Car accident severity

september 26, 2020

Can car accident can be predicted ?

- What factor can or must be used to predict this ?
- What is the type of data we need ?
- How much data ?
- which is the best model ?

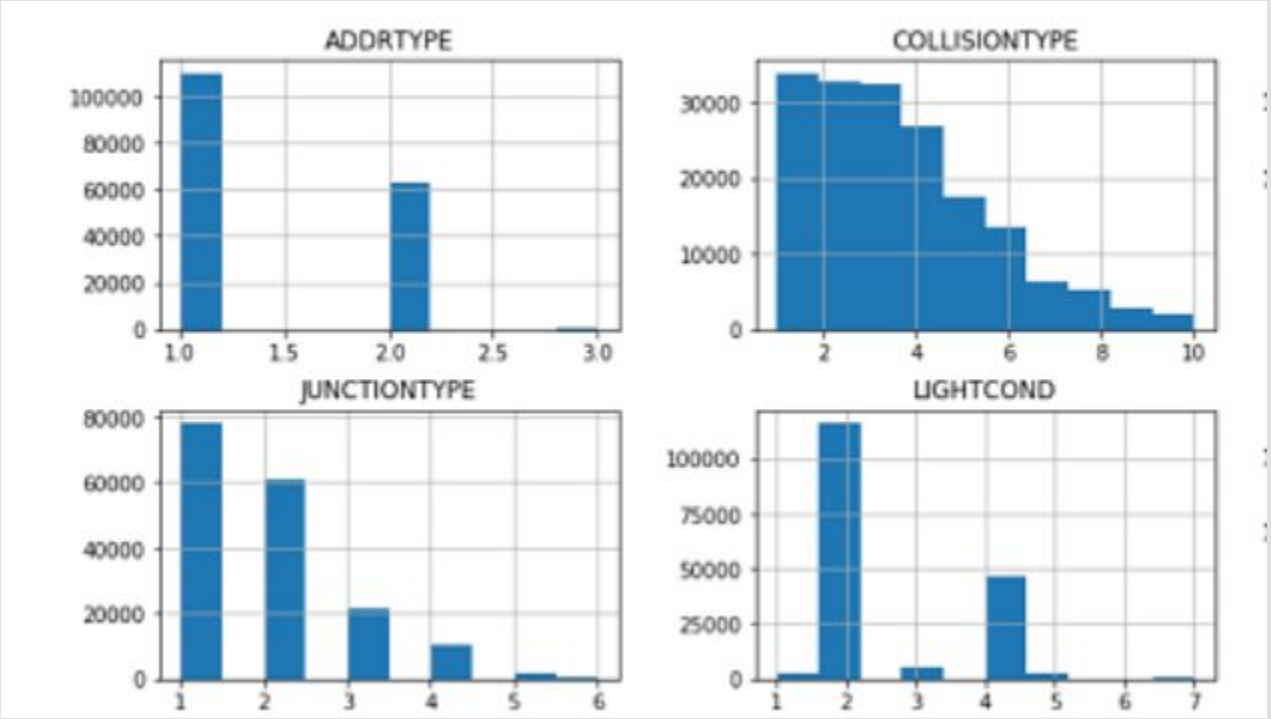
Data acquisition and cleaning

- The data used for this project can be found from here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
- It is the database of accidents in Seattle city.
- The description of the data can be found here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

