Car accident severity in Canada 1999–2017

Applied Data Science Capstone

Database

Data source and selection

As data source for this project we use the national collision database (NCDB) of Canada. It contains all police-reported motor vehicle collisions on public roads. Since Canada is a bilingual country, the data values are all numerical. The rich database contains a total of almost 7 million records. Each row corresponds to one person being involved in a car collision. Hence a collision between two cars A and B, where A has one passenger and B has two passengers, would create three records in the database.

Its 23 rows that are organized in three categories: data elements on collision level, on vehicle level, and on personal level. For the sake of this report, we drop the vehicle and personal level from the data, even though it would be possible to include them as additional features. On collision level, the database consists of the following 12 rows: "Year", "Month", "Day of week", "Collision hour", "Collision severity", "Number of vehicles involved in collision", "Collision configuration", "Roadway configuration", "Weather condition", "Road surface", "Road alignment", "Traffic control", and "Collision case". Obviously, "Collision severity" is a very significant dependent variable. We observe that the data source only distinguishes between car collision with and without death cases.

	C_YEAR	C_MNTH	C_WDAY	C_HOUR	C_SEV	C_VEHS	C_CONF	C_RCFG	C_WTHR	C_RSUR	C_RALN	C_TRAF	C_CASE
0	1999	1	1	20	2	02	34	UU	1	5	3	03	752
1	1999	1	1	20	2	02	34	UU	1	5	3	03	752
2	1999	1	1	20	2	02	34	UU	1	5	3	03	752
3	1999	1	1	08	2	01	01	UU	5	3	6	18	753
4	1999	1	1	08	2	01	01	UU	5	3	6	18	753
5	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
6	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
7	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
8	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
9	1999	1	1	15	2	01	04	UU	1	5	U	UU	932

Figure 1: First rows of the data set after dropping several columns.

Data cleansing, first step

We started with 6.772.563 rows, each value stored as string. Our data preparing should ensure numeric analysis, hence we converted the columns "Year", "Month", "Day of week", "Collision hour", "Number of vehicles involved in collision" to numeric values. For each column, there are values for *unknown* and for *not provided by jurisdiction*. In a first cleaning step, we removed all rows that have such values in a numeric row. This left us still with 6.705.062 records, thus we didn't lose many records. Next, we observed that some values are not stored consistently. For example, the month January is stored either as 1 or 01. We converted the numeric rows to integers and thereby got rid of this inconsistency.

There were still too many categorial variables, for example 21 different collision configurations. We dropped the columns "Collision configuration", "Road alignment", and "Traffic control" to reduce the size of the data, and gave the remaining columns intuitive names. There were 15 different road configuration categories which we simplified by a binary variable, indicating whether the accident occured at an intersection or not. Next, to improve the readability we applied a dictionary to encode the values for weather conditions and road surfaces. Thereby, we simplified the road surface column: there were 12 different categories before, which we compressed to *normal*, *wet*, *snow*, *other*.

	Year	Month	Weekday	Hour	Fatal	Nbr Vehicles	Intersection	Weather	Road Surface
0	1999	1	1	20	0	2	0	Clear	snow
1	1999	1	1	20	0	2	0	Clear	snow
2	1999	1	1	20	0	2	0	Clear	snow
3	1999	1	1	8	0	1	0	Hail	snow
4	1999	1	1	8	0	1	0	Hail	snow
5	1999	1	1	17	0	3	0	Clear	wet
6	1999	1	1	17	0	3	0	Clear	wet
7	1999	1	1	17	0	3	0	Clear	wet
8	1999	1	1	17	0	3	0	Clear	wet
9	1999	1	1	15	0	1	0	Clear	snow

Figure 2: First rows of the data set after the cleaning steps.

Data exploration

Our key dependent variable is the crash severity. We observed that there are 111.302 fatal cases and 6.593.760 cases without any fatality. Hence, with respect to crash severity the data set is very *unbalanced*.

