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

[Project Overview 2](#_Toc511055699)

[Data Preparation 2](#_Toc511055700)

[What was the data source? 2](#_Toc511055701)

[How good was the data quality? 2](#_Toc511055702)

[How we procured source data? 3](#_Toc511055703)

[What tools/code we used to prepare it for analysis? 3](#_Toc511055704)

[What challenges we were facing? 3](#_Toc511055705)

[Analysis 5](#_Toc511055706)

[What trends, correlations and/or patterns we saw in the data? 5](#_Toc511055707)

[Individual analysis and the result we discovered: 5](#_Toc511055708)

[Conclusions 25](#_Toc511055709)

[What we learned from the data set? 25](#_Toc511055710)

[What …. 27](#_Toc511055711)

[Appendices 28](#_Toc511055712)

[A-1 sample dataset 28](#_Toc511055713)

[A-2 Sample Codes 28](#_Toc511055714)

# Project Overview

Our team analyzed the Canadian car accident data from 1999 to 2014 (obtained from Kaggle), and gained some insight from it.

Our team created an open repository on [GitHub](https://github.com/rynho/3250-G3), including all project background information, codes (in Jupyter Note Book format) and reports for easy collaboration and tracking changes between team members.

We provided some suggestions based on our analysis at the end, hope which could be helpful to the audience for avoiding high risk driving conditions in the future.

# Data Preparation

Project group utilized the data from Kaggle, an open platform hosting a lot of real world data sets for predictive modeling and analytics competition. After sense check and cleansing of the data, our team members manipulated the datasets for individual analysis respectively.

## What was the data source?

The data source is from an open dataset hosted on Kaggle: [Canadian Car Accidents 1999-2014](https://www.kaggle.com/tbsteal/canadian-car-accidents-19942014/data). While the [full dataset](https://www.kaggle.com/tbsteal/canadian-car-accidents-19942014/downloads/NCDB_1999_to_2014.csv/1) is provided in CSV file (including header), which includes about 6 million records (15 years’ car collisions registered in Canada from 1999 to 2014), a [pdf file](https://www.kaggle.com/tbsteal/canadian-car-accidents-19942014/downloads/drivingLegend.pdf/1) provided the data dictionary. Each line of the dataset provide one registered collision record of one person involved in one collision, and there are 22 columns in the datasets, including:

* Collision level data elements (***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***, and ***traffic control***);
* Vehicle level data elements (***vehicle sequence number***, ***vehicle type***, and ***vehicle model year***); and
* Person level data elements (***person sequence number***, ***person sex***, ***person age***, ***person position***, ***medical treatment required***, ***safety device used***, and ***road user class***).

## How good was the data quality?

The dataset is generally in good shape, including several missing numbers, and some unknown or not applicable data, which is well defined in the data dictionary. As an example, besides the 17 vehicle types defined in ***vehicle type*** (number in string), there are other four possible 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.

Also, the data provided ***day of week*** instead of date, which brought some difficult to the team when trying to put a time series index for analysis (detail in below sections).

Further, since every data column contains some non-numeric value, pandas treated all columns as strings. Team has to manipulate them when trying to utilize numeric only features, e.g., scatter plot (detail in below sections).

## How we procured source data?

Team downloaded the dataset CSV file from the URL mentioned above, use system command to perform some sense check, then used *pandas.read\_csv* method loading the full dataset into a *DataFrame*. See Appendix **[A-2.R1]**.

## What tools/code we used to prepare it for analysis?

Team was able to perform all the data cleansing and preparation with python. The general cleansing is including:

* Non-numeric check for date time related columns;
* Drop the *nan* values as they represent insignificant samples; and
* Composed a *date* column from ***year*** and ***month*** information for possible usage in analysis.

See **[A-2.R2]** for the sample code.

Note: Team comes up two different ways to generate the date column, and decides to use one or another in individual analysis to compare the difference. See below for detail.

## What challenges we were facing?

Team experienced following challenges in prepare the data, and overcame them after experiment and discussion:

* The dataset didn’t include a ***date*** but a ***day of week***. Team decide to combine ***year*** and ***month*** as the date information, but keep the ***day of week*** in its column for certain analysis.
* All values in dataset are in string type. Team decide to keep them in general dataset, and change the type when make a cloning slice, and proceed with further analysis accordingly.
* Considering some unknown values may represent significant samples, and/or bring meaningful insight, team decide to keep them in general dataset, and treat them differently in individual analysis instead. However, the unknown values in date time columns are dropped, as they represent tiny portion of the data, which is very insignificant and bringing difficulty for most of the analysis.
* Team experiments two different ways to engineering date column, and noted their advantage versus disadvantage. The *pandas.DatetimeIndex* method is faster, however the result includes a date of month value ‘1’, which is not accurately represent the collision record; the *pandas.PeriodIndex* with ‘M’ frequency accurately shows the year and month, but the method is about 10 times slower given the huge dataset our team was using. Team decides to use either way in individual analysis to compare the difference.

# Analysis

## What trends, correlations and/or patterns we saw in the data?

Team analyzed the chances of collisions in different conditions, as well as the relationships / correlations between those conditions. As some of the trend are expectable, the others are really surprised us, e.g., most collisions happened in good weather on straight road.

We addressed the most interesting results in following sections. Please refer to the attached Jupyter Notebooks for all detail, and the analysis not described here.

## Individual analysis and the result we discovered:

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

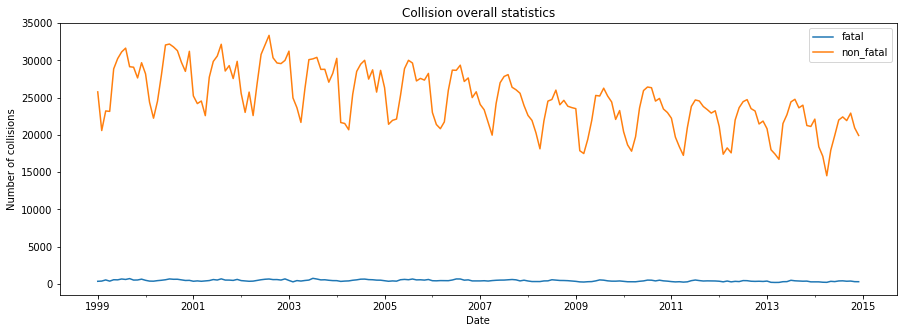
To assess collisions over the time, we created ‘date’ field using *PeriodIndex* function with year and month as arguments. Then we used the created ‘date’ column as an index for dataset (see sample code in **[A-2.B1]**).

