Group 7: Traffic Dataset Cleaning Script

Samuel Geddie + Bryce Emery

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(here)
```

```
here() starts at /home/brycee/_School/456-MATH-Statistical-MethodsII/group
```

```
traffic_accidents.raw <- read.csv(here::here('data/traffic_accidents.csv'))
traffic_accidents.clean <- traffic_accidents.raw %>% janitor::clean_names()
```

Dataset - Traffic Accidents

Dataset Link

Dataset Description

The dataset contains a collection of traffic accidents scraped from the web and includes a large number of useful observed variables (24 columns!) over a large (>200,000) entries. Example observations include details about conditions and qualities of the roadway where the accident occurred, and the type and results of the specific collision(s). On a sour note: the author does not give information on where they were scrapped from, nor the locale these recordings are from–however, given that we are going to be using this data simply for educational exercises this should be fine.

```
summary(traffic_accidents.clean)
```

```
 crash_date        traffic_control_device weather_condition
Length:209306      Length:209306          Length:209306
Class :character   Class :character       Class :character
Mode  :character   Mode  :character       Mode  :character


lighting_condition first_crash_type   trafficway_type     alignment
Length:209306      Length:209306      Length:209306      Length:209306
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


roadway_surface_cond road_defect         crash_type
Length:209306        Length:209306      Length:209306
Class :character     Class :character   Class :character
Mode  :character     Mode  :character   Mode  :character


intersection_related_i    damage          prim_contributory_cause
Length:209306          Length:209306      Length:209306
Class :character       Class :character   Class :character
Mode  :character       Mode  :character   Mode  :character


  num_units      most_severe_injury injuries_total    injuries_fatal
