Situation(s) when most accidents occur, using the [National Collision Database](https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a#wb-auto-6) dataset.

Our hypothesis is that they mostly occur when these conditions are met:  
- Person/driver: Young male.  
- Vehicle: Light duty vehicle (passenger car/van, light duty pick up trucks, etc.)  
- Collision condition: intersection, two vehicles in motion and same direction.  
- Datetime / Weather: Spring, wet, morning

Now let’s see how much we are right.

# Data

The database has data for both drivers and other occupants. We just focus on the data related to drivers.

Number of records (1999-2017): 6,772,563

Number of records for drivers: 4,544,989

|  |  |
| --- | --- |
| **Data element** | **Definition** |
| C\_YEAR | Year |
| C\_MNTH | Month |
| C\_WDAY | Day of week |
| C\_HOUR | Collision hour |
| C\_SEV | Collision severity |
| C\_VEHS | Number of vehicles involved in collision |
| C\_CONF | Collision configuration |
| C\_RCFG | Roadway configuration |
| C\_WTHR | Weather condition |
| C\_RSUR | Road surface |
| C\_RALN | Road alignment |
| C\_TRAF | Traffic control |
| V\_ID | Vehicle sequence number |
| V\_TYPE | Vehicle type |
| V\_YEAR | Vehicle model year |
| P\_ID | Person sequence number |
| P\_SEX | Person sex |
| P\_AGE | Person age |
| P\_PSN | Person position |
| P\_ISEV | Medical treatment required |
| P\_SAFE | Safety device used |
| P\_USER | Road user class |

The columns highlighted in gray are eliminated from this study.

# Dealing with Null and other unknown values:

Number of Null values just from driver’s data:

C\_YEAR 0

C\_MNTH 230 -- assigned random numbers

C\_WDAY 922

C\_HOUR 44962

C\_SEV 0

C\_VEHS 356

C\_CONF 150214

C\_RCFG 367781

C\_WTHR 64391

C\_RSUR 57665

C\_RALN 307462

C\_TRAF 179549

V\_ID 80

V\_TYPE 30270

V\_YEAR 292128

P\_ID 12

P\_SEX 179797

P\_AGE 226420

P\_PSN 0

P\_ISEV 32145

P\_SAFE 500496

P\_USER 74881

C\_CASE 0

Working with a subset of columns and removing rows with null and unknown values still we have a decent number of records for each year:

C\_YEAR

1999 204014

2000 207504

2001 200613

2002 205959

2003 202081

2004 194809

2005 194022

2006 186969

2007 183407

2008 168417

2009 164705

2010 162130

2011 155379

2012 162355

2013 161195

2014 152356

2015 161623

2016 159699

2017 154607

# Observations by looking at each element separately

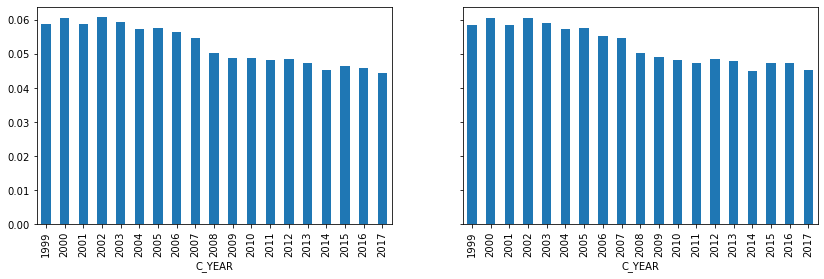
By just looking at each element separately we have the following conclusions:

Accidents mostly occured when these conditions are met:

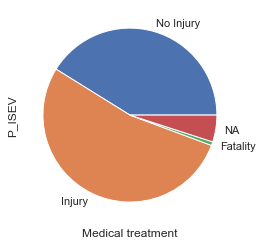
* Young (16-36) male
* Light Duty Vehicle
* In June-Oct, between 15:00 - 17:00 on Fridays
* Weather clear and sunny. Road surface Dry and Normal

They are Rear-end collision and caused more injury that fatality.

