TOTAL ……… /80 PT

EXAM RULES

a) BEFORE starting to solve the problems it is 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 a positive total grade one needs to collect **at least 50%** of points available to collect.

Warsaw, 2020-03-05,

…………………………………..
SIGNATURE

Statistics and Explanatory Data Analysis, final exam 2020-03-05

PROBLEM 1 ……… /10 PTS

You have data on hourly wages and education level. Hourly wages are measured in dollars and education level takes the following values: 1) less than high school, 2) high school, 3) some college; 4) college. The descriptive statistics are given below.

1. For each variable, which location measures would you use to summarize your data? Explain your choice and interpret the values.
2. Based on the output what can you say about the shape of the distribution of the wage data?
3. For each variable, what graphs would you use to represent your data graphically? Explain your choice.
4. What is the percentage of the total sample that earns between \$10 and \$15?

Wages

```
#means
mean(Data$wage)
18.01461
mean(Data$wage, trim=0.1)
14.41197
mean(Data$wage, trim=0.2)
13.89368
winsor.mean(Data$wage, trim=0.1)
14.9712
winsor.mean(Data$wage, trim=0.2)
14.33591
#midrange
(min(Data$wage)+max(Data$wage))/2
1500
#trimean
TMH(Data$wage)
8.21
#mode
names(sort(-table(Data$wage)))[1]
"10"
#median
median(Data$wage)
13.5
#quantiles
quantile(Data$wage, probs=c(0.25, 0.5, 0.75))
    25%     50%     75%
 9.6700 13.5000 19.1925
quantile(Data$wage, probs=c(0.1, 0.2, 0.3, 0.4))
 10%  20%  30%  40%
 7.5  9.0 10.0 12.0
quantile(Data$wage,probs=c(0.5, 0.6, 0.7, 0.8, 0.9))
  50%   60%   70%   80%   90%
13.50 15.10 17.67 21.00 26.92
range(Data$wage)
0 3000
#interquartile range
IQR(Data$wage)
9.522501
#variance and standard deviation
var(Data$wage)
2971.773
sd(Data$wage)
```

```
54.51397
#MAD
mad(Data$wage)
6.6717
#Coefficient of Vartiation
cv(Data$wage)
302.6097
```

Education

```
#means
mean(Data$education)
2.704997
mean(Data$education, trim=0.1)
2.756203
mean(Data$education, trim=0.2)
2.815601
#means
mean(Data$education)
2.704997
mean(Data$education, trim=0.1)
2.756203
mean(Data$education, trim=0.2)
2.815601
winsor.mean(Data$education, trim=0.1)
2.704997
winsor.mean(Data$education, trim=0.2)
2.889312
#midrange
(min(Data$education)+max(Data$education))/2
2.5
#trimean
TMH(Data$education)
2
#mode
names(sort(-table(Data$education)))[1]
"4"
#median
median(Data$education)
3
#quantiles
quantile(Data$education, probs=c(0.25, 0.5, 0.75))
25% 50% 75%
  2   3   4
quantile(Data$education, probs=c(0.1, 0.2, 0.3, 0.4))
10% 20% 30% 40%
  1   2   2   2
quantile(Data$education, probs=c(0.5, 0.6, 0.7, 0.8, 0.9))
50% 60% 70% 80% 90%
  3   3   4   4   4
#range
range(Data$education)
1 4
#interquartile range
IQR(Data$education)
2
#variance and standard deviation
var(Data$education)
1.286283
```

```
sd(Data$education)
1.134144
#MAD
mad(Data$education)
1.4826
#Coefficient of Vartiation
cv(Data$education)
41.92775
```

PROBLEM 2 ……… /10 PTS

Difference in average income for woman and man in 2013 were 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 the claim that income in 2013 was significantly higher for man than for woman?

