

```
In [1]: import pandas as pd
df = pd.read_csv(r'C:\Users\PC-chetan\Downloads\3. Police Data.csv')
df
```

Out[1]:

|       | 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       |
| 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

# Data Exploration

```
In [2]: df.head()
```

Out[2]:

|   | stop_date | stop_time | country_name | driver_gender | driver_age_raw | driver_age | driver_race | viol |
|---|-----------|-----------|--------------|---------------|----------------|------------|-------------|------|
| 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       | Call |
| 4 | 3/14/2005 | 10:00     | NaN          | F             | 1984.0         | 21.0       | White       |      |

```
In [3]: df.columns
```

Out[3]: Index(['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'], dtype='object')

```
In [4]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65535 entries, 0 to 65534
Data columns (total 15 columns):
#   Column                                Non-Null Count  Dtype
---  -
Loading [MathJax]/extensions/Safe.js
65535 non-null  object
```

```

1 stop_time          65535 non-null object
2 country_name       0 non-null    float64
3 driver_gender      61474 non-null object
4 driver_age_raw     61481 non-null float64
5 driver_age         61228 non-null float64
6 driver_race        61475 non-null object
7 violation_raw      61475 non-null object
8 violation          61475 non-null object
9 search_conducted   65535 non-null bool
10 search_type        2479 non-null object
11 stop_outcome       61475 non-null object
12 is_arrested        61475 non-null object
13 stop_duration      61475 non-null object
14 drugs_related_stop 65535 non-null bool
dtypes: bool(2), float64(3), object(10)
memory usage: 6.6+ MB

```

```
In [5]: df.isnull().sum()
```

```

Out[5]: 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 [6]: df.notnull().sum()
```

```

Out[6]: stop_date          65535
stop_time          65535
country_name        0
driver_gender      61474
driver_age_raw     61481
driver_age         61228
driver_race        61475
violation_raw      61475
violation          61475
search_conducted   65535
search_type        2479
stop_outcome       61475
is_arrested        61475
stop_duration      61475
drugs_related_stop 65535
dtype: int64

```

```
In [7]: df.describe()
```

```

Out[7]:
```

|       | country_name | driver_age_raw | driver_age   |
|-------|--------------|----------------|--------------|
| count | 0.0          | 61481.000000   | 61228.000000 |
| mean  | NaN          | 1967.791106    | 34.148984    |

|            | country_name | driver_age_raw | driver_age |
|------------|--------------|----------------|------------|
| <b>std</b> | NaN          | 121.050106     | 12.760710  |
| <b>min</b> | NaN          | 0.000000       | 15.000000  |
| <b>25%</b> | NaN          | 1965.000000    | 23.000000  |
| <b>50%</b> | NaN          | 1978.000000    | 31.000000  |
| <b>75%</b> | NaN          | 1985.000000    | 43.000000  |
| <b>max</b> | NaN          | 8801.000000    | 88.000000  |

In [8]: `df.shape`

Out[8]: (65535, 15)

In [9]: `df.index`

Out[9]: RangeIndex(start=0, stop=65535, step=1)

In [10]: `df.count()`

Out[10]:

|                    |       |
|--------------------|-------|
| stop_date          | 65535 |
| stop_time          | 65535 |
| country_name       | 0     |
| driver_gender      | 61474 |
| driver_age_raw     | 61481 |
| driver_age         | 61228 |
| driver_race        | 61475 |
| violation_raw      | 61475 |
| violation          | 61475 |
| search_conducted   | 65535 |
| search_type        | 2479  |
| stop_outcome       | 61475 |
| is_arrested        | 61475 |
| stop_duration      | 61475 |
| drugs_related_stop | 65535 |

dtype: int64

In [11]: `df.violation.value_counts()`

Out[11]:

|                     |       |
|---------------------|-------|
| Speeding            | 37204 |
| Moving violation    | 11926 |
| Equipment           | 6516  |
| Other               | 3583  |
| Registration/plates | 2243  |
| Seat belt           | 3     |

Name: violation, dtype: int64

In [12]: `df.violation.unique()`

Out[12]: array(['Speeding', 'Other', 'Equipment', 'Moving violation', nan, 'Registration/plates', 'Seat belt'], dtype=object)

In [13]: `df.violation.nunique()`

Out[13]: 6

In [ ]:

# Data Analisys

In [14]:

```
df.head()
```

Out[14]:

|   | stop_date | stop_time | country_name | driver_gender | driver_age_raw | driver_age | driver_race | viol |
|---|-----------|-----------|--------------|---------------|----------------|------------|-------------|------|
| 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       | Call |
| 4 | 3/14/2005 | 10:00     | NaN          | F             | 1984.0         | 21.0       | White       |      |

In [15]:

```
df.violation.unique()
```

Out[15]:

array(['Speeding', 'Other', 'Equipment', 'Moving violation', nan, 'Registration/plates', 'Seat belt'], dtype=object)

In [16]:

```
df[df['driver_race'] == 'White']
```

Out[16]:

|       | 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       |
| ...   | ...       | ...       | ...          | ...           | ...            | ...        | ...         |
| 65528 | 12/6/2012 | 15:39     | NaN          | M             | 1987.0         | 25.0       | White       |
| 65529 | 12/6/2012 | 16:00     | NaN          | M             | 1994.0         | 18.0       | 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       |
| 65534 | 12/7/2012 | 0:30      | NaN          | F             | 1985.0         | 27.0       | White       |

45747 rows × 15 columns

In [17]:

```
df[df['driver_age'] == 20]
```

Out[17]:

|     | 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       |
| 43  | 10/1/2005 | 1:00      | NaN          | M             | 1985.0         | 20.0       | White       |
| 73  | 10/2/2005 | 6:10      | NaN          | M             | 1985.0         | 20.0       | White       |
| 125 | 10/4/2005 | 11:50     | NaN          | M             | 1985.0         | 20.0       | White       |

|       | stop_date  | stop_time | country_name | driver_gender | driver_age_raw | driver_age | driver_race |
|-------|------------|-----------|--------------|---------------|----------------|------------|-------------|
| 137   | 10/4/2005  | 23:10     | NaN          | M             | 1985.0         | 20.0       | White       |
| ...   | ...        | ...       | ...          | ...           | ...            | ...        | ...         |
| 65389 | 11/30/2012 | 15:54     | NaN          | F             | 1992.0         | 20.0       | White       |
| 65472 | 12/4/2012  | 11:20     | NaN          | M             | 1992.0         | 20.0       | White       |
| 65485 | 12/5/2012  | 1:26      | NaN          | M             | 1992.0         | 20.0       | Black       |
| 65513 | 12/6/2012  | 10:10     | NaN          | F             | 1992.0         | 20.0       | White       |
| 65521 | 12/6/2012  | 14:00     | NaN          | F             | 1992.0         | 20.0       | White       |

