Data Analysis of

Canada Car Accidents: 1999 - 2014

Term Project Report

SCS 3250-015, Group 3

**Boris Korotkov**

**Chaewon Kim**

**Patricia Luo**

**Ryan He**

April 2018

Table of Contents

[Project Overview 2](#_Toc511402045)

[Data Preparation 2](#_Toc511402046)

[What was the data source? 2](#_Toc511402047)

[How good was the data quality? 2](#_Toc511402048)

[How did we procure our data? 2](#_Toc511402049)

[What tools/code did we use to prepare for our analysis? 2](#_Toc511402050)

[What challenges did we face? 2](#_Toc511402051)

[Analysis 3](#_Toc511402052)

[What trends, correlations and/or patterns did we see? 3](#_Toc511402053)

[Individual analysis and the result we discovered: 3](#_Toc511402054)

[Conclusion 14](#_Toc511402055)

[Appendices 15](#_Toc511402056)

[A-1 sample dataset 15](#_Toc511402057)

[A-2 Sample Codes 15](#_Toc511402058)

# Project Overview

We analyzed the dataset containing records of collision for car accidents in Canada from 1999 to 2014. Given various information pertaining to collision, vehicle and person, we conducted analyses to understand the trends and patterns of collisions and discover factors and situations that were more prone to accidents.

We created an open repository on [GitHub](https://github.com/rynho/3250-G3) containing background information, codes (via Jupyter Notebook) and reports to allow members to easily collaborate and track changes.

# Data Preparation

## What was the data source?

We selected the [Canadian Car Accidents 1999-2014](https://www.kaggle.com/tbsteal/canadian-car-accidents-19942014/data) dataset (provided by Transport Canada) from Kaggle: an open platform for predictive modeling and analytics competition using real life datasets provided by companies and users.

## How good was the data quality?

The overall quality of the dataset was decent except there were several missing values, unknown/inapplicable values and varying data types. For example, besides the 17 vehicle types defined in ***vehicle type***, there were four nonnumeric values:

* ‘NN’ not applicable, e.g., “dummy” vehicle record created for the pedestrian;
* ‘QQ’ choice other than defined vehicle types;
* ‘UU’ unknown values; and
* ‘XX’ jurisdiction does not provide the data.

Since every column contained some nonnumeric values, Pandas treated all of them as strings, which required manipulation of data when trying to utilize numeric-only features (e.g. scatterplots).

Moreover, the values in the ***c\_wday, Day of week*,** column represented specific days of week (e.g. 1 = Monday, 7 = Sunday) instead of representing days of month, which created some difficulty when trying to use time series index for our analysis (detail in below sections).

## How did we procure our data?

The full dataset, in **csv** format (including header), included around 6 million records of car accidents in Canada over the period of 15 years from 1999 to 2014 and the data dictionary was included in a pdf file. Each row of the dataset represented registered records of each person involved in that specific collision, and there were 22 columns presenting information pertaining to collision, vehicle and person level data elements of each collisions.

We downloaded the dataset (.csv) from the URL mentioned above, used the system command to perform some sense check, then used *pandas.read\_csv* method to load the full dataset into a *DataFrame*. See Appendix **[A-2.1]**.

## What tools/code did we use to prepare for our analysis?

We utilized Python (Pandas, NumPy) for data cleaning and preparing. Some of our data cleaning include the following but are not limited to:

* Converted data type from ‘str’ to ‘int’ for certain numeric columns;
* Converted non-numeric values to NaN and used drop.na() as they do not present significant insights; and
* Composed a ***date*** column combining ***year*** and ***month*** for possible usage in the analysis. See **[A-2.2]** for the sample code.

## What challenges did we face?

Some of the challenges we faced while preparing the data for analysis include the following but are not limited to:

* The dataset did not include ‘day’ part of (YY/MM/DD) but a column ***c\_wday*** representing specific days of week (e.g. 1 = Monday, 7 = Sunday). In order to enhance the accuracy of our analysis, we decided to combine just the ***year*** and ***month*** column to create a time series and kept the ***c\_wday*** for certain analyses.
* We experimented two different ways to engineer the ***date*** column and noted their advantages and disadvantages. We discovered that *pandas.DatetimeIndex* method is faster, but incorporating ***c\_wday*** in the result as day of month produced e.g. 1999/01/01 when it actually represents *1999 January, Monday*. On the contrary, while the *pandas.PeriodIndex* with ‘M’ frequency accurately shows only the year and month, the method is about 10 times slower given the huge dataset. We decided to use one or the other in our individual analyses to compare the difference.
* All values in the dataset were in ‘str’ type. We decided to change the data type when we created a duplicate of the dataset and/or sliced data to proceed with further analysis accordingly.
* Considering that some unknown values may present significant insights into our analysis, we decided to keep them in the general dataset but treat them differently in individual analyses. However, the unknown values in date/time columns are dropped as they not only represent tiny portion of the data and are insignificant, but also complicate the analysis.

# Analysis

## What trends, correlations and/or patterns did we see?

In the following section, we addressed the most interesting findings. However, refer to the attached Jupyter Notebooks for all details, and the analyses not mentioned here. We analyzed the nature of collision records by considering different factors/conditions, as well as the relationships/correlations between those factors/conditions to identify trends and patterns. Before doing each analysis, we sliced the dataset pertaining to only the columns needed in order to maintain original data integrity, make debugging easier, and avoid reloading the huge dataset which takes longer.

## Individual analysis and the result we discovered:

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

To analyze the collision trend over the time, we created a ***date*** column using the pd.*PeriodIndex* function (collision year and month as arguments) and set it as the index of our dataset (see sample code in **[A-2.3]**).

