Data Cleaning
for
Kelowna Weather-Crash Project

Section 1

remove all 'flag' variables

group weather data based on time groups in crash dataset, then join weather data on crash table (1 case = 1 crash, what was the weather like when this crash occurred?)

is 'weather' reliable? may need to correct it if there is precipitation but NA for 'weather'

```
> fullWeather = c()
> for (i in c(2017:2021)){
+   for (j in c(1:12)){
+     temp = subset(read.csv(paste0('../../weatherdata/en_climate_hourly_BC_1123939_',
+       sprintf("%02d", j), '-', i, '_P1H.csv')),
+       select = - c(`Temp.Flag`,
+       `Dew.Point.Temp.Flag`, `Rel.Hum.Flag`,
+       `Precip..Amount.Flag`, `Wind.Dir.Flag`,
+       `Wind.Spd.Flag`, `Visibility.Flag`,
+       `Stn.Press.Flag`, `Hmdx`, `Hmdx.Flag`, `Wind.Chill.Flag`))
+     fullWeather = rbind(fullWeather, temp)
+   }
+ }
> nrow(fullWeather)

[1] 43824

> 24*365*5 + 24 #2020 was a leap year

[1] 43824

> summary(fullWeather)

 Longitude..x.     Latitude..y.    Station.Name     Climate.ID
 Min.   :-119.4   Min.   :49.96   KELOWNA:43824   Min.   :1123939
 1st Qu.:-119.4   1st Qu.:49.96                   1st Qu.:1123939
 Median :-119.4   Median :49.96                   Median :1123939
 Mean   :-119.4   Mean   :49.96                   Mean   :1123939
 3rd Qu.:-119.4   3rd Qu.:49.96                   3rd Qu.:1123939
 Max.   :-119.4   Max.   :49.96                   Max.   :1123939

         Date.Time..LST.       Year          Month             Day
 2017-01-01 00:00:    1   Min.   :2017   Min.   : 1.000   Min.   : 1.00
 2017-01-01 01:00:    1   1st Qu.:2018   1st Qu.: 4.000   1st Qu.: 8.00
 2017-01-01 02:00:    1   Median :2019   Median : 7.000   Median :16.00
 2017-01-01 03:00:    1   Mean   :2019   Mean   : 6.524   Mean   :15.73
 2017-01-01 04:00:    1   3rd Qu.:2020   3rd Qu.:10.000   3rd Qu.:23.00
 2017-01-01 05:00:    1   Max.   :2021   Max.   :12.000   Max.   :31.00
 (Other)         :43818
```

```
  Time..LST.      Temp...C.       Dew.Point.Temp...C.  Rel.Hum....
 00:00  : 1826   Min.   :-28.900   Min.   :-32.800     Min.   : 12.00
 01:00  : 1826   1st Qu.:  0.800   1st Qu.: -2.100     1st Qu.: 52.00
 02:00  : 1826   Median :  7.800   Median :  2.300     Median : 74.00
 03:00  : 1826   Mean   :  8.511   Mean   :  2.043     Mean   : 69.33
 04:00  : 1826   3rd Qu.: 15.700   3rd Qu.:  7.500     3rd Qu.: 89.00
 05:00  : 1826   Max.   : 43.800   Max.   : 19.700     Max.   :100.00
 (Other):32868   NA's   :30        NA's   :29          NA's   :24
 Precip..Amount..mm. Wind.Dir..10s.deg. Wind.Spd..km.h.  Visibility..km.
 Min.   :0.00000     Min.   : 1.00      Min.   : 0.000   Min.   : 0.0
 1st Qu.:0.00000     1st Qu.: 9.00      1st Qu.: 4.000   1st Qu.:16.1
 Median :0.00000     Median :18.00      Median : 5.000   Median :16.1
 Mean   :0.02999     Mean   :19.13      Mean   : 8.415   Mean   :15.1
 3rd Qu.:0.00000     3rd Qu.:33.00      3rd Qu.:11.000   3rd Qu.:16.1
 Max.   :7.10000     Max.   :36.00      Max.   :58.000   Max.   :16.1
 NA's   :24          NA's   :13955      NA's   :52       NA's   :31
 Stn.Press..kPa.   Wind.Chill         Weather
 Min.   :93.73   Min.   :-34.00   Rain    : 2069
 1st Qu.:96.10   1st Qu.:-11.00   Snow    : 1662
