CIND 123: Data Analytics Basic Methods: Assignment-3_F2022

Assignment 3 (10%)

Total 100 Marks

Amarpreet Kaur

CIND 123 D30&501213036]

Instructions

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. Review this website for more details on using R Markdown http://rmarkdown.rstudio.com.

Use RStudio for this assignment. Complete the assignment by inserting your code wherever you see the string "#INSERT YOUR ANSWER HERE".

When you click the **Knit** button, a document (PDF, Word, or HTML format) will be generated that includes both the assignment content as well as the output of any embedded R code chunks.

NOTE: YOU SHOULD NEVER HAVE `install.packages` IN YOUR CODE; OTHERWISE, THE `Knit` OPTION WILL GIVE AN ERROR. COMMENT OUT ALL PACKAGE INSTALLATIONS.

Submit **both** the `rmd` and generated `output` files. Failing to submit both files will be subject to mark deduction. PDF or HTML is preferred.

Sample Question and Solution

```
seq(3, 30, 2)
```

```
##  [1]  3  5  7  9 11 13 15 17 19 21 23 25 27 29
```

```
seq(3, 29, 2)
```

```
##  [1]  3  5  7  9 11 13 15 17 19 21 23 25 27 29
```

Question 1 [30 Pts]

Q1a (10 points)

The midterm and final exam grades of some students are given as `c(92,91,67,72,85,81,53,45)` and `c(87,100,65,81,93,77,55,36)`. Set variables `midterm` and `final` respectively. Then find the least-squares line relating the midterm to the final exam

Does the assumption of a linear relationship appear to be reasonable in this case? Give reasons for your answer as a comment.

```
midterm <- c(92,91,67,72,85,81,53,45)
final <- c(87,100,65,81,93,77,55,36)
relation <- lm(final~midterm)
relation
```

```
##
## Call:
## lm(formula = final ~ midterm)
##
## Coefficients:
## (Intercept)      midterm
##     -10.922        1.163
```

```
summary(relation)
```

```
##
## Call:
## lm(formula = final ~ midterm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.052 -5.617  1.157  5.093  8.204
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.9224    11.3859  -0.959 0.374455
## midterm       1.1628     0.1517   7.662 0.000258 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.976 on 6 degrees of freedom
## Multiple R-squared:  0.9073, Adjusted R-squared:  0.8918
## F-statistic: 58.71 on 1 and 6 DF,  p-value: 0.0002582
```

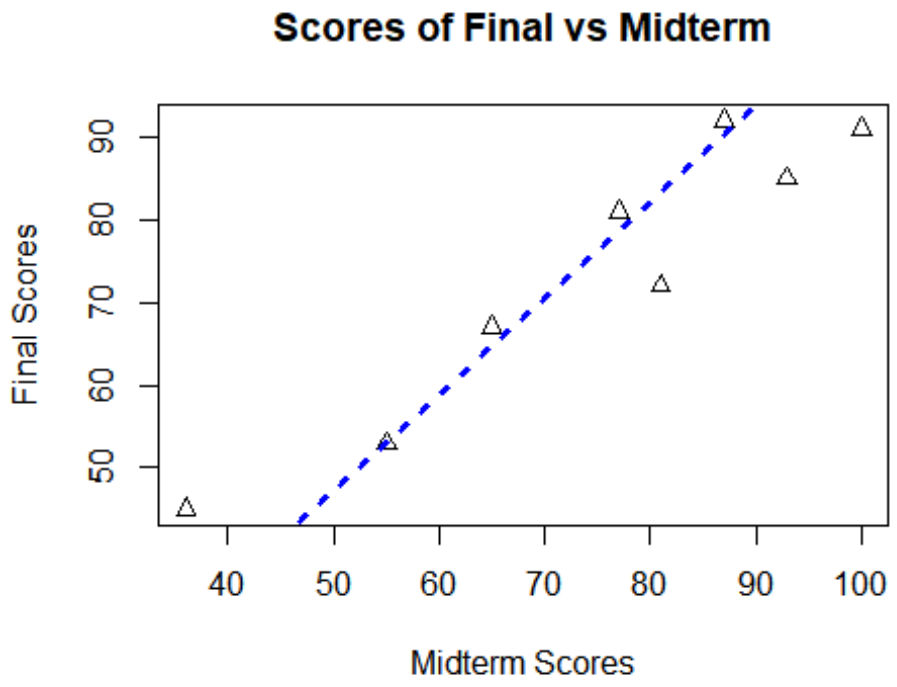
```
#The adjusted R-squared value is 0.9073 (or90%), which is close to 1. Hence,
it is relevant relationship.Standard error is really small, almost
negligible, also p-value is very close to zero. This relationship seems
```

```
reasonable as residuals are evenly distributed.Also, there are three stars
(***), so it shows strong relationship.
```

