Exploratory data Analysis - Crimes in Los Angeles

20/10/2020

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(tinytex)
```

```
# Source : https://data.lacity.org/A-Safe-City/Crime-Data-from-2020-to-Present/2nrs-mtv8

colNames <- c("drNum",   # Division of Records Number: Official file number made up of a 2 digit year,
                         #area ID, and 5 digits
              "dateReported",
              "dateOccurred",
              "timeOccurred", #in 24 hours military time.
              "areaID", # LAPD Divisions - The LAPD has 21 Community Police Stations
              "areaName",
              "rptDistNum", # LAPD Reporting Districts - A four-digit code that represents
                            # a sub-area within a Geographic Area.
              "part12", # TODO: Find what is this ?
              "crimeCode", #Indicates the crime committed. (Same as Crime Code 1)
              "crimeCodeDesc", #Defines the Crime Code provided
              "moCode" , # Modus Operandi: Activities associated with the suspect
                         #in commission of the crime.
              "victAge",
              "victGender",
              "victDescent", # "Descent Code: See "
              "premisCode", # The type of structure, vehicle, or location where the crime took place.
              "premisDesc",  # Defines the Premise Code provided.
              "weaponUsedCode", # The type of weapon used in the crime.
              "weaponUSedDesc", #Defines the Weapon Used Code provided.
              "status" , #status of the case. (IC is the default)"
              "statusDesc", # Defines the Status Code provided."
              "crimeCode1", # Can be removed
              "crimeCode2",
              "crimeCode3",
              "crimeCode4",
              "location",
              "crossStreet",
              "latitude",
              "longtide")


colTypes <- "ccccicciicciffici ccccccccccdd"

#Descent Code:
# A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian
# H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese
# K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian
# V - Vietnamese W - White X - Unknown Z - Asian Indian


crimeLA <- read_csv("../data/Crime_Data_from_2020_to_Present.csv",
                    col_names = colNames,
                    col_types = colTypes,
                    #n_max = 100,
                    skip = 1)
```

Number of observations

```
nrow(crimeLA)
```

```
## [1] 176474
```

Number of variables

```
ncol(crimeLA)
```

```
## [1] 28
```

Variable names and types

```
glimpse(crimeLA)
```

```
## Rows: 176,474
## Columns: 28
## $ drNum          <chr> "010304468", "190101086", "201418201", "191501505", ...
## $ dateReported   <chr> "01/08/2020 12:00:00 AM", "01/02/2020 12:00:00 AM", ...
## $ dateOccurred   <chr> "01/08/2020 12:00:00 AM", "01/01/2020 12:00:00 AM", ...
## $ timeOccurred   <chr> "2230", "0330", "1830", "1730", "0415", "0030", "131...
## $ areaID         <int> 3, 1, 14, 15, 19, 1, 1, 1, 1, 1, 1, 1, 1, 1, 9, 14, ...
## $ areaName       <chr> "Southwest", "Central", "Pacific", "N Hollywood", "M...
## $ rptDistNum     <chr> "0377", "0163", "1454", "1543", "1998", "0163", "016...
## $ part12         <int> 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 2...
## $ crimeCode      <int> 624, 624, 420, 745, 740, 121, 442, 946, 341, 330, 93...
## $ crimeCodeDesc  <chr> "BATTERY - SIMPLE ASSAULT", "BATTERY - SIMPLE ASSAUL...
## $ moCode         <chr> "0444 0913", "0416 1822 1414", "1300 0344 1606 2032"...
## $ victAge        <int> 36, 25, 63, 76, 31, 25, 23, 0, 23, 29, 35, 41, 0, 24...
## $ victGender     <fct> F, M, M, F, X, F, M, X, M, M, M, M, X, F, M, M, NA, ...
## $ victDescent    <fct> B, H, H, W, X, H, H, X, B, A, O, A, X, H, O, O, NA, ...
## $ premisCode     <int> 501, 102, 103, 502, 409, 735, 404, 726, 502, 101, 10...
## $ premisDesc     <chr> "SINGLE FAMILY DWELLING", "SIDEWALK", "ALLEY", "MULT...
## $ weaponUsedCode <int> 400, 500, NA, NA, NA, 500, NA, NA, NA, 306, 511, NA,...
## $ weaponUSedDesc <chr> "STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)", "U...
## $ status         <chr> "AO", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "IC"...
## $ statusDesc     <chr> "Adult Other", "Invest Cont", "Invest Cont", "Invest...
## $ crimeCode1     <chr> "624", "624", "420", "745", "740", "121", "442", "94...
## $ crimeCode2     <chr> NA, NA, NA, "998", NA, "998", "998", "998", "998", N...
## $ crimeCode3     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ crimeCode4     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ location       <chr> "1100 W  39TH                         PL", "700 S  H...
## $ crossStreet    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "OLIVE", NA, NA,...
## $ latitude       <dbl> 34.0141, 34.0459, 33.9813, 34.1685, 34.2198, 34.0452...
## $ longtide       <dbl> -118.2978, -118.2545, -118.4350, -118.4019, -118.446...
```

Summary of the dataset

