**Player attributes study and impact on player ratings in soccer – Under parameters of the UEFA Euro 2018 players.**

***Abstract*-** Soccer is identified widely as an achievement sport. It involves the utilization of talent identification, development and retaining the services of the most talented players and balancing flow of new players. Ratings are based on a unique, comprehensive statistical algorithm, calculated live during the game. There are over 200 raw statistics included in the calculation of a player's rating, weighted according to their influence within the game. The sample in the study consisted of 2,973 players from the UEFA champions league in 2018. Modelling *Age* and *Stamina* to determine variances in their *Overall rating* is the crux of this study. This study analyzed whether there is significant difference in *Overall rating* for players along different *Age* groups and different *Stamina* levels by conducting a Two-way ANOVA with post hoc tests. Players were divided into 5 groups according to their age (Group 1: <=24, Group 2: 25-26, Group 3: 27-28, Group 4: 29-31, Group 5: 32+) and stamina levels were classified into 6 groups. Stamina classes were grouped from 1 to 6 based on the stamina scores which ranged from 15 to 94 with mean of 70.05. A significant main effect for stamina groups (stamina: sig=.000) was found, but no significant main effect for age group (age: sig=.638) was reported. Post-hoc comparisons using the Tukey HSD test indicated the Stamina group 6 (*M*=77.41, *SD*=3.201) was significantly different from the rest of the groups ( Group1: *M*=76.60, *SD*=3.283), ( Group2: *M*=76.06, *SD*=2.620), (Group3: *M*=7.92, *SD*=2.658), ( Group4: *M*=76.07, *SD*=2.681) and ( Group5: *77.08*, *SD*=3.084). The rest of the stamina groups did not differ significantly with each other. Partial Eta Squaredof .011 was reported and classified as small using Cohen’s (1988) criterion. Besides a Two-way Anova, standard linear multiple regression was performed to determine a predictive model to determine *potential rating* of the players. Standardized (beta) for predictors of *Potential rating*: *Age, Composure, Shot-power, Ball control, Jumping and Aggression* are -.563, .420, -.116, .0.89, .067 and -.125 respectively. Unstandardized (B) coefficients for Age, Composure, Shot-power, Ball control, Jumping and Aggression are -.629, .171, -.028, -.022, .025 and -.031 respectively. Standard errors (SE) are .017, .010, .007, .008, .006 and .005 respectively. The model significantly predicted the potential ratings (adjusted *R squared* = 35.8%). However, because of the limited sample size, the study’s relevance remains ambiguous. A challenge for future research lies in the integration of multivariate approaches that expand beyond the pattern of predictive validity across several types of leagues and time periods in addition to a single soccer league and period.

***i)Background and Significance:***

There are well reasoned processes in talent research against premature selection in

talent identification and development programs in soccer. Athletes who are born in different years are hypothesized to have different cognitive and physical development differences. That is why competitive teams tend to select older players more often than youngsters. Personal attributes’ differences effect is extensively described in young and elite team sports such as basketball, volleyball or, ice-hockey, as well as in soccer. Rankings of soccer players and data-driven evaluations of their performance are becoming more and more central in the soccer industry. Many sports companies, websites and television broadcasters and the plethora of online platforms use soccer statistics to compare the performance of professional players, with the purpose of increasing fan engagement through critical analyses, insights and scoring patterns. We are also interested in analytic tools to support tactical analysis and monitor the quality of their players in matches or entire seasons. Soccer scouts are continuously looking for data-driven tools to improve the retrieval of talented players with desired characteristics, based on evaluation criteria that take in account the multi-dimensional nature of soccer performance. Although selecting talents on the entire space of soccer players is unfeasible for humans as it is time consuming, data-driven performance scores could help in selecting a small subset of the best players who meet specific constraints or show some pattern in their performance. It also helps in broadening scouting operations and career opportunities of talented players. Rating players means defining a relation of order between them with respect to some measure of their performance over a sequence of matches. Several data-driven ranking and evaluation algorithms have been proposed in the literature to date, but they suffer from some limitations. This study presents the results of a statistical analysis which takes into concern few determinant variables that are expected to play a vital role in the prediction of player ratings.