First, we looked how the total number of cases evolved during time. As the portion of collisions with fatality is rather small, we created a separate plot for the fatal cases. In general, the numbers are decreasing over the years, as indicated by the linear regression line. We also created plots of the total number of collisions against month, against weekday, and against daily hour.

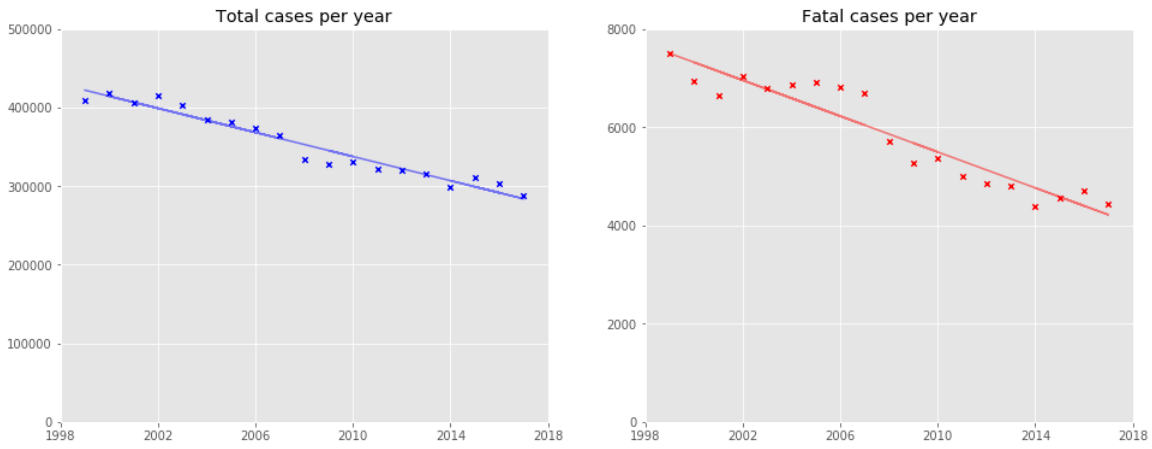

Figure 3: Total cases and total fatal cases per year. The drop around 2008 may be caused by the most recent amendments by the Canadian parliament to the law on drinking and driving.

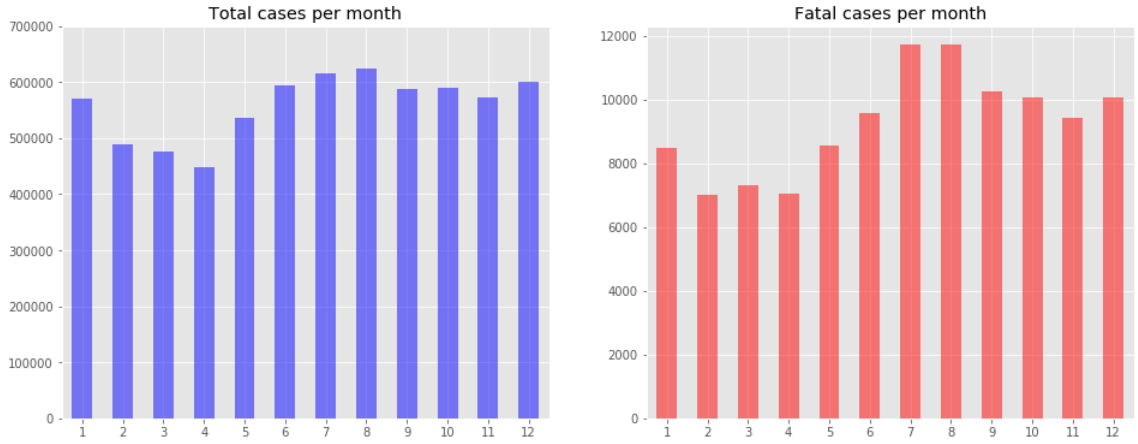

Figure 4: Total cases and total fatal cases per month. We observe that in July and August, the number of fatal collisions is over-average.

