Working on Real Project with Python

(A part of Big Data Analysis)

Police Dataset

Here, The data from a Poilce check post is given The data is available as a csv file.

In [2]:
```python
import pandas as pd
```

In [3]:
```python
data = pd.read_csv('3. Police Data.csv')
data.head()
```

Out[3]:

	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

In [4]:
```python
data.shape
```

Out[4]:
```
(65535, 15)
```

For Data Cleaning

Remove the column that only contains missing values

In [6]:
```python
data.isnull()
```

Out[6]:

	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	False	False	True	False	False	False	False	False	False	False	True	False	False	False	False
1	False	False	True	False	False	False	False	False	False	False	True	False	False	False	False
2	False	False	True	False	False	False	False	False	False	False	True	False	False	False	False
3	False	False	True	False	False	False	False	False	False	False	True	False	False	False	False
4	False	False	True	False	False	False	False	False	False	False	True	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
65530	False	False	True	False	False	False	False	False	False	False	True	False	False	False	False
65531	False	False	True	False	False	False	False	False	False	False	True	False	False	False	False
65532	False	False	True	False	False	False	False	False	False	False	True	False	False	False	False
65533	False	False	True	True	True	True	True	True	True	False	True	True	True	True	False
65534	False	False	True	False	False	False	False	False	False	False	True	False	False	False	False

