Multivariate Analysis

Term Project

Elif Taş
Hakkı Erduran
Orçun Oltulu

Department of Statistics

Jan 19, 2019

Table of contents

1. Overview

2. Index Rating Estimation

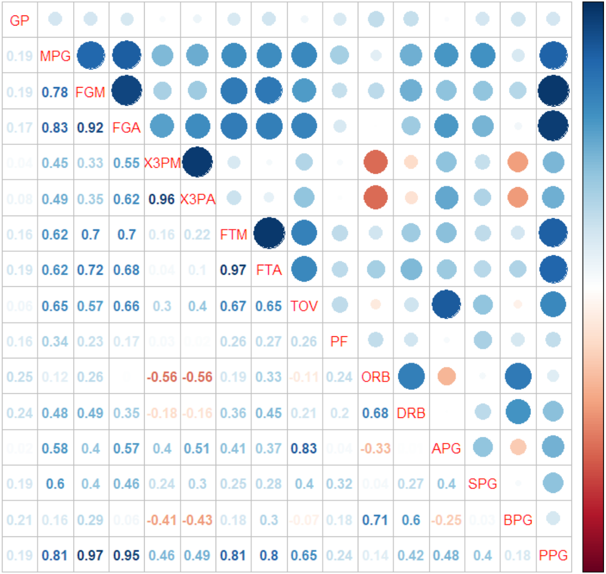
3. Position Prediction

4. MPG Estimation

5. Clustering for Final Four Players

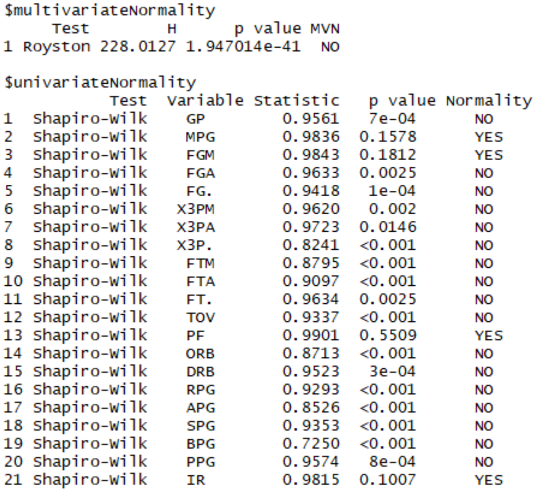
Introduction

Correlation



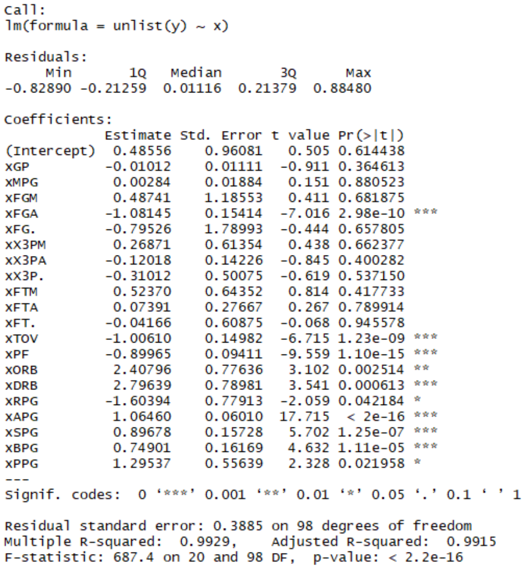
GP															
0.19	MPG														
0.19	0.78	FGM													
0.17	0.83	0.92	FGA												
0.04	0.45	0.33	0.55	X3PM											
0.08	0.49	0.35	0.62	0.96	X3PA										
0.16	0.62	0.7	0.7	0.16	0.22	FTM									
0.19	0.62	0.72	0.68	0.04	0.1	0.97	FTA								
0.06	0.65	0.57	0.66	0.3	0.4	0.67	0.65	TOV							
0.16	0.34	0.23	0.17	0.03	0.02	0.26	0.27	0.26	PF						
0.25	0.12	0.26	0	-0.56	-0.56	0.19	0.33	-0.11	0.24	ORB					
0.24	0.48	0.49	0.35	-0.18	-0.16	0.36	0.45	0.21	0.2	0.68	DRB				
0.02	0.58	0.4	0.57	0.4	0.51	0.41	0.37	0.83	0.04	-0.33	0.01	APG			
0.19	0.6	0.4	0.46	0.24	0.3	0.25	0.28	0.4	0.32	0.04	0.27	0.4	SPG		
0.21	0.16	0.29	0.06	-0.41	-0.43	0.18	0.3	-0.07	0.18	0.71	0.6	-0.25	0.03	BPG	
0.19	0.81	0.97	0.95	0.46	0.49	0.81	0.8	0.65	0.24	0.14	0.42	0.48	0.4	0.18	PPG

