Lab3_YZ_EDA

Yulia Zamriy

March 18, 2018

```
#install.packages("kableExtra")
#install.packages("viridisLite")
#install.packages("viridis")
#install.packages("Hmisc")
library(knitr)
library(kableExtra)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

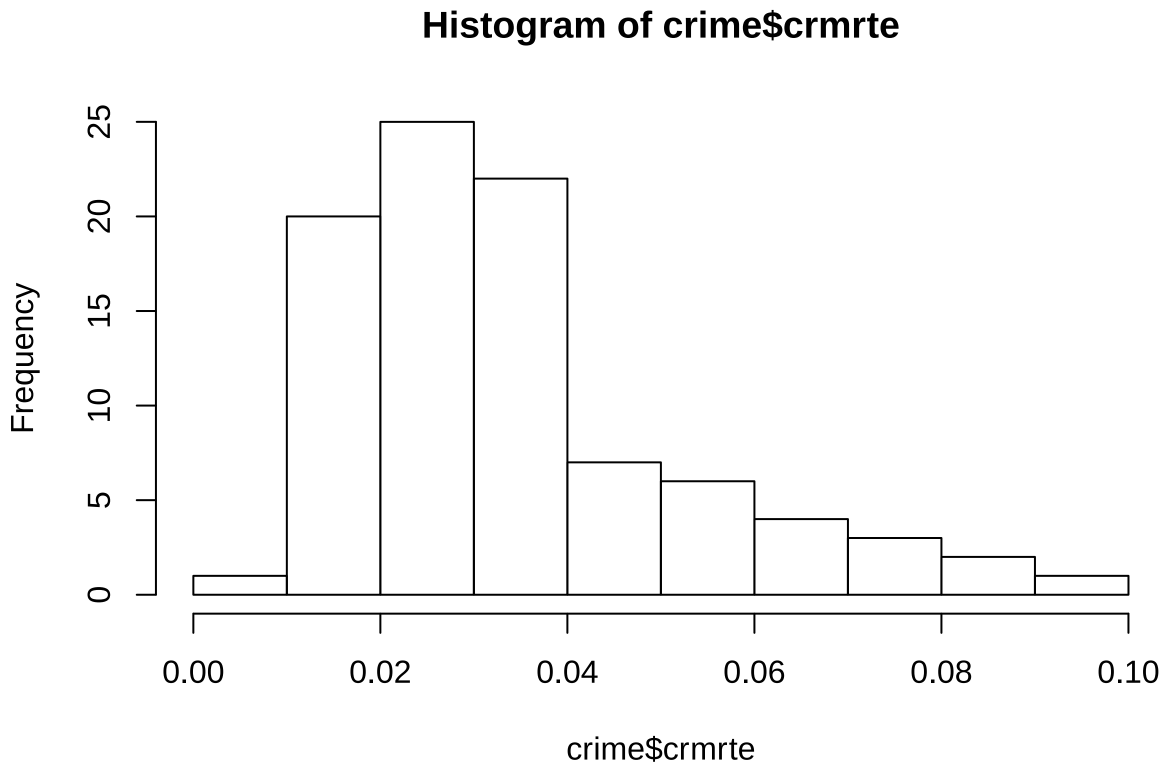
```
library(reshape2)
library(ggplot2)
```

```
#setwd("/home/yulia/Documents/MIDS/W203/Lab_3/")
crime <- read.csv("crime_v2.csv", stringsAsFactors = FALSE)
crime <- na.omit(crime)
```

```
summary(crime$crmrte)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005533 0.020927 0.029986 0.033400 0.039642 0.098966
```

```
hist(crime$crmrte)
```



```
crime$prbconv <- as.numeric(crime$prbconv)
```

```
summary(crime$prbarr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09277 0.20568 0.27095 0.29492 0.34438 1.09091
```

```
summary(crime$prbconv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34541 0.45283 0.55128 0.58886 2.12121
```

```
summary(crime$prbpris)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1500  0.3648  0.4234  0.4108  0.4568  0.6000
```

```
nrow(crime[crime$prbarr >= 1,])
```

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

```
nrow(crime[crime$prbconv >= 1,])
```

```
## [1] 10
```

```
crime$exclude <- 0
crime[crime$prbarr > 1,]$exclude <- 1
crime[crime$prbconv > 1,]$exclude <- 1
table(crime$exclude)
```

```
## 
##  0  1 
## 81 10
```

```
summary(crime$avgsen)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   5.380   7.340   9.100   9.647  11.420  20.700
```

```
summary(crime$polpc)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## 0.0007459 0.0012308 0.0014853 0.0017022 0.0018768 0.0090543
```

```
summary(crime$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.00002 0.54741 0.96226 1.42884 1.56824 8.82765
```

```
summary(crime$taxpc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   25.69   30.66   34.87   38.06   40.95  119.76
```

```
mean(crime$west)
```

```
## [1] 0.2527473
```

```
mean(crime$central)
```

```
## [1] 0.3736264
```

```
mean(crime$urban)
```

```
## [1] 0.08791209
```

```
summary(crime$pctmin80)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.284   9.845  24.312  25.495  38.142  64.348
```

```
summary(crime$wcon)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   193.6   250.8   281.4   285.4   314.8   436.8
```

```
summary(crime$wtuc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   187.6   374.6   406.5   411.7   443.4   613.2
```

```
summary(crime$wtrd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   154.2   190.9   203.0   211.6   225.1   354.7
```

```
summary(crime$wfir)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   170.9   286.5   317.3   322.1   345.4   509.5
```

```
summary(crime$wser)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.7   253.2   275.6   280.5  2177.1
```

```
summary(crime$wmfg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   157.4   288.9   320.2   335.6   359.6   646.9
```

```
summary(crime$wfed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   326.1   400.2   449.8   442.9   478.0   598.0
```

```
summary(crime$wsta)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   258.3   329.3   357.7   357.5   382.6   499.6
```

```
summary(crime$wloc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   239.2   297.3   308.1   312.7   329.2   388.1
```

```
summary(crime$mix)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01961 0.08073 0.10186 0.12884 0.15175 0.46512
```

```
summary(crime$pctymle)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06216 0.07443 0.07771 0.08396 0.08350 0.24871
```

```
crime[crime$wser > 2000,]$exclude <- 1
crime_sub <- subset(crime, exclude == 0)
crime_sub$exclude <- NULL
```

```
# Prepare a .RData for easier sharing and usage.
ind_variables <- c(
  'prbarr', 'prbconv', 'prbpris', 'avgsen',
  'polpc', 'density', 'taxpc', 'west', 'central', 'urban', 'pctmin80', 'wcon',
  'wtuc', 'wtrd', 'wfir', 'wser', 'wmfg', 'wfed', 'wsta', 'wloc', 'mix',
  'pctymle'
)
var_labels <- c(
  'probability of arrest', 'probability of conviction',
  'probability of prison sentence', 'avg. sentence, days',
  'police per capita', 'people per sq. mile', 'tax revenue per capita',
  '=1 if in western N.C.', '=1 if in central N.C.', '=1 if in SMSA',
  'perc. minority, 1980', 'weekly wage, construction',
  'wkly wge, trns, util, commun', 'wkly wge, whlesle, retail trade',
  'wkly wge, fin, ins, real est', 'wkly wge, service industry',
  'wkly wge, manufacturing', 'wkly wge, fed employees',
  'wkly wge, state employees', 'wkly wge, local gov emps',
  'offense mix: face-to-face/other', 'percent young male'
)
impact <- c("Negative" , "Negative", "Negative", "Negative",
            "Negative", "Positive", "Negative",
            "Unclear", "Unclear", "Unclear", "Unclear",
            "Negative","Negative","Negative",
            "Negative", "Negative", "Negative", "Negative",
            "Negative", "Negative", "Unclear","Positive")
control <- c("Yes", "Yes", "Yes", "Yes",
             "Yes", "No", "Yes",
             "No", "No", "No","No",
             "Yes", "Yes", "Yes",
             "Yes", "Yes", "Yes", "Yes",
             "Yes", "Yes", "No", "No")
desc <- data.frame(ind_variables, var_labels, impact, control)
colnames(desc) <- c("Explanatory Variables",
                   "Explanation",
                   "Expected Impact on Crime Rate",
                   "Can Gov Impact on This?")
# col_labels <-  c(ind_variables = "Explanatory Variables",
#                   var_labels = "Explanation")
# desc <- upData(desc, labels = col_labels)
```

```
kable(desc, booktabs = TRUE) %>%
  kable_styling(latex_options = c("scale_down"),
                full_width = FALSE) %>%
  row_spec(0, bold = TRUE) %>%
  column_spec(1, width = "8em") %>%
  column_spec(3, width = "10em") %>%
  column_spec(4, width = "9em")
```

Explanatory Variables	**Explanation**	**Expected Impact on Crime Rate**	**Can Gov Impact on This?**
prbarr	probability of arrest	Negative	Yes
prbconv	probability of conviction	Negative	Yes
prbpris	probability of prison sentence	Negative	Yes
avgsen	avg. sentence, days	Negative	Yes
polpc	police per capita	Negative	Yes
density	people per sq. mile	Positive	No
taxpc	tax revenue per capita	Negative	Yes
west	=1 if in western N.C.	Unclear	No
central	=1 if in central N.C.	Unclear	No
urban	=1 if in SMSA	Unclear	No
pctmin80	perc. minority, 1980	Unclear	No
wcon	weekly wage, construction	Negative	Yes
wtuc	wkly wge, trns, util, commun	Negative	Yes
wtrd	wkly wge, whlesle, retail trade	Negative	Yes
wfir	wkly wge, fin, ins, real est	Negative	Yes
wser	wkly wge, service industry	Negative	Yes
wmfg	wkly wge, manufacturing	Negative	Yes
wfed	wkly wge, fed employees	Negative	Yes
wsta	wkly wge, state employees	Negative	Yes
wloc	wkly wge, local gov emps	Negative	Yes
mix	offense mix: face-to-face/other	Unclear	No
pctymle	percent young male	Positive	No

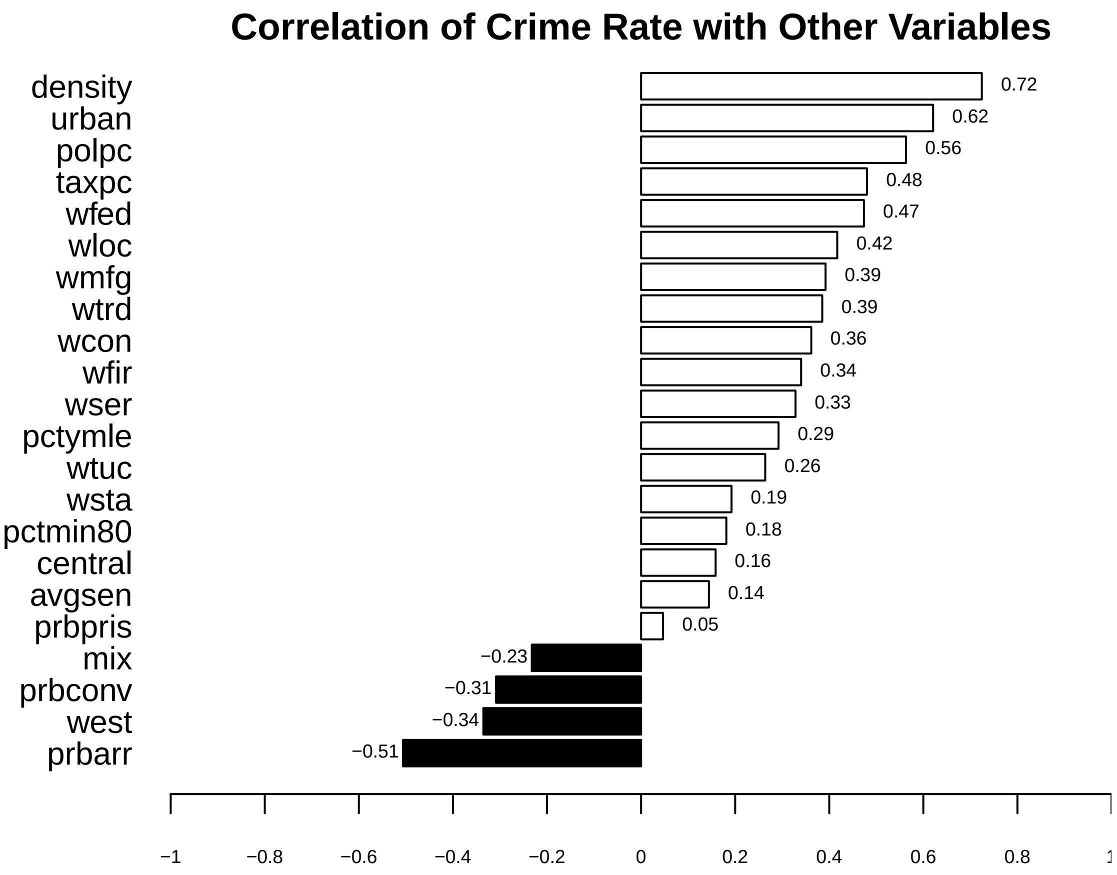
```r
crime_cor <- cor(crime_sub)[3,-c(1,2,3)]
```

```
## Warning in cor(crime_sub): the standard deviation is zero
```

```r
crime_cor <- crime_cor[order(crime_cor)]
negative <- ifelse(crime_cor < 0, 1,0)
```

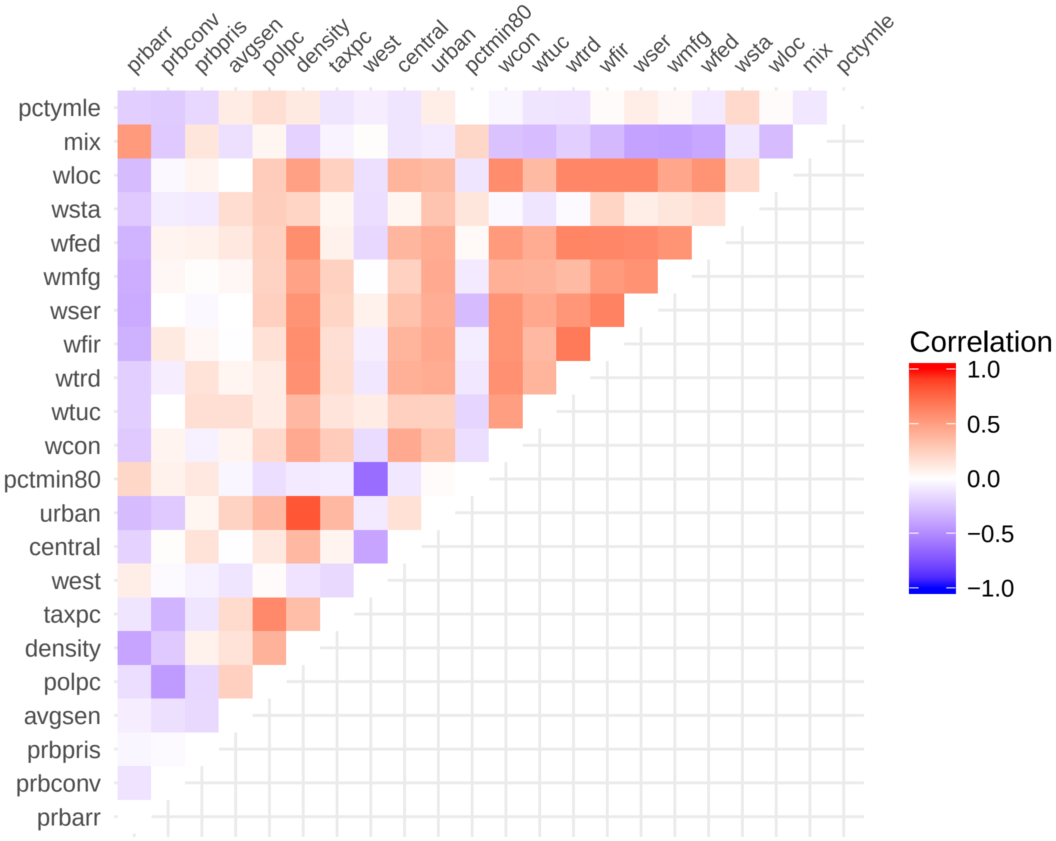
```r
crime_cor_lab <- ifelse(crime_cor < 0, crime_cor-0.15, crime_cor)

par(mar = c(2,8,1,0))
b <- barplot(crime_cor,
        col = negative,
        horiz = TRUE,
        las = 1,
        xaxt = "n",
        xlim = c(-1,1),
        main = "Correlation of Crime Rate with Other Variables")
text(x = crime_cor_lab,
     y = b,
     label = round(crime_cor,2),
     pos = 4,
     cex = 0.6)
axis(1,
     at = seq(-1,1, by = 0.2),
     labels = seq(-1,1, by = 0.2),
     cex.axis = 0.6)
```



```
cor_mat <- round(cor(crime_sub[-c(1:3)]),2)
get_upper_tri <- function(cor_mat){
    cor_mat[lower.tri(cor_mat)]<- NA
    return(cor_mat)
}
cor_mat_upper <- get_upper_tri(cor_mat)
cor_mat_upper2 <- melt(cor_mat_upper, na.rm = TRUE)
cor_mat_upper2[cor_mat_upper2$value == 1,]$value <- 0
```

```
ggplot(data = cor_mat_upper2, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name = "Correlation") +
  theme_minimal() +
  scale_x_discrete(position = "top") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 8, hjust = 0),
        axis.title.x=element_blank(),
        axis.title.y=element_blank()) +
  coord_fixed()
```



```
ind_vars_all <- c("prbarr", "prbconv", "prbpris", "avgsen", "polpc", "density", "taxpc",
               "west", "central", "urban", "pctmin80", "wcon", "wtuc", "wtrd", "wfir",
               "wser", "wmfg", "wfed", "wsta", "wloc", "mix", "pctymle")
ind_vars1 <- c("polpc", "taxpc","wfed","pctymle","avgsen")
```

```r
crmrte_formula1 <- as.formula(paste("crmrte ~", paste(ind_vars1, collapse = "+"), sep = ""))
crmrte_lm1 <- lm(crmrte_formula1, data = crime_sub)
summary(crmrte_lm1)
```

```
## 
## Call:
## lm(formula = crmrte_formula1, data = crime_sub)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.022297 -0.007113 -0.001875  0.005518  0.041679 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -7.550e-02  1.305e-02  -5.788 1.56e-07 ***
## polpc        6.450e+00  3.543e+00   1.820 0.072698 .  
## taxpc        5.527e-04  1.368e-04   4.040 0.000128 ***
## wfed         1.364e-04  2.394e-05   5.698 2.25e-07 ***
## pctymle      2.735e-01  6.222e-02   4.396 3.58e-05 ***
## avgsen      -4.702e-04  6.093e-04  -0.772 0.442709    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.01243 on 75 degrees of freedom
## Multiple R-squared:  0.5921, Adjusted R-squared:  0.5649 
## F-statistic: 21.77 on 5 and 75 DF,  p-value: 2.116e-13
```

```r
crmrte_formula_all <- as.formula(paste("crmrte ~", paste(ind_vars_all, collapse = "+"), sep = ""))
crmrte_lm0 <- lm(crmrte ~ 1,
                 data = crime_sub)
crmrte_lm_all <- lm(crmrte_formula_all,
                 data = crime_sub)
crmrte_lm_step <- step(crmrte_lm0, scope=list(lower=crmrte_lm0, upper=crmrte_lm_all),
                       direction="both",
                       trace = FALSE)
summary(crmrte_lm_step)
```

```
## 
## Call:
## lm(formula = crmrte ~ density + polpc + pctmin80 + prbarr + wsta + 
##     pctymle + taxpc + prbconv + mix + wser + wfed + wloc + central + 
##     wfir + avgsen + wcon + wtrd, data = crime_sub)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.013045 -0.004003 -0.001198  0.003880  0.018825 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.446e-02  1.542e-02   0.938 0.352050    
## density      5.555e-03  7.941e-04   6.995 2.04e-09 ***
## polpc        9.407e+00  2.352e+00   4.000 0.000169 ***
## pctmin80     3.620e-04  5.495e-05   6.588 1.04e-08 ***
## prbarr      -5.667e-02  9.502e-03  -5.964 1.22e-07 ***
## wsta        -5.249e-05  2.193e-05  -2.393 0.019686 *
## pctymle      1.456e-01  4.015e-02   3.625 0.000579 ***
## taxpc        2.413e-04  8.949e-05   2.696 0.008993 **
## prbconv     -9.038e-03  5.884e-03  -1.536 0.129549
## mix         -2.121e-02  1.304e-02  -1.626 0.108938
## wser        -8.551e-05  2.843e-05  -3.007 0.003783 **
## wfed         4.192e-05  2.316e-05   1.810 0.075116 .
## wloc         5.446e-05  4.266e-05   1.277 0.206446
## central     -4.111e-03  1.890e-03  -2.175 0.033420 *
## wfir        -5.645e-05  2.616e-05  -2.158 0.034750 *
## avgsen      -6.384e-04  3.618e-04  -1.765 0.082474 .
## wcon         3.500e-05  2.393e-05   1.462 0.148584
## wtrd         5.199e-05  3.935e-05   1.321 0.191185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006998 on 63 degrees of freedom
## Multiple R-squared:  0.8915, Adjusted R-squared:  0.8622
## F-statistic: 30.44 on 17 and 63 DF,  p-value: < 2.2e-16
```