Min.   : 1.000   Length:209306      Min.   : 0.0000   Min.   :0.000000
1st Qu.: 2.000   Class :character   1st Qu.: 0.0000   1st Qu.:0.000000
```

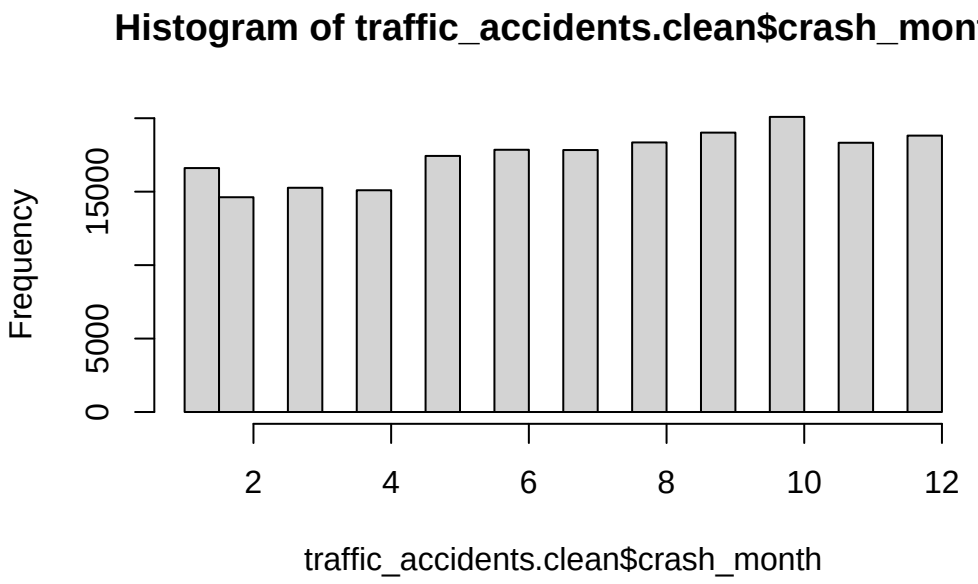
```
 Median : 2.000   Mode  :character   Median : 0.0000   Median :0.000000
 Mean   : 2.063                      Mean   : 0.3827   Mean   :0.001858
 3rd Qu.: 2.000                      3rd Qu.: 1.0000   3rd Qu.:0.000000
 Max.   :11.000                      Max.   :21.0000   Max.   :3.000000
 injuries_incapacitating injuries_non_incapacitating
 Min.   :0.0000          Min.   : 0.0000
 1st Qu.:0.0000          1st Qu.: 0.0000
 Median :0.0000          Median : 0.0000
 Mean   :0.0381          Mean   : 0.2212
 3rd Qu.:0.0000          3rd Qu.: 0.0000
 Max.   :7.0000          Max.   :21.0000
 injuries_reported_not_evident injuries_no_indication   crash_hour
 Min.   : 0.0000               Min.   : 0.000         Min.   : 0.00
 1st Qu.: 0.0000               1st Qu.: 2.000         1st Qu.: 9.00
 Median : 0.0000               Median : 2.000         Median :14.00
 Mean   : 0.1215               Mean   : 2.244         Mean   :13.37
 3rd Qu.: 0.0000               3rd Qu.: 3.000         3rd Qu.:17.00
 Max.   :15.0000               Max.   :49.000         Max.   :23.00
 crash_day_of_week  crash_month
 Min.   :1.000     Min.   : 1.000
 1st Qu.:2.000     1st Qu.: 4.000
 Median :4.000     Median : 7.000
 Mean   :4.144     Mean   : 6.772
 3rd Qu.:6.000     3rd Qu.:10.000
 Max.   :7.000     Max.   :12.000
```

crash_date

This is a combined categorical/numerical variable. It contains the Date and Time of the recorded event in the form `Month/Day/Year Hour:Minute:Second AM/PM`

Note: We will be using Lubridate to handle this variable, and split it into it's composite information (month/day/time). While these are provided, practice with the package seems useful longer term.

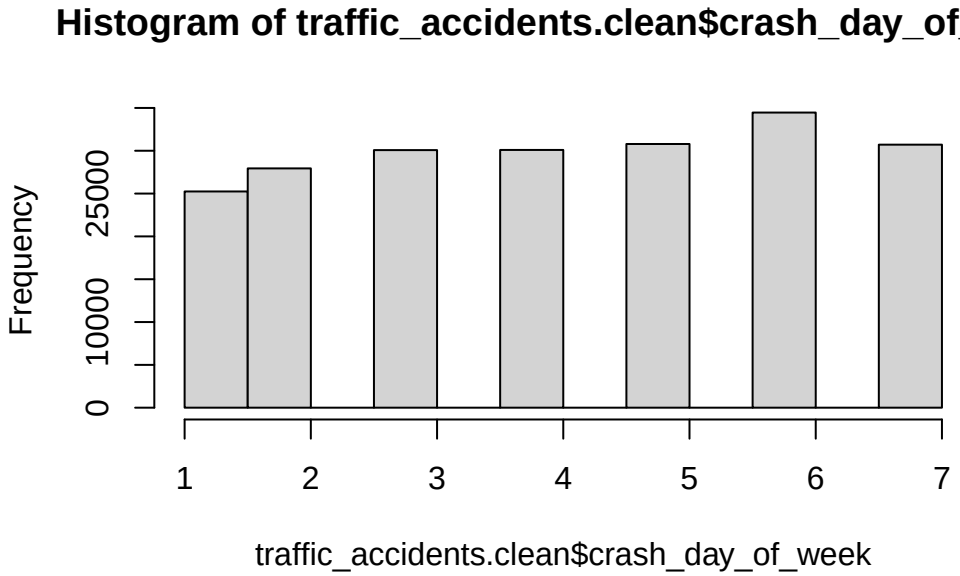
```
## Needs to get split up by lubridate
#table(traffic_accidents.clean$crash_date)