For all tests assume 5% significance level.

```
Tests results:
```

```
Jarque-Bera test for normality

data: Data[Data$Gender == "Male", "Wrkday"]

JB = 0.084323, p-value = 0.9605
```

```
Jarque-Bera test for normality

data: Data[Data$Gender == "Female", "Wrkday"]

JB = 0.66979, p-value = 0.621
```

```
F test to compare two variances

data:  Data$Wrkday by Data$Gender
F = 0.54036, num df = 19, denom df = 19, p-value = 0.1889

alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:  0.2138808 1.3651913
sample estimates: ratio of variances 0.5403593
```

```
t.test(Wrkday ~ Gender, data = Data, conf.i
nt = 0.95, var.equal = FALSE, alternative =
c("greater"))

         Welch Two Sample t-test

data:  Wrkday by Gender
t = 0.76722, df = 34.893, p-value = 0.2241
alternative hypothesis: true difference in
means is greater than 0
95 percent confidence interval:
 -49.59793       Inf
sample estimates:
  mean in group Male mean in group Female
              1287.50              1246.25
```

```
t.test(Wrkday ~ Gender, data = Data, conf.int
 = 0.95, var.equal = TRUE, alternative = c("gr
                   eater"))

Two Sample t-test

data:  Wrkday by Gender
t = 0.76722, df = 38, p-value = 0.2238
alternative hypothesis: true difference in me
ans is greater than 0
95 percent confidence interval:
 -49.39584       Inf
sample estimates:
  mean in group Male mean in group Female
              1287.50              1246.25
```

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

        Exact Wilcoxon rank sum test

data:  Wrkday by Gender
W = 150, p-value = 0.2452
alternative hypothesis: true mu is greater
than 0
```

name, surname, index nr:………

Statistics and Explanatory Data Analysis, final exam 2020-03-05

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 exist 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 depends on occupation type differently in different 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) earnings are significantly higher 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 = 0.98417, p-value = 0.4552
> bartlett.test(Salary ~ interaction(City,Oc
cupation), data=Data)

Bartlett test of homogeneity of variances

data:  Salary by interaction(City, Occupatio
n)
Bartlett's K-squared = 2.3936, df = 7,
p-value = 0.9349
```

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

         Shapiro-Wilk normality test

data:  res2
W = 0.98338, p-value = 0.413
> leveneTest(Salary ~ interaction(City,Occup
ation), data = Data)

Levene's Test for Homogeneity of Variance (center =
median)
      Df F value Pr(>F)
group  7  0.3434  0.931
      69
```

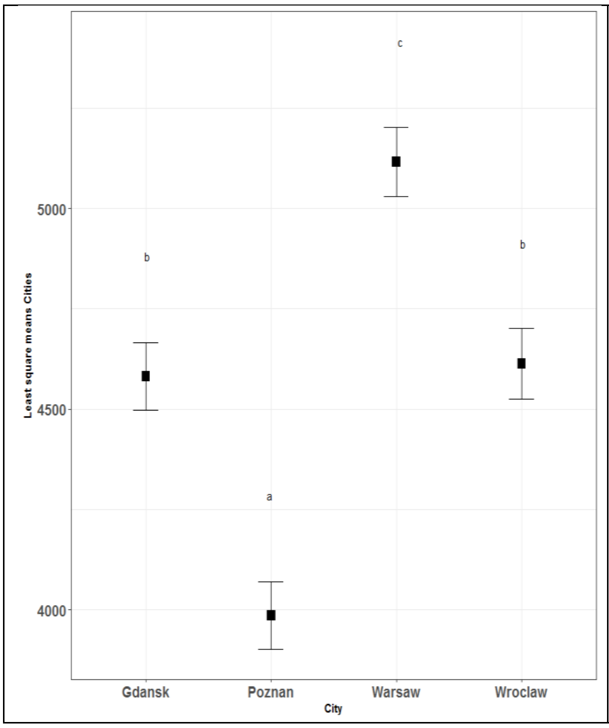
```
Anova Table (Type II tests)

Response: Salary
             Sum Sq Df F value                Pr(>F)
City       12530686  3  191.96 < 0.00000000000000022 ***
Occupation  6529496  1  300.08 < 0.00000000000000022 ***
Residuals   1566644 72
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

