                                                   20.0
                       1:55
                                         Μ
                                                    1985.0
       1/18/2005
                       8:15
                                                    1965.0
                                                                   40.0
```

2	1/23/2005	23:15		М	1972.0	33.0
3	2/20/2005	17:15		М	1986.0	19.0
4	3/14/2005	10:00		F	1984.0	21.0
65530	12/6/2012	17:54		F	1987.0	25.0
65531	12/6/2012	22:22		М	1954.0	58.0
65532	12/6/2012	23:20		М	1985.0	27.0
65533	12/7/2012	0:23	Na	эN	NaN	NaN
65534	12/7/2012	0:30		F	1985.0	27.0
0 1 2 3 4 65530 65531 65532 65533 65534	driver_race White White White White White White White Black NaN White search_conduration		Call ⁻ t/Inspection	Speeding Speeding Speeding Speeding for Service Speeding Speeding Speeding NaN Speeding	0ther	
0 15 Min		False	NaN	Citation	n False	0 -
1 1 15 Min		False	NaN	Citation	n False	0 -
2 15 Min		False	NaN	Citation	n False	0 -
3 30 Min		False	NaN A	rrest Drive	True	16-
4 15 Min		False	NaN	Citation	n False	0 -
65530 15 Min		False	NaN	Citation	n False	0 -
65531		False	NaN	Warning	y False	0 -

15 Min					
65532	False	NaN	Citation	False	0 -
15 Min					
65533	False	NaN	NaN	NaN	
NaN	F-1	NoN	Citation	Fol 66	0
65534 15 Min	False	NaN	Citation	False	0 -
TO LITH					
	drugs related stop				
0	False				
1	False				
2	False				
0 1 2 3 4	False				
4	False				
65530	False				
65531	False				
65532	False				
65533	False				
65534	False				
[65535	rows x 14 columns]				

Question (Based on Filtering + Value Counts)

2. For Speeding, were Men or Women stopped more often?

```
# df[df.Column 1 == 'Element/Value'].Column 2.value counts()
data.head()
   stop_date stop_time driver_gender driver_age_raw driver_age
driver_race \
    1/2/2005
                   1:55
                                                1985.0
                                                              20.0
White
  1/18/2005
                  8:15
                                                1965.0
                                                              40.0
  1/23/2005
                 23:15
                                                1972.0
                                                              33.0
White
   2/20/2005
                 17:15
                                                1986.0
                                                              19.0
White
  3/14/2005
                 10:00
                                                1984.0
                                                              21.0
White
```

```
violation_raw violation search_conducted search_type
stop_outcome \
           Speeding Speeding
                                           False
                                                          NaN
Citation
           Speeding Speeding
                                           False
                                                          NaN
Citation
           Speeding Speeding
                                           False
                                                          NaN
Citation
  Call for Service
                        0ther
                                           False
                                                          NaN
                                                               Arrest
Driver
           Speeding Speeding
                                           False
                                                          NaN
Citation
                              drugs related stop
  is arrested stop duration
        False
                   0-15 Min
                                           False
1
        False
                   0-15 Min
                                           False
2
        False
                   0-15 Min
                                           False
3
         True
                  16-30 Min
                                           False
        False
                   0-15 Min
                                           False
data[data.violation == 'Speeding'].driver gender.value counts()
М
     25517
F
     11686
Name: driver_gender, dtype: int64
```

Question (Groupby)

3. Does gender affect who gets searched during a stop?

dat	a.head()				
	• -	stop_time	driver_gender	driver_age_raw	driver_age
0	ver_race 1/2/2005	1:55	М	1985.0	20.0
Whi					
	1/18/2005	8:15	М	1965.0	40.0
Whi	_				
	1/23/2005	23:15	М	1972.0	33.0
Whi	_				
3	2/20/2005	17:15	М	1986.0	19.0
Whi	te				
4	3/14/2005	10:00	F	1984.0	21.0
Whi	te				

```
violation raw violation search conducted search type
stop outcome
           Speeding Speeding
                                          False
                                                         NaN
Citation
           Speeding Speeding
                                          False
                                                         NaN
Citation
           Speeding Speeding
                                          False
                                                         NaN
Citation
3 Call for Service
                        0ther
                                           False
                                                         NaN
                                                             Arrest
Driver
           Speeding Speeding
                                                         NaN
                                          False
Citation
                             drugs related stop
  is arrested stop duration
        False
                   0-15 Min
                                          False
        False
                   0-15 Min
                                           False
1
2
        False
                   0-15 Min
3
        True
                  16-30 Min
                                           False
        False
                   0-15 Min
                                          False
# df.groupby('Column 1').Column 2.sum()
data.groupby('driver_gender').search_conducted.sum()
driver gender
      366.0
F
М
     2113.0
Name: search conducted, dtype: float64
data.search_conducted.value_counts()
False
         63056
True
          2479
Name: search conducted, dtype: int64
```

Question (mapping + data-type casting)

4. What is the mean stop_duration?

