Exploratory Data Analysis- Stanford Open Policing

Data on Traffic and Pedestrian Stops by Police in Rhode Island

Winnie Liu

Outline

1) Introduction

2) Dataset

3) Preparing The Data (ETL)

    - Address the missing value

    - Fix data type

    - Create date-time index

4) Answering the Questions

Introduction

- This is a dataset on Traffic and Pedestrian Stops by Police in Rhode Island
- This project is made for practicing exploratory data analysis by using pandas in Python
- In this project, I am going to answer the following 5 questions:

  1. Do men or women speed more often?
  2. Does gender affect who gets a ticket for speeding?
  3. Does gender affect whose vehicle is searched?
  4. Are drug-related stops on the rise?
  5. Which year had the least number of stops?

Dataset

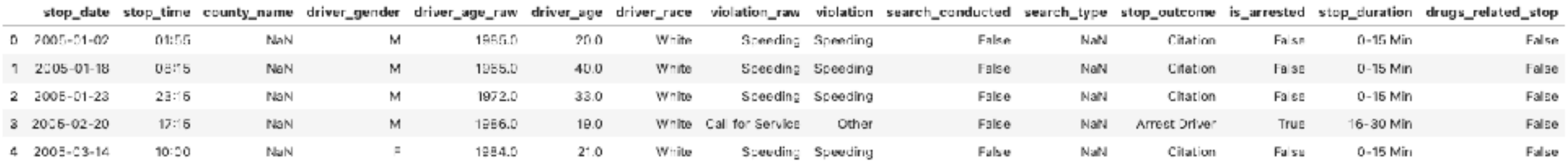
- The data is downloaded from Kaggle
- Data source: https://www.kaggle.com/faressayah/stanford-open-policing-project

Police_project.csv

91741 rows, 15 columns

Column Names:

stop_date, stop_time, county_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

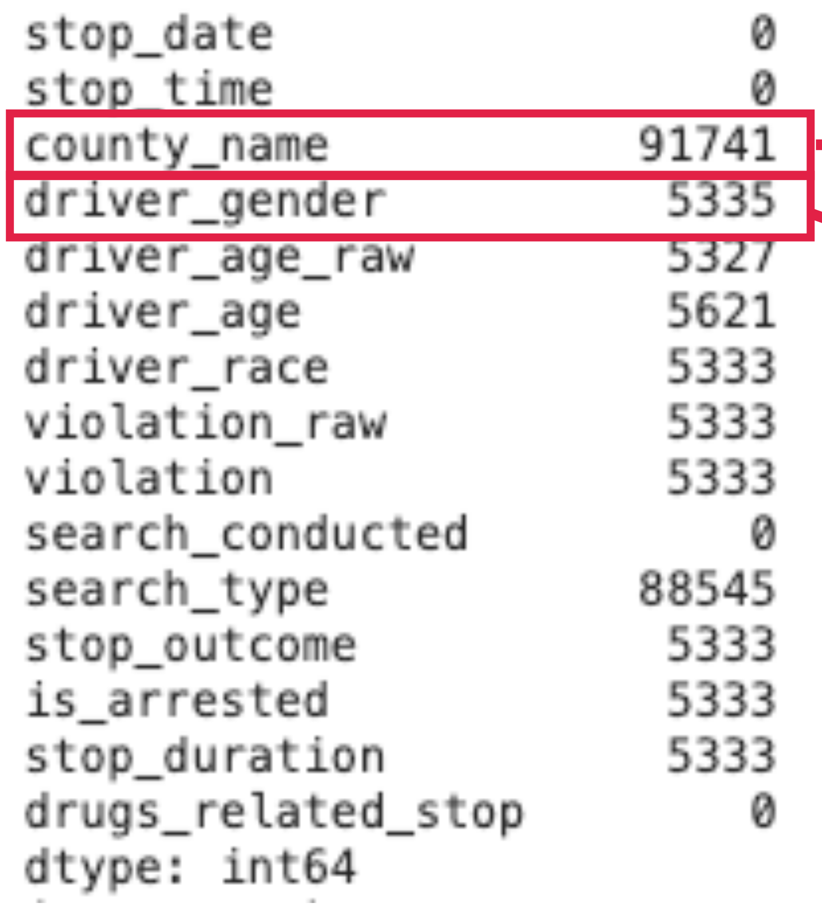


	stop_date	stop_time	county_name	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_conducted	search_type	stop_outcome	is_arrested	stop_duration	drugs_related_stop
0	2005-01-02	01:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
1	2005-01-18	08:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
2	2005-01-23	23:15	NaN	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
3	2005-02-20	17:15	NaN	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	16-30 Min	False
4	2005-03-14	10:00	NaN	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False

Preparing The Data (ETL)

Dealing with missing values

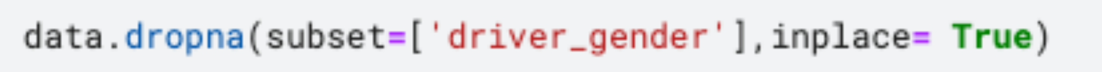
Missing values in the data:



```
stop_date                 0
stop_time                 0
county_name           91741
driver_gender          5335
driver_age_raw         5327
driver_age             5621
driver_race            5333
violation_raw          5333
violation              5333
search_conducted          0
search_type           88545
stop_outcome           5333
is_arrested            5333
stop_duration          5333
drugs_related_stop        0
dtype: int64
```

County_name:
Drop the whole column since it has no value

Driver_gender:
Drop missing value since only a small fraction are missing

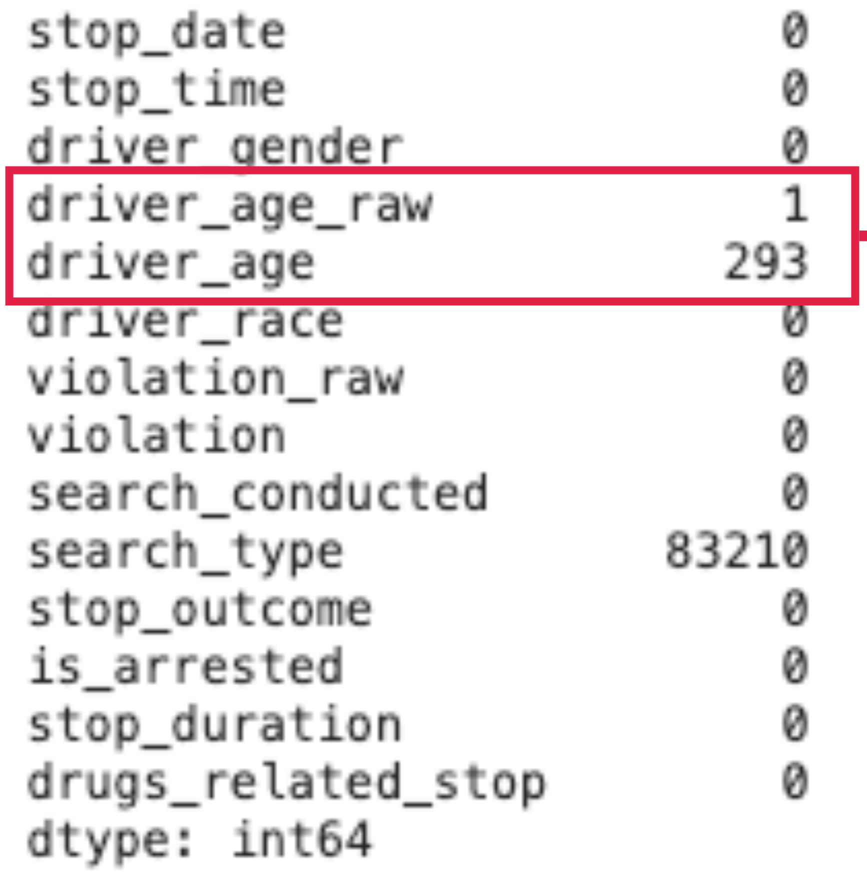


```
data.dropna(subset=['driver_gender'],inplace= True)
```

Preparing The Data (ETL)

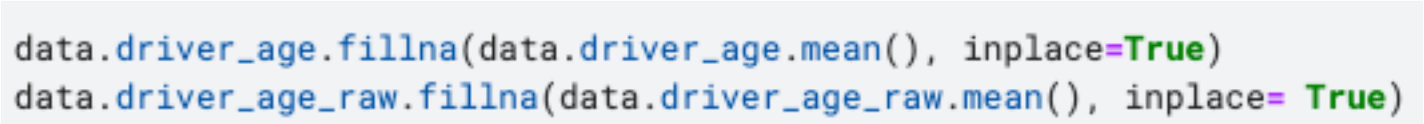
Dealing with missing values

Missing values in the data:



```
stop_date                0
stop_time                0
driver_gender            0
driver_age_raw           1
driver_age             293
driver_race              0
violation_raw            0
violation                0
search_conducted         0
search_type          83210
stop_outcome             0
is_arrested              0
stop_duration            0
drugs_related_stop       0
dtype: int64
```

Driver_age:
Fill missing values by mean

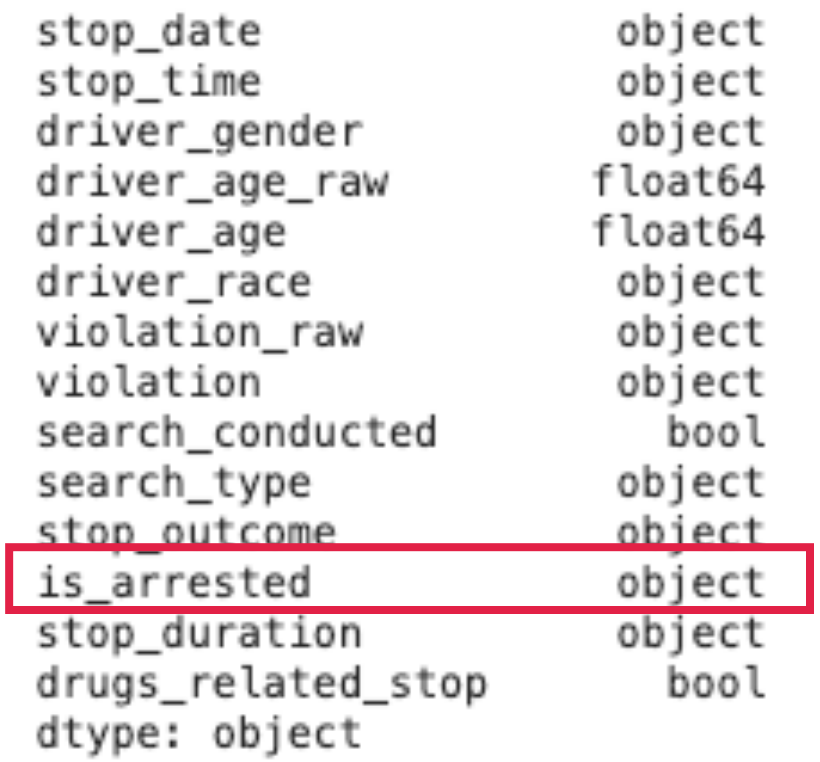


```
data.driver_age.fillna(data.driver_age.mean(), inplace=True)
data.driver_age_raw.fillna(data.driver_age_raw.mean(), inplace= True)
```

Preparing The Data (ETL)

Fix data type

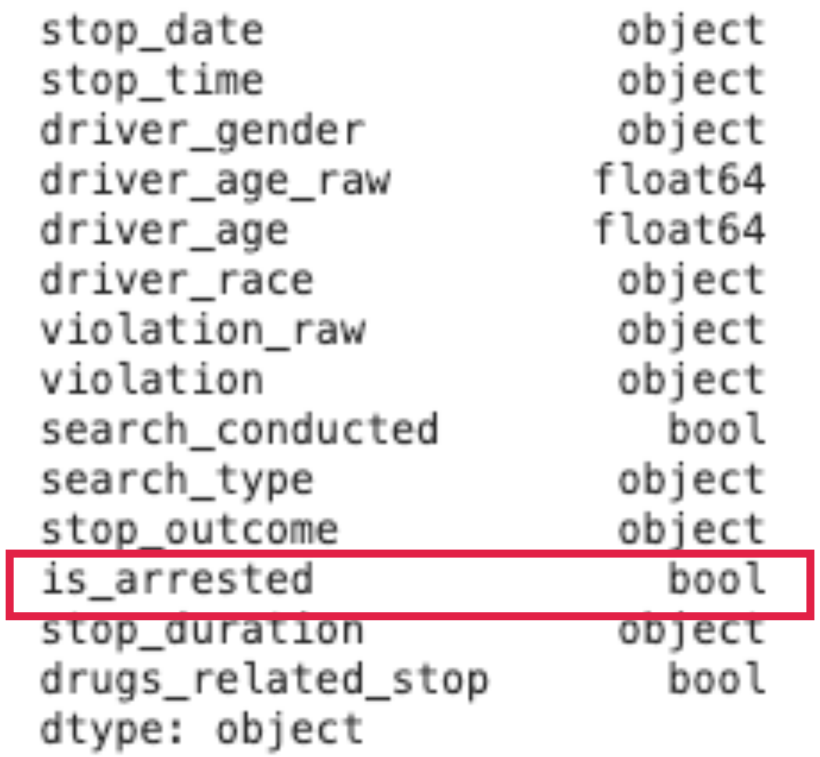
Original



```
stop_date              object
stop_time              object
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
```

Change into boolean

Fixed

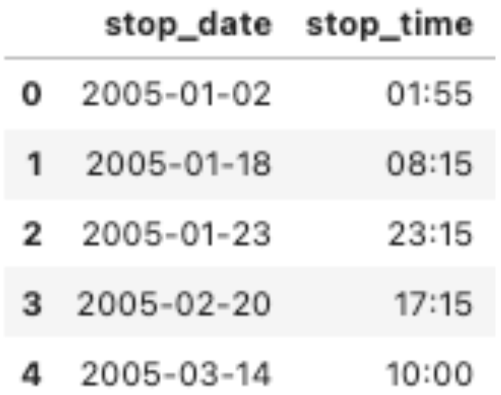


```
stop_date              object
stop_time              object
driver_gender          object
driver_age_raw        float64
driver_age            float64
driver_race            object
violation_raw          object
violation              object
search_conducted         bool
search_type            object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

Preparing The Data (ETL)

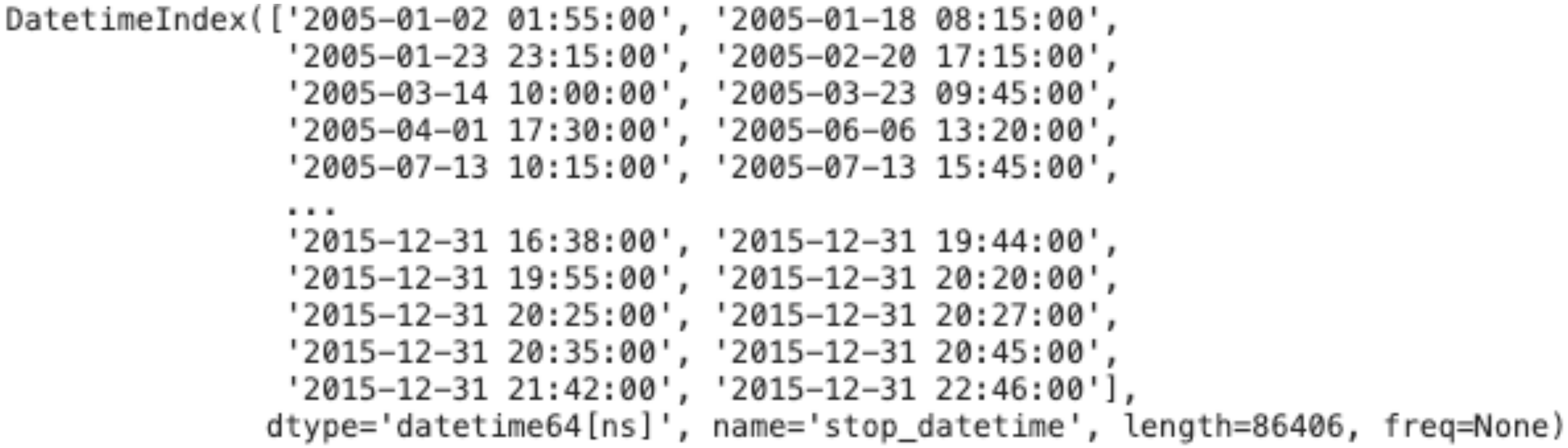
Create a date-time index

- Combine stop_date and stop_time into one column
- Convert it to the date-time format
- Set it as index



	stop_date	stop_time
0	2005-01-02	01:55
1	2005-01-18	08:15
2	2005-01-23	23:15
3	2005-02-20	17:15
4	2005-03-14	10:00

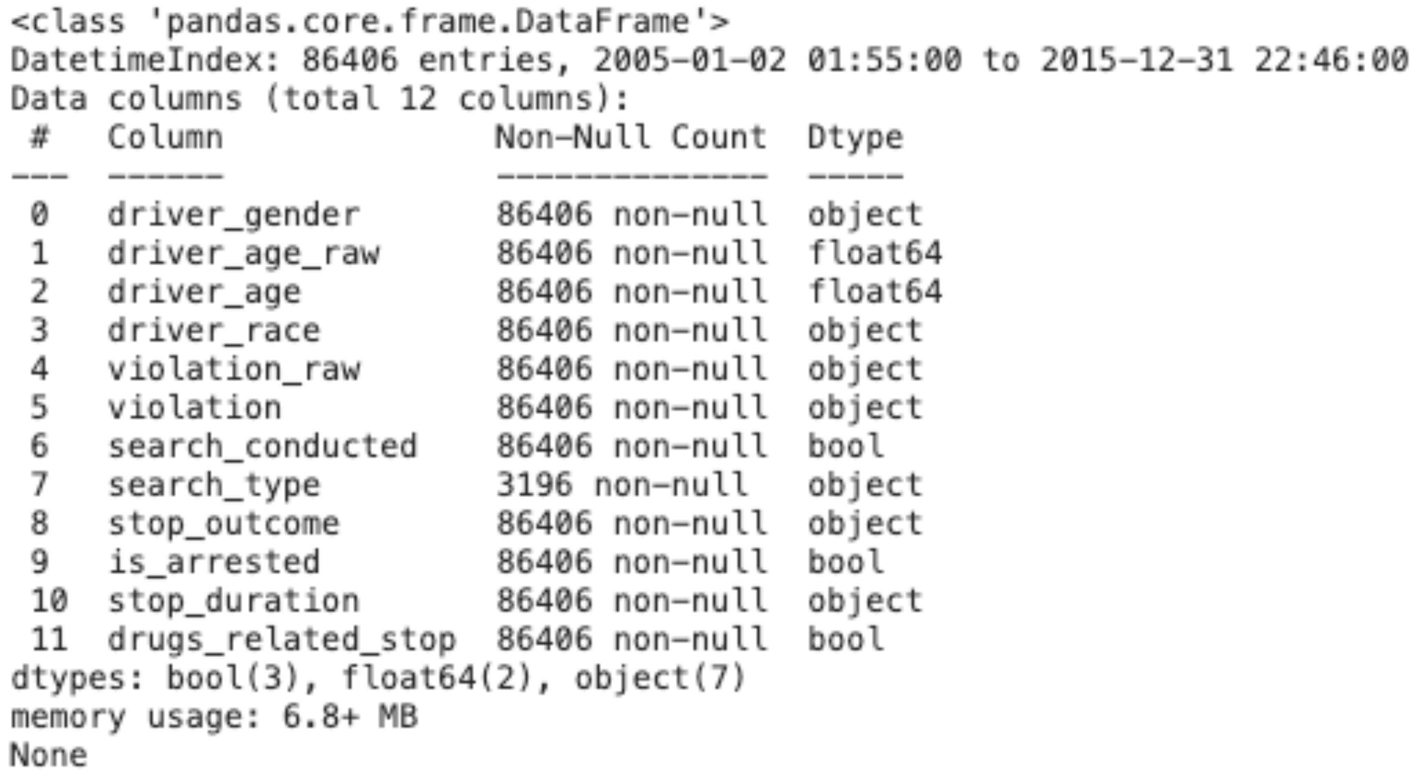
Result:



```
DatetimeIndex(['2005-01-02 01:55:00', '2005-01-18 08:15:00',
               '2005-01-23 23:15:00', '2005-02-20 17:15:00',
               '2005-03-14 10:00:00', '2005-03-23 09:45:00',
               '2005-04-01 17:30:00', '2005-06-06 13:20:00',
               '2005-07-13 10:15:00', '2005-07-13 15:45:00',
               ...
               '2015-12-31 16:38:00', '2015-12-31 19:44:00',
               '2015-12-31 19:55:00', '2015-12-31 20:20:00',
               '2015-12-31 20:25:00', '2015-12-31 20:27:00',
               '2015-12-31 20:35:00', '2015-12-31 20:45:00',
               '2015-12-31 21:42:00', '2015-12-31 22:46:00'],
              dtype='datetime64[ns]', name='stop_datetime', length=86406, freq=None)
```

Preparing The Data (ETL)

View the ETL Result



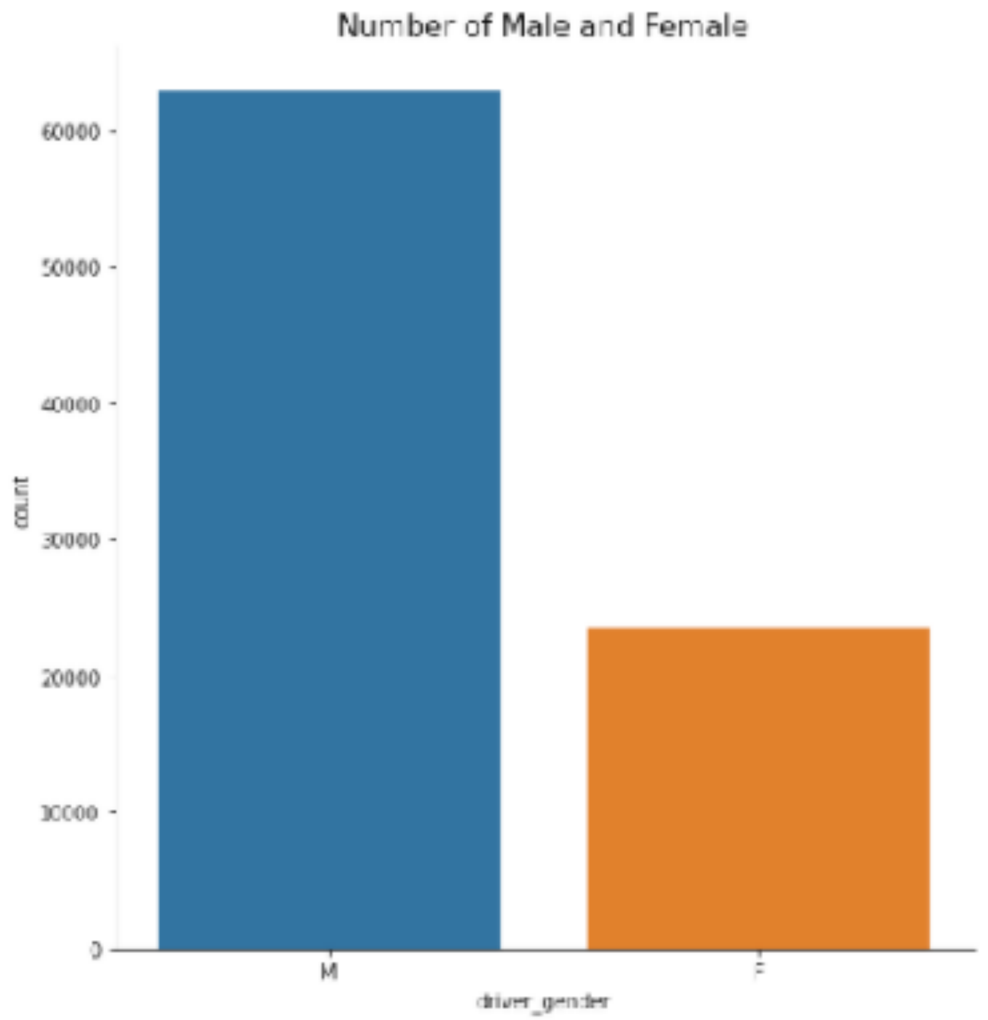
```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 86406 entries, 2005-01-02 01:55:00 to 2015-12-31 22:46:00
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   driver_gender       86406 non-null  object
 1   driver_age_raw      86406 non-null  float64
 2   driver_age          86406 non-null  float64
 3   driver_race         86406 non-null  object
 4   violation_raw       86406 non-null  object
 5   violation           86406 non-null  object
 6   search_conducted    86406 non-null  bool
 7   search_type         3196 non-null   object
 8   stop_outcome        86406 non-null  object
 9   is_arrested         86406 non-null  bool
 10  stop_duration       86406 non-null  object
 11  drugs_related_stop  86406 non-null  bool
dtypes: bool(3), float64(2), object(7)
memory usage: 6.8+ MB
None
```

- ✓ No missing value or duplicates
- ✓ Correct data type
- ✓ Date-time index

Questions To Answer

Q1. Do men or women speed more often?

Compute the numbers of male and female:



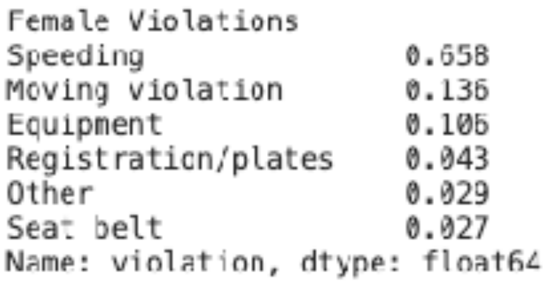
Male: 62,894
Female: 23,510

Responding to this question, there is a non-equalivent distribution of male and female, so we should use fraction in order to take this into account.

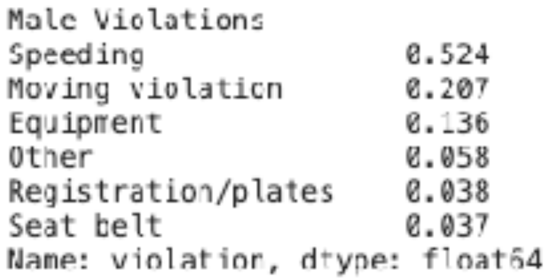
Questions To Answer

Q1. Do men or women speed more often?

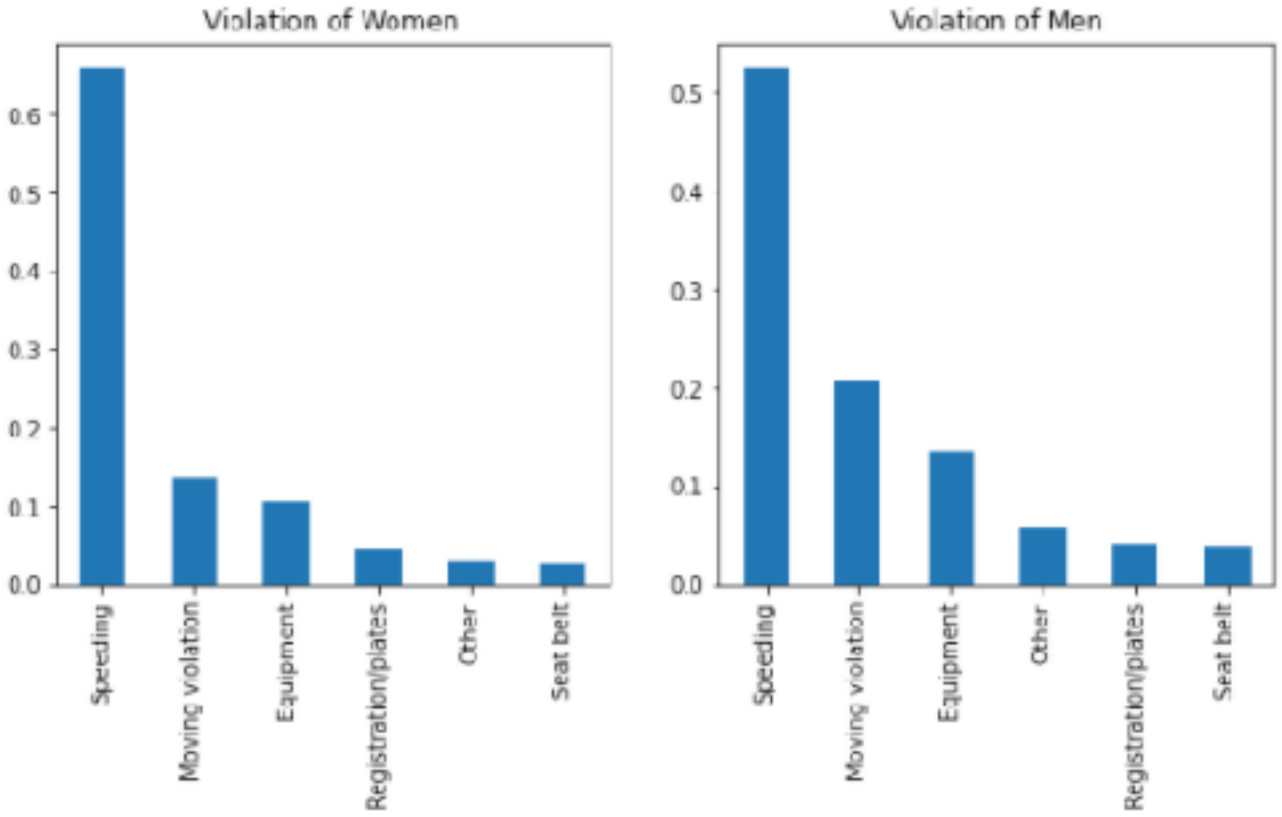
Compute the violations by different genders as proportions:



```
Female Violations
Speeding                0.658
Moving violation        0.136
Equipment               0.106
Registration/plates     0.043
Other                   0.029
Seat belt               0.027
Name: violation, dtype: float64
```



```
Male Violations
Speeding                0.524
Moving violation        0.207
Equipment               0.136
Other                   0.058
Registration/plates     0.038
Seat belt               0.037
Name: violation, dtype: float64
```

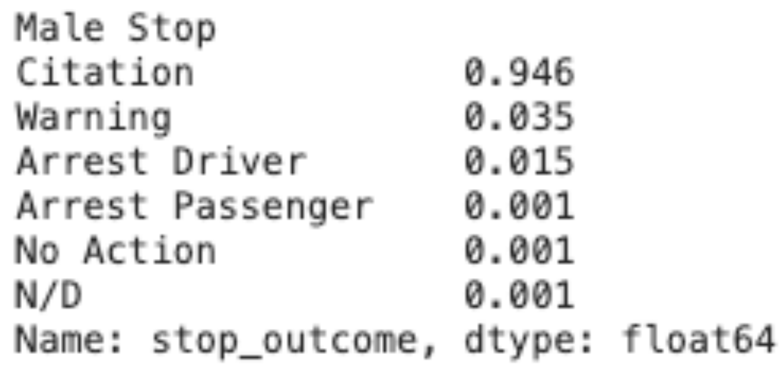


About 2/3 of female traffic stops are for speeding, whereas for males is about half, but we can't conclude that females speed more often than males since we didn't take into account the number of stops or drivers.

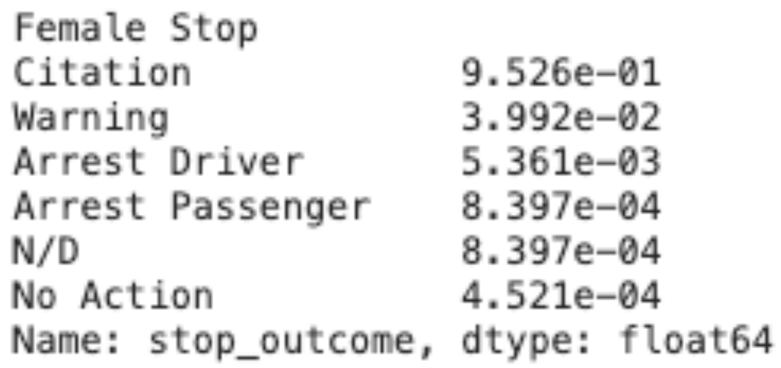
Questions To Answer

Q2. Does gender affect who gets a ticket for speeding?

Compute the stop outcome by different genders as proportions:



```
Male Stop
Citation             0.946
Warning              0.035
Arrest Driver        0.015
Arrest Passenger     0.001
No Action            0.001
N/D                  0.001
Name: stop_outcome, dtype: float64
```



```
Female Stop
Citation             9.526e-01
Warning              3.992e-02
Arrest Driver        5.361e-03
Arrest Passenger     8.397e-04
N/D                  8.397e-04
No Action            4.521e-04
Name: stop_outcome, dtype: float64
```

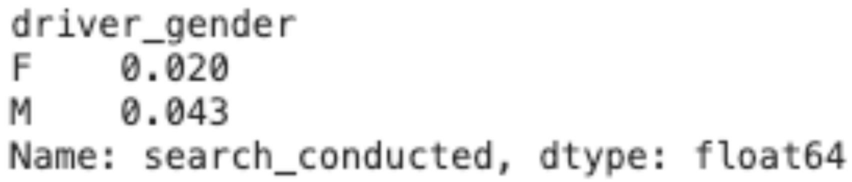
Male: 0.946
Female: 0.953

The numbers are similar for males and females: about 95% stops for speeding result in a ticket. The data doesn't show that gender has an impact on who gets a ticket for speeding.

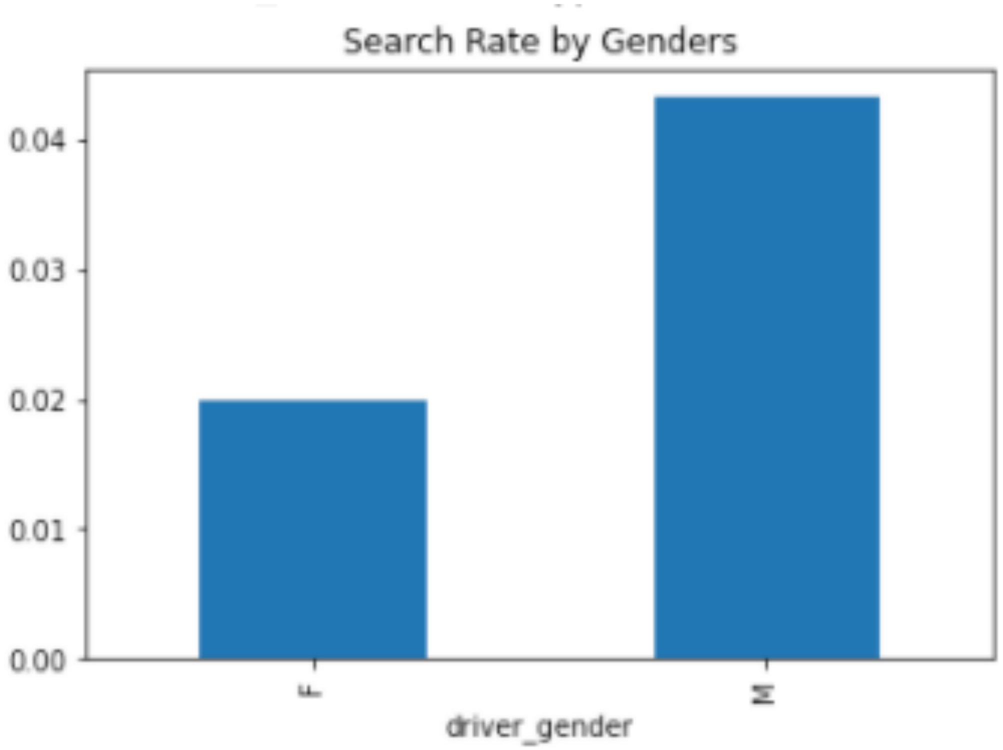
Questions To Answer

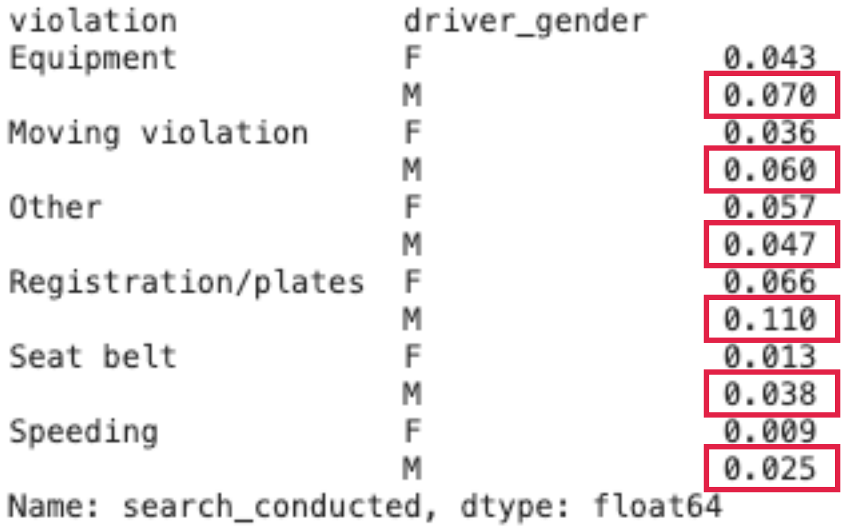
Q3. Does gender affect whose vehicle is searched?

Compute the search rate by different genders and violation types:



```
driver_gender
F    0.020
M    0.043
Name: search_conducted, dtype: float64
```





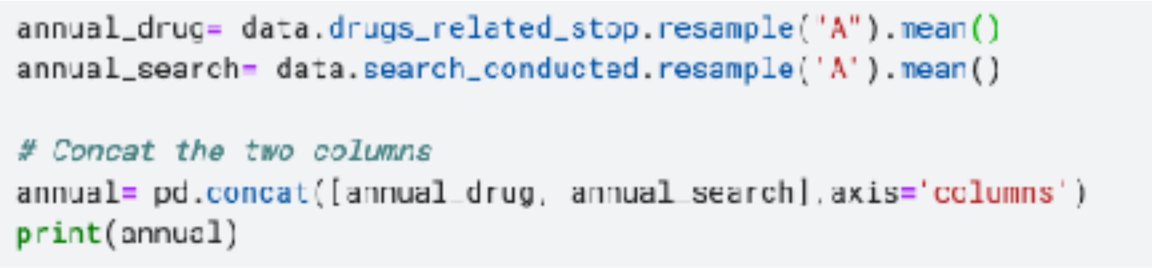
```
violation              driver_gender
Equipment              F                0.043
                       M                0.070
Moving violation       F                0.036
                       M                0.060
Other                  F                0.057
                       M                0.047
Registration/plates    F                0.066
                       M                0.110
Seat belt              F                0.013
                       M                0.038
Speeding               F                0.009
                       M                0.025
Name: search_conducted, dtype: float64
```

It is shown that male drivers are searched more than twice as often as female drivers. (4% and 2% respectively.) Males are searched more than women in different type of violations.

Questions To Answer

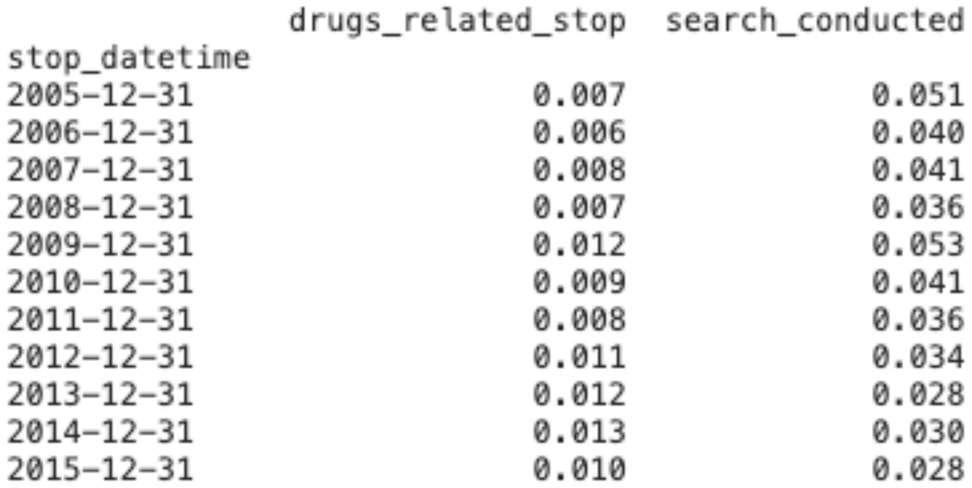
Q4. Are drug-related stops on the rise?

Compute the annual rate of drug-related stops and search:

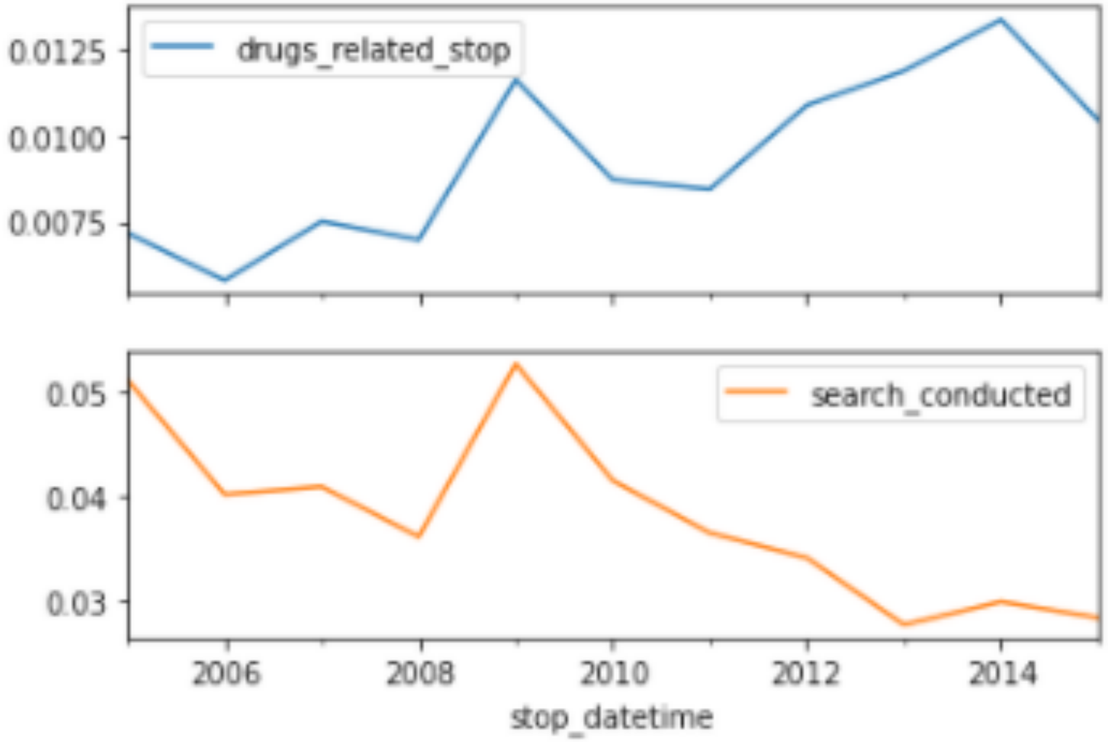


```
annual_drug= data.drugs_related_stop.resample('A").mean()
annual_search= data.search_conducted.resample('A').mean()

# Concat the two columns
annual= pd.concat([annual_drug, annual_search],axis='columns')
print(annual)
```



```
                drugs_related_stop  search_conducted
stop_datetime
2005-12-31                   0.007             0.051
2006-12-31                   0.006             0.040
2007-12-31                   0.008             0.041
2008-12-31                   0.007             0.036
2009-12-31                   0.012             0.053
2010-12-31                   0.009             0.041
2011-12-31                   0.008             0.036
2012-12-31                   0.011             0.034
2013-12-31                   0.012             0.028
2014-12-31                   0.013             0.030
2015-12-31                   0.010             0.028
```

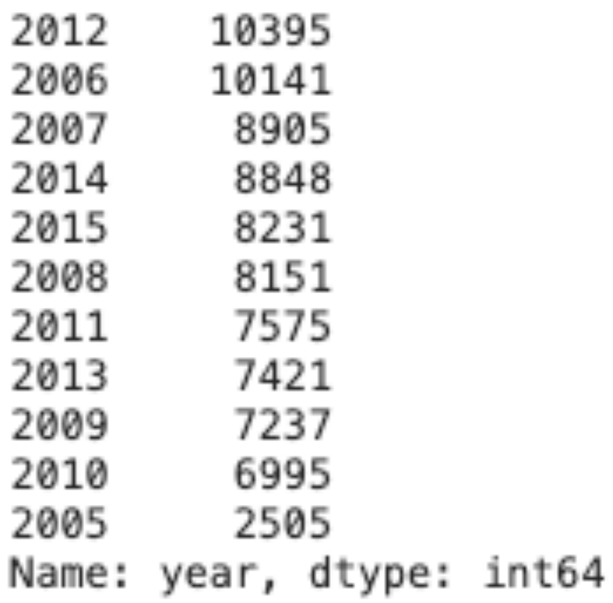


The rate of drug-related stops increased even though the search rate decreased during the 10-year periods.

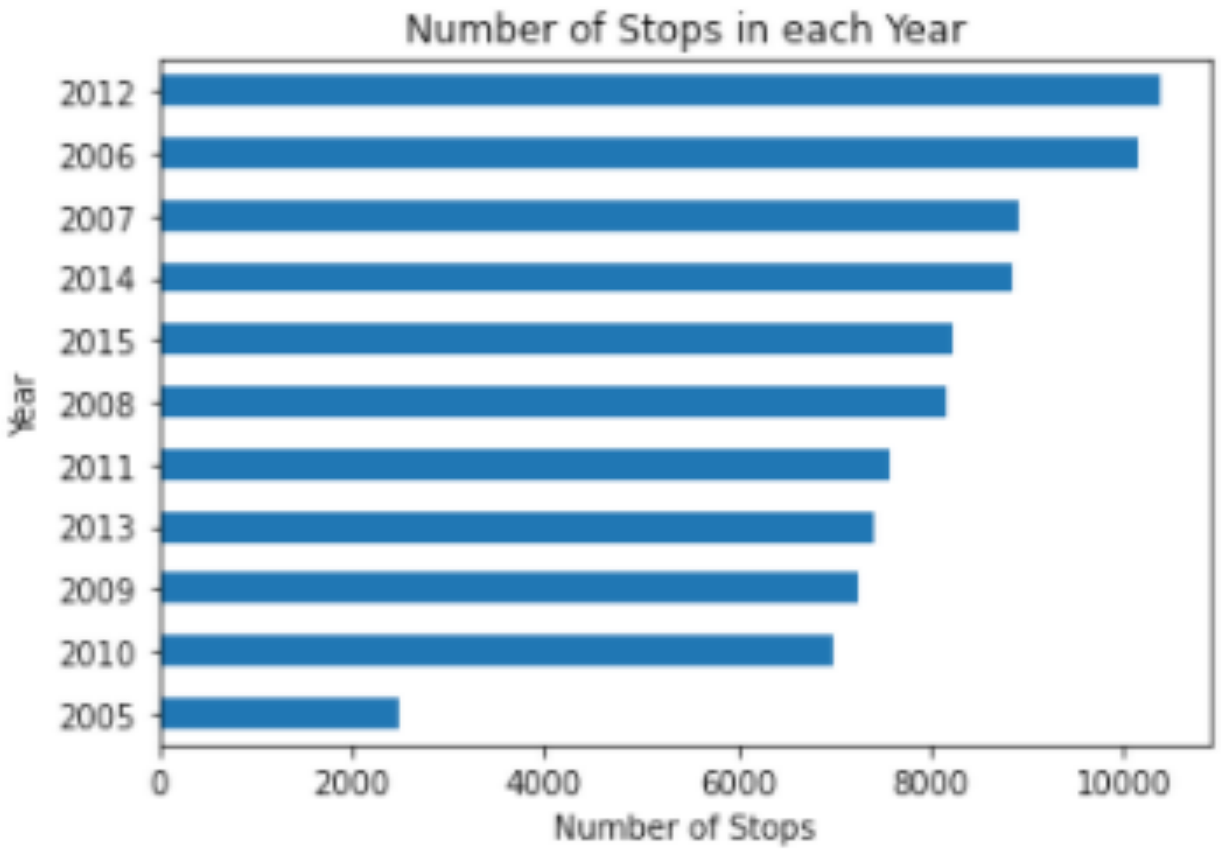
Questions To Answer

Q5. Which year had the least number of stops?

Compute the number of stops in each year:



```
2012    10395
2006    10141
2007     8905
2014     8848
2015     8231
2008     8151
2011     7575
2013     7421
2009     7237
2010     6995
2005     2505
Name: year, dtype: int64
```



As the plot shows, 2005 had the least number of stops.