* 1. **Overall collision diagram**

In order to observe the trend for both fatal and non-fatal collisions, we created two more corresponding columns by using the ***collision severity (c\_sev = 1 fatal , c\_sev =2 non-fatal)*** column. (see sample code in **[A-2.4]**). Then we grouped the ***fatal*** and ***non-fatal*** collisions by ***date*** and summed the number of collisions (see sample code in **[A-2.5]**).



When we plotted the initial result, we noticed that collisions have a seasonal pattern. Nevertheless, it was difficult to identify the exact pattern since the time period was a 15-year scale. Therefore, we created another plot with a smaller range (most recent 5 years, 2010-2014) as shown above. We identified a seasonality for non-fatal collisions as 12 months; the number of collisions increases over time and peaks during mid-year, declines towards first quarter (Q1) and repeats.

While we were able to grasp a 12-month seasonal pattern for non-fatal collisions, the fatal collision line resembled that of a straight line. We assumed that this is due to the very small number of the fatal collisions in comparison to non-fatal ones, so we created a plot just for fatal collisions for the most recent 5 years of data.



The graph above confirmed our assumption that fatal collisions have the seasonal pattern as non-fatal collisions: increases and peaks mid-year and starts to decrease and reaches off-peak in Q1.

* 1. **Fatal vs. Non-fatal collisions: Autocorrelation & Correlation**

To further confirm our fatal and non-fatal collisions’ seasonality statement, we examined the autocorrelation of fatal and non-fatal collisions for the most recent 5 years (see sample code in **[A-2.6]**).



By looking at the autocorrelation diagram (left), we noticed that fatal and non-fatal collisions have a strong autocorrelation with a 12-month lag thus solidifying our result. Furthermore, we used the *df.corr()* to find out how strongly they were correlated. The correlation diagram (right) with the regression line shows that they are strongly correlated; in fact, the correlation coefficient was determined as 0.87. (See sample code in **[A-2.7]**)

1. **Number of vehicles involved in collision (c\_vehs) by Day of week (c\_wday) and Hour (c\_hour)**

***c\_vehs*** column represents the number of vehicles involved in collision. We counted the number of records for certain values in ***c\_vehs*** and confirmed that every vehicle involved in that specific accident had its own record (e.g., there were 58 records for 57 vehicles involved in that collision: the remaining one could be a pedestrian.) ***c\_hour*** represents the time when the collision happened, and we should note that the numbers indicate an interval of time (e.g. ***hour*** 8 represents 8:00-8:59am).

We wanted to observe the relationship between total ***c\_vehs*** over ***c\_wday*** *and* ***c\_hour****.* We grouped the ***c\_vehs*** *by* ***c\_wday*** and the results are shown as below (left). The total number of collisions increases as the days go by, peaks on Friday and decreases towards Sunday. In fact, least number of collisions occurred on Sunday. One of the reasons for this phenomenon is that most working people stay at home on Sundays with their families and enjoy their time off so there are less vehicles on the street. (See sample code in **[A-2.8]**).



To observe the trend of total number of collisions in regard to hours, we grouped the ***c\_vehs*** records by ***c\_hour***. From the graph above (right), we identified two local maxima: one at hour 8, 8:00-8:59AM and the second at hour 16, 16:00-16:59pm. These peaks reflect morning and afternoon rush hours (see sample code in **[A-2.9]**).



The graph on the left shows the trend of number of vehicles involved in collision over time. The maximum number of vehicles involved in a single accident was 77 but since it was hard to observe a pattern, we selected the top 5 most frequent number of vehicles involved in collision to do our analysis. We can see that the overall number of vehicles involved in collision decreased over time and that 2 vehicles were most likely to be involved in a collision followed by 1 and 3 vehicles. The graph on the right shows the trend of number of vehicles involved in collision at different hours. The two peaks are apparent at ***hour*** 8 and ***hour*** 16 which resembles the exact same pattern for the total number of vehicles involved in collision at different hours.

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

The following section will analyse the impact of four different collision factors to collision severity and fatality rate.

* 1. **Road configuration (c\_rcfg)**

In addition to generating a bar graph showing the relationship between collision severity and road configuration, we also included a line plot using the *twinx()* method of a subplot to show the fatal percentage for collisions that happened in each road configuration (the blue curve seen below with the y axis on right).



Most number of fatal and non-fatal collisions occurred at mid-block followed by intersection. While passing or climbing lane had the least number of non-fatal collisions, the fatal percentage was the highest at around 12%. A possible reason could be the difference in speed between vehicles; the vehicle trying to join the passing lane has to speed up since other vehicles travel faster on the passing lane and this puts them in a vulnerable situation. Moreover, the number of fatal collisions for traffic circle was the lowest because drivers are extra cautious since they are aware that both turning and through traffic join onto a one-way circular roadway.

* 1. **Weather (c\_wthr) & Road Surface (c\_rsur)**



From the graph above we can see that surprisingly, neither snow nor rain contributed to more collisions and higher fatality rate. In fact, the majority of collisions actually took place in clear and sunny days. The most dangerous weather for drivers was weather with limited visibility: fog, smog, mist and weather with strong wind. The fatality rate in such weather condition was about 4%.



The analysis done for road surface above brought to our attention that dry and normal road surface does not guarantee a safe trip. Dry road leads in number of non-fatal collisions while fatality rate is not significant (~1.5%). The most life-threatening collisions happen on sandy, dirty and muddy roads with fatality rate 3.5% and 4.5% respectively.