## To simplify completing Project 2, we use the provided date compotent variables
hist(traffic_accidents.clean$crash_month)
```




```r
table(traffic_accidents.clean$crash_month, useNA = 'always')
```

```
    1     2     3     4     5     6     7     8     9    10    11    12  <NA>
16606 14621 15265 15096 17432 17851 17834 18350 19018 20089 18328 18816     0
```

```r
hist(traffic_accidents.clean$crash_day_of_week)
```



```
table(traffic_accidents.clean$crash_day_of_week, useNA = 'always')
```

```
    1     2     3     4     5     6     7  <NA>
25246 27938 30074 30093 30787 34458 30710     0
```

traffic_control_device

This is a categorical variable with 19 possible values. It includes most of the common forms of traffic control encounter in the United States.

With 4455 unknown values, this variable is a candidate for MICE.

```
table(traffic_accidents.clean$traffic_control_device)
```

```
   BICYCLE CROSSING SIGN              DELINEATORS  FLASHING CONTROL SIGNAL
                      11                       17                      150
        LANE USE MARKING              NO CONTROLS              NO PASSING
                     153                    29508                       12
                   OTHER  OTHER RAILROAD CROSSING          OTHER REG. SIGN
                     670                       23                      181
      OTHER WARNING SIGN PEDESTRIAN CROSSING SIGN           POLICE/FLAGMAN
                      95                      247                      104
  RAILROAD CROSSING GATE         RR CROSSING SIGN              SCHOOL ZONE
                      78                       18                       33
       STOP SIGN/FLASHER           TRAFFIC SIGNAL                  UNKNOWN
                   49139                   123944                     4455
                   YIELD
                     468
```

weather_condition

This is a categorical variable with 12 possible values. It covers the weather one might encounter when driving.

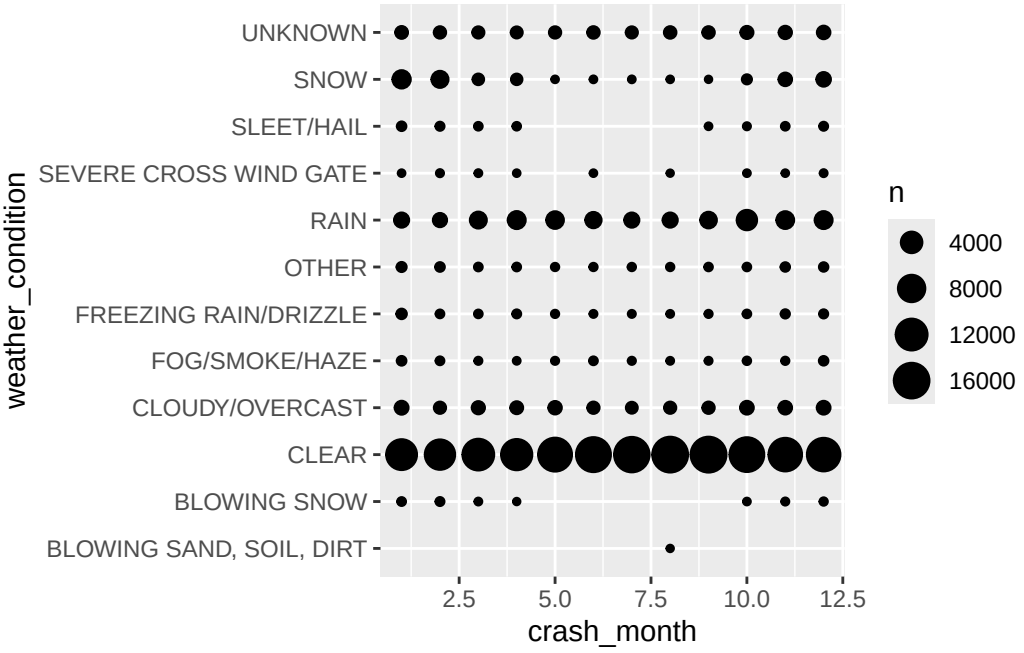
It is a candidate for MICE analysis, especially given the likely correlates relating to the time of year the recorded collision(s) took place (e.g. snow during winter).

However, that same reason marks for concerns of possible confounding. Initial visual exploration plotting the two suggests this exists somewhat, but primarily for snow-related conditions (snow, hail, etc).

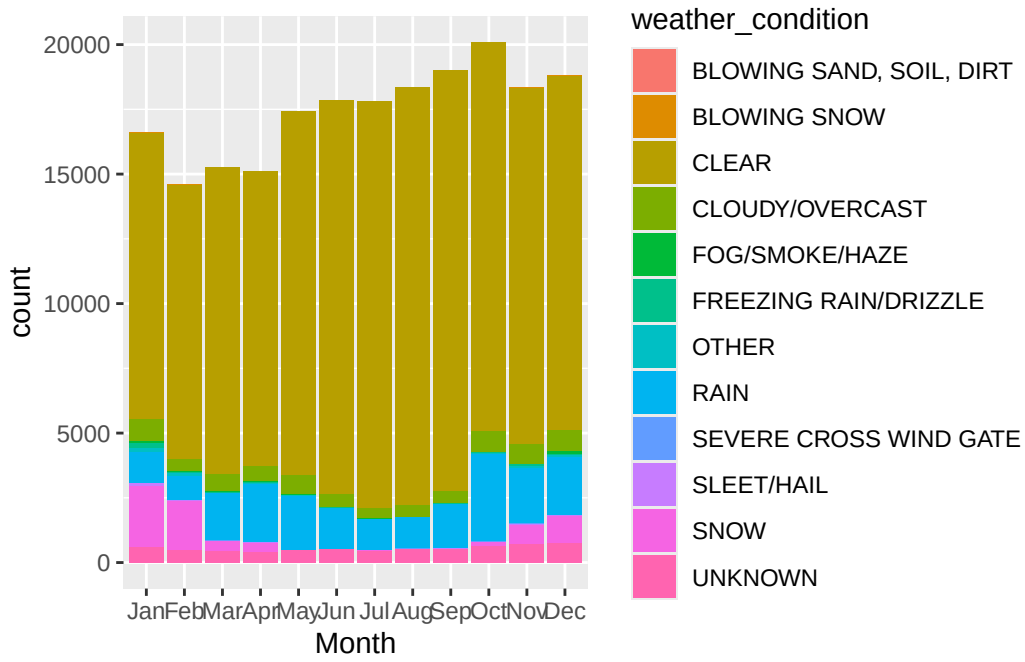
```
table(traffic_accidents.clean$weather_condition)
```