* 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]**).

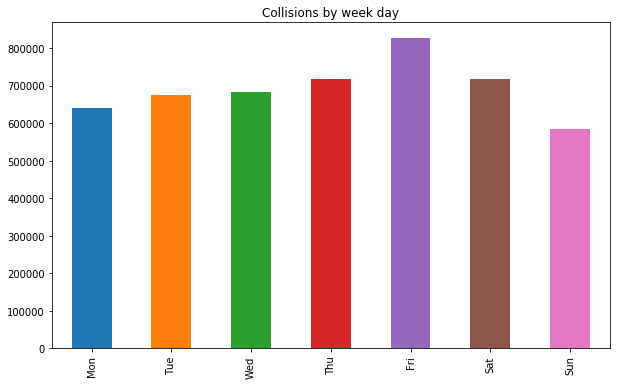


1. **Collision risk by week day and day hour**

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

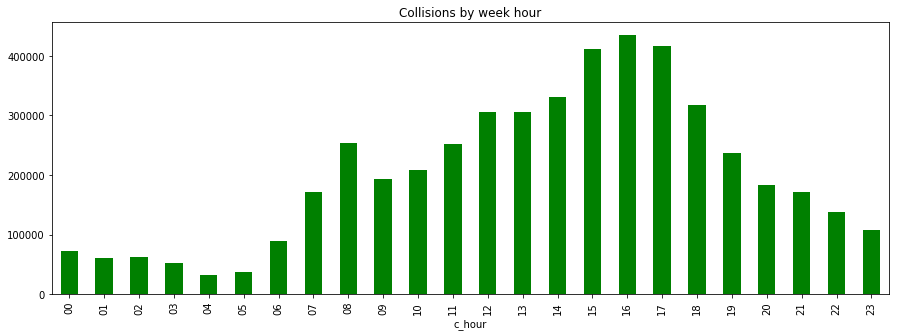
* 1. Collisions by week day

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 of 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 nor 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% and 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.



1. **Road Alignment**

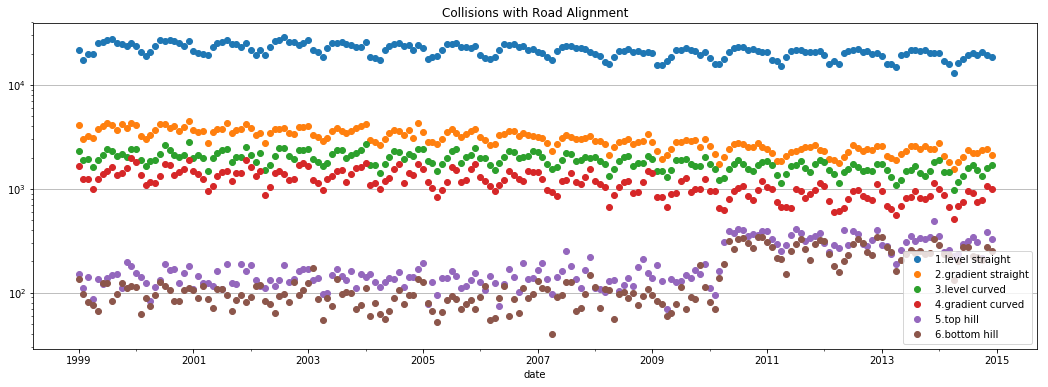
***Road alignment*** (*c\_raln*) is a column indicating road straight/curving, and level/gradient status for certain collision record. We analyzed different relation between it and other variables.

* 1. Road Alignment & Time relation

Before doing each analysis, we took a slice from general dataset to keep the integrity of original data, which ease the debugging, and avoid reload the huge dataset into memory again (see sample code in **[A-2.R3]**).

Because many rows of the dataset may represent collisions in same month, we sliced and grouped the data by time (*date* represent month) and road alignment column (*c\_raln*), then using *size* method aggregated the sum for each possible (*date, c\_raln*) pair into a two level indexed *pandas.Series*, then using *unstack* method to make it into a new *Dataframe* (sample code in **[A-2.R4]**).

The Dataframe is plotted as below (logarithm scale used for showing pattern of small numbers as well as big ones, and grid line for easy reading):

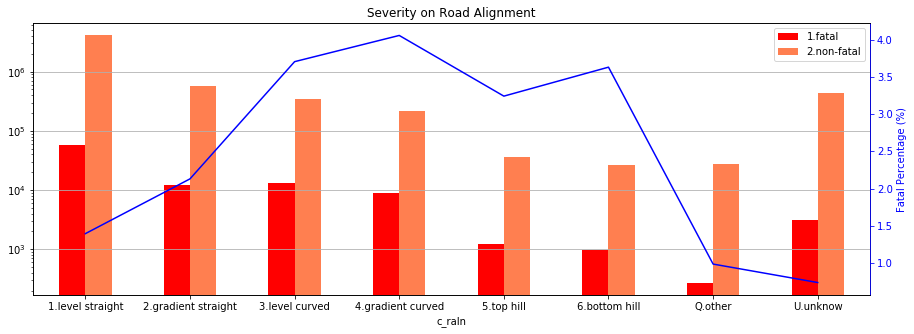


From it we saw most collisions happened on level and straight road, reason could be the vehicle of high speed, and driver not paying enough attention given the good road condition. Also noted clear seasonal pattern, see the auto-correlation analysis for seasonal patterns.