Multivariate Normality



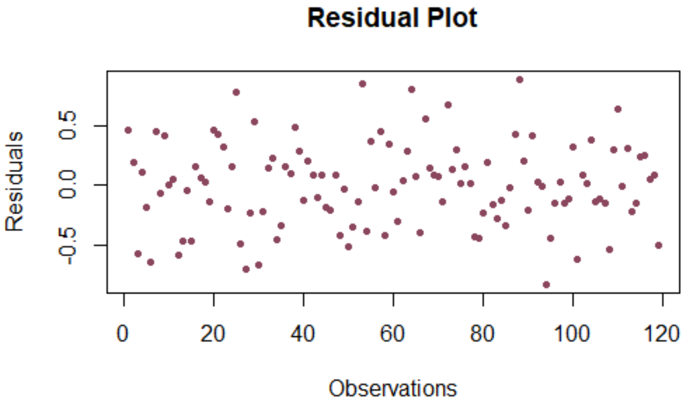
```
$multivariateNormality
     Test        H      p value MVN
1 Royston 228.0127 1.947014e-41  NO

$univariateNormality
           Test  Variable Statistic   p value Normality
1  Shapiro-Wilk    GP        0.9561    7e-04      NO
2  Shapiro-Wilk    MPG       0.9836   0.1578      YES
3  Shapiro-Wilk    FGM       0.9843   0.1812      YES
4  Shapiro-Wilk    FGA       0.9633   0.0025      NO
5  Shapiro-Wilk    FG.       0.9418    1e-04      NO
6  Shapiro-Wilk   X3PM       0.9620    0.002      NO
7  Shapiro-Wilk   X3PA       0.9723   0.0146      NO
8  Shapiro-Wilk   X3P.       0.8241   <0.001      NO
9  Shapiro-Wilk    FTM       0.8795   <0.001      NO
10 Shapiro-Wilk    FTA       0.9097   <0.001      NO
11 Shapiro-Wilk    FT.       0.9634   0.0025      NO
12 Shapiro-Wilk    TOV       0.9337   <0.001      NO
13 Shapiro-Wilk    PF        0.9901   0.5509      YES
14 Shapiro-Wilk    ORB       0.8713   <0.001      NO
15 Shapiro-Wilk    DRB       0.9523    3e-04      NO
16 Shapiro-Wilk    RPG       0.9293   <0.001      NO
17 Shapiro-Wilk    APG       0.8526   <0.001      NO
18 Shapiro-Wilk    SPG       0.9353   <0.001      NO
19 Shapiro-Wilk    BPG       0.7250   <0.001      NO
20 Shapiro-Wilk    PPG       0.9574    8e-04      NO
21 Shapiro-Wilk    IR        0.9815   0.1007      YES
```

Model

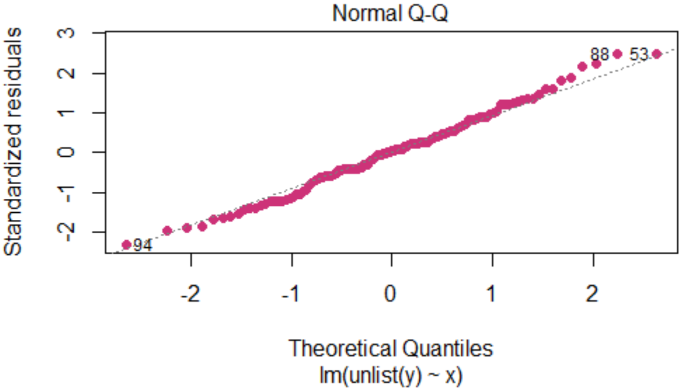


```
Call:
lm(formula = unlist(y) ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.82890 -0.21259  0.01116  0.21379  0.88480 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.48556    0.96081   0.505 0.614438    
xGP         -0.01012    0.01111  -0.911 0.364613    
xMPG         0.00284    0.01884   0.151 0.880523    
xFGM         0.48741    1.18553   0.411 0.681875    
xFGA        -1.08145    0.15414  -7.016 2.98e-10 ***
xFG.        -0.79526    1.78993  -0.444 0.657805    
xX3PM        0.26871    0.61354   0.438 0.662377    
xX3PA       -0.12018    0.14226  -0.845 0.400282    
xX3P.       -0.31012    0.50075  -0.619 0.537150    
xFTM         0.52370    0.64352   0.814 0.417733    
xFTA         0.07391    0.27667   0.267 0.789914    
xFT.        -0.04166    0.60875  -0.068 0.945578    
xTOV        -1.00610    0.14982  -6.715 1.23e-09 ***
xPF         -0.89965    0.09411  -9.559 1.10e-15 ***
xORB         2.40796    0.77636   3.102 0.002514 ** 
xDRB         2.79639    0.78981   3.541 0.000613 ***
xRPG        -1.60394    0.77913  -2.059 0.042184 *  
xAPG         1.06460    0.06010  17.715  < 2e-16 ***
xSPG         0.89678    0.15728   5.702 1.25e-07 ***
xBPG         0.74901    0.16169   4.632 1.11e-05 ***
xPPG         1.29537    0.55639   2.328 0.021958 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3885 on 98 degrees of freedom
Multiple R-squared:  0.9929,    Adjusted R-squared:  0.9915 
F-statistic: 687.4 on 20 and 98 DF,  p-value: < 2.2e-16
```

