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

# Dealing with Null and other unknown values:

Number of Null values just from driver’s data:

C\_YEAR 0

C\_MNTH 230 -- assign 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

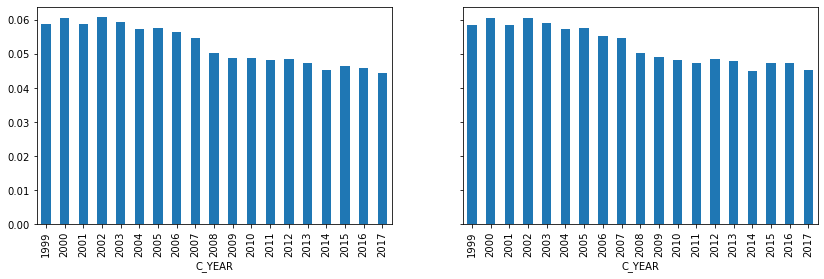
2015 161623

2016 159699

2017 154607

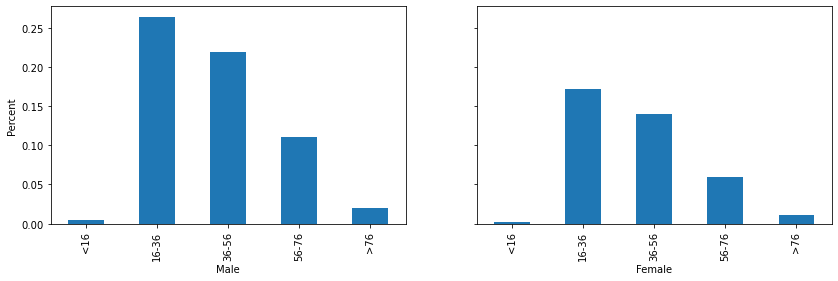
# Observations by looking at each element separately

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



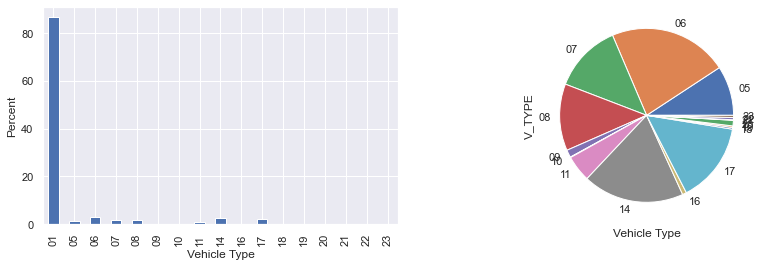
**More Accidents are between**

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

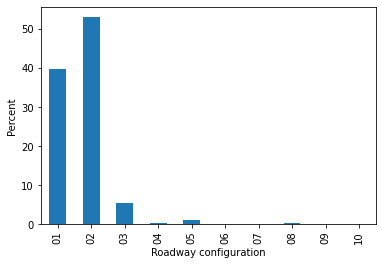


### **V\_Type = 01 - Light Duty Vehicle (Plot 03)**

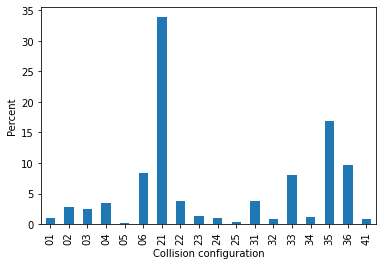
If we remove Light Duty Vehicles as an outliner (temporarily) then we have trucks and vans which are not cargo (06) and after that we have motorcycles (14) and then bicycles (17)



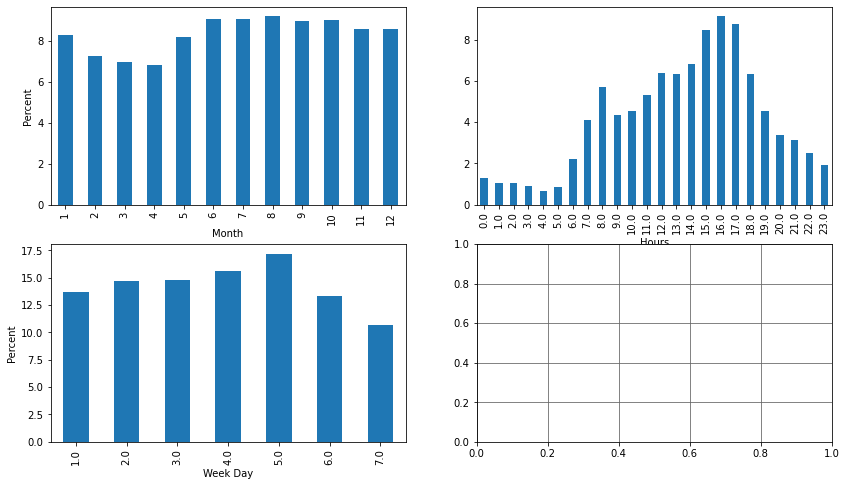
# **C\_RCFG = 02 - Roadway configuration: At an intersection of at least two public roadways (Plot 04)**