Recent talent research in soccer offers several prospective studies investigating the validity of talent predictors. This study builds on often-considered predictors and are recognized as particularly important factors. However, because of a huge variety of study design parameters influencing the research results, current findings may provide an inconsistent picture regarding the validity of results.

***ii)Methods:***

***a) Data collection:***

Sample data of 2973 players has been collected from a population of 8000 players from the information provided in <https://www.uefa.com/>

***b) List of Variables:***

Age, Overall Rating, Potential Rating, Balance, Ball Control, Composure, Crossing, Curve, Dribbling, Finishing, Free Kick Accuracy, Goalkeeper Diving, Handling, Kicking, Positioning, Reflexes, Heading Accuracy, Interception, Jumping, Long passing, Long shot, Marking, Penalties, Positioning, Reactions, Short passing, Shot power, Sliding tackle, Sprint speed, Stamina, Standing tackle, Strength, Vision.

Acceleration, Aggression. Number of rows in SPSS is 2973.

***c) Analytic Methods:***

***Statistical analysis:***

Data were analyzed using SPSS version 26. Regarding difference in *overall rating* for players along different *age* groups and different *stamina* levels, variance in overall rating was determined by a two-way ANOVA including overall rating as dependent variable. Additionally, a corresponding multiple linear regression was conducted incorporating Age, Composure, Shot-power, Ball control, Jumping and Aggression as independent variables and Potential rating as a dependent variable. Post hoc tests (Tukey HSD) were conducted to contrast within group comparisons. *R2* was examined to quantify the amount of variance explained by the linear regression model and refers to the explained variance proportion in the underlying response variables.

***Sample and design:***

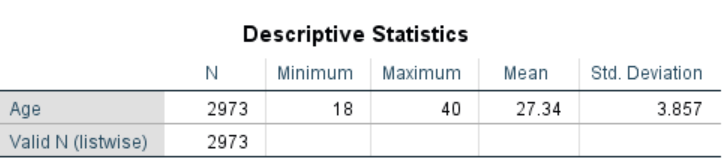
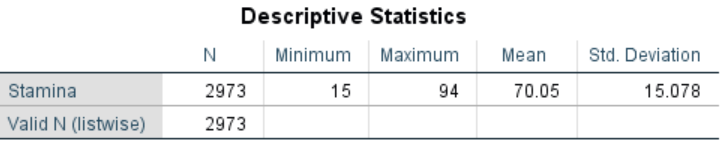
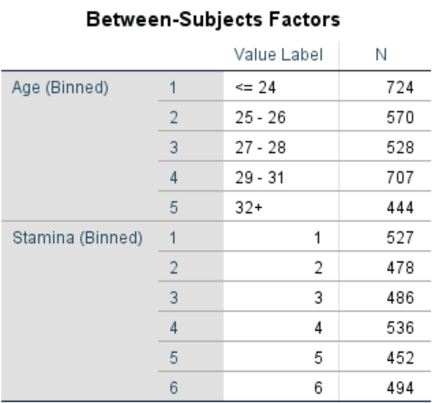
This study investigated the data from N = 2,972 players who belong to one of the four classified groups: <= 24, 25 - 26, 27 – 28, 29 – 31 and 34+. Similarly, Stamina is coded from 1 to 6 based on the reported stamina scores. Variables in the linear regression are all continuous variables. The UEFA Champions League is an annual club football competition organized by the Union of European Football Associations and contested by top-division European clubs, deciding the best team in Europe. Dataset was collected from UEFA Champions League Soccer competition players randomly. Population size is 8000 and the sample size is approximately 3000.