```
BLOWING SAND, SOIL, DIRT             BLOWING SNOW                    CLEAR
                       1                      127                   164700
      CLOUDY/OVERCAST          FOG/SMOKE/HAZE    FREEZING RAIN/DRIZZLE
                    7533                      360                      510
                   OTHER                     RAIN   SEVERE CROSS WIND GATE
                     627                    21703                       32
              SLEET/HAIL                     SNOW                  UNKNOWN
                     308                     6871                     6534
```

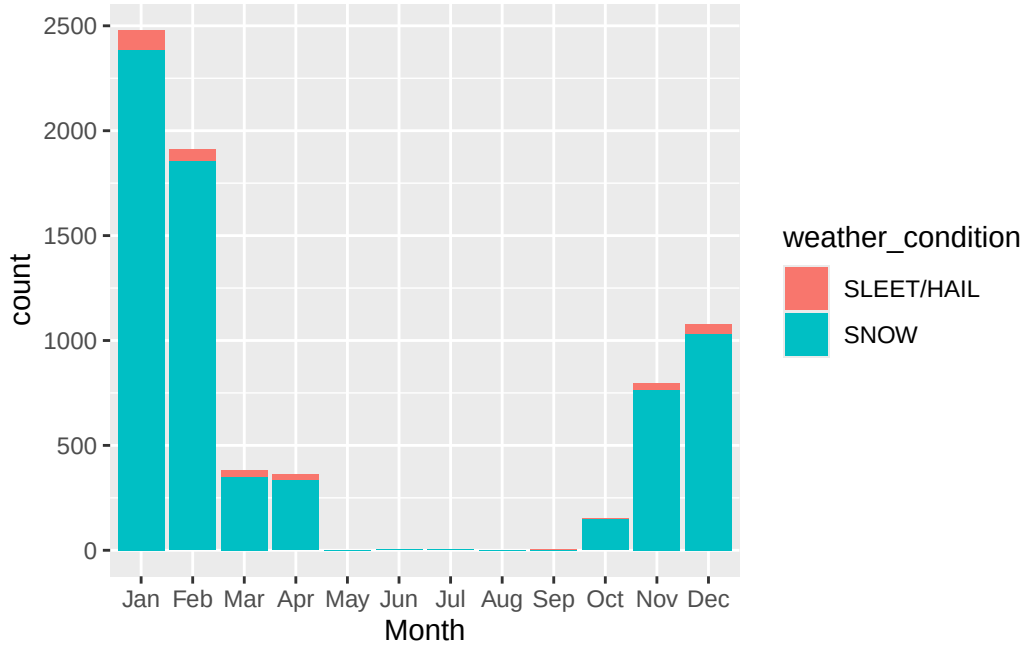
```
ggplot(traffic_accidents.clean, aes(x=crash_month, y=weather_condition))+
  geom_count()
```



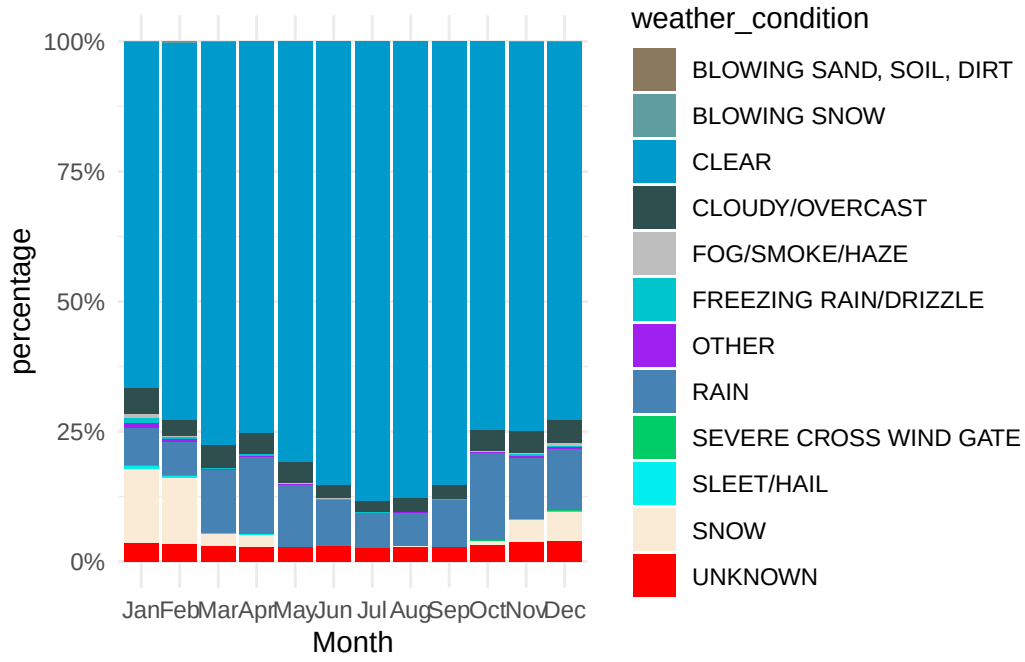

```
traffic_accidents.clean %>%
  mutate(Month = month(mdy_hms(crash_date), label=TRUE)) %>%
  ggplot(aes(x=Month, fill=weather_condition))+
    geom_bar()
```




```
traffic_accidents.clean %>%
  filter(weather_condition == "SNOW" | weather_condition == "SLEET/HAIL") %>%
  mutate(Month = month(mdy_hms(crash_date), label=TRUE)) %>%
  ggplot(aes(x=Month, fill=weather_condition))+
    geom_bar()
```



