Working on Real Project with Python

(A part of Big Data Analysis)

Police Dataset

Here,
The data from a police checj post is given.
This data is available as a CSV file. We are going to analyze this data set using the pandas dataframe.

In [199]:
```
import pandas as pd
```

In [201]:
```
df=pd.read_csv('police dataset.csv')
```

In [202]:
```
df
```

Out[202]:

	stop_date	stop_time	country_name	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_conducted	search_type	stop_outcome	is_arrested	stop_duration	drugs_related_stop
0	1/2/2005	1:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
1	1/18/2005	8:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
2	1/23/2005	23:15	NaN	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
3	2/20/2005	17:15	NaN	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	16-30 Min	False
4	3/14/2005	10:00	NaN	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
65530	12/6/2012	17:54	NaN	F	1987.0	25.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
65531	12/6/2012	22:22	NaN	M	1954.0	58.0	White	Speeding	Speeding	False	NaN	Warning	False	0-15 Min	False
65532	12/6/2012	23:20	NaN	M	1985.0	27.0	Black	Equipment/Inspection Violation	Equipment	False	NaN	Citation	False	0-15 Min	False
65533	12/7/2012	0:23	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False	NaN	NaN	NaN	NaN	False
65534	12/7/2012	0:30	NaN	F	1985.0	27.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False

65535 rows × 15 columns

Instruction (For Data Cleaning)

1. Remove the cloumn that only contains missing valuse

In [207]:
```
df.isnull().sum()
```

Out[207]:
```
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
```

In [208]:
```
df.drop( columns = 'country_name',inplace=True)
```

In [211]:
```
df.sample(4)
```

Out[211]:

	stop_date	stop_time	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_conducted	search_type	stop_outcome	is_arrested	stop_duration	drugs_related_stop
32924	3/1/2009	7:52	M	1972.0	37.0	White	Speeding	Speeding	False	NaN	Citation	False	16-30 Min	False
9848	9/7/2006	0:15	M	1987.0	19.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
3500	1/29/2006	11:00	M	1985.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
12630	12/10/2006	8:31	M	1982.0	24.0	Black	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False

Question (Based on Filtering + Value Counts)

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

In [215]:
```
df.head()
```

Out[215]:

	stop_date	stop_time	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_conducted	search_type	stop_outcome	is_arrested	stop_duration	drugs_related_stop
0	1/2/2005	1:55	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
1	1/18/2005	8:15	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
2	1/23/2005	23:15	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
3	2/20/2005	17:15	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	16-30 Min	False
4	3/14/2005	10:00	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False

In [216]:
```
df[df.violation == 'Speeding'].driver_gender.value_counts()
```

Out[216]:
```
driver_gender
M    25517
F    11686
Name: count, dtype: int64
```

Question (Groupby)

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

In [221]:
```
df.head()
```

Out[221]:

	stop_date	stop_time	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_conducted	search_type	stop_outcome	is_arrested	stop_duration	drugs_related_stop
0	1/2/2005	1:55	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
1	1/18/2005	8:15	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
2	1/23/2005	23:15	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
3	2/20/2005	17:15	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	16-30 Min	False
4	3/14/2005	10:00	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False

In [223]:
```
df.groupby('driver_gender').search_conducted.sum()
```

Out[223]:
```
driver_gender
F     366
M    2113
Name: search_conducted, dtype: int64
```

In [225]:
```
df.search_conducted.value_counts()
```

Out[225]:
```
search_conducted
False    63056
True      2479
Name: count, dtype: int64
```

Question (mapping + data-type casting)

4. What is the mean stop_duration ?

In [229]:
```
df.head()
```

Out[229]:

	stop_date	stop_time	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_conducted	search_type	stop_outcome	is_arrested	stop_duration	drugs_related_stop
0	1/2/2005	1:55	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
1	1/18/2005	8:15	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
2	1/23/2005	23:15	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
3	2/20/2005	17:15	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	16-30 Min	False
4	3/14/2005	10:00	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False

In [231]:
```
df.stop_duration.value_counts()
```

Out[231]:
```
stop_duration
0-15 Min     47379
16-30 Min    11448
30+ Min       2647
2                1
Name: count, dtype: int64
```

In [233]:
```
df['stop_duration']=df['stop_duration'].map( {'0-15 Min':7.5, '16-30 Min': 24,'30+ Min':45})
```

In [235]:
```
df
```

Out[235]:

	stop_date	stop_time	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_conducted	search_type	stop_outcome	is_arrested	stop_duration	drugs_related_stop
0	1/2/2005	1:55	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
1	1/18/2005	8:15	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
2	1/23/2005	23:15	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
3	2/20/2005	17:15	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	24.0	False
4	3/14/2005	10:00	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
65530	12/6/2012	17:54	F	1987.0	25.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
65531	12/6/2012	22:22	M	1954.0	58.0	White	Speeding	Speeding	False	NaN	Warning	False	7.5	False
65532	12/6/2012	23:20	M	1985.0	27.0	Black	Equipment/Inspection Violation	Equipment	False	NaN	Citation	False	7.5	False
65533	12/7/2012	0:23	NaN	NaN	NaN	NaN	NaN	NaN	False	NaN	NaN	NaN	NaN	False
65534	12/7/2012	0:30	F	1985.0	27.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False

65535 rows × 14 columns

In [237]:
```
df['stop_duration'].mean()
```

Out[237]:
```
12.187420698181345
```

Question (Groupby, Describe)

5. Compare the age distributions for each violation

In [241]:
```
df.head()
```

Out[241]:

	stop_date	stop_time	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_conducted	search_type	stop_outcome	is_arrested	stop_duration	drugs_related_stop
0	1/2/2005	1:55	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
1	1/18/2005	8:15	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
2	1/23/2005	23:15	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
3	2/20/2005	17:15	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	24.0	False
4	3/14/2005	10:00	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False

In [245]:
```
df.groupby('violation').driver_age.describe()
```

Out[245]:

violation	count	mean	std	min	25%	50%	75%	max
Equipment	6507.0	31.682957	11.380671	16.0	23.0	28.0	39.0	81.0
Moving violation	11876.0	36.736443	13.258350	15.0	25.0	35.0	47.0	86.0
Other	3477.0	40.362381	12.754423	16.0	30.0	41.0	50.0	86.0
Registration/plates	2240.0	32.656696	11.150780	16.0	24.0	30.0	40.0	74.0
Seat belt	3.0	30.333333	10.214369	23.0	24.5	26.0	34.0	42.0
Speeding	37120.0	33.262581	12.615781	15.0	23.0	30.0	42.0	88.0

End - Project