22-23-SIM-PARTIAL sol

LÍdia Montero

November, 3rd 2022

Contents

1 Boston Housing Data

Load Housing.RData file in your current R or RStudio session. Median value of owner-occupied homes in $1000's (medv) is going to be our numeric target and chas our target factor (chas has to be converted to factor). Use df dataset.

All qüestions account for 1 point (you have to answer 10 out of 15)

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.5
```

```
## Warning: package 'FactoMineR' was built under R version 4.0.5
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
## null device
##           1
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "b"       "lstat"   "medv"
## [15] "f.hcla"
```

1. Some observations have an 'medv' value of 50.0. These data points contain missing or censored values. Since medv is a numeric target, which suitable actions are needed before starting a deeper analysis? Implement those actions in your dataset.

```
# Point 1
summary(df$medv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   17.02   21.20   22.53   25.00   50.00
```

```
ll<-which(df$medv==50);length(ll)
```

```
## [1] 16
```

```
df<-df[-ll,]
# df<-df[-which(row.names(df)=="365"),]
```

2. Determine thresholds for mild and severe outliers for the average number of rooms among homes in the neighborhood. Are there any outliers? Indicate observation id's and atypical numbers for average rooms.

The upper threshold for severe outliers is at 3 times IQR from Q3, thus according to the summary of rm, 8.669 rooms. Lower threshold is 3.79. Mild outliers are those further than 1.5 IQR from/to Q1/Q3: upper 7.62 and lower 4.84. From my point of view, Obs with name "365" is the only severe outlier with 8.78 average rooms per dwelling in the neighborhood, but observation "366" can also be labelled as a severe lower outlier. Once the target is examined for each group defined by the Charles River factor, Observation name "365" seems to be still an outlier. Pay attention original observation "365" is now in row 354 since some records have been removed. Severe outliers correspond to 8.78 and 3.56 rooms.

Quiz grading: all reasonable answers dependent on Question 1 actions have been considered valid.

```
par(mfrow=c(1,2))
ss<-summary(df$rm);ss
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.561   5.881   6.185   6.245   6.578   8.780
```

```
# Upper/lower severe threshold
utso2<-ss[5]+3*(ss[5]-ss[2]);utso2
```

```
## 3rd Qu.
##   8.669
```

```
utsi2<-ss[2]-3*(ss[5]-ss[2]);utsi2
```

```
## 1st Qu.
##    3.79
```

```
# Upper/lower mild threshold
utmo2<-ss[5]+1.5*(ss[5]-ss[2]);utmo2
```

```
## 3rd Qu.
##  7.6235
```

```
utmi2<-ss[2]-1.5*(ss[5]-ss[2]);utmi2
```

```
## 1st Qu.
##  4.8355
```

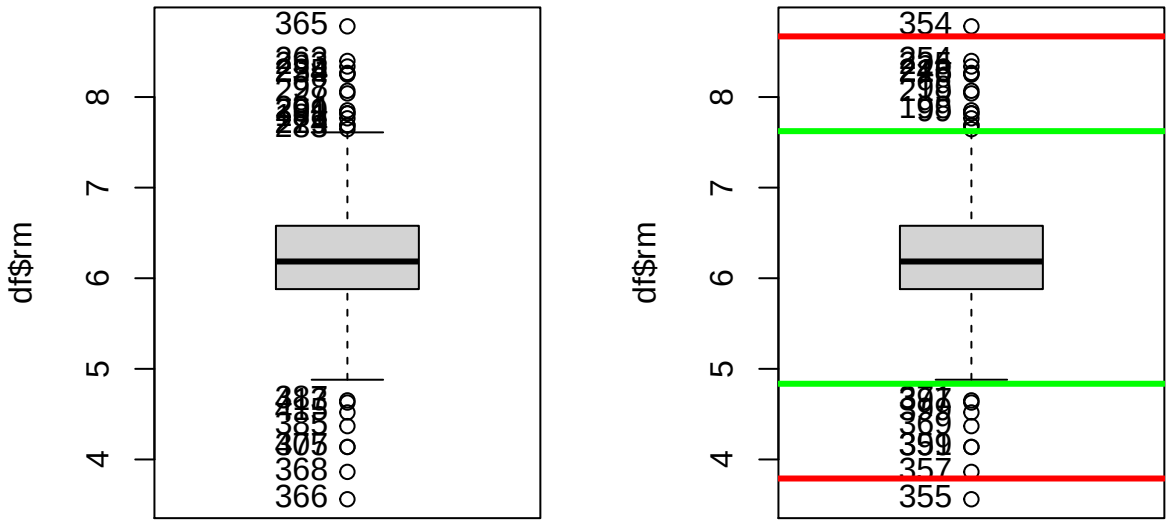
```
Boxplot(df$rm,id=list(n=Inf,labels=row.names(df)))
```

```
##  [1] "366" "368" "375" "385" "387" "407" "413" "415" "98"  "99"  "181" "204"
## [13] "225" "227" "229" "233" "234" "254" "263" "274" "281" "283" "365"
```

```
Boxplot(df$rm)
```

```
##  [1] 355 357 359 369 371 391 397 399 354 254 225 218 246 226  98 219 198  99
```

```
abline(h=utso2,col="red",lwd=3)
abline(h=utsi2,col="red",lwd=3)
abline(h=utmo2,col="green",lwd=3)
abline(h=utmi2,col="green",lwd=3)
```




```
Boxplot(df$rm~df$chas,id=list(n=Inf,labels=row.names(df)),col=heat.colors(2))
```

```
##  [1] "366" "368" "375" "385" "387" "407" "413" "415" "98"  "99"  "181" "203"
## [13] "204" "225" "227" "229" "233" "234" "254" "263" "281" "365"
```

```
abline(h=utso2,col="red",lwd=3)
abline(h=utsi2,col="red",lwd=3)
abline(h=utmo2,col="green",lwd=3)
abline(h=utmi2,col="green",lwd=3)

df[c("365","366"),]
```

```
##        crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lstat
## 365 3.47428  0  18.1    1 0.718 8.780 82.9 1.9047  24 666    20.2 354.55  5.29
## 366 4.55587  0  18.1    0 0.718 3.561 87.9 1.6132  24 666    20.2 354.70  7.12
##     medv f.hcla    f.chas
## 365 21.9      3     River
## 366 27.5      3 Otherwise
```

```
lls<-which((df$rm>utso2)|(df$rm<utsi2));lls
```

```
## [1] 354 355
```

```
df[lls,]
```

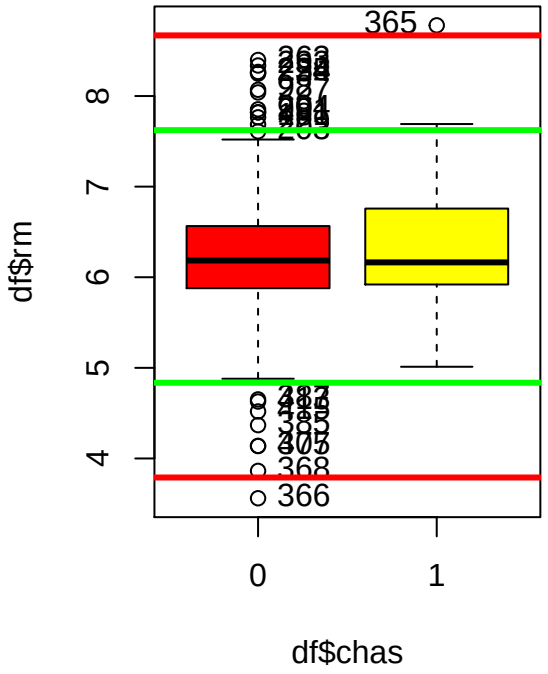
```
##        crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lstat
## 365 3.47428  0  18.1    1 0.718 8.780 82.9 1.9047  24 666    20.2 354.55  5.29
## 366 4.55587  0  18.1    0 0.718 3.561 87.9 1.6132  24 666    20.2 354.70  7.12
##     medv f.hcla    f.chas
```

```
## 365 21.9      3     River
## 366 27.5      3 Otherwise
```

```
llm<-which((df$rm>utmo2)|(df$rm<utmi2));llm
```

```
##  [1]  98  99 177 198 218 219 221 225 226 246 254 264 271 273 354 355 357 359 369
## [20] 371 391 397 399
```

```
par(mfrow=c(1,1))
```



3. Replace by NA those outliers in RM variable detected in Point 2 and use an imputation procedure discussed in class to fill outlier data points. Assess the consistency of imputed value/s.

Observations "365" (in row 354) and "366" have been considered extreme outliers and according to this a NA is set. Using the imputePCA() procedure since it is a numeric variable an imputed value of 5.9934, that can be converted to 6 rooms. One observation does not make any difference to RM distribution and pre/post imputation boxplots are exactly consistent.

```
df[c(354,355),"rm"]<-NA
imres<-imputePCA(df[c(1:3,5:14)])
names(imres)
```

```
## [1] "completeObs" "fittedX"
```

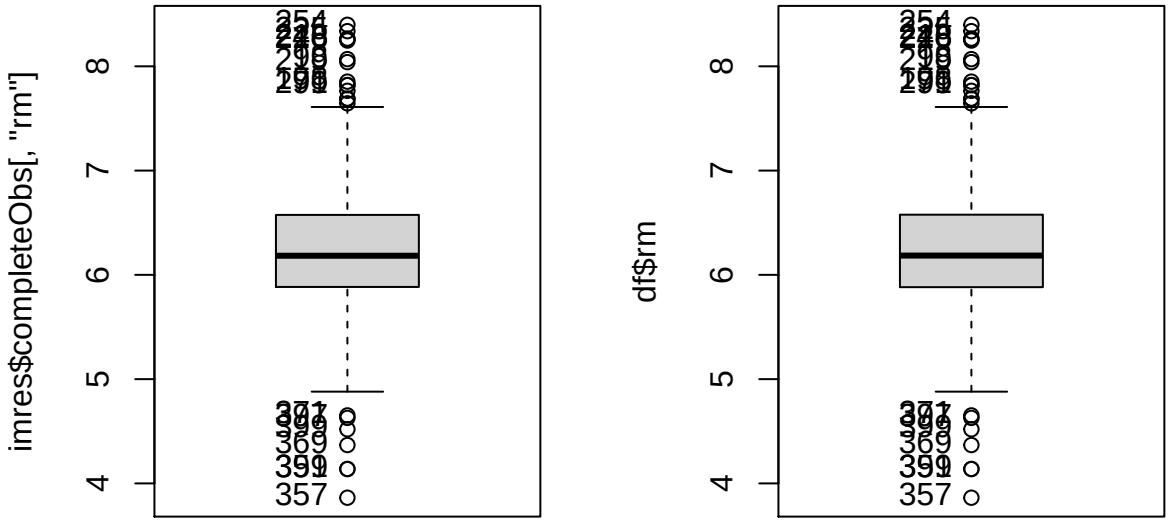
```
imres$completeObs[c(354,355),"rm"] #6.049816 6.049368
```

```
##      365      366
## 6.049816 6.049368
```

```
par(mfrow=c(1,2))
Boxplot(imres$completeObs[,"rm"])
```

```
##  [1] 357 359 369 371 391 397 399 254 225 218 246 226  98 219 198  99 271
```

```
Boxplot(df$rm)
```




```
##  [1] 357 359 369 371 391 397 399 254 225 218 246 226  98 219 198  99 271
```

```
df<-df[-c(354,355),]
```

Remove from dataset those observations with NA in RM variable (room number) in Point 3.

4. Would you expect a neighborhood that has an 'LSTAT' value (percent of lower class workers) of 15 have home prices greater or less than a neighborhood having a 20 'LSTAT' value?

According to Pearson correlation coefficient among LSTAT and MEDV (target), -0.76, increasing LSTAT (lower class workers percentage) implies decreasing home prices (MEDV). Thus, home prices when LSTAT is 15 will be greater than those on neightborhoods with LSTAT value of 20.

```
names(df)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "b"       "lstat"   "medv"
## [15] "f.hcla"  "f.chas"
```

```
cor(df[c(14,6,13,11)])
```

```
##               medv         rm      lstat    ptratio
```

```
## medv     1.0000000  0.7168454 -0.7603735 -0.5208898
## rm       0.7168454  1.0000000 -0.6308106 -0.3033846
## lstat   -0.7603735 -0.6308106  1.0000000  0.3622777
## ptratio -0.5208898 -0.3033846  0.3622777  1.0000000
```

5. Analyse the profile of the numeric target (medv) using condes() method. A detailed explanation of procedure results is requested.

There is no too much to explain. The three variables most positively correlated with home price variable (target, MEDV) are number of rooms (RM), proportion of residential land zoned for lots over 25,000ft2 and weighted distances to Boston employment centres (increasing distance means increasing prices). The three variables most inversely related to MEDV are LSTAT, INDUS and TAX, indicating that increasing lower class workers, industrial use soil and property-tax rate, decreases home prices. Factor chas seems to be no rellevant to stablish differences in mean home prices, but f.hcla has some effect (R2 0.38). Mean medv in f.hcla 1 cluster is 5.95 units over the overall mean, while in f.hcla 3 mean medv is 7.17 units under the overall mean.

```
res.con<- condes(df,14)
res.con$quanti
```

```
##         correlation      p.value
## rm        0.7168454 3.833153e-78
## zn        0.4058147 9.020237e-21
## dis       0.3714416 2.066281e-17
## b         0.3651567 7.720324e-17
## crim     -0.4505361 9.075954e-26
## rad      -0.4819929 9.368217e-30
## age      -0.4946827 1.749327e-31
## ptratio  -0.5208898 2.721787e-35
## nox      -0.5291388 1.466051e-36
## tax      -0.5780299 7.588315e-45
## indus    -0.6032887 1.055626e-49
## lstat    -0.7603735 3.851796e-93
```

```
res.con$quali
```

```
##               R2      p.value
## f.hcla 0.3788561 7.064929e-51
```

```
res.con$category
```

```
##           Estimate      p.value
## f.hcla=1  5.953647 8.035432e-20
## f.hcla=2  1.219825 3.014257e-07
## f.hcla=3 -7.173472 3.108706e-44
```

6. Analyse the profile of the binary target (chas) using a suitable method. A detailed explanation of procedure results is requested.

Mean student-teacher ratio seems to be different in section limiting with Charles river. No additional numeric variable seems to be affected by Charles river binary factor (but nox). Section tracts limiting with Charles river have a teacher-student ratio significantly less than other section tracts. We do not pay attention to chas original numeric variables since f.chas is the target output defined from numeric chas.

```
names(df)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "b"       "lstat"   "medv"
## [15] "f.hcla"  "f.chas"
```

```
res.cat <- catdes(df,num.var=16)
res.cat$quanti.var
```

```
##               Eta2     P-value
## chas    1.00000000 0.000000000
## ptratio 0.01566238 0.005632484
```

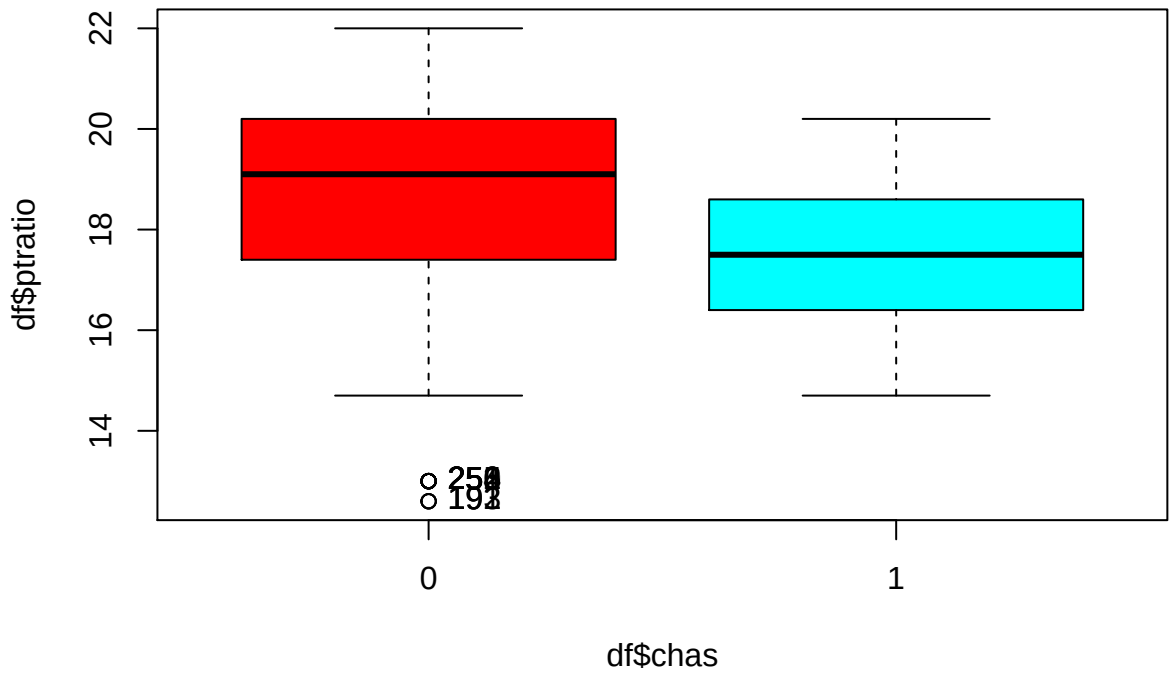
```
res.cat$quanti
```

```
## $Otherwise
##             v.test Mean in category Overall mean sd in category Overall sd
## ptratio   2.761807         18.57826  18.51311475       2.111646  2.1098889
## chas    -22.068076          0.00000   0.05737705       0.000000  0.2325617
##               p.value
## ptratio  5.748246e-03
## chas    6.405927e-108
##
## $River
##            v.test Mean in category Overall mean sd in category Overall sd
## chas    22.068076          1.00000   0.05737705       0.000000  0.2325617
## ptratio -2.761807         17.44286  18.51311475       1.764821  2.1098889
##               p.value
## chas    6.405927e-108
## ptratio  5.748246e-03
```

```
tapply(df$ptratio,df$chas,summary)
```

```
## $`0`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.10   18.58   20.20   22.00
##
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.70   16.40   17.50   17.44   18.60   20.20
```

```
Boxplot(df$ptratio~df$chas, col=rainbow(2))
```




```
##  [1] "191" "192" "193" "250" "251" "252" "253" "254" "255" "256"
```

7. Discuss whether a normal distribution would be a reasonable distribution for medv target.

Shapiro-Wilk test shows a very low pvalue, thus H0 Normally distributed medv data is clearly rejected. Histogram based assessment also discards normally distributed data.

```
shapiro.test(df$medv)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$medv
## W = 0.95914, p-value = 2.183e-10
```

```
shapiro.test(log(df$medv))
```

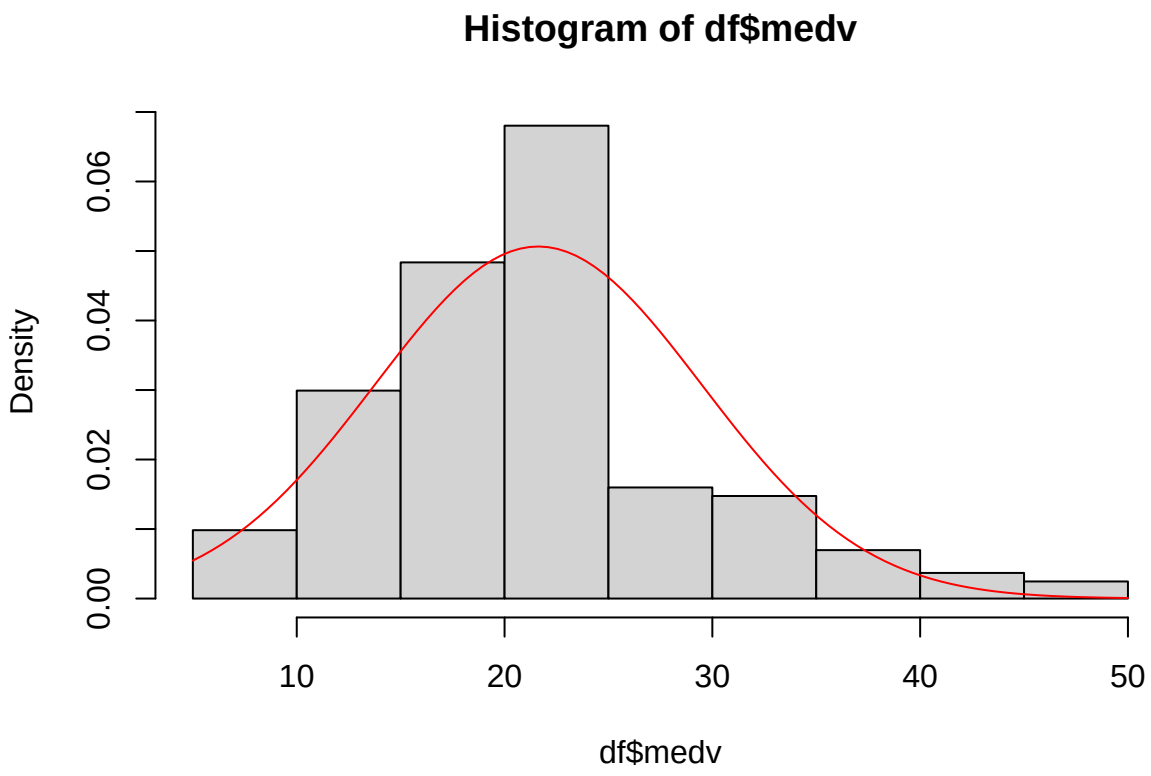
```
##
##  Shapiro-Wilk normality test
##
## data:  log(df$medv)
## W = 0.97204, p-value = 4.93e-08
```

```
hist(df$medv,freq=F,10)
mm <- mean(df$medv);ss <- sd(df$medv);mm;ss
```

```
## [1] 21.62336
```

```
## [1] 7.876935
```

```
curve(dnorm(x,mm,ss),col="red",add=T)
```



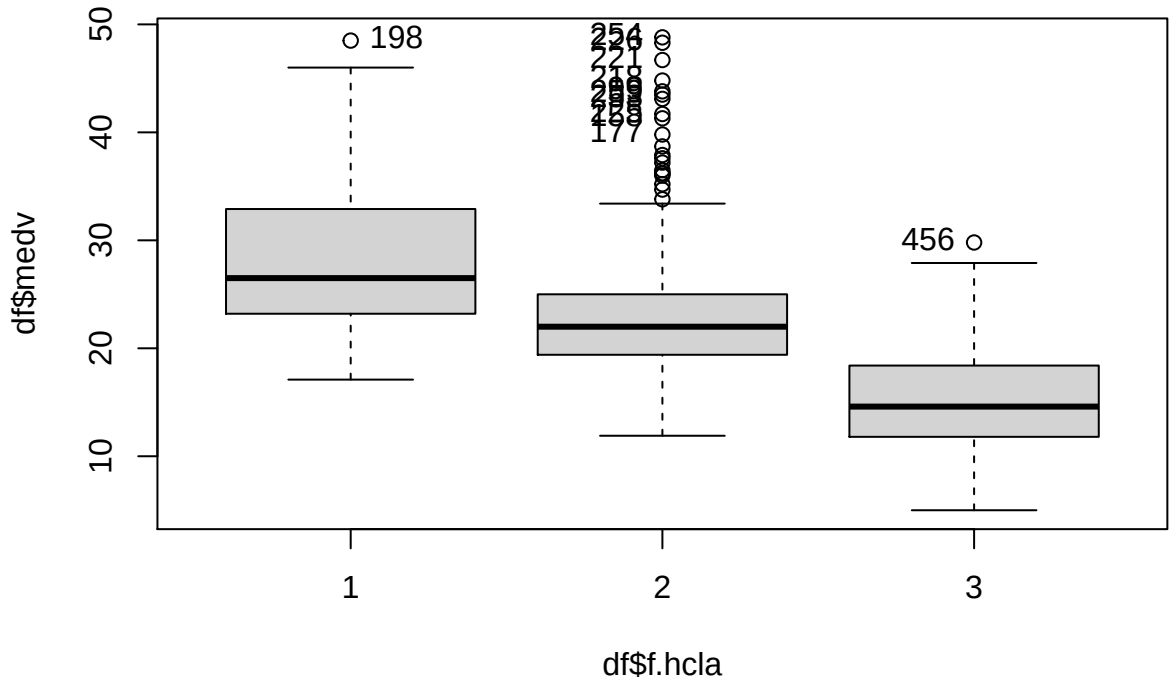
8. Is there variance homogeneity in the medv target groups defined by f.hcla clusters?

Since normal distribution for target variable (medv) can not be taken, a non-parametric test for assessing variance homogeneity across clusters defined by f.hcla has to be used. Fligner-Killeen test has been discussed in the course. Null hypothesis states variance homogeneity and p value is 0.0048 so at 5% or 1% significance level, H0 can be rejected and thus variance in groups defined by f.hcla levels can not be considered to be equal, there exists at least one group showing a remarkable different variance from the rest.

```
fligner.test(df$medv,df$f.hcla)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  df$medv and df$f.hcla
## Fligner-Killeen:med chi-squared = 10.695, df = 2, p-value = 0.004759
```

```
Boxplot(df$medv~df$f.hcla)
```




```
##  [1] "198" "254" "226" "221" "218" "99"  "259" "253" "225" "158" "177" "456"
```

```
tapply(df$medv, df$f.hcla, sd)
```

```
##        1        2        3
## 6.903295 6.687963 4.930611
```

9. Mean medv target can be considered to be the equal across groups defined by f.hcla cluster? Use a two.sided test at 99% confidence.

Since normal distribution for target variable (medv) can not be taken, a non-parametric test for assessing mean homogeneity across clusters defined by f.hcla has to be used. Kruskal-Wallis test has been discussed in the course. Null hypothesis states mean homogeneity and p value is almost 0.0 so at any significance level, H0 can be rejected and thus means in groups defined by f.hcla levels can not be considered to be equal, there exists at least one group showing a remarkable different mean from the rest. Using pairwise Wilcoxon tests medv means in groups failed to be equal for any pair of groups.

```
kruskal.test(df$medv,df$f.hcla)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$medv and df$f.hcla
## Kruskal-Wallis chi-squared = 219.1, df = 2, p-value < 2.2e-16
```

```
tapply(df$medv, df$f.hcla, summary)
```

```
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   17.10   23.23   26.50   28.18   32.85   48.50
##
## $`2`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.90   19.40   22.00   23.45   25.00   48.80
##
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   11.80   14.60   15.06   18.40   29.80
```

```
pairwise.wilcox.test(df$medv,df$f.hcla)
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  df$medv and df$f.hcla
##
##   1       2
## 2 4.9e-10 -
## 3 < 2e-16 < 2e-16
##
## P value adjustment method: holm
```

10. State and test one.sided hypothesis to assess whether medv is greater for f.hclas 1 than for class 3 or the opposite at 99% confidence.

Considering results for the pairwise Wilcoxon tests discussed in the course, using alternative 'less', mean medv in group 2 is less than mean in group 1 and the same applies for group 3 at 99% confidence.

```
pairwise.wilcox.test(df$medv,df$f.hcla, alternative="less")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  df$medv and df$f.hcla
##
##   1       2
## 2 2.5e-10 -
## 3 < 2e-16 < 2e-16
##
## P value adjustment method: holm
```

```
pairwise.wilcox.test(df$medv,df$f.hcla, alternative="greater")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  df$medv and df$f.hcla
##
##   1 2
## 2 1 -
## 3 1 1
##
## P value adjustment method: holm
```

11. The standard deviation of medv in f.hcla 1 should not exceed 10,000.$For the sample in f.hcla 1 in your dataset, calc$ **against the alternative that it is.**

We have to select the subset of observations in cluster 1 for medv. Units are included in thousands of dollars. Thus, the test should be set for variance to be less than 100 (since the standard deviations threshold is 10). pvalue for the alternative hypothesis is 1-6.826296e-06 and thus standard deviation can be considered to be under 10 000 $.

```
tapply(df$medv, df$f.hcla, sd)
```

```
##        1        2        3
## 6.903295 6.687963 4.930611
```

```
tapply(df$medv, df$f.hcla, var)
```

```
##        1        2        3
## 47.65549 44.72884 24.31092
```

```
ll<-which(df$f.hcla =="1");length(ll)
```

```
## [1] 90
```

```
library(EnvStats)
```

```
## Warning: package 'EnvStats' was built under R version 4.0.5
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following object is masked from 'package:car':
##
##     qqPlot
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## The following object is masked from 'package:base':
##
##     print.default
```

```
varTest(df$medv[ll],sigma.squared=100,conf.level=0.99,alternative="less")
```

```
## $statistic
## Chi-Squared
##    42.41338
##
## $parameters
## df
## 89
##
## $p.value
## [1] 6.826296e-06
##
## $estimate
## variance
## 47.65549
##
## $null.value
## variance
##      100
##
## $alternative
```

```
## [1] "less"
##
## $method
## [1] "Chi-Squared Test on Variance"
##
## $data.name
## [1] "df$medv[ll]"
##
## $conf.int
##      LCL      UCL
##  0.00000 69.61224
## attr(,"conf.level")
## [1] 0.99
##
## attr(,"class")
## [1] "htestEnvStats"
```

12. Figure out the 99% upper threshold for medv in f.hcla 1 population variance. Normal distribution for medv is assumed to hold.

Upper threshold at 99% confidence based on variance for group 1 can be obtained directly from theoretical formulae or using varTest() method in EnvStats library, Threshold is 69.61 (one-sided point of view) or 72.64 according to two-sided point of view (any of both approaches have been taken as correct).

```
varTest(df$medv[ll],conf.level=0.99,alternative="less")
```

```
## $statistic
## Chi-Squared
##    4241.338
##
## $parameters
## df
## 89
##
## $p.value
## [1] 1
##
## $estimate
## variance
## 47.65549
##
## $null.value
## variance
##        1
##
## $alternative
## [1] "less"
##
## $method
## [1] "Chi-Squared Test on Variance"
##
## $data.name
## [1] "df$medv[ll]"
##
## $conf.int
##      LCL      UCL
```

```
##  0.00000 69.61224
## attr(,"conf.level")
## [1] 0.99
##
## attr(,"class")
## [1] "htestEnvStats"
```

```
89*47.65549/qchisq(0.01,89)
```

```
## [1] 69.61225
```

```
varTest(df$medv[ll],conf.level=0.99)
```

```
## $statistic
## Chi-Squared
##    4241.338
##
## $parameters
## df
## 89
##
## $p.value
## [1] 0
##
## $estimate
## variance
## 47.65549
##
## $null.value
## variance
##        1
##
## $alternative
## [1] "two.sided"
##
## $method
## [1] "Chi-Squared Test on Variance"
##
## $data.name
## [1] "df$medv[ll]"
##
## $conf.int
##      LCL      UCL
## 33.36844 72.63963
## attr(,"conf.level")
## [1] 0.99
##
## attr(,"class")
## [1] "htestEnvStats"
```

```
89*47.65549/qchisq(0.005,89)
```

```
## [1] 72.63964
```

13. Build a 99% two-sided confidence interval for the difference in the mean of medv between f.hcla 1 and 3. Assume that equal variances in the population medv does not hold and normal distribution of medv (to simplify the calculations), but justify if these assumptions are critical.

A 99% CI for mean medv difference between group 1 and 3 is [10.97,15.29]. Normal distribution does not hold, but it is not critical, independence among sample is critical.

```
ll3 <- which(df$f.hcla=="3");length(ll3)
```

```
## [1] 157
```

```
t.test(df$medv[ll],df$medv[ll3],conf.level=0.99, var.equal=F,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$medv[ll] and df$medv[ll3]
## t = 15.868, df = 141.75, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  10.96719 15.28705
## sample estimates:
## mean of x mean of y
##  28.18444  15.05732
```

14. Determine a 99% confidence interval for the population proportion that favors Riverside in front of Otherwise. Test the null hypothesis that selecting Riverside and Otherwise zones has equal probability.

A two-sided test based on normal approximation is set. A 99% CI for Riverside preference is [0.036, 0.091]. Null hypothesis stating equal preference vs H1 not equal shows a pvalue of 0 and thus it can be rejected: Riverside locations are not as likely as Otherwise in the sample.

```
prop.table(table(df$f.chas))
```

```
##
##  Otherwise      River
## 0.94262295 0.05737705
```

```
prop.test(x=28, n=488,p=0.5,conf.level=0.99, correct=F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  28 out of 488, null probability 0.5
## X-squared = 382.43, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.03573292 0.09089562
## sample estimates:
##          p
## 0.05737705
```

15. A new survey considered 300 people, 110 prefer Riverside to Otherwise locations. Determine a 99% confidence interval for the difference in the population proportion that favors Riverside in front of other areas accounting the two sources. Test the null hypothesis that selecting Riverside zones has a greater probability in the survey.

Current sample shows 28 out of 488 units indicating Riverside preference, while new survey data indicates 110 out of 300. A 99% CI two-sided for proportion 1 minus proportion 2 Riverside choice lies between [-0.386,-0.233] thus 0 is not contained. Testing null hypothesis of equal Riverside share versus current data probability less than survey probability indicates a pvalue almost 0, H0 is rejected and thus current sample proportion is less than survey proportion.

```r
prop.test( c(28, 110), c(488, 300), correct = F, conf.level = 0.99)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(28, 110) out of c(488, 300)
## X-squared = 123.03, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.3859137 -0.2326656
## sample estimates:
##     prop 1     prop 2
## 0.05737705 0.36666667
```

```r
prop.test( c(28, 110), c(488, 300), correct = F, conf.level = 0.99, alternative = "less")
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(28, 110) out of c(488, 300)
## X-squared = 123.03, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 99 percent confidence interval:
##  -1.000000 -0.240087
## sample estimates:
##     prop 1     prop 2
## 0.05737705 0.36666667
```

Do not forget to knit your .Rmd file to .pdf (or to word and afterwards to pdf) before posting it on the ATENEA platform