```r
color.values <- c('navajowhite4', 'cadetblue', 'deepskyblue3', 'darkslategrey',
                  'grey', 'turquoise3', 'purple', 'steelblue',
                  'springgreen3', 'cyan2', 'antiquewhite', 'red')
# .groups="drop" method
traffic_accidents.clean %>%
  mutate(Month = month(mdy_hms(crash_date), label=TRUE)) %>%
  group_by(Month, weather_condition) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(Month) %>%
  mutate(percentage = n/ sum(n)) %>%
  ggplot(aes(x=Month, y=percentage, fill=weather_condition)) +
    geom_bar(stat ="identity", position="stack")+
    scale_y_continuous(labels = scales::percent_format(), limits  =c(0,1))+
    theme_minimal()+
    scale_fill_manual(values=color.values)
```

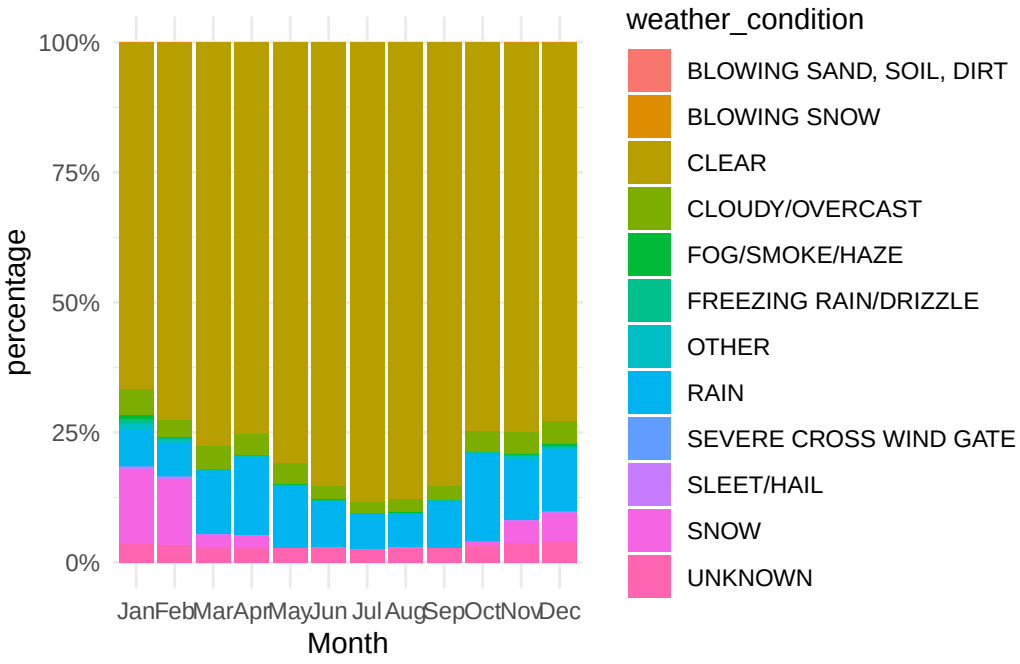



```r
## Last Try ChatGPT


traffic_accidents.clean %>%
  mutate(Month = month(mdy_hms(crash_date), label=TRUE)) %>%
  group_by(Month, weather_condition) %>%
  summarise(n = n()) %>%
```

```
  mutate(percentage = n/ sum(n)) %>%
  print() %>%
  ggplot(aes(x=Month, y=percentage, fill=weather_condition)) +
    #geom_area(alpha=0.6 , size=1, colour="black")+
    geom_bar(stat ="identity", position="stack")+
    scale_y_continuous(labels = scales::percent_format(), limits  =c(0,1))+
    #geom_point()+
    theme_minimal()
```

```
`summarise()` has grouped output by 'Month'. You can override using the
`.groups` argument.
```

```
# A tibble: 121 x 4
# Groups:   Month [12]
   Month weather_condition          n percentage
   <ord> <chr>                  <int>      <dbl>
 1 Jan   BLOWING SNOW              41   0.00247
 2 Jan   CLEAR                  11021   0.664
 3 Jan   CLOUDY/OVERCAST          844   0.0508
 4 Jan   FOG/SMOKE/HAZE           103   0.00620
 5 Jan   FREEZING RAIN/DRIZZLE    191   0.0115
 6 Jan   OTHER                    145   0.00873
 7 Jan   RAIN                    1197   0.0721
 8 Jan   SEVERE CROSS WIND GATE     2   0.000120
 9 Jan   SLEET/HAIL                95   0.00572
10 Jan   SNOW                    2385   0.144
# i 111 more rows
```