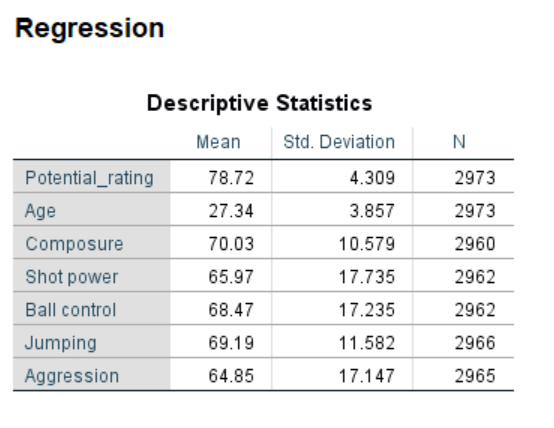
***Measures:***

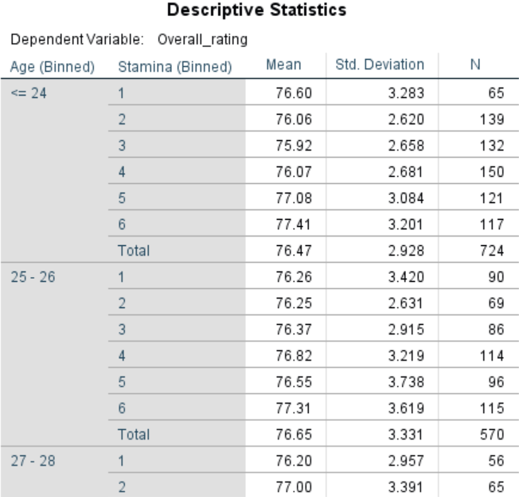
Dependent variables: Overall rating and potential rating. Predictors: Players attributes were gathered which included Age, Stamina, Balance, Ball Control, Composure, Crossing, Curve, Dribbling, Finishing, Free Kick Accuracy, Goalkeeper Diving, Handling, Kicking, Positioning, Reflexes et.al. Descriptive and inferential statistics for the diagnostics was used in the analysis process.