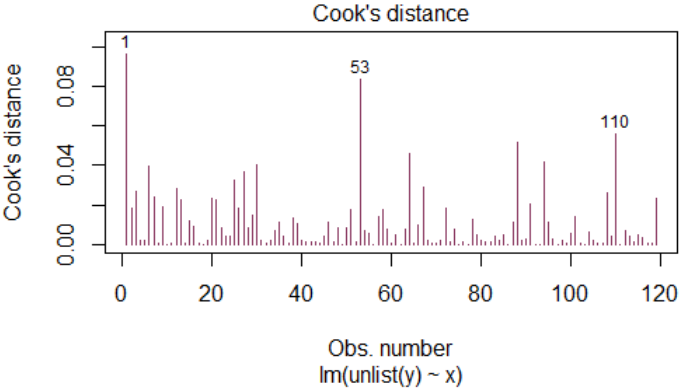
Model Diagnostics



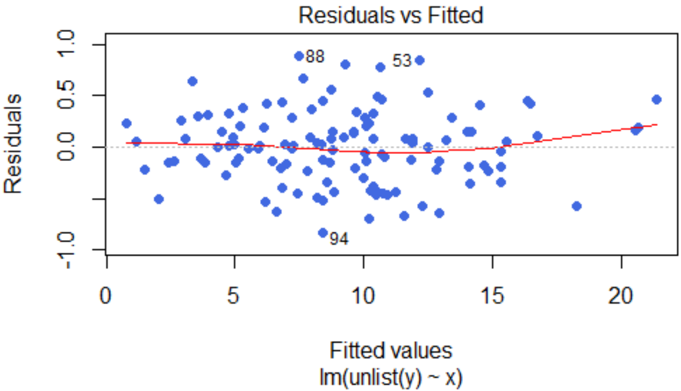
Normality of Residuals



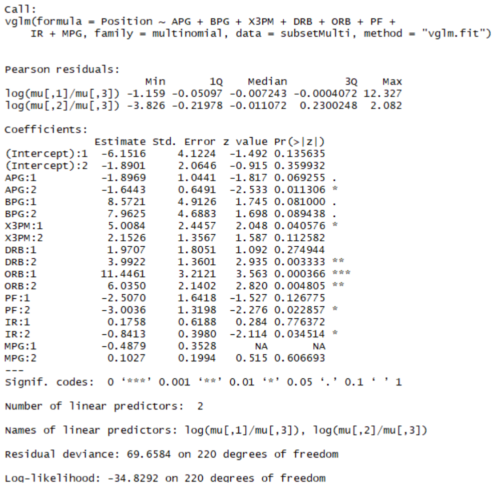
Cook Distance (Outlier Test)



Heteroscedasticity of Residuals (Constant Variance Test)

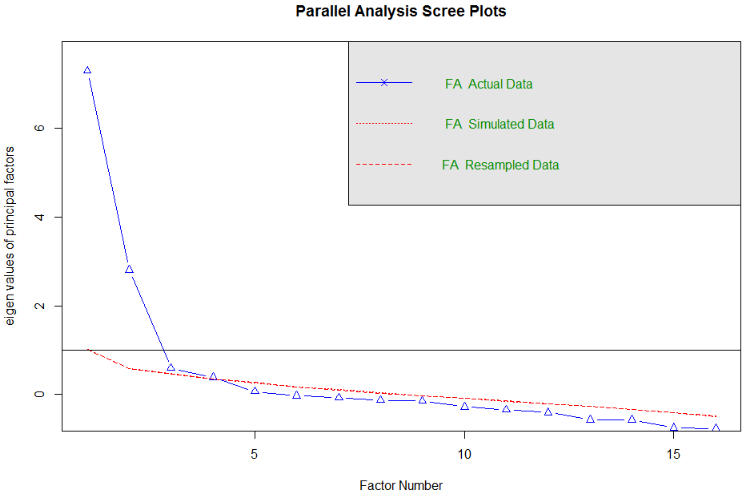


Multinomial Regression

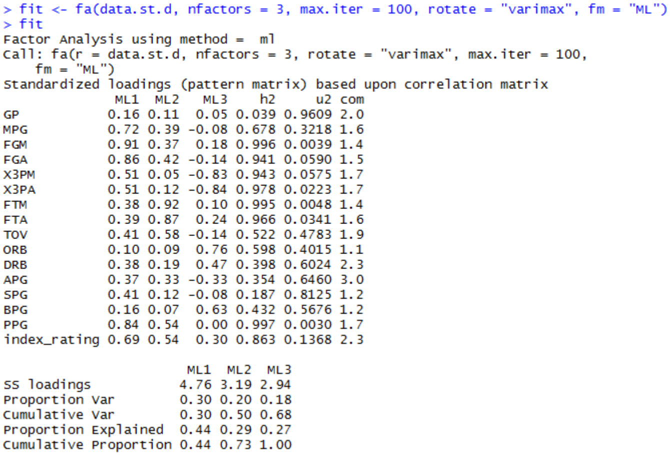


```
Call:
vglm(formula = Position ~ APG + BPG + X3PM + DRB + ORB + PF +
    IR + MPG, family = multinomial, data = subsetMulti, method = "vglm.fit")


Pearson residuals:
                      Min       1Q    Median         3Q    Max
log(mu[,1]/mu[,3]) -1.159 -0.05097 -0.007243 -0.0004072 12.327
log(mu[,2]/mu[,3]) -3.826 -0.21978 -0.011072  0.2300248  2.082

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -6.1516     4.1224  -1.492 0.135635
(Intercept):2  -1.8901     2.0646  -0.915 0.359932
APG:1          -1.8969     1.0441  -1.817 0.069255 .
APG:2          -1.6443     0.6491  -2.533 0.011306 *
BPG:1           8.5721     4.9126   1.745 0.081000 .
BPG:2           7.9625     4.6883   1.698 0.089438 .
X3PM:1          5.0084     2.4457   2.048 0.040576 *
X3PM:2          2.1526     1.3567   1.587 0.112582
DRB:1           1.9707     1.8051   1.092 0.274944
DRB:2           3.9922     1.3601   2.935 0.003333 **
ORB:1          11.4461     3.2121   3.563 0.000366 ***
ORB:2           6.0350     2.1402   2.820 0.004805 **
PF:1           -2.5070     1.6418  -1.527 0.126775
PF:2           -3.0036     1.3198  -2.276 0.022857 *
IR:1            0.1758     0.6188   0.284 0.776372
IR:2           -0.8413     0.3980  -2.114 0.034514 *
MPG:1          -0.4879     0.3528      NA       NA
MPG:2           0.1027     0.1994   0.515 0.606693
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  2

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Residual deviance: 69.6584 on 220 degrees of freedom

Log-likelihood: -34.8292 on 220 degrees of freedom
```

