APPLIED DATA SCIENCE CAPSTONE: PREDICTING CAR ACCIDENT SEVERITY

BUHAI DIANA-VICTORIA

COURSERA 2020 PROJECT

INTRODUCTION

The demand for vehicles rises consistently. Consequently, so does the number of vehicles on the road and the probability of involvement in traffic jams or car accidents.

By collecting and analysing relevant data on car accidents, we aim to establish to what extent can:

1) weather conditions
2) the total number of people involved in the collision or the number pedestrians
3) road conditions or light conditions
4) driving under the influence of drugs or alcohol (DUI)

help us predict car accident severity.

Stakeholders: traffic participants, traffic police (authorities).

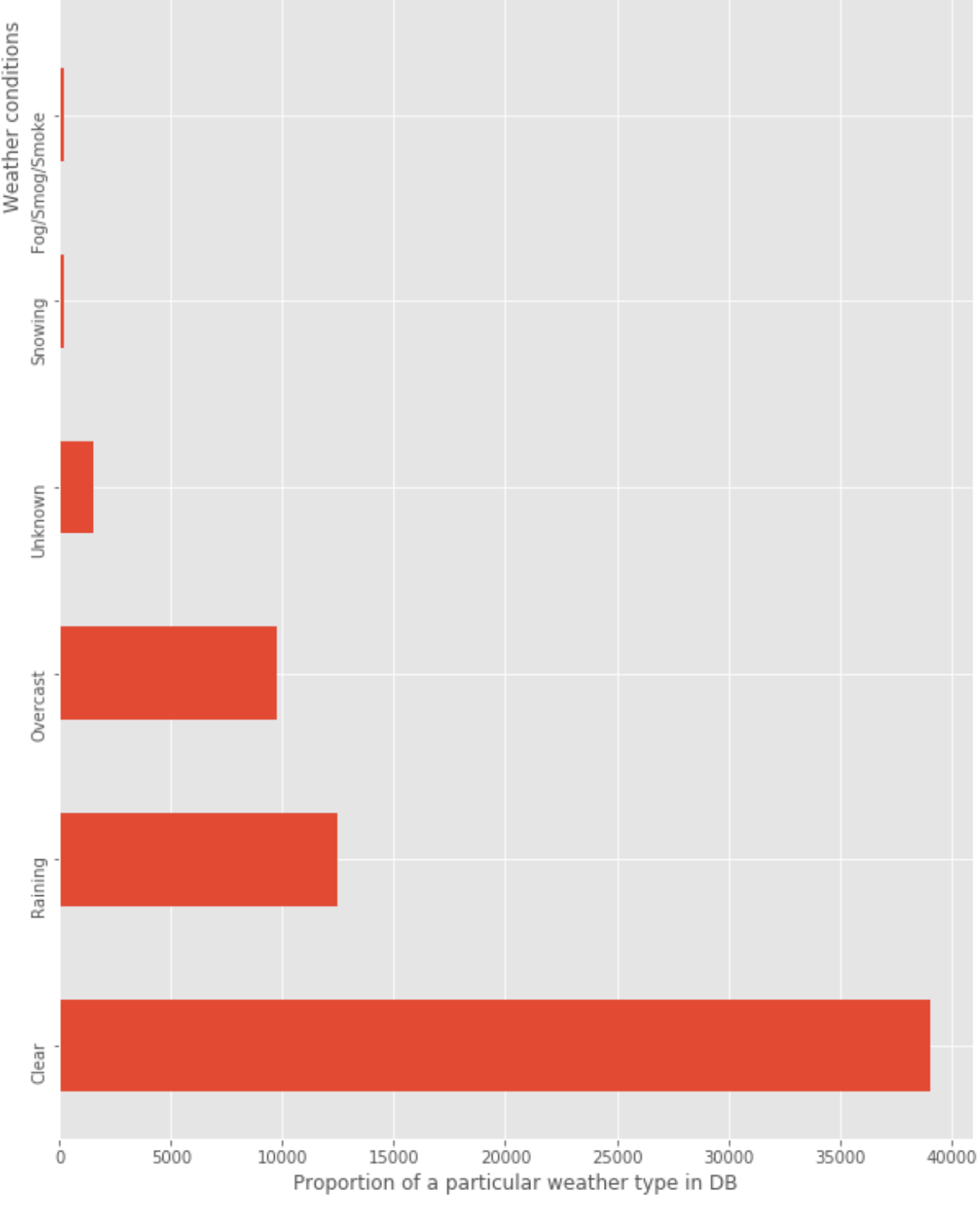
DATA SOURCES

In order to answer the proposed questions, we shall use data from the SDOT Traffic Management Division, Traffic Records Group.

Here is the provided link to database: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The metadata is to be found here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

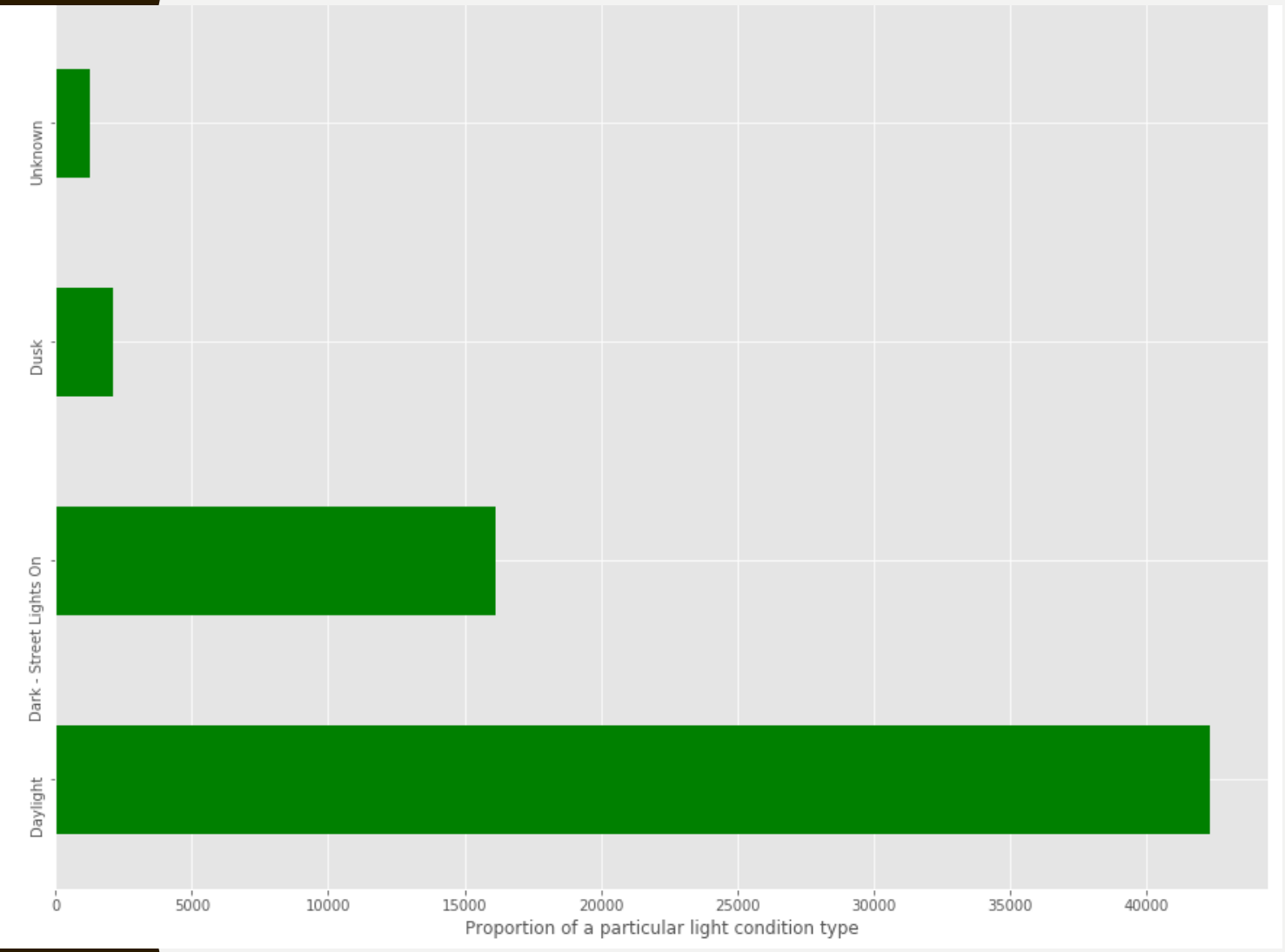
The database contains 37 attributes, including the severity code, the number of persons involved in the car accident, the number of pedestrians, road and light conditions, information on driving under the influence of drugs or alcohol. The database contains 194673 records.



WEATHER CONDITIONS AMONG CAR ACCIDENTS DATA

Clear, raining, and overcast weather types are the most common among the available data.

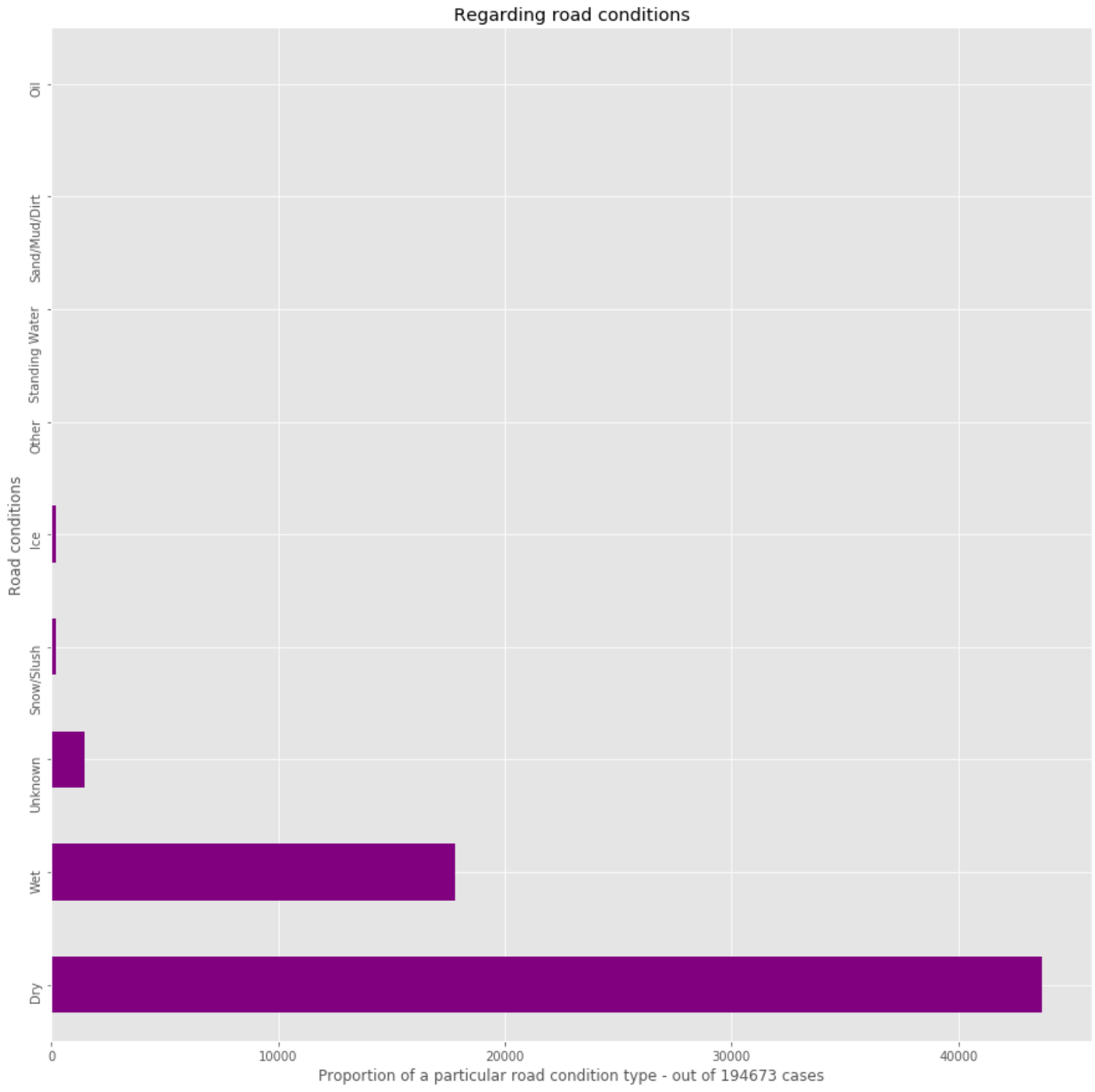
*Bar plot was cropped for the purpose of this presentation. Please refer to the full report for the figure.



LIGHT CONDITIONS AMONG CAR ACCIDENT DATA.

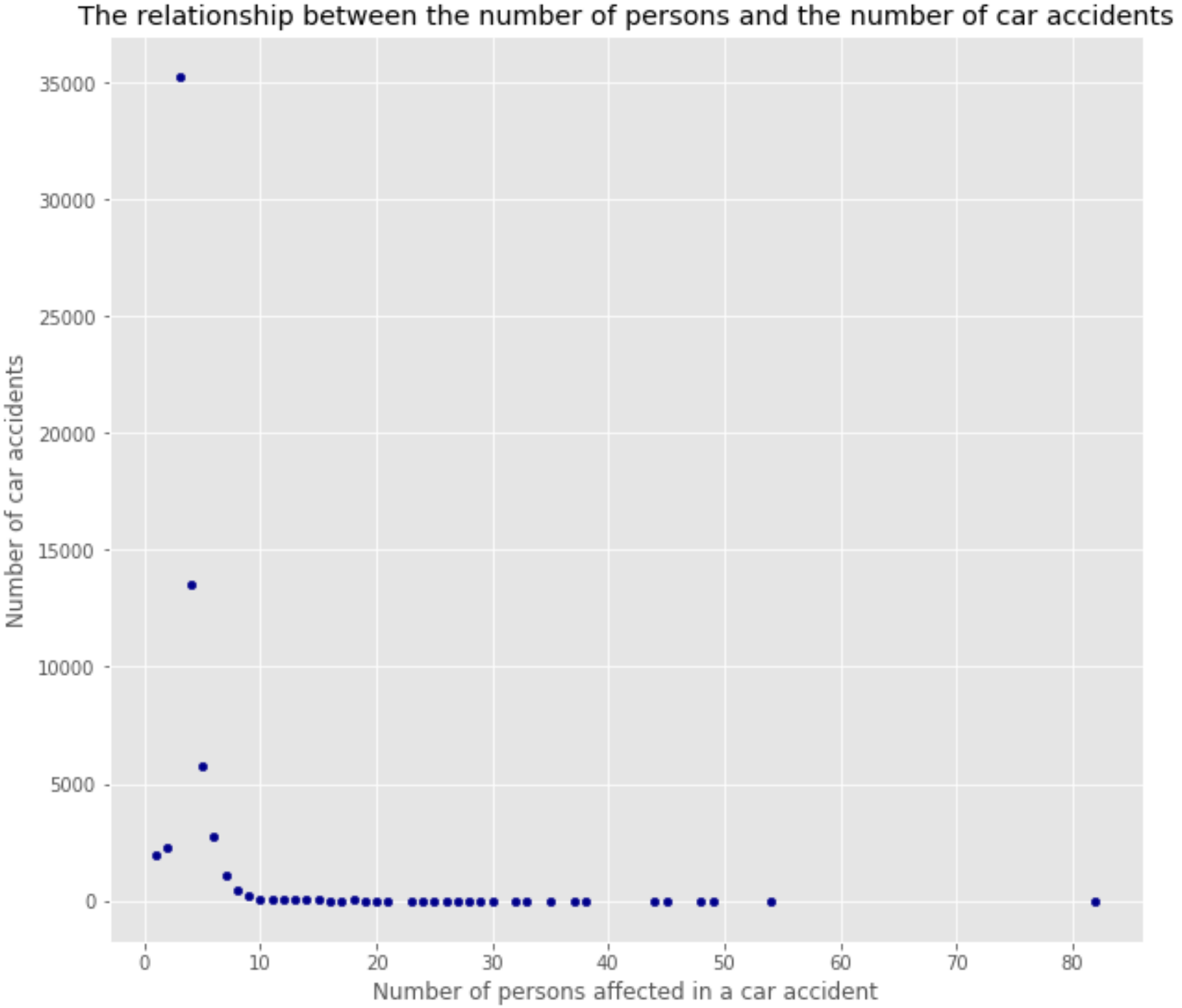
The most common light conditions are: daylight, dark with street lights on, and dusk.

*Bar plot was cropped for the purpose of this presentation. Please refer to the full report for the figure.

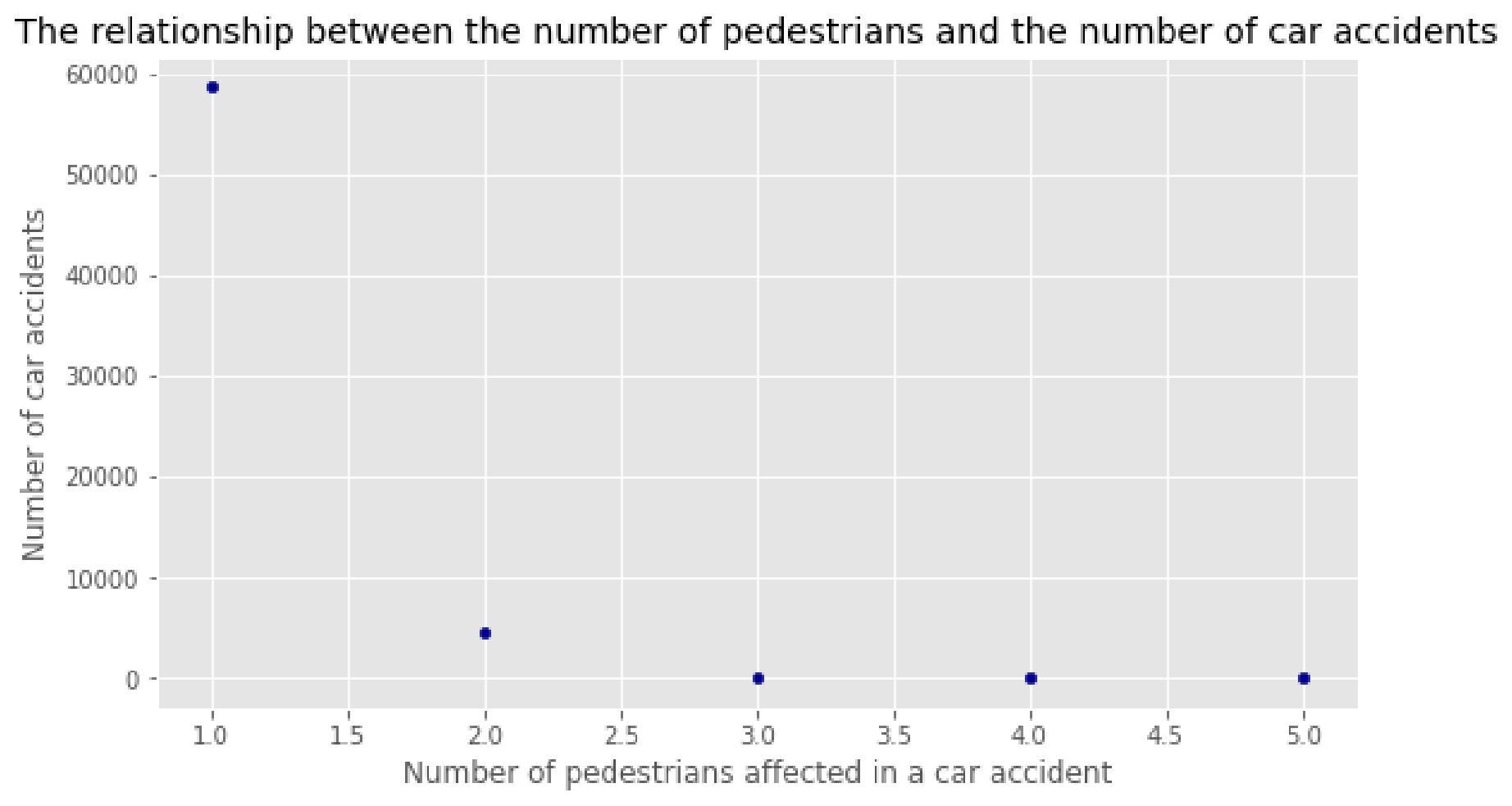


ROAD CONDITIONS AMONG CAR ACCIDENT DATA

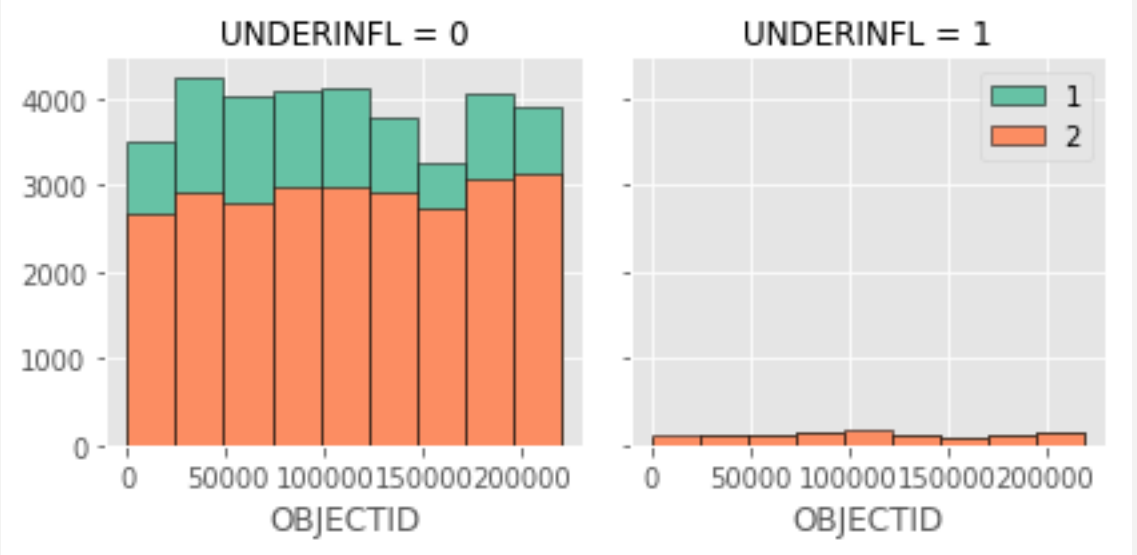
Dry, wet, and unknown road conditions are the most common.



The majority of cases involve under 10 persons.



One can tell the most the accidents involve one pedestrian.



REPRESENTATION OF DUI CASES

With green are represented the property damage cases (1) and with orange the injury cases (2). Additionally, one notices that most cases do not involve DUI. In fact, 3.15% are dui-caused car accidents (2231 cases).

Attribute (in correlation with severity code)	Spearman coefficient	p value	Interpretation
DUI	0.030	1.468	not correlated; not statistically significant
No. of persons	0.079	6.383	not correlated; not statistically significant
No. of pedestrians	0.270	0.0	slightly correlated; probably statistically significant
Raining (weather condition)	0.006	0.103	not correlated; not statistically significant
Dark - Street Lights Off (light conditions)	-0.0002	0.952	not correlated; not statistically significant
Standing water (road conditions)	-0.004	0.235	not correlated; not statistically significant
Daylight (weather condition)	0.017	1.332	not correlated; not statistically significant
Sand/Mud/Dirt (road conditions)	-0.001	0.733	not correlated; not statistically significant

INFERENTIAL STATISTICAL TESTING

The set level of statistical significance was $p<0.05$. A Spearman coefficient of 1 or -1 (or close to those values) denote a correlation.

