Statistics and Explanatory Data Analysis, final exam 2024-05-02

TOTAL ……… /50 PT

EXAM RULES

a) All solutions have to be solved at paper sheets using **handwriting**.
b) One has to solve **all problems**.
c) Exam lasts **90 minutes**.
d) The exam is an **open book exam**.
e) To obtain a positive total grade one needs to collect **at least 50%** of points available to collect.
f) 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.

Ethical Statement.

I hereby declare that I will comply with the examination rules established by the examiner and resulting from rules of studying (see above). I declare that during the exam I will not use any unauthorized examination aids and that I will not communicate with other people. I am aware that non-compliance with these rules is an expression of dishonesty towards all participants of the examination process and the whole academic community and at the same time may result in disciplinary penalty.
In case of doubts please refer to the Rules of Study at the University of Warsaw and statements of the Head of the Educational Unit (EUH).

Warsaw, 2024-05-02,

……………………………….
SIGNATURE

Statistics and Explanatory Data Analysis, final exam 2024-05-02

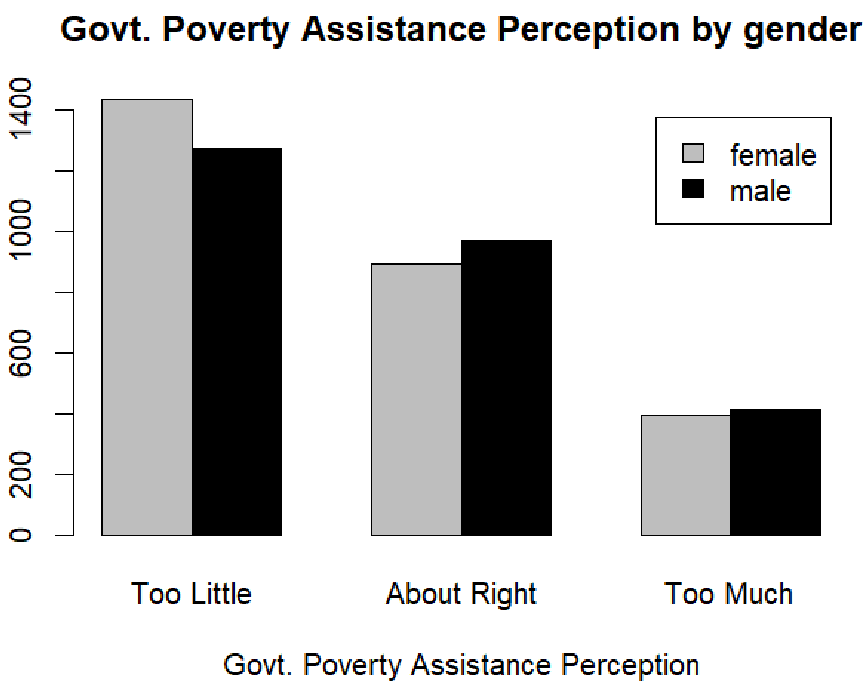
Problem 1…/20 PTS

You have the data from World Value Survey. Following are the variables along with their labels.

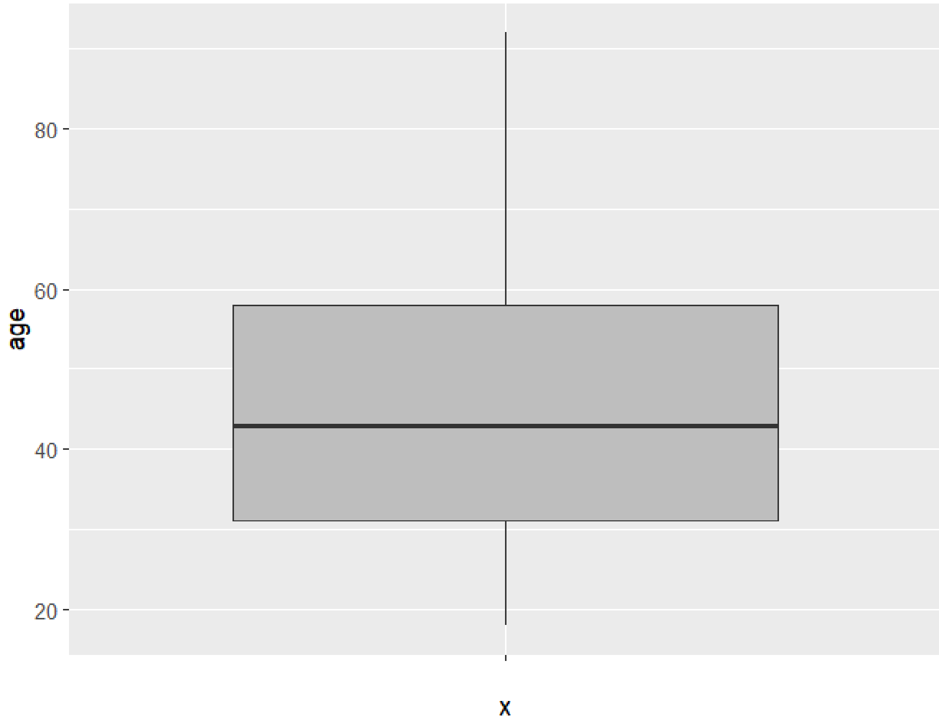
- **Poverty**: "Do you think that what the government is doing for people in poverty in this country is about the right amount, too much, or little?" (ordered): Too Little, About Right, Too Much.
- **Religion**: Member of a religion: no or yes.
- **Degree**: Held a university degree: no or yes.
- **Country**: Australia, Norway, Sweden, or USA.
- **Age**: in years.
- **Gender**: male or female

```
> str(WVS)
'data.frame':	5381 obs. of  6 variables:
 $ poverty : Ord.factor w/ 3 levels "Too Little"<"About Right"<..: 1 2 1 3 1 2 3 1 1 1 ...
 $ religion: Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ degree  : Factor w/ 2 levels "no","yes": 1 1 1 2 2 1 1 1 1 1 ...
 $ country : Factor w/ 4 levels "Australia","Norway",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ age     : int  44 40 36 25 39 80 48 32 74 30 ...
 $ gender  : Factor w/ 2 levels "female","male": 2 1 1 1 2 1 1 2 1 2 ...
```

1. For each variable, what graphs would you use to represent your data graphically? Explain your choice. **(3 PTS)**.
2. Based on the bar chart below, what is the 1) overall pattern you observe about governments' poverty assistance perception 2) and with respect to gender of the respondent? **(3 PTS)**.