```
## Reference
#data <- data  %>%
#  group_by(time, group) %>%
#  summarise(n = sum(value)) %>%
#  mutate(percentage = n / sum(n))
```

lighting_condition

This is a Categorical Variable with six possible values: "DARKNESS", "DARKNESS, LIGHTED ROAD", "DAWN", "DAYLIGHT", and "UNKNOWN".

One concern is that the lighting condition could be confounded by month and time. We want to ensure that it's not so simple as that. Which unfortunately it appears to be.

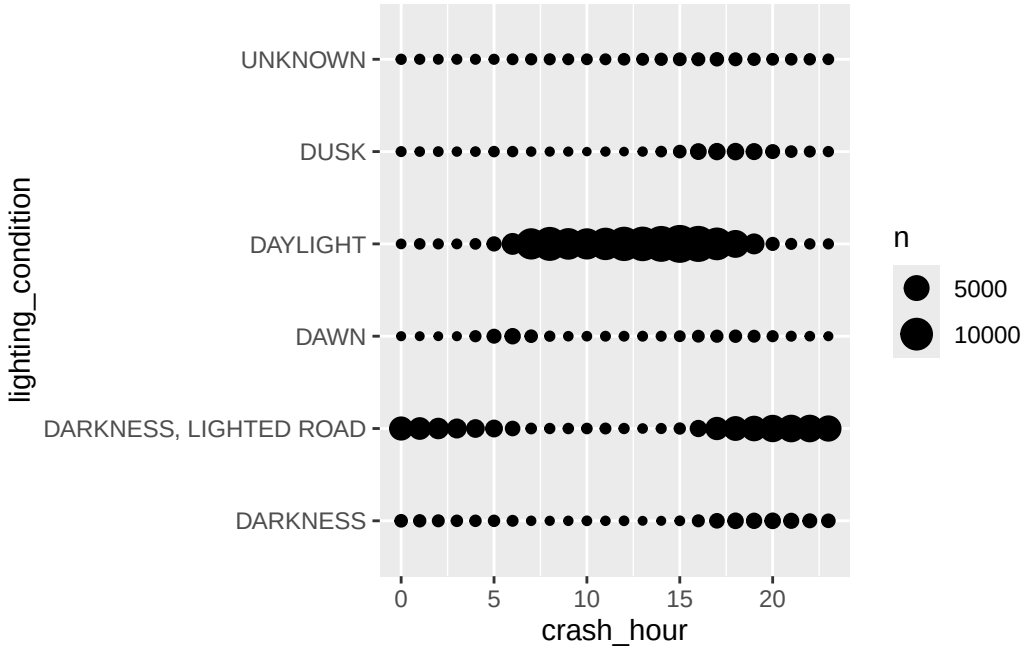
We've included it in this report for transparency of work, and on the potential that the categories of "DARKNESS" and "DARKNESS, LIGHTED ROAD" could be of value at a later date. Perhaps adding texture to the time of day.

```
table(traffic_accidents.clean$lighting_condition)
```

```
            DARKNESS DARKNESS, LIGHTED ROAD                         DAWN
                7436                   53378                         3724
```

```
DAYLIGHT                 DUSK                 UNKNOWN
  134109                 6323                    4336
```

```
ggplot(traffic_accidents.clean, aes(x=crash_hour, y=lighting_condition))+
  geom_count()
```



first_crash_type

This is a Categorical variable with 18 possible values.

Note that it is different from "Primary Cause" in that it is describing what was colliding with upon recorded collision(s).

```
table(traffic_accidents.clean$first_crash_type)
```

```
                  ANGLE                  ANIMAL
                  52250                      77
            FIXED OBJECT                HEAD ON
                   4742                    1790
      OTHER NONCOLLISION            OTHER OBJECT
                    249                     759
             OVERTURNED     PARKED MOTOR VEHICLE
```

```
                        96                         4893
                PEDALCYCLIST                   PEDESTRIAN
                        5337                         8996
                    REAR END                 REAR TO FRONT
                       42018                         1157
                REAR TO REAR                  REAR TO SIDE
                          49                          773
SIDESWIPE OPPOSITE DIRECTION     SIDESWIPE SAME DIRECTION
                        1839                        20116
                       TRAIN                      TURNING
                           8                        64157
```

trafficway_type

This is a Categorical Variable with 20 possible values describing various intersections that may be encountered by motorists.

```
table(traffic_accidents.clean$trafficway_type)
```

```
                          ALLEY                 CENTER TURN LANE
                            741                             2862
DIVIDED - W/MEDIAN (NOT RAISED)       DIVIDED - W/MEDIAN BARRIER
                          34221                            10720
                        DRIVEWAY              FIVE POINT, OR MORE
                            143                             1119
                       FOUR WAY                  L-INTERSECTION
                           49057                              127
                     NOT DIVIDED                    NOT REPORTED
                           77753                              581
                        ONE-WAY                           OTHER
                           12341                             4757
                     PARKING LOT                            RAMP
                             448                              375
                      ROUNDABOUT                  T-INTERSECTION
                             149                             9233
                   TRAFFIC ROUTE                         UNKNOWN
                             776                             1060
       UNKNOWN INTERSECTION TYPE                  Y-INTERSECTION
                            1885                              958
```

roadway_surface_cond

This is a Categorical Variable with 7 possible values: “DRY”, “ICE”, “SNOW OR SLUSH”, “WET”, “SAND, MUD, DIRT”, “OTHER”, AND “UNKNOWN”.

This is a candidate for MICE analysis.

```
table(traffic_accidents.clean$roadway_surface_cond)
```

```
            DRY             ICE           OTHER SAND, MUD, DIRT   SNOW OR SLUSH
         155905            1303             438              40            6203
        UNKNOWN             WET
          12509           32908
```

road_defect

This is a Categorical Variable with 7 possible values: “DEBRIS ON ROADWAY”, “SHOULDER DEFECT”, “NO DEFECTS”, “WORN SURFACE”, “RUT, HOLES”, “OTHER”, AND “UNKNOWN”

This is a candidate for MICE analysis.

