# Lab 8

*Johnathan Hsu*

*October 27, 2017*

Using your own dataset (which may include more than one table) carry out the following data cleaning steps. Knit together the PDF document and commit both the Lab 8 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

**Before you begin: as many of you have large datasets, you're going to want to select only the variables you're interested in utilizing for this project (ideally no more than twenty columns but perhaps much smaller) so you don't have R Studio's memory working on the entire dataset. The example code provided below can be modified to allow you to subset your data to only the variables you wish to use. First, read in your complete dataset and save it as `data`. Then, add the names of the variables you wish to use for your poster project to the `select` function, separated by commas. Run the two lines of code to save this new, smaller version of your data to `data_subset`. Use this smaller dataset to complete the rest of the lab**

```
# Adding dplyr to the library for select function
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.3.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```
# loading the csv file.
data_95_96 <- read.csv("clean_data/initial_data/95-96.csv")

# add year to the data in case I want to merge later
data_95_96$YEAR <- 1996
# Selecting the variables that I'm using

data_subset_95_96 <- data_95_96 %>%
  select(MONRACE, YEAR, MONSEX, EDUCATN, AGE, STATMIN, STATMAX, DISPOSIT, TOTPRISN, XFOLSOR, XCRHISSR,
```

1. To get a feel for its structure, look at the class, dimensions, column names, structure, and basic summary statistics of your data.

I will test the 1995-1996 data, and uniformize the variables to factors as needed

```
class(data_subset_95_96)
```

```
## [1] "data.frame"
```

```
dim(data_subset_95_96)
```

```
## [1] 42436    13
```

```r
colnames(data_subset_95_96)
```

```
##  [1] "MONRACE"  "YEAR"     "MONSEX"   "EDUCATN"  "AGE"      "STATMIN" 
##  [7] "STATMAX"  "DISPOSIT" "TOTPRISN" "XFOLSOR"  "XCRHISSR" "DISTRICT"
## [13] "MONCIRC"
```

```r
str(data_subset_95_96)
```

```
## 'data.frame':    42436 obs. of  13 variables:
##  $ MONRACE : Factor w/ 6 levels "American Indian or Alaskan Native",..: 3 3 3 6 6 6 6 6 6 6 ...
##  $ YEAR    : num  1996 1996 1996 1996 1996 ...
##  $ MONSEX  : Factor w/ 3 levels "Female","Male",..: 1 2 2 2 2 1 1 2 2 2 ...
##  $ EDUCATN : Factor w/ 29 levels "Associate degree (AA)",..: 24 13 23 4 24 13 4 7 7 13 ...
##  $ AGE     : Factor w/ 69 levels "16","17","18",..: 5 39 6 10 20 6 32 6 30 8 ...
##  $ STATMIN : Factor w/ 28 levels "0","1","12","120",..: 4 1 1 18 1 1 1 1 1 18 ...
##  $ STATMAX : Factor w/ 124 levels "100","105","108",..: 69 31 30 61 45 76 30 76 76 61 ...
##  $ DISPOSIT: Factor w/ 6 levels "Both guilty plea and trial",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ TOTPRISN: Factor w/ 358 levels "1","10","100",..: 24 53 333 214 135 242 1 268 53 324 ...
##  $ XFOLSOR : Factor w/ 57 levels "1","10","11",..: 16 4 20 16 7 7 5 53 2 16 ...
##  $ XCRHISSR: Factor w/ 8 levels "1","2","3","4",..: 3 4 3 3 1 5 1 1 3 4 ...
##  $ DISTRICT: int  23 45 25 4 70 70 70 41 46 42 ...
##  $ MONCIRC : int  4 6 4 1 9 9 9 5 6 5 ...
```

```r
summary(data_subset_95_96)
```

```
##                                MONRACE           YEAR     
##  American Indian or Alaskan Native:  681   Min.   :1996  
##  Asian or Pacific Islander        : 1265   1st Qu.:1996  
##  Black                            :12562   Median :1996  
##  Missing or Indeterminable        : 2027   Mean   :1996  
##  Other                            :   57   3rd Qu.:1996  
##  White/Caucasian                  :25844   Max.   :1996  
##                                                          
##                       MONSEX                                EDUCATN     
##  Female                   : 6540   High school graduate           : 7728  
##  Male                     :35893   Some college                   : 6451  
##  Missing or Indeterminable:    3   GED (General Education Diploma): 3873  
##                                    Nine yr school completed       : 2761  
##                                    Ten yr school completed        : 2671  
##                                    Eleven yr school compl         : 2624  
##                                    (Other)                        :16328  
##       AGE                            STATMIN     
##  25     : 1724   0                        :29394  
##  24     : 1713   120                      : 5871  
##  26     : 1694   60                       : 5615  
##  27     : 1687   180                      :  570  
##  23     : 1625   240                      :  359  
##  31     : 1592   Missing or Indeterminable:  195  
##  (Other):32401   (Other)                  :  432  
##              STATMAX                                DISPOSIT    
##  60               : 7454   Both guilty plea and trial   :   32  
##  Life imprisonment: 7091   Guilty plea                  :38842  
##  240              : 5083   Jury trial                   : 3522  
##  120              : 5024   Missing or Indeterminable    :    6  
##  480              : 4603   Nolo contendere              :   27  
##  12               : 2262   Trial by judge or bench trial:    7
##  (Other)          :10919
##                            TOTPRISN                          XFOLSOR
##  No prison or - 1 month ordered: 8839   21                       : 3097
##  24                            : 2263   Missing or Indeterminable: 2829
##  12                            : 1703   10                       : 2509
##  60                            : 1424   13                       : 2024
##  18                            : 1421   23                       : 1926
##  30                            : 1352   4                        : 1633
##  (Other)                       :25434   (Other)                  :28418
##     XCRHISSR        DISTRICT        MONCIRC
##  1      :22205   Min.   : 0.00   Min.   : 0.000
##  3      : 5006   1st Qu.:22.00   1st Qu.: 4.000
##  2      : 4283   Median :41.00   Median : 6.000
##  6      : 3599   Mean   :43.68   Mean   : 6.445
##  9      : 2795   3rd Qu.:70.00   3rd Qu.: 9.000
##  4      : 2695   Max.   :96.00   Max.   :11.000
##  (Other): 1853
```

2. Preview the first and last 15 rows of your data. Is you dataset tidy? If not, what principles of tidy data does it seem to be violating?

```
# Calling head and tail, the data appears to be tidy
head(data_subset_95_96, 15)
```

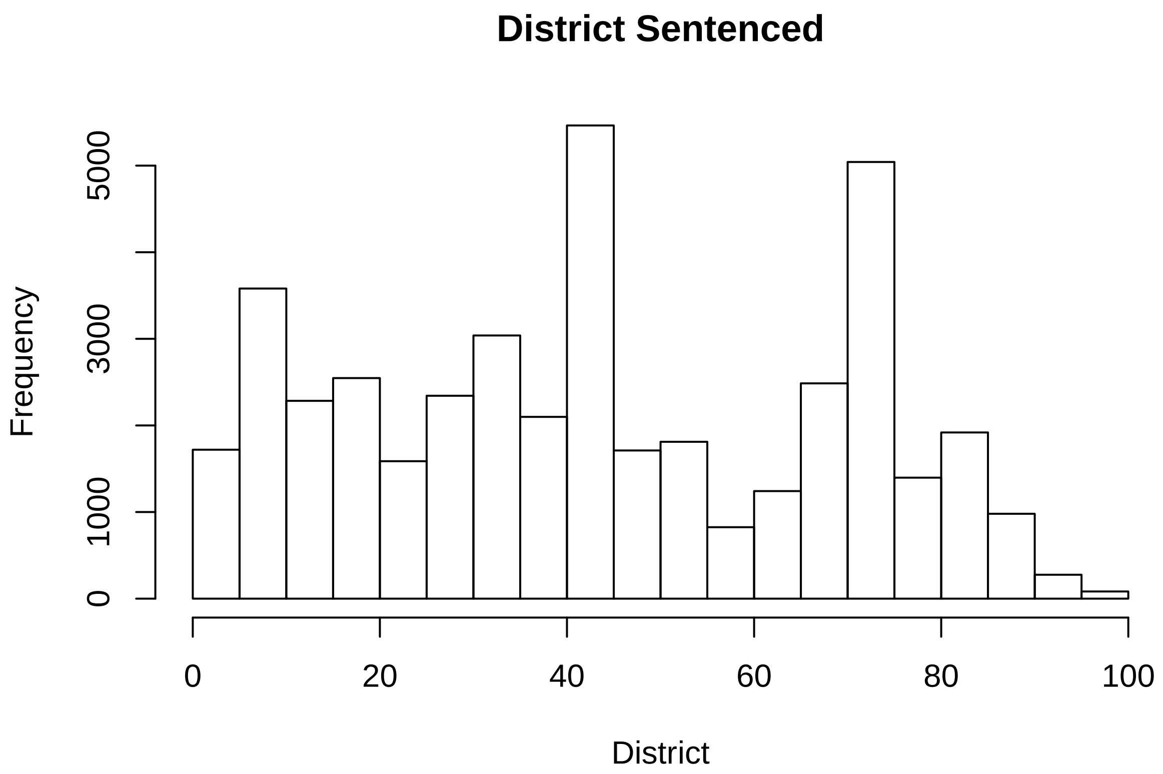
```
##                       MONRACE YEAR MONSEX                         EDUCATN
## 1                       Black 1996 Female         Ten yr school completed
## 2                       Black 1996   Male        Nine yr school completed
## 3                       Black 1996   Male Some trade or vocational school
## 4             White/Caucasian 1996   Male          Eleven yr school compl
## 5             White/Caucasian 1996   Male         Ten yr school completed
## 6             White/Caucasian 1996 Female        Nine yr school completed
## 7             White/Caucasian 1996 Female          Eleven yr school compl
## 8             White/Caucasian 1996   Male GED (General Education Diploma)
## 9             White/Caucasian 1996   Male GED (General Education Diploma)
## 10            White/Caucasian 1996   Male        Nine yr school completed
## 11            White/Caucasian 1996   Male            High school graduate
## 12                      Black 1996 Female          Eleven yr school compl
## 13            White/Caucasian 1996   Male            High school graduate
## 14 Missing or Indeterminable 1996   Male          Eleven yr school compl
## 15            White/Caucasian 1996   Male          Eleven yr school compl
##    AGE STATMIN STATMAX    DISPOSIT                       TOTPRISN
## 1   20     120     540 Guilty plea                            120
## 2   54       0     252 Guilty plea                             15
## 3   21       0     240 Guilty plea                             87
## 4   25      60     480 Guilty plea                             36
## 5   35       0      36 Guilty plea                             24
## 6   21       0      60 Guilty plea                             42
## 7   47       0     240 Guilty plea                              1
## 8   21       0      60 Guilty plea                              5
## 9   45       0      60 Guilty plea                             15
## 10  23      60     480 Guilty plea                             80
## 11  22       0     120 Guilty plea                              5
## 12  32       0      12 Guilty plea No prison or - 1 month ordered
## 13  32       0     480 Guilty plea                            151
## 14  26       0      12 Guilty plea No prison or - 1 month ordered
## 15  29       0     120 Guilty plea No prison or - 1 month ordered
##                      XFOLSOR XCRHISSR DISTRICT MONCIRC
## 1                         23        3       23       4
## 2                         12        4       45       6
## 3                         27        3       25       4
## 4                         23        3        4       1
## 5                         15        1       70       9
## 6                         15        5       70       9
## 7                         13        1       70       9
## 8                          9        1       41       5
## 9                         10        3       46       6
## 10                        23        4       42       5
## 11                         9        1       35       5
## 12 Missing or Indeterminable        9        8       2
## 13                        32        3        8       2
## 14 Missing or Indeterminable        9        8       2
## 15                        10        1        8       2
```

```
tail(data_subset_95_96, 15)
```

```
##                                 MONRACE YEAR MONSEX
## 42422                             Black 1996   Male
## 42423                   White/Caucasian 1996 Female
## 42424 American Indian or Alaskan Native 1996   Male
## 42425                             Black 1996   Male
## 42426                             Black 1996   Male
## 42427                   White/Caucasian 1996   Male
## 42428                   White/Caucasian 1996   Male
## 42429                   White/Caucasian 1996   Male
## 42430                   White/Caucasian 1996   Male
## 42431                   White/Caucasian 1996   Male
## 42432         Missing or Indeterminable 1996   Male
## 42433                             Black 1996   Male
## 42434         Asian or Pacific Islander 1996   Male
## 42435                   White/Caucasian 1996   Male
## 42436                   White/Caucasian 1996   Male
##                               EDUCATN AGE STATMIN STATMAX    DISPOSIT
## 42422            High school graduate  30      60     480 Guilty plea
## 42423           Associate degree (AA)  36       0      12 Guilty plea
## 42424          Eleven yr school compl  29       0      12 Guilty plea
## 42425          Eleven yr school compl  22       0     240 Guilty plea
## 42426 GED (General Education Diploma)  22       0     240 Guilty plea
## 42427        Nine yr school completed  25       0      24 Guilty plea
## 42428           Eight yr school compl  25       0      24 Guilty plea
## 42429       Missing or Indeterminable  28       0      24 Guilty plea
## 42430       Missing or Indeterminable  26       0      24 Guilty plea
## 42431           Associate degree (AA)  19       0      12 Guilty plea
## 42432       Missing or Indeterminable  32       0      24 Guilty plea
## 42433                    Some college  35       0     240 Guilty plea
## 42434            High school graduate  35       0      60 Guilty plea
## 42435                    Some college  34       0     240 Guilty plea
## 42436            High school graduate  34       0      60 Guilty plea
##                             TOTPRISN                   XFOLSOR XCRHISSR
## 42422      Missing or Indeterminable Missing or Indeterminable        9
## 42423 No prison or - 1 month ordered                         9        1
## 42424 No prison or - 1 month ordered Missing or Indeterminable        9
## 42425                            66                         21        5
## 42426      Missing or Indeterminable Missing or Indeterminable        9
## 42427                            24 Missing or Indeterminable        9
## 42428                            24 Missing or Indeterminable        9
## 42429                            24 Missing or Indeterminable        9
## 42430                            24 Missing or Indeterminable        9
## 42431 No prison or - 1 month ordered Missing or Indeterminable        9
## 42432                            24                         21        6
## 42433                           151 Missing or Indeterminable        9
## 42434 No prison or - 1 month ordered Missing or Indeterminable        9
## 42435                            28                         22        1
## 42436 No prison or - 1 month ordered Missing or Indeterminable        9
##       DISTRICT MONCIRC
## 42422       47       6
## 42423       78       9
## 42424       88      10
## 42425       78       9
## 42426       39       5
## 42427       79       9
## 42428       79       9
## 42429       79       9
## 42430       79       9
## 42431       10       2
## 42432       74       9
## 42433       73       9
## 42434       73       9
## 42435       73       9
## 42436       73       9
```

3. Create a histogram for at least two variables you plan to focus on for your study. Describe what these plots show you about these variables.

```
# Calling histogram for number of sentences per district for the data from 1995-1996.
hist(data_subset_95_96$DISTRICT, main = "District Sentenced", xlab = "District")
```

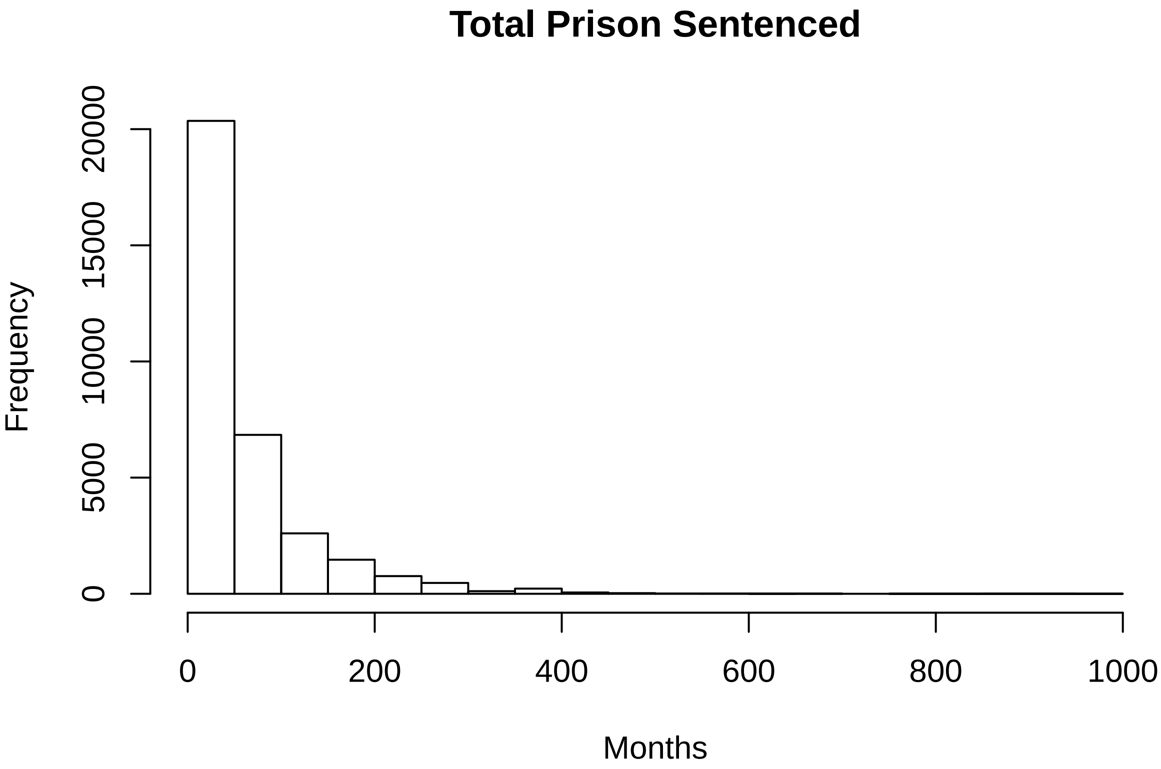


```
# Calling historgram for number for the length of the sentence. Note that because the letters are
# currently factors (e.g. less than 1 day of imprisonment), I will be switching them from factors
# to numeric, and excluding these observations (hence the "NAs introduced by coercion warning).

data_subset_95_96$TOTPRISN <- as.numeric(paste(data_subset_95_96$TOTPRISN))
```

```
## Warning: NAs introduced by coercion
```

```
hist(data_subset_95_96$TOTPRISN, main = "Total Prison Sentenced", xlab = "Months")
```



4. Create at least one bivariate plot showing the relationship between two variables of interest. What does/do the(se) plot(s) tell you about the association between these two variables?

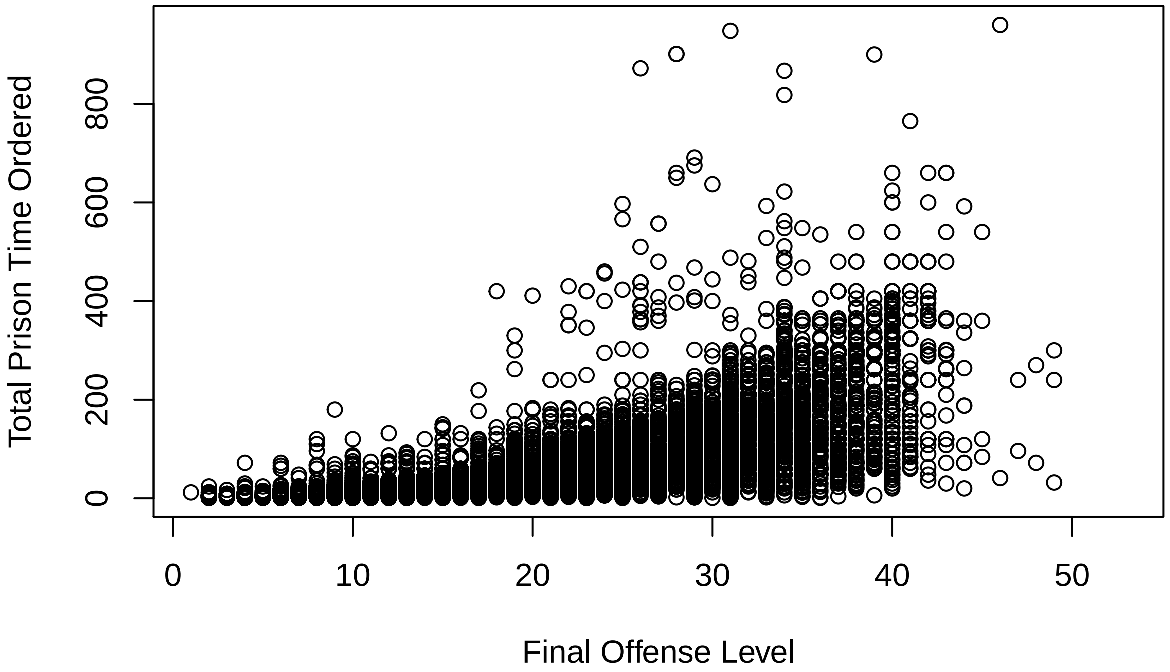
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
data_subset_95_96$XFOLSOR <- as.numeric(paste(data_subset_95_96$XFOLSOR))
```

```
## Warning: NAs introduced by coercion
```

```
plot(y = data_subset_95_96$TOTPRISN, x = data_subset_95_96$XFOLSOR, xlab = "Final Offense Level", ylab =
```



5. Load the `tidyr` package. Do all of your columns correspond to variables? Do any columns represent multiple variables? If your answer is yes to either question, carry out the appropriate `tidyr` function (`gather()` or `spread()` respectively) to tidy your data. # Not applicable.

My data columns appear to be fine.

```
install.packages("tidyr", repos = "http://cran.us.r-project.org")
```

```
## Installing package into '/Users/hsujohnathan/Library/R/3.3/library'
## (as 'lib' is unspecified)

##
## The downloaded binary packages are in
##  /var/folders/_v/bp2t8nmd09xb1l7j2mqg88l40000gn/T//RtmpeUdsda/downloaded_packages
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.3.2
```

6. Do any columns need to be separated into two or more? Do any columns need to be combined into one? If so, carry out the appropriate the appropriate `tidyr` function (`separate()` or `unite()` respectively) to tidy your data.

Not applicable.

**At this stage each row in your data should represent one observation, each column should be a variable, and each table should be observational unit.**

7. What is the class of each of the variables in your analysis? Are these classes appropriate for the type of measurement they purport to capture? Explain your reasoning.

This is an issue with my dataset. Currently, a lot of these variables are in factors (because there are special cases). I will be introducing NAs through coersion.

8. Do any of your variables need to be coerced into a different data type? If so, carry out the appropriate coercion methods below. (This includes transformation of any date objects using the `lubridate` package)

Nope.

9. Are there any strings you need to manipulate for your analysis? If so, use the appropriate function from the `stringr` package.

Nope.

10. Do you have any missing values in your dataset? How many and how are they coded? **Be sure to look out for specific codebook values for missing values (i.e. `-1` for `NA`) as well as empty strings or other software-specific values for `NA`.** Don't worry about removing NAs yet - we'll tackle this question later once discern whether they're random or systematically distributed.

These are different every year. There are a bunch of NAs in the dataset though.

```r
summary(data_subset_95_96)
```

```
##                                MONRACE           YEAR     
##  American Indian or Alaskan Native:  681   Min.   :1996  
##  Asian or Pacific Islander        : 1265   1st Qu.:1996  
##  Black                            :12562   Median :1996  
##  Missing or Indeterminable        : 2027   Mean   :1996  
##  Other                            :   57   3rd Qu.:1996  
##  White/Caucasian                  :25844   Max.   :1996  
##                                                          
##                         MONSEX                                 EDUCATN     
##  Female                   : 6540   High school graduate           : 7728  
##  Male                     :35893   Some college                   : 6451  
##  Missing or Indeterminable:    3   GED (General Education Diploma): 3873  
##                                    Nine yr school completed       : 2761  
##                                    Ten yr school completed        : 2671  
##                                    Eleven yr school compl         : 2624  
##                                    (Other)                        :16328  
##       AGE                             STATMIN     
##  25     : 1724   0                        :29394  
##  24     : 1713   120                      : 5871  
##  26     : 1694   60                       : 5615  
##  27     : 1687   180                      :  570  
##  23     : 1625   240                      :  359  
##  31     : 1592   Missing or Indeterminable:  195  
##  (Other):32401   (Other)                  :  432  
##                STATMAX                               DISPOSIT    
##  60               : 7454   Both guilty plea and trial   :   32  
##  Life imprisonment: 7091   Guilty plea                  :38842  
##  240              : 5083   Jury trial                   : 3522  
##  120              : 5024   Missing or Indeterminable    :    6  
##  480              : 4603   Nolo contendere              :   27  
##  12               : 2262   Trial by judge or bench trial:    7  
##  (Other)          :10919                                        
##     TOTPRISN          XFOLSOR         XCRHISSR        DISTRICT    
##  Min.   :  1.00   Min.   : 1.00   1      :22205   Min.   : 0.00  
##  1st Qu.: 15.00   1st Qu.:10.00   3      : 5006   1st Qu.:22.00  
##  Median : 34.00   Median :18.00   2      : 4283   Median :41.00  
##  Mean   : 59.29   Mean   :18.44   6      : 3599   Mean   :43.68  
##  3rd Qu.: 72.00   3rd Qu.:25.00   9      : 2795   3rd Qu.:70.00  
##  Max.   :960.00   Max.   :53.00   4      : 2695   Max.   :96.00  
##  NA's   :9483     NA's   :3034    (Other): 1853                   
##     MONCIRC
##  Min.   : 0.000
##  1st Qu.: 4.000
##  Median : 6.000
##  Mean   : 6.445
##  3rd Qu.: 9.000
##  Max.   :11.000
##
```

11. Are there any special values in your dataset? If so, what are they and how do you think they got there? *The presence of special values is less likely if you haven't performed any data manipulation yet so you should remember to return to this step each time you carry out a mathematical transformation of any values in your dataset.*

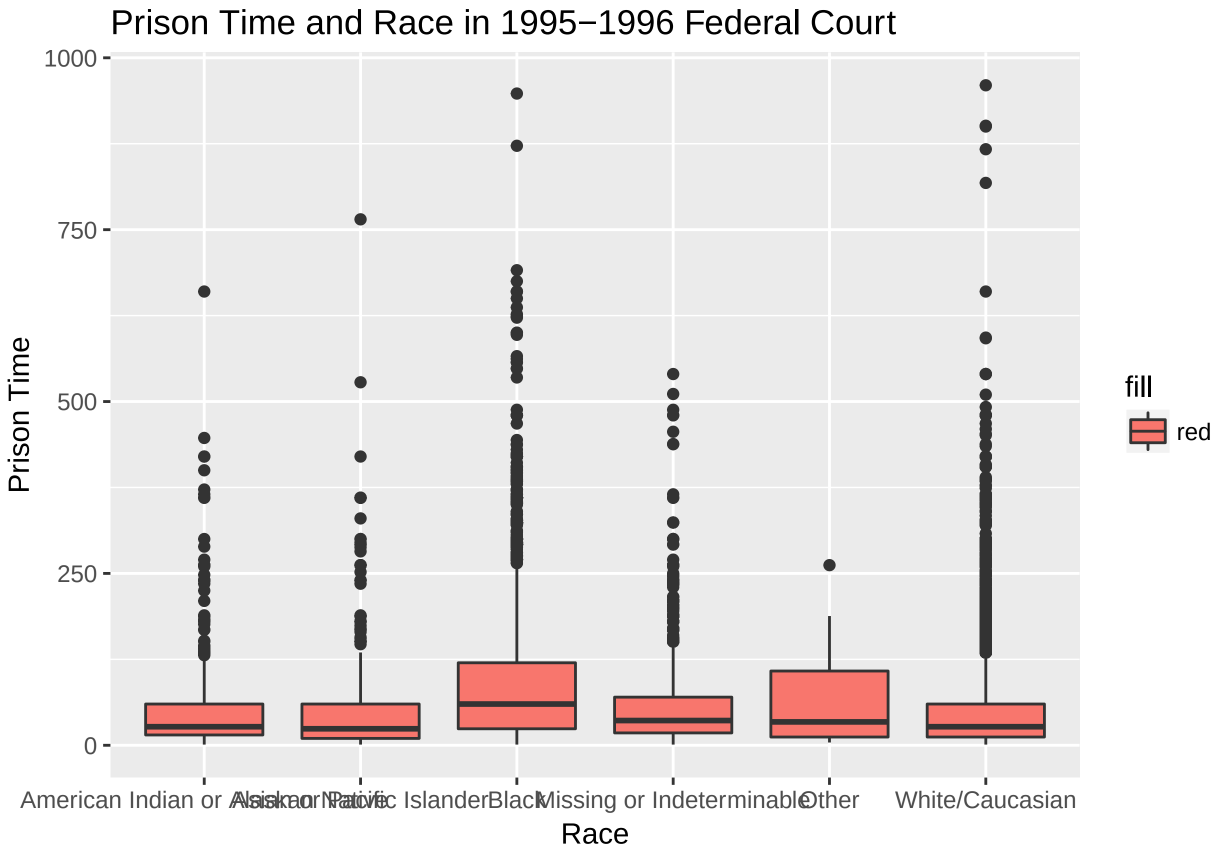
# There are some years where specific values such as time served etc. which I converted them to NAs.

12. Create a boxplot of your data (you can create an individual boxplot for each variable if there are too many variables in your dataset to meaningfully visualize them all in one plot). Are there any outliers? If so, what are they and to which variable do they correspond? Do any of these outliers seem like obvious errors? If so, why?

```
library(ggplot2)

ggplot(data_subset_95_96, aes(x = MONRACE, y = TOTPRISN, fill = "red")) + labs(x = "Race", y = "Prison
```

```
## Warning: Removed 9483 rows containing non-finite values (stat_boxplot).
```



13. For any outliers and/or obvious errors, what do you think is the best way to handle them (i.e. remove them entirely, run analyses including and excluding them and compare the results, manually change them to an appropriate measure of center, or something else?).

I have chosen to exclude this completely because they would mess up my model if I code them to - say -1.