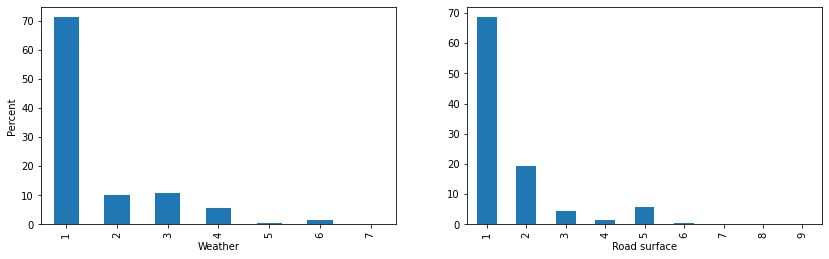
# **C\_CONF = 21 rear-end collision (Plot 05)**



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



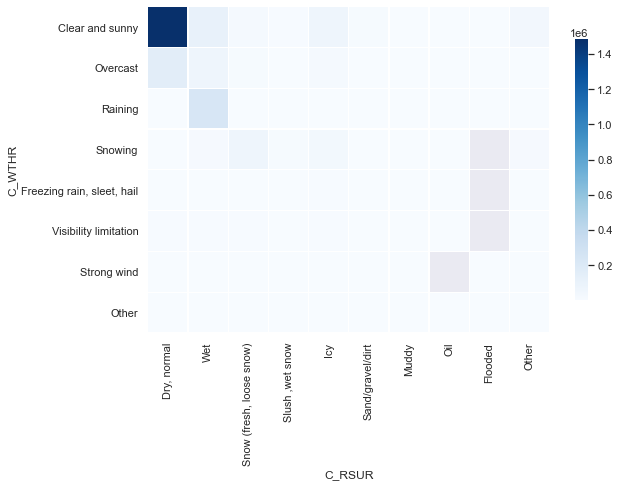
## **Weather clear and sunny. Road surface Dry and Normal (Plot 07)**



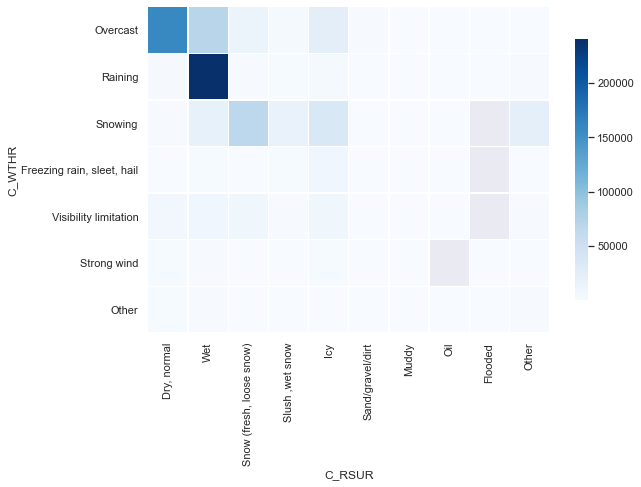
# Correlations

1. Correlation between weather and road surface (Plot 8)

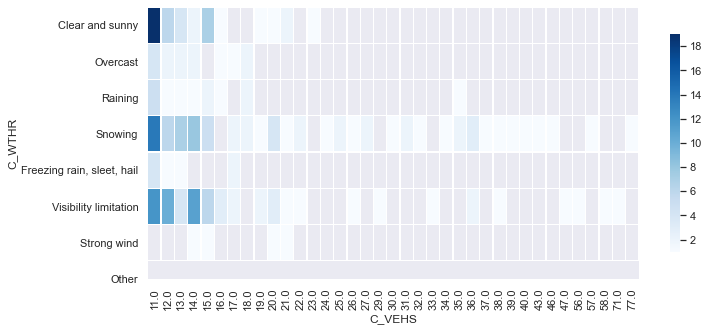
Clear and Normal weather + Dry, normal surface have the most collisions, after removing Clear and Dry weather other relations are clearer.



After removing Clear and Sunny



1. Weather and number of Vehicles (Plot 9)



# Looking at seasons

We have seasonality in Month