Factor Analysis

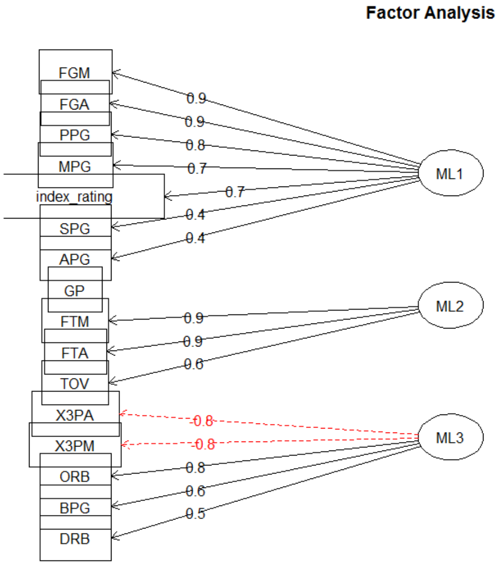


Factor Analysis

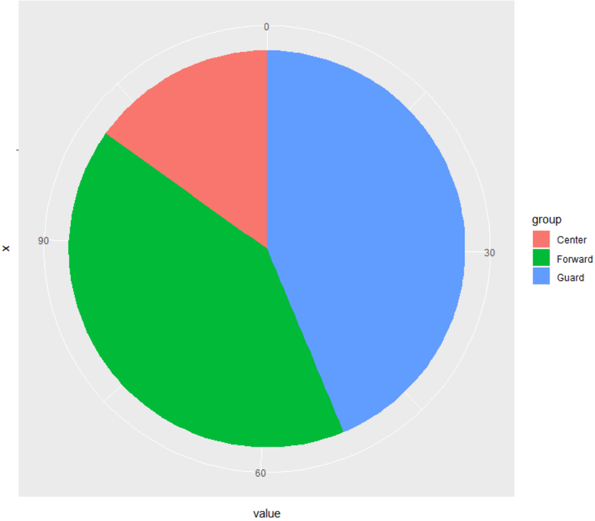


```
> fit <- fa(data.st.d, nfactors = 3, max.iter = 100, rotate = "varimax", fm = "ML")
> fit
Factor Analysis using method =  ml
Call: fa(r = data.st.d, nfactors = 3, rotate = "varimax", max.iter = 100,
    fm = "ML")
Standardized loadings (pattern matrix) based upon correlation matrix
              ML1  ML2   ML3    h2     u2 com
GP           0.16 0.11  0.05 0.039 0.9609 2.0
MPG          0.72 0.39 -0.08 0.678 0.3218 1.6
FGM          0.91 0.37  0.18 0.996 0.0039 1.4
FGA          0.86 0.42 -0.14 0.941 0.0590 1.5
X3PM         0.51 0.05 -0.83 0.943 0.0575 1.7
X3PA         0.51 0.12 -0.84 0.978 0.0223 1.7
FTM          0.38 0.92  0.10 0.995 0.0048 1.4
FTA          0.39 0.87  0.24 0.966 0.0341 1.6
TOV          0.41 0.58 -0.14 0.522 0.4783 1.9
ORB          0.10 0.09  0.76 0.598 0.4015 1.1
DRB          0.38 0.19  0.47 0.398 0.6024 2.3
APG          0.37 0.33 -0.33 0.354 0.6460 3.0
SPG          0.41 0.12 -0.08 0.187 0.8125 1.2
BPG          0.16 0.07  0.63 0.432 0.5676 1.2
PPG          0.84 0.54  0.00 0.997 0.0030 1.7
index_rating 0.69 0.54  0.30 0.863 0.1368 2.3

                       ML1  ML2  ML3
SS loadings           4.76 3.19 2.94
Proportion Var        0.30 0.20 0.18
Cumulative Var        0.30 0.50 0.68
Proportion Explained  0.44 0.29 0.27
Cumulative Proportion 0.44 0.73 1.00
```

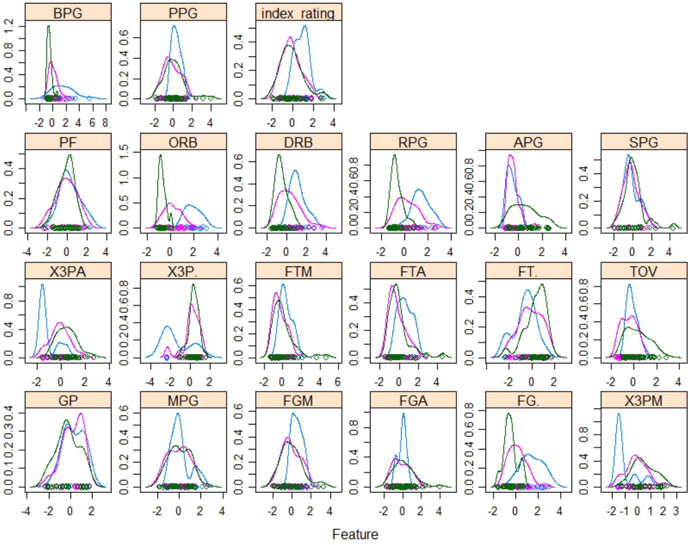
Factor Analysis



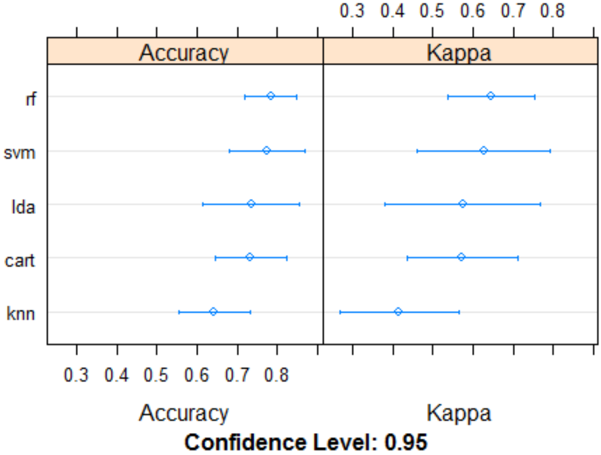
Classification On The Basis Of Position



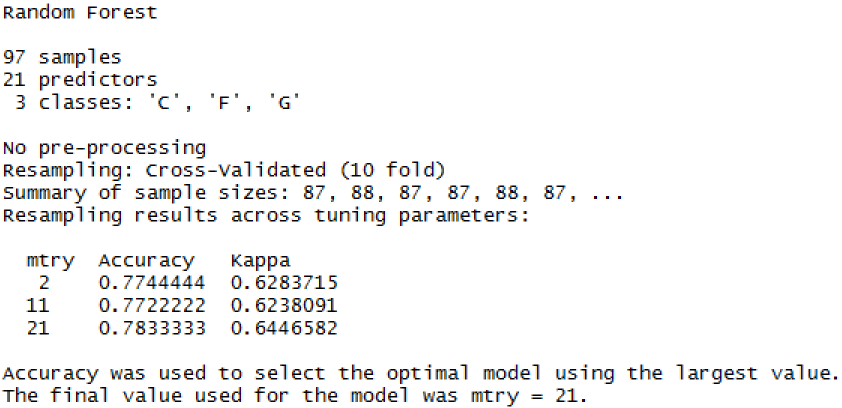
Classification On The Basis Of Position



Classification On The Basis Of Position

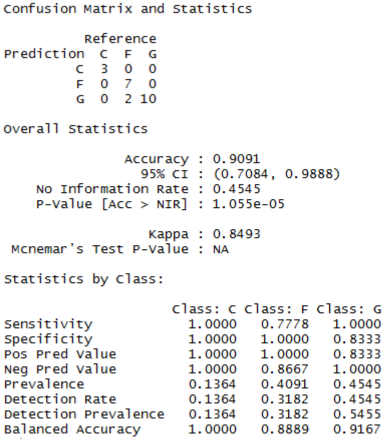


Classification On The Basis Of Position



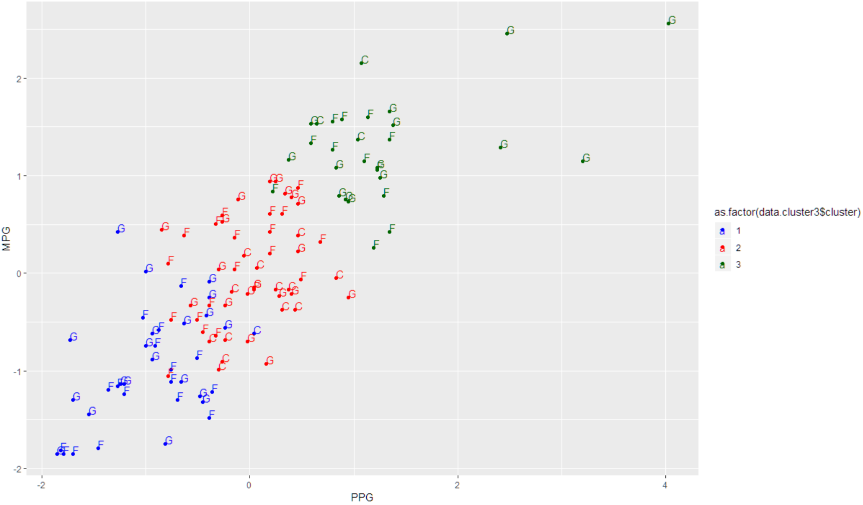
```
Random Forest

97 samples
21 predictors
 3 classes: 'C', 'F', 'G'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 87, 88, 87, 87, 88, 87, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.7744444  0.6283715
  11    0.7722222  0.6238091
  21    0.7833333  0.6446582

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 21.
```

Classification On The Basis Of Position

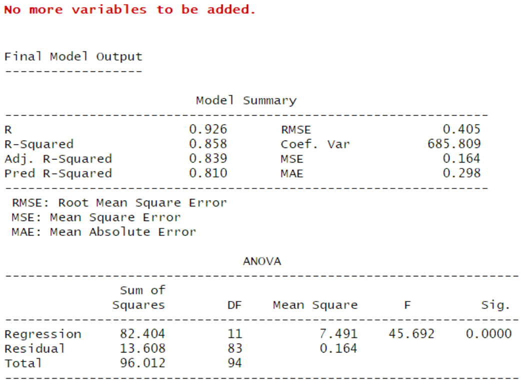


```
Confusion Matrix and Statistics

          Reference
Prediction  C  F  G
         C  3  0  0
         F  0  7  0
         G  0  2 10

Overall Statistics

               Accuracy : 0.9091
                 95% CI : (0.7084, 0.9888)
    No Information Rate : 0.4545
    P-Value [Acc > NIR] : 1.055e-05

                  Kappa : 0.8493
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: C Class: F Class: G
Sensitivity            1.0000   0.7778   1.0000
Specificity            1.0000   1.0000   0.8333
Pos Pred Value         1.0000   1.0000   0.8333
Neg Pred Value         1.0000   0.8667   1.0000
Prevalence             0.1364   0.4091   0.4545
Detection Rate         0.1364   0.3182   0.4545
Detection Prevalence   0.1364   0.3182   0.5455
Balanced Accuracy      1.0000   0.8889   0.9167
```

K-Means Clustering

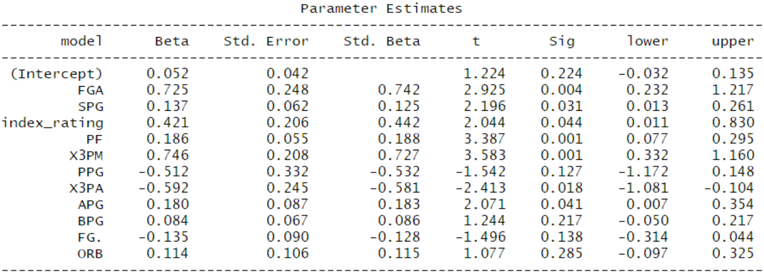


Regression Model For MPG



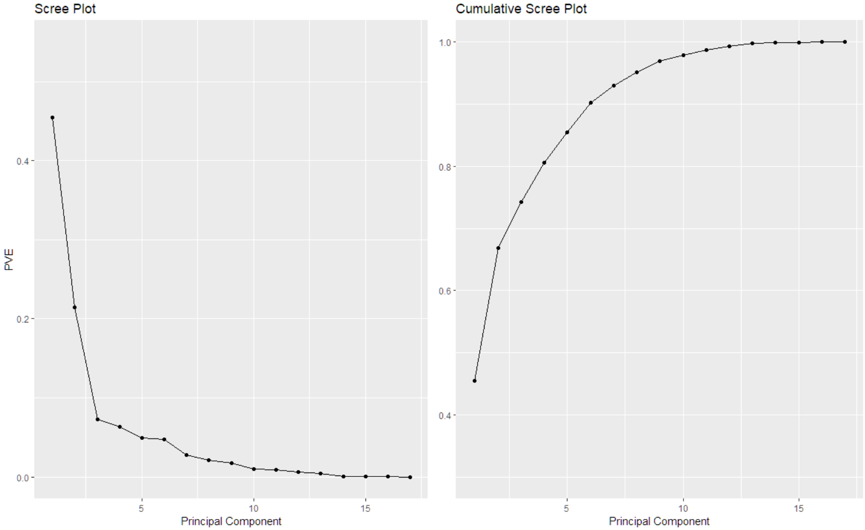
```
No more variables to be added.


Final Model Output
------------------

                        Model Summary
----------------------------------------------------------------
R                       0.926       RMSE                  0.405
R-Squared               0.858       Coef. Var           685.809
Adj. R-Squared          0.839       MSE                   0.164
Pred R-Squared          0.810       MAE                   0.298
----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                               ANOVA
--------------------------------------------------------------------
                Sum of
               Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------
Regression      82.404        11          7.491    45.692    0.0000
Residual        13.608        83          0.164
Total           96.012        94
--------------------------------------------------------------------
```

