

```
import pandas as pd
data = pd.read_csv('file (1).csv')
```

data

	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_dura
0	1/2/2005	1:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15
1	1/18/2005	8:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15
2	1/23/2005	23:15	NaN	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15
3	2/20/2005	17:15	NaN	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	16-30
4	3/14/2005	10:00	NaN	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15
...
65530	12/6/2012	17:54	NaN	F	1987.0	25.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15
65531	12/6/2012	22:22	NaN	M	1954.0	58.0	White	Speeding	Speeding	False	NaN	Warning	False	0-15
65532	12/6/2012	23:20	NaN	M	1985.0	27.0	Black	Equipment/Inspection Violation	Equipment	False	NaN	Citation	False	0-15
65533	12/7/2012	0:23	NaN	NaN	NaN	NaN	NaN	NaN	NaN	False	NaN	NaN	NaN	NaN
65534	12/7/2012	0:30	NaN	F	1985.0	27.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15

65535 rows × 15 columns

data.head(2)

	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
0	1/2/2005	1:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
1	1/18/2005	8:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	


data.shape

(65535, 15)

data.columns

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')

data.isnull()




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

65535 rows x 15 columns




```
data.isnull().sum()
```



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


```
data.notnull()
```



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

65535 rows × 15 columns

```
data.notnull().sum()
```



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	

```
data.describe()
```



	country_name	driver_age_raw	driver_age
count	0.0	61481.000000	61228.000000
mean	NaN	1967.791106	34.148984
std	NaN	121.050106	12.760710
min	NaN	0.000000	15.000000
25%	NaN	1965.000000	23.000000
50%	NaN	1978.000000	31.000000
75%	NaN	1985.000000	43.000000
max	NaN	8801.000000	88.000000


```
data.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65535 entries, 0 to 65534
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype
---  -
0   stop_date              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
```

✓ Remove the column that only contains missing values:

```
data.isnull()
```




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

65535 rows x 15 columns



```
data.isnull().sum()
```



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

```
data.drop( columns = 'country_name')
```

	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_re:
0	1/2/2005	1:55	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
1	1/18/2005	8:15	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
2	1/23/2005	23:15	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
3	2/20/2005	17:15	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	16-30 Min	
4	3/14/2005	10:00	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
...
65530	12/6/2012	17:54	F	1987.0	25.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
65531	12/6/2012	22:22	M	1954.0	58.0	White	Speeding	Speeding	False	NaN	Warning	False	0-15 Min	
65532	12/6/2012	23:20	M	1985.0	27.0	Black	Equipment/Inspection Violation	Equipment	False	NaN	Citation	False	0-15 Min	
65533	12/7/2012	0:23	NaN	NaN	NaN	NaN	NaN	NaN	False	NaN	NaN	NaN	NaN	
65534	12/7/2012	0:30	F	1985.0	27.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	

data.head(2)

	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	dru
0	1/2/2005	1:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
1	1/18/2005	8:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	

▼ For speeding were women or men stopped more often?

data.head(5)

	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	dru
0	1/2/2005	1:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
1	1/18/2005	8:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
2	1/23/2005	23:15	NaN	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
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	
4	3/14/2005	10:00	NaN	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	

data[data['violation'] == 'Speeding']



	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
0	1/2/2005	1:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min
1	1/18/2005	8:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min
2	1/23/2005	23:15	NaN	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min
4	3/14/2005	10:00	NaN	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min
6	4/1/2005	17:30	NaN	M	1969.0	36.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min
...
65527	12/6/2012	15:26	NaN	F	1981.0	31.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min
65529	12/6/2012	16:00	NaN	M	1994.0	18.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min
65530	12/6/2012	17:54	NaN	F	1987.0	25.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min
65531	12/6/2012	22:22	NaN	M	1954.0	58.0	White	Speeding	Speeding	False	NaN	Warning	False	0-15 Min
65534	12/7/2012	0:30	NaN	F	1985.0	27.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min

37204 rows × 15 columns

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



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

✓ Does gender affect who gets searched during a stop?

```
data.head(5)
```



	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	dr
0	1/2/2005	1:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
1	1/18/2005	8:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
2	1/23/2005	23:15	NaN	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
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	
4	3/14/2005	10:00	NaN	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	

```
data.groupby('driver_gender').search_conducted.sum()
```



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

```
data['search_conducted'].sum()
```

↗ 2479

✓ what is the mean stop duration?

data.head(5)

	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	dru
0	1/2/2005	1:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
1	1/18/2005	8:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
2	1/23/2005	23:15	NaN	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	
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	
4	3/14/2005	10:00	NaN	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	

data['stop_duration'].value_counts()

```

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

```

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

data.head(5)


	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	dru
0	1/2/2005	1:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	
1	1/18/2005	8:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	
2	1/23/2005	23:15	NaN	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	
3	2/20/2005	17:15	NaN	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	24.0	
4	3/14/2005	10:00	NaN	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	

data['stop_duration'].mean()

↗ 12.187420698181345


✓ Compare the age distributions for each violations

data.head(2)



	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	dru
0	1/2/2005	1:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	
1	1/18/2005	8:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	7.5	

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



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