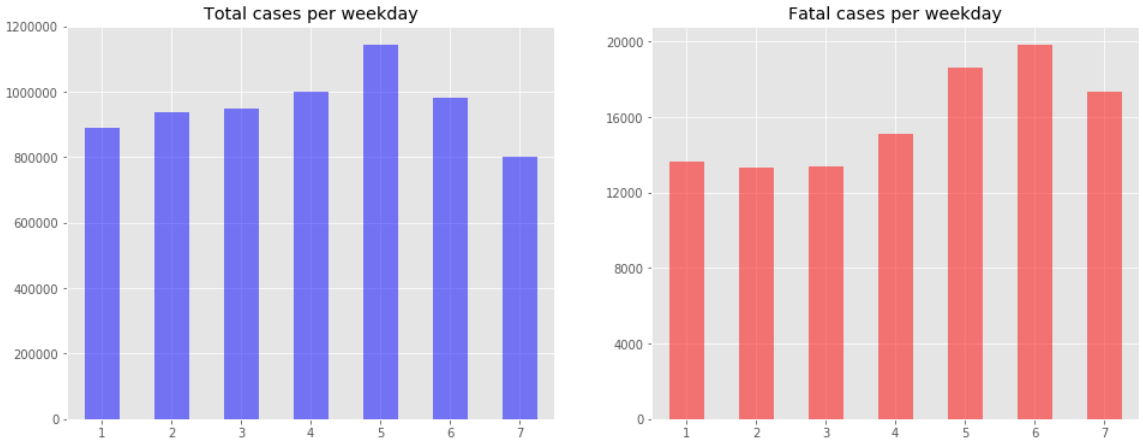

Figure 5: Total cases and total fatal cases per weekday. We see that most collisions occur on Fridays and the least on Sundays. However, the most fatal collisions happen on Fridays and Saturdays.

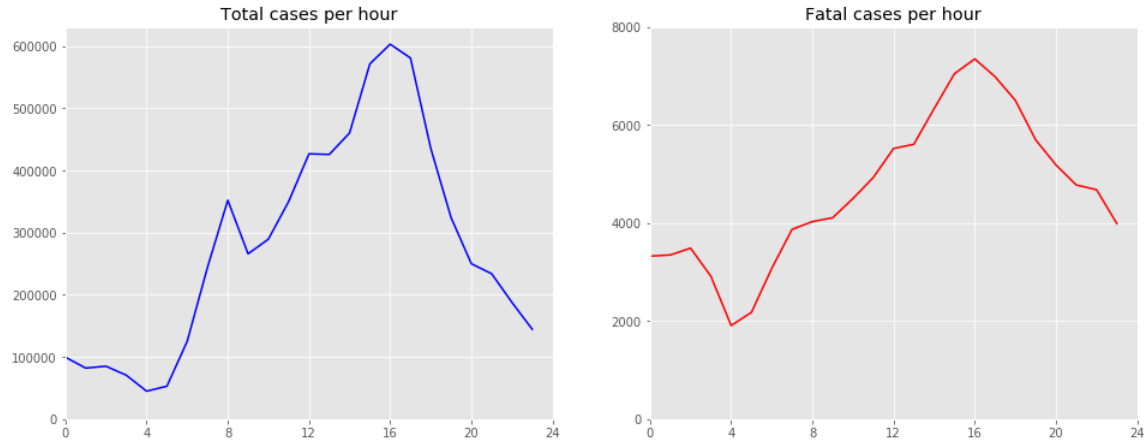

Figure 6: Total cases and total fatal cases per hour. Clearly, there is a peak around 4 PM. At night, the fraction of fatal cases is higher than at day.