2783 rows × 15 columns

In [18]:

df.isnull().sum()

Out[18]:

stop\_date0  
stop\_time0  
country\_name65535  
driver\_gender4061  
driver\_age\_raw4054  
driver\_age4307  
driver\_race4060  
violation\_raw4060  
violation4060  
search\_conducted0  
search\_type63056  
stop\_outcome4060  
is\_arrested4060  
stop\_duration4060  
drugs\_related\_stop0  
dtype: int64

In [19]:

df.columns

Out[19]:

Index(['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'],  
dtype='object')

In [20]:

df.rename(columns= {'violation': 'vio'} , inplace =True)

In [21]:

df

Out[21]:

|   | 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       |

|              | stop_date | stop_time | country_name | driver_gender | driver_age_raw | driver_age | driver_race |
|--------------|-----------|-----------|--------------|---------------|----------------|------------|-------------|
| <b>65530</b> | 12/6/2012 | 17:54     | NaN          | F             | 1987.0         | 25.0       | White       |
| <b>65531</b> | 12/6/2012 | 22:22     | NaN          | M             | 1954.0         | 58.0       | White       |
| <b>65532</b> | 12/6/2012 | 23:20     | NaN          | M             | 1985.0         | 27.0       | Black       |
| <b>65533</b> | 12/7/2012 | 0:23      | NaN          | NaN           | NaN            | NaN        | NaN         |
| <b>65534</b> | 12/7/2012 | 0:30      | NaN          | F             | 1985.0         | 27.0       | White       |

65535 rows × 15 columns

In [22]: `df.driver_age.mean()`

Out[22]: 34.14898412491017

In [23]: `df.driver_age.var()`

Out[23]: 162.83572552652538

In [24]: `df[df['vio'] == 'Equipment']`

Out[24]:

|              | stop_date | stop_time | country_name | driver_gender | driver_age_raw | driver_age | driver_race |
|--------------|-----------|-----------|--------------|---------------|----------------|------------|-------------|
| <b>5</b>     | 3/23/2005 | 9:45      | NaN          | M             | 1982.0         | 23.0       | Black       |
| <b>17</b>    | 7/19/2005 | 23:30     | NaN          | M             | 1979.0         | 26.0       | White       |
| <b>35</b>    | 9/30/2005 | 0:15      | NaN          | M             | 1978.0         | 27.0       | White       |
| <b>40</b>    | 10/1/2005 | 0:00      | NaN          | M             | 1988.0         | 17.0       | White       |
| <b>41</b>    | 10/1/2005 | 0:00      | NaN          | M             | 1988.0         | 17.0       | White       |
| ...          | ...       | ...       | ...          | ...           | ...            | ...        | ...         |
| <b>65520</b> | 12/6/2012 | 12:43     | NaN          | M             | 1953.0         | 59.0       | White       |
| <b>65525</b> | 12/6/2012 | 15:09     | NaN          | M             | 1980.0         | 32.0       | White       |
| <b>65526</b> | 12/6/2012 | 15:22     | NaN          | M             | 1982.0         | 30.0       | White       |
| <b>65528</b> | 12/6/2012 | 15:39     | NaN          | M             | 1987.0         | 25.0       | White       |
| <b>65532</b> | 12/6/2012 | 23:20     | NaN          | M             | 1985.0         | 27.0       | Black       |

6516 rows × 15 columns

In [25]: `df[(df['stop_time'] < '23:20' ) & (df['driver_age'] > 23 )]`

Out[25]:

|  | stop_date | stop_time | country_name | driver_gender | driver_age_raw | driver_age | driver_race |
|--|-----------|-----------|--------------|---------------|----------------|------------|-------------|
|--|-----------|-----------|--------------|---------------|----------------|------------|-------------|

|       | stop_date | stop_time | country_name | driver_gender | driver_age_raw | driver_age | driver_race |
|-------|-----------|-----------|--------------|---------------|----------------|------------|-------------|
| 2     | 1/23/2005 | 23:15     | NaN          | M             | 1972.0         | 33.0       | White       |
| 6     | 4/1/2005  | 17:30     | NaN          | M             | 1969.0         | 36.0       | White       |
| 8     | 7/13/2005 | 10:15     | NaN          | M             | 1970.0         | 35.0       | Black       |
| 9     | 7/13/2005 | 15:45     | NaN          | M             | 1970.0         | 35.0       | White       |
| 10    | 7/13/2005 | 16:20     | NaN          | M             | 1979.0         | 26.0       | Asian       |
| ...   | ...       | ...       | ...          | ...           | ...            | ...        | ...         |
| 65527 | 12/6/2012 | 15:26     | NaN          | F             | 1981.0         | 31.0       | White       |
| 65528 | 12/6/2012 | 15:39     | NaN          | M             | 1987.0         | 25.0       | 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       |
| 65534 | 12/7/2012 | 0:30      | NaN          | F             | 1985.0         | 27.0       | White       |

30739 rows × 15 columns

In [26]:

df.head()

Out[26]:

|   | stop_date | stop_time | country_name | driver_gender | driver_age_raw | driver_age | driver_race | viol |
|---|-----------|-----------|--------------|---------------|----------------|------------|-------------|------|
| 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       | Call |
| 4 | 3/14/2005 | 10:00     | NaN          | F             | 1984.0         | 21.0       | White       |      |

In [27]:

df.groupby('driver\_age').mean()

Out[27]:

|            | country_name | driver_age_raw | search_conducted | drugs_related_stop |
|------------|--------------|----------------|------------------|--------------------|
| driver_age |              |                |                  |                    |
| 15.0       | NaN          | 1992.600000    | 0.200000         | 0.000000           |
| 16.0       | NaN          | 1992.088235    | 0.088235         | 0.029412           |
| 17.0       | NaN          | 1991.144766    | 0.031180         | 0.008909           |
| 18.0       | NaN          | 1990.211896    | 0.044610         | 0.013383           |
| 19.0       | NaN          | 1989.463568    | 0.049414         | 0.016332           |
| ...        | ...          | ...            | ...              | ...                |
| 83.0       | NaN          | 1926.000000    | 0.000000         | 0.000000           |
| 84.0       | NaN          | 1925.666667    | 0.000000         | 0.000000           |
| 85.0       | NaN          | 1927.000000    | 0.000000         | 0.000000           |
| 86.0       | NaN          | 1923.166667    | 0.000000         | 0.000000           |
| 88.0       | NaN          | 1920.500000    | 0.000000         | 0.000000           |