* 1. Road Alignment & Collision Severity

***Collision severity*** (*c\_sev*) is a column indicating a collision is fatal or not.

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



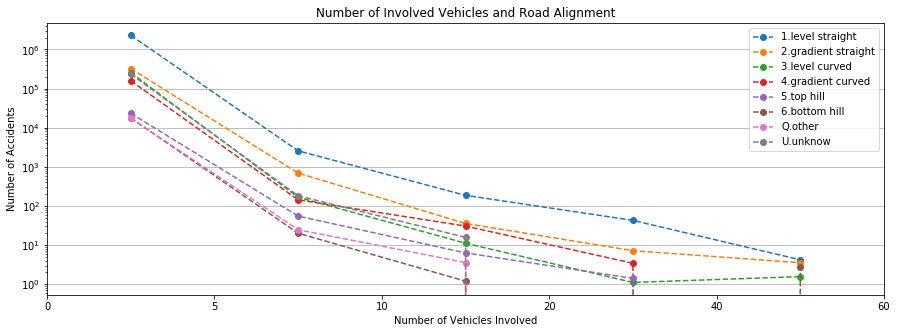
Fatal injury represent a very small portion of all accident (less than 4% in any given cases). It's also obvious that the level-straight road has less fatal injury, though the total number is much more than other cases; instead, the uneven condition (straight or curved) may produce severe accident. This could be caused by limited sight.

* 1. Road Alignment & Number of Vehicles Involved in Collision

***Number of vehicles involved in collision*** (*c\_vehs*) is a column indicating how many vehicles involved in certain records. In this analysis, all unknown value in this column are dropped as they didn’t provide insight.

To gain better understanding, we count the records for certain values in *c\_vehs*, and confirmed that every vehicle involved in one accident has its own record (e.g., there are 58 records in the dataset have 57 vehicles involved, the outstanding one could be a pedestrian), thus we calculated the portion of each accident, and put them in a new column in order to project the number of accidents that involved certain number of vehicles, by each road alignment (sample code in **[A-2.R5]**).

We also binned the *c\_vehs* value into several bins to show the trend instead of focusing on each certain number, and plot in a cleaner graph below (sample code in **[A-2.R5]**):

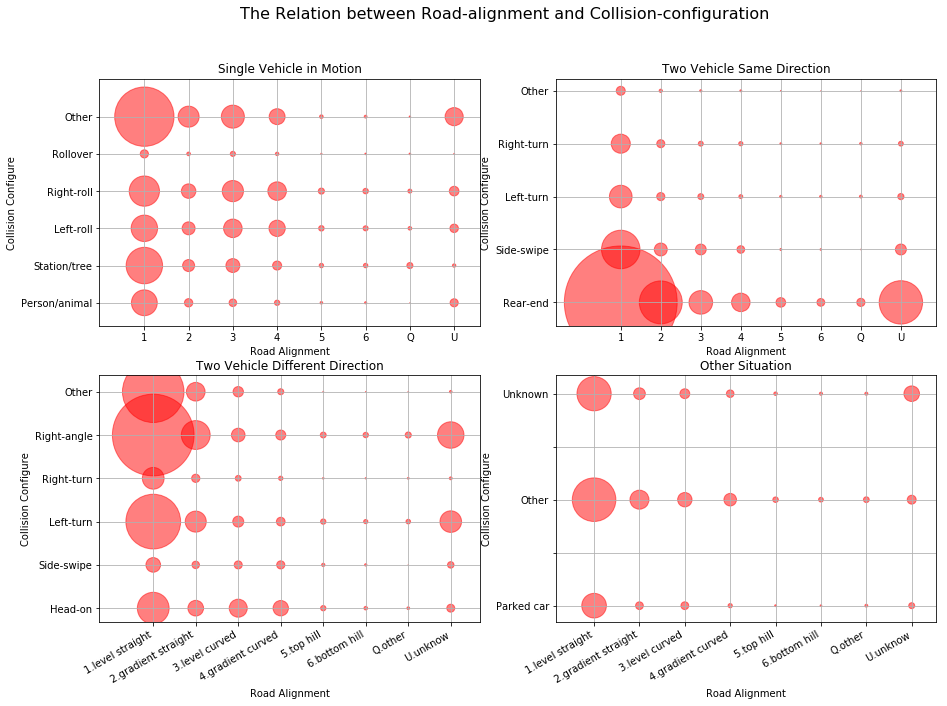


Majority accident are involving less than 5 cars. As the plots are very similar for different road alignments, no specific relationship observed between the two variables.

* 1. Road Alignment & Collision Configuration

***Collision configuration*** (*c\_conf*) column indicates one or two vehicle involved in collision, their motion, direction and relative position. As we want to plot them in scatter graph, which only supports numeric axis, all nonnumeric value of involved columns are replaced, and type is changed to integer.

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



The size of each dot indicate the collision numbers in each setting, same scale factor used for all subplots, which makes the number in each plot comparable to its neighbour. In all cases, clear trend shows that the collision chance increases when road-alignment getting better, due to higher speed and less caution. Except the unknown or "other" conditions, we can see in each cases:

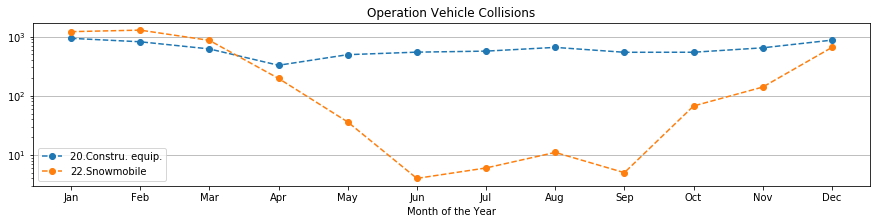
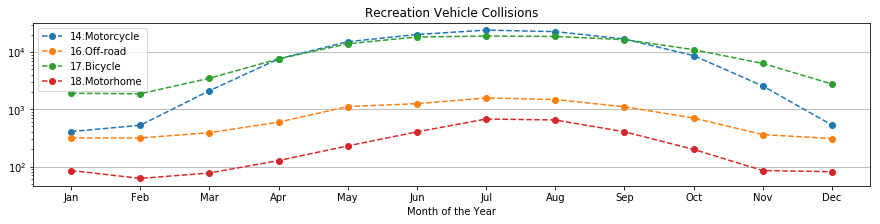
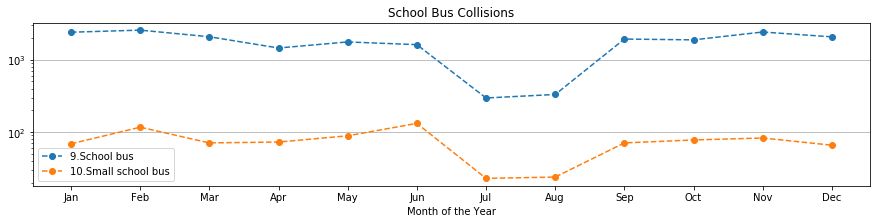
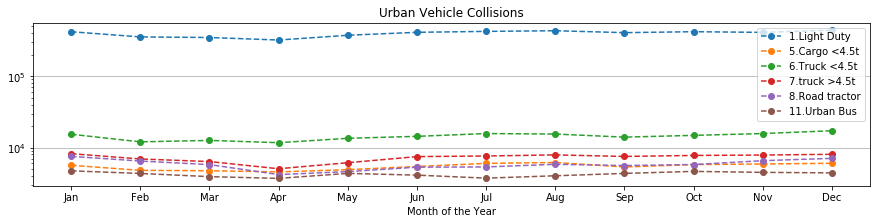
* Most collision for one car accident is hitting a static object, followed closely by right and left roll on to shoulder.
* For two car in same direction, most accidents are rear-end collision, which clearly increase along road-alignment.
* For two car in different direction, right-angle collision is huge, guess is in intersection. Left turn also post a lot risk, but right-turn is very low, even lower than head-on collision. This may be a result of the inherent complexity of left turn.

1. **Vehicle Type**

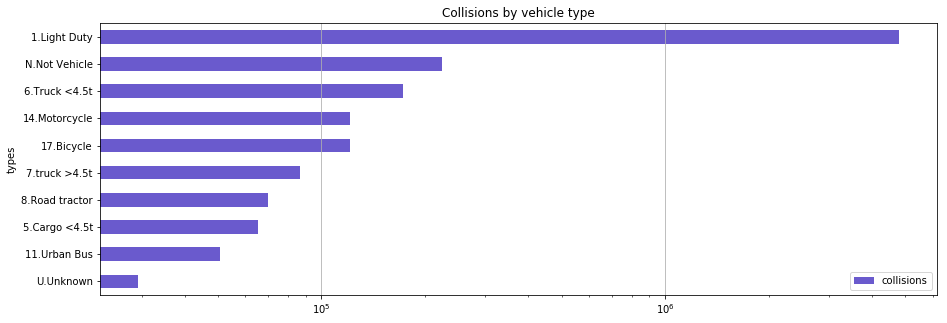
***Vehicle type*** (*v\_type*) indicates the vehicle type of the collision record, e.g., truck, school bus, bicycle, etc. Following analysis is focused on this column.

* 1. Vehicle Type & Time Relation

Because there are 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 vehicle types. Below are several insightful ones among them.



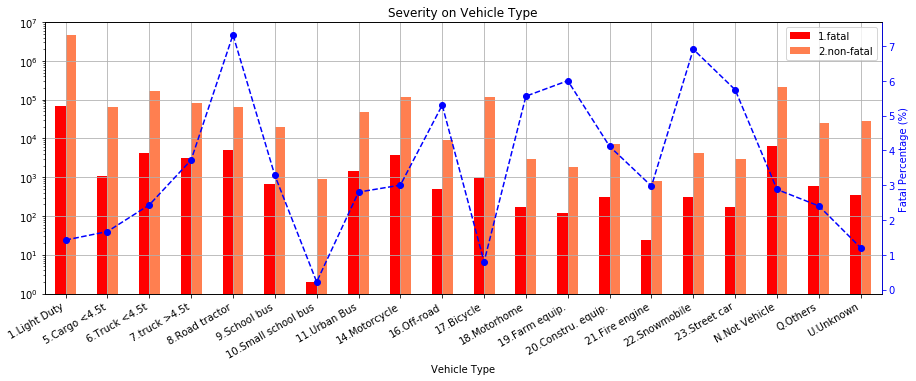
We also plotted an overall summarize showing the vehicle type involved in most collisions, see below (only top 10 types, logarithm scale for x-axis).



By looking at the plots above, clear pattern shown in different vehicle types:

* Vast majority collisions are caused by light duty vehicles, followed by light truck, motorcycle and bicycle. This is due to the dominant population of the personal vehicles, and the unprofessional nature of the drivers.
* The collisions distribution for urban vehicles (light duty, trucks, road tractor, and bus) is flat across all seasons, without showing obvious pattern, indicating urban traffic being busy for all time.
* School bus involved collisions happened much less in July and August, due to summer vacation.
* More accident related to bicycle, motorcycle, off-road vehicle and motorhome happened from April to October, as they are used for outdoor activities during summer time; in opposite, snow-mobile collisions are more from October to April.
  1. Vehicle Type & Severity

Similar as the analysis for ***road alignment*** and ***severity***, we plotted the collision number by severity in a bar chart, with the corresponding fatality percentage on another y-axis on right. To make the multiple columned graph easier for reading, we plot the percentage line with marker (indicate value) and dashed connection.



From the figure above we see most fatal collisions are caused by light-duty vehicles, followed by motorcycle, road tractor, and trucks. However, the highest percentage of fatal accidents is caused by road tractor and Snowmobile (about 7%), and followed by street car, farm equipment and motorhome (around 6%). Also noticed that small school bus produces least fatal collisions, and bicycle number is also non-surprisingly low.