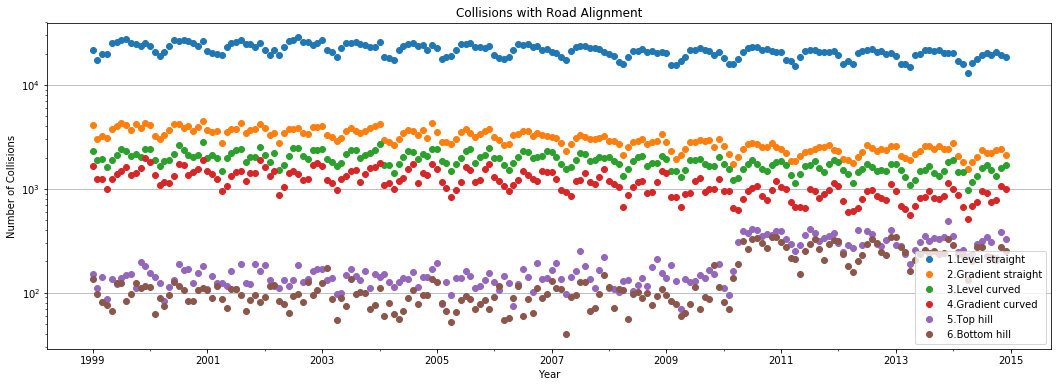
* 1. **Traffic Control (c\_traf)**

Traffic controls are designed to inform, guide and control traffic to enhance the safety of motorists and pedestrians. As we can see from the graph below, most number of collisions happened when traffic signals were fully operated, and no traffic control was present. The highest fatality rate occurred when collisions happened in reduced speed zone and railway crossing with signs only, around 7% and 8% respectively. While high fatality rate near railway signs can be explained by collision with trains, the high collision rate in reduced speed zone was not easy to comprehend. One possible explanation could be that because it is a reduced speed zone, there are pedestrians trying to cross without traffic lights and this increases the chances of getting in accidents and the reduced speed zone signs are usually present near schools and day care services.



1. **Road Alignment (c\_raln)** 
   1. **Road Alignment (c\_raln) & Time relation (date)**

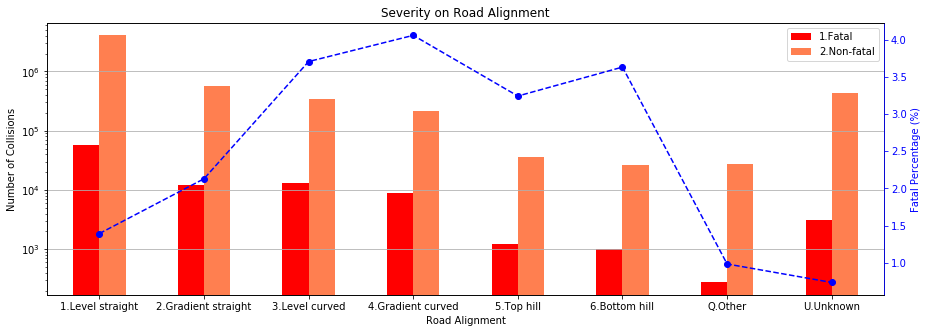
We sliced and grouped the data by ***date*** and ***c\_raln***, used the *size* method to aggregate the sum for each possible **(*date, c\_raln*)** pair into a two-level indexed *pandas.Series* and used the *unstack* method to transform into a new *Dataframe* (sample code in **[A-2.10]**).



From the graph above, we saw that most collisions happened on *Straight and level* road alignment. A possible reasoning could be that drivers were not cautious enough and drove at a high speed given the easy road.

* 1. **Road Alignment (c\_raln) & Collision Severity (c\_sev)**

The graph below shows the relationship between collision severity and road alignment and presents the fatality rate at each road alignment.



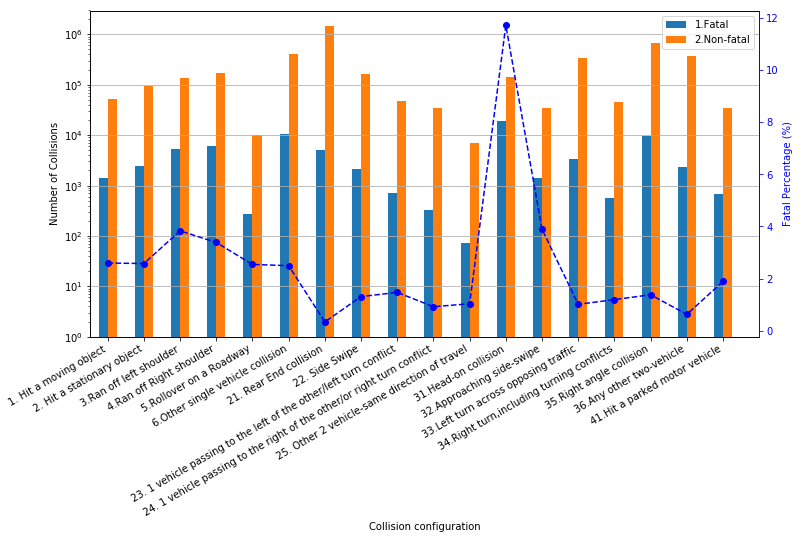
Generally, there were less fatal accidents compared to non-fatal accidents (less than 4% in any given cases) and to no surprise, the Straight-level road alignment had a lower fatality rate, even though the total number of collisions was much higher than that of other cases. On the other hand, the curved alignment led to more severe accidents; this could be caused by limited sight.

1. **Collision configuration (c\_conf)**
   1. **Collision configuration (c\_conf) & Number of vehicles (c\_vehs)**



***c\_conf*** column indicates how one or two vehicles were arranged during a collision: their motion, traveling direction and relative position. From the graph above, we saw that the most number of vehicles were involved in a rear end collision followed by a right-angle collision.