FINAL FUTURE SET FOR MACHINE LEARNING MODELS

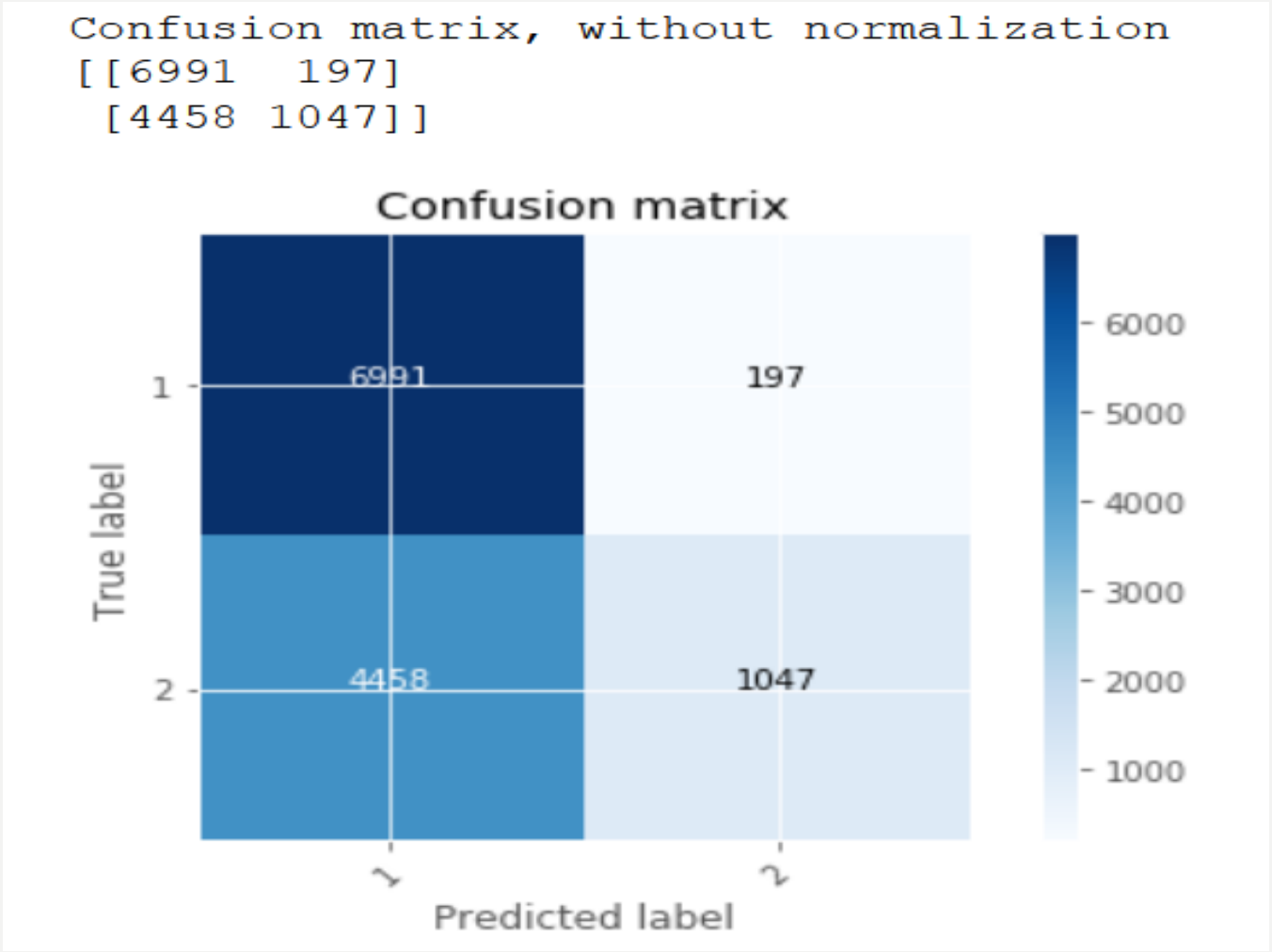
```
Index(['WEATHER', 'UNDERINFL', 'PERSONCOUNT', 'PEDCOUNT', 'LIGHTCOND',
       'ROADCOND', 'Blowing Sand/Dirt', 'Clear', 'Fog/Smog/Smoke', 'Other',
       'Overcast', 'Partly Cloudy', 'Raining', 'Severe Crosswind',
       'Sleet/Hail/Freezing Rain', 'Snowing', 'Unknown',
       'Dark - No Street Lights', 'Dark - Street Lights Off',
       'Dark - Street Lights On', 'Dark - Unknown Lighting', 'Dawn',
       'Daylight', 'Dusk', 'Other', 'Unknown', 'Dry', 'Ice', 'Oil', 'Other',
       'Sand/Mud/Dirt', 'Snow/Slush', 'Standing Water', 'Unknown', 'Wet'],
      dtype='object')
```

The following attributes were one-hot coded:

- Weather conditions
- Road conditions
- Light conditions

SUPPORT VECTOR MACHINE – CONFUSION MATRIX

```
Confusion matrix, without normalization
[[6991  197]
 [4458 1047]]
```



ACCURACY OF SVM AND LOGISTIC REGRESSION MODELS

	Jaccard's index	F1-score	Log Loss
Support Vector Machine	0.633	0.559	-
Logistic Regression	0.633	0.565	0.632

The values for F1-score and Jaccard's index are comparable with those obtained through support vector machine. Log Loss of 0.632 in a case of binary classification (severity code of 1 or 2) showcases a considerable level of uncertainty or entropy of the model.

DISCUSSION: RECOMMENDATIONS

- Choose / provide balanced set of data
- Increase the number of attributes for feature sets (overfitting!)

CONCLUSION

Our project highlighted the fact…

that the number of involved pedestrians is slightly correlated with car accidents' severity code;

that in the large majority of cases the drivers were not driving under the influence of drugs or alcohol;

that most car accidents seem to occur having clear weather and daylight.