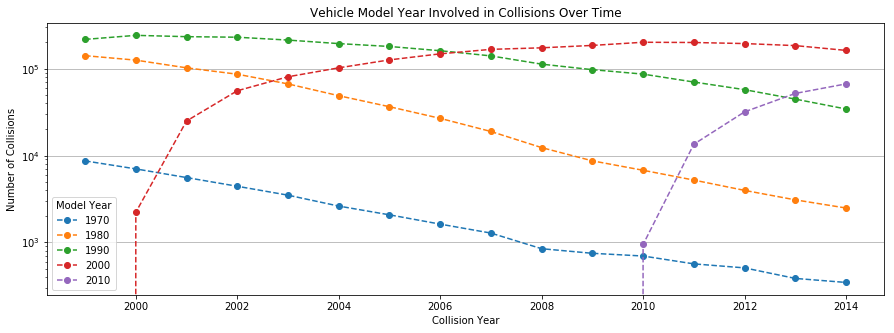
1. **Vehicle Model Year**

***Vehicle model year*** (*v\_year*) indicates the year of the vehicle model registered for each collision.

* 1. Vehicle Model Year & Time Relation

By sense checking the data, team realized that all values other than years in *v\_year* would not bring insight to this analysis, thus we dropped all of them. We also changed all values of ***vehicle model year*** (*v\_year*) and collision ***year*** (*c\_year*) into integer for easy binning.

After cut the *v\_year* into 12 bins (every 10 year from 1900 to 2020), we plotted the collisions of top 5 involved model years across the available collision years (1999 to 2014).



As the data covers collisions from 1999 to 2014, most cars involved in collisions are made in recent thirty years. Clear trend observed that the model getting involved in collision peaks in its 10 years' age, which could be a result of the market share growth and the reliability drops for old vehicles.

Apart from some external environment factors, individual’s characteristics also are a significant factor to influence the likelihood of a crash. Therefore, the following analysis will pay more attention to some person-level data elements, separately focusing on person sex, age, and road user class.

1. **Person sex**

***Person sex*** (*p\_sex*) is a column shows the sex of the individual who is involved in a traffic accident. For convenience, all unknown and missing information variables (which include not applicable elements and those are not provided by jurisdiction) will be dropped.

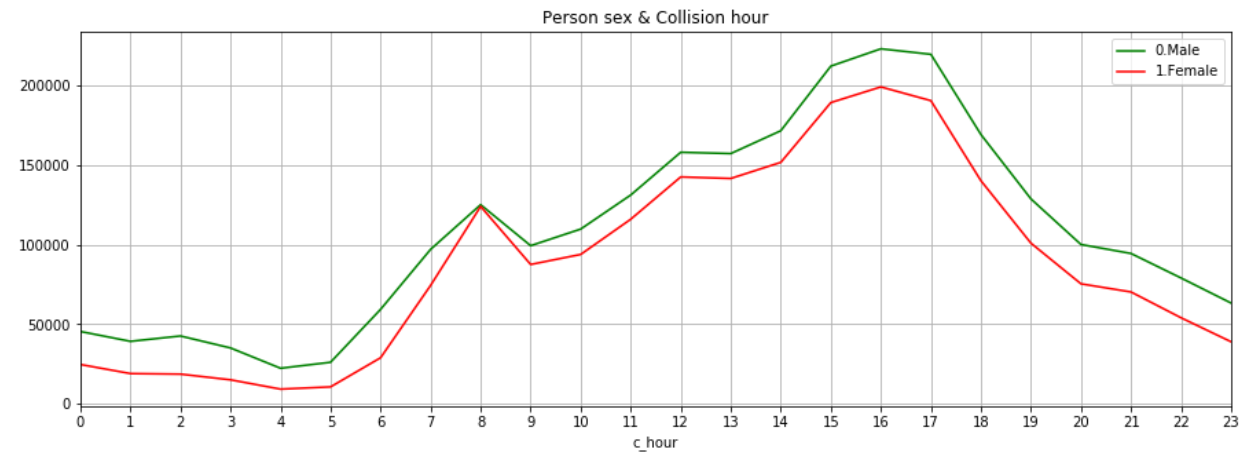
In this section, we will analysis three factors to person sex: collision hour; medical treatment required; collision configuration.

* 1. Person Sex & Collision Hour

***Collision Hour*** (*c\_hour*) is a column simply shows when the collision happened. Number from 0 to 23 correspond to 24 hours. For example, 0 indicates form midnight to 0:59; 14 represents from 14:00 to 14:59.

In the whole sample, we grouped the data by sex (*p\_sex*), and then using *count* method to account the number for each sex. So, data includes 2, 608, 529 men and 2,116, 502 women. Therefore, we might expect that man will more likely to involve in a collision.

In order to analyze the relation with collision hour, we need slice and group the data by sex (*p\_sex*) and collision hour column (*c\_hour*). Then using *size* and *unstack* method to aggregate the result. In the end, we plot a line chart with grid line for data visualization purpose:

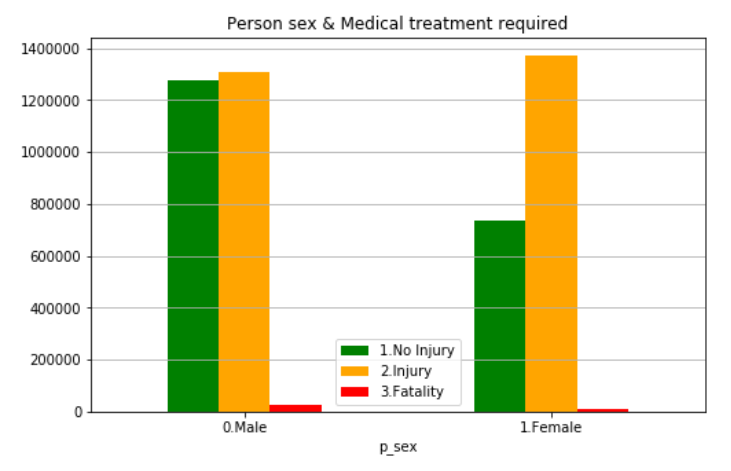


Our expectation is confirmed by the line chart, men are more likely to involve in a collision accident at all times. However, the collision times of male and female are very close to each other at 8 am, around 125047 and 123961 times. Also, for both male and female, the riskiest time period is form 3 pm to 5 pm, and then the risk has a monotonically decreasing trend.

* 1. Person Sex & Medical Treatment Required

***Medical Treatment Required*** (*p\_isev*) is a column indicating the degree of casualty. After dropping all unknown and missing information elements, dataset indicated three levels: *No injury*, *Injury* and *Fatality*.

By using the same method as before, we can easily aggregate the result into a two-level index table (*p\_sex* and *p\_isev*), and then plotting a bar chart.