65535 rows × 15 columns

In [7]:
```python
data.isnull().sum()
```

Out[7]:
```
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 [9]:
```python
data.drop(columns = 'country_name' , inplace = True)
```

In [10]:
```python
data
```

Out[10]:

	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
65530	12/6/2012	17:54	F	1987.0	25.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
65531	12/6/2012	22:22	M	1954.0	58.0	White	Speeding	Speeding	False	NaN	Warning	False	0-15 Min	False
65532	12/6/2012	23:20	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	False	NaN	NaN	NaN	NaN	False
65534	12/7/2012	0:30	F	1985.0	27.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False

65535 rows × 14 columns

Based on Filtering and value plus

For speeding,were Men or Women stopped more often?

In [11]:
```python
data.head()
```

Out[11]:

	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 [14]:
```python
data[data.violation == 'Speeding'].driver_gender.value_counts()
```

Out[14]:
```
M    25517
F    11686
Name: driver_gender, dtype: int64
```

Groupby

Does gender affect who gets searched during a stop?

In [15]:
```python
data.groupby('driver_gender').search_conducted.sum()
```

Out[15]:
```
driver_gender
F     366
M    2113
Name: search_conducted, dtype: int64
```

In [16]:
```python
data.search_conducted.value_counts()
```

Out[16]:
```
False    63056
True      2479
Name: search_conducted, dtype: int64
```

Mapping and Data-Type casting

What is the maen stop_duration?

In [17]:
```python
data.stop_duration.value_counts()
```

Out[17]:
```
0-15 Min     47379
16-30 Min    11448
30+ Min       2647
2                1
Name: stop_duration, dtype: int64
```

In [18]:
```python
data['stop_duration'] = data['stop_duration'].map({'0-15 Min': 7.5 , '16-30 Min' : 24 , '30+ Min' : 45})
```

In [19]:
```python
data
```

Out[19]:

	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 [20]:
```python
data['stop_duration'].mean()
```

Out[20]:
```
12.187420698181345
```

Groupby , Describe

Compare the age distribution for each violation

In [21]:
```python
data.groupby('violation').driver_age.describe()
```

Out[21]:

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 [ ]:

In [ ]:

In [ ]:

In [ ]:

Matplotlib

In [23]:
```python
import numpy as np
import matplotlib.pyplot as plt
```

In [24]:
```python
data.head()
```

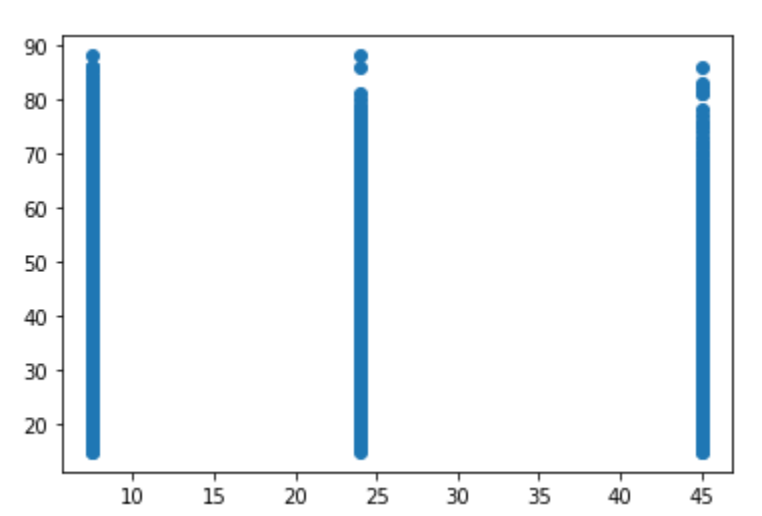
Out[24]:

	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

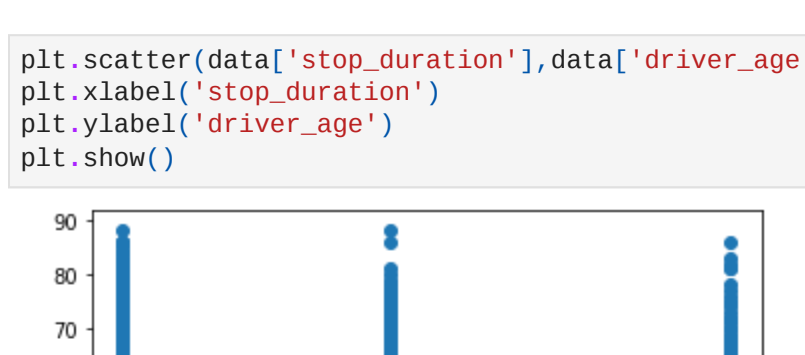
Scatterplot

In [38]:
```python
plt.scatter(data['stop_duration'],data['driver_age'])
```

Out[38]:
```
<matplotlib.collections.PathCollection at 0x2131ff1bf40>
```

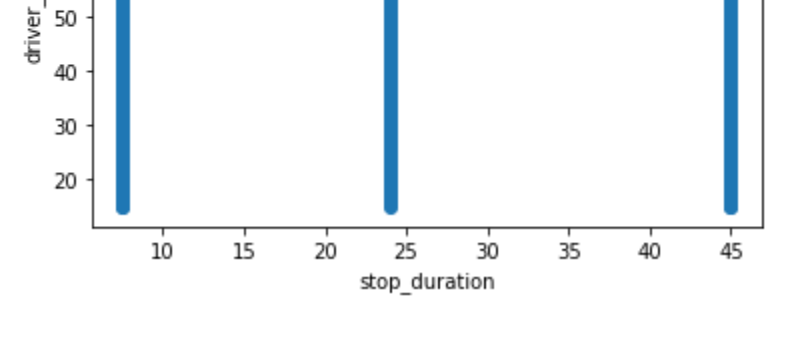


In [39]:
```python
plt.scatter(data['stop_duration'],data['driver_age'])
plt.xlabel('stop_duration')
plt.ylabel('driver_age')
plt.show()
```



Create Histogram on Driver_age

In [35]:
```python
plt.hist(data['driver_age'])
plt.show()
```



In [61]:
```python
data.head()
```

Out[61]:

	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	1:55	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
1	8:15	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
2	23:15	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False
3	17:15	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	24.0	False
4	10:00	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	False

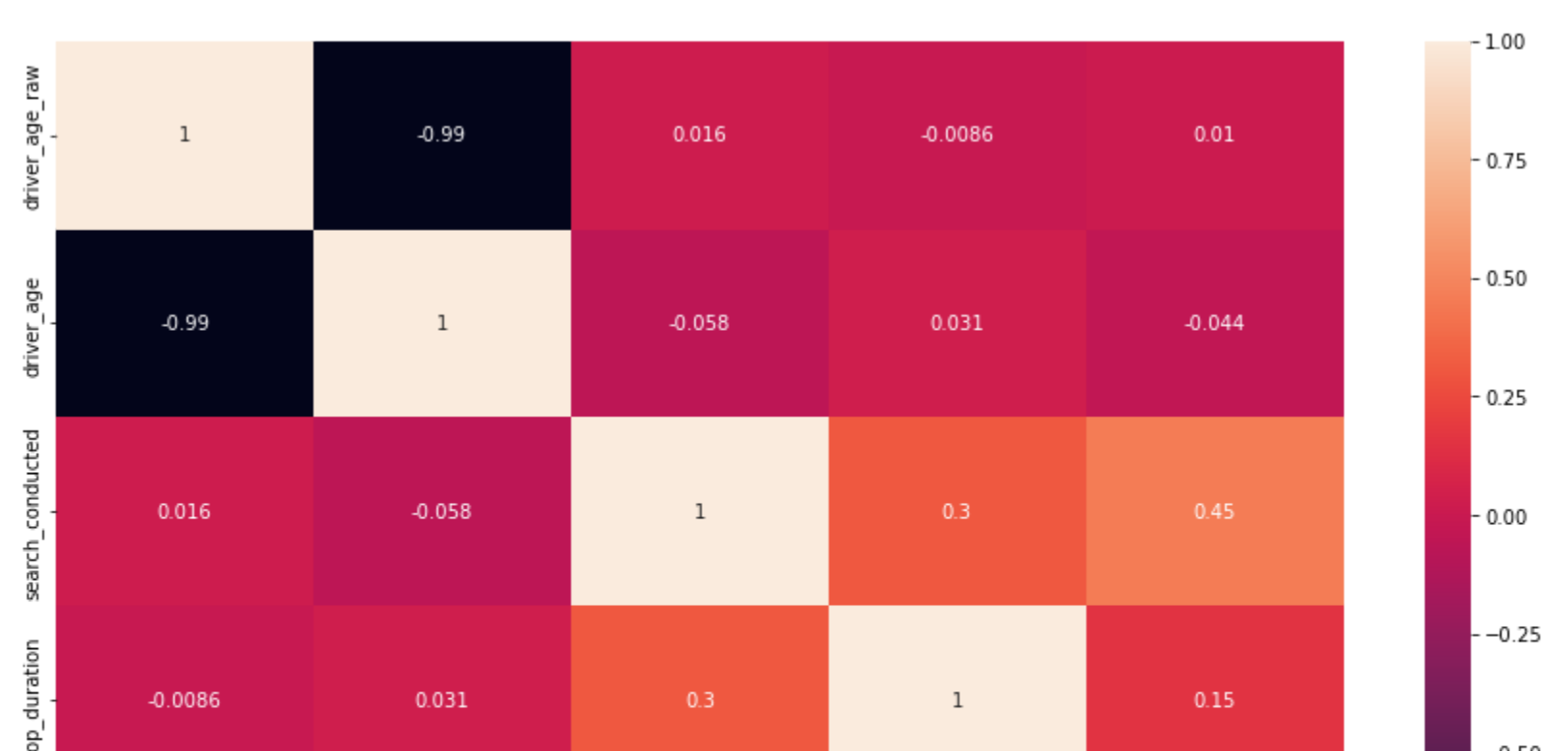
In [62]:
```python
data.shape
```

Out[62]:
```
(65535, 13)
```

In [63]:
```python
import seaborn as sns
```

In [64]:
```python
plt.figure(figsize = (15 , 9))
sns.heatmap(data.corr() , annot = True)
```

Out[64]:
```
<AxesSubplot:>
```



In [92]:
```python
data.head()
```

Out[92]:

	driver_age_raw	driver_age	violation_raw	violation	stop_outcome	is_arrested	stop_duration	drugs_related_stop
0	1985.0	20.0	Speeding	Speeding	Citation	False	7.5	False
1	1965.0	40.0	Speeding	Speeding	Citation	False	7.5	False
2	1972.0	33.0	Speeding	Speeding	Citation	False	7.5	False
3	1986.0	19.0	Call for Service	Other	Arrest Driver	True	24.0	False
4	1984.0	21.0	Speeding	Speeding	Citation	False	7.5	False

In [93]:
```python
data.drop(columns = 'violation',inplace = True)
```

In [94]:
```python
data
```

Out[94]:

	driver_age_raw	driver_age	violation_raw	stop_outcome	is_arrested	stop_duration	drugs_related_stop
0	1985.0	20.0	Speeding	Citation	False	7.5	False
1	1965.0	40.0	Speeding	Citation	False	7.5	False
2	1972.0	33.0	Speeding	Citation	False	7.5	False
3	1986.0	19.0	Call for Service	Arrest Driver	True	24.0	False
4	1984.0	21.0	Speeding	Citation	False	7.5	False
...	...	...	...	...	...	...	...
65530	1987.0	25.0	Speeding	Citation	False	7.5	False
65531	1954.0	58.0	Speeding	Warning	False	7.5	False
65532	1985.0	27.0	Equipment/Inspection Violation	Citation	False	7.5	False
65533	NaN	NaN	NaN	NaN	NaN	NaN	False