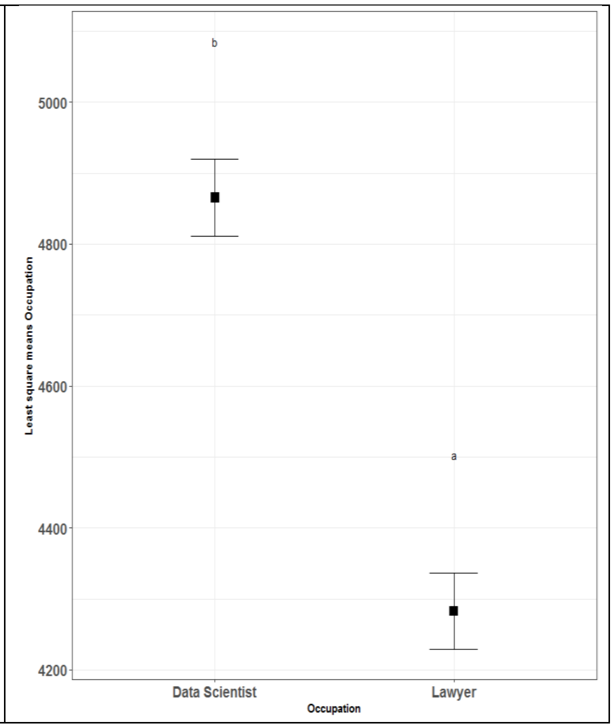
```
Anova Table (Type III tests)

Response: Salary
                   Sum Sq Df    F value                Pr(>F)
(Intercept)     237617252  1 10476.2801 < 0.00000000000000022 ***
City              5977484  3    87.8469 < 0.00000000000000022 ***
Occupation        1718738  1    75.7773     0.000000000001029 ***
City:Occupation      1624  3     0.0239                 0.995
Residuals         1565020 69
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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

DV:  Salary
Observations:  77
D:  0.9999869
MS total:  500.5
                Df  Sum Sq      H p.value
City             3 21837.0 43.631 0.00000
Occupation       1 11626.4 23.230 0.00000
City:Occupation  3   483.1  0.965 0.80965
Residuals       69  4091.1
```