If we focus on No injure and Injure now, we can conclude that female is more likely to injure in an accident, compared to male. In another word, women are more likely to have a serious accident. Plotted in this scale, it is hard to get information about fatality. Therefore, we will plot these same data on a log scale:



By comparing the red bars, we can see that men have greater likelihood to die in an accident. Numerically, men are twice as likely to die as woman.

* 1. Person Sex & Collision Configuration

***Collision Configuration*** (*c\_conf*) is a variable to describe the type of a traffic accident. Dataset includes 4 major categories, which are single vehicle in motion, two vehicles in motion with the same direction, two vehicles in motion with the different direction, and one vehicle hits a parked motor vehicle. Also, the specific type of collision is indicated within each major category. Therefore, more details of each of them will be provided as following.

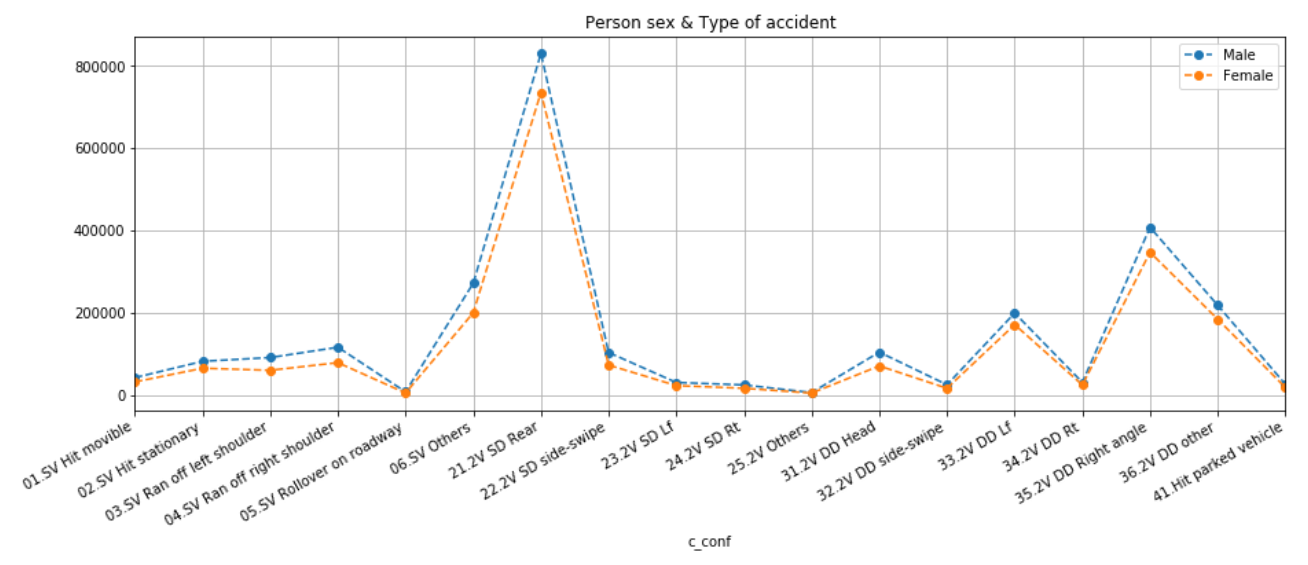
There are six types of collision when only one vehicle is involved. The first type is hitting a moving object, such as an animal. The second type is hitting a stationary object, like a tree. The third one is vehicle ran off left shoulder, including rollover in the left ditch. Corresponding with the previous one, the fourth one is it ran off right shoulder, which include rollover in the right ditch. Rollover on roadway is the fifth type, and the last one is others.

When two vehicles are moving in the same direction, five types of collision might occur. The first one is rear-end collision. The second one is side swipe, which means the sides of vehicles hit one another. The third type is one vehicle passing to the left of the other, or left turn conflict. The fourth one is exactly the opposite of the previous one. The last one is other types.

Six kinds of crashes might happen when two vehicles are moving in the different direction. The first one is head-on collision. The second type is approaching side-swipe. The third and fourth types are left turn or right turn across opposing traffic. The fifth one is right-angle collision. The last one is other types.

The final category only includes one type which two vehicles are involved in, but one hits a parked motor vehicle.

Now, the similar method will be used to analyze the relation between sex and collision configuration, which is shown as below.



It is obvious that male always has higher collision risk than female. Basically, male and female have a similar pattern among all kinds of collision configurations. Which means accident types do not have a significant sexual differentiation. Rear-end collision is the most frequent type of the traffic accidents for both man and woman, and Right-angle side collision is ranked as the second. Additionally, Rollover on roadway is the least common type of accident.

1. **Person age**

***Person age*** (*p\_age*) is numerical column, the age is from 0 to 99 years old. It is noteworthy that 0 means less then 1 years old and 99 represents 99 years or older. All other numbers, between 1 and 98, correspond to 1 to 98 years old. In doer to focus on generation differentiation instead of each age, we binned person age into 10 bins. For example, 10 represents the age from 11 to 20; 20 represents the age from 21 to 30 and so on. For analysis purpose, we still choose to drop all unknown and missing information data.

* 1. Person Age & Collision Hour

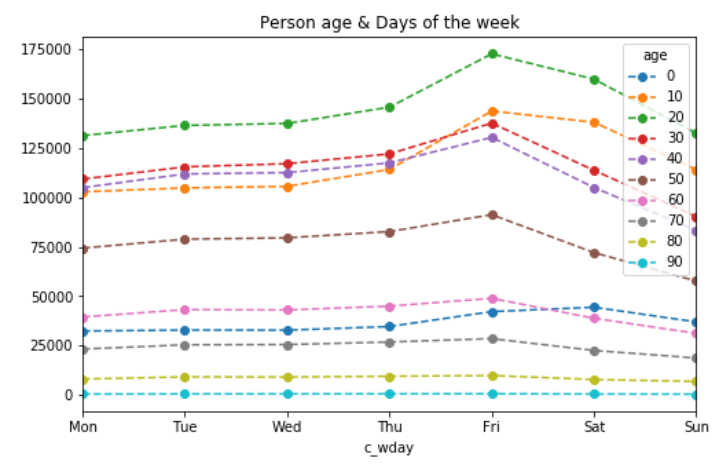
Like in “Person Sex & Collision Hour” analyze, we defined x-axis as collision hour in one day. Also, 10 different-color lines represent 10 age-groups.



