Police Data Analisys

January 24, 2024

1 Working on Real Project with python

(A Part Of Big Data Analysis)

Police Dataset

2 Here,

the data from a ploice check post is given. This data is available as a CSV file. We are going to analyze this data set using the pandas DataFrame.

[13]:
```
import pandas as pd
pol=pd.read_csv(r"C:\Users\Dell\Desktop\Police data in pyton\Police data.csv")
pol.head(1)
```

[13]:
```
    stop_date stop_time  country_name driver_gender  …  stop_outcome
is_arrested stop_duration drugs_related_stop
0  01-02-2005     01:55           NaN             M  …      Citation
False      0-15 Min             False

[1 rows x 15 columns]
```

[4]:
```
pol.dtypes
```

[4]:
```
stop_date             object
stop_time             object
country_name         float64
driver_gender         object
driver_age_raw       float64
driver_age           float64
driver_race           object
violation_raw         object
violation             object
search_conducted        bool
search_type           object
stop_outcome          object
is_arrested           object
stop_duration         object
```

```
drugs_related_stop       bool
dtype: object
```

[7]:
```
pol.shape
```

[7]:
```
(65535, 15)
```

[8]:
```
pol.count()
```

[8]:
```
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
```

[9]:
```
pol.index
```

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

[11]:
```
pol.nunique()
```

[11]:
```
stop_date             2651
stop_time             1432
country_name             0
driver_gender            2
driver_age_raw          93
driver_age              73
driver_race              5
violation_raw           12
violation                6
search_conducted         2
search_type             23
stop_outcome             6
is_arrested              2
stop_duration            4
drugs_related_stop       2
dtype: int64
```

[12]:
```
pol.head(5)
```

[12]:
```
    stop_date stop_time  country_name driver_gender  …   stop_outcome
is_arrested stop_duration drugs_related_stop
0  01-02-2005     01:55           NaN             M  …       Citation
False      0-15 Min             False
1   1/18/2005     08:15           NaN             M  …       Citation
False      0-15 Min             False
2   1/23/2005     23:15           NaN             M  …       Citation
False      0-15 Min             False
3   2/20/2005     17:15           NaN             M  …  Arrest Driver
True     16-30 Min             False
4   3/14/2005     10:00           NaN             F  …       Citation
False      0-15 Min             False

[5 rows x 15 columns]
```

3 Instruction (For Data Cleaning)

1 . Remove the columns that only contains mssing values

[33]:
```
pol.isnull().sum()
```

[33]:
```
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
```

```
                                        pol.drop(columns="country_name",inplace=True)   ##
```

[36]:
```
pol.isnull().sum()
```

[36]:
```
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
```

[37]:
```
pol
```

[37]:
```
        stop_date stop_time driver_gender  driver_age_raw  …   stop_outcome
is_arrested stop_duration drugs_related_stop
0      01-02-2005     01:55             M          1985.0  …       Citation
False      0-15 Min              False
1       1/18/2005     08:15             M          1965.0  …       Citation
False      0-15 Min              False
2       1/23/2005     23:15             M          1972.0  …       Citation
False      0-15 Min              False
3       2/20/2005     17:15             M          1986.0  …  Arrest Driver
True      16-30 Min             False
4       3/14/2005     10:00             F          1984.0  …       Citation
False      0-15 Min              False
…           …       …           …              …  …            …
…           …                 …
65530  12-06-2012     17:54             F          1987.0  …       Citation
False      0-15 Min              False
65531  12-06-2012     22:22             M          1954.0  …        Warning
False      0-15 Min              False
65532  12-06-2012     23:20             M          1985.0  …       Citation
False      0-15 Min              False
65533  12-07-2012     00:23           NaN             NaN  …            NaN
NaN           NaN             False
65534  12-07-2012     00:30             F          1985.0  …       Citation
False      0-15 Min              False

[65535 rows x 14 columns]
```

[32]:
```
pol
```

[32]:
```
        stop_date stop_time driver_gender  driver_age_raw  …   stop_outcome
is_arrested stop_duration drugs_related_stop
0      01-02-2005     01:55             M          1985.0  …       Citation
False      0-15 Min              False
1       1/18/2005     08:15             M          1965.0  …       Citation
```

```
False      0-15 Min              False
2       1/23/2005     23:15             M          1972.0  …       Citation
False      0-15 Min              False
3       2/20/2005     17:15             M          1986.0  …  Arrest Driver
True     16-30 Min             False
4       3/14/2005     10:00             F          1984.0  …       Citation
False      0-15 Min              False
…           …       …           …              …  …            …
…           …               …
65530   12-06-2012     17:54             F          1987.0  …       Citation
False      0-15 Min              False
65531   12-06-2012     22:22             M          1954.0  …        Warning
False      0-15 Min              False
65532   12-06-2012     23:20             M          1985.0  …       Citation
False      0-15 Min              False
65533   12-07-2012     00:23           NaN             NaN  …            NaN
NaN           NaN             False
65534   12-07-2012     00:30             F          1985.0  …       Citation
False      0-15 Min              False

[65535 rows x 14 columns]
```

4 Question (Based on Filtering + Values counts)

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

[46]:
```
import pandas as pd
pl=pd.read_csv(r"C:\Users\Dell\Desktop\987\Police data.csv")
pl.head(1)
```

[46]:
```
    stop_date stop_time  country_name driver_gender  …  stop_outcome
is_arrested stop_duration drugs_related_stop
0  01-02-2005     01:55           NaN             M  …      Citation
False      0-15 Min              False

[1 rows x 15 columns]
```

[47]:
```
pl.violation.value_counts()
```

[47]:
```
violation
Speeding               37204
Moving violation       11926
Equipment               6516
Other                   3583
Registration/plates     2243
Seat belt                  3
Name: count, dtype: int64
```

[56]:
```
pl[pl.violation=="Speeding"].driver_gender.value_counts()    # Answer
```

[56]:
```
driver_gender
M    25517
F    11686
Name: count, dtype: int64
```

[59]:
```

```

[59]:
```
array(['Speeding', 'Other', 'Equipment', 'Moving violation', nan,
       'Registration/plates', 'Seat belt'], dtype=object)
```

5 Question (Group by)

3. Does gender affect who gets searched a stop

[60]:
```
pl
```

[60]:
```
        stop_date stop_time  country_name driver_gender  …   stop_outcome
is_arrested stop_duration drugs_related_stop
0      01-02-2005     01:55           NaN             M  …       Citation
False      0-15 Min              False
1       1/18/2005     08:15           NaN             M  …       Citation
False      0-15 Min              False
2       1/23/2005     23:15           NaN             M  …       Citation
False      0-15 Min              False
3       2/20/2005     17:15           NaN             M  …  Arrest Driver
True     16-30 Min             False
4       3/14/2005     10:00           NaN             F  …       Citation
False      0-15 Min              False
…           …       …           …            …  …            …
…           …                 …
65530  12-06-2012     17:54           NaN             F  …       Citation
False      0-15 Min              False
65531  12-06-2012     22:22           NaN             M  …        Warning
False      0-15 Min              False
65532  12-06-2012     23:20           NaN             M  …       Citation
False      0-15 Min              False
65533  12-07-2012     00:23           NaN           NaN  …            NaN
NaN           NaN             False
65534  12-07-2012     00:30           NaN             F  …       Citation
False      0-15 Min              False

[65535 rows x 15 columns]
```

[65]:
```
pl.search_conducted.value_counts()
```

[65]:
```
search_conducted
False    63056
True      2479
Name: count, dtype: int64
```

[63]:
```
pl.groupby("driver_gender").search_conducted.sum()  # answer
```

[63]:
```
driver_gender
F     366
M    2113
Name: search_conducted, dtype: int64
```

[71]:
```
pl[pl.search_conducted==True].driver_gender.value_counts()  # Answer
```

[71]:
```
driver_gender
M    2113
F     366
Name: count, dtype: int64
```

[75]:
```
pl[pl.search_conducted==True].driver_gender.count()
```

[75]:
```
2479
```

6 Question (mapping + data +type casting)

4 . What is the Mean stop _duration?

[145]:
```
import pandas as pd
pls=pd.read_csv(r"C:\Users\Dell\Desktop\898\Police2.csv")
pls.head(5)
```

[145]:
```
    stop_date stop_time  country_name driver_gender  …   stop_outcome
is_arrested stop_duration drugs_related_stop
0  01-02-2005     01:55           NaN             M  …       Citation
False      0-15 Min              False
1   1/18/2005     08:15           NaN             M  …       Citation
False      0-15 Min              False
2   1/23/2005     23:15           NaN             M  …       Citation
False      0-15 Min              False
3   2/20/2005     17:15           NaN             M  …  Arrest Driver
True     16-30 Min             False
4   3/14/2005     10:00           NaN             F  …       Citation
False      0-15 Min              False

[5 rows x 15 columns]
```

[149]:
```
pls.stop_duration.value_counts()
```

[149]:
```
stop_duration
0-15 Min     47379
16-30 Min    11448
30+ Min       2647
2                1
Name: count, dtype: int64
```

[150]:
```
pls.stop_duration.map({"0-15 Min":7.5,"16-30 Min":23,"30+ Min":45})
```

[150]:
```
0         7.5
1         7.5
2         7.5
3        23.0
4         7.5
         …
65530     7.5
65531     7.5
65532     7.5
65533     NaN
65534     7.5
Name: stop_duration, Length: 65535, dtype: float64
```

[152]:
```
pls.stop_duration=pls.stop_duration.map({"0-15 Min":7.5,"16-30 Min":23,"30+␣
↪Min":45})
```

[154]:
```
pls.stop_duration.mean()
```

[154]:
```
12.001195627419722
```

[157]:
```
pls.dtypes
```

[157]:
```
stop_date              object
stop_time              object
country_name          float64
driver_gender          object
driver_age_raw        float64
driver_age            float64
driver_race            object
violation_raw          object
violation              object
search_conducted         bool
search_type            object
stop_outcome           object
is_arrested            object
stop_duration         float64
drugs_related_stop       bool
dtype: object
```

7 Question (Group by ,Describe)

5. Compare the age distributions for each violation.

[159]:
```
pls.head()
```

[159]:
```
    stop_date stop_time  country_name driver_gender  …   stop_outcome
is_arrested stop_duration drugs_related_stop
0  01-02-2005     01:55           NaN             M  …       Citation
False           7.5              False
1   1/18/2005     08:15           NaN             M  …       Citation
False           7.5              False
2   1/23/2005     23:15           NaN             M  …       Citation
False           7.5              False
3   2/20/2005     17:15           NaN             M  …  Arrest Driver
True          23.0              False
4   3/14/2005     10:00           NaN             F  …       Citation
False           7.5              False

[5 rows x 15 columns]
```

[160]:
```
pls.violation.unique()
```

[160]:
```
array(['Speeding', 'Other', 'Equipment', 'Moving violation', nan,
       'Registration/plates', 'Seat belt'], dtype=object)
```

[163]:
```
pls.groupby("violation").driver_age.describe()     # Answer
```

[163]:

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

[ ]:
```

```