Regression Model For MPG

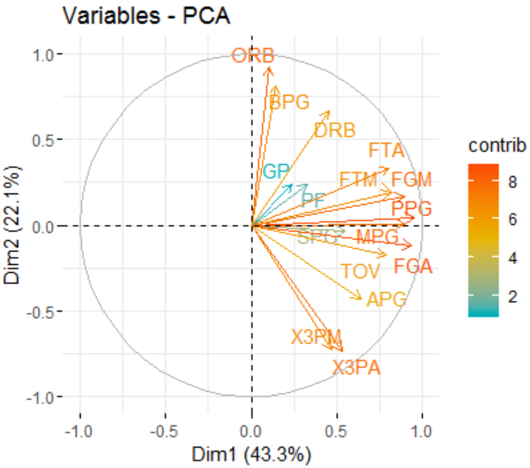


```
                              Parameter Estimates
------------------------------------------------------------------------------------
        model        Beta    Std. Error    Std. Beta       t        Sig      lower     upper
------------------------------------------------------------------------------------
  (Intercept)       0.052         0.042                 1.224     0.224    -0.032     0.135
          FGA       0.725         0.248        0.742    2.925     0.004     0.232     1.217
          SPG       0.137         0.062        0.125    2.196     0.031     0.013     0.261
 index_rating       0.421         0.206        0.442    2.044     0.044     0.011     0.830
           PF       0.186         0.055        0.188    3.387     0.001     0.077     0.295
         X3PM       0.746         0.208        0.727    3.583     0.001     0.332     1.160
          PPG      -0.512         0.332       -0.532   -1.542     0.127    -1.172     0.148
         X3PA      -0.592         0.245       -0.581   -2.413     0.018    -1.081    -0.104
          APG       0.180         0.087        0.183    2.071     0.041     0.007     0.354
          BPG       0.084         0.067        0.086    1.244     0.217    -0.050     0.217
          FG.      -0.135         0.090       -0.128   -1.496     0.138    -0.314     0.044
          ORB       0.114         0.106        0.115    1.077     0.285    -0.097     0.325
------------------------------------------------------------------------------------
```

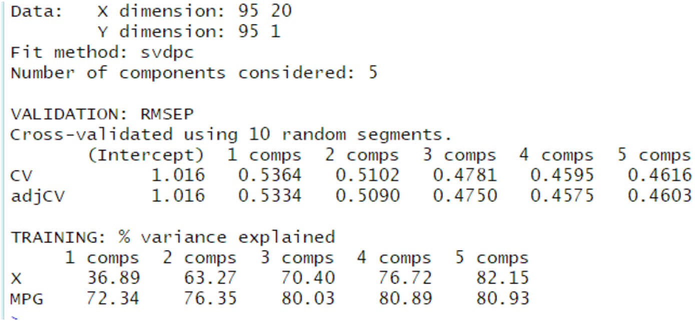
Principle Component Analysis



Principle Component Analysis

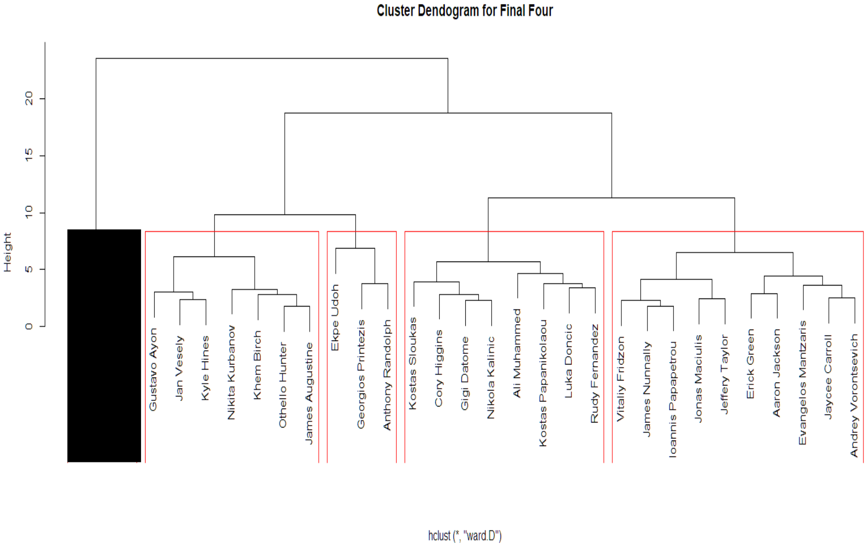


Principle Component Regression For MPG

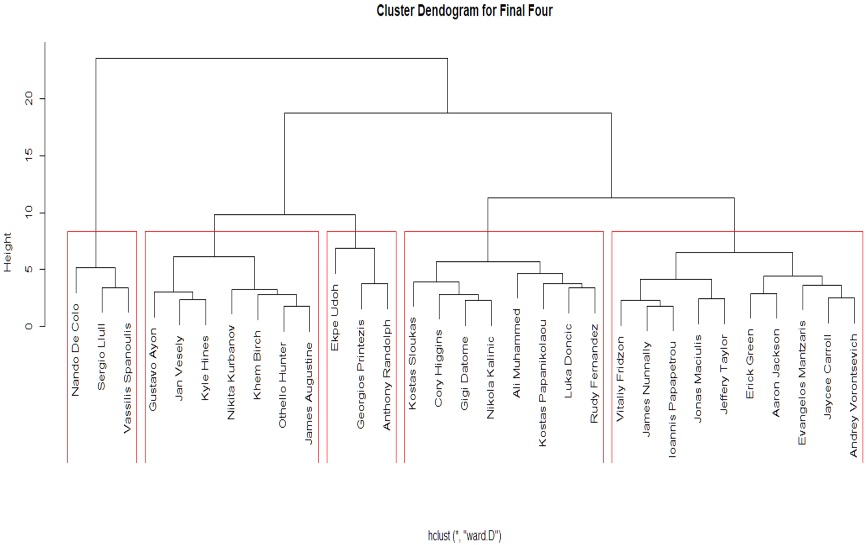


```
Data:    X dimension: 95 20
         Y dimension: 95 1
Fit method: svdpc
Number of components considered: 5

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV           1.016   0.5364   0.5102   0.4781   0.4595   0.4616
adjCV        1.016   0.5334   0.5090   0.4750   0.4575   0.4603

TRAINING: % variance explained
     1 comps  2 comps  3 comps  4 comps  5 comps
X      36.89    63.27    70.40    76.72    82.15
MPG    72.34    76.35    80.03    80.89    80.93
```