73 rows × 4 columns

```
In [28]: df.groupby('driver_age').min()

C:\Users\PC-CHE~1\AppData\Local\Temp\ipykernel_6736\3430904074.py:1: FutureWarning: Dropping invalid columns in DataFrameGroupBy.min is deprecated. In a future version, a TypeError will be raised. Before calling .min, select only columns which should be valid for the function.
df.groupby('driver_age').min()
```

Out[28]:

|            | stop_date  | stop_time | country_name | driver_age_raw | search_conducted | is_arrested | dr  |
|------------|------------|-----------|--------------|----------------|------------------|-------------|-----|
| driver_age |            |           |              |                |                  |             |     |
| 15.0       | 11/10/2010 | 0:40      | NaN          | 1991.0         | False            | False       |     |
| 16.0       | 1/10/2006  | 0:01      | NaN          | 1989.0         | False            | False       |     |
| 17.0       | 1/1/2008   | 0:00      | NaN          | 1988.0         | False            | False       |     |
| 18.0       | 1/10/2007  | 0:00      | NaN          | 1987.0         | False            | False       |     |
| 19.0       | 1/1/2007   | 0:00      | NaN          | 1986.0         | False            | False       |     |
| ...        | ...        | ...       | ...          | ...            | ...              | ...         | ... |
| 83.0       | 3/19/2007  | 10:50     | NaN          | 1924.0         | False            | False       |     |
| 84.0       | 1/24/2009  | 11:14     | NaN          | 1925.0         | False            | False       |     |
| 85.0       | 2/22/2012  | 13:58     | NaN          | 1927.0         | False            | False       |     |
| 86.0       | 11/17/2009 | 10:07     | NaN          | 1922.0         | False            | False       |     |
| 88.0       | 6/20/2011  | 12:05     | NaN          | 1918.0         | False            | False       |     |

73 rows × 7 columns

```
In [29]: df.groupby('driver_age').max()

C:\Users\PC-CHE~1\AppData\Local\Temp\ipykernel_6736\1046300893.py:1: FutureWarning: Dropping invalid columns in DataFrameGroupBy.max is deprecated. In a future version, a TypeError will be raised. Before calling .max, select only columns which should be valid for the function.
df.groupby('driver_age').max()
```

Out[29]:

|            | stop_date | stop_time | country_name | driver_age_raw | search_conducted | is_arrested | dr  |
|------------|-----------|-----------|--------------|----------------|------------------|-------------|-----|
| driver_age |           |           |              |                |                  |             |     |
| 15.0       | 9/30/2006 | 1:30      | NaN          | 1995.0         | True             | True        |     |
| 16.0       | 9/16/2008 | 7:20      | NaN          | 1996.0         | True             | True        |     |
| 17.0       | 9/9/2007  | 9:51      | NaN          | 1995.0         | True             | True        |     |
| 18.0       | 9/9/2012  | 9:59      | NaN          | 1994.0         | True             | True        |     |
| 19.0       | 9/9/2012  | 9:58      | NaN          | 1993.0         | True             | True        |     |
| ...        | ...       | ...       | ...          | ...            | ...              | ...         | ... |
| 83.0       | 9/28/2011 | 8:20      | NaN          | 1928.0         | False            | False       |     |
| 84.0       | 9/21/2011 | 7:04      | NaN          | 1927.0         | False            | False       |     |
| 85.0       | 2/22/2012 | 13:58     | NaN          | 1927.0         | False            | False       |     |
| 86.0       | 4/5/2008  | 18:32     | NaN          | 1925.0         | False            | False       |     |
| 88.0       | 9/6/2006  | 12:15     | NaN          | 1923.0         | False            | False       |     |



```
In [30]: df.isnull().sum()
```

```
Out[30]: 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
vio               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.drop(columns= ['country_name'], inplace =True)
```

```
In [32]: df.columns
```

```
Out[32]: Index(['stop_date', 'stop_time', 'driver_gender', 'driver_age_raw',
               'driver_age', 'driver_race', 'violation_raw', 'vio', 'search_conducted',
               'search_type', 'stop_outcome', 'is_arrested', 'stop_duration',
               'drugs_related_stop'],
              dtype='object')
```

```
In [33]: df.isnull().sum()
```

```
Out[33]: stop_date          0
stop_time          0
driver_gender      4061
driver_age_raw     4054
driver_age        4307
driver_race        4060
violation_raw      4060
vio               4060
search_conducted   0
search_type        63056
stop_outcome       4060
is_arrested        4060
stop_duration      4060
drugs_related_stop 0
dtype: int64
```

```
In [34]: df.driver_gender.value_counts()
```

```
Out[34]: M    45164
F     16310
Name: driver_gender, dtype: int64
```

```
In [35]: df.driver_gender.isnull().sum()
```

```
Out[35]: 4061
```

```
df.driver_gender.fillna('F', inplace = True)
```

```
In [37]: df.driver_gender.isnull().sum()
```

```
Out[37]: 0
```

```
In [38]: df.isnull().sum()
```

```
Out[38]: stop_date      0
stop_time      0
driver_gender   0
driver_age_raw  4054
driver_age      4307
driver_race     4060
violation_raw   4060
vio             4060
search_conducted 0
search_type     63056
stop_outcome     4060
is_arrested      4060
stop_duration    4060
drugs_related_stop 0
dtype: int64
```

```
In [39]: df['driver_age_raw'][df.driver_age_raw.min()]
```

```
Out[39]: 1985.0
```

```
In [40]: df[df['driver_age_raw'] == 1985]
```

```
Out[40]:
```

|       | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|-------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
| 0     | 1/2/2005  | 1:55      | M             | 1985.0         | 20.0       | White       | Spe                    |
| 43    | 10/1/2005 | 1:00      | M             | 1985.0         | 20.0       | White       | Spe                    |
| 73    | 10/2/2005 | 6:10      | M             | 1985.0         | 20.0       | White       | Spe                    |
| 125   | 10/4/2005 | 11:50     | M             | 1985.0         | 20.0       | White       | Spe                    |
| 137   | 10/4/2005 | 23:10     | M             | 1985.0         | 20.0       | White       | Spe                    |
| ...   | ...       | ...       | ...           | ...            | ...        | ...         | ...                    |
| 65499 | 12/5/2012 | 23:13     | M             | 1985.0         | 27.0       | Black       | Equipment/Inspe<br>Vio |
| 65509 | 12/6/2012 | 7:13      | M             | 1985.0         | 27.0       | White       | Spe                    |
| 65519 | 12/6/2012 | 11:40     | F             | 1985.0         | 27.0       | Hispanic    | Spe                    |
| 65532 | 12/6/2012 | 23:20     | M             | 1985.0         | 27.0       | Black       | Equipment/Inspe<br>Vio |
| 65534 | 12/7/2012 | 0:30      | F             | 1985.0         | 27.0       | White       | Spe                    |

