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             M          1985.0
1  1/18/2005      8:15           NaN             M          1965.0
2  1/23/2005     23:15           NaN             M          1972.0
3  2/20/2005     17:15           NaN             M          1986.0
4  3/14/2005     10:00           NaN             F          1984.0

   driver_age driver_race     violation_raw violation
search_conducted  \
0        20.0       White          Speeding  Speeding
False
1        40.0       White          Speeding  Speeding
False
2        33.0       White          Speeding  Speeding
False
3        19.0       White  Call for Service     Other
False
4        21.0       White          Speeding  Speeding
False

  search_type   stop_outcome is_arrested stop_duration
drugs_related_stop
0         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                 0
country_name          65535
driver_gender          4061
driver_age_raw         4054
driver_age             4307
driver_race            4060
violation_raw          4060
violation              4060
search_conducted          0
search_type           63056
stop_outcome           4060
is_arrested            4060
stop_duration          4060
drugs_related_stop        0
dtype: int64
```

```
data.drop( columns = 'country_name', inplace = True)

data
```

```
       stop_date stop_time driver_gender  driver_age_raw
driver_age  \
0       1/2/2005      1:55             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     17:15              M          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              M          1954.0         58.0
65532  12/6/2012     23:20              M          1985.0         27.0
65533  12/7/2012      0:23            NaN             NaN          NaN
65534  12/7/2012      0:30              F          1985.0         27.0


      driver_race                    violation_raw  violation  \
0           White                         Speeding   Speeding
1           White                         Speeding   Speeding
2           White                         Speeding   Speeding
3           White                Call for Service      Other
4           White                         Speeding   Speeding
...           ...                              ...        ...
65530       White                         Speeding   Speeding
65531       White                         Speeding   Speeding
65532       Black  Equipment/Inspection Violation  Equipment
65533         NaN                              NaN        NaN
65534       White                         Speeding   Speeding

       search_conducted search_type   stop_outcome is_arrested
stop_duration  \
0                 False         NaN       Citation       False      0-
15 Min
1                 False         NaN       Citation       False      0-
15 Min
2                 False         NaN       Citation       False      0-
15 Min
3                 False         NaN  Arrest Driver        True     16-
30 Min
4                 False         NaN       Citation       False      0-
15 Min
...                 ...         ...            ...         ...
...
65530             False         NaN       Citation       False      0-
15 Min
65531             False         NaN        Warning       False      0-
```

```
15 Min
65532              False          NaN       Citation        False      0-
15 Min
65533              False          NaN            NaN          NaN
NaN
65534              False          NaN       Citation        False      0-
15 Min

       drugs_related_stop
0                   False
1                   False
2                   False
3                   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  \
0   1/2/2005      1:55             M          1985.0        20.0
White
1  1/18/2005      8:15             M          1965.0        40.0
White
2  1/23/2005     23:15             M          1972.0        33.0
White
3  2/20/2005     17:15             M          1986.0        19.0
White
4  3/14/2005     10:00             F          1984.0        21.0
White
```

```
      violation_raw violation  search_conducted search_type
stop_outcome  \
0          Speeding  Speeding             False         NaN
Citation
1          Speeding  Speeding             False         NaN
Citation
2          Speeding  Speeding             False         NaN
Citation
3  Call for Service     Other             False         NaN  Arrest
Driver
4          Speeding  Speeding             False         NaN
Citation

  is_arrested stop_duration  drugs_related_stop
0       False      0-15 Min               False
1       False      0-15 Min               False
2       False      0-15 Min               False
3        True     16-30 Min               False
4       False      0-15 Min               False
```

```
data[data.violation == 'Speeding'].driver_gender.value_counts()
```

```
M    25517
F    11686
Name: driver_gender, dtype: int64
```

Question (Groupby)

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

```
data.head()
```

```
   stop_date stop_time driver_gender  driver_age_raw  driver_age
driver_race  \
0   1/2/2005      1:55             M          1985.0        20.0
White
1  1/18/2005      8:15             M          1965.0        40.0
White
2  1/23/2005     23:15             M          1972.0        33.0
White
3  2/20/2005     17:15             M          1986.0        19.0
White
4  3/14/2005     10:00             F          1984.0        21.0
White
```

```
      violation_raw violation  search_conducted search_type
stop_outcome  \
0          Speeding  Speeding             False         NaN
Citation
1          Speeding  Speeding             False         NaN
Citation
2          Speeding  Speeding             False         NaN
Citation
3  Call for Service     Other             False         NaN  Arrest
Driver
4          Speeding  Speeding             False         NaN
Citation

  is_arrested stop_duration  drugs_related_stop
0       False      0-15 Min               False
1       False      0-15 Min               False
2       False      0-15 Min               False
3        True     16-30 Min               False
4       False      0-15 Min               False
```

```
# df.groupby('Column_1').Column_2.sum()

data.groupby('driver_gender').search_conducted.sum()
```

```
driver_gender
F     366.0
M    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()
```

```
   stop_date stop_time driver_gender  driver_age_raw  driver_age
driver_race  \
0   1/2/2005      1:55             M          1985.0        20.0
White
1  1/18/2005      8:15             M          1965.0        40.0
White
2  1/23/2005     23:15             M          1972.0        33.0
White
3  2/20/2005     17:15             M          1986.0        19.0
White
4  3/14/2005     10:00             F          1984.0        21.0
White

      violation_raw violation  search_conducted search_type
stop_outcome  \
0          Speeding  Speeding             False         NaN
Citation
1          Speeding  Speeding             False         NaN
Citation
2          Speeding  Speeding             False         NaN
Citation
3  Call for Service     Other             False         NaN  Arrest
Driver
4          Speeding  Speeding             False         NaN
Citation

  is_arrested stop_duration  drugs_related_stop
0       False      0-15 Min               False
1       False      0-15 Min               False
2       False      0-15 Min               False
3        True     16-30 Min               False
4       False      0-15 Min               False

data.stop_duration.value_counts()

Series([], Name: stop_duration, dtype: int64)

data['stop_duration']= data['stop_duration'].map( {'0-15 Min' : 7.5 ,
'16-30 Min' : 24 , '30+ Min' : 45 })

data

       stop_date stop_time driver_gender  driver_age_raw
driver_age  \
0       1/2/2005      1:55             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     17:15             M          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              M          1954.0         58.0

65532  12/6/2012     23:20              M          1985.0         27.0

65533  12/7/2012      0:23            NaN             NaN          NaN

65534  12/7/2012      0:30              F          1985.0         27.0


      driver_race                    violation_raw  violation  \
0           White                         Speeding   Speeding
1           White                         Speeding   Speeding
2           White                         Speeding   Speeding
3           White                Call for Service      Other
4           White                         Speeding   Speeding
...           ...                              ...        ...
65530       White                         Speeding   Speeding
65531       White                         Speeding   Speeding
65532       Black  Equipment/Inspection Violation  Equipment
65533         NaN                              NaN        NaN
65534       White                         Speeding   Speeding

       search_conducted search_type   stop_outcome is_arrested
stop_duration  \
0                 False         NaN       Citation       False
7.5
1                 False         NaN       Citation       False
7.5
2                 False         NaN       Citation       False
7.5
3                 False         NaN  Arrest Driver        True
NaN
4                 False         NaN       Citation       False
7.5
...                 ...         ...            ...         ...
...
65530             False         NaN       Citation       False
7.5
65531             False         NaN        Warning       False
7.5
65532             False         NaN       Citation       False
7.5
65533             False         NaN            NaN         NaN
```

```
NaN
65534              False          NaN       Citation        False
7.5

       drugs_related_stop  sto_duration
0                   False           7.5
1                   False           7.5
2                   False           7.5
3                   False           NaN
4                   False           7.5
...                   ...           ...
65530               False           7.5
65531               False           7.5
65532               False           7.5
65533               False           NaN
65534               False           7.5

[65535 rows x 15 columns]
```

```
data['stop_duration'].mean()
```

```
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  \
0   1/2/2005      1:55             M          1985.0        20.0
White
1  1/18/2005      8:15             M          1965.0        40.0
White
2  1/23/2005     23:15             M          1972.0        33.0
White
3  2/20/2005     17:15             M          1986.0        19.0
White
4  3/14/2005     10:00             F          1984.0        21.0
White

      violation_raw violation  search_conducted search_type
```

```
stop_outcome  \
0          Speeding  Speeding              False          NaN
Citation
1          Speeding  Speeding              False          NaN
Citation
2          Speeding  Speeding              False          NaN
Citation
3  Call for Service     Other              False          NaN  Arrest
Driver
4          Speeding  Speeding              False          NaN
Citation

  is_arrested  stop_duration  drugs_related_stop  sto_duration
0       False            7.5               False           7.5
1       False            7.5               False           7.5
2       False            7.5               False           7.5
3        True            NaN               False           NaN
4       False            7.5               False           7.5
```

```
data.groupby('violation').driver_age.describe()
```

```
                        count       mean        std   min   25%   50%
75%  \
violation

Equipment              6507.0  31.682957  11.380671  16.0  23.0  28.0
39.0
Moving violation      11876.0  36.736443  13.258350  15.0  25.0  35.0
47.0
Other                  3477.0  40.362381  12.754423  16.0  30.0  41.0
50.0
Registration/plates    2240.0  32.656696  11.150780  16.0  24.0  30.0
40.0
Seat belt                 3.0  30.333333  10.214369  23.0  24.5  26.0
34.0
Speeding              37120.0  33.262581  12.615781  15.0  23.0  30.0
42.0

                        max
violation
Equipment              81.0
Moving violation       86.0
Other                  86.0
Registration/plates    74.0
Seat belt              42.0
Speeding               88.0
```

```
a = [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15]
```

```
import numpy as np
np.mean(a)
```

```
8.0
```