Results for Anova test without interactions

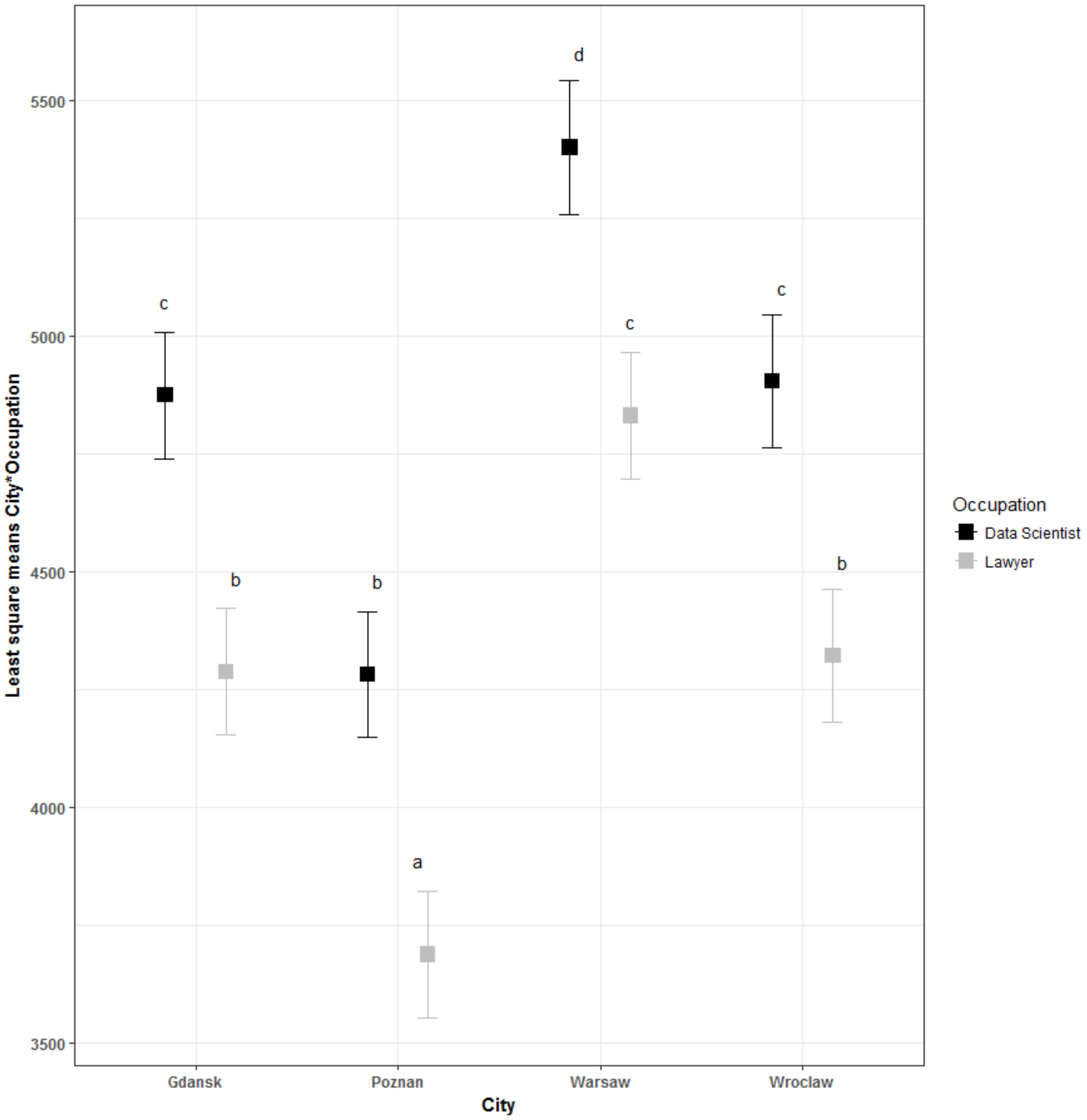
```
> lsCity$contrasts
 contrast             estimate       SE df t.ratio p.value
 Gdansk - Poznan     595.85000 46.64648 72  12.774  <.0001
 Gdansk - Warsaw    -534.77640 47.26455 72 -11.315  <.0001
 Gdansk - Wroclaw    -31.88333 47.92470 72  -0.665  0.9098
 Poznan - Warsaw   -1130.62640 47.26455 72 -23.921  <.0001
 Poznan - Wroclaw   -627.73333 47.92470 72 -13.098  <.0001
 Warsaw - Wroclaw    502.89307 48.52650 72  10.363  <.0001

> CLDCity = cld(lsCity, alpha   = 0.05,   Letters = letters,   adjust  = "tukey")
> CLDCity
 City      lsmean       SE df lower.CL upper.CL .group
 Poznan  3985.600 32.98404 72 3901.335 4069.865  a
 Gdansk  4581.450 32.98404 72 4497.185 4665.715   b
 Wroclaw 4613.333 34.76824 72 4524.510 4702.157   b
 Warsaw  5116.226 33.85249 72 5029.742 5202.710    c

> lsOccupation <- lsmeans(model2, pairwise ~ Occupation, adjust = "tukey")
> lsOccupation$contrasts
 contrast                 estimate       SE df t.ratio p.value
 Data Scientist - Lawyer 582.6033 33.63195 72  17.323  <.0001

> CLDOccupation = cld(lsOccupation,  alpha   = 0.05,    Letters = letters,  adjust  = "tukey")
> CLDOccupation
 Occupation       lsmean       SE df lower.CL upper.CL .group
 Lawyer         4282.851 23.63271 72 4228.873 4336.829  a
 Data Scientist 4865.454 23.94553 72 4810.762 4920.146   b
 significance level used: alpha = 0.05
```