2630 rows × 14 columns

```
In [41]: df.driver_age_raw.fillna(1985, inplace= True)
```

In [42]: `df.isnull().sum()`

Out[42]:

|                    |       |
|--------------------|-------|
| stop_date          | 0     |
| stop_time          | 0     |
| driver_gender      | 0     |
| driver_age_raw     | 0     |
| driver_age         | 4307  |
| driver_race        | 4060  |
| violation_raw      | 4060  |
| vio                | 4060  |
| search_conducted   | 0     |
| search_type        | 63056 |
| stop_outcome       | 4060  |
| is_arrested        | 4060  |
| stop_duration      | 4060  |
| drugs_related_stop | 0     |
| dtype:             | int64 |

In [43]: `df1 = df.groupby('driver_age').min()`

C:\Users\PC-CHE~1\AppData\Local\Temp\ipykernel\_6736\1525749074.py:1: FutureWarning: Dropping invalid columns in DataFrameGroupBy.min is deprecated. In a future version, a TypeError will be raised. Before calling .min, select only columns which should be valid for the function.  
`df1 = df.groupby('driver_age').min()`

In [44]: `df1`

Out[44]:

|            | stop_date  | stop_time | driver_gender | driver_age_raw | search_conducted | is_arrested | dr  |
|------------|------------|-----------|---------------|----------------|------------------|-------------|-----|
| driver_age |            |           |               |                |                  |             |     |
| 15.0       | 11/10/2010 | 0:40      | F             | 1991.0         | False            | False       |     |
| 16.0       | 1/10/2006  | 0:01      | F             | 1989.0         | False            | False       |     |
| 17.0       | 1/1/2008   | 0:00      | F             | 1988.0         | False            | False       |     |
| 18.0       | 1/10/2007  | 0:00      | F             | 1987.0         | False            | False       |     |
| 19.0       | 1/1/2007   | 0:00      | F             | 1986.0         | False            | False       |     |
| ...        | ...        | ...       | ...           | ...            | ...              | ...         | ... |
| 83.0       | 3/19/2007  | 10:50     | M             | 1924.0         | False            | False       |     |
| 84.0       | 1/24/2009  | 11:14     | F             | 1925.0         | False            | False       |     |
| 85.0       | 2/22/2012  | 13:58     | M             | 1927.0         | False            | False       |     |
| 86.0       | 11/17/2009 | 10:07     | M             | 1922.0         | False            | False       |     |
| 88.0       | 6/20/2011  | 12:05     | M             | 1918.0         | False            | False       |     |

73 rows × 7 columns

In [45]: `df.driver_age.fillna(15, inplace = True)`

In [46]: `df.driver_age.isnull().sum()`

Out[46]: 0

Out[47]:

|                    |       |
|--------------------|-------|
| stop_date          | 0     |
| stop_time          | 0     |
| driver_gender      | 0     |
| driver_age_raw     | 0     |
| driver_age         | 0     |
| driver_race        | 4060  |
| violation_raw      | 4060  |
| vio                | 4060  |
| search_conducted   | 0     |
| search_type        | 63056 |
| stop_outcome       | 4060  |
| is_arrested        | 4060  |
| stop_duration      | 4060  |
| drugs_related_stop | 0     |

dtype: int64

In [48]:

```
df.groupby('driver_race').count()
```

Out[48]:

|             | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | violation_raw | vio   |
|-------------|-----------|-----------|---------------|----------------|------------|---------------|-------|
| driver_race |           |           |               |                |            |               |       |
| Asian       | 1639      | 1639      | 1639          | 1639           | 1639       | 1639          | 1639  |
| Black       | 8267      | 8267      | 8267          | 8267           | 8267       | 8267          | 8267  |
| Hispanic    | 5611      | 5611      | 5611          | 5611           | 5611       | 5611          | 5611  |
| Other       | 211       | 211       | 211           | 211            | 211        | 211           | 211   |
| White       | 45747     | 45747     | 45747         | 45747          | 45747      | 45747         | 45747 |

In [49]:

```
df.driver_race.fillna('Other', inplace = True)
```

In [50]:

```
df
```

Out[50]:

|       | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|-------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
| 0     | 1/2/2005  | 1:55      | M             | 1985.0         | 20.0       | White       | Spe                    |
| 1     | 1/18/2005 | 8:15      | M             | 1965.0         | 40.0       | White       | Spe                    |
| 2     | 1/23/2005 | 23:15     | M             | 1972.0         | 33.0       | White       | Spe                    |
| 3     | 2/20/2005 | 17:15     | M             | 1986.0         | 19.0       | White       | Call for Se            |
| 4     | 3/14/2005 | 10:00     | F             | 1984.0         | 21.0       | White       | Spe                    |
| ...   | ...       | ...       | ...           | ...            | ...        | ...         | ...                    |
| 65530 | 12/6/2012 | 17:54     | F             | 1987.0         | 25.0       | White       | Spe                    |
| 65531 | 12/6/2012 | 22:22     | M             | 1954.0         | 58.0       | White       | Spe                    |
| 65532 | 12/6/2012 | 23:20     | M             | 1985.0         | 27.0       | Black       | Equipment/Inspe<br>Vio |
| 65533 | 12/7/2012 | 0:23      | F             | 1985.0         | 15.0       | Other       |                        |
| 65534 | 12/7/2012 | 0:30      | F             | 1985.0         | 27.0       | White       | Spe                    |

