Police Dataset

- The data from a Police Check Post is given.
- You have to analyze the data using the Pandas DataFrame

In [24]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [25]:
```
df =pd.read_csv('D:\Download in D-Drive\police.csv')
```

In [27]:
```
df.head()
```

Out[27]:

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

- Instruction (For Data Cleaning)

1.Remove the column that only contains missing values

In [20]:
```
# df.isnull().sum()
# df.drop( columns = 'column_name', inplace =True)
```

In [32]:
```
df.isnull().sum()
```

Out[32]:
```
stop_date               0
stop_time               0
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 [31]:
```
df
```

Out[31]:

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

65535 rows × 14 columns

- For Speeding(Based on filtering + Value Counts) # For speeding,were Men or Women stopped more often ?

In [33]:
```
# df[df.columns_name =='Elements/Value'].column_name.value_counts()
df[df.violation == 'Speeding'].driver_gender.value_counts()
```

Out[33]:
```
M    25517
F    11686
Name: driver_gender, dtype: int64
```

In [ ]:
```

```

- question(groupby) # Does gender affect who gets searched during a stop?

In [34]:
```
df
```

Out[34]:

	stop_date	stop_time	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw
0	1/2/2005	1:55	M	1985.0	20.0	White	Speeding

	stop_date	stop_time	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw
1	1/18/2005	8:15	M	1965.0	40.0	White	Speeding
2	1/23/2005	23:15	M	1972.0	33.0	White	Speeding
3	2/20/2005	17:15	M	1986.0	19.0	White	Call for Service
4	3/14/2005	10:00	F	1984.0	21.0	White	Speeding
...	...	...	...	...	...	...	...
65530	12/6/2012	17:54	F	1987.0	25.0	White	Speeding
65531	12/6/2012	22:22	M	1954.0	58.0	White	Speeding
65532	12/6/2012	23:20	M	1985.0	27.0	Black	Equipment/Inspection Violation
65533	12/7/2012	0:23	NaN	NaN	NaN	NaN	NaN
65534	12/7/2012	0:30	F	1985.0	27.0	White	Speeding

65535 rows × 14 columns

In [42]:
```
# df.groupby('column_1').column_2.sum()
df.groupby('driver_race').search_conducted.sum()
```

Out[42]:
```
driver_race
Asian         36
Black        616
Hispanic     407
Other          1
White       1419
Name: search_conducted, dtype: int64
```

In [43]:
```
df.groupby('driver_gender').search_conducted.sum()
```

Out[43]:
```
driver_gender
F     366
M    2113
Name: search_conducted, dtype: int64
```

In [46]:
```
df.search_conducted.value_counts()
```

Out[46]:
```
False    63056
True      2479
Name: search_conducted, dtype: int64
```

In [48]:
```
df.is_arrested.value_counts()
```

Out[48]:
```
False    59215
True      2260
Name: is_arrested, dtype: int64
```

In [47]:
```
df.groupby('driver_gender').is_arrested.sum()
```

Out[47]:
```
driver_gender
F     464
M    1796
Name: is_arrested, dtype: int64
```

- question (mapping + stop_duration) # what is the mean stop duration

In [53]:
```
# df['column_name'] =df['column_name'].map({old:new , old:new})
# df['column_name'].mean()
```

In [54]:
```
df.stop_duration.value_counts()
```

Out[54]:
```
0-15 Min     47379
16-30 Min    11448
30+ Min       2647
2                1
Name: stop_duration, dtype: int64
```

In [56]:
```
df['stop_duration'] =df['stop_duration'].map({'0-15 Min': 7.5,'16-30 Min': 24 ,'30+ Min
```

In [57]:
```
df
```

Out[57]:

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

65535 rows × 14 columns

In [60]:
```
print(" The mean value of stop duration is:",df['stop_duration'].mean())
```

```
The mean value of stop duration is: 12.187420698181345
```

- Question (Groupby,describe) # Compare the age distributions for each violation

In [61]:
```
# df.groupby('column_1').column_2.decribe()
```

In [63]:
```
df.groupby('violation').driver_age.describe()
```

Out[63]:

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

In [64]:
```
df['driver_age'].mean()
```

Out[64]:
```
34.14898412491017
```

In [ ]:
```

```