Data Analysis of Canada Car Accident during 1999 - 2014

Term Project Report

Project Group #3, 3250-15

**Boris Korotkov**

**Chaewon Kim**

**Patricia Luo**

**Ryan He**

School of Continuing Studies, University of Toronto

April 2018

Table of Contents

[Data Preparation 2](#_Toc508318085)

[What was the data source? 2](#_Toc508318086)

[How good was the data quality? 2](#_Toc508318087)

[What did you need to do to procure it? 2](#_Toc508318088)

[What tools or code did you need to use to prepare it for analysis? 2](#_Toc508318089)

[What challenges did you face? 3](#_Toc508318090)

[Analysis 4](#_Toc508318091)

[What trends, correlations and/or patterns do you see in the data? 4](#_Toc508318092)

[What …. 4](#_Toc508318093)

[Conclusions 5](#_Toc508318094)

[What did you learn about your data set? 5](#_Toc508318095)

[What …. 5](#_Toc508318096)

[Appendices 6](#_Toc508318097)

[A-1 sample dataset 6](#_Toc508318098)

[A-2 Program Code 7](#_Toc508318099)

# Analysis

## Individual analysis and the result we discovered:

1. **Collision trend, correlation and seasonality**

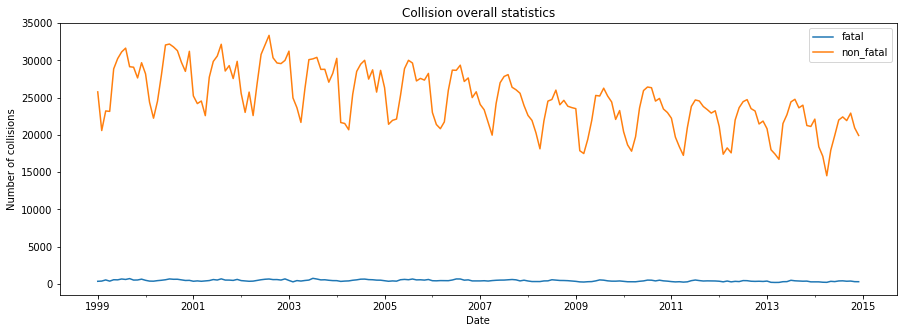
To assess collisions over the time, we created ‘date’ field using *PeriodIndex* function with year and month as arguments. Then we used the created ‘date’ column as an index for dataset (see sample code in **[A-2.B1]**).

* 1. Overall collision diagram

Since we wanted to see a trend to both, fatal and non-fatal collisions we had to create two more columns. One with fatal collisions and one with non-fatal ones. We used Numpy function *Where* to filter *c\_sev* column that includes information about collision severity to *fatal* column if value is equal to ‘1’ and to *non\_fatal* column if value is equal to ‘2’. As per data dictionary, ‘1’ value in *c\_sev* column represent a collision producing at least one fatality, and ‘2’ is used for collision producing non-fatal injury (see sample code in **[A-2.B2]**).

Then we grouped ‘fatal’ and ‘non\_fatal’ columns by date and summarized collision number (see sample code in **[A-2.B3]**).

The Dataframe was plotted as below:



We noticed that collisions have a seasonal pattern, but it was difficult to identify a period within 15 year scale. Thus, we created a smaller diagram for the most recent 5 years and identified a seasonality for non-fatal collisions as 12 months. The collision peak is in the middle of a year and off-peak is in the first quarter (Q1).



While we were able to see a seasonality patter for non-fatal collisions, the fatal ones are represented by a flat line. We assumed that the number of fatal collisions is very small in comparison with non-fatal ones and created a plot for fatal collisions only for the most recent 5 years of data.



The created plot confirmed our assumption. The fatal collisions have the same seasonal pattern as non-fatal collisions, e.g. a peak is in the middle of a year and off-peak in Q1.

* 1. Collisions autocorrelation

To confirm the seasonality assumption we decided to look at , the autocorrelation of fatal and non-fatal collisions was performed for the recent 5 years (see sample code in **[A-2.B4]**).

As per Autocorrelation diagram, we noticed that fatal and non-fatal collisions have strong autocorrelation with 12 month period.



* 1. Fatal and non-fatal collision correlation

We have already seen from the Autocorrelation plot, that fatal and non-fatal diagrams have similar shape and seasonality. However, it was interesting to know how strong they are correlated. We used dataframe *corr()* function to calculate correlation and Seaborn library to visualize the correlation. The correlation coefficient was determined as 0.87 and correlation diagram with Linear Regression line is below (see sample code in **[A-2.B5]**).

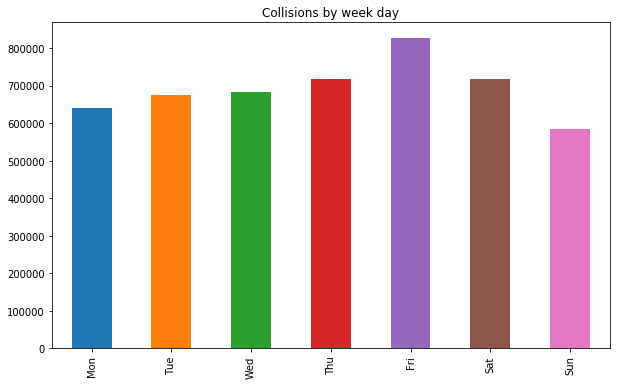


1. **Collision risk by week day and day hour**

One insight we were interested to get, it is collision correlation to week day and day hour.

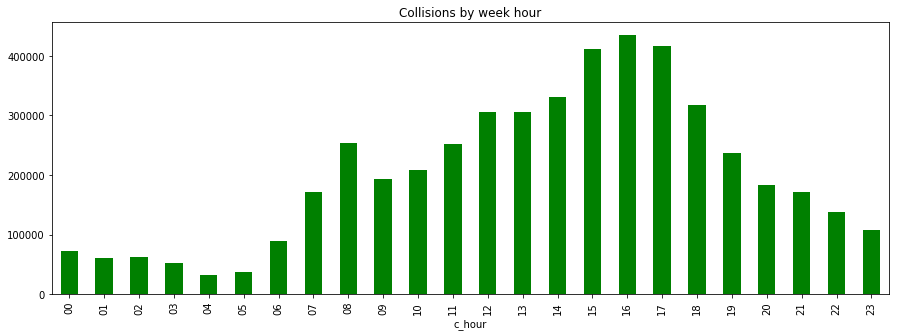
* 1. Collisions by week day

Since we found that fatal and non-fatal collisions are strongly correlated, we analysed collision correlation to week day for the whole collision population. We grouped collisions by *c\_wday* field and created bar type plot. We noticed that overall collision distribution over a week is similar with a small peak of Friday and downturn od Sunday (see sample code in **[A-2.B6]**).



* 1. Collisions by day hour

To get the collision distribution over day hours we grouped records by *c\_hour* field and created bar type plot. Hourly plot has two humps. One is at 8am and the second is at 4pm. These peaks reflect morning and afternoon rush hours (see sample code in **[A-2.B7]**).



1. **Road configuration, road surface, weather and traffic control**

The further analysis is similar in nature and will demonstrate the impact of four different collision factors to collision and fatality rate.

* 1. Road configuration

In addition to generate a bar plot showing the relation between collision severity and road configuration, we also plot a second axis using *twinx()* method of a subplot to show the fatal percentage for a collision happened in certain road configuration (the blue curve below with the y axis on right).



While mid-block and intersection collisions represent the biggest portion of collisions, the passing or climbing lane collisions have the highest fatality rate (~12%). The possible reason could be a speed difference between vehicles at the collision moment and a vulnerable position of vehicle that was trying to change the lane.

* 1. Weather

Surprisingly, neither snow or rainy weather hit the record in overall statistic and fatality rate. The majority of collisions had place in a clear and sunny day. The most dangerous weather for drivers was weather with limited visibility like as fog, smog and mist and weather with strong wind. The fatality rate in such weather condition was about 4%.



* 1. Road Surface

The analysis done for road surface brought to our attention that dry and normal road surface can not guarantee a safe trip. Dry road leads in the statistics while fatality rate is not significant (~1.5%). The most life-threatening collisions happen on sandy, dirty and muddy roads. The fatality rate there is between 3.5% to 4.5%.



* 1. Traffic Control

The traffic controls are designed to limit the number of collisions and make our commute safer. As we can see from the graph, the number of collision almost the same when traffic signals were fully operated and no traffic control presented. The most critical collision type from fatality perspective were collisions happened in reduced speed zone and near railway traffic controls (7%-8%). While high fatality rate near railway signals can be explained by collision with train, the high collision rate in reduced speed zone was not easy to understand. One possible explanation could be a collision with minors, because the reduced speed zone signs are installed near schools and day care services.



# Conclusions

## What did you learn about your data set?

1. **Collision trend, correlation and seasonality**

From the analysis we performed above, following observations:

* Collisions number was declining over 15 year period
* Fatal and non-fatal collisions are auto-correlated and strongly correlated between themselves.
* The collisions have seasonality pattern. The peak happens is the middle of a year and off-peak is in the first quarter.

1. **Collision risk by week day and day hour**

From the analysis we performed above, following observations:

* Overall collision distribution over a week is similar with a small peak of Friday and downturn od Sunday
* There are two peaks of collision risk. One is at 8am and the second is at 4pm. These peaks reflect morning and afternoon rush hours

1. **Road configuration, road surface, weather and traffic control**

From the analysis we performed above, following observations:

* Mid-block and intersection collisions represent the biggest portion of collisions.
* The collisions with passing or climbing lane have the highest fatality rate ~12%. The possible reason could be a speed difference between vehicles at the collision moment and a vulnerable position of vehicle that was trying to change the lane.
* The majority of collisions had place in a clear and sunny day.
* The most dangerous weather for drivers was weather with limited visibility like as fog, smog and mist and weather with strong wind. The fatality rate in such weather condition was about 4%.
* The most often collisions occurred on dry and normal road surface.
* The most life-threatening collisions happen on sandy, dirty and muddy roads. The fatality rate there is between 3.5% to 4.5%.
* The present of traffic control signs and signals was not a defining factor of collision. The proportion of collisions in both cases was similar. The most critical collision type from fatality perspective were collisions happened in reduced speed zone and near railway traffic controls (7%-8%). While high fatality rate near railway signals can be explained by collision with train, the high collision rate in reduced speed zone was not easy to understand. One possible explanation could be a collision with minors, because the reduced speed zone signs are installed near schools and day care services.

# Appendices

## A-2 Program Code

1. **[A-2.B1]** create new column ‘date’ with data type as ‘period[M]’ and set this column as an index (see line 7 and 8 in “G3\_Project\_Boris.ipynb”).

|  |
| --- |
| df['date'] = pd.PeriodIndex(df['c\_year'].map(str) + '-' + df['c\_mnth'], freq='M')  df=df.set\_index('date') |

1. **[A-2.B2]** create ‘fatal’ and ‘non\_fatal’ columns based on c\_sev column value (see line 9 in “G3\_Project\_Boris.ipynb”).

|  |
| --- |
| df['fatal']=np.where(df['c\_sev']==1,1,0)  df['non\_fatal']=np.where(df['c\_sev']==2,1,0) |

1. **[A-2.B3]** group ‘fatal’ and ‘non\_fatal’ columns by date (see line 10 in “G3\_Project\_Boris.ipynb”).

|  |
| --- |
| df\_1=df.groupby('date')['fatal','non\_fatal'].sum() |

1. **[A-2.B4]** Autocorrelation plot (see line 15 in “G3\_Project\_Boris.ipynb”).

|  |
| --- |
| plt.figure(figsize=(15,5))  for c in df\_1.columns:  autocorrelation\_plot(df\_1[c][-60:],label=c) |

1. **[A-2.B5]** correlation plot andcorrelation coefficientdetermination (see line 16 and 17 in “G3\_Project\_Boris.ipynb”).

|  |
| --- |
| sns.regplot('fatal', 'non\_fatal', data=df\_1)  corr = df\_1.corr() |

1. **[A-2.B6]** group collisions by week day and create a bar plot (see line 19 in “G3\_Project\_Boris.ipynb”).

|  |
| --- |
| by\_weekday = df.groupby('c\_wday')['c\_sev'].count()  by\_weekday.index = ['Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun']  plot2 = by\_weekday.plot(kind='bar',title='Collisions by week day') |

1. **[A-2.B7]** group collisions by day hour day and create a bar plot (see line 20 in “G3\_Project\_Boris.ipynb”).

|  |
| --- |
| plt.figure(figsize=(15,5))  by\_hour = df.groupby('c\_hour')['c\_sev'].count()  plot3 = by\_hour.plot(kind='bar',title='Collisions by week hour',color='G') |