3. By visualizing the box plot of variable „age", what do you infer about:
   a) the proportion of the data falls within the interquartile range (IQR)? **(1 PT)**.
   b) whether the box appear symmetrical, or is there asymmetry? **(1 PT)**.
   c) the range of the dataset, as indicated by the whiskers of the boxplot? **(1 PT)**.
   d) why the interquartile range may be a better measure of spread than the range. (**1 PT)**.



4. Difference in means between DAX index in two periods 'First' and 'Second' are investigated.
   a) Decide which test from two-samples tests is the most appropriate for checking whether prices from the first period are equal to prices in the second period. **(5 PTS)**.
   b) Is there enough evidence to support a claim that prices in both periods are not significantly different? **(5 PTS)**.

For all tests assume 5% significance level.

```
describe(db.all[db.all$Period=='First',"DAX"])
vars n    mean   sd median trimmed   mad     min    max range  skew kurtosis  se
X1    1 20 1737.59 9.83 1738.1 1738.26 11.68 1714.77 1753.1 38.33 -0.44    -0.71 2.2

describe(db.all[db.all$Period=='Second',"DAX"])
vars n    mean    sd  median trimmed   mad     min     max range skew kurtosis   se
X1    1 20 1765.94 28.73 1758.08 1765.56 36.87 1719.92 1812.33 92.41 0.16    -1.49 6.42
```

```
        Shapiro-Wilk normality test

data:  db.all[db.all$Period == "First", "DAX"]
W = 0.94491, p-value = 0.2964
```

```
        Shapiro-Wilk normality test

data:  db.all[db.all$Period == "Second", "DAX"]
W = 0.92868, p-value = 0.1456
```

```
        F test to compare two variances

data:  db.all$DAX by db.all$Period
F = 0.11695, num df = 19, denom df = 19, p-value = 1.996e-05
```

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

        Welch Two Sample t-test

data:  DAX by Period
t = -4.1753, df = 23.384, p-value = 0.9998

>   t.test(DAX ~ Period,  db.all, conf.int = 0.95,
+   var.equal = FALSE, alternative=("two.sided"))

        Welch Two Sample t-test

data:  DAX by Period
t = -4.1753, df = 23.384, p-value = 0.0003535

>  t.test(DAX ~ Period,  db.all, conf.int = 0.95,
+   var.equal = TRUE, alternative=("greater"))

        Two Sample t-test

data:  DAX by Period
t = -4.1753, df = 38, p-value = 0.9999

> t.test(DAX ~ Period,  db.all, conf.int = 0.95,
+   var.equal = TRUE, alternative=("two.sided"))

        Two Sample t-test

data:  DAX by Period
t = -4.1753, df = 38, p-value = 0.0001673

> wilcox.exact(DAX ~ Period, db.all, conf.int = 0.95,
+       exact=TRUE, alternative=("greater"))

        Exact Wilcoxon rank sum test

data:  DAX by Period
W = 70, p-value = 0.9999

>   wilcox.exact(DAX ~ Period, db.all, conf.int = 0.95,
+      exact=TRUE, alternative=("two.sided"))

        Exact Wilcoxon rank sum test

data:  DAX by Period
W = 70, p-value = 0.000251
```

PROBLEM 2 ……… /20 PTS

Data Scientist analysed efficiency in classification problem of 3 algorithms (LightGBM, XGBoost and Neural Network) for 3 datasets. She considered 3 different datasets with 3 different targets – Probability of Default (CreditScoring), Propensity to Buy a life insurance (PropesityToBuy) and having Covid infection by a patient (Covid). To assess whether there exists a difference in discrimination power between aforementioned algorithms and datasets AUC scores for bootstrapped samples (interpedently bootstrapped for every Model) using ANOVA with (model) and without (model2) interactions & Scheirer-Ray-Hare tests were performed:

- model <- lm(AUC ~ Model + Target + Model:Target, data = Data)
- model2 <- lm(AUC ~ Model + Target, data = Data),
- model3<-scheirerRayHare(AUC ~ Model+Target,   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.
    a. Choose and interpret results of appropriate diagnostic tests. **(4 PTS)**
    b. Choose and interpret results of appropriate ANOVA/Scheirer-Ray-Hare test. **(3 PTS)**
    c. Explain how AUC Score depends on Models and Targets types. Are effects of Model and Target independent? **(3 PTS)**
2. Based on pairwise analysis provide an answer for questions:
    a. Is there a dataset that has statistically the highest average of AUC for each of the model? Explain your decision based on appropriate statistical test results (particular p-value or common letter approach). **(3 PTS)**
    b. Which Model(-s) is (are) the best for CreditScoring, which for the PropesityToBuy and which for Covid datasets? Explain your decision based on appropriate statistical test results (particular p-value or common letter approach). **(3 PTS)**
    c. Is there a Model that could be recommended as the best? Explain your decision based on appropriate statistical test results (particular p-value or common letter approach). **(2 PTS)**
    d. Is there a Model which may be removed from the consideration, as for all Targets there is a statically better model (in terms of AUC Score)? Explain your decision based on appropriate statistical test results (particular p-value or common letter approach). **(2 PTS)**

Statistics and Explanatory Data Analysis, final exam 2024-05-02

```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.99215, p-value = 0.3979
```