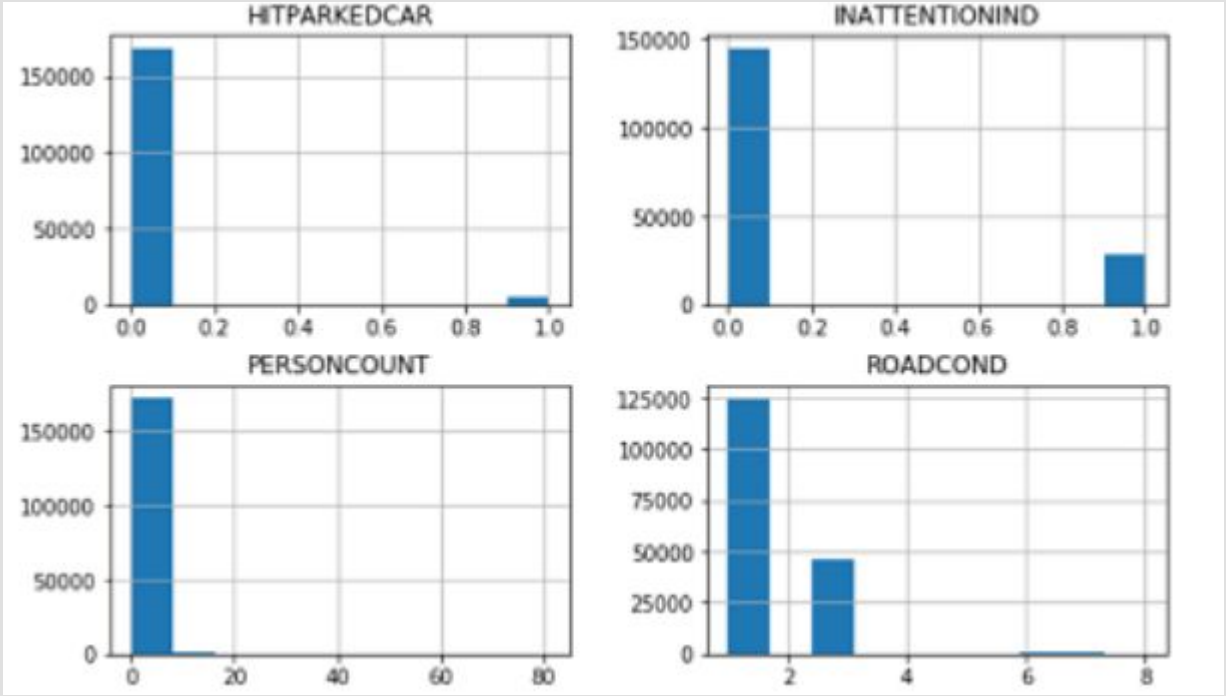
Correlation matrix

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	VEHCOUNT	JUNCTIONTYPE	INATTENTIONIND	UNDERINFL
SEVERITYCODE	1.000000	0.168967	0.153360	0.114616	-0.047699	0.082422	0.030811	0.032387
ADDRTYPE	0.168967	1.000000	0.025514	0.047121	-0.072367	0.209585	-0.096458	-0.055898
COLLISIONTYPE	0.153360	0.025514	1.000000	-0.045447	-0.296308	0.026931	-0.017740	-0.006697
PERSONCOUNT	0.114616	0.047121	-0.045447	1.000000	0.383231	0.049028	0.069815	0.015799
VEHCOUNT	-0.047699	-0.072367	-0.296308	0.383231	1.000000	0.001458	0.081871	0.009675
JUNCTIONTYPE	0.082422	0.209585	0.026931	0.049028	0.001458	1.000000	0.001482	-0.044127
INATTENTIONIND	0.030811	-0.096458	-0.017740	0.069815	0.081871	0.001482	1.000000	-0.033350
UNDERINFL	0.032387	-0.055898	-0.006697	0.015799	0.009675	-0.044127	-0.033350	1.000000
WEATHER	-0.005868	-0.012489	0.003326	-0.009141	-0.041934	-0.018454	0.003899	-0.001494
ROADCOND	-0.001235	-0.003310	-0.004160	-0.002322	0.023816	-0.004633	-0.033590	0.007717
LIGHTCOND	-0.029804	-0.032144	-0.017913	0.000944	0.018826	-0.059012	-0.054505	0.233664
SPEEDING	0.027587	-0.071925	0.006763	-0.009934	-0.023665	-0.028210	-0.056597	0.089300
HITPARKEDCAR	-0.085961	-0.116806	-0.099756	-0.041096	0.047573	-0.107552	0.017722	0.025327