65535 rows × 14 columns

In [51]:

```
df.isnull().sum()
```

Out[51]:

|                    |       |
|--------------------|-------|
| stop_date          | 0     |
| stop_time          | 0     |
| driver_gender      | 0     |
| driver_age_raw     | 0     |
| driver_age         | 0     |
| driver_race        | 0     |
| violation_raw      | 4060  |
| vio                | 4060  |
| search_conducted   | 0     |
| search_type        | 63056 |
| stop_outcome       | 4060  |
| is_arrested        | 4060  |
| stop_duration      | 4060  |
| drugs_related_stop | 0     |

dtype: int64

In [52]:

```
df.violation_raw.value_counts()
```

Out[52]:

|                                  |       |
|----------------------------------|-------|
| Speeding                         | 37204 |
| Other Traffic Violation          | 11926 |
| Equipment/Inspection Violation   | 6516  |
| Special Detail/Directed Patrol   | 2383  |
| Registration Violation           | 2243  |
| Call for Service                 | 793   |
| Violation of City/Town Ordinance | 170   |
| Motorist Assist/Courtesy         | 138   |
| APB                              | 51    |
| Suspicious Person                | 38    |
| Warrant                          | 10    |
| Seatbelt Violation               | 3     |

Name: violation\_raw, dtype: int64

In [53]:

```
df.violation_raw.fillna('Seatbelt Violation', inplace=True)
```

In [54]:

```
df
```

Out[54]:

|       | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|-------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
| 0     | 1/2/2005  | 1:55      | M             | 1985.0         | 20.0       | White       | Spe                    |
| 1     | 1/18/2005 | 8:15      | M             | 1965.0         | 40.0       | White       | Spe                    |
| 2     | 1/23/2005 | 23:15     | M             | 1972.0         | 33.0       | White       | Spe                    |
| 3     | 2/20/2005 | 17:15     | M             | 1986.0         | 19.0       | White       | Call for Se            |
| 4     | 3/14/2005 | 10:00     | F             | 1984.0         | 21.0       | White       | Spe                    |
| ...   | ...       | ...       | ...           | ...            | ...        | ...         |                        |
| 65530 | 12/6/2012 | 17:54     | F             | 1987.0         | 25.0       | White       | Spe                    |
| 65531 | 12/6/2012 | 22:22     | M             | 1954.0         | 58.0       | White       | Spe                    |
| 65532 | 12/6/2012 | 23:20     | M             | 1985.0         | 27.0       | Black       | Equipment/Inspe<br>Vio |
| 65533 | 12/7/2012 | 0:23      | F             | 1985.0         | 15.0       | Other       | Seatbelt Vio           |
| 65534 | 12/7/2012 | 0:30      | F             | 1985.0         | 27.0       | White       | Spe                    |

65535 rows × 14 columns

In [55]:

```
Loading [MathJax]/extensions/Safe.js )
```

Out[55]:

|                    |       |
|--------------------|-------|
| stop_date          | 0     |
| stop_time          | 0     |
| driver_gender      | 0     |
| driver_age_raw     | 0     |
| driver_age         | 0     |
| driver_race        | 0     |
| violation_raw      | 0     |
| vio                | 4060  |
| search_conducted   | 0     |
| search_type        | 63056 |
| stop_outcome       | 4060  |
| is_arrested        | 4060  |
| stop_duration      | 4060  |
| drugs_related_stop | 0     |

dtype: int64

```
In [56]: df.vio.value_counts()
```

Out[56]:

|                     |       |
|---------------------|-------|
| Speeding            | 37204 |
| Moving violation    | 11926 |
| Equipment           | 6516  |
| Other               | 3583  |
| Registration/plates | 2243  |
| Seat belt           | 3     |

Name: vio, dtype: int64

```
In [57]: df.vio.fillna('Seat belt', inplace=True)
```

```
In [58]: df
```

Out[58]:

|       | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|-------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
| 0     | 1/2/2005  | 1:55      | M             | 1985.0         | 20.0       | White       | Spe                    |
| 1     | 1/18/2005 | 8:15      | M             | 1965.0         | 40.0       | White       | Spe                    |
| 2     | 1/23/2005 | 23:15     | M             | 1972.0         | 33.0       | White       | Spe                    |
| 3     | 2/20/2005 | 17:15     | M             | 1986.0         | 19.0       | White       | Call for Se            |
| 4     | 3/14/2005 | 10:00     | F             | 1984.0         | 21.0       | White       | Spe                    |
| ...   | ...       | ...       | ...           | ...            | ...        | ...         |                        |
| 65530 | 12/6/2012 | 17:54     | F             | 1987.0         | 25.0       | White       | Spe                    |
| 65531 | 12/6/2012 | 22:22     | M             | 1954.0         | 58.0       | White       | Spe                    |
| 65532 | 12/6/2012 | 23:20     | M             | 1985.0         | 27.0       | Black       | Equipment/Inspe<br>Vio |
| 65533 | 12/7/2012 | 0:23      | F             | 1985.0         | 15.0       | Other       | Seatbelt Vio           |
| 65534 | 12/7/2012 | 0:30      | F             | 1985.0         | 27.0       | White       | Spe                    |

65535 rows × 14 columns

```
In [59]: df.isnull().sum()
```

Out[59]:

|                |   |
|----------------|---|
| stop_date      | 0 |
| stop_time      | 0 |
| driver_gender  | 0 |
| driver_age_raw | 0 |

```
driver_age      0
driver_race     0
violation_raw   0
vio             0
search_conducted 0
search_type     63056
stop_outcome    4060
is_arrested     4060
stop_duration   4060
drugs_related_stop 0
dtype: int64
```

```
In [60]: df.search_type.value_counts()
```

```
Out[60]: Incident to Arrest      1074
Probable Cause      579
Inventory           173
Reasonable Suspicion 149
Protective Frisk     126
Incident to Arrest,Inventory 104
Incident to Arrest,Probable Cause 79
Probable Cause,Reasonable Suspicion 38
Incident to Arrest,Protective Frisk 31
Incident to Arrest,Inventory,Probable Cause 26
Probable Cause,Protective Frisk 23
Inventory,Probable Cause 14
Inventory,Protective Frisk 11
Incident to Arrest,Inventory,Protective Frisk 11
Incident to Arrest,Probable Cause,Protective Frisk 10
Protective Frisk,Reasonable Suspicion 10
Incident to Arrest,Reasonable Suspicion 8
Inventory,Reasonable Suspicion 4
Incident to Arrest,Probable Cause,Reasonable Suspicion 3
Incident to Arrest,Inventory,Reasonable Suspicion 2
Inventory,Probable Cause,Protective Frisk 2
Incident to Arrest,Protective Frisk,Reasonable Suspicion 1
Inventory,Probable Cause,Reasonable Suspicion 1
Name: search_type, dtype: int64
```

```
In [61]: df.drop(columns = ['search_type'], inplace = True)
```

```
In [62]: df.columns
```

```
Out[62]: Index(['stop_date', 'stop_time', 'driver_gender', 'driver_age_raw',
               'driver_age', 'driver_race', 'violation_raw', 'vio', 'search_conducted',
               'stop_outcome', 'is_arrested', 'stop_duration', 'drugs_related_stop'],
              dtype='object')
```

```
In [63]: df.isnull().sum()
```

```
Out[63]: stop_date      0
stop_time      0
driver_gender   0
driver_age_raw  0
driver_age      0
driver_race     0
violation_raw   0
vio             0
search_conducted 0
stop_outcome    4060
is_arrested     4060
```

stop\_duration 4060  
drugs\_related\_stop 0  
dtype: int64

```
In [64]: df.stop_outcome.value_counts()
```