```
> shapiro.test(res2)

        Shapiro-Wilk normality test

data:  res2
W = 0.98868, p-value = 0.1356
```

```
> bartlett.test(AUC ~ interaction(M
odel,Target), data=Data)

        Bartlett test of homogeneity
of variances

data:  AUC by interaction(Model, Ta
rget)
Bartlett's K-squared = 9.8748, df =
8, p-value = 0.2739
```

```
> leveneTest(AUC ~ interaction(Mode
l,Target), data = Data)
Levene's Test for Homogeneity of Va
riance (center = median)
       Df F value Pr(>F)
group   8   1.257 0.2688
      181
```

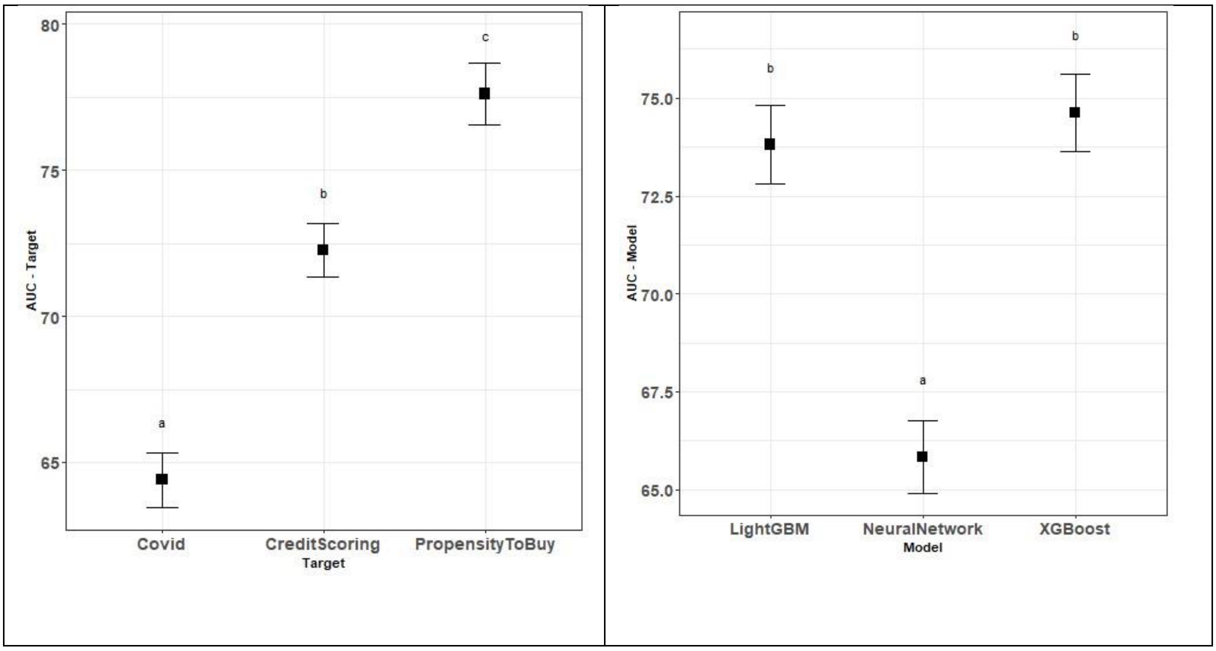
```
> Anova(model,type = "II")
Anova Table (Type II tests)

Response: AUC
             Sum Sq  Df  F value                Pr(>F)    
Model        2840.6   2 143.1224 < 0.00000000000000022 ***
Target       5134.6   2 258.7033 < 0.00000000000000022 ***
Model:Target  301.6   4   7.5982            0.00001111 ***
Residuals    1796.2 181                                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> Anova(model,type = "III")
Anova Table (Type III tests)

Response: AUC
            Sum Sq  Df   F value                Pr(>F)    
(Intercept)  85747   1 8640.5392 < 0.00000000000000022 ***
Model          756   2   38.0902   0.00000000000001561 ***
Target        1350   2   68.0379 < 0.00000000000000022 ***
Model:Target   302   4    7.5982   0.00001111114676908 ***
Residuals     1796 181                                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> scheirerRayHare(AUC ~ Model+Target,    data = Data)


