CAPSTONE PROJECT IBM DATA SCIENCE

Understanding the impact of various factors on the severity of accidents

Andre Holz

CONTENT

INTRODUCTION / BUSINESS PROBLEM

According to the “Global status report on road safety 2018” from the WHO (source:https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/) highlights that about 1.35 people aged 5-9 years have been killed in an road traffic accident.

To reduce this number in future we are looking at the association of external factors as weather condition and road condition or driver related factors such as inattention and under influence of drugs and/or alcohol.

The WHO would be interested in the outcome of this analysis to recommend the most effective actions to reduce the number of facilities for the future in the US. Depend on the outcome they would be able to decide to:

a.) *Focus on the infrastructure investment program of a country*

b.) *Focus on the training and information distribution to vehicle drivers*

DATA

For the analysis of the problem raised earlier I would use the data set from the “SDOT Traffic Management Division, Traffic Records Group” which is including all types of collision from 2014 onwards.

I would focus on these attributes from that data source to be able to quicker work with the data set:

	DESCRIPTION
OBJECTID	ESRI unique identifier
SEVERITYCODE	A code that corresponds to the severity of the collision
INATTENTIONIND	Whether or not collision was due to inattention.
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.

These attributes are likely to have an impact on the severity of the accident. The data only includes collisions provided by SPD and thus in the US, but there can be very likely generalized to other OECD countries as well. To confirm this, I would use similar data set from some other OECD countries to identify, if the result conclusion would be the same. This is not part of this assessment.

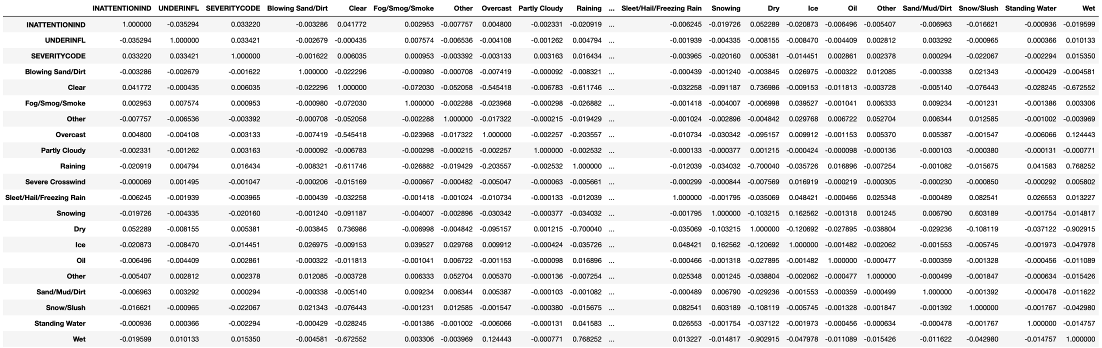
Then I applied some pre-processing and wrangling to the data. The plan was to normalize them so that I would be able to do easier analysis in the next phase. So, I got rid of categorical data and converted them via dummy data function to numerical data. Furthermore, I have drops lines where the values have been unknown for some of the attributes. In addition I also did some data cleansing as not all value have been in the same format in an attribute.

METHOLOGOGY

I have decided to use a correlation analysis between the various attributes and would like to understand the correlation of the severity of an accident and the various factors that may impact this.