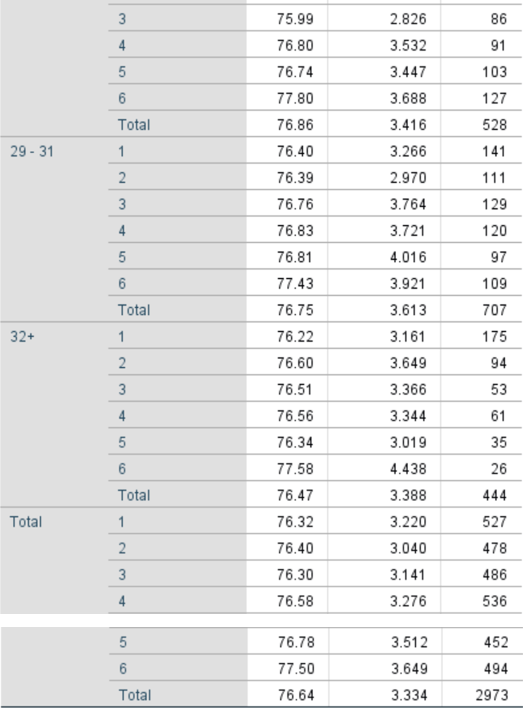
***iii)Results and Analysis:***

**Two-way ANOVA**



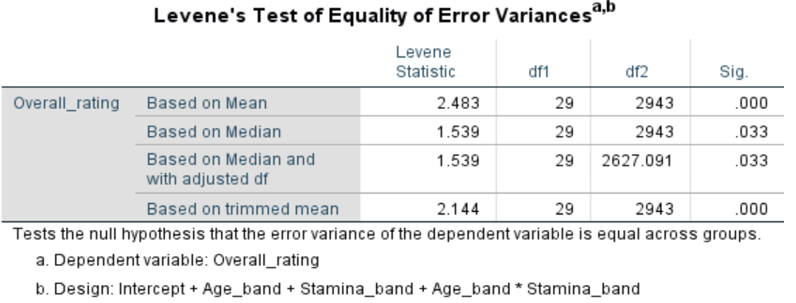




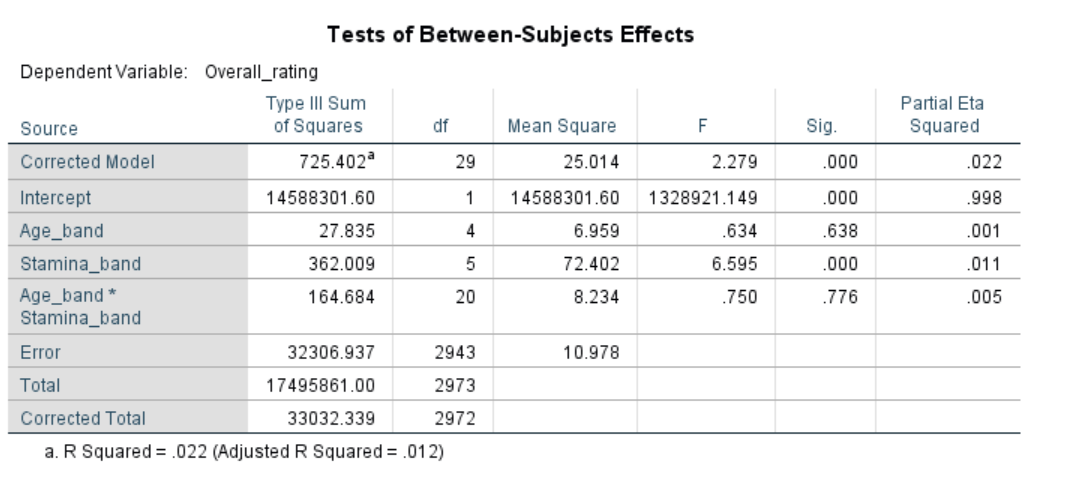


**1. For Two- way ANOVA:**

• Descriptive statistics:Mean scores, standard deviations and N for each subgroup of Age and Stamina can be referred. Age has been classified into five groups: <= 24, 25 - 26, 27 – 28, 29 – 31 and 34+, with mean age being 27.34. Similarly, Stamina is coded from 1 to 6 based on the reported stamina scores ranging from 15 to 94 with mean of 70.05.



• Levene’s test: It provides a test of one of the assumptions underlying analysis of variance. A significant result (Sig. value less than .05) has been reported across all given statistics. It suggests that the variance of overall rating across different groups of age and stamina is not equal.

• Tests of Between-Subjects Effects: The main output from two-way ANOVA is the table labelled below:

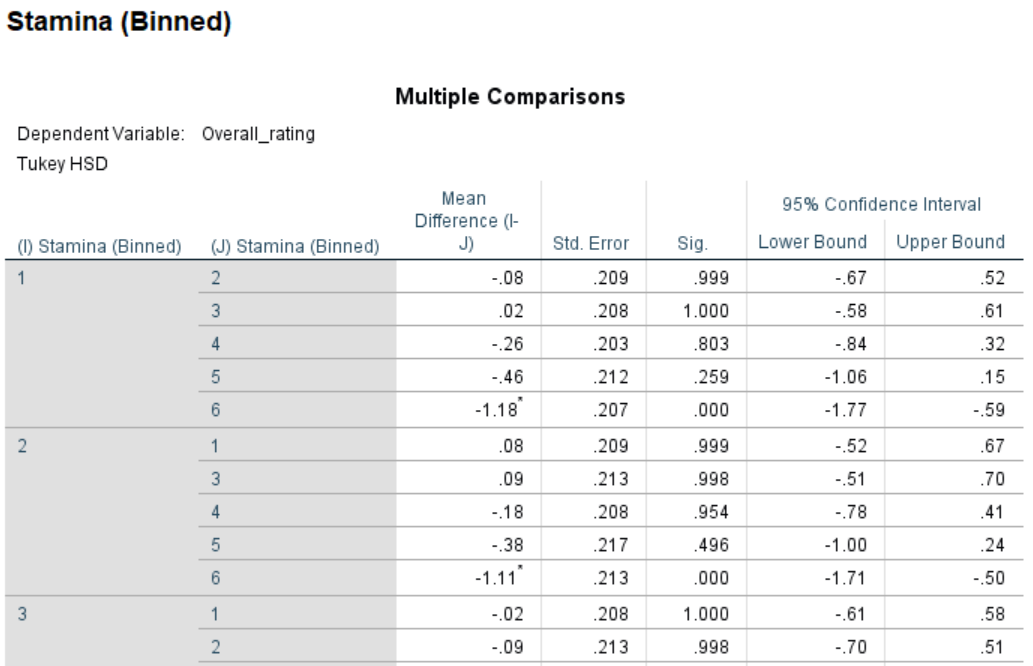
• Interaction effects: The influence of stamina on overall rating (depends on which age group players belong) and is checked for the possibility of an interaction effect. If we find a significant interaction effect, we cannot easily and simply interpret the main effects. This is because, in order to describe the influence of one of the independent variables, we need to specify the level of the other independent variable. The interaction effect (Age\_band \* Stamina\_band) is not significant (.776). This indicates that there is no significant difference in the effect of stamina on overall ratings for the given age groups.

