Preparing ADM's first data set

Matt Richey/Jaime Davila

2/13/2022

Introduction

Let's do a quick tour of some introductory concepts and programming. Our data will come from the Minneapolis Open Data Portal. In particular, we'll look at the Police Incidents from 2016 and 2017.

- Police Incidents 2016
- Police Incidents 2017

The goal is to try to predict the number of "incidents" as a function of the time of year, taking care to account for the effects of the day of the week. We will use one year (2016) as the "training" data and then see how it works on the other year (2017).

Set up

Define some variables for the file names.

```
dataFile16 <- "~/Mscs 341 S22/Class/Data/Police_Incidents_2016.csv"
dataFile17 <- "~/Mscs 341 S22/Class/Data/Police_Incidents_2017.csv"
```

Now we can load the data.

```
data16.df0 <- read_csv(dataFile16,
                       col_types= (cols(ReportedDate=col_datetime(format="%Y/%m/%d %H:%M:%S")))) %>%
  select(ReportedDate,Time,UCRCode,Precinct) %>%
  mutate(year=2016)

data17.df0 <- read_csv(dataFile17) %>%
  select(ReportedDate,Time,UCRCode,Precinct) %>%
  mutate(year=2017)

## Row  bind these data
data.df0 <- bind_rows(data16.df0,data17.df0)
```

Take a peek at how the UCR Codes are distributed.

```
# Notice the use of with, which is equivalent to
# table(data.df0$year, data.df0$UCRCode)
with(data.df0,table(year,UCRCode))
```

```
##       UCRCode
## year      01    03    04    05    06    07    08    10
##   2016    24   281  1694  1726  3321 11176  1849    84
##   2017    24   342  1714  1785  3717 12230  2161   112
```

Now we have to do some data wrangling. The key fields are:

- ReportedDate: the Date+Time that the report was filed.
- Time: the approximate time of the incident.
- UCRCode: These describe the type of incident
  - 1 = MURDER
  - 3 = RAPE
  - 4 = ROBBERY
  - 5 = ASSAULT
  - 6 = BURGLARY
  - 7 = LARCENY
  - 8 = AUTO THEFT
  - 10 = ARSON

An annoying feature of the this data is the `ReportedDate` field contains the day and time the incident was reported. The `Time` field contains the time the incident occurred (at least as described by the victim). The `ReportDate` could on, say, Wednesday, but the actual incident was on Tuesday. We need to correct for that.

As well, we want to include information on the day of the week and the day of the year.

```
data.df <- data.df0 %>%
  mutate(ReportedDateTime=ReportedDate,
         ReportDate=date(ReportedDateTime),
         ReportHour=hour(ReportedDateTime),
         IncidentTime=Time,
         IncidentHour=hour(IncidentTime),
         ##Correct for previous day
         IncidentDate=ReportDate-(IncidentHour > ReportHour),
         ##decimal value of time.
         time=hour(IncidentTime)+minute(IncidentTime)/60,
         ##########################
         ##correct the year.
         year=factor(year(IncidentDate)),
         ## day of year
         yday=yday(IncidentDate),
         ##month, of course
         month=month(IncidentDate),
          ## day of week (1=Monday)
         wday=wday(IncidentDate, label=TRUE),
         ## week of the year
         week=week(IncidentDate),
         )%>%
         select(wday,week,yday,month,year,UCRCode,Precinct)
```

Finally, let's tally the number of incidents by the day of the week.

Now we can start counting......

```
dataWeekDay.df <- data.df %>%
  ## WeekDay, YearDay, and Year
  group_by(wday,week,year) %>%
  summarize(tot=n())
```

And let's save so that we can use it in class

```
write_csv(dataWeekDay.df,"~/Mscs 341 S22/Class/Data/police_incidents.mn.csv")
```