```r
library(readr)
library(lubridate)
library(dplyr)
library(ggplot2)

# read the data
setwd("~/Documents/2019/1USC Courses/ISE 535 DMining/hw/hw1")
df0 = read_csv("crime.csv",col_names=TRUE)
#
# browse the data
names(df0)
```

```
##  [1] "Report Number"               "Occurred Date"
##  [3] "Occurred Time"               "Reported Date"
##  [5] "Reported Time"               "Crime Subcategory"
##  [7] "Primary Offense Description" "Precinct"
##  [9] "Sector"                      "Beat"
## [11] "Neighborhood"
```

```r
#
#
df = select(df0,ODate = "Occurred Date",RDate = "Reported Date",Category = "Crime Subcategory",
                description = 'Primary Offense Description',Sector,Beat,Neighborhood)
head(df)
```

```
## # A tibble: 6 x 7
##   ODate    RDate   Category        description    Sector Beat  Neighborhood
##   <chr>    <chr>   <chr>           <chr>          <chr>  <chr> <chr>
## 1 12/13/1... 12/13/... DUI             DUI-LIQUOR     G      G2    CENTRAL AREA/SQU...
## 2 06/15/1... 06/15/... FAMILY OFFENSE... CHILD-OTHER    Q      Q2    QUEEN ANNE
## 3 01/01/1... 01/25/... SEX OFFENSE-OT... SEXOFF-OTHER   N      N2    NORTHGATE
## 4 06/01/1... 09/09/... SEX OFFENSE-OT... SEXOFF-OTHER   <NA>   <NA>  UNKNOWN
## 5 01/01/1... 08/11/... SEX OFFENSE-OT... SEXOFF-OTHER   <NA>   <NA>  UNKNOWN
## 6 12/16/1... 12/16/... BURGLARY-RESID... BURGLARY-FORC... R      R3    LAKEWOOD/SEWARD ...
```

```r
#
dim(df)
```

```
## [1] 481376      7
```

```r
#
# 1) number of neighbors and number of crime categories
#
length(table(df$Neighborhood))
```

```
## [1] 59
```

```r
# there are 59 neighborhoods
#
table(df$Category)
```

```
##
##                   AGGRAVATED ASSAULT               AGGRAVATED ASSAULT-DV
##                                13954                                6307
##                               ARSON                 BURGLARY-COMMERCIAL
##                                1009                               21274
##  BURGLARY-COMMERCIAL-SECURE PARKING                BURGLARY-RESIDENTIAL
```

```
##                                1042                                43908
## BURGLARY-RESIDENTIAL-SECURE PARKING                          CAR PROWL
##                                7667                               137766
##                  DISORDERLY CONDUCT                                  DUI
##                                 245                                11849
##           FAMILY OFFENSE-NONVIOLENT                               GAMBLE
##                                6372                                   17
##                            HOMICIDE                  LIQUOR LAW VIOLATION
##                                 250                                 1588
##                           LOITERING                  MOTOR VEHICLE THEFT
##                                  82                                40362
##                            NARCOTIC                           PORNOGRAPHY
##                               16428                                 154
##                         PROSTITUTION                                RAPE
##                                3503                                1758
##                  ROBBERY-COMMERCIAL                  ROBBERY-RESIDENTIAL
##                                4103                                 988
##                      ROBBERY-STREET                    SEX OFFENSE-OTHER
##                               11096                                5783
##                     THEFT-ALL OTHER                         THEFT-BICYCLE
##                               49624                               10206
##                      THEFT-BUILDING                       THEFT-SHOPLIFT
##                               19718                               44768
##                            TRESPASS                              WEAPON
##                               14733                                4560
```

```
#



length(table(df$Category))
```

```
## [1] 30
```

```
# there are 30 crime identified categories
#
# 2) number of crimes in each neighborhood
#
table(df$Neighborhood)
```

```
##
##                   ALASKA JUNCTION                             ALKI
##                              6378                             2335
##                     BALLARD NORTH                    BALLARD SOUTH
##                             10155                            14031
##                          BELLTOWN                       BITTERLAKE
##                             14437                             9227
##                   BRIGHTON/DUNLAP                     CAPITOL HILL
##                              6608                            28296
##          CENTRAL AREA/SQUIRE PARK CHINATOWN/INTERNATIONAL DISTRICT
##                             11361                            13627
##           CLAREMONT/RAINIER VISTA                     COLUMBIA CITY
##                              2303                             3126
##               COMMERCIAL DUWAMISH         COMMERCIAL HARBOR ISLAND
##                               294                              152
##               DOWNTOWN COMMERCIAL                 EASTLAKE - EAST
```

```
##                           45127                           807
##                  EASTLAKE - WEST                 FAUNTLEROY SW
##                            3253                          2176
##                      FIRST HILL                       FREMONT
##                            12826                         9057
##                        GENESEE                     GEORGETOWN
##                            1462                          5425
##                       GREENWOOD                    HIGH POINT
##                            10570                         3204
##                    HIGHLAND PARK                  HILLMAN CITY
##                            5113                          2484
##   JUDKINS PARK/NORTH BEACON HILL                     LAKECITY
##                            4163                          12474
##             LAKEWOOD/SEWARD PARK                  MADISON PARK
##                            3238                          1690
##                   MADRONA/LESCHI                     MAGNOLIA
##                            6083                          7152
##                  MID BEACON HILL                   MILLER PARK
##                            5490                          3156
##             MONTLAKE/PORTAGE BAY                       MORGAN
##                            3478                          4138
##                      MOUNT BAKER                    NEW HOLLY
##                            6321                          3001
##                    NORTH ADMIRAL              NORTH BEACON HILL
##                            4577                          8610
##                   NORTH DELRIDGE                    NORTHGATE
##                            3299                          28480
##                    PHINNEY RIDGE                  PIGEON POINT
##                            4331                           561
##                   PIONEER SQUARE                    QUEEN ANNE
##                            8124                          25172
##                    RAINIER BEACH                  RAINIER VIEW
##                            5470                          5117
##                ROOSEVELT/RAVENNA   ROXHILL/WESTWOOD/ARBOR HEIGHTS
##                            17673                         7542
##                       SANDPOINT                   SLU/CASCADE
##                            9859                          21630
##                            SODO              SOUTH BEACON HILL
##                            7781                          2086
##                   SOUTH DELRIDGE                   SOUTH PARK
##                            1787                          3676
##                      UNIVERSITY                      UNKNOWN
##                            19167                         3026
##                     WALLINGFORD
##                            9190
```

```
# most dangeous neighborhodd
index = which.max(table(df$Neighborhood))
table(df$Neighborhood)[index]
```

```
## DOWNTOWN COMMERCIAL
##               45127
```

```
#
# 3) pipe
```

```
#
df3 = read_csv("crime.csv",col_names=TRUE) %>%
  select(ODate = "Occurred Date",RDate = "Reported Date",Category = "Crime Subcategory",
         description = 'Primary Offense Description',Sector,Beat,Neighborhood)%>%
  summarize(n_distinct(Neighborhood),n_distinct(Category))
df3
```

```
## # A tibble: 1 x 2
##   `n_distinct(Neighborhood)` `n_distinct(Category)`
##                        <int>                  <int>
## 1                         59                     31
```

```
# there are 59 neighborhoods and 30 crime categories
#
# use n() function to get a count of number of records in each group
#
df3 = read_csv("crime.csv",col_names=TRUE) %>%
  select(ODate = "Occurred Date",RDate = "Reported Date",Category = "Crime Subcategory",
         description = 'Primary Offense Description',Sector,Beat,Neighborhood) %>%
  group_by(Neighborhood) %>%
  summarize(category=n()) %>%
  filter(category == max(category))
df3
```

```
## # A tibble: 1 x 2
##   Neighborhood        category
##   <chr>                  <int>
## 1 DOWNTOWN COMMERCIAL    45127
```

```
#
# work out of the pipe to find most dangerous
names(df3)[2] = 'crimes'
index = which.max(df3$crimes)
df3[index,]
```

```
## # A tibble: 1 x 2
##   Neighborhood        crimes
##   <chr>                <int>
## 1 DOWNTOWN COMMERCIAL  45127
```

```
#
# 4) most frequent crime category in Queen Anne
#
dfQA = filter(df,Neighborhood == 'QUEEN ANNE')
dtemp = group_by(dfQA,Category)
dfmax = summarize(dtemp,crimes = n())
head(dfmax)
```

```
## # A tibble: 6 x 2
##   Category                           crimes
##   <chr>                               <int>
## 1 AGGRAVATED ASSAULT                    402
## 2 AGGRAVATED ASSAULT-DV                 189
## 3 ARSON                                  45
## 4 BURGLARY-COMMERCIAL                  1359
## 5 BURGLARY-COMMERCIAL-SECURE PARKING    125
## 6 BURGLARY-RESIDENTIAL                 2193
```

```r
#
# 30 categories, each row has n. of crimes by category
#
# find most frequent category
index = which.max(dfmax$crimes)
dfmax[index,]
```

```
## # A tibble: 1 x 2
##   Category  crimes
##   <chr>      <int>
## 1 CAR PROWL  10115
```

```r
#
# or
dfsort = arrange(dfmax,desc(crimes))
head(dfsort)
```

```
## # A tibble: 6 x 2
##   Category             crimes
##   <chr>                 <int>
## 1 CAR PROWL             10115
## 2 MOTOR VEHICLE THEFT    2284
## 3 THEFT-ALL OTHER        2225
## 4 BURGLARY-RESIDENTIAL   2193
## 5 BURGLARY-COMMERCIAL    1359
## 6 THEFT-SHOPLIFT         1265
```

```r
#
# 5) crimes per month
#
# extract month
df5 = df
head(df5)
```

```
## # A tibble: 6 x 7
##   ODate    RDate    Category        description    Sector Beat  Neighborhood
##   <chr>    <chr>    <chr>           <chr>          <chr>  <chr> <chr>
## 1 12/13/1... 12/13/... DUI             DUI-LIQUOR     G      G2    CENTRAL AREA/SQU...
## 2 06/15/1... 06/15/... FAMILY OFFENSE... CHILD-OTHER    Q      Q2    QUEEN ANNE
## 3 01/01/1... 01/25/... SEX OFFENSE-OT... SEXOFF-OTHER   N      N2    NORTHGATE
## 4 06/01/1... 09/09/... SEX OFFENSE-OT... SEXOFF-OTHER   <NA>   <NA>  UNKNOWN
## 5 01/01/1... 08/11/... SEX OFFENSE-OT... SEXOFF-OTHER   <NA>   <NA>  UNKNOWN
## 6 12/16/1... 12/16/... BURGLARY-RESID... BURGLARY-FORC... R      R3    LAKEWOOD/SEWARD ...
```

```r
df5$RDate = as.Date(df5$RDate,format = '%m/%d/%Y')
df5$RMonth = month(df5$RDate)
head(df5)
```

```
## # A tibble: 6 x 8
##   ODate  RDate      Category     description  Sector Beat  Neighborhood   RMonth
##   <chr>  <date>     <chr>        <chr>        <chr>  <chr> <chr>           <dbl>
## 1 12/13... 2008-12-13 DUI          DUI-LIQUOR   G      G2    CENTRAL AREA/...     12
## 2 06/15... 2010-06-15 FAMILY OFFE... CHILD-OTHER  Q      Q2    QUEEN ANNE           6
## 3 01/01... 2012-01-25 SEX OFFENSE... SEXOFF-OTHER N      N2    NORTHGATE            1
## 4 06/01... 2013-09-09 SEX OFFENSE... SEXOFF-OTHER <NA>   <NA>  UNKNOWN              9
## 5 01/01... 2016-08-11 SEX OFFENSE... SEXOFF-OTHER <NA>   <NA>  UNKNOWN              8
```

```
## 6 12/16... 1975-12-16 BURGLARY-RE... BURGLARY-FO... R      R3    LAKEWOOD/SEWA...     12
```

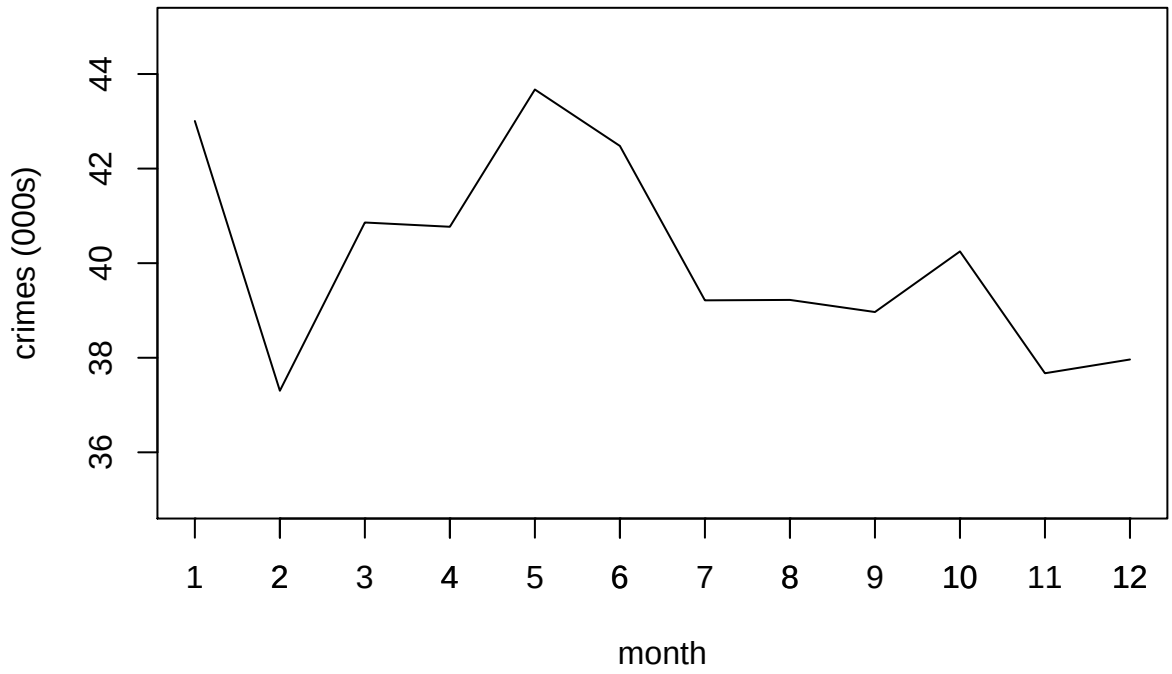
```
#
# group by month
dtemp = group_by(df5,RMonth)
dmonth = summarise(dtemp,n=n())
head(dmonth)
```

```
## # A tibble: 6 x 2
##   RMonth     n
##    <dbl> <int>
## 1      1 43006
## 2      2 37302
## 3      3 40860
## 4      4 40770
## 5      5 43672
## 6      6 42479
```

```
# most dangerous
index = which.max(dmonth$n)
month.abb[index]
```

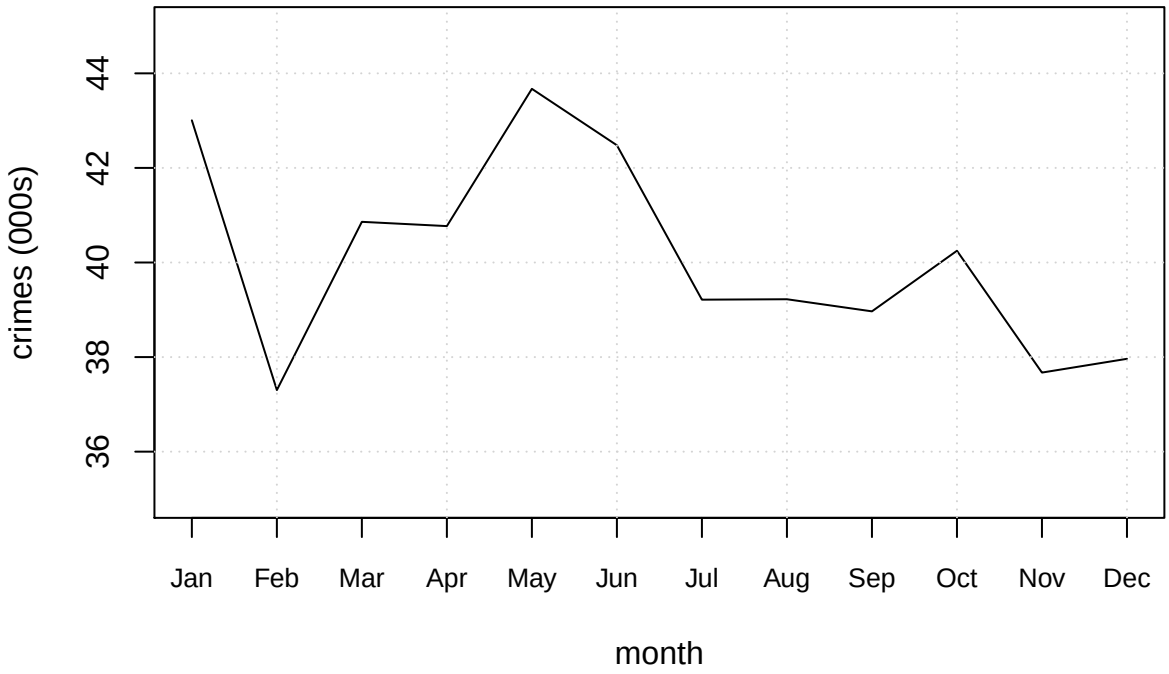
```
## [1] "May"
```

```
#
# 6) lineplot crimes vs time
#
aux = dmonth$n/1000
plot(aux~dmonth$RMonth,type = 'l',xlab='month',ylab='crimes (000s)',ylim=c(35,45))
for(i in 1:12) axis(1, i)
```




```
# or make plot with no x-axis tick marks, as given by xaxt
plot(aux~dmonth$RMonth,type = 'l',xlab='month',ylab='crimes (000s)',ylim=c(35,45), xaxt='n')
# add month.abb as tick marks scaled 80% font size
axis(1, at = 1:12, label = month.abb[dmonth$RMonth],cex.axis = 0.8)
```

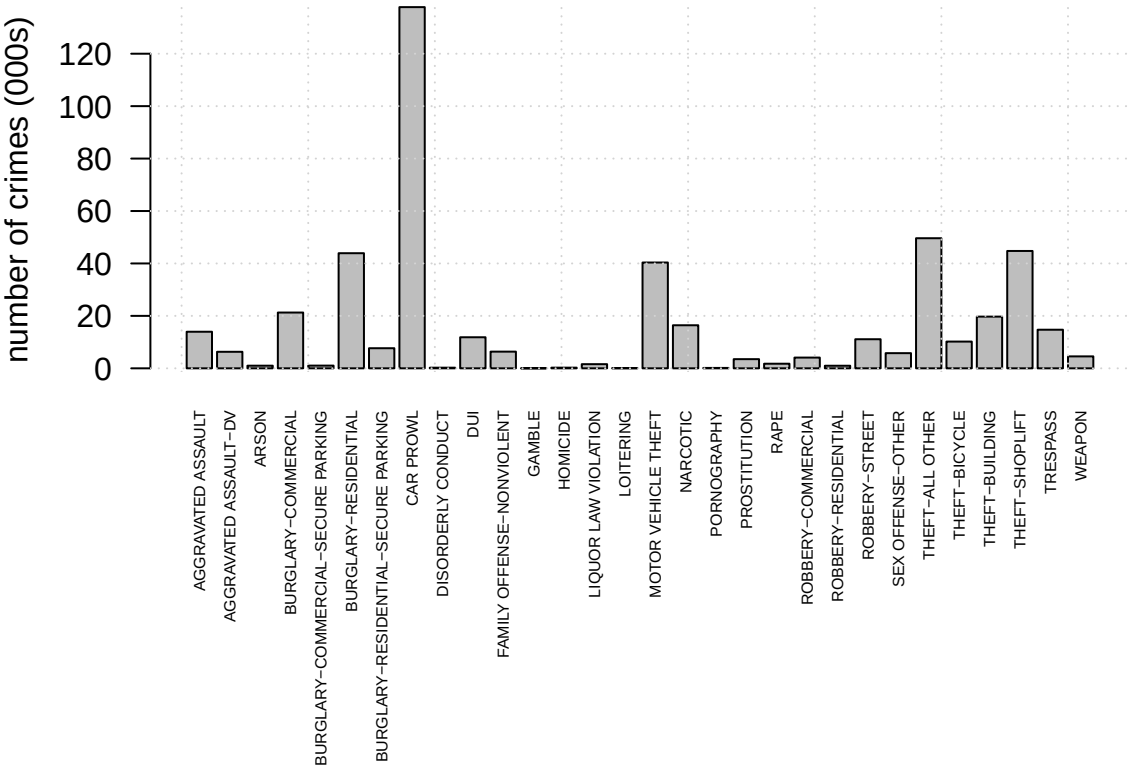
```
grid()
```




```
#
# 8) vertical boxplots
#
dtemp = group_by(df,Category)
dfby_category = summarize(dtemp,crimes = n())
dfby_category2 = arrange(dfby_category,desc(crimes))
head(dfby_category2)
```

```
## # A tibble: 6 x 2
##   Category             crimes
##   <chr>                 <int>
## 1 CAR PROWL            137766
## 2 THEFT-ALL OTHER       49624
## 3 THEFT-SHOPLIFT        44768
## 4 BURGLARY-RESIDENTIAL  43908
## 5 MOTOR VEHICLE THEFT   40362
## 6 BURGLARY-COMMERCIAL   21274
```

```
dfby_category2$crimes = dfby_category2$crimes/1000
# adjust margins to fit xaxis labels
par(mar=c(11,6,2,1))
barplot(crimes~Category,dfby_category2,las=2,cex.names=0.5,
        ylab='number of crimes (000s)',xlab='')
grid()
```


number of crimes (000s)
0
20
40
60
80
100
120
AGGRAVATED ASSAULT
AGGRAVATED ASSAULT−DV
ARSON
BURGLARY−COMMERCIAL
BURGLARY−COMMERCIAL−SECURE PARKING
BURGLARY−RESIDENTIAL
BURGLARY−RESIDENTIAL−SECURE PARKING
CAR PROWL
DISORDERLY CONDUCT
DUI
FAMILY OFFENSE−NONVIOLENT
GAMBLE
HOMICIDE
LIQUOR LAW VIOLATION
LOITERING
MOTOR VEHICLE THEFT
NARCOTIC
PORNOGRAPHY
PROSTITUTION
RAPE
ROBBERY−COMMERCIAL
ROBBERY−RESIDENTIAL
ROBBERY−STREET
SEX OFFENSE−OTHER
THEFT−ALL OTHER
THEFT−BICYCLE
THEFT−BUILDING
THEFT−SHOPLIFT
TRESPASS
WEAPON