 Median :96.52   Median : -6.00   Haze    : 1187
 Mean   :96.55   Mean   : -8.22   Fog     :  873
 3rd Qu.:96.96   3rd Qu.: -4.00   Rain,Fog:  233
 Max.   :99.34   Max.   : -1.00   (Other) :  246
 NA's   :34      NA's   :36294    NA's    :37554
```

```
> fullCrash = subset(read.csv('../../crashdata/Southern Interior_Full Data_data.csv'),
+           select = - c(`Crash.Breakdown.2`, `Region`,
+                        `Municipality.Name..ifnull.`))
> summary(fullCrash)
```

```
 Date.Of.Loss.Year Animal.Flag              Crash.Severity  Cyclist.Flag
 Min.   :2017      No :54118   CASUALTY CRASH      :11473   No :55725
 1st Qu.:2018      Yes: 2018   PROPERTY DAMAGE ONLY:44663   Yes:  411
 Median :2019
 Mean   :2019
 3rd Qu.:2020
 Max.   :2021

    Day.Of.Week                Derived.Crash.Configuration Heavy.Veh.Flag
 FRIDAY   :9316   REAR END                    :13024        No :54085
 MONDAY   :8024   SINGLE VEHICLE              :12495        Yes: 2051
 SATURDAY :6753   UNDETERMINED                :11453
 SUNDAY   :5464   SIDE IMPACT                 :11369
 THURSDAY :9061   CONFLICTED                  : 3068
 TUESDAY  :8728   SIDE SWIPE - SAME DIRECTION: 1833
 WEDNESDAY:8790   (Other)                     : 2894
 Intersection.Crash  Month.Of.Year   Motorcycle.Flag Parked.Vehicle.Flag
 No :31677           JULY    : 5152   No :55673       No :38567
 Yes:24459           DECEMBER: 5095   Yes:  463       Yes:17569
                     JANUARY : 5063
                     AUGUST  : 4907
                     OCTOBER : 4836
```

```
                  JUNE    : 4735
                  (Other) :26348
 Parking.Lot.Flag Pedestrian.Flag  Street.Full.Name..ifnull.
 No :37189        No :55790       HWY 97         : 6019
 Yes:18947        Yes:  346       HARVEY AVE     : 3168
                                  HWY 33         : 2064
                                  GORDON DR      : 1921
                                  LAKESHORE RD   : 1372
                                  SPRINGFIELD RD: 1326
                                  (Other)        :40266
     Time.Category      Municipality.Name Road.Location.Description
 12:00-14:59:14870   KELOWNA     :45943   UNKNOWN      : 2211
 15:00-17:59:14473   WEST KELOWNA:10193   HWY 97       : 1961
 09:00-11:59:11021                        HARVEY AVE   : 1670
 18:00-20:59: 5805                        LAKESHORE RD:  932
 06:00-08:59: 5489                        LOUIE DR     :  751
 21:00-23:59: 2684                        HWY 33       :  715
 (Other)    : 1794                        (Other)      :47896
       Street.Full.Name Metric.Selector Total.Crashes   Total.Victims
 HWY 97         : 6019   Min.   :1.000   Min.   :1.000   Min.   :0.0000
 HARVEY AVE     : 3168   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000
 HWY 33         : 2064   Median :1.000   Median :1.000   Median :0.0000
 GORDON DR      : 1921   Mean   :1.006   Mean   :1.006   Mean   :0.2917
 LAKESHORE RD   : 1372   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.0000
 SPRINGFIELD RD: 1326   Max.   :3.000   Max.   :3.000   Max.   :9.0000
 (Other)        :40266

> #save(alldata, file = "../rda_files/all_data.rda")
```