TOTAL ……… /50 PT

EXAM RULES

a) BEFORE starting to solve the problems you are required to sign **all sheets** of the exam (on top in the header) and below the exam rules. Signing below the exam rules means its acceptance. Only students who accept the exam rules can take part in it.

b) One has to solve **all problems**.

c) Exam lasts **90 minutes**.

d) Each noticed attempt of cheating means immediate turning out of the exam, information to the Dean and a request for disciplinary measures to the University Disciplinary Commission. Above consequences apply also to writing the exam after its time is over.

e) To obtain passing final grade one needs to collect **at least 50%** of points.

Warsaw, 2020-03-02,

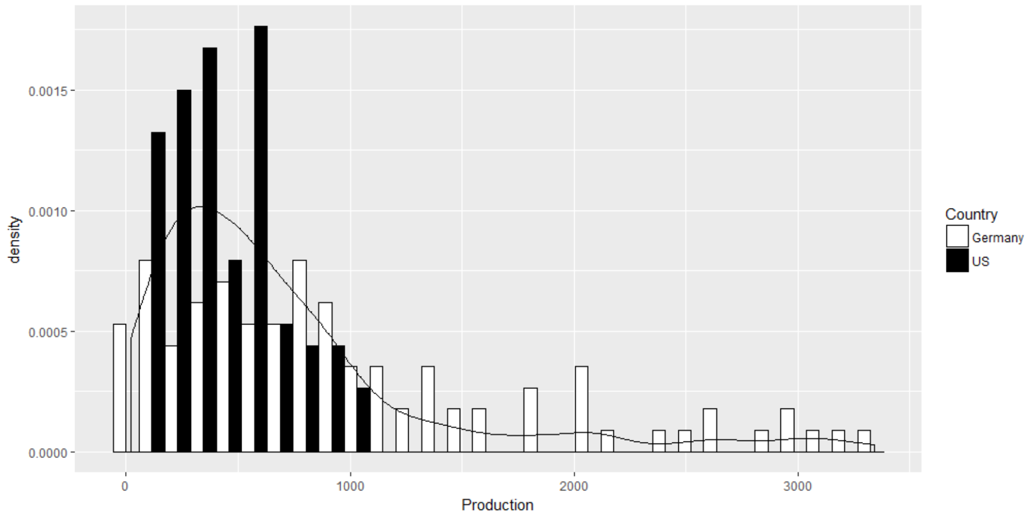
…………………………………..
SIGNATURE

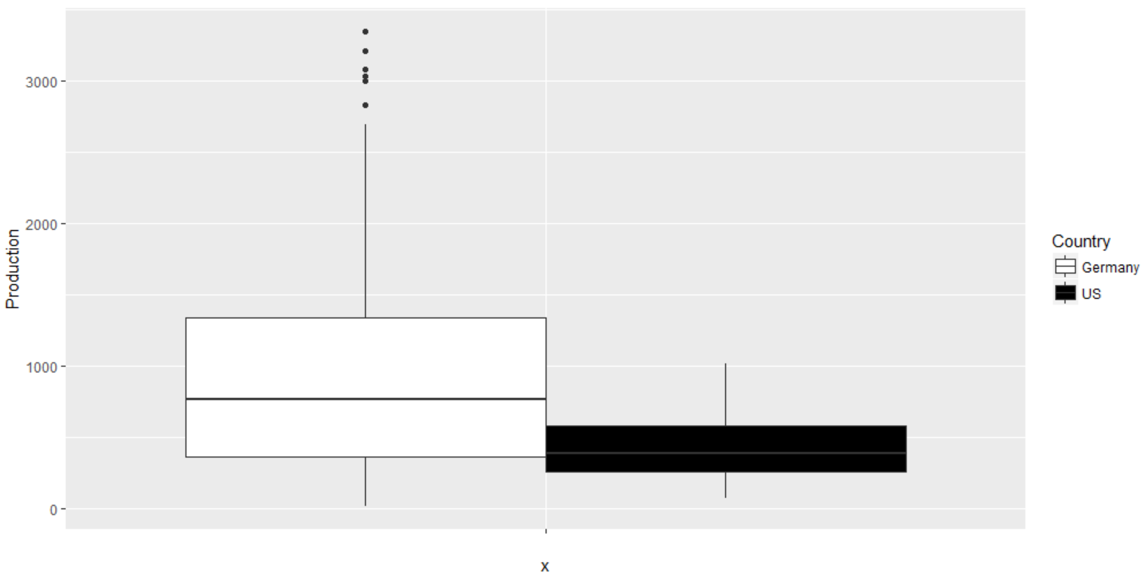
Statistics and Explanatory Data Analysis, final exam 2020-03-02

PROBLEM 1 ……… /10 PTS

You have data on breweries' productions in the US and Germany. There are 100 observations of the amount of beer produced for the US, and 100 observations of the amount produced in Germany. While in the US the beer production is measured in pints, in Germany it is measured in litres. You have ran graphical and descriptive analysis of your data. The output is given below.

1. Which location measures would you use to summarize your data for each country? Choose the measure/s and interpret the value/s.
2. In which country is the variation in the amount of beer produced smaller? Choose relevant measure/s and interpret the value/s.
3. Calculate the percentage of the total sample that approximately falls into the interval of 350 - 765 (for each country).





Statistics and Explanatory Data Analysis, final exam 2020-03-02

```
###US####
#means
mean(us$Production)
[1] 441.0101
mean(us$Production, trim=0.1)
[1] 423.7037
mean(us$Production, trim=0.2)
[1] 414.377
winsor.mean(us$Production, trim=0.1)
[1] 431.8889
winsor.mean(us$Production, trim=0.2)
[1] 414.4848
#midrange
(min(us$Production)+max(us$Production))/2
[1] 547.5
#trimean
TMH(us$Production)
[1] 243.25
#mode
names(sort(-table(us$Production)))[1]
[1] "77"
#median
median(us$Production)
[1] 388
#quantiles
quantile(us$Production, probs=c(0.25, 0.5, 0.7
5))
  25%   50%   75% 
255.0 388.0 583.5 
quantile(us$Production, probs=c(0.1, 0.2, 0.3,
0.4))
  10%   20%   30%   40% 
142.6 214.4 283.0 350.6 
quantile(us$Production, probs=c(0.5, 0.6, 0.7,
0.8, 0.9))
  50%   60%   70%   80%   90% 
388.0 483.0 554.0 615.2 796.0 
quantile(us$Production, probs=c(0.01, 0.24, 0.
31))
    1%    24%    31% 
 76.96 251.04 284.14 
quantile(us$Production, probs=c(0.56, 0.67, 0.
88, 0.99))
    56%     67%     88%     99% 
 453.44  544.30  764.80 1015.10 
#range
range(us$Production)
[1]   75 1020
#interquartile range
IQR(us$Production)
[1] 328.5
#variance and standard deviation
var(us$Production)
[1] 59950.64
sd(us$Production)
[1] 244.8482
#MAD
mad(us$Production)
[1] 255.0072
```

```
###Germany####
#means
mean(germany$Production)
[1] 979.6869
mean(germany$Production, trim=0.1)
[1] 865.2222
mean(germany$Production, trim=0.2)
[1] 793.7541
winsor.mean(germany$Production, trim=0.1)
[1] 917.4141
winsor.mean(germany$Production, trim=0.2)
[1] 839.6263
#midrange
(min(germany$Production)+max(germany$Produc
tion))/2
[1] 1681.5
#trimean
TMH(germany$Production)
[1] 375.25
#mode
names(sort(-table(germany$Production)))[1]
[1] "141"
#median
median(germany$Production)
[1] 763
#quantiles
quantile(germany$Production, probs=c(0.25,
0.5, 0.75))
   25%    50%    75% 
 359.5  763.0 1336.5 
quantile(germany$Production, probs=c(0.1, 0
.2, 0.3, 0.4))
  10%   20%   30%   40% 
129.0 289.0 456.4 610.4 
quantile(germany$Production, probs=c(0.5, 0
.6, 0.7, 0.8, 0.9))
   50%    60%    70%    80%    90% 
 763.0  878.8 1154.4 1539.8 2193.6 
quantile(germany$Production, probs=c(0.01,
0.24, 0.31))
    1%    24%    31% 
 32.72 346.96 468.36 
quantile(germany$Production, probs=c(0.56,
0.67, 0.88, 0.99))
    56%     67%     88%     99% 
 842.04 1051.42 2090.28 3216.60 
#range
range(germany$Production)
[1]   19 3344
#interquartile range
IQR(germany$Production)
[1] 977
#variance and standard deviation
var(germany$Production)
[1] 699780.4
sd(germany$Production)
[1] 836.5288
#MAD
mad(germany$Production)
```

```
#Coefficient of Variation
cv(us$Production)
[1] 55.51986
#Coefficient of assymetry
skewness(us$Production)
[1] 0.546767
#kurtosis
kurtosis(us$Production)
[1] -0.5119056
```

```
[1] 658.2744
#Coefficient of Variation
cv(germany$Production)
[1] 85.38736
#Coefficient of assymetry
skewness(germany$Production)
[1] 1.144767
#kurtosis
kurtosis(us$Production)
[1] -0.5119056
```

Statistics and Explanatory Data Analysis, final exam 2020-03-02

PROBLEM 2 ……… /10 PTS

Difference in means between closing prices of FTSE index in two periods are investigated.

1. Decide which test from two-samples tests is the most appropriate. Make your decision based on the results of relevant analyses and tests.
2. Is there enough evidence to support a claim that the mean price from the first period is higher than from the second period?

For all tests assume 5% significance level.

```
$First
   vars  n    mean    sd median trimmed   mad    min    max range  skew kurtosis   se
X1    1 20 2702.97 19.11 2702.8 2704.62 13.12 2659.8 2737.8    78 -0.62     0.17 4.27

$Second
   vars  n    mean    sd median trimmed   mad    min    max range skew kurtosis    se
X1    1 20 2613.03 54.69 2608.9  2611.9 70.05 2532.6 2705.9 173.3 0.16    -1.32 12.23
```

```
        Shapiro-Wilk normality test

data:  db.all[db.all$Period == "First", "FTSE"]
W = 0.93607, p-value = 0.2019
```

```
        Shapiro-Wilk normality test

data:  db.all[db.all$Period == "Second", "FTSE"]
W = 0.9507, p-value = 0.3778
```

```
        F test to compare two variances

data:  db.all$FTSE by db.all$Period
F = 0.12213, num df = 19, denom df = 19,
p-value = 0.00002785
```

```
t.test(FTSE ~ Period,  db.all, conf.int = 0.95, var.equal = FALSE, alternative=("greater"))

        Welch Two Sample t-test

data:  FTSE by Period
t = 6.9435, df = 23.573, p-value = 0.0000001943
alternative hypothesis: true difference in means is greater than 0
```

```
t.test(FTSE ~ Period,  db.all, conf.int = 0.95, var.equal = FALSE, alternative=("less"))

        Welch Two Sample t-test

data:  FTSE by Period
t = 6.9435, df = 23.573, p-value = 1
alternative hypothesis: true difference in means is less than 0
```

```
t.test(FTSE ~ Period,  db.all, conf.int = 0.95, var.equal = T, alternative=("greater"))

        Two Sample t-test

data:  FTSE by Period
t = 6.9435, df = 38, p-value = 0.00000001467
alternative hypothesis: true difference in means is greater than 0
```

```
t.test(FTSE ~ Period,  db.all, conf.int = 0.95, var.equal = T, alternative=("less"))

        Two Sample t-test

data:  FTSE by Period
t = 6.9435, df = 38, p-value = 1
alternative hypothesis: true difference in means is less than 0
```

Statistics and Explanatory Data Analysis, final exam 2020-03-02

```
wilcox.exact(FTSE ~ Period, db.all, conf.int = 0.95
        ,exact=T, alternative=("greater"))

        Exact Wilcoxon rank sum test

data:  FTSE by Period
W = 377, p-value = 0.00000004155
alternative hypothesis: true mu is greater than 0
```

```
wilcox.exact(FTSE ~ Period, db.all, conf.int = 0.95
        ,exact=T, alternative=("less"))

        Exact Wilcoxon rank sum test

data:  FTSE by Period
W = 377, p-value = 1
alternative hypothesis: true mu is less than 0
```

name, surname, index nr:……

Statistics and Explanatory Data Analysis, final exam 2020-03-02

PROBLEM 3 ……… /20 PTS

Difference between salaries for Data Scientists and Lawyers were analysed in 4 polish cities: Gdansk, Poznan, Warsaw and Wroclaw. To assess whether there exists a difference between salaries ANOVA with (model) and without (model2) interactions & Scheirer-Ray-Hare tests were performed:

- model <- lm(Salary ~ City + Occupation + City:Occupation, data = Data),
- model2 <- lm(Salary ~ City + Occupation, data = Data),
- scheirerRayHare(Salary ~ City+Occupation, data = Data).

For all tests assume 5% significance level.

1. Decide which test from aforementioned is the most appropriate. Make your decision based on the results of relevant analyses and tests.
2. Is there enough evidence to support a claim that salaries differ for occupation and city of living? Make your decision based on the results of relevant analyses and tests.
3. Based on pairwise analysis provide an answer for questions:
   a. In which city(-ies) Data Scientists earn significantly more than in the other cities?
   b. In which city(-ies) Lawyers earn significantly more than in the other cities?
   c. In which city(-ies) Lawyers earn significantly more than Data Scientists?

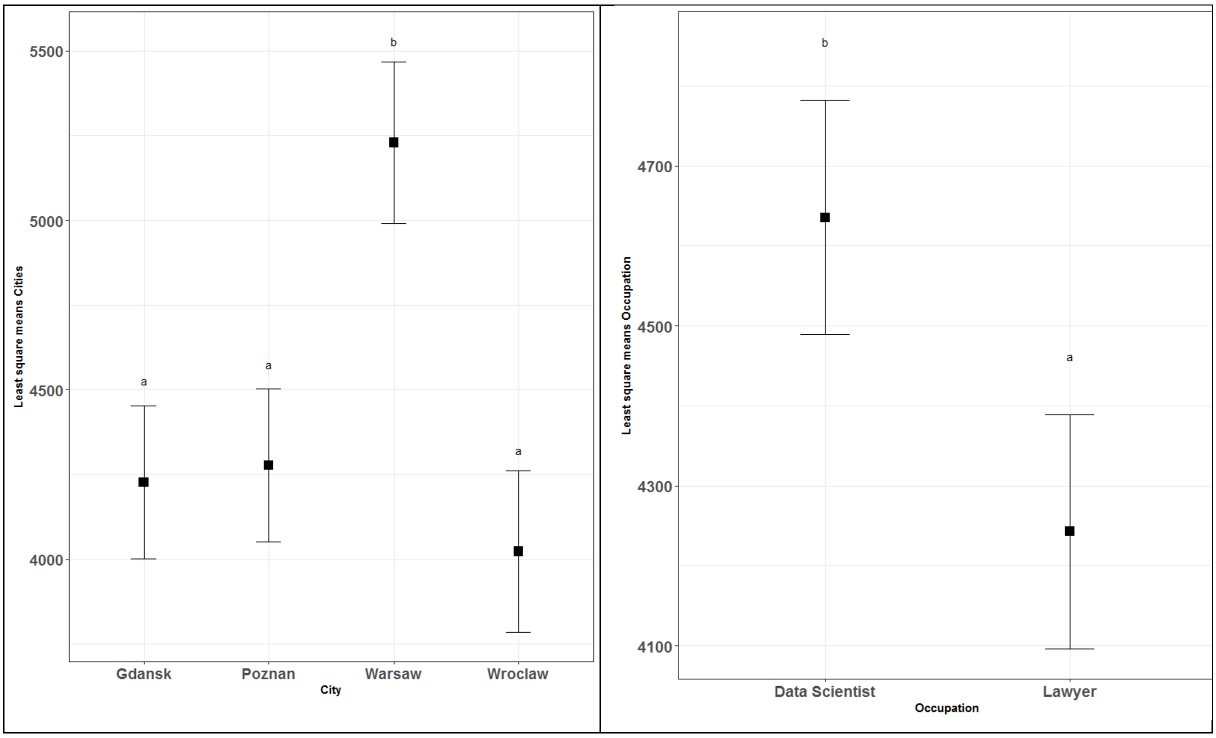
```
> res<- residuals(model)
> plotNormalHistogram(res)
> shapiro.test(res)

Shapiro-Wilk normality test

data:  res
W = 0.98332, p-value = 0.4193

> bartlett.test(Salary ~ interaction(City,Occupation), data=Data)

Bartlett test of homogeneity of variances

data:  Salary by interaction(City, Occupation)
Bartlett's K-squared = 5.0943, df = 7,
p-value = 0.6485
```

```
> res2<- residuals(model2)
> plotNormalHistogram(res2)
> shapiro.test(res2)

Shapiro-Wilk normality test

data:  res2
W = 0.98385, p-value = 0.4467

> leveneTest(Salary ~ interaction(City,Occupation), data = Data)

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  7  0.7284 0.6484
```

```
Anova Table (Type II tests)

Response: Salary
                  Sum Sq Df F value           Pr(>F)    
City            15763264  3  33.705 0.000000000000183 ***
Occupation       3055620  1  19.601 0.000035470960746 ***
City:Occupation  2585358  3   5.528          0.001863 ** 
Residuals       10600815 68                              
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Anova Table (Type III tests)

Response: Salary
                   Sum Sq Df  F value                Pr(>F)    
(Intercept)     212161572  1 1360.932 < 0.00000000000000022 ***
City             12634094  3   27.014      0.00000000001284 ***
Occupation        2869789  1   18.409      0.00005776924582 ***
City:Occupation   2585358  3    5.528              0.001863 ** 

scheirerRayHare(Salary ~ City+Occupation,    data = Data)

DV:  Salary 
Observations:  76 
D:  0.999918 
MS total:  487.6667 

                Df  Sum Sq      H  p.value
City             3 16764.3 34.379 0.000000
Occupation       1  2676.3  5.488 0.019142
City:Occupation  3  2937.7  6.025 0.110423
Residuals       68 14193.7                
```

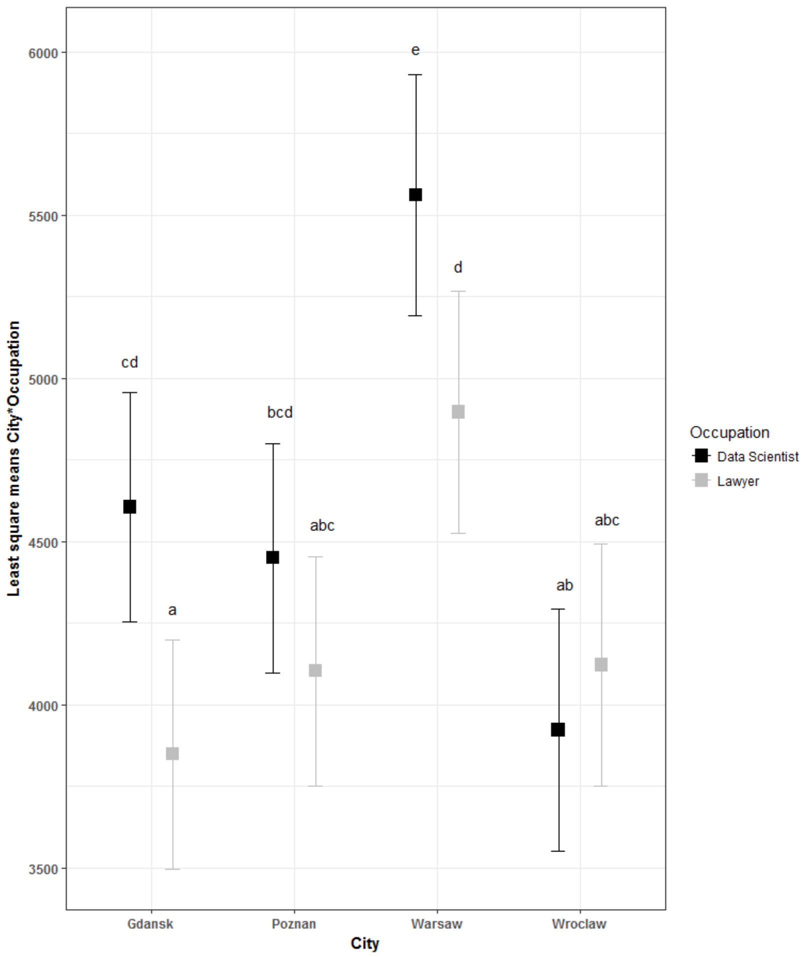
Results for Anova test without interactions

```
> lsCity$contrasts
 contrast          estimate       SE df t.ratio p.value
 Gdansk - Poznan    -49.7500 124.8577 68  -0.398  0.9784
 Gdansk - Warsaw  -1001.7556 128.2790 68  -7.809  <.0001
 Gdansk - Wroclaw   204.2444 128.2790 68   1.592  0.3899
 Poznan - Warsaw   -952.0056 128.2790 68  -7.421  <.0001
 Poznan - Wroclaw   253.9944 128.2790 68   1.980  0.2056
 Warsaw - Wroclaw  1206.0000 131.6115 68   9.163  <.0001

Conf-level adjustment: sidak method for 2 estimates
significance level used: alpha = 0.05
> CLDCity = cld(lsCity, alpha   = 0.05,   Letters = letters,   adjust  = "tukey")
> CLDCity
 City      lsmean       SE df lower.CL upper.CL .group
 Wroclaw 4023.056 93.06340 68 3784.945 4261.166  a
 Gdansk  4227.300 88.28769 68 4001.409 4453.191  a
 Poznan  4277.050 88.28769 68 4051.159 4502.941  a
 Warsaw  5229.056 93.06340 68 4990.945 5467.166   b
> lsOccupation$contrasts
 contrast                 estimate       SE df t.ratio p.value
 Data Scientist - Lawyer 392.6361 90.70698 68   4.329  0.0001

Results are averaged over the levels of: City
> CLDOccupation = cld(lsOccupation,  alpha   = 0.05,    Letters = letters,  adjust  = "tukey")
> CLDOccupation
 Occupation       lsmean       SE df lower.CL upper.CL .group
 Lawyer         4242.797 64.13952 68 4096.117 4389.477  a
 Data Scientist 4635.433 64.13952 68 4488.753 4782.113   b

Conf-level adjustment: sidak method for 2 estimates
significance level used: alpha = 0.05
```



Results for ANOVA test with interactions

```
> leastsquare = lsmeans(model, pairwise ~ City + Occupation, adjust = "tukey")
> CLD = cld(leastsquare,  alpha   = 0.05,  Letters = letters,  adjust  = "tukey")
> ###  Remove spaces in .group
> CLD$.group=gsub(" ", "", CLD$.group)
> CLD
 City    Occupation       lsmean       SE df lower.CL upper.CL .group
 Gdansk  Lawyer         3848.500 124.8577 68 3497.139 4199.861 a
 Wroclaw Data Scientist 3923.556 131.6115 68 3553.188 4293.923 ab
 Poznan  Lawyer         4103.800 124.8577 68 3752.439 4455.161 abc
 Wroclaw Lawyer         4122.556 131.6115 68 3752.188 4492.923 abc
 Poznan  Data Scientist 4450.300 124.8577 68 4098.939 4801.661 bcd
 Gdansk  Data Scientist 4606.100 124.8577 68 4254.739 4957.461 cd
 Warsaw  Lawyer         4896.333 131.6115 68 4525.966 5266.701 d
 Warsaw  Data Scientist 5561.778 131.6115 68 5191.411 5932.145 e

Confidence level used: 0.95
Conf-level adjustment: sidak method for 8 estimates
P value adjustment: tukey method for comparing a family of 8 estimates
significance level used: alpha = 0.05
```



Statistics and Explanatory Data Analysis, final exam 2020-03-02

Results for Scheirer-Ray-Hare test without interactions

```
> DTCity = dunnTest(Salary ~ City, data=Data, method="bh")
> DTCity
        Comparison          Z          P.unadj           P.adj
1  Gdansk - Poznan -0.2720883 0.78555412077122 0.7855541207712
2  Gdansk - Warsaw -4.3639089 0.00001277587892 0.0000383276368
3  Poznan - Warsaw -4.0990776 0.00004147999468 0.0000829599894
4 Gdansk - Wroclaw  1.2889231 0.19742480643401 0.2369097677208
5 Poznan - Wroclaw  1.5537545 0.12024299802476 0.1803644970371
6 Warsaw - Wroclaw  5.5096992 0.00000003594475 0.0000002156685

> DTOccupation = t.test(Salary ~ Occupation, data=Data)
> DTOccupation

         Welch Two Sample t-test
data:  Salary by Occupation
t = 2.7948, df = 70.142, p-value = 0.006693
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 114.8509 687.2017
sample estimates:
mean in group Data Scientist        mean in group Lawyer
                    4629.789                    4228.763
```

Results for Scheirer-Ray-Hare test with interactions

```
> DTAll = dunnTest(Salary ~ interaction(City,Occupation), data=Data, method="bh")
> DTAll

                                       Comparison           Z         P.unadj          P.adj
1            Gdansk.Data Scientist - Gdansk.Lawyer  2.73910416 0.0061606851587 0.01724991844
2    Gdansk.Data Scientist - Poznan.Data Scientist  0.49617783 0.6197689488853 0.66744348341
3           Gdansk.Lawyer - Poznan.Data Scientist -2.24292633 0.0249015656489 0.05363414140
4            Gdansk.Data Scientist - Poznan.Lawyer  1.85813536 0.0631497951897 0.11787961769
5                   Gdansk.Lawyer - Poznan.Lawyer -0.88096881 0.3783347039311 0.48151689591
6            Poznan.Data Scientist - Poznan.Lawyer  1.36195752 0.1732112910633 0.25525874472
7    Gdansk.Data Scientist - Warsaw.Data Scientist -2.39610331 0.0165704192610 0.03866431161
8           Gdansk.Lawyer - Warsaw.Data Scientist -5.06215153 0.0000004145513 0.00001160744
9    Poznan.Data Scientist - Warsaw.Data Scientist -2.87904735 0.0039887844518 0.01240955163
10          Poznan.Lawyer - Warsaw.Data Scientist -4.20467946 0.0000261452427 0.00024402226
11           Gdansk.Data Scientist - Warsaw.Lawyer -1.10934765 0.2672802380755 0.35637365077
12                  Gdansk.Lawyer - Warsaw.Lawyer -3.77539586 0.0001597537542 0.00089462102
13           Poznan.Data Scientist - Warsaw.Lawyer -1.59229169 0.1113191801748 0.18334923793
14                  Poznan.Lawyer - Warsaw.Lawyer -2.91792379 0.0035237045437 0.01409481817
15           Warsaw.Data Scientist - Warsaw.Lawyer  1.25417428 0.2097786662499 0.29369013275
16 Gdansk.Data Scientist - Wroclaw.Data Scientist  2.63593266 0.0083906381091 0.02135798791
17         Gdansk.Lawyer - Wroclaw.Data Scientist -0.03011556 0.9759748928515 0.97597489285
18 Poznan.Data Scientist - Wroclaw.Data Scientist  2.15298862 0.0313195728543 0.06263914571
19         Poznan.Lawyer - Wroclaw.Data Scientist  0.82735651 0.4080350215820 0.49673828714
20 Warsaw.Data Scientist - Wroclaw.Data Scientist  4.90462197 0.0000009360744 0.00001310504
21         Warsaw.Lawyer - Wroclaw.Data Scientist  3.65044769 0.0002617836206 0.00122165690
22          Gdansk.Data Scientist - Wroclaw.Lawyer  1.85292815 0.0638926574483 0.11181215053
23                 Gdansk.Lawyer - Wroclaw.Lawyer -0.81312007 0.4161492225070 0.48550742626
24          Poznan.Data Scientist - Wroclaw.Lawyer  1.36998411 0.1706918612213 0.26552067301
25                 Poznan.Lawyer - Wroclaw.Lawyer  0.04435200 0.9646238195230 1.00000000000
26          Warsaw.Data Scientist - Wroclaw.Lawyer  4.14144358 0.0000345126774 0.00024158874
27                 Warsaw.Lawyer - Wroclaw.Lawyer  2.88726930 0.0038860146744 0.01360105136
28         Wroclaw.Data Scientist - Wroclaw.Lawyer -0.76317839 0.4453570137629 0.49879985541
```

PROBLEM 4 ……… /10 PTS

You have data on 47 regions, whose structure is presented below. There are 5 variables: Density, which represents the population density, Mean.Educ, which represents population's average years of education, Pop.Read, which measures whether the majority of the population can read or not (dummy variable), Pop.Write, which measures whether the majority of the population can write or not (dummy variable), and GDP, which is the measure of the GDP per capita. You want to test the association between different variables in the dataset.

1. Which measures would you use to evaluate the association between each pair of the five variables?
2. You want to test the significance on the relation between the Density (population density) and Pop.Read (whether the majority of population can read). Based on the output from the tests below decide which test is the most appropriate and interpret its output (assume 5% significance level).

```
'data.frame':   47 obs. of  5 variables:
 1 $ Density  : Factor w/ 3 levels "high","low","medium": 1 1 2 1 3 3 ...
 2 $ Mean.Educ: num  10.7 12 6 10 15 12 8 9 15 8 ...
 3 $ Pop.Read : int  0 1 1 1 1 1 0 0 1 1 ...
 4 $ Pop.Write: int  0 0 1 1 1 0 0 0 0 1 ...
 5 $ GDP      : int  48027 5262 23314 21563 698 16721 5560 44956 4385 ...
```

```
Cochran-Armitage test for trend

data:  Data
Z = -0.080776, dim = 3, p-value = 0.9356
alternative hypothesis: two.sided
```

```
Asymptotic Linear-by-Linear Association Test

data:  Pop.Read by Density (High < Medium < Small)
Z = -0.079912, p-value = 0.9363
alternative hypothesis: two.sided
```

```
Call:corr.test(x = Data.num, method = "pearson")
Correlation matrix
            Density.num  Read
Density.num        1.00 -0.21
Read              -0.21  1.00
Sample Size
[1] 47
Probability values (Entries above the diagonal are adjusted for multiple te
sts.)
            Density.num Read
Density.num        0.00 0.94
Read               0.94 0.00
```

```
Call:corr.test(x = Data.num, method = "spearman")
Correlation matrix
            Density.num  Read
Density.num        1.00 -0.23
Read              -0.23  1.00
Sample Size
[1] 47
Probability values (Entries above the diagonal are adjusted for multiple te
sts.)
            Density.num Read
Density.num        0.00 0.93
Read               0.93 0.00
```

```
Call:corr.test(x = Data.num, method = "kendall")
Correlation matrix
            Density.num  Read
Density.num        1.00 -0.21
Read              -0.21  1.00
Sample Size
[1] 47
Probability values (Entries above the diagonal are adjusted for multiple te
sts.)
            Density.num Read
Density.num        0.00 0.93
Read               0.93 0.00
```