Furthermore, we investigated the proportions of weather conditions and road surfaces. Here, it turned out that for the fatal cases, the proportions were rather similar, so we only illustrate the numbers for the overall collisions. The majority of collisions happens at clear weather, on normal road conditions. Minor but still significant weather conditions are "cloudy", "raining", and "snowing". For the sake of overview, we grouped all other weather conditions ("hail", "fog", "storm", "other", "unknown") to a single category.

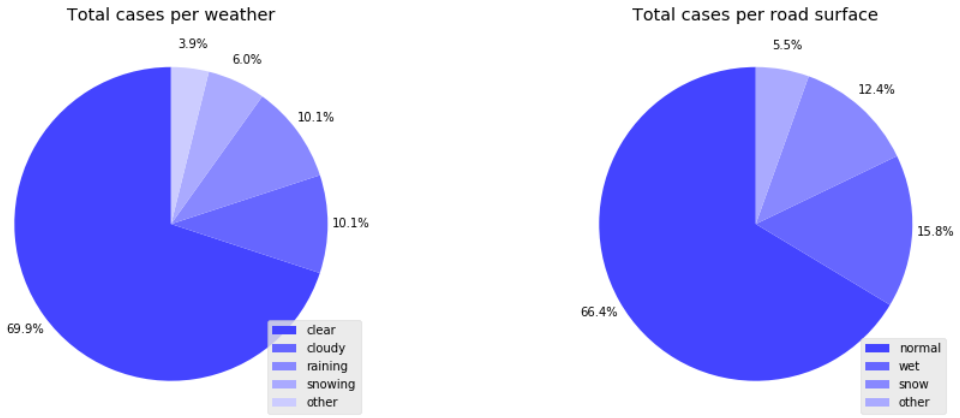

Figure 7: The total number of collision cases per weather category and per road surface category.

Data cleansing, second step

As discussed above, our data set is highly imbalanced. In fact, only 1.69% of the records are marked as fatal severity. Directly using this data would be highly problematic: a trivial algorithm could always predict "not fatal", achieving a brillant Jaccard-score of 0.9831. There are different ways to handle unbalanced datasets: oversampling fatal cases, undersampling cases without fatality, using modified evaluation scores, or using weighted cost functions in the algorithms. Since the data set is rather large and we suffered from performance and memory issues, we decided to undersample cases without fatality.

More precisely, we sampled an 0.1-fraction of the fatal cases and from the non-fatal records a random subset of fraction 0.002. This reduced the dataset to a total of 24.318 records, 11.130 being fatal cases, leaving us with a rather balanced dataset. Finally, we replaced the two columns for the road surface and the weather by dummy 0-1-columns. This variant is more suitable when training machine learning algorithms.

	Year	Month	Weekday	Hour	Fatal	Nbr Vehicles	Intersection	Clear	Cloudy	Fog	Hail	Raining	Snowing	Stormy	Unknown	normal	other	snow	wet
2610316	2005	6	2	17	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
4346196	2010	4	4	9	0	1	1	0	0	0	0	0	1	0	0	0	1	0	0
624422	2000	7	4	7	0	3	1	1	0	0	0	0	0	0	0	1	0	0	0
3668039	2008	3	5	10	0	3	1	0	1	0	0	0	0	0	0	1	0	0	0
2062666	2003	12	5	12	0	3	1	1	0	0	0	0	0	0	0	1	0	0	0
6073014	2015	9	2	16	0	2	1	1	0	0	0	0	0	0	0	1	0	0	0
5521204	2013	11	4	10	0	2	1	1	0	0	0	0	0	0	0	1	0	0	0
3057329	2006	7	7	2	0	2	0	1	0	0	0	0	0	0	0	1	0	0	0
4630273	2011	2	2	7	0	11	0	0	0	0	0	0	1	0	0	0	0	1	0
826725	2000	12	6	16	0	2	1	1	0	0	0	0	0	0	0	1	0	0	0

Figure 8: First rows of the reduced data set with dummy variables for the weather and road surface categories.