**Accidents are generally declined over the years (plot 01)**

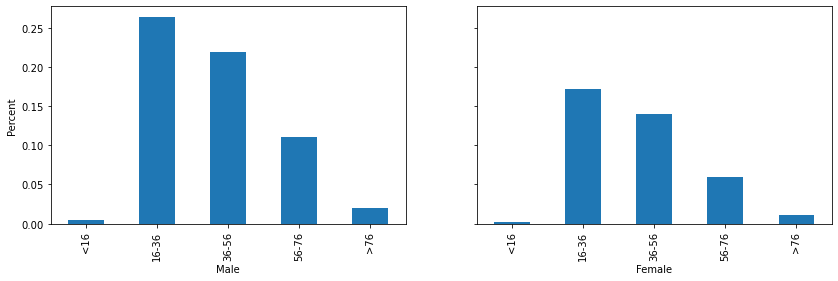


**Accidents mostly caused injury (plot 08)**

****

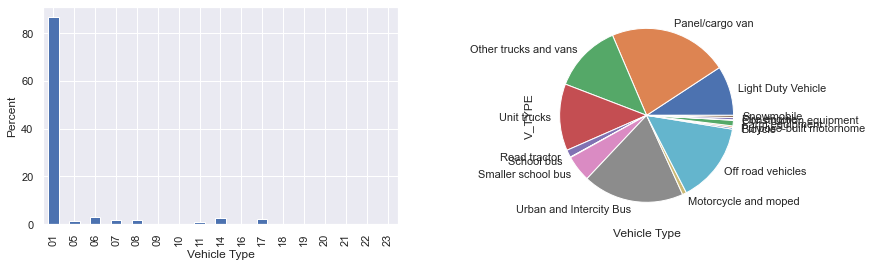
**More Accidents are between**

1. **Young male (Plot 02)**

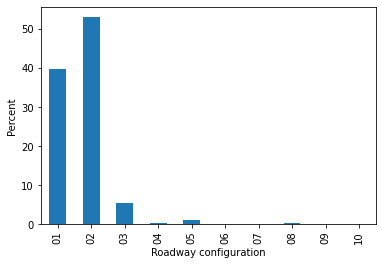


### **Light Duty Vehicle - V\_Type = 01 (Plot 03)**

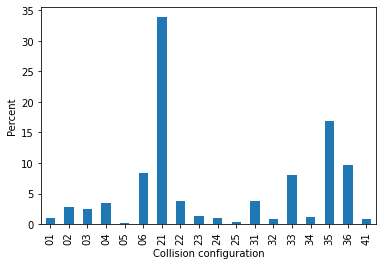
If we remove “Light Duty Vehicles” as an outliner (temporarily) then we have trucks and vans which are not cargo (06) and after that we have motorcycles (14) and then bicycles (17)



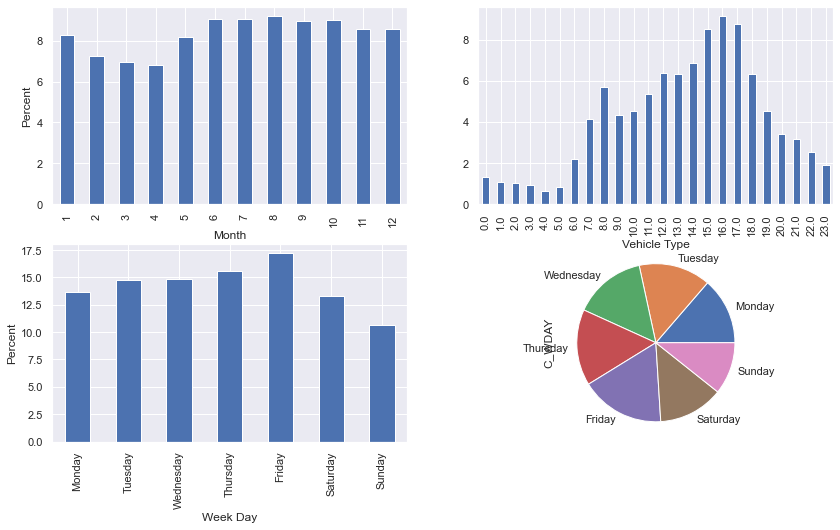
# **Roadway configuration: At an intersection of at least two public roadways C\_RCFG = 02 (Plot 04)**



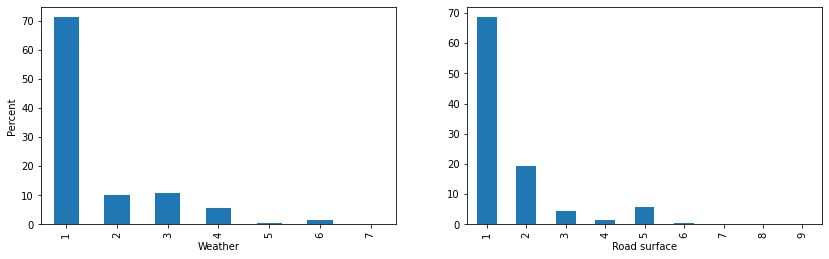
# **Rear-end collision - C\_CONF = 21 (Plot 05)**