Q1b (10 points)

Plot the midterm as a function of the final exam grades using a scatterplot and graph the least-square line on the same plot.

```
plot(final, midterm, pch=02, xlab="Midterm Scores", ylab="Final Scores",
main="Scores of Final vs Midterm")
abline(coefficients(relation), lwd=3, lty=3, col="blue")
```



Q1c (10 points)

Use the regression line to predict the midterm grade when the final exam grade is 88

```
#Method 1
#From above model 1(a)
#line equation
final<- -10.9224 + 1.1628*midterm
#midterm drived from above equation
midterm<- (final+10.9224)/1.1628
final<- 88
midterm

## [1] 92 91 67 72 85 81 53 45
```

```r
# method 2--- create midterm function from final by recreating the euation
from below model
midterm <- c(92,91,67,72,85,81,53,45)
final <- c(87,100,65,81,93,77,55,36)
relation2 <- lm(midterm~final)
relation2
```

```
##
## Call:
## lm(formula = midterm ~ final)
##
## Coefficients:
## (Intercept)        final
##     15.3140       0.7803
```

```r
summary(relation2)
```

```
##
## Call:
## lm(formula = midterm ~ final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5169 -3.4676 -0.6873  2.5979  8.8014
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.3140     7.8263   1.957 0.098138 .
## final         0.7803     0.1018   7.662 0.000258 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.715 on 6 degrees of freedom
## Multiple R-squared:  0.9073, Adjusted R-squared:  0.8918
## F-statistic: 58.71 on 1 and 6 DF,  p-value: 0.0002582
```

```r
#Using Regression Line
midterm <- 15.3140+0.7803*88
midterm
```

```
## [1] 83.9804
```

Question 2 [45 Pts]

Please load the Crime data by running the following chunk of code.

You can read more about at this link - https://rdrr.io/cran/plm/man/Crime.html

```r
Crime = read.csv("https://r-data.pmagunia.com/system/files/datasets/dataset-
28105.csv")
```

Q2a (5 points)