              Df Sum Sq      H p.value
Model          2 154280 51.016 0.00000
Target         2 297522 98.382 0.00000
Model:Target   4  11118  3.677 0.45155
Residuals    181 106371
```

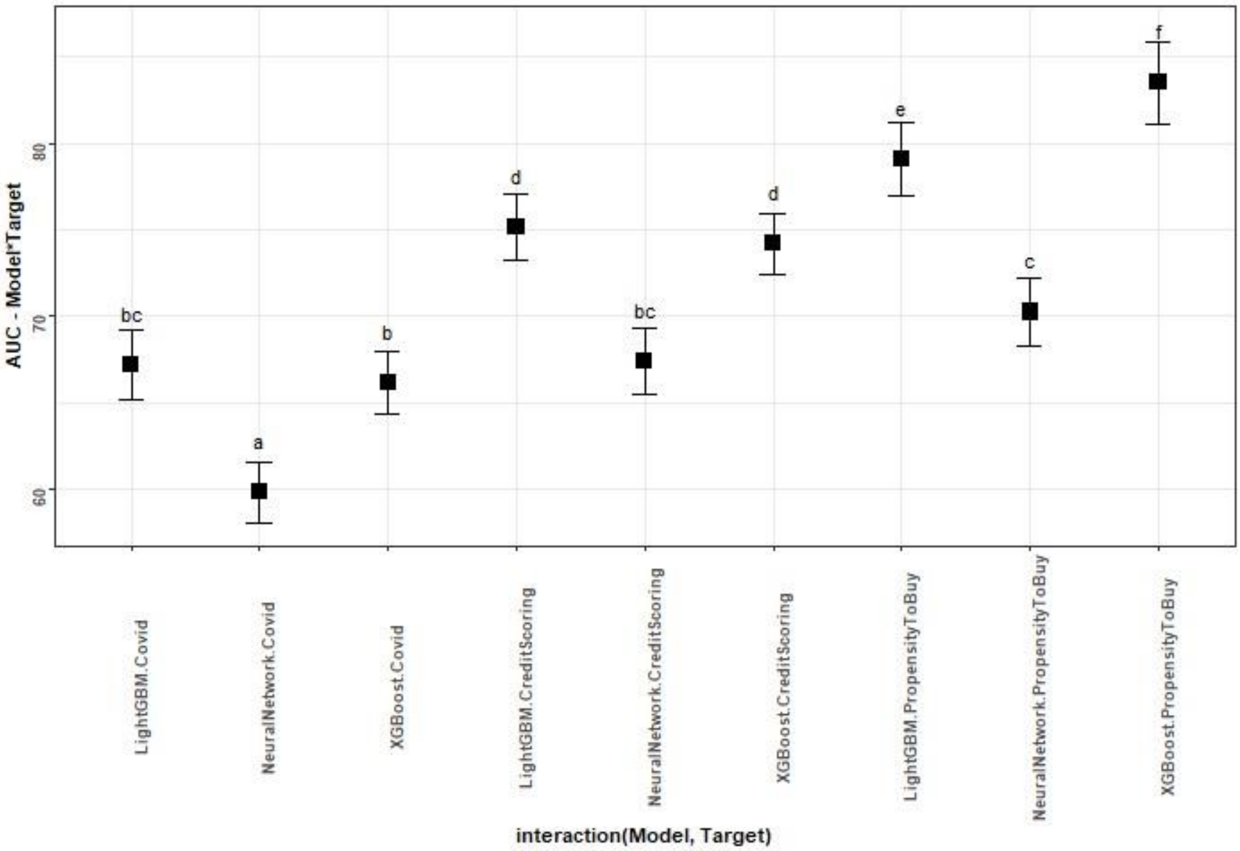
Results for Anova test without interactions

```
> lsModel <- lsmeans::lsmeans(model, pairwise ~ Model, adjust = "tuk
ey")
> lsModel$contrasts

 contrast                    estimate    SE  df t.ratio p.value
 LightGBM - NeuralNetwork        7.97 0.566 181  14.080  <.0001
 LightGBM - XGBoost             -0.81 0.583 181  -1.389  0.3486
 NeuralNetwork - XGBoost        -8.78 0.561 181 -15.657  <.0001


> CLDModel = cld(lsModel[[1]], alpha   = 0.05,   Letters = letters,
adjust  = "tukey")
> CLDModel

 Model          lsmean    SE  df lower.CL upper.CL .group
 NeuralNetwork    65.8 0.384 181     64.9     66.8  a
 LightGBM         73.8 0.416 181     72.8     74.8   b
 XGBoost          74.6 0.409 181     73.6     75.6   b


> lsTarget <- lsmeans(model, pairwise ~ Target, adjust = "tukey")
> lsTarget$contrasts

 contrast                           estimate    SE  df t.ratio p.value
 Covid - CreditScoring                 -7.87 0.540 181 -14.580  <.0001
 Covid - PropensityToBuy              -13.23 0.585 181 -22.598  <.0001
 CreditScoring - PropensityToBuy       -5.35 0.584 181  -9.167  <.0001

> CLDTarget = cld(lsTarget[[1]],  alpha   = 0.05,    Letters = lette
rs,  adjust  = "tukey")
> CLDTarget
 Target            lsmean    SE  df lower.CL upper.CL .group
 Covid               64.4 0.383 181     63.5     65.3  a
 CreditScoring       72.3 0.381 181     71.3     73.2   b
 PropensityToBuy     77.6 0.443 181     76.6     78.7    c
```



Results for ANOVA test with interactions