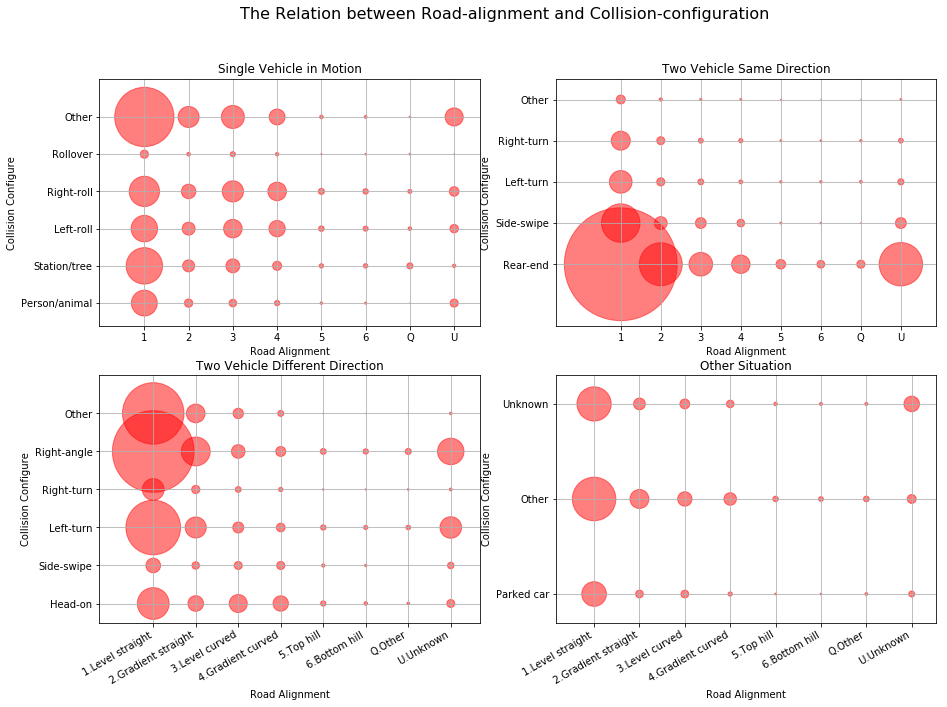
* 1. **Collision configuration (c\_conf) & severity (c\_sev)**

****

We examined the relationship between ***c\_conf*** and ***c\_sev***and we saw thatwhile rear-end non-fatal collision was the most common, its fatality rate was the lowest. The number for right angle non-fatal collisions followed rear-end collisions. On the other hand, head-on collisions had the most number of fatal collisions and the highest fatal percentage at around 12%.

* 1. **Collision configuration (c\_conf) & Road alignment (c\_raln)**

Considering the complexity of collision configurations, we grouped them into four categories by number of vehicles and their directions and plotted them into four subplots.

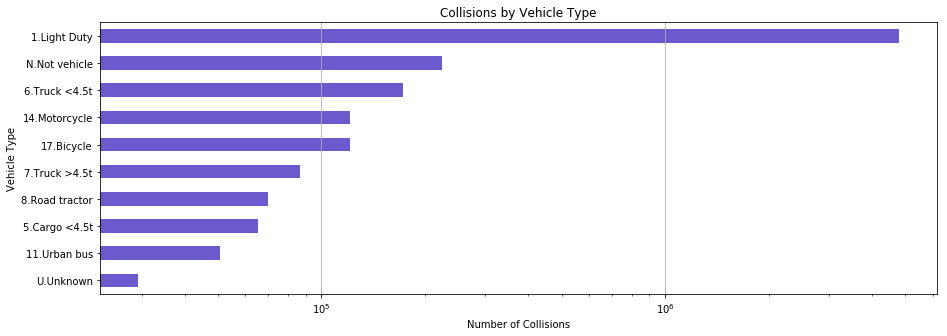


The size of each bubble indicates the number of collisions in each setting and the scale for all subplots were equal, which makes the results comparable. In all cases, there was a clear trend that as road-alignment improved, the collision chances increased due to higher speed and less caution. Disregarding the unknown or "other" conditions, we can see from each case that:

* Most collisions for one car accident occurred from hitting a static object, followed closely by right and left roll on to shoulder.
* For two cars travelling in the same direction, rear-end collision was the most common and we can see from the scatter plot above that the possibility increases moving towards left on road-alignment.
* For two cars traveling in different direction, right-angle collision is the most frequent, most likely at an intersection. Left turn also seems to be risky while right turn shows very low risk: a result of the inherent complexity of left turn.

1. **Vehicle Type (v\_type)**

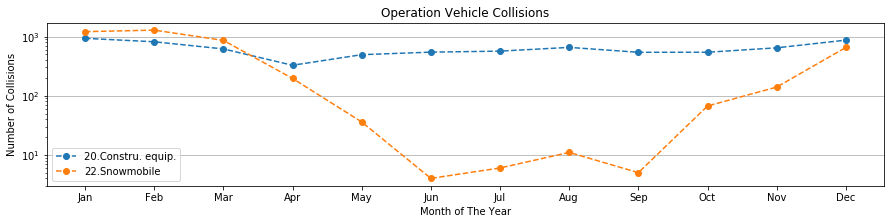
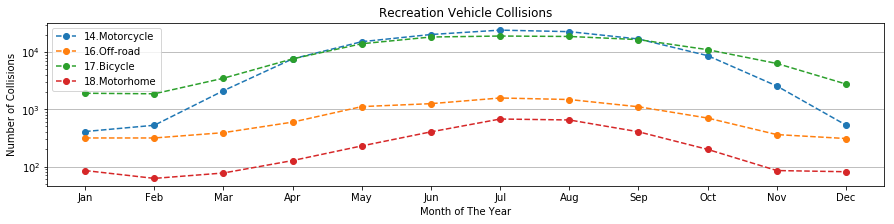
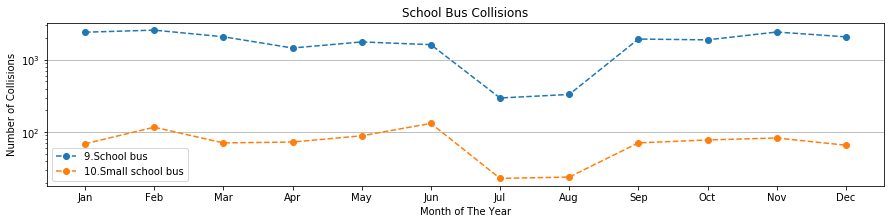
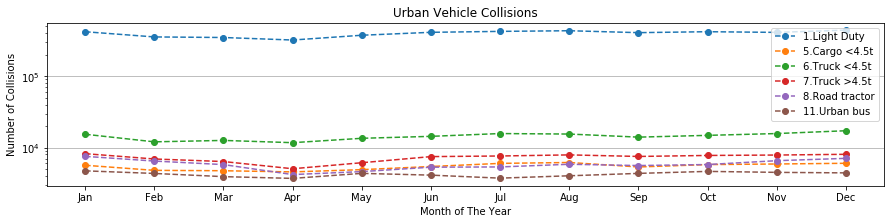
***v*\_type** indicates the vehicle types involved in accidents e.g., truck, school bus, bicycle, etc.



From the graph above, we can see that vast majority of collisions were caused by light duty vehicles, followed by light truck, motorcycle and bicycle.

* 1. **Vehicle Type (v\_type) & Time Relation (date)**

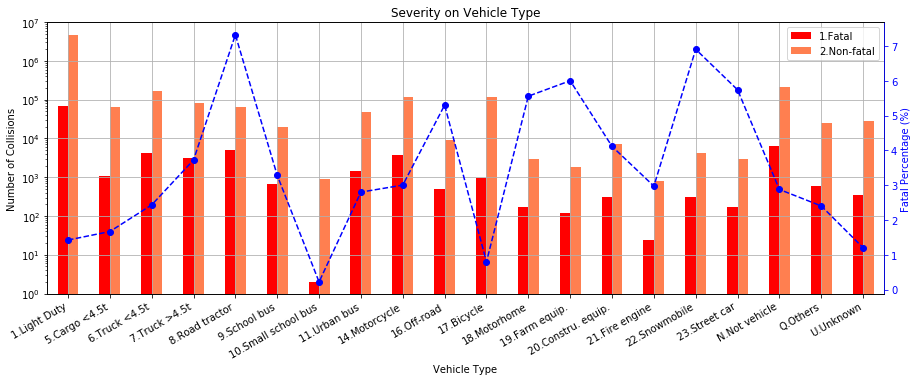
Because there were around 20 different types of vehicles in the dataset, we grouped similar types, and plotted them into a set of graphs, which focused mostly on the seasonal pattern observed for each vehicle types.



By looking at the plots above, different vehicle types showed different observations:

* The number of collisions for urban vehicles (light duty, trucks, road tractor, and bus) is flat across all seasons, indicating that urban vehicles are in duty all throughout the season.
* School bus collisions decreased in July and August due to summer vacation.
* More accidents related to bicycle, motorcycle, off-road vehicle and motorhome occurred from April to October, as they are used for outdoor activities during warmer temperature, whereas more snow-mobile collisions were prevalent from October to April.
  1. **Vehicle Type (v\_type) & Severity (c\_sev)**

Similar to the analysis for *road alignment* and*collision severity*, we plotted the results for fatal and non-fatal collisions by different vehicle types with their corresponding fatality rates with a blue dotted line.



From the figure above, we see that most fatal and non-fatal collisions occurred with light-duty vehicles, followed by trucks, motorcycle, and bicycle. However, accidents with the highest fatality rate were caused by road tractor and snowmobile at around 7%, followed by street car, farm equipment and motorhome at around 6%. It is also noticeable that small school bus has the least fatal collisions as well as that of bicycle.

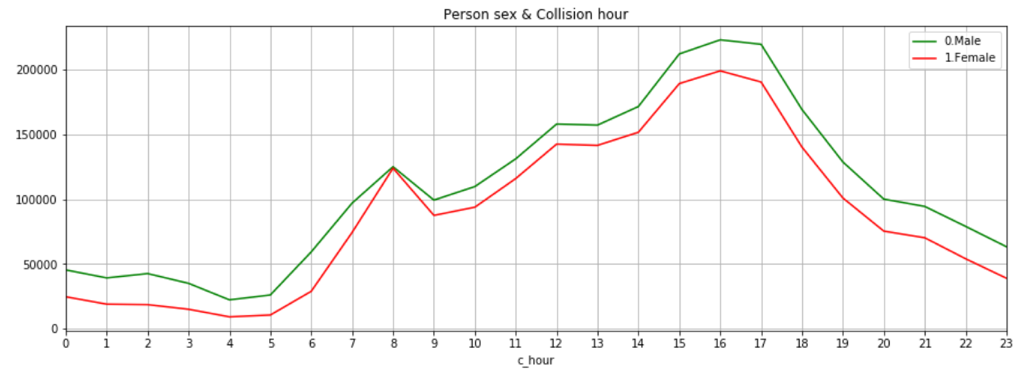
Apart from some external factors and conditions, human factors also play a significant role in contributing to the likelihood of a collision. Therefore, the following analysis will pay more attention to some person-level data elements, focusing on person sex, age, position, and safety device used.

1. **Person sex (p\_sex)**

***p\_sex*** indicates the gender of the individuals involved in accidents. For convenience, all unknown and missing values were dropped.

* 1. **Person Sex (p\_sex) & Collision Hour (c\_hour)**

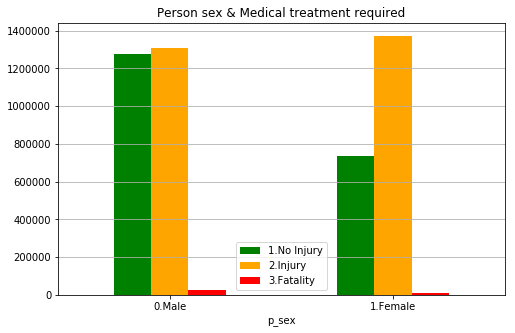
We grouped the data by ***p\_sex*** and ***c\_hour*** by using the count method and we found that 2,608,529 males and 2,116, 502 females in total were involved in accidents over the 15 years. Consequently, we assumed that males were more likely to be involved in an accident than females. The graph below shows the relation of gender involved in collision and collision hour.



The graph above follows the observed pattern of the ***hour*** plots discussed earlier on in the report and our assumption was confirmed that at all hours, males were more likely to be involved in collisions than females.

* 1. **Person Sex (p\_sex) & Medical Treatment Required (p\_isev)**

***p\_isev*** indicates the degree of casualty for each collisions. After dropping all unknown and missing values, the data indicated three levels of casualty: *No injury*, *Injury* and *Fatality*. We used the same method as earlier by aggregating the result into a two-level index table ***p\_sex*** and ***p\_isev***, and plotted a bar graph.



By looking at the graph above (left), we can conclude that females are more likely to get injured rather than experience no injury in an accident. Since we could not see the results for fatality well, we rescaled the graph into a logarithm scale. We have to note that *fatality* refers to died immediately or within the time limit. Consequently, when we look at the graph above (right) we can see that males were more likely to die from collisions than women.

1. **Person age (p\_age)**

***p\_age*** represents age of people involved in collisions. 0 indicates less than 1 years old and 99 indicates 99 years or older. We binned people into 10 age groups e.g. 10 represents 11 to 20 years, 20 represents 21 to 30 and so on to enhance visualizing the results of our analysis (see **[A-2.11]**).

* 1. **Person Age (p\_age) & Collision Hour (c\_hour)**



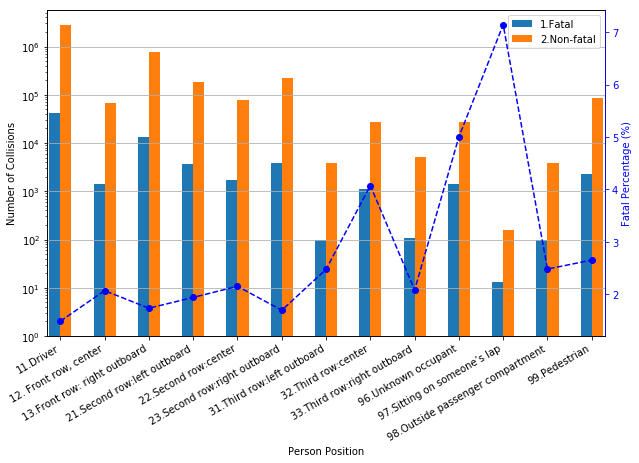
The 10 different color lines in the graph above represent 10 age groups. We can see that the age group most prone to accident is 21 to 30 years old. Similar to our ***hour*** plots earlier in the report, there are two peaks at ***hour 8*** and at **hour *17***. In the morning from 8:00 am to 11:59am, three age-groups 21 to 30, 31 to 40 and 41 to 50 resembled a similar pattern; decreases towards ***hour 9*** and increases towards ***hour 12***. We also observed that age groups involved work force are most likely to get into accidents during rush hour 8:00-8:59am and 16:00-17:59pm.

* 1. **Person Age (p\_age) & Medical Treatment Required (p\_isev)**



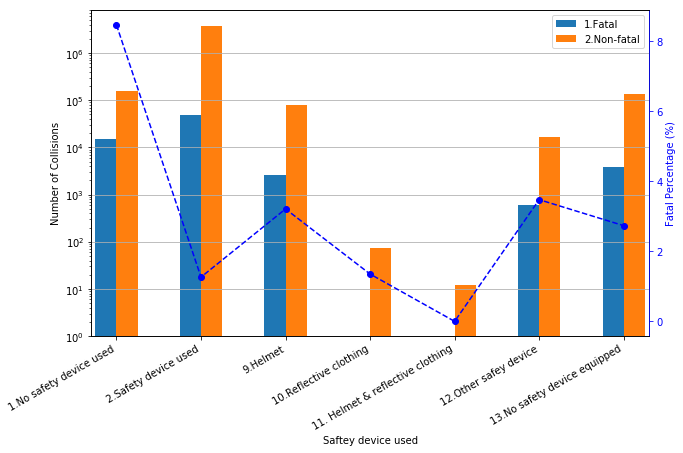
From the graph above, young adults aged 10 to 20 and 20 to 30 show the highest number of injuries but the 20 to 30 age group had the highest fatality numbers during collisions. Those aged above 90 have lower overall casualty results compared to that of other age groups. Statistically, from 1999 to 2014 only 185 people above the age of 90 died in accidents whereas 7064 young adults died.