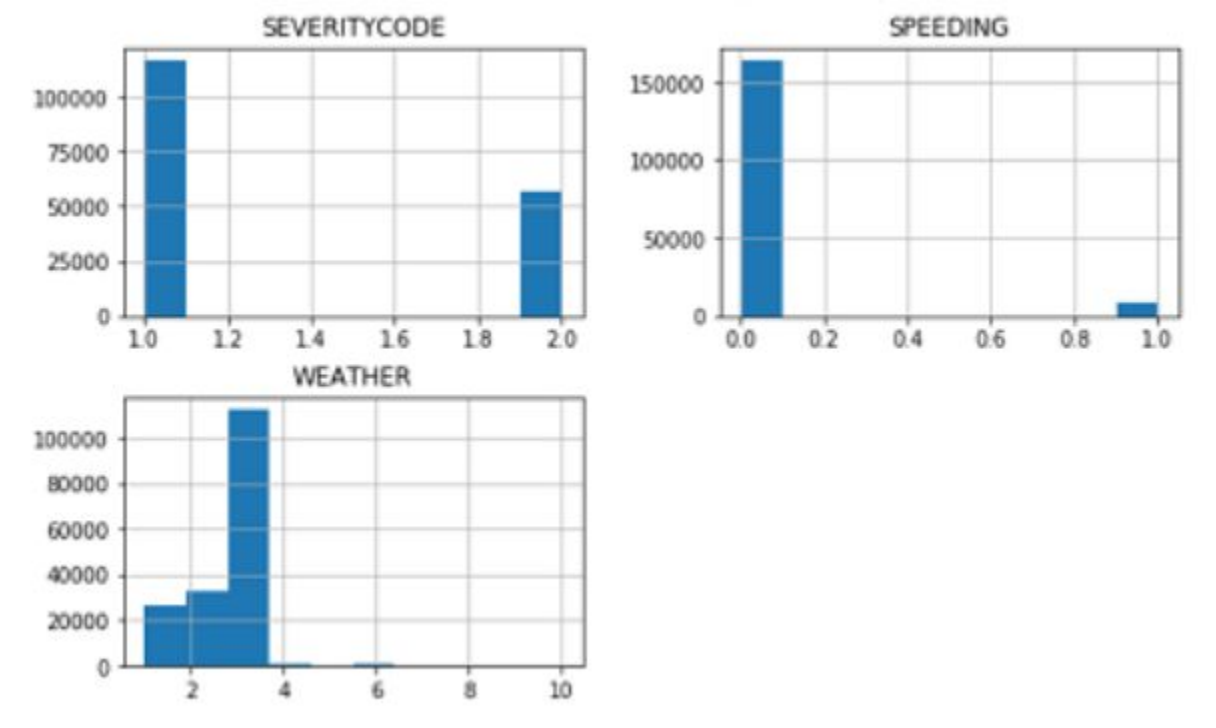
Distribution of data



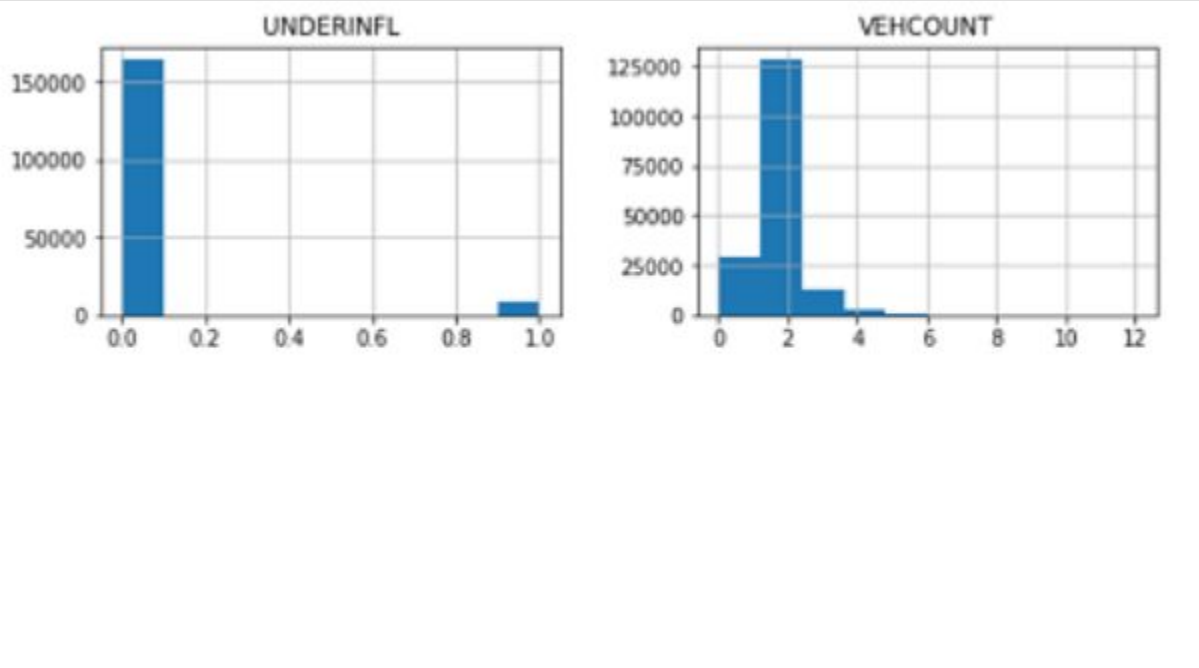