```
> leastsquare = lsmeans(model, pairwise ~ Model + Target, adjust = "tukey")
> leastsquare$contrasts
 contrast                                                    estimate    SE  df t.ratio p.value
 LightGBM Covid - NeuralNetwork Covid                           7.345 0.951 181   7.725  <.0001
 LightGBM Covid - XGBoost Covid                                 1.008 0.967 181   1.042  0.9811
 LightGBM Covid - LightGBM CreditScoring                       -7.984 0.987 181  -8.092  <.0001
 LightGBM Covid - NeuralNetwork CreditScoring                  -0.256 0.997 181  -0.257  1.0000
 LightGBM Covid - XGBoost CreditScoring                        -7.019 0.951 181  -7.382  <.0001
 LightGBM Covid - LightGBM PropensityToBuy                    -11.919 1.052 181 -11.333  <.0001
 LightGBM Covid - NeuralNetwork PropensityToBuy                -3.080 0.997 181  -3.088  0.0578
 LightGBM Covid - XGBoost PropensityToBuy                     -16.323 1.110 181 -14.711  <.0001
 NeuralNetwork Covid - XGBoost Covid                           -6.336 0.892 181  -7.106  <.0001
 NeuralNetwork Covid - LightGBM CreditScoring                 -15.329 0.913 181 -16.797  <.0001
 NeuralNetwork Covid - NeuralNetwork CreditScoring             -7.601 0.924 181  -8.223  <.0001
 NeuralNetwork Covid - XGBoost CreditScoring                  -14.363 0.874 181 -16.440  <.0001
 NeuralNetwork Covid - LightGBM PropensityToBuy               -19.264 0.983 181 -19.605  <.0001
 NeuralNetwork Covid - NeuralNetwork PropensityToBuy          -10.425 0.924 181 -11.279  <.0001
 NeuralNetwork Covid - XGBoost PropensityToBuy                -23.668 1.044 181 -22.664  <.0001
 XGBoost Covid - LightGBM CreditScoring                        -8.992 0.930 181  -9.671  <.0001
 XGBoost Covid - NeuralNetwork CreditScoring                   -1.264 0.941 181  -1.343  0.9168
 XGBoost Covid - XGBoost CreditScoring                         -8.027 0.892 181  -9.002  <.0001
 XGBoost Covid - LightGBM PropensityToBuy                     -12.927 0.999 181 -12.945  <.0001
 XGBoost Covid - NeuralNetwork PropensityToBuy                 -4.089 0.941 181  -4.343  0.0008
 XGBoost Covid - XGBoost PropensityToBuy                      -17.332 1.059 181 -16.360  <.0001
 LightGBM CreditScoring - NeuralNetwork CreditScoring           7.728 0.961 181   8.041  <.0001
 LightGBM CreditScoring - XGBoost CreditScoring                 0.965 0.913 181   1.058  0.9793
 LightGBM CreditScoring - LightGBM PropensityToBuy             -3.935 1.017 181  -3.868  0.0047
 LightGBM CreditScoring - NeuralNetwork PropensityToBuy         4.904 0.961 181   5.103  <.0001
 LightGBM CreditScoring - XGBoost PropensityToBuy              -8.339 1.077 181  -7.743  <.0001
 NeuralNetwork CreditScoring - XGBoost CreditScoring           -6.763 0.924 181  -7.317  <.0001
 NeuralNetwork CreditScoring - LightGBM PropensityToBuy       -11.663 1.028 181 -11.348  <.0001
 NeuralNetwork CreditScoring - NeuralNetwork PropensityToBuy   -2.824 0.972 181  -2.905  0.0943
 NeuralNetwork CreditScoring - XGBoost PropensityToBuy        -16.067 1.087 181 -14.782  <.0001
 XGBoost CreditScoring - LightGBM PropensityToBuy              -4.900 0.983 181  -4.987  <.0001
 XGBoost CreditScoring - NeuralNetwork PropensityToBuy          3.939 0.924 181   4.261  0.0011
 XGBoost CreditScoring - XGBoost PropensityToBuy               -9.304 1.044 181  -8.910  <.0001
 LightGBM PropensityToBuy - NeuralNetwork PropensityToBuy       8.839 1.028 181   8.600  <.0001
 LightGBM PropensityToBuy - XGBoost PropensityToBuy            -4.404 1.137 181  -3.874  0.0046
 NeuralNetwork PropensityToBuy - XGBoost PropensityToBuy      -13.243 1.087 181 -12.184  <.0001

> CLD = cld(leastsquare[[1]],  alpha   = 0.05,  Letters = letters,  adjust  = "tukey")
> CLD
 Model         Target          lsmean    SE  df lower.CL upper.CL .g
roup
 NeuralNetwork Covid             59.8 0.618 181     58.1     61.6 a
 XGBoost       Covid             66.2 0.643 181     64.4     68.0 b
 LightGBM      Covid             67.2 0.723 181     65.2     69.2 bc
 NeuralNetwork CreditScoring     67.4 0.687 181     65.5     69.4 bc
 NeuralNetwork PropensityToBuy   70.3 0.687 181     68.3     72.2 c
 XGBoost       CreditScoring     74.2 0.618 181     72.5     75.9 d
 LightGBM      CreditScoring     75.2 0.672 181     73.3     77.0 d
 LightGBM      PropensityToBuy   79.1 0.764 181     77.0     81.2 e
 XGBoost       PropensityToBuy   83.5 0.842 181     81.1     85.9 f
```


AUC - Model*Target
60
70
80
bc
a
b
d
bc
d
e
c
f
LightGBM.Covid
NeuralNetwork.Covid
XGBoost.Covid
LightGBM.CreditScoring
NeuralNetwork.CreditScoring
XGBoost.CreditScoring
LightGBM.PropensityToBuy
NeuralNetwork.PropensityToBuy
XGBoost.PropensityToBuy
interaction(Model, Target)

Statistics and Explanatory Data Analysis, final exam 2024-05-02

Results for SRH test without interactions

```
> DTModel = dunnTest(AUC ~ Model, data=Data, method="bh")

              Comparison          Z            P.unadj              P.adj
1 LightGBM - NeuralNetwork  6.3809734 0.0000000001759659 0.0000000005278978
2       LightGBM - XGBoost  0.5804809 0.5615903474410326 0.5615903474410326
3  NeuralNetwork - XGBoost -5.9445359 0.0000000027724115 0.0000000041586173
```

```
> DTTarget = dunnTest(AUC ~ Target, data=Data)

                   Comparison         Z                            P.unadj                              P.adj
1         Covid - CreditScoring -7.177855 0.000000000000708135113933081 0.00000000000141627022786616
2        Covid - PropensityToBuy -9.441345 0.00000000000000000003680259 0.00000000000000000001104078
3 CreditScoring - PropensityToBuy -2.786795 0.005323216474333368995741633 0.00532321647433336899574163
```

Results for SRH test with interactions