Display the first 5 rows of the Crime data, the names of all the variables, and a descriptive summary of each variable

```
head(Crime, n=5)
```

```
##   county year    crmrte   prbarr  prbconv  prbpris avgsen     polpc
density
## 1      1   81 0.0398849 0.289696 0.402062 0.472222   5.61 0.0017868
2.307159
## 2      1   82 0.0383449 0.338111 0.433005 0.506993   5.59 0.0017666
2.330254
## 3      1   83 0.0303048 0.330449 0.525703 0.479705   5.80 0.0018358
2.341801
## 4      1   84 0.0347259 0.362525 0.604706 0.520104   6.89 0.0018859
2.346420
## 5      1   85 0.0365730 0.325395 0.578723 0.497059   6.55 0.0019244
2.364896
##      taxpc  region smsa  pctmin     wcon      wtuc     wtrd     wfir
wser
## 1 25.69763 central   no 20.2187 206.4803  333.6209 182.3330 272.4492
215.7335
## 2 24.87425 central   no 20.2187 212.7542  369.2964 189.5414 300.8788
231.5767
## 3 26.45144 central   no 20.2187 219.7802 1394.8030 196.6395 309.9696
240.1568
## 4 26.84235 central   no 20.2187 223.4238  398.8604 200.5629 350.0863
252.4477
## 5 28.14034 central   no 20.2187 243.7562  358.7830 206.8827 383.0707
261.0861
##     wmfg   wfed   wsta   wloc       mix   pctymle
## 1 229.12 409.37 236.24 231.47 0.0999179 0.0876968
## 2 240.33 419.70 253.88 236.79 0.1030491 0.0863767
## 3 269.70 438.85 250.36 248.58 0.0806787 0.0850909
## 4 281.74 459.17 261.93 264.38 0.0785035 0.0838333
## 5 298.88 490.43 281.44 288.58 0.0932486 0.0823065
```

```
names(Crime)
```

```
##  [1] "county"  "year"    "crmrte"  "prbarr"  "prbconv" "prbpris" "avgsen"
##  [8] "polpc"   "density" "taxpc"   "region"  "smsa"    "pctmin"  "wcon"
## [15] "wtuc"    "wtrd"    "wfir"    "wser"    "wmfg"    "wfed"    "wsta"
## [22] "wloc"    "mix"     "pctymle"
```

```
summary(Crime)
```

```
##      county           year         crmrte              prbarr
##  Min.   :  1.0   Min.   :81   Min.   :0.001812   Min.   :0.05882
##  1st Qu.: 51.0   1st Qu.:82   1st Qu.:0.018352   1st Qu.:0.21790
##  Median :103.0   Median :84   Median :0.028441   Median :0.27824
```

```
##  Mean   :100.6   Mean   :84   Mean   :0.031588   Mean   :0.30737
##  3rd Qu.:151.0   3rd Qu.:86   3rd Qu.:0.038406   3rd Qu.:0.35252
##  Max.   :197.0   Max.   :87   Max.   :0.163835   Max.   :2.75000
##     prbconv           prbpris           avgsen           polpc
##  Min.   : 0.06838   Min.   :0.1489   Min.   : 4.220   Min.   :0.0004585
##  1st Qu.: 0.34769   1st Qu.:0.3744   1st Qu.: 7.160   1st Qu.:0.0011913
##  Median : 0.47437   Median :0.4286   Median : 8.495   Median :0.0014506
##  Mean   : 0.68862   Mean   :0.4255   Mean   : 8.955   Mean   :0.0019168
##  3rd Qu.: 0.63560   3rd Qu.:0.4832   3rd Qu.:10.197   3rd Qu.:0.0018033
##  Max.   :37.00000   Max.   :0.6786   Max.   :25.830   Max.   :0.0355781
##     density           taxpc           region            smsa
##  Min.   :0.1977   Min.   : 14.30   Length:630         Length:630
##  1st Qu.:0.5329   1st Qu.: 23.43   Class :character   Class :character
##  Median :0.9526   Median : 27.79   Mode  :character   Mode  :character
##  Mean   :1.3861   Mean   : 30.24
##  3rd Qu.:1.5078   3rd Qu.: 33.27
##  Max.   :8.8277   Max.   :119.76
##      pctmin            wcon             wtuc             wtrd
##  Min.   : 1.284   Min.   :  65.62   Min.   :  28.86   Min.   :  16.87
##  1st Qu.:10.005   1st Qu.: 201.66   1st Qu.: 317.60   1st Qu.: 168.05
##  Median :24.852   Median : 236.46   Median : 358.20   Median : 185.48
##  Mean   :25.713   Mean   : 245.67   Mean   : 406.10   Mean   : 192.82
##  3rd Qu.:38.223   3rd Qu.: 269.69   3rd Qu.: 411.02   3rd Qu.: 204.82
##  Max.   :64.348   Max.   :2324.60   Max.   :3041.96   Max.   :2242.75
##       wfir              wser              wmfg             wfed
##  Min.   :  3.516   Min.   :   1.844   Min.   :101.8   Min.   :255.4
##  1st Qu.:235.705   1st Qu.: 191.319   1st Qu.:234.0   1st Qu.:361.5
##  Median :264.423   Median : 216.475   Median :271.6   Median :404.0
##  Mean   :272.059   Mean   : 224.671   Mean   :285.2   Mean   :403.9
##  3rd Qu.:302.440   3rd Qu.: 247.155   3rd Qu.:320.0   3rd Qu.:444.6
##  Max.   :509.466   Max.   :2177.068   Max.   :646.9   Max.   :598.0
##       wsta              wloc              mix              pctymle
##  Min.   :173.0   Min.   :163.6   Min.   :0.002457   Min.   :0.06216
##  1st Qu.:258.2   1st Qu.:226.8   1st Qu.:0.075324   1st Qu.:0.07859
##  Median :289.4   Median :253.1   Median :0.102089   Median :0.08316
##  Mean   :296.9   Mean   :258.0   Mean   :0.139396   Mean   :0.08897
##  3rd Qu.:331.5   3rd Qu.:289.3   3rd Qu.:0.149009   3rd Qu.:0.08919
##  Max.   :548.0   Max.   :388.1   Max.   :4.000000   Max.   :0.27436
```

Q2b (5 points)

Calculate the mean, variance and standard deviation of the weekly wage in construction (wcon) by omitting the missing values, if any.

```
Crime$Wcon = na.omit(Crime$wcon)
cat('The mean is:', mean(Crime$Wcon))

## The mean is: 245.6661

cat('\n\nThe variance is:', var(Crime$Wcon))
```

```
##
##
## The variance is: 14880.03

cat('\n\nThe standard deviation is:', sd(Crime$Wcon))

##
##
## The standard deviation is: 121.9837
```

Q2c-1 (5 points)

Use people per square mile (density) and police per capita (polpc) to build a linear regression model to predict tax per capita (taxpc).

Q2c-2 (5 points)

How can you draw a conclusion from the results? (Note: Full marks requires comment on the predictors)

```
# Q2c-1
per_capita1 <- lm(Crime$taxpc~Crime$density+Crime$polpc)
per_capita1

##
## Call:
## lm(formula = Crime$taxpc ~ Crime$density + Crime$polpc)
##
## Coefficients:
##   (Intercept)  Crime$density    Crime$polpc
##        27.051          1.626        487.517

summary((per_capita1))

##
## Call:
## lm(formula = Crime$taxpc ~ Crime$density + Crime$polpc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.735  -6.431  -2.532   3.481  89.924
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    27.0508     0.6998  38.654  < 2e-16 ***
## Crime$density   1.6261     0.3094   5.256 2.02e-07 ***
## Crime$polpc   487.5170   162.8628   2.993  0.00287 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.16 on 627 degrees of freedom
```

```
## Multiple R-squared:  0.05343,    Adjusted R-squared:  0.05041
## F-statistic:  17.7 on 2 and 627 DF,  p-value: 3.338e-08

# Q2c-2

#It appears that the intercept is equal to ~27.0508, and the coefficients for
'density' and 'polpc' (with a value of 'yes') are  1.6261 and 487.5170,
respectively. This indicates that as the density and polpc of people per
square mile increases.Both the independent variables have a positive
correlation with tax per capita.Finally, the adjusted R-squared value is
0.05343 (or 5%), this is very low, and indicates that the regression model
explains a small percentage of the variation of the response data.Also, means
is less than median hence it is Left Skewed Distribution.
```

Q2d (5 points)

Based on the output of your model, write an equation using the intercept and coefficients of density when polpc. Then, use the equation for a case with density of 0.4 and polpc of 0.0015 to predict its tax per capita.

```
#Summary
per_capita1 <- lm(Crime$taxpc~Crime$density+Crime$polpc)
summary((per_capita1))

##
## Call:
## lm(formula = Crime$taxpc ~ Crime$density + Crime$polpc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.735  -6.431  -2.532   3.481  89.924
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    27.0508     0.6998  38.654  < 2e-16 ***
## Crime$density   1.6261     0.3094   5.256 2.02e-07 ***
## Crime$polpc   487.5170   162.8628   2.993  0.00287 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.16 on 627 degrees of freedom
## Multiple R-squared:  0.05343,    Adjusted R-squared:  0.05041
## F-statistic:  17.7 on 2 and 627 DF,  p-value: 3.338e-08

#Equations
#predictdensity <- 27.0508+1.6261 *"density" + 487.5170*"polpc"
#Predicted  tax per capita with `density` of 0.4 and `polpc` of 0.0015
density<- 0.4
polpc<- 0.0015
predicttaxpc <- 27.0508+1.6261 *density+487.5170*polpc
predicttaxpc
```

```
## [1] 28.43252
```

Q2e-1 (5 points)

Find Pearson correlation between crimes committed per person (crmrte) and the probability of arrest (prbarr); and between percentage minority in 1980 (pctmin) and police per capita (polpc).

Q2e-2 (5 points)

What conclusions can you draw? Write your reasons as comments.

```
#Q2e-1

cor(Crime$crmrte,Crime$prbarr, method = "pearson")

## [1] -0.3585528

cor(Crime$pctmin,Crime$polpc, method = "pearson")

## [1] 0.03168164

#Q2e-2

#Both the correlations are related weakly. Crime committed per person and
probability of arrest are negatively correlated, hence criminal are not
caught. However percentage minority and police per capita are positively
correlated.
```

Q2f-1 (5 points)

Display the correlation matrix of the following variables: - crimes committed per person (crmrte), - probability of arrest (prbarr), - probability of conviction (prbconv), - police per capita (polpc), - percentage minority in 1980 (pctmin).

Q2f-2 (5 points)

Write what conclusion you can draw as comments. (answer not included, Please grade on student's comments)

```
# Q2f-1

cor(Crime[c("crmrte", "prbarr", "prbconv", "polpc", "pctmin")])

##             crmrte     prbarr    prbconv      polpc     pctmin
## crmrte   1.0000000 -0.3585528 -0.1130327 0.18482644 0.16902095
## prbarr  -0.3585528  1.0000000  0.0355689 0.29058128 0.10005025
## prbconv -0.1130327  0.0355689  1.0000000 0.44963500 0.10507694
## polpc    0.1848264  0.2905813  0.4496350 1.00000000 0.03168164
## pctmin   0.1690210  0.1000503  0.1050769 0.03168164 1.00000000
```

```
# Q2f-2

#It is clear that mostly variables are positively correlated (each increases
with an increase in the other) with the exception of  'crmrte' with 'prbarr'
and 'prbconv', which makes sense, as the number of police officers per person
will decrease as the count of people per square mile increases also the crime
rate increase.
```

Question 3 [25 Pts]

This question makes use of package "ISwR". Please load airquality dataset as following:

```
# or install.packages("ISwR")
library(ISwR)
data(airquality)
str(airquality)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

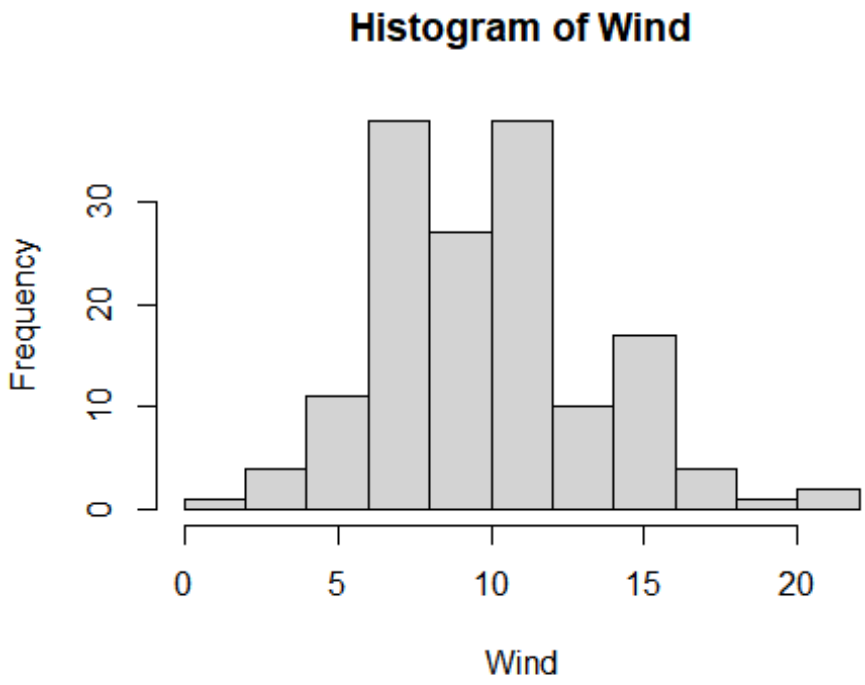
Q3a (7.5 points)

Use a histogram to assess the normality of the Wind variable, then explain why it appears (to some extent) normally distributed.

```
#Summary
summary(airquality)
```

```
##      Ozone           Solar.R           Wind             Temp
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
##  NA's   :37       NA's   :7
##      Month            Day
##  Min.   :5.000   Min.   : 1.0
##  1st Qu.:6.000   1st Qu.: 8.0
##  Median :7.000   Median :16.0
##  Mean   :6.993   Mean   :15.8
##  3rd Qu.:8.000   3rd Qu.:23.0
##  Max.   :9.000   Max.   :31.0
##
```

```r
#Histogram of the "Wind"
Wind <- airquality$Wind
hist(Wind)
```

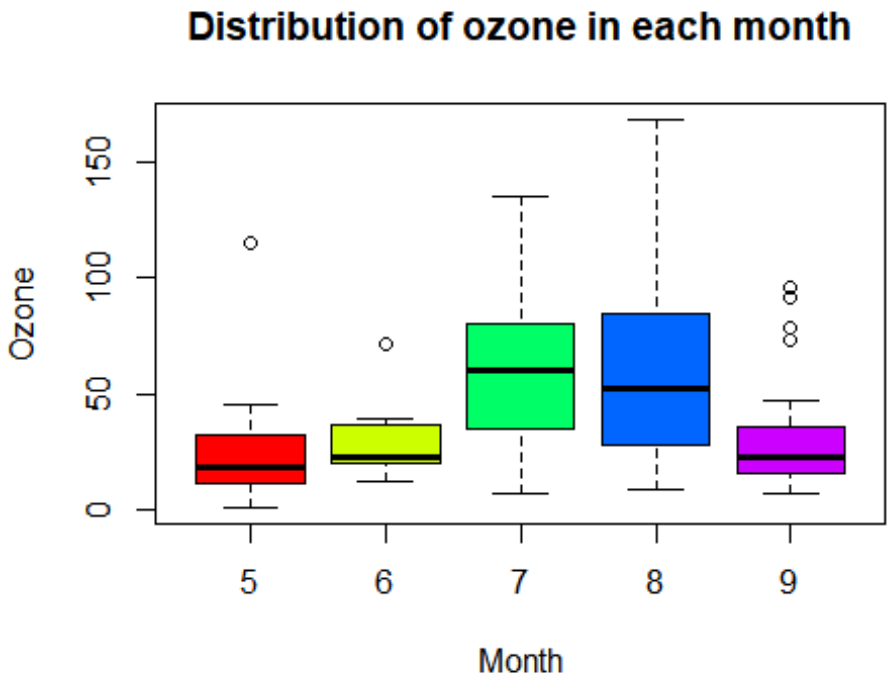



```r
# Shape is more and less similar to normal bell curve, but this is not exact normal distributed, because  the sample size is not large enough for the data points to appear as normally distributed (i.e., easily affected by extreme values).Wind is positively skewed, as mean is greater than the median.
```

Q3b (7.5 points)

Create a comparison boxplot that shows the distribution of Ozone in each month. Use different colours for each month.

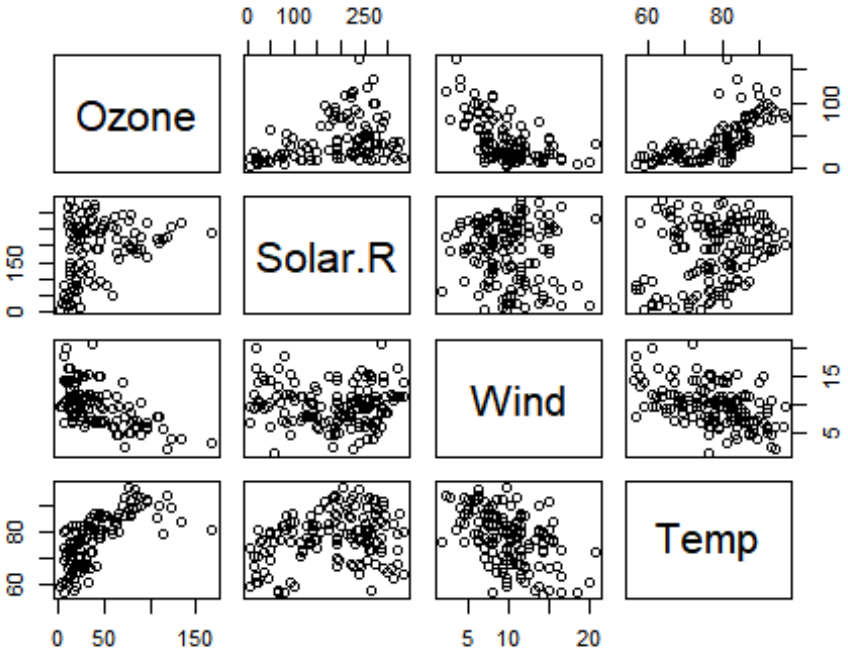
```r
boxplot(Ozone~Month,data = airquality, main="Distribution of ozone in each month", col=rainbow(length(unique(airquality$Month))))
```



Q3c (10 points)

Create a pairwise matrix of scatterplots of all the numeric attributes in the `airquality` dataset (i.e, `Ozone`, `Solar.R`, `Wind` and `Temp`.) (Hint: investigate `pairs()` function)

```
pairs(airquality[1:4])
```




```
# From plot we can have some of the following observations:
# Ozone and Solar have positive weak relationship(slope +ve).
# Ozone and Wind have negative moderate relationship(slope -ve).
# Ozone and Temp have positive strong relationship(slope +ve).
# Solar and wind have no relationship (slope almost equal to zero).
# solar and Wind have no relationship (slope almost equal to zero).
# Wind and Temp have moderate negative relationship(slope -ve).
```

** End of Assignment **