1. **Person position (p\_psn) vs. Collision severity (c\_sev)**



***p\_psn*** indicates where the person was positioned when the collision happened. There were 17 categories in total but unknown (e.g. runaway cars), inapplicable (e.g. “dummy” person record created for parked cars) and missing categories were dropped to simplify the analysis. By looking at the graph above (left), we can see that people that were driving were most likely to be involved in both fatal and non-fatal accidents followed by people seated in the front row: right outboard. However, drivers actually had the lowest fatality rate whereas people sitting on someone’s lap had the lowest number of collisions but the highest fatality rate at 7%. This emphasizes the common knowledge of the importance of wearing a seat belt and using a **safety device, *p\_safe*.**

1. **Safety device used (p\_safe) vs. Collision severity (c\_sev)**

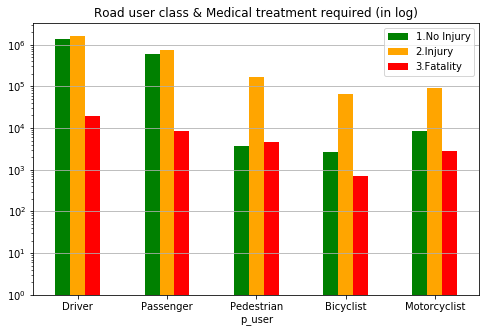


There were 13 categories for ***p\_safe*** column but similar to ***p\_psn***, the unknown, inapplicable and missing categories were dropped for better analysis. By looking at the graph above (right), surprisingly, the most number of fatal and non-fatal collisions happened when safety device was used. It was then followed by no safety device used and no safety device equipped (e.g. buses). However, we can note that the fatality rate for collisions with safety device used is low while the fatality rate is the highest at 8% when no safety device was used. We can also note that for motorcyclists, bicyclists, snowmobilers, all-terrain vehicle riders, the number of collisions as well as fatality rate decreases when both helmet and reflective clothing were used: instead of either just helmet or reflective clothing.

1. **Road user (p\_user)**

**Road User (p\_user) vs. Medical Treatment Required (p\_isev)**

***p\_user*** represents those who were using the road such as drivers, pedestrians, cyclists or motorists who was involved when accidents occurred. Dataset included 5 road users (driver, passenger, pedestrian, bicyclist, motorcyclist and unknown) with codes from 1 to 6. For consistency, the unknown values were dropped.



By looking at the bar graph above, we can see that the number of casualties for all three degrees is the highest for drivers. Drivers are most likely to experience no injuries while they are also most likely to die from an accident. Passengers followed the lead by having high numbers for all three degrees of casualty. Among all 5 user groups, bicyclists had the lowest figure for all three degrees for casualties. However, we can also note that when bicyclists are involved in collisions, the possibility of getting injured is much higher than experiencing no injury at all. Motorcyclists’ fatality figure was smaller than that of pedestrians. This could possibly be explained by the use of safety devices. As pedestrians are usually not equipped with safety devices they are more vulnerable when they get into accidents. On the other hand, from the ***p\_safe*** *and* ***c\_sev*** graph earlier motorcyclists equipped with helmet and reflective clothing have a lower fatal percentage and are more protected than pedestrians.

# Conclusion

For this term project, our group analyzed the dataset containing records of collision for car accidents in Canada from 1999 to 2014. Given the immense amount of data, unknown and inapplicable values, and incompatible data type pertaining to our analysis, major data cleaning was required to prepare the dataset for analysis. We allocated different columns to analyze among the four of us and as we put the results together in this report, we saw how different factors relate to each other.

Over 15 years, the overall number of collisions decreased, and we witnessed a seasonality pattern for collisions; number of collisions increased over time and reached its peak mid-year and decreased till it reached off-peak in the first quarter. When we dove in to analyse the trend over a span of week and hour basis, we saw that most number of collisions happened on Fridays and showed downturn on Sundays. Moreover, there were two peaks for the hourly analysis: one at ***hour*** 8*(8:00-8:59am)* and the other at ***hour*** 16 *(16:00-16:59pm)* reflecting rush hours in the morning and in the afternoon.

In terms of road configuration, while passing or climbing lane had the least number of non-fatal collisions, the fatal percentage was the highest at around 12%. We concluded that this could be due to the difference in speed between vehicles; the vehicle trying to join the passing lane has to speed up since other vehicles travel faster on the passing lane and this puts them in a vulnerable situation. One of the surprising findings was that the majority of collisions took place in clear and sunny days on dry and normal road surface. In fact, the most dangerous weather for drivers was weather with limited visibility: fog, smog, mist and weather with strong wind. When we observed the influence of traffic controls, whether or not traffic controls were present did not make a difference in number of collisions. We observed that the highest fatality rate occurred with railway signals with no gate.

It was interesting to see that for two cars travelling in the same direction the most common collision configuration was rear-end collision while for two cars traveling in different directions, right-angle collision was the most frequent: most likely at an intersection. This result supported our finding that the most frequent number of vehicles involved in collision over time was 2 vehicles. While rear-end collision was the most common, it obviously made sense that head-on collision had the highest fatality rate.

Among different vehicle types, we saw that light duty vehicles were most likely to be involved in collisions followed by light truck, motorcycle and bicycle. There were clear seasonal patterns observed for certain vehicle types: school buses showed less number of collisions during summer break while vehicles for outdoor activities (e.g. bicycle, motorcycle, off-road vehicle) showed an increase in number of collisions during warmer periods of the year.