```
> DTAll = dunnTest(AUC ~ interaction(Model,Target), data=Data, method="bh")
> DTAll
                                              Comparison          Z P.unadj P.adj
1                     LightGBM.Covid - LightGBM.CreditScoring -3.9555903   0.000 0.000
2                   LightGBM.Covid - LightGBM.PropensityToBuy -5.2625550   0.000 0.000
3            LightGBM.CreditScoring - LightGBM.PropensityToBuy -1.6042886   0.109 0.130
4                       LightGBM.Covid - NeuralNetwork.Covid  2.8946652   0.004 0.007
5                LightGBM.CreditScoring - NeuralNetwork.Covid  7.2924505   0.000 0.000
6              LightGBM.PropensityToBuy - NeuralNetwork.Covid  8.4338216   0.000 0.000
7               LightGBM.Covid - NeuralNetwork.CreditScoring -0.1327124   0.894 0.894
8        LightGBM.CreditScoring - NeuralNetwork.CreditScoring  3.9229746   0.000 0.000
9      LightGBM.PropensityToBuy - NeuralNetwork.CreditScoring  5.2562456   0.000 0.000
10          NeuralNetwork.Covid - NeuralNetwork.CreditScoring -3.1209716   0.002 0.003
11            LightGBM.Covid - NeuralNetwork.PropensityToBuy -1.6915798   0.091 0.113
12     LightGBM.CreditScoring - NeuralNetwork.PropensityToBuy  2.3051201   0.021 0.030
13   LightGBM.PropensityToBuy - NeuralNetwork.PropensityToBuy  3.7433986   0.000 0.000
14        NeuralNetwork.Covid - NeuralNetwork.PropensityToBuy -4.8032586   0.000 0.000
15 NeuralNetwork.CreditScoring - NeuralNetwork.PropensityToBuy -1.5993641   0.110 0.127
16                            LightGBM.Covid - XGBoost.Covid  0.4482820   0.654 0.692
17                     LightGBM.CreditScoring - XGBoost.Covid  4.6635182   0.000 0.000
18                   LightGBM.PropensityToBuy - XGBoost.Covid  5.9764845   0.000 0.000
19                       NeuralNetwork.Covid - XGBoost.Covid -2.6000662   0.009 0.015
20                NeuralNetwork.CreditScoring - XGBoost.Covid  0.6013190   0.548 0.597
21              NeuralNetwork.PropensityToBuy - XGBoost.Covid  2.2531352   0.024 0.034
22                     LightGBM.Covid - XGBoost.CreditScoring -3.8440237   0.000 0.000
23              LightGBM.CreditScoring - XGBoost.CreditScoring  0.2715052   0.786 0.808
24            LightGBM.PropensityToBuy - XGBoost.CreditScoring  1.9131084   0.056 0.072
25                NeuralNetwork.Covid - XGBoost.CreditScoring -7.3331382   0.000 0.000
26         NeuralNetwork.CreditScoring - XGBoost.CreditScoring -3.8111406   0.000 0.000
27       NeuralNetwork.PropensityToBuy - XGBoost.CreditScoring -2.1288536   0.033 0.044
28                      XGBoost.Covid - XGBoost.CreditScoring -4.5849125   0.000 0.000
29                   LightGBM.Covid - XGBoost.PropensityToBuy -5.9691856   0.000 0.000
30            LightGBM.CreditScoring - XGBoost.PropensityToBuy -2.5261288   0.012 0.017
31          LightGBM.PropensityToBuy - XGBoost.PropensityToBuy -0.9575344   0.338 0.381
32              NeuralNetwork.Covid - XGBoost.PropensityToBuy -8.9778679   0.000 0.000
33       NeuralNetwork.CreditScoring - XGBoost.PropensityToBuy -5.9717717   0.000 0.000
34     NeuralNetwork.PropensityToBuy - XGBoost.PropensityToBuy -4.5412569   0.000 0.000
35                    XGBoost.Covid - XGBoost.PropensityToBuy -6.6611945   0.000 0.000
36             XGBoost.CreditScoring - XGBoost.PropensityToBuy -2.8425243   0.004 0.007
```

Results for SRH test without interactions

```
> DTModel = dunnTest(AUC ~ Model, data=Data, method="bh")

              Comparison         Z          P.unadj           P.adj
1 LightGBM - NeuralNetwork  6.3809734 0.0000000001759659 0.0000000005278978
2      LightGBM - XGBoost  0.5804809 0.5615903474410326 0.5615903474410326
3  NeuralNetwork - XGBoost -5.9445359 0.0000000027724115 0.0000000041586173
```

```
> DTTarget = dunnTest(AUC ~ Target, data=Data)

                  Comparison        Z                  P.unadj                  P.adj
1        Covid - CreditScoring -7.177855 0.000000000000708135113933081 0.0000000000014162702
2786616
2       Covid - PropensityToBuy -9.441345 0.000000000000000000003680259 0.00000000000000000
001104078
3 CreditScoring - PropensityToBuy -2.786795 0.005323216474333368995741633 0.00532321647433
336899574163
```

Results for SRH test with interactions