Out[64]: Citation 56564  
Arrest Driver 2009  
Warning 1889  
N/D 423  
No Action 339  
Arrest Passenger 251  
Name: stop\_outcome, dtype: int64

```
In [65]: df.stop_duration.value_counts()
```

Out[65]: 0-15 Min 47379  
16-30 Min 11448  
30+ Min 2647  
2 1  
Name: stop\_duration, dtype: int64

```
In [66]: df.is_arrested.value_counts()
```

Out[66]: False 59215  
True 2260  
Name: is\_arrested, dtype: int64

```
In [67]: df.stop_outcome.fillna('Arrest Passenger', inplace=True)
```

```
In [68]: df.stop_duration.fillna(2, inplace=True)
```

```
In [69]: df.is_arrested.fillna(True , inplace=True)
```

```
In [70]: df
```

Out[70]:

|       | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|-------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
| 0     | 1/2/2005  | 1:55      | M             | 1985.0         | 20.0       | White       | Spe                    |
| 1     | 1/18/2005 | 8:15      | M             | 1965.0         | 40.0       | White       | Spe                    |
| 2     | 1/23/2005 | 23:15     | M             | 1972.0         | 33.0       | White       | Spe                    |
| 3     | 2/20/2005 | 17:15     | M             | 1986.0         | 19.0       | White       | Call for Se            |
| 4     | 3/14/2005 | 10:00     | F             | 1984.0         | 21.0       | White       | Spe                    |
| ...   | ...       | ...       | ...           | ...            | ...        | ...         |                        |
| 65530 | 12/6/2012 | 17:54     | F             | 1987.0         | 25.0       | White       | Spe                    |
| 65531 | 12/6/2012 | 22:22     | M             | 1954.0         | 58.0       | White       | Spe                    |
| 65532 | 12/6/2012 | 23:20     | M             | 1985.0         | 27.0       | Black       | Equipment/Inspe<br>Vio |
| 65533 | 12/7/2012 | 0:23      | F             | 1985.0         | 15.0       | Other       | Seatbelt Vio           |
| 65534 | 12/7/2012 | 0:30      | F             | 1985.0         | 27.0       | White       | Spe                    |



65535 rows × 13 columns

```
In [71]: df.isnull().sum()
```

```
Out[71]: stop_date      0
stop_time      0
driver_gender   0
driver_age_raw  0
driver_age      0
driver_race     0
violation_raw   0
vio            0
search_conducted 0
stop_outcome     0
is_arrested      0
stop_duration    0
drugs_related_stop 0
dtype: int64
```

```
In [72]: df
```

| Out[72]: |       | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|----------|-------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
|          | 0     | 1/2/2005  | 1:55      | M             | 1985.0         | 20.0       | White       | Spe                    |
|          | 1     | 1/18/2005 | 8:15      | M             | 1965.0         | 40.0       | White       | Spe                    |
|          | 2     | 1/23/2005 | 23:15     | M             | 1972.0         | 33.0       | White       | Spe                    |
|          | 3     | 2/20/2005 | 17:15     | M             | 1986.0         | 19.0       | White       | Call for Se            |
|          | 4     | 3/14/2005 | 10:00     | F             | 1984.0         | 21.0       | White       | Spe                    |
|          | ...   | ...       | ...       | ...           | ...            | ...        | ...         |                        |
|          | 65530 | 12/6/2012 | 17:54     | F             | 1987.0         | 25.0       | White       | Spe                    |
|          | 65531 | 12/6/2012 | 22:22     | M             | 1954.0         | 58.0       | White       | Spe                    |
|          | 65532 | 12/6/2012 | 23:20     | M             | 1985.0         | 27.0       | Black       | Equipment/Inspe<br>Vio |
|          | 65533 | 12/7/2012 | 0:23      | F             | 1985.0         | 15.0       | Other       | Seatbelt Vio           |
|          | 65534 | 12/7/2012 | 0:30      | F             | 1985.0         | 27.0       | White       | Spe                    |

65535 rows × 13 columns

```
In [73]: df[df['vio'].isin(['Speeding', 'seat belt'])]
```

| Out[73]: |       | 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      |
|          | 4     | 3/14/2005 | 10:00     | F             | 1984.0         | 21.0       | White       | Speeding      |
|          | 6     | 4/1/2005  | 17:30     | M             | 1969.0         | 36.0       | White       | Speeding      |
|          | ...   | ...       | ...       | ...           | ...            | ...        | ...         | ...           |
|          | 65527 | 12/6/2012 | 15:26     | F             | 1981.0         | 31.0       | White       | Speeding      |

|              | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation_raw |
|--------------|-----------|-----------|---------------|----------------|------------|-------------|---------------|
| <b>65529</b> | 12/6/2012 | 16:00     | M             | 1994.0         | 18.0       | White       | Speeding      |
| <b>65530</b> | 12/6/2012 | 17:54     | F             | 1987.0         | 25.0       | White       | Speeding      |
| <b>65531</b> | 12/6/2012 | 22:22     | M             | 1954.0         | 58.0       | White       | Speeding      |
| <b>65534</b> | 12/7/2012 | 0:30      | F             | 1985.0         | 27.0       | White       | Speeding      |