The graph shows that the most dangerous population is from 21 to 30 years old. Also, the youth whose age is from 11 to 20 are more likely to involve in a collision at night, especially from 9 pm to 1 am. At morning (8 am to 11 am), three age-groups (21 to 30; 31 to 40; 41 to 50) has the similar number of accident. In the end, graph shows the working people share the same pattern, which is the rush hour are more likely to collide.

* 1. Person Age & Day of week

***Day of week*** (*c\_wday*) is a variable, which contains a series of numbers from 1 to 7. Therefore, 1 to 7 means Monday to Sunday.

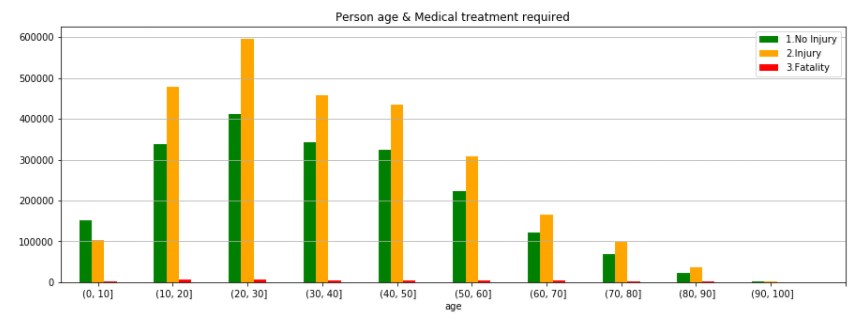


The plot shows that during the whole week, young adults (from 21 to 31) are the riskiest population. Risk decrease with age increases, for example the age from 41 to 50 are less likely to involve in an accident, and compared with whose age is from 31 to 40. Also, in-accident probabilities of the youth are significantly increasing over the weekend. Which is realistic, because the most of teenagers or young drivers are still in school.

In addition, the most of age-groups, like the teenager, young adult and adult, point out most collisions happened on Friday. Also, teenagers and young adults involved collisions happened much less on Monday. In contrast, adults are less likely being involved in an accident on Sunday.

* 1. Person Age & Medical Treatment Required

Considering the normal scale first.



Except for whose age below 10, other age groups get the similar conclusion that people are more likely to injure in collision, and the likelihood decreases with age. Same as before, the graph shows the youth are the most dangerous group. The injuries of young adults are around twice of those of whose age between 51 and 60.

Now, changing to a log scale with the same data.

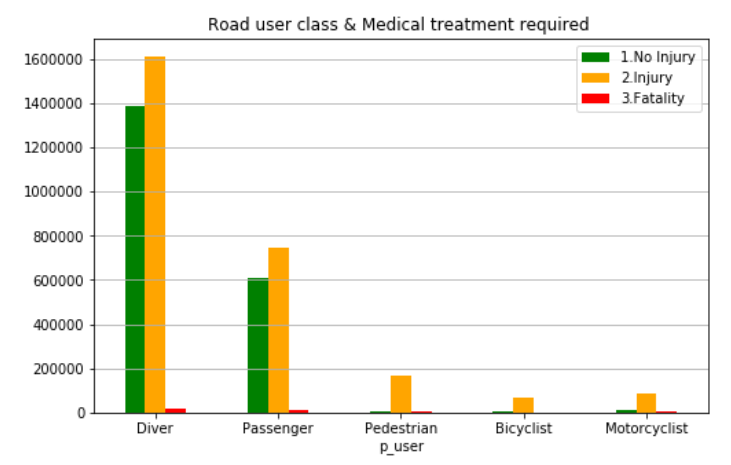


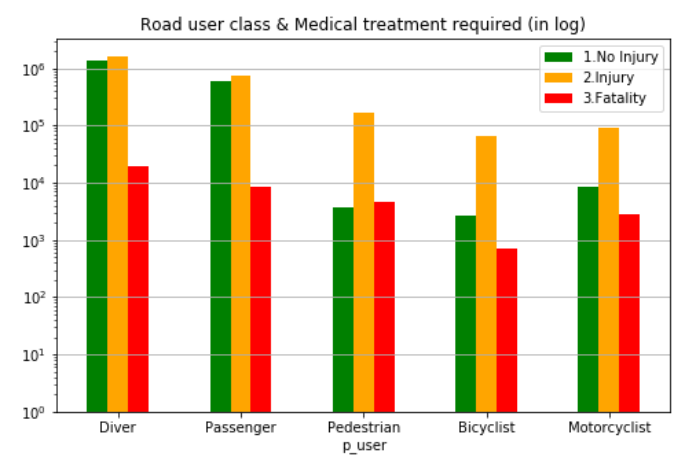
In contrast with the huge gap in injuries, young adults’ deaths are not outstanding, although it is the maximum. Additionally, people who above 90 are less likely to die in traffic casualty then all other age groups. Numerically, from 1999 to 2014, only 185 old people (above 90) died in an accident, compared with 7064 young adults died.

1. **Road user class**

***Road user class*** (*p\_user*) shows when accident happened who is involved. Dataset provides 5 road users, with codes from 1 to 6. Thus, 1 is motor vehicle driver; 2 is motor vehicle passenger, 3 is pedestrian; 4 is Bicyclist; 5 is Motorcyclist; and the last one includes others and unknown. For consistency, the last elements will be dropped.