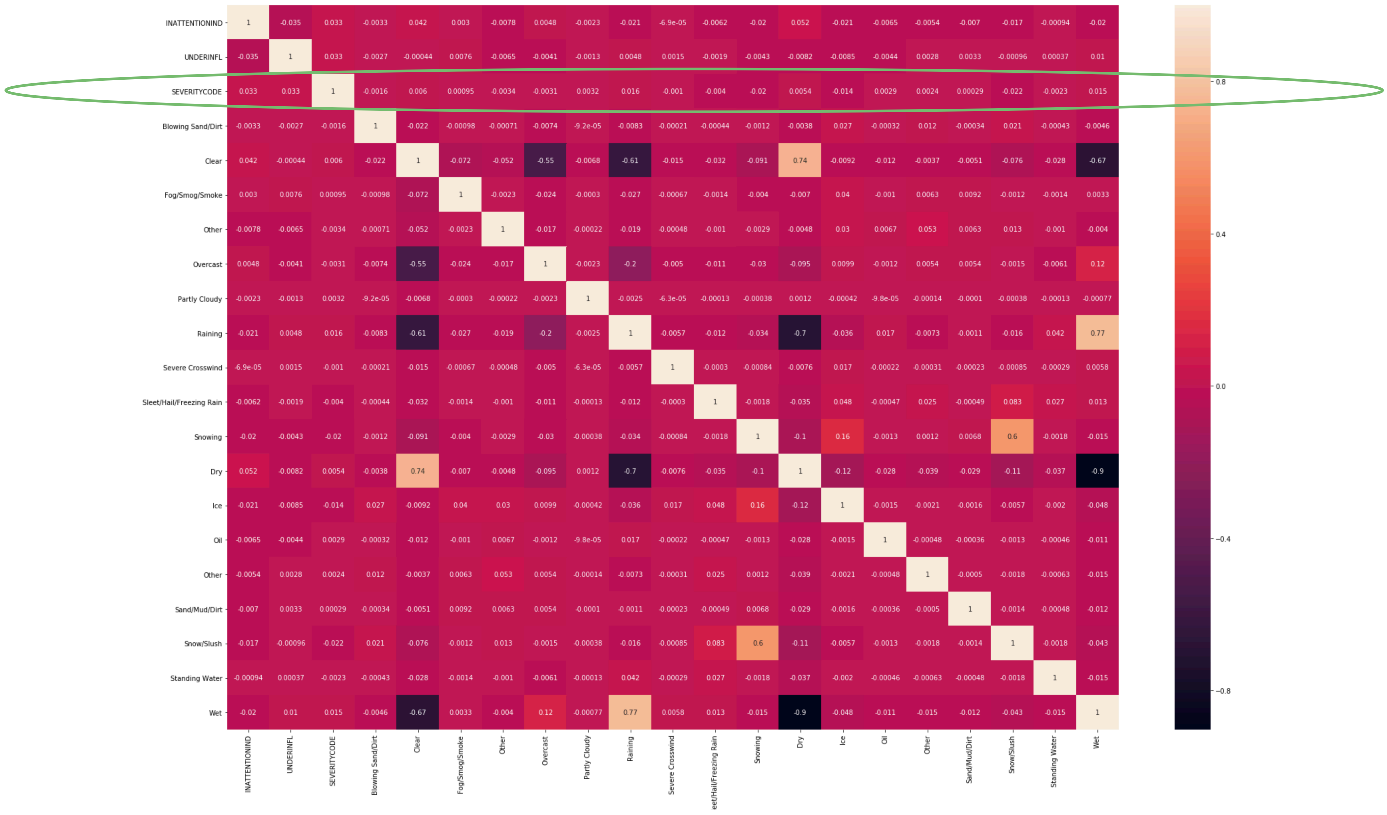
RESULTS

This is the correlation with has been calculated:



	INATTENTIONIND	UNDERINFL	SEVERITYCODE	Blowing Sand/Dirt	Clear	Fog/Smog/Smoke	Other	Overcast	Partly Cloudy	Raining	...	Sleet/Hail/Freezing Rain	Snowing	Dry	Ice	Oil	Other	Sand/Mud/Dirt	Snow/Slush	Standing Water	Wet
INATTENTIONIND	1.000000	-0.035294	0.033220	-0.003286	0.041772	0.002953	-0.007757	0.004800	-0.002331	-0.020919	...	-0.006245	-0.019726	0.052289	-0.020873	-0.006496	-0.005407	-0.006963	-0.016621	-0.000936	-0.019599
UNDERINFL	-0.035294	1.000000	0.033421	-0.002679	-0.000435	0.007574	-0.006536	-0.004108	-0.001262	0.004794	...	-0.001939	-0.004335	-0.008155	-0.008470	-0.004409	0.002812	0.003292	-0.000965	0.000366	0.010133
SEVERITYCODE	0.033220	0.033421	1.000000	-0.001622	0.006035	0.000953	-0.003392	-0.003133	0.003163	0.016434	...	-0.003965	-0.020160	0.005381	-0.014451	0.002861	0.002378	0.000294	-0.022067	-0.002294	0.015350
Blowing Sand/Dirt	-0.003286	-0.002679	-0.001622	1.000000	-0.022296	-0.000980	-0.000708	-0.007419	-0.000092	-0.008321	...	-0.000439	-0.001240	-0.003845	0.026975	-0.000322	0.012085	-0.000338	0.021343	-0.000429	-0.004581
Clear	0.041772	-0.000435	0.006035	-0.022296	1.000000	-0.072030	-0.052058	-0.545418	-0.006783	-0.611746	...	-0.032258	-0.091187	0.736986	-0.009153	-0.011813	-0.003728	-0.005140	-0.076443	-0.028245	-0.672552
Fog/Smog/Smoke	0.002953	0.007574	0.000953	-0.000980	-0.072030	1.000000	-0.002288	-0.023968	-0.000298	-0.026882	...	-0.001418	-0.004007	-0.006998	0.039527	-0.001041	0.006333	0.009234	-0.001231	-0.001386	0.003306
Other	-0.007757	-0.006536	-0.003392	-0.000708	-0.052058	-0.002288	1.000000	-0.017322	-0.000215	-0.019429	...	-0.001024	-0.002896	-0.004842	0.029768	0.006722	0.052704	0.006344	0.012585	-0.001002	-0.003969
Overcast	0.004800	-0.004108	-0.003133	-0.007419	-0.545418	-0.023968	-0.017322	1.000000	-0.002257	-0.203557	...	-0.010734	-0.030342	-0.095157	0.009912	-0.001153	0.005370	0.005387	-0.001547	-0.006066	0.124443
Partly Cloudy	-0.002331	-0.001262	0.003163	-0.000092	-0.006783	-0.000298	-0.000215	-0.002257	1.000000	-0.002532	...	-0.000133	-0.000377	0.001215	-0.000424	-0.000098	-0.000136	-0.000103	-0.000380	-0.000131	-0.000771
Raining	-0.020919	0.004794	0.016434	-0.008321	-0.611746	-0.026882	-0.019429	-0.203557	-0.002532	1.000000	...	-0.012039	-0.034032	-0.700040	-0.035726	0.016896	-0.007254	-0.001082	-0.015675	0.041583	0.768252
Severe Crosswind	-0.000069	0.001495	-0.001047	-0.000206	-0.015169	-0.000667	-0.000482	-0.005047	-0.000063	-0.005661	...	-0.000299	-0.000844	-0.007569	0.016919	-0.000219	-0.000305	-0.000230	-0.000850	-0.000292	0.005802
Sleet/Hail/Freezing Rain	-0.006245	-0.001939	-0.003965	-0.000439	-0.032258	-0.001418	-0.001024	-0.010734	-0.000133	-0.012039	...	1.000000	-0.001795	-0.035069	0.048421	-0.000466	0.025348	-0.000489	0.082541	0.026553	0.013227
Snowing	-0.019726	-0.004335	-0.020160	-0.001240	-0.091187	-0.004007	-0.002896	-0.030342	-0.000377	-0.034032	...	-0.001795	1.000000	-0.103215	0.162562	-0.001318	0.001245	0.006790	0.603189	-0.001754	-0.014817
Dry	0.052289	-0.008155	0.005381	-0.003845	0.736986	-0.006998	-0.004842	-0.095157	0.001215	-0.700040	...	-0.035069	-0.103215	1.000000	-0.120692	-0.027895	-0.038804	-0.029236	-0.108119	-0.037122	-0.902915
Ice	-0.020873	-0.008470	-0.014451	0.026975	-0.009153	0.039527	0.029768	0.009912	-0.000424	-0.035726	...	0.048421	0.162562	-0.120692	1.000000	-0.001482	-0.002062	-0.001553	-0.005745	-0.001973	-0.047978
Oil	-0.006496	-0.004409	0.002861	-0.000322	-0.011813	-0.001041	0.006722	-0.001153	-0.000098	0.016896	...	-0.000466	-0.001318	-0.027895	-0.001482	1.000000	-0.000477	-0.000359	-0.001328	-0.000456	-0.011089
Other	-0.005407	0.002812	0.002378	0.012085	-0.003728	0.006333	0.052704	0.005370	-0.000136	-0.007254	...	0.025348	0.001245	-0.038804	-0.002062	-0.000477	1.000000	-0.000499	-0.001847	-0.000634	-0.015426
Sand/Mud/Dirt	-0.006963	0.003292	0.000294	-0.000338	-0.005140	0.009234	0.006344	0.005387	-0.000103	-0.001082	...	-0.000489	0.006790	-0.029236	-0.001553	-0.000359	-0.000499	1.000000	-0.001392	-0.000478	-0.011622
Snow/Slush	-0.016621	-0.000965	-0.022067	0.021343	-0.076443	-0.001231	0.012585	-0.001547	-0.000380	-0.015675	...	0.082541	0.603189	-0.108119	-0.005745	-0.001328	-0.001847	-0.001392	1.000000	-0.001767	-0.042980
Standing Water	-0.000936	0.000366	-0.002294	-0.000429	-0.028245	-0.001386	-0.001002	-0.006066	-0.000131	0.041583	...	0.026553	-0.001754	-0.037122	-0.001973	-0.000456	-0.000634	-0.000478	-0.001767	1.000000	-0.014757
Wet	-0.019599	0.010133	0.015350	-0.004581	-0.672552	0.003306	-0.003969	0.124443	-0.000771	0.768252	...	0.013227	-0.014817	-0.902915	-0.047978	-0.011089	-0.015426	-0.011622	-0.042980	-0.014757	1.000000