Agglomerative Hierarchical Clustering

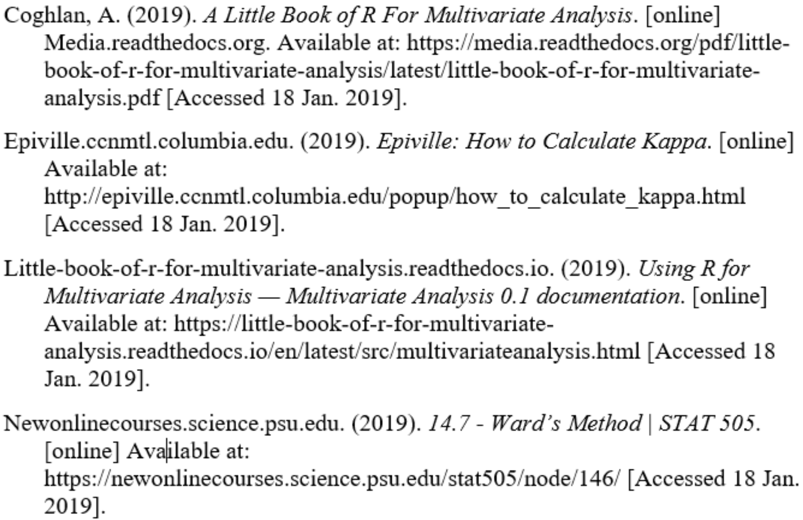


Agglomerative Hierarchical Clustering



Questions?

References I



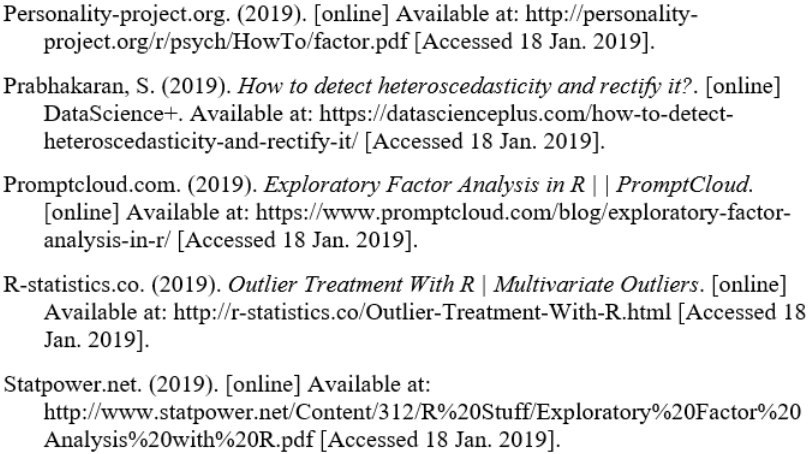
Coghlan, A. (2019). *A Little Book of R For Multivariate Analysis*. [online] Media.readthedocs.org. Available at: https://media.readthedocs.org/pdf/little-book-of-r-for-multivariate-analysis/latest/little-book-of-r-for-multivariate-analysis.pdf [Accessed 18 Jan. 2019].

Epiville.ccnmtl.columbia.edu. (2019). *Epiville: How to Calculate Kappa*. [online] Available at: http://epiville.ccnmtl.columbia.edu/popup/how_to_calculate_kappa.html [Accessed 18 Jan. 2019].

Little-book-of-r-for-multivariate-analysis.readthedocs.io. (2019). *Using R for Multivariate Analysis — Multivariate Analysis 0.1 documentation*. [online] Available at: https://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/src/multivariateanalysis.html [Accessed 18 Jan. 2019].

Newonlinecourses.science.psu.edu. (2019). *14.7 - Ward's Method | STAT 505*. [online] Available at: https://newonlinecourses.science.psu.edu/stat505/node/146/ [Accessed 18 Jan. 2019].

References II



Personality-project.org. (2019). [online] Available at: http://personality-project.org/r/psych/HowTo/factor.pdf [Accessed 18 Jan. 2019].

Prabhakaran, S. (2019). *How to detect heteroscedasticity and rectify it?*. [online] DataScience+. Available at: https://datascienceplus.com/how-to-detect-heteroscedasticity-and-rectify-it/ [Accessed 18 Jan. 2019].

Promptcloud.com. (2019). *Exploratory Factor Analysis in R | | PromptCloud*. [online] Available at: https://www.promptcloud.com/blog/exploratory-factor-analysis-in-r/ [Accessed 18 Jan. 2019].

R-statistics.co. (2019). *Outlier Treatment With R | Multivariate Outliers*. [online] Available at: http://r-statistics.co/Outlier-Treatment-With-R.html [Accessed 18 Jan. 2019].

Statpower.net. (2019). [online] Available at: http://www.statpower.net/Content/312/R%20Stuff/Exploratory%20Factor%20Analysis%20with%20R.pdf [Accessed 18 Jan. 2019].