37204 rows × 13 columns

In [74]: `df[~(df['driver_age'] == 15)]`

|              | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|--------------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
| <b>0</b>     | 1/2/2005  | 1:55      | M             | 1985.0         | 20.0       | White       | Spe                    |
| <b>1</b>     | 1/18/2005 | 8:15      | M             | 1965.0         | 40.0       | White       | Spe                    |
| <b>2</b>     | 1/23/2005 | 23:15     | M             | 1972.0         | 33.0       | White       | Spe                    |
| <b>3</b>     | 2/20/2005 | 17:15     | M             | 1986.0         | 19.0       | White       | Call for Se            |
| <b>4</b>     | 3/14/2005 | 10:00     | F             | 1984.0         | 21.0       | White       | Spe                    |
| <b>...</b>   | ...       | ...       | ...           | ...            | ...        | ...         |                        |
| <b>65529</b> | 12/6/2012 | 16:00     | M             | 1994.0         | 18.0       | White       | Spe                    |
| <b>65530</b> | 12/6/2012 | 17:54     | F             | 1987.0         | 25.0       | White       | Spe                    |
| <b>65531</b> | 12/6/2012 | 22:22     | M             | 1954.0         | 58.0       | White       | Spe                    |
| <b>65532</b> | 12/6/2012 | 23:20     | M             | 1985.0         | 27.0       | Black       | Equipment/Inspe<br>Vio |
| <b>65534</b> | 12/7/2012 | 0:30      | F             | 1985.0         | 27.0       | White       | Spe                    |

61223 rows × 13 columns

In [75]: `df`

|              | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|--------------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
| <b>0</b>     | 1/2/2005  | 1:55      | M             | 1985.0         | 20.0       | White       | Spe                    |
| <b>1</b>     | 1/18/2005 | 8:15      | M             | 1965.0         | 40.0       | White       | Spe                    |
| <b>2</b>     | 1/23/2005 | 23:15     | M             | 1972.0         | 33.0       | White       | Spe                    |
| <b>3</b>     | 2/20/2005 | 17:15     | M             | 1986.0         | 19.0       | White       | Call for Se            |
| <b>4</b>     | 3/14/2005 | 10:00     | F             | 1984.0         | 21.0       | White       | Spe                    |
| <b>...</b>   | ...       | ...       | ...           | ...            | ...        | ...         |                        |
| <b>65530</b> | 12/6/2012 | 17:54     | F             | 1987.0         | 25.0       | White       | Spe                    |
| <b>65531</b> | 12/6/2012 | 22:22     | M             | 1954.0         | 58.0       | White       | Spe                    |
| <b>65532</b> | 12/6/2012 | 23:20     | M             | 1985.0         | 27.0       | Black       | Equipment/Inspe<br>Vio |
| <b>65533</b> | 12/7/2012 | 0:23      | F             | 1985.0         | 15.0       | Other       | Seatbelt Vio           |
| <b>65534</b> | 12/7/2012 | 0:30      | F             | 1985.0         | 27.0       | White       | Spe                    |

65535 rows × 13 columns

```
In [76]: df['driver_age']=df['driver_age'].apply(lambda x:x+3)
```

```
In [77]: df
```

Out[77]:

|       | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|-------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
| 0     | 1/2/2005  | 1:55      | M             | 1985.0         | 23.0       | White       | Spe                    |
| 1     | 1/18/2005 | 8:15      | M             | 1965.0         | 43.0       | White       | Spe                    |
| 2     | 1/23/2005 | 23:15     | M             | 1972.0         | 36.0       | White       | Spe                    |
| 3     | 2/20/2005 | 17:15     | M             | 1986.0         | 22.0       | White       | Call for Se            |
| 4     | 3/14/2005 | 10:00     | F             | 1984.0         | 24.0       | White       | Spe                    |
| ...   | ...       | ...       | ...           | ...            | ...        | ...         |                        |
| 65530 | 12/6/2012 | 17:54     | F             | 1987.0         | 28.0       | White       | Spe                    |
| 65531 | 12/6/2012 | 22:22     | M             | 1954.0         | 61.0       | White       | Spe                    |
| 65532 | 12/6/2012 | 23:20     | M             | 1985.0         | 30.0       | Black       | Equipment/Inspe<br>Vio |
| 65533 | 12/7/2012 | 0:23      | F             | 1985.0         | 18.0       | Other       | Seatbelt Vio           |
| 65534 | 12/7/2012 | 0:30      | F             | 1985.0         | 30.0       | White       | Spe                    |

65535 rows × 13 columns

```
In [78]: df
```

Out[78]:

|       | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|-------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
| 0     | 1/2/2005  | 1:55      | M             | 1985.0         | 23.0       | White       | Spe                    |
| 1     | 1/18/2005 | 8:15      | M             | 1965.0         | 43.0       | White       | Spe                    |
| 2     | 1/23/2005 | 23:15     | M             | 1972.0         | 36.0       | White       | Spe                    |
| 3     | 2/20/2005 | 17:15     | M             | 1986.0         | 22.0       | White       | Call for Se            |
| 4     | 3/14/2005 | 10:00     | F             | 1984.0         | 24.0       | White       | Spe                    |
| ...   | ...       | ...       | ...           | ...            | ...        | ...         |                        |
| 65530 | 12/6/2012 | 17:54     | F             | 1987.0         | 28.0       | White       | Spe                    |
| 65531 | 12/6/2012 | 22:22     | M             | 1954.0         | 61.0       | White       | Spe                    |
| 65532 | 12/6/2012 | 23:20     | M             | 1985.0         | 30.0       | Black       | Equipment/Inspe<br>Vio |
| 65533 | 12/7/2012 | 0:23      | F             | 1985.0         | 18.0       | Other       | Seatbelt Vio           |
| 65534 | 12/7/2012 | 0:30      | F             | 1985.0         | 30.0       | White       | Spe                    |

65535 rows × 13 columns

```
In [79]: df.driver_age = df.driver_age.apply(lambda x: x-3)
```

|          |              |                  |                  |                      |                       |                   |                    |                        |
|----------|--------------|------------------|------------------|----------------------|-----------------------|-------------------|--------------------|------------------------|
| Out[80]: |              | <b>stop_date</b> | <b>stop_time</b> | <b>driver_gender</b> | <b>driver_age_raw</b> | <b>driver_age</b> | <b>driver_race</b> | <b>violation</b>       |
|          | <b>0</b>     | 1/2/2005         | 1:55             | M                    | 1985.0                | 20.0              | White              | Spe                    |
|          | <b>1</b>     | 1/18/2005        | 8:15             | M                    | 1965.0                | 40.0              | White              | Spe                    |
|          | <b>2</b>     | 1/23/2005        | 23:15            | M                    | 1972.0                | 33.0              | White              | Spe                    |
|          | <b>3</b>     | 2/20/2005        | 17:15            | M                    | 1986.0                | 19.0              | White              | Call for Se            |
|          | <b>4</b>     | 3/14/2005        | 10:00            | F                    | 1984.0                | 21.0              | White              | Spe                    |
|          | <b>...</b>   | ...              | ...              | ...                  | ...                   | ...               | ...                |                        |
|          | <b>65530</b> | 12/6/2012        | 17:54            | F                    | 1987.0                | 25.0              | White              | Spe                    |
|          | <b>65531</b> | 12/6/2012        | 22:22            | M                    | 1954.0                | 58.0              | White              | Spe                    |
|          | <b>65532</b> | 12/6/2012        | 23:20            | M                    | 1985.0                | 27.0              | Black              | Equipment/Inspe<br>Vio |
|          | <b>65533</b> | 12/7/2012        | 0:23             | F                    | 1985.0                | 15.0              | Other              | Seatbelt Vio           |
|          | <b>65534</b> | 12/7/2012        | 0:30             | F                    | 1985.0                | 27.0              | White              | Spe                    |