## **June-Oct! Good weather?! Between 15:00 - 17:00 (Plot 06)**



## **Weather clear and sunny. Road surface Dry and Normal!! (Plot 07)**

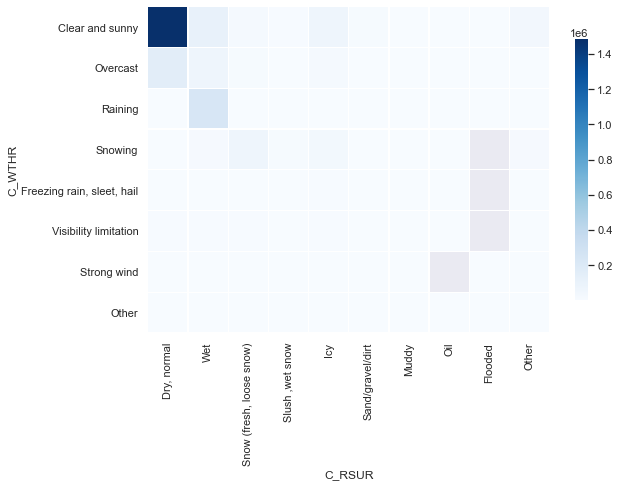


# Correlations

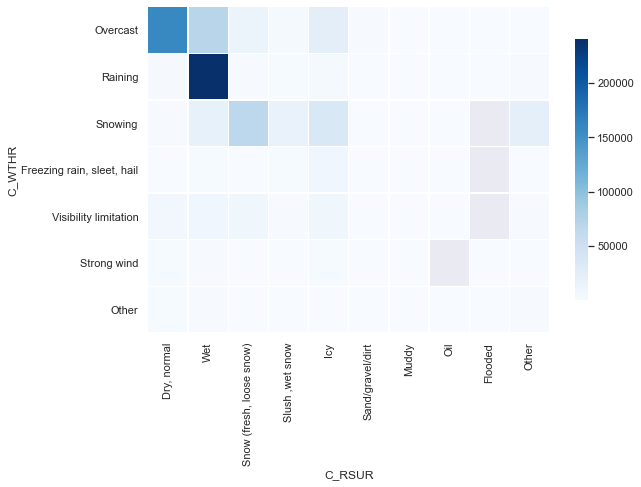
Looking at correlations positively supports our observation from individual factors:

1. Correlation between weather and road surface (Plot 10)

Clear and Normal weather + Dry, normal surface have the most collisions, after removing Clear and Dry weather other relations are clearer.

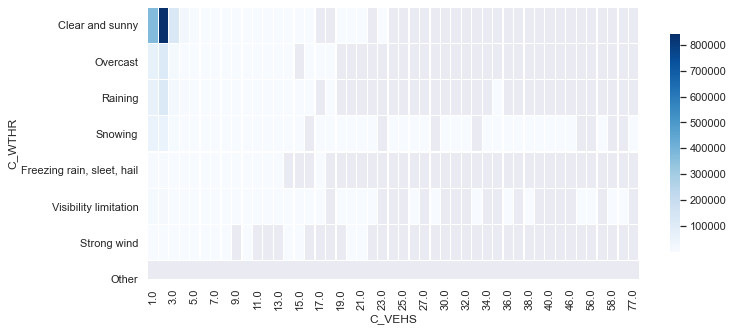


After removing Clear and Sunny



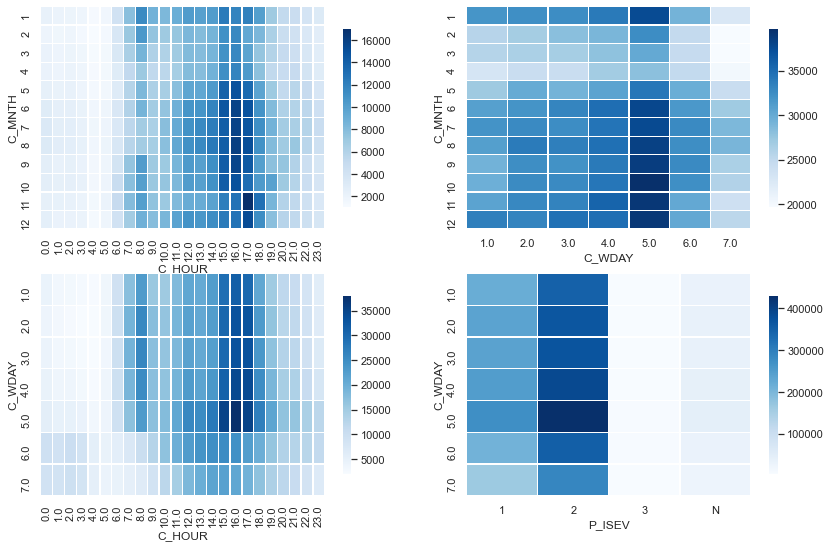
1. Weather and number of Vehicles (Plot 11)

Here again there was a big outliner which makes the rest of data insignificant. In this study we just focus on the outliner since our hypothesis is about finding the situations when most accidents happen.