Distribution of data



Distribution of data

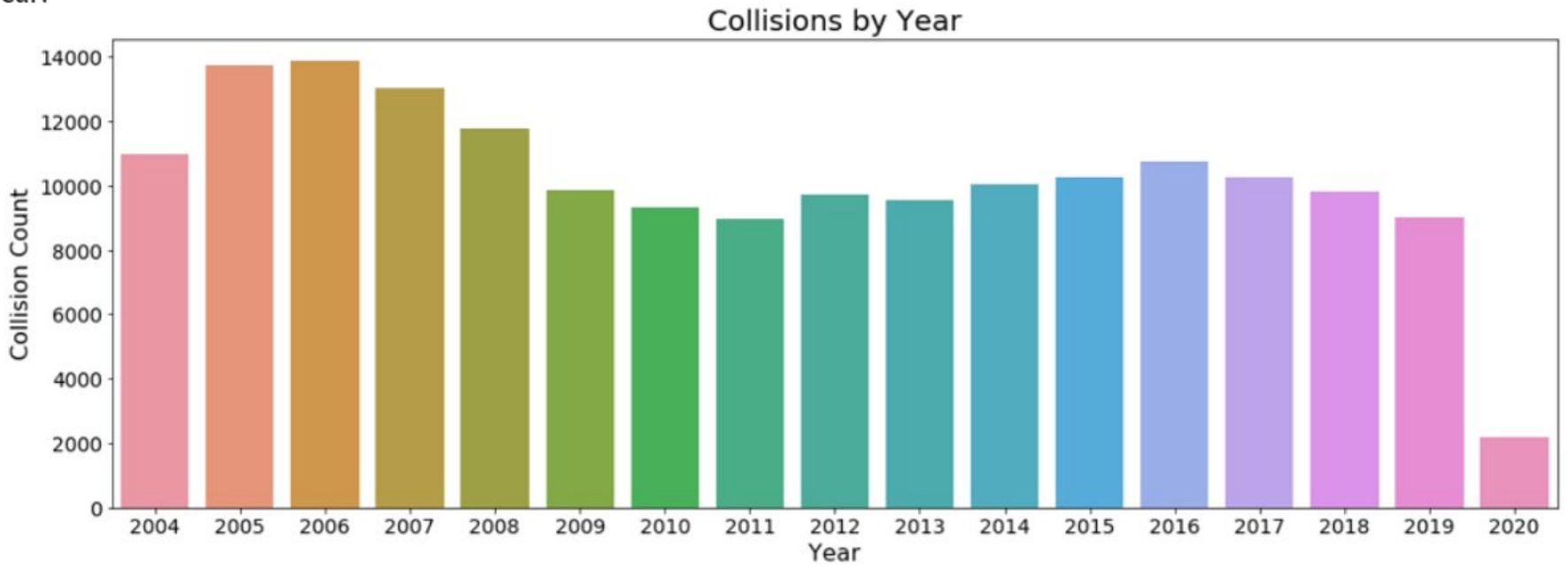