```
# df['Column_name'] = df['Column_name'].map( { old:new , old:new} )
# df['Column_name'].mean()
data.head()
```

<pre>stop_date stop_time d driver race \</pre>	river_gender dri	iver_age_raw	driver_age					
$0 1/\overline{2}/2005 1:55$	М	1985.0	20.0					
White 1 1/18/2005 8:15	М	1965.0	40.0					
White 2 1/23/2005 23:15	М	1972.0	33.0					
White 3 2/20/2005 17:15	М	1986.0	19.0					
White 4 3/14/2005 10:00 White	F	1984.0	21.0					
<pre>violation_raw viola stop outcome \</pre>	ation search_cor	nducted search	_type					
0 Speeding Spee	eding	False	NaN					
1 Speeding Spee	eding	False	NaN					
2 Speeding Spee	eding	False	NaN					
	Other	False	NaN Arrest					
4 Speeding Spee Citation	eding	False	NaN					
is_arrested stop_durat: 0 False 0-15	Min Min Min Min	ed_stop False False False False False						
data.stop_duration.value	_counts()							
Series([], Name: stop_du	ration, dtype: ir	nt64)						
<pre>data['stop_duration'] = data['stop_duration'].map({'0-15 Min' : 7.5 , '16-30 Min' : 24 , '30+ Min' : 45 })</pre>								
data								
<pre>stop_date stop_time driver_age \</pre>	me driver_gender	driver_age_r	aw					
0 1/2/2005 1:5	55 M	1985	.0 20.0					
1 1/18/2005 8:	15 M	1965	.0 40.0					
2 1/23/2005 23:	15 M	1972	.0 33.0					
3 2/20/2005 17:3	15 M	1986	.0 19.0					

65530 3 65531 3 65532 3	3/14/2005 12/6/2012	10:00		F	1984.0	21.0
65530 3 65531 3 65532 3	 12/6/2012					
65531 3 65532 3	12/6/2012			•		
65532		17:54		F	1987.0	25.0
	12/6/2012	22:22		М	1954.0	58.0
	12/6/2012	23:20		М	1985.0	27.0
65533	12/7/2012	0:23	Na	N	NaN	NaN
65534	12/7/2012	0:30		F	1985.0	27.0
di 0 1 2 3 4	river_race White White White White White			Speeding Speeding Speeding Speeding For Service Speeding	violation \ Speeding Speeding Speeding Other Speeding	
65530 65531 65532 65533 65534	White White Black NaN White	Equipment/	Inspection	Speeding Speeding Violation NaN Speeding	Speeding Speeding Equipment NaN Speeding	
	search_cond ration \	ucted searc	h_type s	top_outcome	e is_arrested	
0 7.5		False	NaN	Citation	n False	
7.5 1 7.5		False	NaN	Citation	n False	
2		False	NaN	Citation	n False	
7.5 3		False	NaN Ar	rest Driver	True	
NaN 4		False	NaN	Citation	n False	
7.5 						
 65530		False	NaN	Citation	n False	
7.5 65531		False	NaN	Warning	y False	
7.5 65532		False	NaN	Citation	n False	
7.5 65533		False	NaN	NaN	l NaN	

NaN 65534	False	NaN	Citation	False					
7.5									
Θ	drugs_related_stop False	sto_duration 7.5							
0 1 2 3 4	False	7.5							
2	False	7.5							
3 1	False False	NaN 7.5							
	14636	,							
65530	False	7.5							
65531	False	7.5							
65532 65533	False False	7.5 NaN							
65534	False	7.5							
[65535 rows x 15 columns]									
data[ˈ	<pre>data['stop_duration'].mean()</pre>								
9.484218206532603									

Question (Groupby, Describe)

5. Compare the age distributions for each violation

```
# df.groupby('Column_1').Column_2.describe()
data.head()
   stop_date stop_time driver_gender driver_age_raw
                                                      driver age
driver_race \
    1/2/2005
                  1:55
                                               1985.0
                                                              20.0
White
  1/18/2005
                  8:15
                                               1965.0
                                                              40.0
White
  1/23/2005
                 23:15
                                               1972.0
                                                              33.0
White
3 2/20/2005
                 17:15
                                               1986.0
                                                              19.0
White
  3/14/2005
                 10:00
                                               1984.0
                                                              21.0
White
      violation_raw violation search_conducted search_type
```

stop_outco						
0	Speeding	Speeding		False	NaN	
Citation 1	Speeding	Speeding		False	NaN	
Citation						
2 Citation	Speeding	Speeding		False	NaN	
	or Service	0ther		False	NaN	l Arrest
Driver						
4 Citation	Speeding	Speeding		False	NaN	
1 Fa 2 Fa 3	sted stop_ alse alse alse True alse	duration 7.5 7.5 7.5 NaN 7.5	drugs_rela	ted_stop s False False False False False		ion 7.5 7.5 7.5 NaN 7.5
data.group	oby('violat	ion').dri	ver_age.des	cribe()		
		count	mean	std	min	25% 50%
75% \ violation						
Equipment 39.0		6507.0	31.682957	11.380671	16.0 2	23.0 28.0
Moving vi	olation	11876.0	36.736443	13.258350	15.0 2	25.0 35.0
47.0 Other		3477.0	40.362381	12.754423	16.0 3	80.0 41.0
50.0						
Registrat:	ion/plates	2240.0	32.656696	11.150780	16.0 2	4.0 30.0
Seat belt		3.0	30.333333	10.214369	23.0 2	4.5 26.0
34.0		27120 0	22 262501	10 615701	15 0 0	20 20 0
Speeding 42.0		37120.0	33.262581	12.615781	15.0 2	23.0 30.0
violation		max				
Equipment Moving vio Other	olation ion/plates	81.0 86.0 86.0 74.0 42.0 88.0				
a = [1,2,3]	3,4,5,6,7,8	,9,10,11,	12,13,14,15]		

import numpy as np np.mean(a)	
8.0	
	_
	_
	—
	_