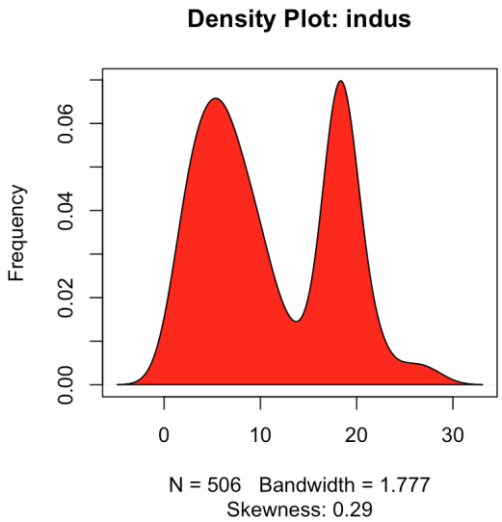
Jake Kruse
Geog 560
Lab 1

(10 pts in total) Use the "Boston" data (from the "MASS" package) that contains the Housing Values in Suburbs of Boston (506 towns and 14 attributes.) to answer the following questions. Figures/plots that may help answer the questions could be included in the document. The data attribute descriptions can be found in page 20 – page 21:
https://cran.r-project.org/web/packages/MASS/MASS.pdf

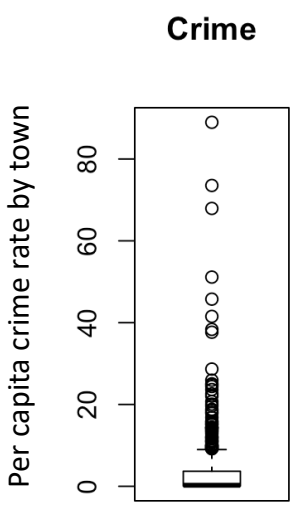
(1) (2 pts) Which of the 14 variables has the most symmetric distribution (Hint: check the skewness for all variables)?

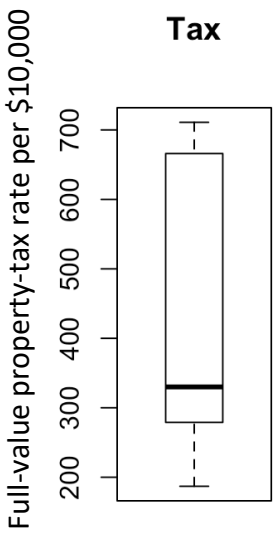
'indus' variable has the skewness value closest to zero; the peaks are approximately even in size and distribution about the mean.



(2) (2 pts) Do any of the suburbs of Boston appear to have particularly high crime rates (outliers)? Extreme high tax rates? (Hint: Boxplot)

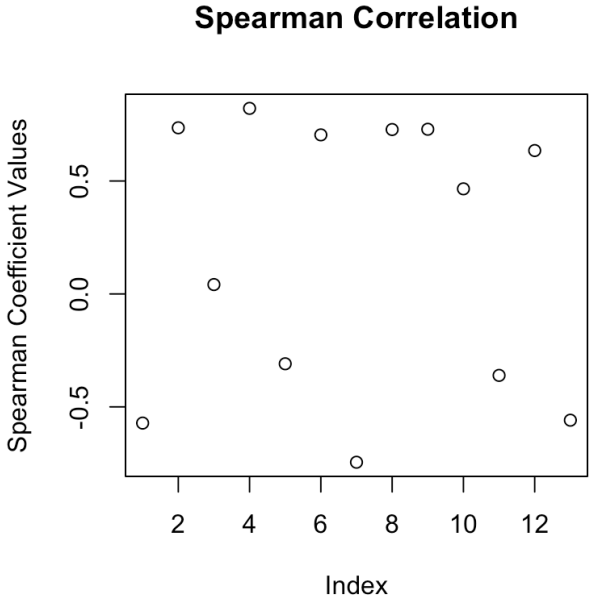
64 of the Boston suburbs had outlying crime values, while 0 of the Boston suburbs had outlying tax rates, as can be seen from the following boxplots.

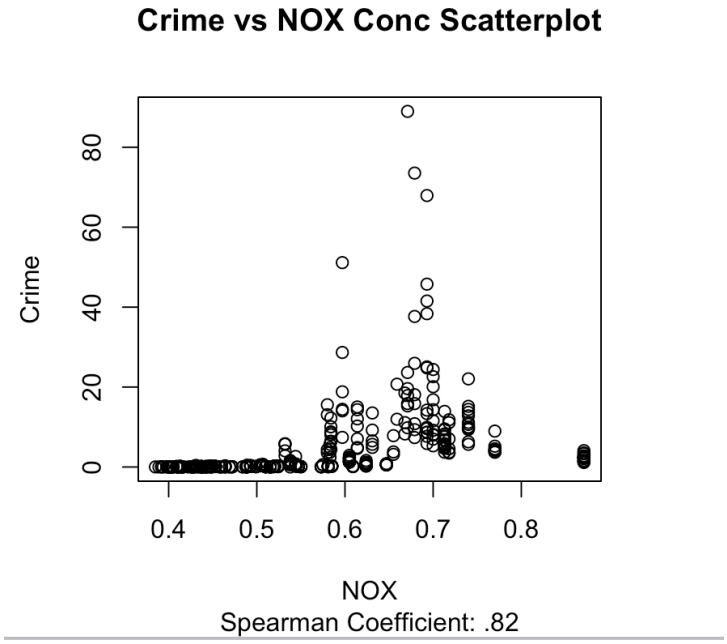




(3) (2 pts) Are any of the 14 predictors (independent variables) associated with per capita crime rate (dependent variable)? If so, explain the relationship and comment on your findings.

Scatterplots of each variable against crime showed that none of the relationships were linear. As such, the Spearman Rank Correlation Coefficient was used to find significant correlations (here anything with a correlation greater than an absolute value of .75 was considered). Nitrous Oxides Concentration and Distance to Boston Employment Centers fell into this category, with .82 and -.75 correlation coefficients. The first of the following three graphs is a visualization of the Spearman Coefficients for all comparisons to crime; this was used to identify Spearman Correlations with absolute values of >.75.



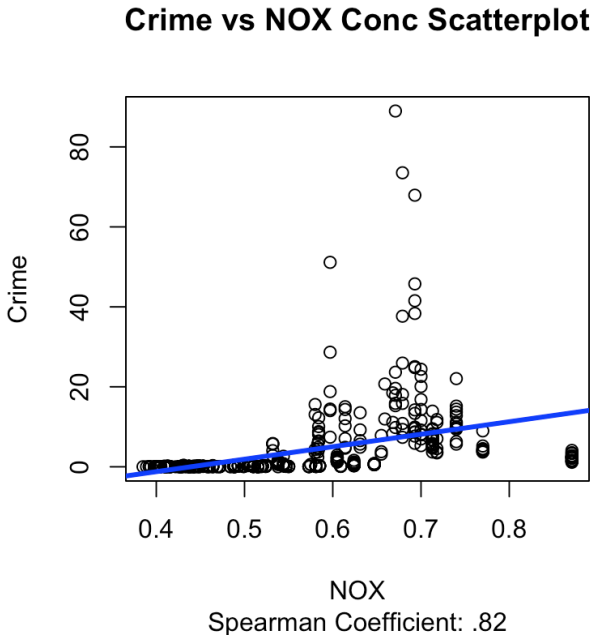


(4) (2 pts) Assume that a safe town requires lower than 3% average crime rate (with unknown population variance), please use the hypothesis testing method to judge whether the suburbs of Boston in this dataset are statistically safe or not (Hint: use the variable “crim”).

An alpha value of 0.05 was chosen ahead of time, as well as a Null hypothesis of pop mean for crime= 3% and an Alternate Hypothesis of pop mean for crime <3%. A one-sample t-test yielded a p-value of 0.95, such that the Null was rejected; mean crime for all the Boston suburbs is probably higher than 3%, and thus on the whole Boston suburbs should not be considered safe.

(5) (2 pts) Choose the predictor that has the largest positive correlation relationship with per capita crime rate and examine their linear regression model. Comment on your findings.

NOX concentration was the best positive predictor of Crime Rates. The R-squared value was .1756, which was irrelevant because a linear model does not seem to fit the data. Perhaps a quadratic model would be more appropriate.



R-code

```
##Jake Kruse
##Geog 560, Lab 1
if (!require("MASS")) install.packages('MASS')
if (!require("moments")) install.packages('moments')
if (!require("e1071")) install.packages('e1017')
library('moments')
library("MASS")
library('e1071')
```

```
##detach("package:e1071", unload=TRUE) #load an installed package before usage for the caculation of Skewness
data("Boston")
head(Boston, 3)
tail(Boston,3)
summary(Boston)

##1
##calculate skewness of variables
crimskew = skewness(Boston$crim)
znskew = skewness(Boston$zn)
indusskew = skewness(Boston$indus)
chasskew = skewness(Boston$chas)
noxskew = skewness(Boston$nox)
rmskew = skewness(Boston$rm)
ageskew = skewness(Boston$age)
disskew = skewness(Boston$dis)
radskew = skewness(Boston$rad)
taxskew = skewness(Boston$tax)
ptratioskew = skewness(Boston$ptratio)
blackskew = skewness(Boston$black)
lstatskew = skewness(Boston$lstat)
medvskew = skewness(Boston$medv)

a = c(crimskew, znskew, indusskew, chasskew, noxskew,
      rmskew, ageskew, disskew, radskew, taxskew, ptratioskew, blackskew, lstatskew, medvskew)
absa = abs(a)
which(a==min(absa))
a[which(a==min(absa))]
##indus variable has the skewness value closest to zero

#density plot of variable with lowest abs skew
par(mar=c(5,5,5,5)) #margins of figure
plot(density(Boston$indus), main="Density Plot: indus", ylab="Frequency",
sub=paste("Skewness:", round(indusskew, 2)))  # density plot for 'speed'
polygon(density(Boston$indus), col="red") #draw a polygon on the density plot

##2
summary(Boston$crim)
boxplot.stats(Boston$crim)$out
boxplot.stats(Boston$tax)$out
par(mfrow=c(1, 2))  # divide the graph area in 1 row, 2 columns
boxplot(Boston$crim, main="Crime")
```

```
boxplot(Boston$tax, main="Tax")

##3
##scatterplots to determine if linear relationship
plot(Boston$zn, Boston$crim) # not L
plot(Boston$crim,Boston$indus) # not L
plot(Boston$crim,Boston$chas) # not L
plot(Boston$crim,Boston$nox)# not L
plot(Boston$crim,Boston$rm) # not L
plot(Boston$crim,Boston$age)  # not L
plot(Boston$crim,Boston$dis) #not L
plot(Boston$crim,Boston$rad)  # not L
plot(Boston$crim,Boston$tax)  # not L
plot(Boston$crim,Boston$ptratio)# not L
plot(Boston$crim,Boston$black) # not L
plot(Boston$crim,Boston$lstat) # not L
plot(Boston$crim,Boston$medv) # not L
plot(Boston$crim,Boston$crim) # not L

# calculate Person's correlation between speed and distance
c=cor(Boston$zn, Boston$crim)
d=cor(Boston$indus, Boston$crim)
e=cor(Boston$chas, Boston$crim)
f=cor(Boston$nox, Boston$crim)
g=cor(Boston$rm, Boston$crim)
h=cor(Boston$age, Boston$crim)
i=cor(Boston$dis, Boston$crim)
j=cor(Boston$rad, Boston$crim)
k=cor(Boston$tax, Boston$crim)
l=cor(Boston$ptratio, Boston$crim)
m=cor(Boston$black, Boston$crim)
n=cor(Boston$lstat, Boston$crim)
o=cor(Boston$medv, Boston$crim)
pearcorlis = c(c,d,e,f,g,h,i,j,k,l,m,n,o)
#abspear = abs(pearcorlis)
#which(pearcorlis==min(abspear))


par(mfrow=c(1, 1)) #margins of figure
plot(pearcorlis, main="Pearson Correlation", ylab = "Pearson Correlation Values")

#plot(Boston$rad, Boston$crim, main = "Index of Acc to Radial Hwys\n vs Crime Rate per
Capita",
  #   xlab = "Index of Acc to Radial Hwys",
```

```
   # ylab = "Crime Rate per Capita", xaxt="n")
#axis(1,at = seq(0, 200, by = 5))

# calculate Spearman's rank correlation
p=cor(Boston$zn, Boston$crim, method="spearman")
q=cor(Boston$indus, Boston$crim, method="spearman")
r=cor(Boston$chas, Boston$crim, method="spearman")
s=cor(Boston$nox, Boston$crim, method="spearman")
t=cor(Boston$rm, Boston$crim, method="spearman")
u=cor(Boston$age, Boston$crim, method="spearman")
v=cor(Boston$dis, Boston$crim, method="spearman" )
w=cor(Boston$rad, Boston$crim, method="spearman")
x=cor(Boston$tax, Boston$crim, method="spearman")
y=cor(Boston$ptratio, Boston$crim, method="spearman")
z=cor(Boston$black, Boston$crim, method="spearman" )
aa=cor(Boston$lstat, Boston$crim, method="spearman")
bb=cor(Boston$medv, Boston$crim, method="spearman")

spearcor = c(p,q,r,s,t,u,v,w,x,y,z,aa,bb)
max(spearcor)

par(mfrow=c(1, 1)) #margins of figure
plot(spearcor, main="Spearman Correlation", ylab = "Spearman Coefficient Values")
plot(Boston$nox, Boston$crim, main = "Crime vs NOX Conc Scatterplot", xlab = "NOX", ylab =
"Crime", sub = paste("Spearman Coefficient: .82"))
plot(Boston$dis, Boston$crim, main = "Distance to Boston Employment Centers\nvs Crime
Scatterplot", xlab = "Distance", ylab = "Crime", sub = paste("Spearman Coefficient: -.75"))

##4
# One-sample t-test: it is used to test hypotheses about the mean value of a population from
which a sample is drawn.
mean(Boston$crim)
t.test(Boston$crim, mu=3, alternative = "less")

##5
# linear regression analysis
linearMod <- lm(Boston$crim~Boston$nox)  # build linear regression model on full data
print(linearMod) # present out the resulting model
summary(linearMod)  # linear regression model summary
# We can then draw a trend line with abline function with the linear regression model
help(abline) # search for the requiremnent of a method
plot(Boston$nox, Boston$crim, main = "Crime vs NOX Conc Scatterplot", xlab = "NOX", ylab =
"Crime", sub = paste("Spearman Coefficient: .82"))
abline(linearMod,  col="blue", lwd = 3) # col: line color; lwd: line width
```