```
> DTAll = dunnTest(AUC ~ interaction(Model,Target), data=Data, method="bh")
> DTAll
                                   Comparison         Z P.unadj P.adj
1                LightGBM.Covid - LightGBM.CreditScoring -3.9555903   0.000 0.000
2              LightGBM.Covid - LightGBM.PropensityToBuy -5.2625550   0.000 0.000
3       LightGBM.CreditScoring - LightGBM.PropensityToBuy -1.6042886   0.109 0.130
4                 LightGBM.Covid - NeuralNetwork.Covid  2.8946652   0.004 0.007
5          LightGBM.CreditScoring - NeuralNetwork.Covid  7.2924505   0.000 0.000
6        LightGBM.PropensityToBuy - NeuralNetwork.Covid  8.4338216   0.000 0.000
7          LightGBM.Covid - NeuralNetwork.CreditScoring -0.1327124   0.894 0.894
8   LightGBM.CreditScoring - NeuralNetwork.CreditScoring  3.9229746   0.000 0.000
9 LightGBM.PropensityToBuy - NeuralNetwork.CreditScoring  5.2562456   0.000 0.000
10    NeuralNetwork.Covid - NeuralNetwork.CreditScoring -3.1209716   0.002 0.003
11       LightGBM.Covid - NeuralNetwork.PropensityToBuy -1.6915798   0.091 0.113
12 LightGBM.CreditScoring - NeuralNetwork.PropensityToBuy  2.3051201   0.021 0.030
13 LightGBM.PropensityToBuy - NeuralNetwork.PropensityToBuy  3.7433986   0.000 0.000
14   NeuralNetwork.Covid - NeuralNetwork.PropensityToBuy -4.8032586   0.000 0.000
15 NeuralNetwork.CreditScoring - NeuralNetwork.PropensityToBuy -1.5993641   0.110 0.127
16                  LightGBM.Covid - XGBoost.Covid  0.4482820   0.654 0.692
17           LightGBM.CreditScoring - XGBoost.Covid  4.6635182   0.000 0.000
18         LightGBM.PropensityToBuy - XGBoost.Covid  5.9764845   0.000 0.000
19              NeuralNetwork.Covid - XGBoost.Covid -2.6000662   0.009 0.015
20       NeuralNetwork.CreditScoring - XGBoost.Covid  0.6013190   0.548 0.597
21     NeuralNetwork.PropensityToBuy - XGBoost.Covid  2.2531352   0.024 0.034
22           LightGBM.Covid - XGBoost.CreditScoring -3.8440237   0.000 0.000
23    LightGBM.CreditScoring - XGBoost.CreditScoring  0.2715052   0.786 0.808
24  LightGBM.PropensityToBuy - XGBoost.CreditScoring  1.9131084   0.056 0.072
25       NeuralNetwork.Covid - XGBoost.CreditScoring -7.3331382   0.000 0.000
```

26	NeuralNetwork.CreditScoring - XGBoost.CreditScoring	-3.8111406	0.000	0.000
27	NeuralNetwork.PropensityToBuy - XGBoost.CreditScoring	-2.1288536	0.033	0.044
28	XGBoost.Covid - XGBoost.CreditScoring	-4.5849125	0.000	0.000
29	LightGBM.Covid - XGBoost.PropensityToBuy	-5.9691856	0.000	0.000
30	LightGBM.CreditScoring - XGBoost.PropensityToBuy	-2.5261288	0.012	0.017
31	LightGBM.PropensityToBuy - XGBoost.PropensityToBuy	-0.9575344	0.338	0.381
32	NeuralNetwork.Covid - XGBoost.PropensityToBuy	-8.9778679	0.000	0.000
33	NeuralNetwork.CreditScoring - XGBoost.PropensityToBuy	-5.9717717	0.000	0.000
34	NeuralNetwork.PropensityToBuy - XGBoost.PropensityToBuy	-4.5412569	0.000	0.000
35	XGBoost.Covid - XGBoost.PropensityToBuy	-6.6611945	0.000	0.000
36	XGBoost.CreditScoring - XGBoost.PropensityToBuy	-2.8425243	0.004	0.007

Statistics and Explanatory Data Analysis, final exam 2024-05-02

PROBLEM 3 ……… / 10 PTS

You are working with astronomical dataset. Each column unravels a distinct facet of celestial phenomena, providing an exhaustive exploration of key parameters essential for unraveling the cosmic mysteries. The temperature column immerses us in the thermal intricacies of stars, unveiling the nuanced variations in their heat emissions. Luminosity, a cornerstone of celestial understanding, discloses the radiant energy output, enabling a profound comprehension of a star's brilliance within the vast cosmic tapestry. The radius column serves as a cosmic ruler, delineating the spatial dimensions of these celestial entities, offering a profound grasp of their structural characteristics.

Absolute magnitude, a standardized measure of brightness, facilitates comparative analyses, shedding light on the intrinsic luminosity of diverse celestial bodies. The star type column categorizes these celestial actors, providing a systematic taxonomy crucial for discerning their roles within the cosmic narrative. Simultaneously, the spectral class and color columns paint a vivid portrait of the visual signatures of these stellar entities, offering nuanced insights into their chemical composition, temperature, and evolutionary stages.

This comprehensive data compilation is an invaluable resource, not merely for researchers and astronomers but also for enthusiasts seeking a deeper and more nuanced understanding of the cosmos. It serves as a reservoir of knowledge, fostering a symbiotic relationship between scientific inquiry and the innate human curiosity that propels us ever further into the boundless expanse of the universe.

As a beginner astronomer you want to test for basic associacions and correlations between the stars. Based on the results of the statistical analysis below and your knowledge, anwser the following questions assuming 5% confidence level. Remember to precisely justify your anwsers:

1. Which measure(s) of association would you choose to compare between star color and its type? **(… PTS)**
2. Is there a statistically significant association between star color and its spectral class? **(… PTS)**
3. Is there enough evidence to support a claim that the star absolute magnitude is positively correlated with its temperature? **(… PTS)**

name, surname, index nr:……

Statistics and Explanatory Data Analysis, final exam 2024-05-02

```
> describe(stars)
                       vars   n      mean        sd  median trimmed     mad      min       max     range  skew kurtosis       se
Temperature..K.           1 240  10497.46   9552.43 5776.00 8777.02 4341.05  1939.00  40000.00  38061.00  1.31     0.80   616.61
Luminosity.L.Lo.          2 240 107188.36 179432.24    0.07 67496.04    0.10     0.00 849420.00 849420.00  2.04     4.29 11582.30
Radius.R.Ro.              3 240    237.16    517.16    0.76  105.70    1.12     0.01   1948.50   1948.49  1.92     1.96    33.38
Absolute.magnitude.Mv.    4 240      4.38     10.53    8.31    4.54   13.16   -11.92     20.06     31.98 -0.12    -1.66     0.68
Star.type                 5 240      2.50      1.71    2.50    2.50    2.22     0.00      5.00      5.00  0.00    -1.28     0.11
Star.color*               6 240      2.54      1.09    3.00    2.47    1.48     1.00      5.00      4.00  0.20    -0.29     0.07
Spectral.Class*           7 240      4.76      2.09    6.00    4.92    1.48     1.00      7.00      6.00 -0.64    -1.28     0.13
> str(stars)
'data.frame':   240 obs. of  7 variables:
 $ Temperature..K.       : int  3068 3042 2600 2800 1939 2840 2637 2600 2650 2700 ...
 $ Luminosity.L.Lo.      : num  0.0024 0.0005 0.0003 0.0002 0.000138 0.00065 0.00073 0.0004 0.00069 0.00018 ...
 $ Radius.R.Ro.          : num  0.17 0.154 0.102 0.16 0.103 ...
 $ Absolute.magnitude.Mv.: num  16.1 16.6 18.7 16.6 20.1 ...
 $ Star.type             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Star.color            : chr  "Red" "Red" "Red" "Red" ...
 $ Spectral.Class        : chr  "M" "M" "M" "M" ...
 > table2 <- table(stars[c('Star.type', 'Star.color')])
 > table3 <- stars[c('Temperature..K.', 'Luminosity.L.Lo.', 'Radius.R.Ro.', 'Absolute.magnitude.Mv.', 'Star.type')]
 >
 > assocstats(table1)
                    X^2 df P(> X^2)
 Likelihood Ratio 504.51 24        0
 Pearson          571.98 24        0