1. Month and Hours, Month and Week days[,](http://localhost:8888/notebooks/Desktop/DataScience/1/Assignments/Group/grp17assignment.ipynb#Month-and-Week-days) Week days and Hours and Week days and Severity

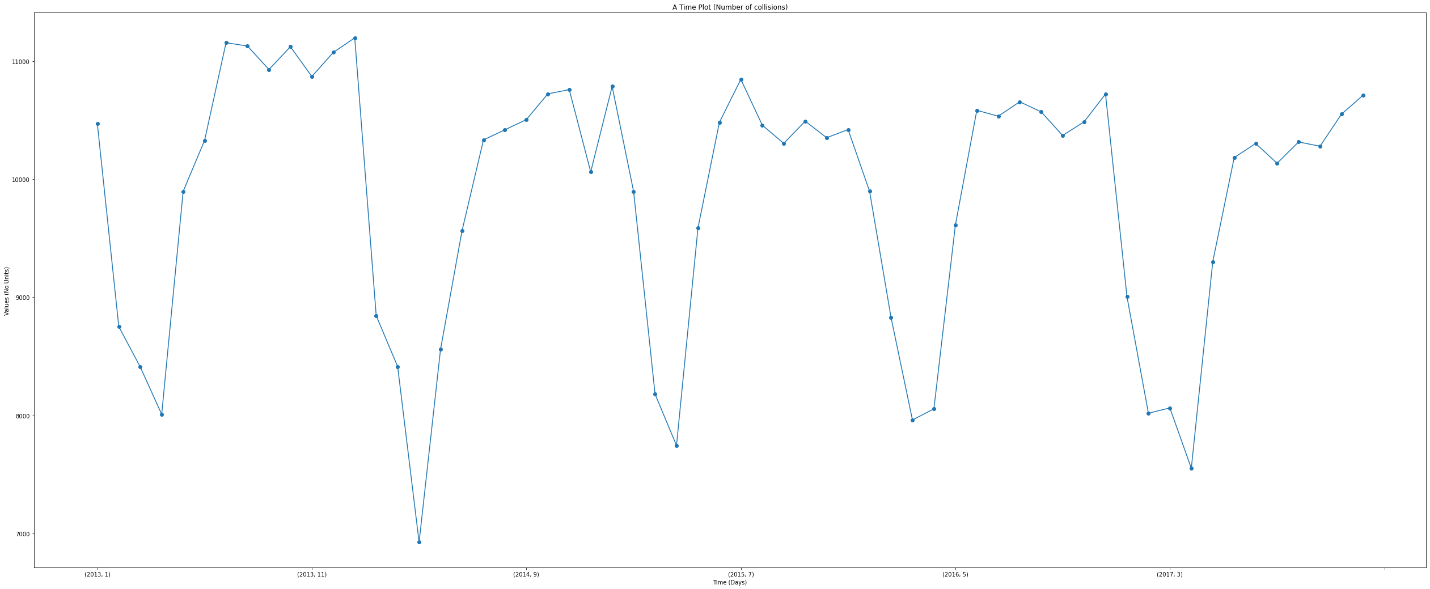
We see that correlations are supporting the single elements observations.



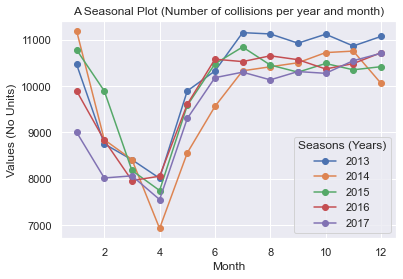
# Time Series

We observe seasonality in years.

Plot 20



Plot 21



To look at the data as time series we needed to have a datetime index. A new datetime column in engineered from C\_YEAR, C\_MNTH, C\_HOUR and since we don’t have the day of the month, the week day is used.

ct.C\_WDAY = ct.C\_WDAY.astype(np.int64)

ct=ct.rename(columns={"C\_YEAR": "year", "C\_MNTH": "month", "C\_WDAY": "day", "C\_HOUR": "hour"})

ct['dtindex'] = pd.to\_datetime(ct[['year', 'month', 'day', 'hour']])

ct=ct.set\_index('dtindex')

ct.drop(['year', 'month', 'day', 'hour'],axis=1,inplace=True)

Now are data looks like this, which C\_CASE is the total number of accidents for each datetime.