I thought it may be better to also do a visualization of the data:



Important for me is only the correlation of the severity of accidents SEVERITYCODE in relation to all other attributes.

DISCUSSION

Looking at the heatmap I was a bit surprised that the correction in general was not as huge as a I have expected.

The highest correlations (TOP 5) have been:

- *Inattention (0.033)*
- *Influence of drugs & alcohol (0.033)*
- *Raining (0.016)*
- *Wet (0.015)*
- *Clear (0.006)*

We see that – noting that all correlations are not too strong – that there is a huge gap after the 4th and 5th position. I will continue on the TOP 4 for further analysis, because I think the correlation is more visible.

When I recap my origin question: are external or driver related factors have more correlation to the severity of accidents, I need to cluster the attributes in external and driver factors:

- *Inattention (0.033)* ➔ *Driver*
- *Influence of drugs & alcohol (0.033)* ➔ *Driver*
- *Raining (0.016)* ➔ *External*
- *Wet (0.015)* ➔ *External*

It is very obvious that the driver is much more impacting the severity of accidents and external factors are only secondary.

The correlation of inattention and influence of drugs & alcohol have the same correlation, so there is no clear distinction between these driver related attributes.

For the external factors there is a clear correlation of raining weather and thus resulting in wet streets condition.

CONCLUSION

Driver related factors have a higher impact on the severity of an accident with inattention and influence of drugs & alcohol being on par.

External factors have a lower impact with rainy weather and thus wet roads having the strongest influence on the severity of accidents.

As a result, we should focus on the ‘training and information distribution to vehicle drivers’ focusing on these topics:

- *Stay awake – inattention may kill you!*
- *Don’t drink and drive*
- *Rainy weather? Slow down!*