We extended our analysis to person level elements and discovered some interesting insights. The number of men involved in collisions was greater than that of women, but women were more prone to injuries. We also observed that age groups involved work force were most likely to get into accidents during rush hours. We also found that people that were driving were most likely to be involved in both fatal and non-fatal accidents followed by people seated in the front row: right outboard. However, drivers actually had the lowest fatality rate whereas people sitting on someone’s lap had the highest fatality rate at 7%. This emphasized the importance of wearing a seat belt and using safety devices. Unfortunately, we discovered that even with safety device used, the number of fatal and non-fatal collisions was the highest. Yet, we can note that the fatality rate for collisions with safety device used is low while the fatality rate is the highest at 8% when no safety device was used; the importance of safety device still holds true. We can also note that for motorcyclists, bicyclists, snowmobilers and all-terrain vehicle riders, the number of collisions as well as fatality rate decreases when both helmet and reflective clothing were used: instead of either just helmet or reflective clothing on its own. This is also why pedestrians are more vulnerable and are likely to get injured in accidents since they are usually not equipped with safety devices like motorcyclists.

By working on this project, we were able to gain some insights about what factors contributed to more collisions and how severe the collisions were. Even after dropping the missing, unknown, and inapplicable values our dataset was adequate for a good analysis and show how each factors were related to each other. When analyzing the results, we were able to apply our common knowledge as drivers and pedestrians which also made it easier for us to understand the results.

# Appendices

## A-1 sample dataset

* Data dictionary:



* Sample data:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C\_YEAR** | **C\_MNTH** | **C\_WDAY** | **C\_HOUR** | **C\_SEV** | **C\_VEHS** | **C\_CONF** | **C\_RCFG** | **C\_WTHR** | **C\_RSUR** | **C\_RALN** |
| 1999 | 1 | 1 | 20 | 2 | 2 | 34 | UU | 1 | 5 | 3 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C\_TRAF** | **V\_ID** | **V\_TYPE** | **V\_YEAR** | **P\_ID** | **P\_SEX** | **P\_AGE** | **P\_PSN** | **P\_ISEV** | **P\_SAFE** | **P\_USER** |
| 3 | 1 | 6 | 1990 | 1 | M | 41 | 11 | 1 | UU | 1 |

## A-2 Sample Codes

**[A-2.1]** load the data into a pandas.DataFrame (in “G3\_Project\_Ryan.ipynb”).

|  |
| --- |
| accident = pd.read\_csv('NCDB\_1999\_to\_2014.csv', engine = 'python')  *#replace header to lower case*  accident.columns = accident.columns.map(lambda x: x[:].lower()) |

**[A-2.2]** clean and prepare the data for analysis (in “G3\_Project\_Ryan.ipynb”).

|  |
| --- |
| *#Check the non-numeric values in date-time columns*  accident[accident['c\_mnth'].str.contains('[^0-9]')| accident['c\_wday'].str.contains('[^0-9]')| accident['c\_hour'].str.contains('[^0-9]')]  *#Make a copy "df" for further analysis, avoid mess up the original data "accident".*  *#Remove all special values (unknown to us) in date-time columns, prepare for using time series as index.*  df = accident[:]  df[['c\_mnth','c\_wday','c\_hour']]=df[['c\_mnth','c\_wday','c\_hour']].\  replace('[^0-9]+',np.nan,regex=True)replace('[^0-9]+',np.nan,regex=True)  df.dropna(axis=0,subset=['c\_mnth','c\_wday','c\_hour'],inplace=True) |

**[A-2.3]** create new column ‘date’ with data type as ‘period[M]’ and set this column as an index (in “G3\_Project\_Boris.ipynb”).

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

**[A-2.4]** create ‘fatal’ and ‘non\_fatal’ columns based on c\_sev column value (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) |

**[A-2.5]** group ‘fatal’ and ‘non\_fatal’ columns by date (in “G3\_Project\_Boris.ipynb”).

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

**[A-2.6]** 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) |

**[A-2.7]** correlation plot and correlation coefficient determination (in “G3\_Project\_Boris.ipynb”).

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

**[A-2.8]** group collisions by week day and create a bar plot (in “G3\_Project\_Chaewon.ipynb”).

|  |
| --- |
| *#Total number of vehicles involved in collision in regards to weekday*  weekday=['Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun']  w\_day= df3.groupby('c\_wday')['c\_vehs'].size()  *#Visualize the result*  w\_day.plot(kind='barh', color='c',figsize=(10,6),rot =0, title='Total Number of vehicles involved in collision by Weekday'); |

**[A-2.9]** group collisions by day hour day and create a bar plot (in “G3\_Project\_Chaewon.ipynb”).

|  |
| --- |
| *#Total Number of vehicles involved in collision by hour*  hour=df4.groupby('c\_hour')['c\_vehs'].size()  hour.plot(kind='bar',title='Total Number of vehicles involved in collision by hour',color='G',rot=0) |

**[A-2.10]** aggregate by a pair of columns, and reform into a DataFrame (in “G3\_Project\_Ryan.ipynb”).

|  |
| --- |
| *#Aggregate the events by month into a summarized dataframe.*  df2 = df1.groupby(['date','c\_raln']).size().unstack() |

**[A-2.11]** bin the ages into 10 groups, and plot accordingly (in “G3\_Project\_Ryan.ipynb”).

|  |
| --- |
| bins = list(range(0,101,10)) |
| df8['age'] = pd.cut(df8.index,bins,labels=bins[:-1])  df9 = df8.groupby('age').sum() |
| df9.plot(kind='bar',rot=0, color=['g','orange','r'],figsize=(15,5),title='Person age & Medical treatment required')  plt.legend(['1.No Injury','2.Injury','3.Fatality'])  plt.xticks(range(0,11),['(0, 10]',' (10, 20]',' (20, 30]',' (30, 40]',' (40, 50]',' (50, 60]',' (60, 70]',' (70, 80]',' (80, 90]',' (90, 100]'])  plt.grid(axis='y'); |