Distribution of data



Distribution of collision

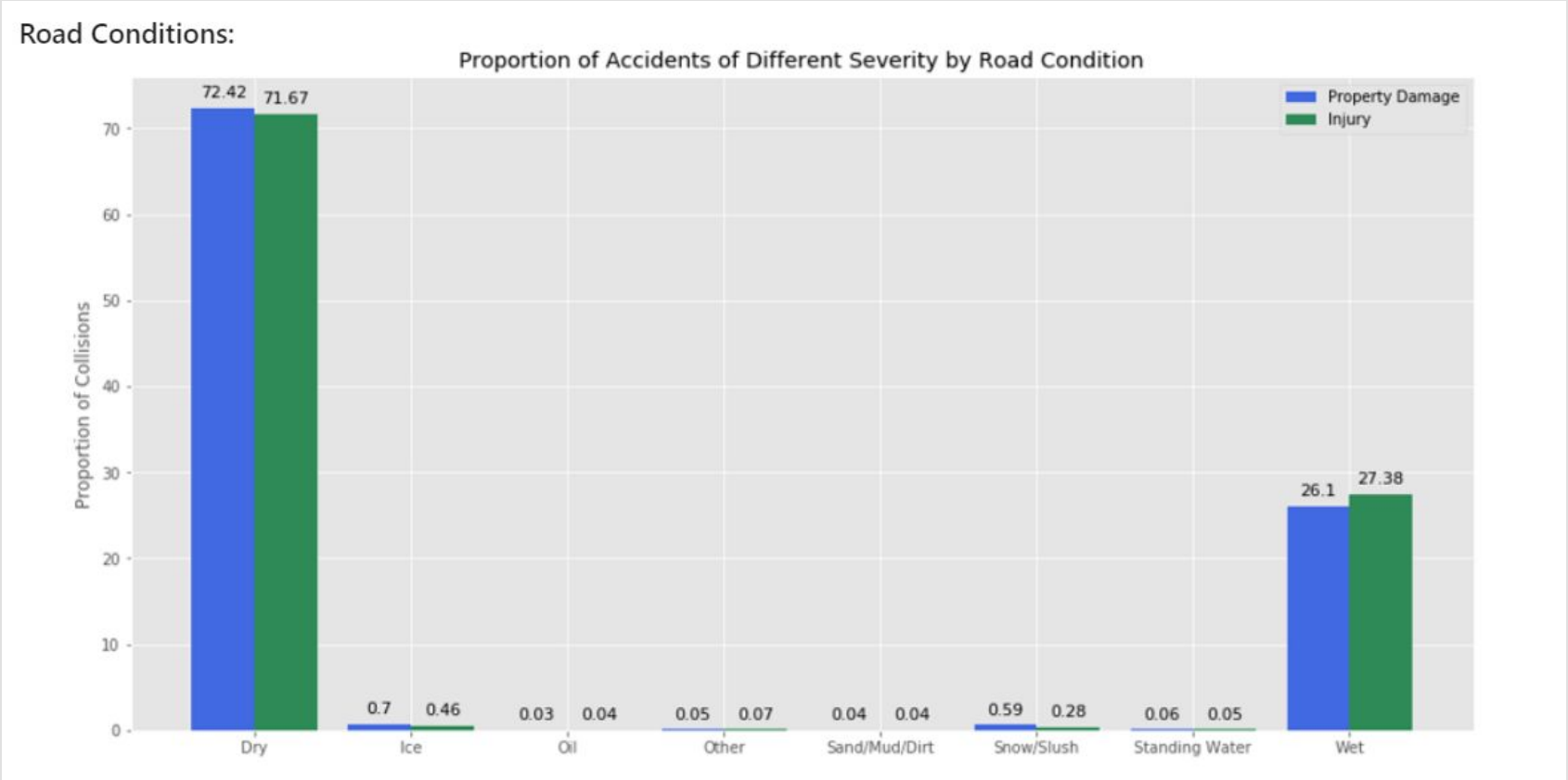
Year:



¶

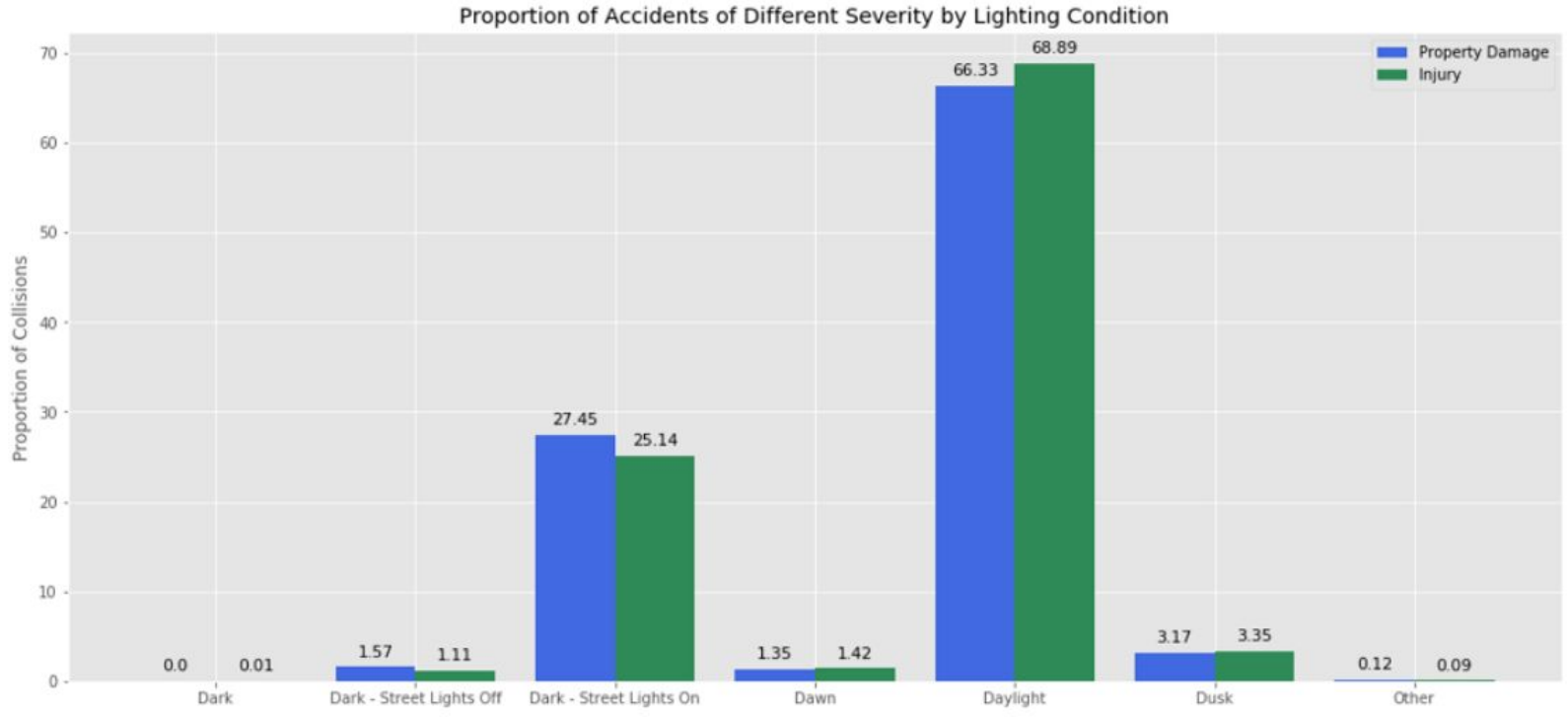
Proportion of severity of collisions (Road)

Road Conditions:



Proportion of severity of collisions (Light)

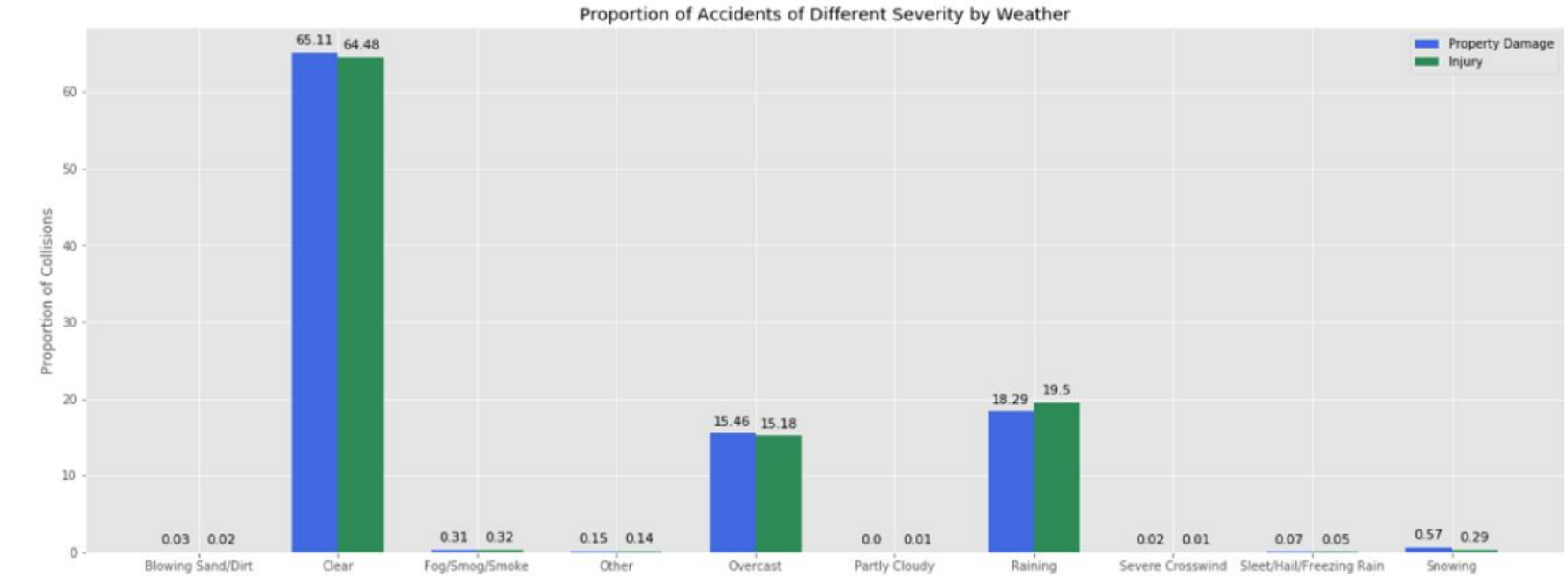
Light Conditions:



¶

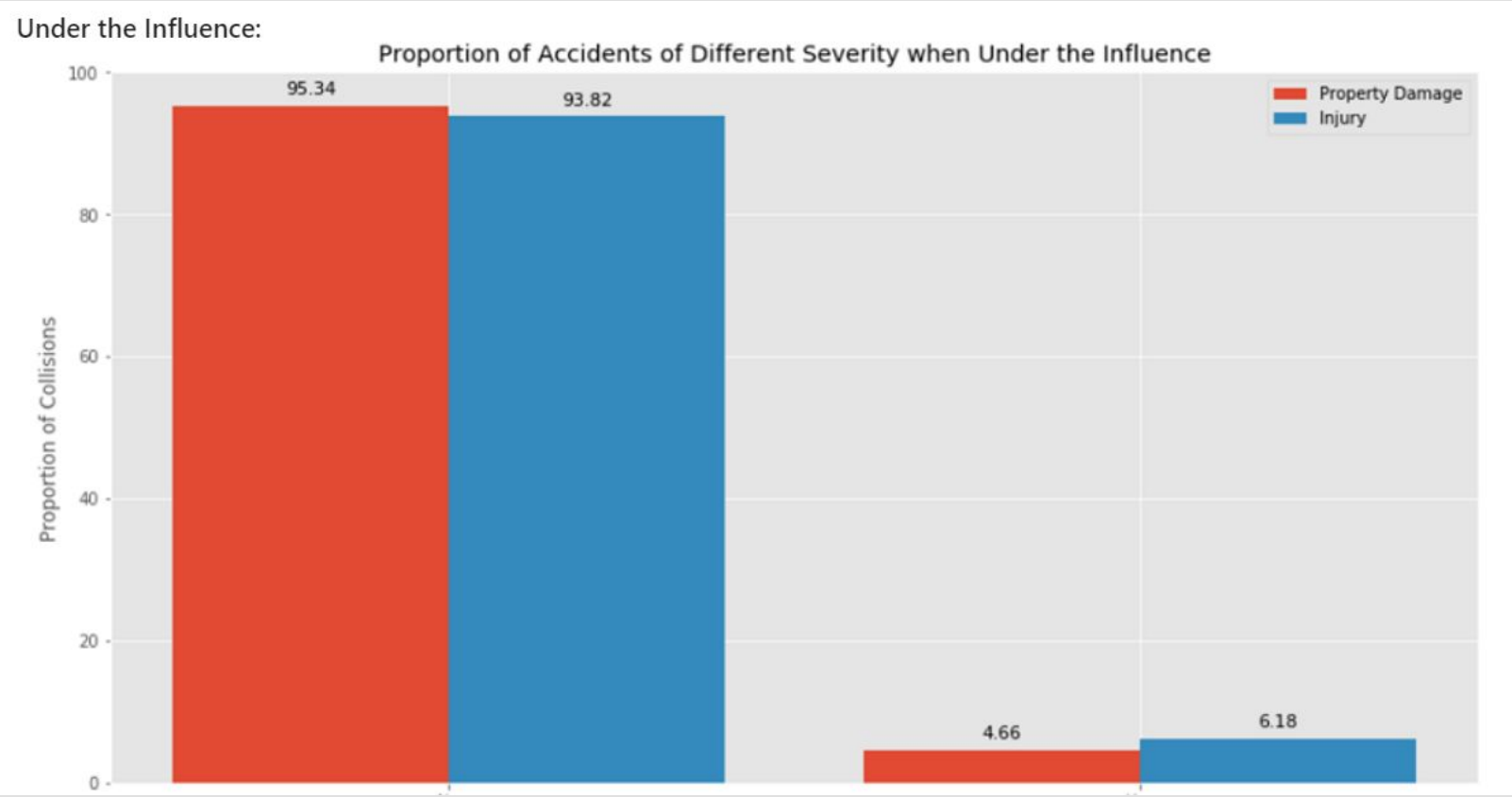
Proportion of severity of collisions (Weather)