Results for ANOVA test with interactions

```
> CLD
 City    Occupation       lsmean       SE df lower.CL upper.CL .group
 Poznan  Lawyer         3688.800 47.62505 69 3554.840 3822.760 a
 Poznan  Data Scientist 4282.400 47.62505 69 4148.440 4416.360 b
 Gdansk  Lawyer         4288.300 47.62505 69 4154.340 4422.260 b
 Wroclaw Lawyer         4322.778 50.20121 69 4181.572 4463.984 b
 Warsaw  Lawyer         4831.600 47.62505 69 4697.640 4965.560 c
 Gdansk  Data Scientist 4874.600 47.62505 69 4740.640 5008.560 c
 Wroclaw Data Scientist 4903.889 50.20121 69 4762.683 5045.095 c
 Warsaw  Data Scientist 5400.111 50.20121 69 5258.905 5541.317 d
```



Results for Scheirer-Ray-Hare test without interactions

```
       Comparison          Z            P.unadj             P.adj
1  Gdansk - Poznan  3.5443921 0.00039351959724042 0.0007870391944808
2  Gdansk - Warsaw -3.0824597 0.00205297503076595 0.0030794625461489
3  Poznan - Warsaw -6.5811158 0.00000000004669308 0.0000000002801585
4 Gdansk - Wroclaw -0.1834421 0.85445114900160035 0.8544511490016004
5 Poznan - Wroclaw -3.6333000 0.00027981935886571 0.0008394580765971
6 Warsaw - Wroclaw  2.8210752 0.00478629826515237 0.0057435579181828
>
> DTOccupation = t.test(Salary ~ Occupation, data=Data)
> DTOccupation

	Welch Two Sample t-test

data:  Salary by Occupation
t = 5.7518, df = 74.988, p-value = 0.0000001807
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 371.4813 765.1421
sample estimates:
mean in group Data Scientist        mean in group Lawyer
                    4850.158                    4281.846
```

Results for Scheirer-Ray-Hare test with interactions

```
> DTAll = dunnTest(Salary ~ interaction(City,Occupation), data=Data, method="bh")
> DTAll
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Benjamini-Hochberg method.

                             Comparison          Z         P.unadj          P.adj
1           Gdansk.Data Scientist - Gdansk.Lawyer  2.9485455 0.0031927315344778 0.008126952996853
2   Gdansk.Data Scientist - Poznan.Data Scientist  3.0634888 0.0021877240653391 0.006806252647722
3           Gdansk.Lawyer - Poznan.Data Scientist  0.1149433 0.9084900652381284 0.908490065238128
4           Gdansk.Data Scientist - Poznan.Lawyer  4.8975840 0.0000009702215116 0.000009055400775
5                   Gdansk.Lawyer - Poznan.Lawyer  1.9490385 0.0512908223106427 0.075586474984105
6           Poznan.Data Scientist - Poznan.Lawyer  1.8340952 0.0666398575979077 0.093295800637071
7   Gdansk.Data Scientist - Warsaw.Data Scientist -1.7997699 0.0718969708494953 0.095862627799327
8           Gdansk.Lawyer - Warsaw.Data Scientist -4.6696734 0.0000030167899369 0.000016894023646
9   Poznan.Data Scientist - Warsaw.Data Scientist -4.7815510 0.0000017394788236 0.000012176351765
10          Poznan.Lawyer - Warsaw.Data Scientist -6.5667282 0.0000000000514328 0.000000001440119
11          Gdansk.Data Scientist - Warsaw.Lawyer  0.2698669 0.7872626720025946 0.918473117336360
12                  Gdansk.Lawyer - Warsaw.Lawyer -2.6786786 0.0073913300257135 0.013797149381332
13          Poznan.Data Scientist - Warsaw.Lawyer -2.7936219 0.0052121379172367 0.011226143206356
14                  Poznan.Lawyer - Warsaw.Lawyer -4.6277171 0.0000036971848957 0.000017253529513
15          Warsaw.Data Scientist - Warsaw.Lawyer  2.0624391 0.0391659521851358 0.064508627128459
16 Gdansk.Data Scientist - Wroclaw.Data Scientist -0.1351179 0.8925186862309953 0.961173969787226
17         Gdansk.Lawyer - Wroclaw.Data Scientist -3.0050213 0.0026556225477259 0.007435743133632
18 Poznan.Data Scientist - Wroclaw.Data Scientist -3.1168989 0.0018276424458317 0.006396748560411
19         Poznan.Lawyer - Wroclaw.Data Scientist -4.9020761 0.0000009482908067 0.000013276071293
20 Warsaw.Data Scientist - Wroclaw.Data Scientist  1.6225022 0.1046958725101554 0.133249292285652
21         Warsaw.Lawyer - Wroclaw.Data Scientist -0.3977870 0.6907872067952079 0.840958338707210
22         Gdansk.Data Scientist - Wroclaw.Lawyer  2.7455950 0.0060401277477915 0.012080255495583
23                 Gdansk.Lawyer - Wroclaw.Lawyer -0.1243084 0.9010710689993708 0.934444071554903
24         Poznan.Data Scientist - Wroclaw.Lawyer -0.2361860 0.8132883221710988 0.910882920831631
25                 Poznan.Lawyer - Wroclaw.Lawyer -2.0213632 0.0432421764518804 0.067265607814036
26         Warsaw.Data Scientist - Wroclaw.Lawyer  4.4302737 0.0000094113547665 0.000037645419066
27                 Warsaw.Lawyer - Wroclaw.Lawyer  2.4829259 0.0130308202010697 0.022803935351872
28        Wroclaw.Data Scientist - Wroclaw.Lawyer  2.8077716 0.0049885596533003 0.011639972524367
```