```
table(traffic_accidents.clean$road_defect)
```

```
DEBRIS ON ROADWAY        NO DEFECTS             OTHER        RUT, HOLES
              139            171730               912               741
  SHOULDER DEFECT           UNKNOWN      WORN SURFACE
              358             34426              1000
```

intersection_related_i

This is a Categorical Binary Variable, with the values ‘Y’ and ‘N’. It indicates whether the recorded collision(s) is related to an occuring at an intersection.

Potentially of interest is whether this correlates with pedestrian injury. Given that intersections house the most common pedestrian related traffic control, being crosswalks.

```
table(traffic_accidents.clean$intersection_related_i)
```

```
    N      Y
 9982 199324
```

prim_contributory_cause

This is a Categorical Variable with 40 possible values. It lists the primary cause of the recorded collision(s) and provides a highly granular list of possibilities.

These descriptions differ from `first_crash_type` in that they mostly speak in terms of human error..

```
table(traffic_accidents.clean$prim_contributory_cause)
```

```
                                                                     ANIMAL
                                                                         49
                                            BICYCLE ADVANCING LEGALLY ON RED LIGHT
                                                                         32
                                                CELL PHONE USE OTHER THAN TEXTING
                                                                        254
                                                 DISREGARDING OTHER TRAFFIC SIGNS
                                                                       1099
                                                       DISREGARDING ROAD MARKINGS
                                                                        336
                                                           DISREGARDING STOP SIGN
                                                                       6749
                                                     DISREGARDING TRAFFIC SIGNALS
                                                                      14591
                                                          DISREGARDING YIELD SIGN
                                                                        132
                                                DISTRACTION - FROM INSIDE VEHICLE
                                                                       1275
                                               DISTRACTION - FROM OUTSIDE VEHICLE
                                                                        760
     DISTRACTION - OTHER ELECTRONIC DEVICE (NAVIGATION DEVICE, DVD PLAYER, ETC.)
                                                                         93
                                                  DRIVING ON WRONG SIDE/WRONG WAY
                                                                       1188
                                              DRIVING SKILLS/KNOWLEDGE/EXPERIENCE
                                                                       5048
                                                    EQUIPMENT - VEHICLE CONDITION
                                                                        952
                               EVASIVE ACTION DUE TO ANIMAL, OBJECT, NONMOTORIST
                                                                        284
                                                 EXCEEDING AUTHORIZED SPEED LIMIT
                                                                        403
```

EXCEEDING SAFE SPEED FOR CONDITIONS
441
FAILING TO REDUCE SPEED TO AVOID CRASH
10676
FAILING TO YIELD RIGHT-OF-WAY
42914
FOLLOWING TOO CLOSELY
19084
HAD BEEN DRINKING (USE WHEN ARREST IS NOT MADE)
133
IMPROPER BACKING
2340
IMPROPER LANE USAGE
6462
IMPROPER OVERTAKING/PASSING
8302
IMPROPER TURNING/NO SIGNAL
12643
MOTORCYCLE ADVANCING LEGALLY ON RED LIGHT
7
NOT APPLICABLE
5241
OBSTRUCTED CROSSWALKS
48
OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER
1868
PASSING STOPPED SCHOOL BUS
17
PHYSICAL CONDITION OF DRIVER
779
RELATED TO BUS STOP
79
ROAD CONSTRUCTION/MAINTENANCE
298
ROAD ENGINEERING/SURFACE/MARKING DEFECTS
179
TEXTING
72
TURNING RIGHT ON RED
435
UNABLE TO DETERMINE
58316
UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)

```
                                                                    860
                VISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.)
                                                                   1793
                                                                WEATHER
                                                                   3074
```

Injuries

Each of the following is a Numerical Variable, whose values relate to counting form of injury resultant from the recorded collision(s).

Do to the relative infrequency of injuries among the recorded collision(s)s, it will be preemptively noted that the median value of all of these variables is 0 (yay).

injuries_total

This variable counts the total injuries of the recorded collision(s). The values range from 0 to 21 total injuries, with a mean of 0.38.

```
table(traffic_accidents.clean$injuries_total)
```

```
     0      1      2      3      4      5      6      7      8      9     10
154789  38378  10447   3505   1338    488    212     80     30     14      7
    11     12     13     15     16     17     19     21
     5      3      1      4      1      1      1      2
```

```
summary(traffic_accidents.clean$injuries_total)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3827  1.0000 21.0000
```

injuries_fatal

This variable counts the number of fatalities resultant from the recorded collision(s). The values range from 0 to 3 fatilies, with a mean of 0.001858.

Of note, extremely few of the recorded collision(s)s resulted in any fatalities (hooray), only occurring in 351/209306 or ~0.001680% of recorded collision(s)s.

```r
table(traffic_accidents.clean$injuries_fatal)
```

```
     0      1      2      3
208955    317     30      4
```

```r
summary(traffic_accidents.clean$injuries_fatal)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000000 0.000000 0.000000 0.001858 0.000000 3.000000
```

injuries_incapacitating

This variable counts the number of incapacitating injuries resultant from the recorded collision(s). The values range from 0 to 7 injuries, with a mean of 0.001858.

```r
table(traffic_accidents.clean$injuries_incapacitating)
```

```
     0      1      2      3      4      5      6      7
202672   5682    683    182     62     19      4      2
```

```r
summary(traffic_accidents.clean$injuries_incapacitating)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.0381  0.0000  7.0000
```

```r
## Add some sort of count here
```

injuries_non_incapacitating

This variable counts the number of non-incapacitating resultant from the recorded collision(s). The values range from 0 to 21 injuries, with a mean of 0.2212.

```r
table(traffic_accidents.clean$injuries_non_incapacitating)
```

```
     0      1      2      3      4      5      6      7      8      9     10
176306  24413   5688   1828    667    232    106     33     15      5      5
    11     12     13     14     16     18     19     21
     1      1      1      1      1      1      1      1
```

```
summary(traffic_accidents.clean$injuries_non_incapacitating)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2212  0.0000 21.0000
```

injuries_reported_not_evident

This is a Numerical Variable that counts the number of injuries that were reported but not visibly evident. The values range from 0 to 11 injuries, with a mean of 0.1215.

```
table(traffic_accidents.clean$injuries_reported_not_evident)
```

```
     0      1      2      3      4      5      6      7      8      9     10
190619  14029   3302    904    289    105     29     15      7      3      2
    11     15
     1      1
```

```
summary(traffic_accidents.clean$injuries_reported_not_evident)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1215  0.0000 15.0000
```

injuries_no_indication

This is a Numerical Variable that counts the number of incidents where no injuries were reported among parties involved in the recorded collision(s). The values range from 0 to 49, with a median of 2, and a mean of 2.244.

```
table(traffic_accidents.clean$injuries_no_indication)
```

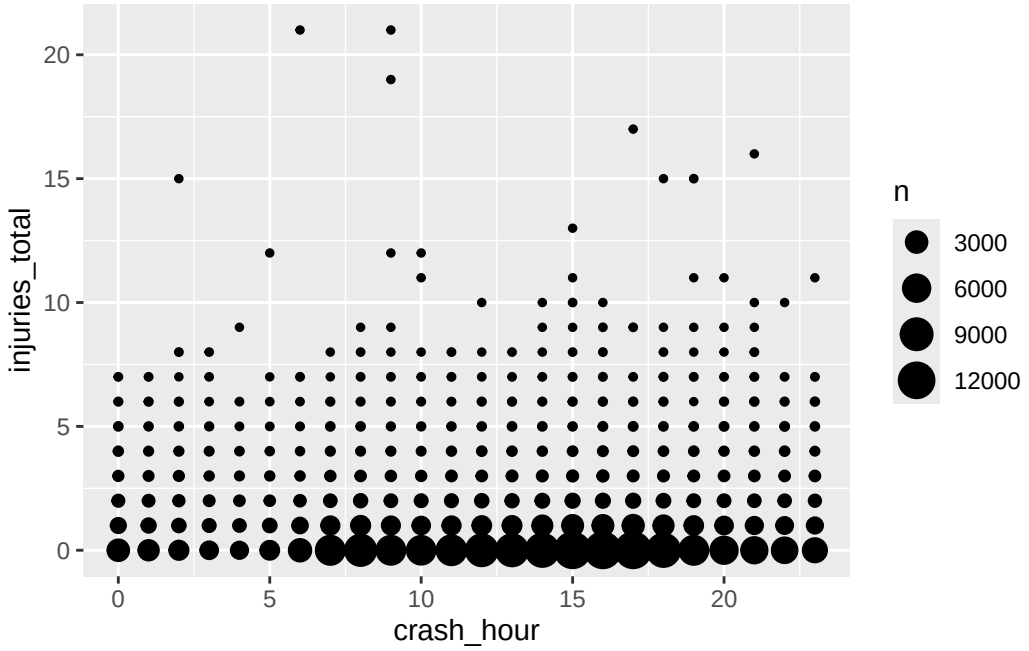
```
    0     1     2     3     4     5     6     7     8     9    10
 6229 36148 109130 34350 13453  5781  2458   945   395   188    86
   11    12    13    14    15    16    17    18    19    20    21
   32    24    12    16     8     8     5     1     4     4     2
   22    25    26    27    28    29    30    31    32    34    35
    1     1     3     4     3     2     1     1     1     1     1
   36    37    39    42    46    49
    2     1     1     2     1     1
```

```
summary(traffic_accidents.clean$injuries_no_indication)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   2.000   2.244   3.000  49.000
```

Data Exploration

```
ggplot(traffic_accidents.clean, aes(x=crash_hour, y=injuries_total))+
  geom_count()
```



```r
(holidays <- mdy(c("December 25 2000", "December 31, 2000")))
```

```
[1] "2000-12-25" "2000-12-31"
```