In [1]:

```
import pandas as pd
```

In [3]:

```
police = pd.read_csv("Police Dataset.csv")
police
```

Out[3]:

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

65535 rows × 15 columns

1. Instruction (For Data Cleaning) - Remove the column that only contains missing values

In [6]:

```
police.isnull().sum()
```

Out[6]:

```
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 [8]:

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

In [10]:

```
police
```

Out[10]:

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

65535 rows × 14 columns

2. Question (Based on Filtering + Value Counts) - For Speeding , were Men or Women stopped more often ?

In [13]:

```
police.head(1)
```

Out[13]:

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

In [19]:

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

Out[19]:

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

3. Question (Groupby) - Does gender affect who gets searched during a stop ? Question (mapping + data-type casting)

In [22]:

```
police.head(1)
```

Out[22]:

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

In [24]:

```
police.groupby('driver_gender')['search_conducted'].sum()
```

Out[24]:

```
driver_gender
F     366
M    2113
Name: search_conducted, dtype: int64
```

In [26]:

```
police['search_conducted'].value_counts()
```

Out[26]:

```
search_conducted
False    63056
True      2479
Name: count, dtype: int64
```

4. Question (mapping + data-type casting) - What is the mean stop_duration ?

In [29]:

```
police.stop_duration.value_counts()
```

Out[29]:

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

In [33]:

```
police['stop_duration'] = police['stop_duration'].map({'0-15 Min' : 7.5, '16-30 Min' : 2
```

In [35]:

```
police
```

Out[35]:

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

65535 rows × 14 columns

5. Question (Groupby , Describe) - Compare the age distributions for each violation.

In [38]:

```
police.head(1)
```

Out[38]:

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

In [64]:

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

Out[64]:

	count	mean	std	min	25%	50%	75%	max
violation								
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