PROBLEM 4 ……… /10 PTS

You have data on employment status of 1,000 individuals in two points in time: before and after the crisis of 2008. The employment status is defined as: 1) being employed; 2) being unemployed. You want to test the hypothesis that crisis is associated with worse labor market situation, i.e. <u>that the number of individuals who became unemployed after the crisis is greater than the number of individuals who found a job after the crisis.</u>

1. Decide which test from the tests below is the most appropriate for testing the hypothesis.
2. Is there enough evidence to support a claim that crisis is associated with worse labor market situation?

For all tests assume 5% significance level.

The distribution of the data

```
Matrix

                  After.Employed After.Unemploed
Before.Employed              400             300
Before.Unemployed            100             200


prop.table(Matrix)


                  After.Employed After.Unemploed
Before.Employed              0.4             0.3
Before.Unemployed            0.1             0.2


Before=c(0.7, 0.3)
After=c(0.5, 0.5)
```

Tests

```
chisq.test(x = before, p = after)

Chi-squared test for given probabilities

data:  before
X-squared = 1429.7, df = 1, p-value < 0.00000000000000022


GTest(x=before,  p=after, correct="none")

Log likelihood ratio (G-test) goodness of fit test

data:  before
G = 1540, X-squared df = 1, p-value < 0.00000000000000022


mcnemar.test(Matrix)
```

```
McNemar's Chi-squared test
data:  Matrix
McNemar's chi-squared = 99.002, df = 1, p-value < 0.00000000000000022

StuartMaxwellTest(Matrix)

Stuart-Maxwell test

data:  Matrix
chi-squared = 100, df = 1, p-value < 0.00000000000000022

fisher.test(Matrix)
Fisher's Exact Test for Count Data

data:  Matrix
p-value = 0.000000000006015
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.991381 3.578600
sample estimates:
odds ratio
  2.664003

chisq.test(Matrix)

Pearson's Chi-squared test with Yates' continuity correction

data:  Matrix
X-squared = 46.671, df = 1, p-value = 0.000000000008394
```

PROBLEM 1 ……… /10 PTS

PROBLEM 2 ……… /10 PTS

PROBLEM 3 ……… /20 PTS

PROBLEM 4 ……… /10 PTS