C\_CASE

dtindex

1999-01-01 00:00:00 10

1999-01-01 01:00:00 9

1999-01-01 02:00:00 8

1999-01-01 03:00:00 5

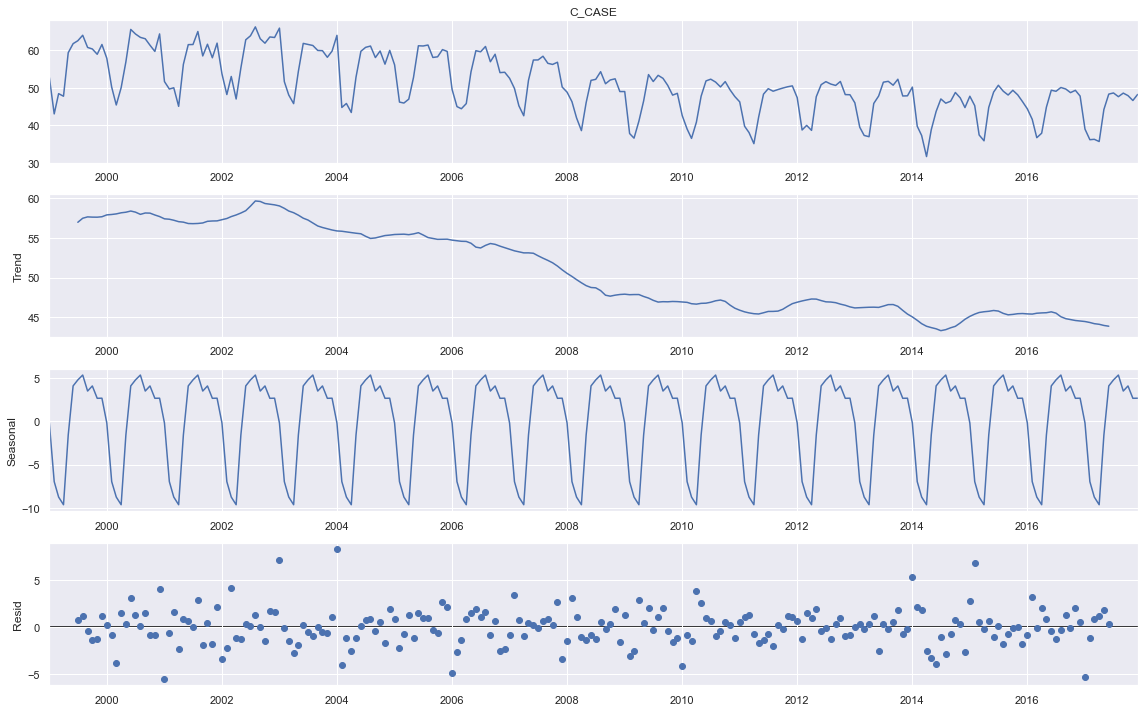
1999-01-01 04:00:00 6

Then we resample the data to get average number of accidents per month

y = ct['C\_CASE'].resample('MS').mean()

The we use decompose from statsmodels to separate trend, seasonality and noise.

Plot 22



Base on seasonality and Trend we can see the same pattern will repeat in the future year as previous years and the total accidents will go down.

# Forecasting with machine learning

To test what we learnt in machine learning so far, the following columns were picked:

Month, Hour, Weather and count of accidents. For counts we created a boxed column to make it categorical:

clsb2 = clsb1.groupby(['C\_MNTH', 'C\_HOUR', 'C\_WTHR']).C\_CASE.nunique().round(0).to\_frame()

clsb2 = clsb2.reset\_index()

bin = [1, 2, 5, 10, 20, 300, 100000]

bin\_labels = ['1-2', '2-5', '5-10', '10-20','20-300', '>300']

clsb2['C\_COUNT'] = pd.cut(clsb2['C\_CASE'], bin, right = False, labels = bin\_labels);

clsb2['C\_COUNT'].value\_counts() # It is not very balanced but we continue

clsb2.head()

independent columns are (X): Month, Hour and Weather and the dependent column is Count. It was observed adding more columns to independent would drop the accuracy score.

We used both KNN and Random Forest just to compare. The accuracy for each model was around 0.8 which means these models can predict the number of accidents for the time of the month and the weather with 80% accuracy.