 Phi-Coefficient   : NA
 Contingency Coeff.: 0.839
 Cramer's V        : 0.772
 > lbl_test(table1)

        Asymptotic Linear-by-Linear Association Test

 data:  Star.color (ordered) by Spectral.Class (A < B < F < G < K < M < O)
 Z = -1.2123, p-value = 0.2254
 alternative hypothesis: two.sided


 > chisq_test(table1)

        Asymptotic Pearson Chi-Squared Test

 data:  Star.color by Spectral.Class (A, B, F, G, K, M, O)
 chi-squared = 571.98, df = 24, p-value < 0.00000000000000022

 > corr.test(table1, use="pairwise", method = "pearson", adjust = "bonferroni")
 Call:corr.test(x = table1, use = "pairwise", method = "pearson", adjust = "bonferroni")
 Correlation matrix
              Blue Blue-White   Red White Yellow-White
 Blue          1.00       0.13 -0.24 -0.32        -0.29
 Blue-White    0.13       1.00 -0.24  0.23        -0.29
 Red          -0.24      -0.24  1.00 -0.28        -0.20
 White        -0.32       0.23 -0.28  1.00         0.30
 Yellow-White -0.29      -0.29 -0.20  0.30         1.00
 Sample Size
 [1] 7
```

```
Probability values (Entries above the diagonal are adjusted for multiple tests.)
             Blue Blue-White  Red White Yellow-White
Blue         0.00       1.00 1.00  1.00            1
Blue-White   0.78       0.00 1.00  1.00            1
Red          0.60       0.60 0.00  1.00            1
White        0.49       0.62 0.54  0.00            1
Yellow-White 0.53       0.52 0.66  0.51            0

 To see confidence intervals of the correlations, print with the short=FALSE option
> assocstats(table2)
                    X^2 df P(> X^2)
Likelihood Ratio 300.55 20        0
Pearson          280.43 20        0

Phi-Coefficient   : NA 
Contingency Coeff.: 0.734 
Cramer's V        : 0.54 
> lbl_test(table2)


	Asymptotic Linear-by-Linear Association Test

data:  Star.color (ordered) by Star.type (0 < 1 < 2 < 3 < 4 < 5)
Z = -4.4771, p-value = 0.000007568
alternative hypothesis: two.sided

> chisq_test(table2)

	Asymptotic Pearson Chi-Squared Test

data:  Star.color by Star.type (0, 1, 2, 3, 4, 5)
chi-squared = 280.43, df = 20, p-value < 0.00000000000000022

> corr.test(table2, use="pairwise", method = "pearson", adjust = "bonferroni")
Call:corr.test(x = table2, use = "pairwise", method = "pearson", adjust = "bonferroni")
Correlation matrix 
              Blue Blue-White   Red White Yellow-White
Blue          1.00      -0.10 -0.57  0.09        -0.13
Blue-White   -0.10       1.00 -0.75  0.61         0.93
Red          -0.57      -0.75  1.00 -0.61        -0.68
White         0.09       0.61 -0.61  1.00         0.35
Yellow-White -0.13       0.93 -0.68  0.35         1.00
Sample Size 
[1] 6
Probability values (Entries above the diagonal are adjusted for multiple tests.) 
             Blue Blue-White  Red White Yellow-White
Blue         0.00       1.00 1.00   1.0         1.00
Blue-White   0.84       0.00 0.85   1.0         0.07
Red          0.23       0.09 0.00   1.0         1.00
White        0.87       0.20 0.20   0.0         1.00
Yellow-White 0.81       0.01 0.14   0.5         0.00

 To see confidence intervals of the correlations, print with the short=FALSE option
> corr.test(table3, use="pairwise", method = "kendall", adjust = "bonferroni")
Call:corr.test(x = table3, use = "pairwise", method = "kendall", adjust = "bonferroni")
Correlation matrix 
```

Statistics and Explanatory Data Analysis, final exam 2024-05-02

```
Call:corr.test(x = table3, use = "pairwise", method = "kendall", adjust = "bonferroni")
Correlation matrix
                      Temperature..K. Luminosity.L.Lo. Radius.R.Ro. Absolute.magnitude.Mv. Star.type
Temperature..K.                  1.00             0.35         0.20                  -0.37      0.42
Luminosity.L.Lo.                 0.35             1.00         0.71                  -0.71      0.68
Radius.R.Ro.                     0.20             0.71         1.00                  -0.69      0.67
Absolute.magnitude.Mv.          -0.37            -0.71        -0.69                   1.00     -0.85
Star.type                        0.42             0.68         0.67                  -0.85      1.00
Sample Size
[1] 240
Probability values (Entries above the diagonal are adjusted for multiple tests.)
                      Temperature..K. Luminosity.L.Lo. Radius.R.Ro. Absolute.magnitude.Mv. Star.type
Temperature..K.                     0                0         0.02                      0         0
Luminosity.L.Lo.                    0                0         0.00                      0         0
Radius.R.Ro.                        0                0         0.00                      0         0
Absolute.magnitude.Mv.              0                0         0.00                      0         0
Star.type                           0                0         0.00                      0         0

 To see confidence intervals of the correlations, print with the short=FALSE option
```