* 1. Road User & Medical Treatment Required

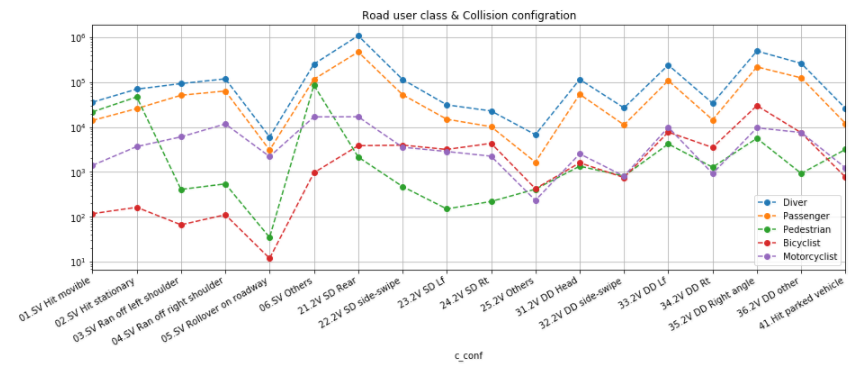




The major class involved in an accident is drivers, they are twice likely to get hurt or die than passengers. In particular, for each three degrees of casualty, the cases of driver are more than the sum of the rest of road users. Also, among all 5 categories, bicyclists are the safest, because of the lowest injuries and deaths.

Additionally, as far as drivers and passengers is concerned, not injury and injury separately account for half of total cases since fatality only accounts for about 0.06% in both road users. Whereas, a collision has above 85% possibility to cause injury when either pedestrians, bicyclists or motorcyclists are involved in.

* 1. Road User & Collision Configuration



In the end, it is no surprise that diver and passenger have the identical pattern, both of them are more likely to involve in a rear-end collision, and the second common type is right-angle collision. Single vehicle rollover on roadway causes pedestrian or bicyclists to injure or die is infrequent. However, pedestrians are more likely to involve in any other single vehicle collision configuration; for bicyclists, they are more likely to get hurt in a right-angle collision. Among all types of accidents, Motorcyclists get a middling ranking.

# Conclusions

## What we learned from the 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% and 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.

1. **Road Alignment and Accident**

From the analysis we performed above, following observations:

* Most collisions happened in level and aligned road. Drivers should pay more attention when drive on good road condition, and follow speed limit.
* However, bad road alignment may post higher risk of fatal accident. Drivers need to anticipate possible outcome when the view is not so wide, and always be cautious to avoid serious injury as possible.
* Rear-end collision is the most popular collision type. Keep enough space to the car in front of you, especially on well aligned road.
* Right-angle collision and left-turn collision are also popular, wait one more second before turning your wheel, or rushing into the intersection.

1. **Vehicle Type and Accident**

Clear seasonal patterns observed for certain types:

* School bus involved in collisions out of summer vacation (July and August);
* Vehicles for out-door activities (i.e., bicycle, motorcycle, off-road vehicle and motorhome) involved in collisions mostly during summer time (April to October);
* In opposite, snow-mobile collisions are more from October to April.

While light duty vehicles, light trucks, motorcycles and bicycles involved in most of the collisions, most fatal collision ratio observed from following types:

* Road tractor and Snowmobile is around 7%;
* Street car, farm equipment and motorhome is about 6%.

If you see those vehicle when you drive, try stay away from them.

1. **Vehicle Model Year and Accident**

As the collision rate peaks in a vehicle model’s 10th year, maybe you should consider a new car when it get into that age to reduce the risk of an accident.

1. **Person Sex and Accident**

Based on previous analysis, some observations can be concluded as follow:

* Men are more likely to involve in a collision accident than women.
* Female is more likely to injure in an accident.
* However, men are more likely being involved in a fatal accident. Numerically, men are twice as likely to die as woman.
* Accident configurations and collision hour do not have a significant sexual differentiation.

1. **Person Age and Accident**

* The most dangerous population is from 21 to 30 years old, called young adults, and the in-accident probabilities decreases with age increases.
* The youth whose age is from 11 to 20 are more likely to involve in a collision at night, especially from 9 pm to 1 am.
* People who are in working age share the same tendency, which is the rush hour are more likely to collide.
* As far as the teenager, young adult and adult are concerned, most collisions happened on Friday.
* Also, in-accident probabilities of the youth are significantly increasing over the weekend. In contrast, adults are less likely being involved in an accident on Sunday.
* The injuries of young adults are around twice of those of whose age between 51 and 60. However, all cases show that fatality of young adults are not outstanding. Which means young adults are more likely being involved in a not fatal crash.

1. **Road User Class and Accident**

* The major road user class involved in an accident is drivers, they are twice likely to get hurt or die than passengers.
* Bicyclists are the safest, because of the lowest injuries and deaths.
* A collision has above 85% possibility to cause injury when either pedestrians, bicyclists or motorcyclists are involved in.
* Drivers and passengers are more likely to involve in a rear-end collision. Pedestrians are more likely to involve in other types of single vehicle collision. For bicyclists, they are more likely to get hurt in a right-angle collision.
* Also, single vehicle rollover on roadway causes pedestrian or bicyclists to injure or die is infrequent.

## What ….

Sample sample sample sample sample

* List
* List
* List

Sample sample sample sample sample

# 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

1. **[A-2.B1]** 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') |

1. **[A-2.B2]** 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) |

1. **[A-2.B3]** group ‘fatal’ and ‘non\_fatal’ columns by date (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 (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 (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 (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') |

1. **[A-2.R1]** 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()) |

1. **[A-2.R2]** 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) |

1. **[A-2.R3]** make a copy of the dataset for easy debugging and faster processing (in “G3\_Project\_Ryan.ipynb”).

|  |
| --- |
| *#Slice the columns I would analyse from original dataset.*  dfa = df.loc[:,['date','c\_year','c\_mnth','c\_wday','c\_hour',  'c\_sev','c\_vehs','c\_conf','c\_wthr','c\_raln',  'v\_type','v\_year']]  *# Make a slice only focusing on c\_raln: road alignment.*  df1 = dfa.loc[:,['date','c\_raln']] |

1. **[A-2.R4]** aggregate by a pair, 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() |

1. **[A-2.R5]** calculate accident portion for each collision, and bin the numbers of vehicle involved in each collision (in “G3\_Project\_Ryan.ipynb”).

|  |
| --- |
| *#Calculate accident portion of each event, considering 58 records that have 57 vehicle*  *# involved, that should be just one accident, every relevant record should be only about 1/57.*  df1['acc']= 1/df1.c\_vehs  *#Categorize number of involved vehicles into bins.*  bins = [0,5,10,20,40,60]  df1['cats'] = pd.cut(df1.c\_vehs,bins) |