65535 rows × 13 columns

In [81]: `df.groupby('driver_gender').count()`

|          |                      |                  |                  |                       |                   |                    |                      |            |
|----------|----------------------|------------------|------------------|-----------------------|-------------------|--------------------|----------------------|------------|
| Out[81]: |                      | <b>stop_date</b> | <b>stop_time</b> | <b>driver_age_raw</b> | <b>driver_age</b> | <b>driver_race</b> | <b>violation_raw</b> | <b>vio</b> |
|          | <b>driver_gender</b> |                  |                  |                       |                   |                    |                      |            |
|          | <b>F</b>             | 20371            | 20371            | 20371                 | 20371             | 20371              | 20371                | 20371      |
|          | <b>M</b>             | 45164            | 45164            | 45164                 | 45164             | 45164              | 45164                | 45164      |

In [82]: `df[df['vio'] == 'Speeding'].driver_gender.value_counts()`

Out[82]: M 25517  
F 11687  
Name: driver\_gender, dtype: int64

In [83]: `df.groupby('driver_gender').search_conducted.sum()`

Out[83]: driver\_gender  
F 366  
M 2113  
Name: search\_conducted, dtype: int64

In [84]: `df.stop_duration.unique()`

Out[84]: array(['0-15 Min', '16-30 Min', '30+ Min', 2, '2'], dtype=object)

In [85]: `df['stop_duration'] = df['stop_duration'].map({'0-15 Min': 7.5, '16-30 Min': 24, '30+ Min'`

In [86]: `df`

Out[86]: **stop\_date** **stop\_time** **driver\_gender** **driver\_age\_raw** **driver\_age** **driver\_race** **violation**

|       | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation              |
|-------|-----------|-----------|---------------|----------------|------------|-------------|------------------------|
| 0     | 1/2/2005  | 1:55      | M             | 1985.0         | 20.0       | White       | Spe                    |
| 1     | 1/18/2005 | 8:15      | M             | 1965.0         | 40.0       | White       | Spe                    |
| 2     | 1/23/2005 | 23:15     | M             | 1972.0         | 33.0       | White       | Spe                    |
| 3     | 2/20/2005 | 17:15     | M             | 1986.0         | 19.0       | White       | Call for Se            |
| 4     | 3/14/2005 | 10:00     | F             | 1984.0         | 21.0       | White       | Spe                    |
| ...   | ...       | ...       | ...           | ...            | ...        | ...         |                        |
| 65530 | 12/6/2012 | 17:54     | F             | 1987.0         | 25.0       | White       | Spe                    |
| 65531 | 12/6/2012 | 22:22     | M             | 1954.0         | 58.0       | White       | Spe                    |
| 65532 | 12/6/2012 | 23:20     | M             | 1985.0         | 27.0       | Black       | Equipment/Inspe<br>Vio |
| 65533 | 12/7/2012 | 0:23      | F             | 1985.0         | 15.0       | Other       | Seatbelt Vio           |
| 65534 | 12/7/2012 | 0:30      | F             | 1985.0         | 27.0       | White       | Spe                    |

65535 rows × 13 columns

In [87]:

df.stop\_duration.mean()

Out[87]: 12.187420698181345

In [88]:

df.groupby('driver\_age').vio.describe()

Out[88]:

|            | count | unique | top              | freq |
|------------|-------|--------|------------------|------|
| driver_age |       |        |                  |      |
| 15.0       | 4312  | 6      | Seat belt        | 4055 |
| 16.0       | 34    | 5      | Speeding         | 18   |
| 17.0       | 449   | 5      | Speeding         | 338  |
| 18.0       | 1345  | 6      | Speeding         | 980  |
| 19.0       | 2388  | 5      | Speeding         | 1655 |
| ...        | ...   | ...    | ...              | ...  |
| 83.0       | 2     | 2      | Speeding         | 1    |
| 84.0       | 3     | 1      | Speeding         | 3    |
| 85.0       | 1     | 1      | Moving violation | 1    |
| 86.0       | 6     | 3      | Speeding         | 3    |
| 88.0       | 2     | 1      | Speeding         | 2    |

73 rows × 4 columns

In [90]:

df.sort\_values(by= ['driver\_age'])

Out[90]:

|       | stop_date | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation_raw         |
|-------|-----------|-----------|---------------|----------------|------------|-------------|-----------------------|
| 23620 | 2/5/2008  | 17:43     | F             | 1985.0         | 15.0       | Other       | Seatbelt<br>Violation |

|       | stop_date  | stop_time | driver_gender | driver_age_raw | driver_age | driver_race | violation_raw           |
|-------|------------|-----------|---------------|----------------|------------|-------------|-------------------------|
| 36431 | 7/28/2009  | 0:24      | F             | 1985.0         | 15.0       | Other       | Seatbelt Violation      |
| 36426 | 7/27/2009  | 21:53     | F             | 1985.0         | 15.0       | Other       | Seatbelt Violation      |
| 7190  | 6/5/2006   | 20:46     | F             | 1985.0         | 15.0       | Other       | Seatbelt Violation      |
| 36413 | 7/27/2009  | 3:20      | F             | 1985.0         | 15.0       | Other       | Seatbelt Violation      |
| ...   | ...        | ...       | ...           | ...            | ...        | ...         | ...                     |
| 23847 | 2/15/2008  | 14:34     | M             | 1922.0         | 86.0       | White       | Speeding                |
| 38468 | 11/17/2009 | 15:33     | M             | 1923.0         | 86.0       | White       | Other Traffic Violation |
| 48329 | 3/20/2011  | 18:32     | M             | 1925.0         | 86.0       | White       | Call for Service        |
| 9829  | 9/6/2006   | 12:15     | M             | 1918.0         | 88.0       | White       | Speeding                |
| 50377 | 6/20/2011  | 12:05     | M             | 1923.0         | 88.0       | White       | Speeding                |

65535 rows × 13 columns

In [ ]: