22-23-SIM-PARTIAL Template

LÍdia Montero

November, 3rd 2022

Contents

1 Boston Housing Data

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

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

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Warning in register(): Can't find generic 'scale_type' in package ggplot2 to
## register S3 method.
```

```
##
## 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
```

```
## corrplot 0.92 loaded
```

1. Determine thresholds for mild and severe outliers for wage. Are there any outliers? Indicate observation id's and atypical number/s.

Severe outliers are in the range (Q1-3IQR, Q3+3IQR) and mild outliers are in the range (Q1-1.5IQR, Q3+1.5IQR), where IQR is Q3-Q1.

We can see that there are not severe outliers. There are many mild outliers.

```
summary(df)
```

```
##     gender        ethnicity        score        fcollege   mcollege      home
## male  :2139   other   :3050   Min.   :28.95   no :3753   no :4088   no : 852
## female:2600   afam    : 786   1st Qu.:43.92   yes: 986   yes: 651   yes:3887
##               hispanic: 903   Median :51.19
##                               Mean   :50.89
##                               3rd Qu.:57.77
##                               Max.   :72.81
## urban          unemp            wage          distance         tuition
## no :3635   Min.   : 1.400   Min.   : 6.590   Min.   : 0.000   Min.   :0.2575
## yes:1104   1st Qu.: 5.900   1st Qu.: 8.850   1st Qu.: 0.400   1st Qu.:0.4850
##            Median : 7.100   Median : 9.680   Median : 1.000   Median :0.8245
##            Mean   : 7.597   Mean   : 9.501   Mean   : 1.803   Mean   :0.8146
##            3rd Qu.: 8.900   3rd Qu.:10.150   3rd Qu.: 2.500   3rd Qu.:1.1270
##            Max.   :24.900   Max.   :12.960   Max.   :20.000   Max.   :1.4042
##    education      income        region
## Min.   :12.00   low :3374   other:3796
## 1st Qu.:12.00   high:1365   west : 943
## Median :13.00
## Mean   :13.81
## 3rd Qu.:16.00
## Max.   :18.00
```

```
summa <- summary(df$wage)
summa
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.590   8.850   9.680   9.501  10.150  12.960
```

```
sev_out_ut <- summa[5]+ 3*(summa[5]- summa[2])
sev_out_ut
```

```
## 3rd Qu.
##   14.05
```

```
sev_out_lt <- summa[2] -3*(summa[5]- summa[2])
sev_out_lt
```

```
##  1st Qu.
## 4.950003
```

```
mild_out_ut <- summa[5] + 1.5*(summa[5]-summa[2])
mild_out_ut
```
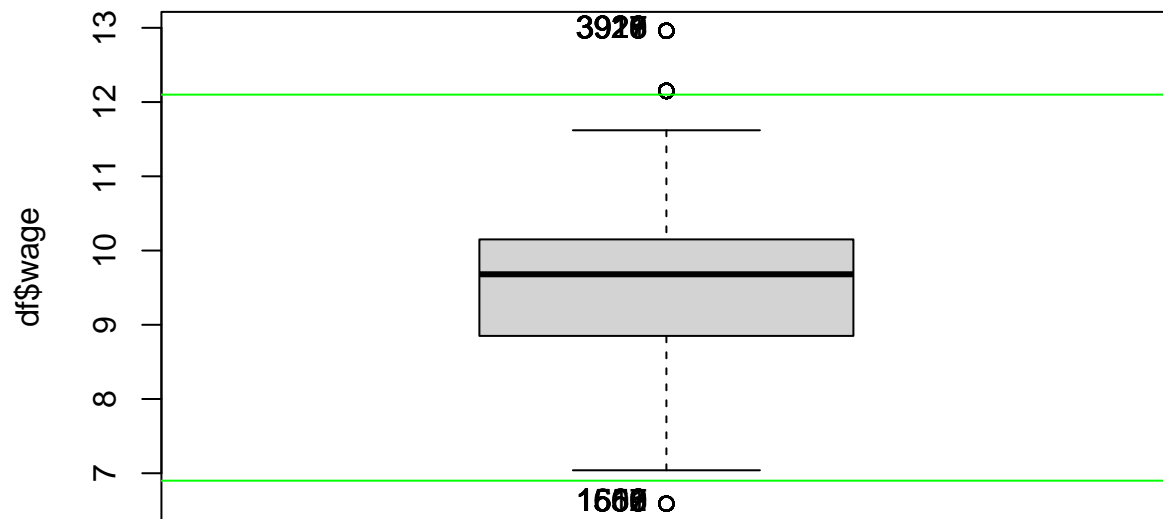
```
## 3rd Qu.
##    12.1
```

```
mild_out_lt <- summa[2] -1.5*(summa[5]- summa[2])
mild_out_lt
```

```
##  1st Qu.
## 6.900002
```

```
Boxplot(df$wage)
```

```
##  [1] 1555 1556 1557 1558 1559 1560 1561 1602 1609 1610 3916 3917 3918 3919 3920
## [16] 3921 3926 3927 3928 3929
```

```
abline(h=sev_out_ut, col = "red")
abline(h=sev_out_lt, col = "red")
abline(h=mild_out_ut, col = "green")
abline(h=mild_out_lt, col = "green")
```

```
llsev <- which((df$wage > sev_out_ut) | (df$wage < sev_out_lt))
llsev # no severe outliers
```

```
## integer(0)
```

```
llmild <- which((df$wage > mild_out_ut) | (df$wage < mild_out_lt))
llmild
```

```
##    [1] 1555 1556 1557 1558 1559 1560 1561 1602 1609 1610 1611 1612 1613 1614 1615
##   [16] 1616 1623 1624 1625 1626 1627 1654 1660 1661 1662 1663 1664 1665 1666 1667
##   [31] 1668 2051 2052 2053 2054 2055 2059 2060 2061 2062 2063 2064 2065 2066 2077
##   [46] 2078 2079 2080 2081 2082 2091 2092 2093 2094 2095 2096 2097 2098 2099 2100
##   [61] 2101 2102 2103 2123 2124 2125 2126 2127 2128 2129 2130 2131 2132 2133 2134
##   [76] 2135 2136 2137 2138 2139 2140 2141 2142 2143 2144 2145 2146 2164 2165 2166
##   [91] 2167 2168 2169 2170 2171 2172 2173 2174 2175 2176 2177 2178 2179 2180 2201
##  [106] 2202 2203 2204 2205 2206 2207 2240 2241 2242 2243 2244 2279 2280 2281 2282
##  [121] 2283 2284 2285 2286 2287 2288 2289 2290 2291 2292 2293 2294 2295 2296 2297
##  [136] 2298 2299 2300 2301 2302 2303 2304 2323 2324 2325 2326 2327 2328 2329 2330
##  [151] 2331 2332 2333 2354 2355 2356 2357 2358 2359 2360 2361 2362 3284 3285 3307
##  [166] 3312 3313 3314 3315 3316 3317 3318 3458 3459 3460 3461 3462 3463 3464 3465
##  [181] 3466 3467 3468 3469 3470 3595 3596 3597 3598 3599 3600 3645 3646 3647 3648
##  [196] 3669 3670 3671 3672 3673 3674 3675 3676 3677 3678 3679 3680 3916 3917 3918
##  [211] 3919 3920 3921 3926 3927 3928 3929 3930 3931 3932 3933 3934 3935 3936 3937
##  [226] 3938 3939 3940 3958 3959 3960 3966 3967 3968 3969 3970 3971 3974 3975 3976
##  [241] 3977 3978 3979 3980 3981 3987 3988 3989 3990 3991 3992 4596 4597 4598 4599
##  [256] 4600 4601 4602 4625 4626 4627 4628 4629 4630 4631 4632 4633 4634 4635 4636
##  [271] 4637 4638
```

2. Replace by NA mild outliers in wage variable detected in Point 1 and use an imputation procedure discussed in class to fill outlier data points. Assess the consistency of imputed value/s.

I can use imputeMCA to do the imputation.

```
df[llmild, "wage"] <- NA
# library(missMDA)
# imp_res<-imputeMCA(df,method="EM")
# imp_res$completeObs
#
# Boxplot(df[imp_res$completeObs,])
```

Remove from dataset those observations with NA in wage, those labelled as mild outliers.

3. Would you expect a family paying tuition fees of 900$ to have a shorter or longer college distance than a family paying 270$?

We can use the cor to see the correlation. We can see that tuition and distance and inversely correlated so when the distance goes up the tuition fee goes down. So, a family paying $900 has a shorter distance than a family paying 270 for tuition fees.

```
# removal of mild outliers from dataset
df <- df[-llmild,]

cor(df[c(3,8:11)])
```

```
##               score        unemp        wage      distance      tuition
## score     1.00000000 -0.03761297 0.09630618 -0.05916241  0.1269400
## unemp    -0.03761297  1.00000000 0.15616707  0.31770842  0.1235313
## wage      0.09630618  0.15616707 1.00000000  0.01215047  0.2931288
## distance -0.05916241  0.31770842 0.01215047  1.00000000 -0.1004843
## tuition   0.12694001  0.12353135 0.29312877 -0.10048425  1.0000000
```

```
corrplot(cor(df[c(3,8:11)]))
```
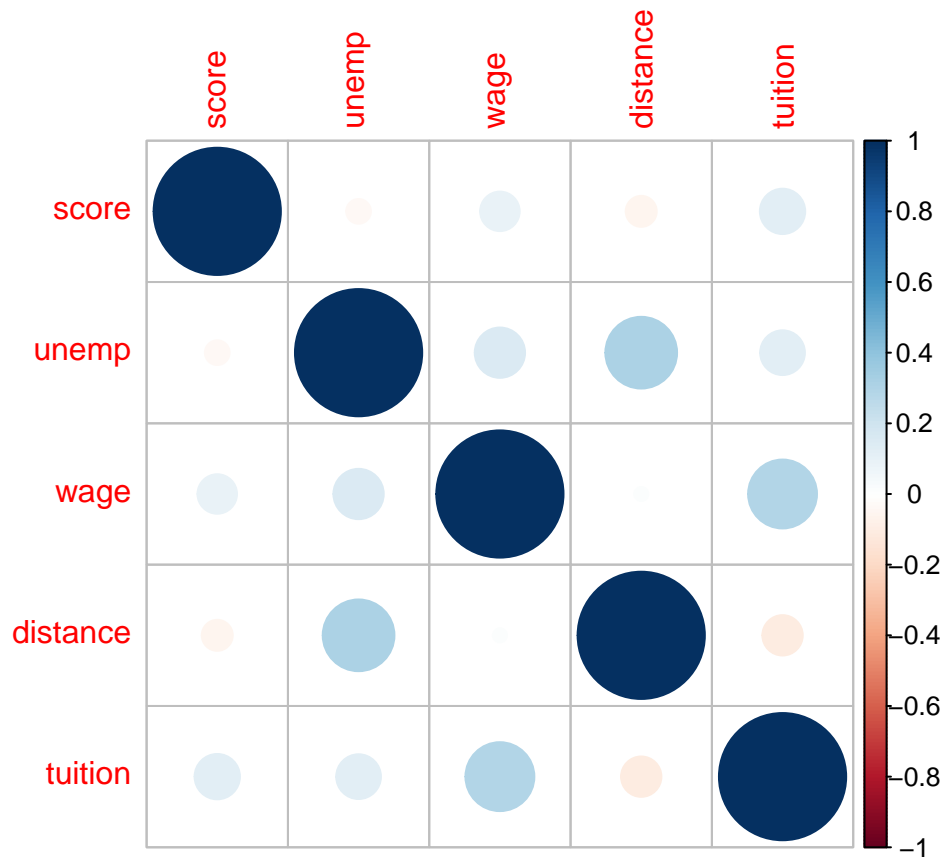

4. Analyse the profile of the numeric target (distance) using condes() method. A detailed explanation of procedure results is requested.

Using condes we can see that unemp is positive correlated to distance. And that score, education and tuition are negatively correlated to distance.

The categorical variables most correlated to distance are urban and ethnicity. If the school is not in an urban area is the most correlated to distance follow by the students who father is not a college graduated.

```
res.con = condes(df,10)
res.con$quanti
```

```
##            correlation       p.value
## unemp       0.31770842 2.563300e-105
## score      -0.05916241  7.605090e-05
## education  -0.08622410  7.834560e-09
## tuition    -0.10048425  1.683974e-11
```

```
res.con$quali
```

```
##                  R2        p.value
## urban     0.085772732 4.599764e-89
## ethnicity 0.012944454 2.346583e-13
## fcollege  0.010030340 1.963762e-11
## income    0.006562367 5.905276e-08
## mcollege  0.005728826 4.093197e-07
## region    0.004517220 6.928709e-06
```

```
res.con$category
```

```
##                     Estimate     p.value
## urban=no           0.7845767 4.599764e-89
## fcollege=no        0.2830955 1.963762e-11
## income=low         0.2054104 5.905276e-08
## mcollege=no        0.2525074 4.093197e-07
## region=west        0.1929407 6.928709e-06
## ethnicity=hispanic 0.3441993 7.682709e-05
## ethnicity=other    0.1330294 2.085174e-02
## region=other      -0.1929407 6.928709e-06
## mcollege=yes      -0.2525074 4.093197e-07
## income=high       -0.2054104 5.905276e-08
## fcollege=yes      -0.2830955 1.963762e-11
## ethnicity=afam    -0.4772287 4.735870e-13
## urban=yes         -0.7845767 4.599764e-89
```

5. Analyse the profile of the categorical target (income) using a suitable method. A detailed explanation of procedure results is requested when profiling low income (high income may be omitted).

For this we can use catdes.

When the income is low we can see that the parents are not college graduated and that the family does not own a house. When the income is high it is the other way around.

The numerical variables most associated are education and score.

```
res.cat=catdes(df,13)
res.cat$test.chi2
```

```
##              p.value df
## fcollege  5.016368e-128  1
## mcollege  1.202661e-62   1
## ethnicity 6.261733e-32   2
## home      1.819516e-19   1
## urban     3.003815e-07   1
## gender    2.596250e-04   1
```

```
res.cat$category
```

```
## $low
```

```
##                       Cla/Mod   Mod/Cla   Global         p.value        v.test
## fcollege=no          80.03941 88.622195 79.51645 2.162287e-117   23.033310
## mcollege=no          76.26197 91.832918 86.47862   1.106595e-56   15.865033
## home=no              84.58738 21.726933 18.44638   3.393648e-21    9.449835
## ethnicity=hispanic   83.42728 23.067332 19.85673   4.128504e-19    8.933470
## ethnicity=afam       81.20805 18.859102 16.67786   1.188209e-10    6.440831
## urban=yes            77.98683 25.841646 23.79673   1.921228e-07    5.206802
## gender=female        74.04238 56.639651 54.93620   2.660478e-04    3.646296
## gender=male          69.10084 43.360349 45.06380   2.660478e-04   -3.646296
## urban=no             69.88837 74.158354 76.20327   1.921228e-07   -5.206802
## home=yes             68.92671 78.273067 81.55362   3.393648e-21   -9.449835
## ethnicity=other      65.71429 58.073566 63.46541   2.294132e-34  -12.224705
## mcollege=yes         43.37748  8.167082 13.52138   1.106595e-56  -15.865033
## fcollege=yes         39.89071 11.377805 20.48355 2.162287e-117  -23.033310
##
## $high
##                       Cla/Mod  Mod/Cla   Global         p.value        v.test
## fcollege=yes         60.10929 43.68546 20.48355 2.162287e-117   23.033310
## mcollege=yes         56.62252 27.16442 13.52138   1.106595e-56   15.865033
## ethnicity=other      34.28571 77.20413 63.46541   2.294132e-34   12.224705
## home=yes             31.07329 89.91263 81.55362   3.393648e-21    9.449835
## urban=no             30.11163 81.41382 76.20327   1.921228e-07    5.206802
## gender=male          30.89916 49.40429 45.06380   2.660478e-04    3.646296
## gender=female        25.95762 50.59571 54.93620   2.660478e-04   -3.646296
## urban=yes            22.01317 18.58618 23.79673   1.921228e-07   -5.206802
## ethnicity=afam       18.79195 11.11994 16.67786   1.188209e-10   -6.440831
## ethnicity=hispanic   16.57272 11.67593 19.85673   4.128504e-19   -8.933470
## home=no              15.41262 10.08737 18.44638   3.393648e-21   -9.449835
## mcollege=no          23.73803 72.83558 86.47862   1.106595e-56  -15.865033
## fcollege=no          19.96059 56.31454 79.51645 2.162287e-117  -23.033310
```

res.cat$quanti.var

```
##                 Eta2       P-value
## education 0.047172646 7.717142e-49
## score     0.030197245 1.272546e-31
## unemp     0.014955329 2.364714e-16
## distance  0.006562367 5.905276e-08
## tuition   0.001649722 6.627419e-03
## wage      0.001586851 7.750996e-03
```

res.cat$quanti

```
## $low
##           v.test Mean in category Overall mean sd in category Overall sd
## unemp    8.172545        7.5929863    7.3941571      2.6761165  2.5952927
## distance 5.413643        1.8957294    1.7799418      2.4042040  2.2815827
## wage    -2.662119        9.3365992    9.3659750      1.1773209  1.1771346
## tuition -2.714343        0.7943633    0.8030668      0.3380959  0.3420502
## score  -11.612962       49.8757045   50.8232595      8.6505638  8.7041239
## education -14.514580     13.5676434   13.8110589      1.7245948  1.7889879
##              p.value
## unemp    3.019502e-16
```

```
## distance   6.175517e-08
## wage        7.765051e-03
## tuition     6.640728e-03
## score       3.541299e-31
## education 9.795568e-48
##
## $high
##             v.test Mean in category Overall mean sd in category Overall sd
## education 14.514580      14.4312947   13.8110589      1.8003633  1.7889879
## score     11.612962      53.2376807   50.8232595      8.3673778  8.7041239
## tuition    2.714343       0.8252436    0.8030668      0.3509506  0.3420502
## wage       2.662119       9.4408261    9.3659750      1.1733399  1.1771346
## distance  -5.413643       1.4849087    1.7799418      1.9027244  2.2815827
## unemp     -8.172545       6.8875298    7.3941571      2.3005497  2.5952927
##               p.value
## education 9.795568e-48
## score     3.541299e-31
## tuition   6.640728e-03
## wage      7.765051e-03
## distance  6.175517e-08
## unemp     3.019502e-16
```

6. Discuss whether a normal distribution would be a reasonable distribution for distance target.

To see if a normal distribution is reasonable we can see it graphically with an histogram and we can use the shapiro test which null hypothesis is that there is normal distribution.

From the histogram we can see that it does not follow a normal distribution. The result for shapiro test is a p-value «< 0 so we reject the null hypothesis. Thus it does not follow a normal distribution.

```
mm <- mean(df$distance)
ss <- sd(df$distance)
hist(df$distance)
```

Histogram of df$distance
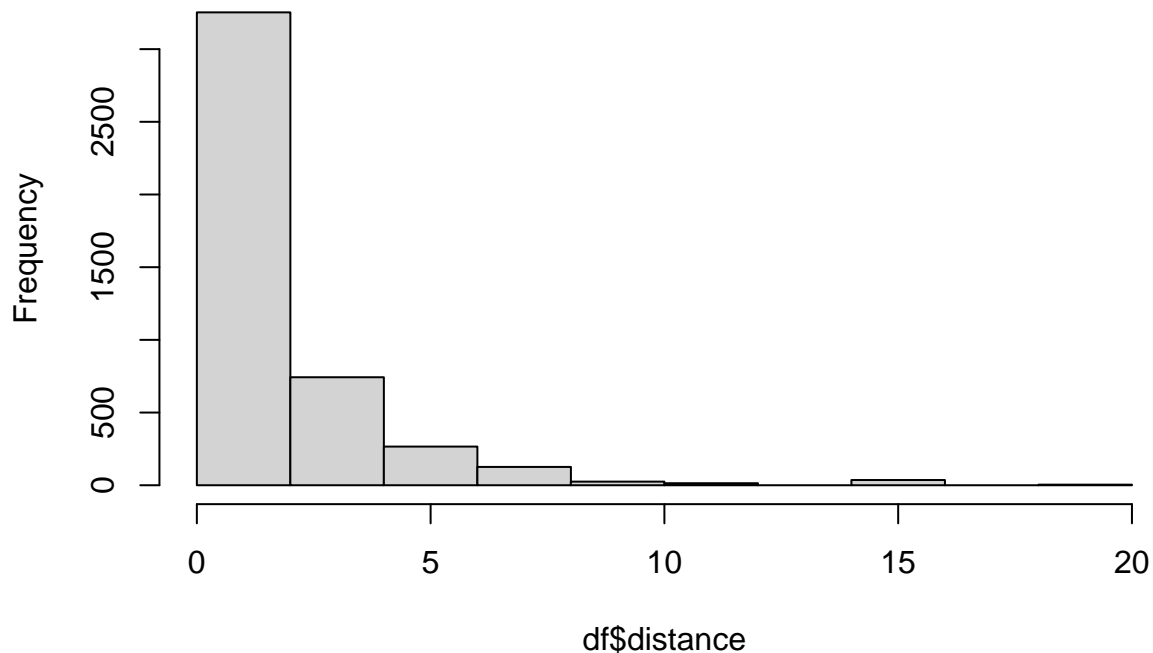

```
shapiro.test(df$distance)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$distance
## W = 0.68916, p-value < 2.2e-16
```

7. Is there variance homogeneity in the distance target groups defined by ethnicity levels? Assess race characteristics.

Since we saw that it does not follow a normal distribution we can use a non parametric test. We can use fligner test. The null hypothesis is that there is variance homogeneity. We get a p-value < 2.2e-16 so we reject the null hypothesis. So there is at least one group with variance significantly different.

```
tapply(df$distance, df$ethnicity, sd)
```

```
##    other     afam hispanic
## 2.076306 1.684117 3.120025
```

```
Boxplot(df$distance~df$ethnicity)
```
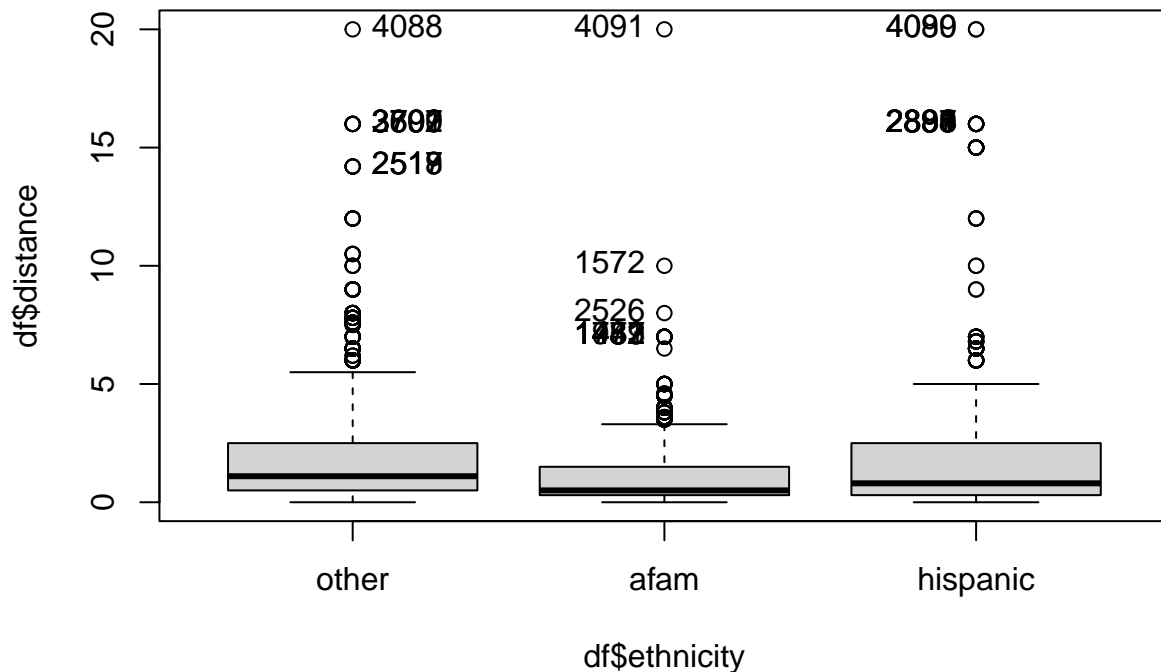
```
##  [1] "4088" "2892" "3697" "3699" "3700" "3701" "3702" "2517" "2518" "2519"
## [11] "4091" "1572" "2526" "732"  "942"  "1457" "1479" "1481" "1482" "1483"
## [21] "4089" "4090" "2887" "2888" "2889" "2890" "2891" "2893" "2894" "2895"
```

```
fligner.test(df$distance~df$ethnicity)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  df$distance by df$ethnicity
## Fligner-Killeen:med chi-squared = 133.46, df = 2, p-value < 2.2e-16
```

8. Distance target can be considered to be the equal across groups defined by ethnicity levels? Use a two.sided test at 99% confidence.

We can use pairwise.wilcox test to asses this. We see that it is not equal among the groups.

```
pairwise.wilcox.test(df$distance,df$ethnicity, exact=F)
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  df$distance and df$ethnicity
##
```

```
##          other    afam
## afam       < 2e-16 -
## hispanic 0.0012  4.2e-07
##
## P value adjustment method: holm
```

9. State and test one.sided hypothesis to assess whether distance is greater for the afroamerican group than for the Hispanic group or the opposite at 99% confidence.

10. The standard deviation of distance for the afam group should not exceed 2 miles. For the sample in afam group in your dataset, calculate the standard deviation of distance assuming a normal distribution. Stating any assumptions, you need (write them), test at the 1% level the null hypothesis that the population standard deviation is not larger than 2 miles against the alternative that it is.

Distance is in 10 miles. So $10/2 = 5$

To test this we can use varTest. H0: ss = 5^2 H1: ss > 5^2

```
llafam <- df[df$ethnicity == "afam",]
varTest(llafam$distance, alternative="greater",conf.level =0.99 , sigma.squared = 5^2)
```

```
## $statistic
## Chi-Squared
##      84.40681
##
## $parameters
##   df
## 744
##
## $p.value
## [1] 1
##
## $estimate
## variance
##   2.83625
##
## $null.value
## variance
##        25
##
## $alternative
## [1] "greater"
##
## $method
## [1] "Chi-Squared Test on Variance"
##
## $data.name
## [1] "llafam$distance"
##
## $conf.int
##      LCL       UCL
## 2.522112       Inf
## attr(,"conf.level")
## [1] 0.99
```

```
##
## attr(,"class")
## [1] "htestEnvStats"
```

11. Figure out the 99% upper threshold for distance in afam ethnicity subpopulation variance. Normal distribution for distance is assumed to hold.

Upper threshold is 3.210785.

```
varTest(llafam$distance, alternative="less",conf.level =0.99 , sigma.squared = 5^2)
```

```
## $statistic
## Chi-Squared
##     84.40681
##
## $parameters
##  df
## 744
##
## $p.value
## [1] 9.634074e-211
##
## $estimate
## variance
##   2.83625
##
## $null.value
## variance
##       25
##
## $alternative
## [1] "less"
##
## $method
## [1] "Chi-Squared Test on Variance"
##
## $data.name
## [1] "llafam$distance"
##
## $conf.int
##      LCL      UCL
## 0.000000 3.210785
## attr(,"conf.level")
## [1] 0.99
##
## attr(,"class")
## [1] "htestEnvStats"
```

12. Build a 99% two-sided confidence interval for the difference in the mean of distance between Hispanic and Afroamerican ethnicity groups. Assume that equal variances in the population distance does not hold and normal distribution of distance (to simplify the calculations) does hold, but justify if these assumptions are critical

13. Determine a 99% confidence interval for the population proportion that represents a low income. Test the null hypothesis that low and high income groups have equal probability

We can use prop.test to test this. We get a p-value $< 2.2e\text{-}16$ so we reject the null hypothesis

```
low_income <- df[df$income == "low",]
prop.test( x= 3208, n = 4467 , p=0.5, conf.level = 0.99)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  3208 out of 4467, null probability 0.5
## X-squared = 849.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.7003888 0.7352706
## sample estimates:
##         p
## 0.7181554
```

14. A new survey considered 1000 people, 660 were classified in the low income group. Determine a 99% confidence interval for the difference in the population proportion of low income accounting for the two sources. Test the null hypothesis that having a low income has a lower incidence in the survey than in the original sample

I can use the prop.test to test this. We get a p-value of 0.999 so we fail to reject the null hypothesis.

```
prop.test( c(3208, 660), c(4467, 1000) ,alternative = "less", conf.level = 0.99)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(3208, 660) out of c(4467, 1000)
## X-squared = 13.074, df = 1, p-value = 0.9999
## alternative hypothesis: less
## 99 percent confidence interval:
##  -1.00000000  0.09697269
## sample estimates:
##    prop 1    prop 2
## 0.7181554 0.6600000
```

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