```
summary(crimeLA)
```

```
##     drNum           dateReported       dateOccurred       timeOccurred
##  Length:176474      Length:176474      Length:176474      Length:176474
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      areaID        areaName          rptDistNum            part12
##  Min.   : 1.00   Length:176474      Length:176474      Min.   :1.000
##  1st Qu.: 6.00   Class :character   Class :character   1st Qu.:1.000
##  Median :11.00   Mode  :character   Mode  :character   Median :1.000
##  Mean   :10.81                                         Mean   :1.415
##  3rd Qu.:16.00                                         3rd Qu.:2.000
##  Max.   :21.00                                         Max.   :2.000
##
##    crimeCode     crimeCodeDesc         moCode             victAge
##  Min.   :110.0   Length:176474      Length:176474      Min.   :  0.00
##  1st Qu.:330.0   Class :character   Class :character   1st Qu.: 10.00
##  Median :510.0   Mode  :character   Mode  :character   Median : 31.00
##  Mean   :512.2                                         Mean   : 29.96
##  3rd Qu.:627.0                                         3rd Qu.: 46.00
##  Max.   :956.0                                         Max.   :120.00
##
##  victGender    victDescent      premisCode     premisDesc
##  F   :63566   H      :54927   Min.   :101.0   Length:176474
##  M   :75332   W      :37182   1st Qu.:101.0   Class :character
##  X   :14844   B      :25180   Median :203.0   Mode  :character
##  H   :   15   X      :16513   Mean   :291.7
##  NA's:22717   O      :14268   3rd Qu.:501.0
##               (Other): 5685   Max.   :971.0
##               NA's   :22719   NA's   :3
##  weaponUsedCode   weaponUSedDesc        status           statusDesc
##  Min.   :101.0    Length:176474      Length:176474      Length:176474
##  1st Qu.:311.0    Class :character   Class :character   Class :character
##  Median :400.0    Mode  :character   Mode  :character   Mode  :character
##  Mean   :365.3
##  3rd Qu.:400.0
##  Max.   :516.0
##  NA's   :110506
##   crimeCode1         crimeCode2         crimeCode3         crimeCode4
##  Length:176474      Length:176474      Length:176474      Length:176474
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    location         crossStreet           latitude        longtide
##  Length:176474      Length:176474      Min.   : 0.00   Min.   :-118.7
##  Class :character   Class :character   1st Qu.:34.01   1st Qu.:-118.4
##  Mode  :character   Mode  :character   Median :34.06   Median :-118.3
##                                        Mean   :33.90   Mean   :-117.8
##                                        3rd Qu.:34.16   3rd Qu.:-118.3
##                                        Max.   :34.33   Max.   :   0.0
##
```

Missing Values

- NA (Not Avaialble) values

```
colSums(is.na(crimeLA))/nrow(crimeLA)
```

```
##          drNum   dateReported   dateOccurred   timeOccurred         areaID
##   0.000000e+00   0.000000e+00   0.000000e+00   0.000000e+00   0.000000e+00
##       areaName     rptDistNum         part12      crimeCode  crimeCodeDesc
##   0.000000e+00   0.000000e+00   0.000000e+00   0.000000e+00   0.000000e+00
##         moCode        victAge     victGender   victDescent     premisCode
##   1.345184e-01   0.000000e+00   1.287272e-01   1.287385e-01   1.699967e-05
##     premisDesc weaponUsedCode weaponUSedDesc         status     statusDesc
##   3.456600e-04   6.261886e-01   6.261886e-01   0.000000e+00   0.000000e+00
##     crimeCode1     crimeCode2     crimeCode3     crimeCode4       location
##   1.133311e-05   9.171436e-01   9.971384e-01   9.999150e-01   0.000000e+00
##    crossStreet       latitude       longtide
##   8.194975e-01   0.000000e+00   0.000000e+00
```

```
sum(complete.cases(crimeLA))
```

```
## [1] 2
```

Data Processing

```
# Date

crimeLA <- crimeLA %>%
  mutate(dateReported = date(mdy_hms(dateReported)),
         dateOccurred = date(mdy_hms(dateOccurred)),
         timeOccurred =  format(strptime(timeOccurred,format = "%H%M"),’%H:%M’),
         hour = hour(hm(timeOccurred)), ## Extract hour from time
         )
```

1. Who is most vulnerable to be a victim of crime ?

Age distribution of victims

```
crimeLA %>% select(victAge) %>% table()
```

```
## .
##     0     2     3     4     5     6     7     8     9    10    11    12    13
## 43191    77    90   106   120   112   116   113   126   144   190   287   374
##    14    15    16    17    18    19    20    21    22    23    24    25    26
##   506   642   695   801  1107  3029  1988  2153  2592  2951  3167  3435  3585
##    27    28    29    30    31    32    33    34    35    36    37    38    39
##  3641  3748  3876  4105  3768  3589  3416  3462  3778  3188  3093  3076  2888
##    40    41    42    43    44    45    46    47    48    49    50    51    52
##  2796  2675  2458  2485  2349  2247  2123  2139  2177  2207  2631  2056  1943
##    53    54    55    56    57    58    59    60    61    62    63    64    65
##  1852  1918  1963  1803  1755  1646  1591  1551  1378  1379  1270  1188  1041
##    66    67    68    69    70    71    72    73    74    75    76    77    78
##   976   785   740   673   635   567   555   486   377   311   271   278   236
##    79    80    81    82    83    84    85    86    87    88    89    90    91
##   196   196   143   146   150   122   109    90    53    70    49    53    41
##    92    93    94    95    96    97    98    99   120
##    27    26    19    22    18    14    14    79     1
```

```
crimeLA[which(crimeLA$victAge == 0),"victAge"] <- NA
```

```
summary(crimeLA$victAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
##    2.00   28.00   37.00   39.68   50.00  120.00   43191
```

```
crimeLA %>%  filter(victAge == 120)
```

```
## # A tibble: 1 x 29
##   drNum dateReported dateOccurred timeOccurred areaID areaName rptDistNum part12
##   <chr> <date>       <date>       <chr>         <int> <chr>    <chr>       <int>
## 1 2008~ 2020-04-19   2020-04-19   21:45             8 West LA  0889            1
## # ... with 21 more variables: crimeCode <int>, crimeCodeDesc <chr>,
## #   moCode <chr>, victAge <int>, victGender <fct>, victDescent <fct>,
## #   premisCode <int>, premisDesc <chr>, weaponUsedCode <int>,
## #   weaponUSedDesc <chr>, status <chr>, statusDesc <chr>, crimeCode1 <chr>,
## #   crimeCode2 <chr>, crimeCode3 <chr>, crimeCode4 <chr>, location <chr>,
## #   crossStreet <chr>, latitude <dbl>, longtide <dbl>, hour <dbl>
```

Gender distribution of victims

```
crimeLA %>% select(victGender) %>%  table()
```

```
## .
##     F     M     X     H
## 63566 75332 14844    15
```

```
crimeLA %>%
  group_by(victGender) %>%
  summarise(count = n(),
            precentage = (count/nrow(crimeLA))*100 )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 3
##   victGender count precentage
##   <fct>      <int>      <dbl>
## 1 F          63566   36.0
## 2 M          75332   42.7
## 3 X          14844    8.41
## 4 H             15    0.00850
## 5 <NA>       22717   12.9
```

Distribution of victims' descent

```
crimeLA %>%  group_by(victDescent) %>%
  summarise(count= n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 20 x 2
##    victDescent count
##    <fct>       <int>
##  1 B           25180
##  2 H           54927
##  3 W           37182
##   4 X            16513
##   5 A             3912
##   6 O            14268
##   7 C              330
##   8 F              384
##   9 K              649
##  10 I               77
##  11 V               95
##  12 J              132
##  13 Z               25
##  14 P               31
##  15 U               19
##  16 S                8
##  17 D                5
##  18 G               14
##  19 L                4
##  20 <NA>         22719
```

Create Age groups

children, teenager, adult and elderly person

```
 crimeLA <- crimeLA %>%
  mutate(victAgeGroup = case_when(victAge <= 12 ~ ’children’,
                                  victAge >= 13   & victAge <= 19 ~ ’teenager’,
                                  victAge >=20 & victAge <=60 ~ ’adult’,
                                  victAge > 60 ~ ’elderlyPerson’))
```

```
crimeLA %>%  group_by(victAgeGroup) %>%
  summarise(count= n())
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 5 x 2
##   victAgeGroup   count
##   <chr>          <int>
## 1 adult         109864
## 2 children        1481
## 3 elderlyPerson  14784
## 4 teenager        7154
## 5 <NA>           43191
```

```
crimeLA %>% filter(!is.na(victAgeGroup), !is.na(victGender) , !is.na(victDescent) ) %>%
  group_by(victAgeGroup,victGender,victDescent) %>%
  summarise(count=n(),
            avgAge = mean(victAge),
            medianAge = median(victAge)) %>%
  arrange(desc(count))
```

```
## ‘summarise()‘ regrouping output by ’victAgeGroup’, ’victGender’ (override with ‘.groups‘ argument)
```

```
## # A tibble: 120 x 6
## # Groups:   victAgeGroup, victGender [14]
##    victAgeGroup  victGender victDescent count avgAge medianAge
##    <chr>         <fct>      <fct>       <int>  <dbl>     <dbl>
##  1 adult         M          H           23495   37.3        36
##  2 adult         F          H           22308   35.9        34
##  3 adult         M          W           15674   39.7        38
##  4 adult         F          W           12105   38.2        36
##  5 adult         F          B           11431   36.7        34
##  6 adult         M          B            9028   38.8        37
##  7 adult         M          O            5990   39.0        38
##  8 adult         F          O            3928   38.2        37
##  9 elderlyPerson M          W            3200   69.0        67
## 10 elderlyPerson F          W            2595   70.7        69
## # ... with 110 more rows
```

2. What are the most likely places to be a victim of crime?

```
crimeLA %>%
 group_by(premisDesc) %>%
 summarise(count= n()) %>%
 arrange(desc(count))   %>%
 head(n=10)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 10 x 2
##    premisDesc                                   count
##    <chr>                                        <int>
##  1 STREET                                       45512
##  2 SINGLE FAMILY DWELLING                       28963
##  3 MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC) 21016
##  4 PARKING LOT                                  13038
##  5 SIDEWALK                                      8615
##  6 OTHER BUSINESS                                7993
##  7 VEHICLE, PASSENGER/TRUCK                      6143
##  8 GARAGE/CARPORT                                3666
##  9 DRIVEWAY                                      3618
## 10 RESTAURANT/FAST FOOD                          2315
```

```
crimeLA %>% filter(!is.na(victAgeGroup), !is.na(victGender) , !is.na(victDescent), premisDesc=="SINGLE 
  group_by(victAgeGroup,victGender,victDescent) %>%
  summarise(count=n(),
            avgAge = mean(victAge),
            medianAge = median(victAge)) %>%
  arrange(desc(count))
```

```
## `summarise()` regrouping output by 'victAgeGroup', 'victGender' (override with `.groups` argument)
```

```
## # A tibble: 84 x 6
## # Groups:   victAgeGroup, victGender [12]
##    victAgeGroup  victGender victDescent count avgAge medianAge
##    <chr>         <fct>      <fct>       <int>  <dbl>     <dbl>
##  1 adult         F          H            5537   36.3        35
##  2 adult         M          H            3282   38.3        37
##  3 adult         M          W            3077   41.5        41
##  4 adult         F          W            2903   40.7        40
##  5 adult         F          B            2635   38.2        37
##  6 adult         M          B            1384   40.2        39
##  7 adult         M          O            1066   40.7        40
##  8 elderlyPerson F          W            1014   72.1        70
##  9 adult         F          O            1002   39.8        39
## 10 elderlyPerson M          W             997   70.8        69
## # ... with 74 more rows
```

3. What are the most common crimes in the city of Los Angeles ?

```
 crimeLA %>%
  filter(str_detect(location, "LOS ANGELES")) %>%
  group_by(crimeCodeDesc)%>%
  summarise(count=n())%>%
  arrange(desc(count))%>%
  head(10)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 10 x 2
##    crimeCodeDesc                                           count
##    <chr>                                                   <int>
##  1 BURGLARY FROM VEHICLE                                      55
##  2 VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)    44
##  3 BATTERY - SIMPLE ASSAULT                                   42
##  4 BURGLARY                                                   26
##  5 THEFT PLAIN - PETTY ($950 & UNDER)                         24
##  6 VANDALISM - MISDEAMEANOR ($399 OR UNDER)                   20
##  7 VEHICLE - STOLEN                                           20
##  8 ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT             17
##  9 BATTERY POLICE (SIMPLE)                                    16
## 10 ROBBERY                                                    13
```

4. What is the most dangerous day of the week ?

```
crimeLA = crimeLA %>%
  mutate(weekday = case_when(
    wday(dateReported)==1 ~ "Sunday",
    wday(dateReported)==2 ~ "Monday",
    wday(dateReported)==3 ~ "Tuesday",
    wday(dateReported)==4 ~ "Wednesday",
    wday(dateReported)==5 ~ "Thursday",
    wday(dateReported)==6 ~ "Friday",
    wday(dateReported)==7 ~ "Saturday"
  ))
```

```
crimeLA %>%
  group_by(weekday) %>%
  summarise(count=n()) %>%
  arrange(desc(count)) %>%
  head(1)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 1 x 2
##   weekday count
##   <chr>   <int>
## 1 Monday  27410
```

Monday is the most dangerous day of the week.

5. What is the most dangerous time to be on the street ? Does it change with day of the week ?

Group by only with time

```
crimeLA %>%
  group_by(timeOccurred) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(5)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 2
##   timeOccurred count
##   <chr>        <int>
## 1 12:00         6395
## 2 18:00         5182
## 3 17:00         4952
## 4 20:00         4814
## 5 19:00         4412
```

Group with weekday and time

```
n = crimeLA %>%
  group_by(weekday, timeOccurred) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
```

```
## `summarise()` regrouping output by 'weekday' (override with `.groups` argument)
```

```
for (i in c("Sunday","Monday","Tuesday","Wednesday","Thursday","Friday","Saturday")) {
n %>% filter(weekday == i) %>%
  head(1) %>%
  print(head(1))
}
```

```
## # A tibble: 1 x 3
## # Groups:   weekday [1]
##   weekday timeOccurred count
##   <chr>   <chr>        <int>
## 1 Sunday  20:00          658
## # A tibble: 1 x 3
## # Groups:   weekday [1]
##   weekday timeOccurred count
##   <chr>   <chr>        <int>
## 1 Monday  12:00         1025
## # A tibble: 1 x 3
## # Groups:   weekday [1]
##   weekday timeOccurred count
##   <chr>   <chr>        <int>
## 1 Tuesday 12:00         1069
## # A tibble: 1 x 3
## # Groups:   weekday [1]
##   weekday   timeOccurred count
##   <chr>     <chr>        <int>
## 1 Wednesday 12:00         1066
## # A tibble: 1 x 3
## # Groups:   weekday [1]
##   weekday  timeOccurred count
##   <chr>    <chr>        <int>
## 1 Thursday 12:00         1017
## # A tibble: 1 x 3
## # Groups:   weekday [1]
##   weekday timeOccurred count
##   <chr>   <chr>        <int>
## 1 Friday  12:00          920
## # A tibble: 1 x 3
## # Groups:   weekday [1]
##   weekday  timeOccurred count
##   <chr>    <chr>        <int>
## 1 Saturday 12:00          676
```

Considering with the day of the week the most dangerous time is changing. But 12:00 is the most dangerous time of all the days in the week except Sunday. For Sunday the most dangerous time is 20:00