65534	1985.0	27.0	Speeding	Citation	False	7.5	False

65535 rows × 7 columns

In [95]:
```python
data.drop(columns = 'violation_raw' , inplace = True)
```

In [96]:
```python
data
```

Out[96]:

	driver_age_raw	driver_age	stop_outcome	is_arrested	stop_duration	drugs_related_stop
0	1985.0	20.0	Citation	False	7.5	False
1	1965.0	40.0	Citation	False	7.5	False
2	1972.0	33.0	Citation	False	7.5	False
3	1986.0	19.0	Arrest Driver	True	24.0	False
4	1984.0	21.0	Citation	False	7.5	False
...	...	...	...	...	...	...
65530	1987.0	25.0	Citation	False	7.5	False
65531	1954.0	58.0	Warning	False	7.5	False
65532	1985.0	27.0	Citation	False	7.5	False
65533	NaN	NaN	NaN	NaN	NaN	False
65534	1985.0	27.0	Citation	False	7.5	False

65535 rows × 6 columns

In [100…
```python
data.drop(['stop_outcome'],axis=1 , inplace = True)
data.head()
```

Out[100]:

	driver_age_raw	driver_age	is_arrested	stop_duration	drugs_related_stop
0	1985.0	20.0	False	7.5	False
1	1965.0	40.0	False	7.5	False
2	1972.0	33.0	False	7.5	False
3	1986.0	19.0	True	24.0	False
4	1984.0	21.0	False	7.5	False

In [103…
```python
data.drop(['is_arrested' , 'drugs_related_stop'],axis=1 , inplace = True)
data.head()
```

Out[103]:

	driver_age_raw	driver_age	stop_duration
0	1985.0	20.0	7.5
1	1965.0	40.0	7.5
2	1972.0	33.0	7.5
3	1986.0	19.0	24.0
4	1984.0	21.0	7.5

CREATE X and Y from the data set

In [105…
```python
x = data.iloc[:, :-1]
y = data.iloc[: , -1]
```

Divide the dataset into train , test , spilt

In [106…
```python
from sklearn.model_selection import train_test_split
x_train , x_test , y_train , y_test = train_test_split(x , y ,test_size = 0.25)
x_train.shape , x_test.shape
```

Out[106]:
```
((49151, 2), (16384, 2))
```

Feature scaling

In [108…
```python
data
```

Out[108]:

	driver_age_raw	driver_age	stop_duration
0	1985.0	20.0	7.5
1	1965.0	40.0	7.5
2	1972.0	33.0	7.5
3	1986.0	19.0	24.0
4	1984.0	21.0	7.5
...	...	...	...
65530	1987.0	25.0	7.5
65531	1954.0	58.0	7.5
65532	1985.0	27.0	7.5
65533	NaN	NaN	NaN
65534	1985.0	27.0	7.5

65535 rows × 3 columns

In [109…
```python
from sklearn.preprocessing import StandardScaler

ss = StandardScaler()
ss.fit_transform(x_train)
ss.fit_transform(x_test)
```

Out[109]:
```
array([[ 0.13740734, -0.87260212],
       [ 0.14419654, -0.87260212],
       [ 0.07630453, -0.32480747],
       ...,
       [ 0.02199092,  0.53601268],
       [ 0.06272612, -0.16829472],
       [ 0.11703973, -0.87260212]])
```

In [ ]: