

Files

+

..

sample_data

RTA Dataset.csv.zip

+ Code + Text

```
[1] import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
```


```
[2] import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
[3] df = pd.read_csv("/content/RTA Dataset.csv.zip")
```

```
df.head()
```

	Time	Day_of_week	Age_band_of_driver	Sex_of_driver	Educational_level	Vehicle_
0	17:02:00	Monday	18-30	Male	Above high school	
1	17:02:00	Monday	31-50	Male	Junior high school	
2	17:02:00	Monday	18-30	Male	Junior high school	
3	1:06:00	Sunday	18-30	Male	Junior high school	
4	1:06:00	Sunday	18-30	Male	Junior high school	

5 rows x 32 columns

✓  `df.sample(5)`

	Time	Day_of_week	Age_band_of_driver	Sex_of_driver	Educational_level	Vehic
8716	20:05:00	Friday	18-30	Male	Writing & reading	
6845	0:17:00	Sunday	Under 18	Male	Junior high school	
650	10:40:00	Thursday	Over 51	Male	High school	
6726	15:42:00	Thursday	18-30	Male	Junior high school	
5300	15:59:00	Monday	18-30	Male	Junior high school	

5 rows x 32 columns

✓ [6] df.shape

(12316, 32)

```
[7] df.columns
```

```
Index(['Time', 'Day_of_week', 'Age_band_of_driver', 'Sex_of_driver',
      'Educational_level', 'Vehicle_driver_relation', 'Driving_experience',
      'Type_of_vehicle', 'Owner_of_vehicle', 'Service_year_of_vehicle',
      'Defect_of_vehicle', 'Area_accident_occured', 'Lanes_or_Medians',
      'Road_allignment', 'Types_of_Junction', 'Road_surface_type',
      'Road_surface_conditions', 'Light_conditions', 'Weather_conditions',
```

Files



{x}



..
sample_data
RTA Dataset.csv.zip

+ Code + Text

```
[7]: index(['Time', 'Day_of_week', 'Age_band_of_driver', 'Sex_of_driver',  
'Educational_level', 'Vehicle_driver_relation', 'Driving_experience',  
'Type_of_vehicle', 'Owner_of_vehicle', 'Service_year_of_vehicle',  
'Defect_of_vehicle', 'Area_accident_occured', 'Lanes_or_Medians',  
'Road_allignment', 'Types_of_Junction', 'Road_surface_type',  
'Road_surface_conditions', 'Light_conditions', 'Weather_conditions',  
'Type_of_collision', 'Number_of_vehicles_involved',  
'Number_of_casualties', 'Vehicle_movement', 'Casualty_class',  
'Sex_of_casualty', 'Age_band_of_casualty', 'Casualty_severity',  
'Work_of_casualty', 'Fitness_of_casualty', 'Pedestrian_movement',  
'Cause_of_accident', 'Accident_severity'],  
      dtype='object')
```

```
df.describe(include="all")
```



	Time	Day_of_week	Age_band_of_driver	Sex_of_driver	Educational_level	Veh
count	12316	12316	12316	12316	11575	
unique	1074	7	5	3	7	
top	15:30:00	Friday	18-30	Male	Junior high school	
freq	120	2041	4271	11437	7619	
mean	NaN	NaN	NaN	NaN	NaN	
std	NaN	NaN	NaN	NaN	NaN	
min	NaN	NaN	NaN	NaN	NaN	
25%	NaN	NaN	NaN	NaN	NaN	
50%	NaN	NaN	NaN	NaN	NaN	
75%	NaN	NaN	NaN	NaN	NaN	



Files



{x}



sample_data



RTA Dataset.csv.zip



<>

Disk  81.54 GB available

+ Code + Text



0s

[8]

max

NaN

NaN

NaN

NaN

NaN

11 rows × 32 columns

4



0s



df.dtypes



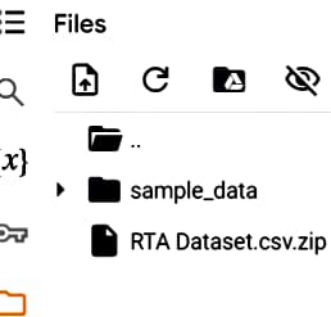
Time	object
Day_of_week	object
Age_band_of_driver	object
Sex_of_driver	object
Educational_level	object
Vehicle_driver_relation	object
Driving_experience	object
Type_of_vehicle	object
Owner_of_vehicle	object
Service_year_of_vehicle	object
Defect_of_vehicle	object
Area_accident_occured	object
Lanes_or_Medians	object
Road_alignment	object
Types_of_Junction	object
Road_surface_type	object
Road_surface_conditions	object
Light_conditions	object
Weather_conditions	object
Type_of_collision	object
Number_of_vehicles_involved	int64
Number_of_casualties	int64
Vehicle_movement	object
Casualty_class	object
Sex_of_casualty	object



```
+ Code + Text
[9] type_of_collision      object
    Number_of_vehicles_involved  int64
    Number_of_casualties      int64
    Vehicle_movement      object
    Casualty_class      object
    Sex_of_casualty      object
    Age_band_of_casualty      object
    Casualty_severity      object
    Work_of_casualty      object
    Fitness_of_casualty      object
    Pedestrian_movement      object
    Cause_of_accident      object
    Accident_severity      object
    dtype: object
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12316 entries, 0 to 12315
Data columns (total 32 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                ---
0   Time                                12316 non-null  object
1   Day_of_week                        12316 non-null  object
2   Age_band_of_driver                12316 non-null  object
3   Sex_of_driver                     12316 non-null  object
4   Educational_level                  11575 non-null  object
5   Vehicle_driver_relation            11737 non-null  object
6   Driving_experience                 11487 non-null  object
7   Type_of_vehicle                    11366 non-null  object
8   Owner_of_vehicle                   11834 non-null  object
9   Service_year_of_vehicle            8388 non-null   object
10  Defect_of_vehicle                  7889 non-null   object
11  Area_accident_occured              12077 non-null  object
```



+ Code + Text

```
0
✓ 0s 1 Day_of_week 12316 non-null object
2 Age_band_of_driver 12316 non-null object
3 Sex_of_driver 12316 non-null object
4 Educational_level 11575 non-null object
5 Vehicle_driver_relation 11737 non-null object
6 Driving_experience 11487 non-null object
7 Type_of_vehicle 11366 non-null object
8 Owner_of_vehicle 11834 non-null object
9 Service_year_of_vehicle 8388 non-null object
10 Defect_of_vehicle 7889 non-null object
11 Area_accident_occured 12077 non-null object
12 Lanes_or_Medians 11931 non-null object
13 Road_alignment 12174 non-null object
14 Types_of_Junction 11429 non-null object
15 Road_surface_type 12144 non-null object
16 Road_surface_conditions 12316 non-null object
17 Light_conditions 12316 non-null object
18 Weather_conditions 12316 non-null object
19 Type_of_collision 12161 non-null object
20 Number_of_vehicles_involved 12316 non-null int64
21 Number_of_casualties 12316 non-null int64
22 Vehicle_movement 12008 non-null object
23 Casualty_class 12316 non-null object
24 Sex_of_casualty 12316 non-null object
25 Age_band_of_casualty 12316 non-null object
26 Casualty_severity 12316 non-null object
27 Work_of_casualty 9118 non-null object
28 Fitness_of_casualty 9681 non-null object
29 Pedestrian_movement 12316 non-null object
30 Cause_of_accident 12316 non-null object
31 Accident_severity 12316 non-null object
dtypes: int64(2), object(30)
memory usage: 3.0+ MB
```

Files



..
sample_data
RTA Dataset.csv.zip



< >



Disk 81.54 GB available

+ Code + Text

✓ [11] # convert the 'Date' column to datetime format

```
df['Time'] = pd.to_datetime(df['Time'])
```

```
<ipython-input-11-e69021690a5a>:2: UserWarning: Could not infer format, so each element  
df['Time'] = pd.to_datetime(df['Time'])
```

✓ [12] df.duplicated()

```
0      False  
1      False  
2      False  
3      False  
4      False  
...  
12311   False  
12312   False  
12313   False  
12314   False  
12315   False  
Length: 12316, dtype: bool
```

✓ [13] df.duplicated().sum()

```
0
```

+ Code

+ Text

✓ [14] df.groupby('Accident_severity').size()

```
Accident_severity  
Fatal injury      158  
Serious Injuriy  1743
```




Files



+ Code + Text



..

sample_data



RTA Dataset.csv.zip



Time	0
Day_of_week	0
Age_band_of_driver	0
Sex_of_driver	0
Educational_level	741
Vehicle_driver_relation	579
Driving_experience	829
Type_of_vehicle	950
Owner_of_vehicle	482
Service_year_of_vehicle	3928
Defect_of_vehicle	4427
Area_accident_occured	239
Lanes_or_Medians	385
Road_allignment	142
Types_of_Junction	887
Road_surface_type	172
Road_surface_conditions	0
Light_conditions	0
Weather_conditions	0
Type_of_collision	155
Number_of_vehicles_involved	0
Number_of_casualties	0
Vehicle_movement	308
Casualty_class	0
Sex_of_casualty	0
Age_band_of_casualty	0
Casualty_severity	0
Work_of_casualty	3198
Fitness_of_casualty	2635
Pedestrian_movement	0
Cause_of_accident	0
Accident_severity	0
dtype: int64	





Files

..
sample_data

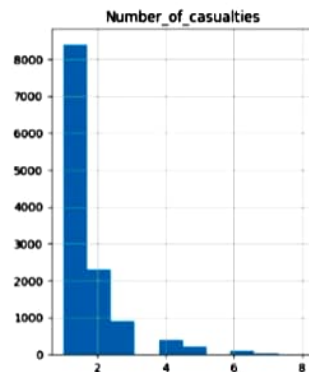
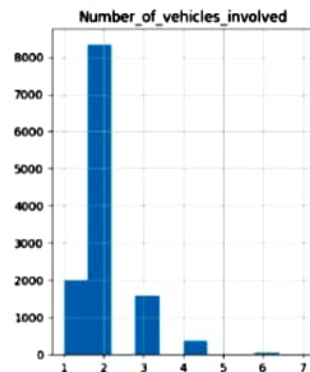
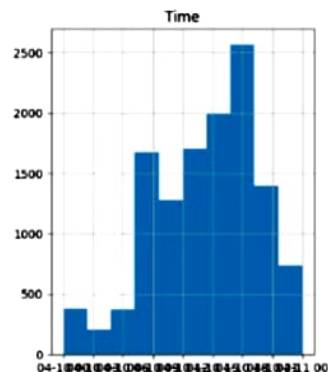
RTA Dataset.csv.zip



+ Code + Text

```
[16] <ipython-input-16-83145e50fd97> in <cell line: 2>()
      1 df.hist(layout=(1,6), figsize=(30,5))
----> 2 plt.show()
```

NameError: name 'plt' is not defined



Next steps: [Explain error](#)

```
✓ [17] df['Number_of_casualties'].value_counts()
```

```
Number_of_casualties
1      8397
2      2290
3       909
4       394
5       207
6        89
```

Disk  81.54 GB available