Weather Conditions:



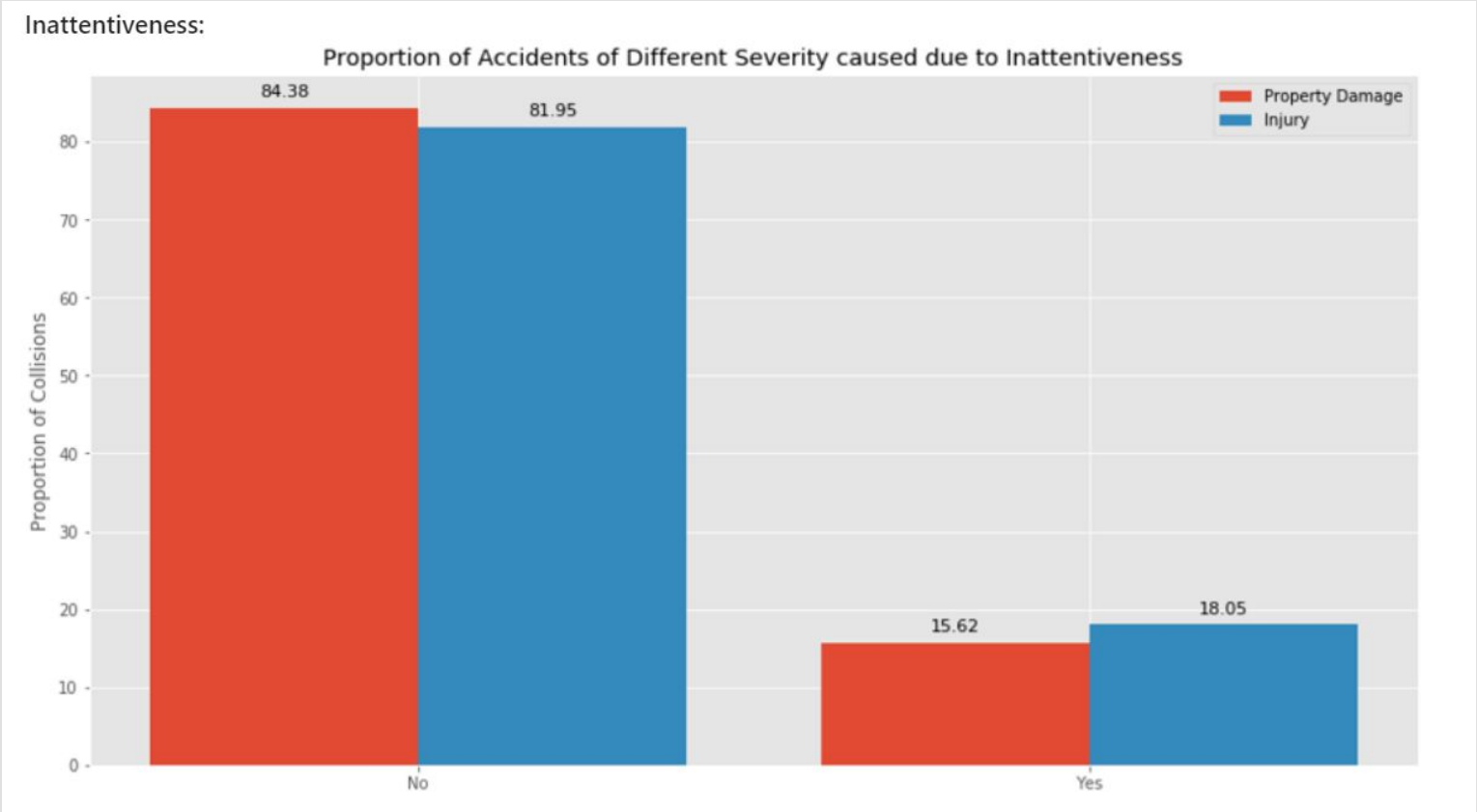
¶

Proportion of severity of collisions (Under the influence)



Proportion of severity of collisions (Inattentiveness)

Inattentiveness:



Conclusion

From now on data must be store as balanced. Cause now the data is unbalanced so it more difficult to build a best model for predicting.