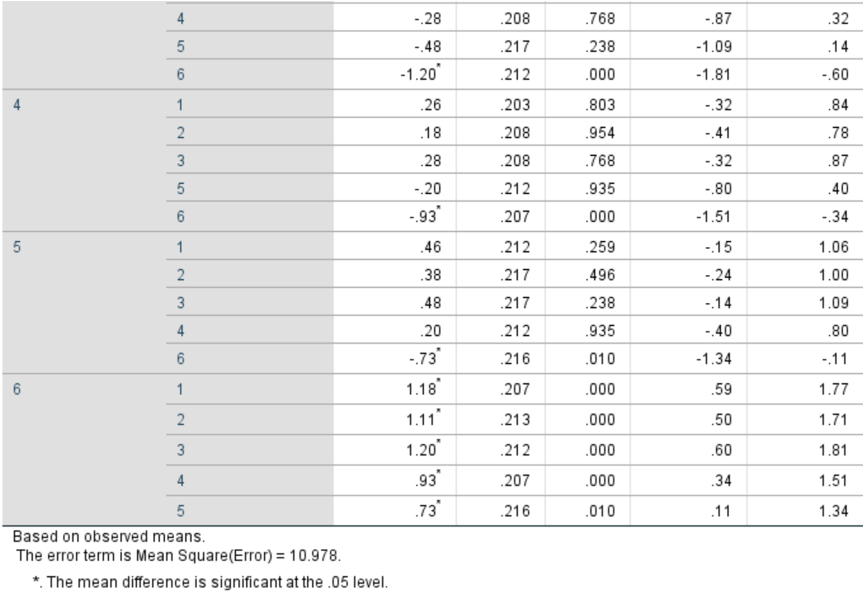
• Main effects: We did not have a significant interaction effect; therefore, we can interpret the main effects. If the value is less than or equal to .05, then there is a significant main effect for that independent variable. We can report a significant main effect for stamina group (stamina: sig=.000), but no significant main effect for age group (age: sig=.638). This means that players across several age groups do not differ in terms of their overall ratings, but there is a difference in scores for players with different stamina groups.

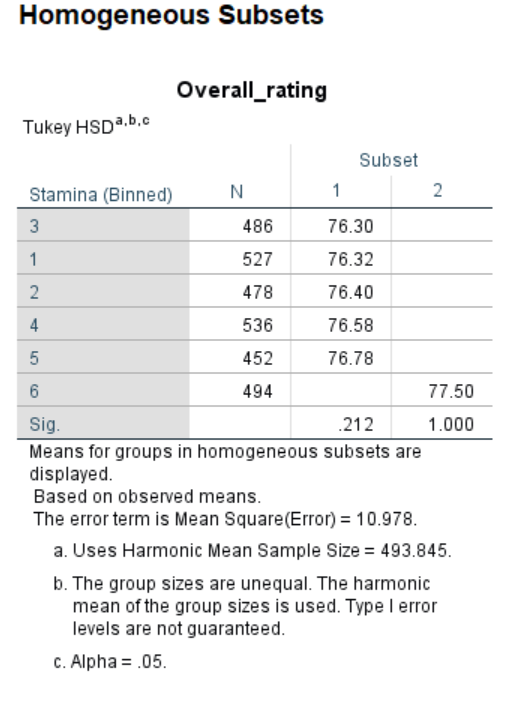
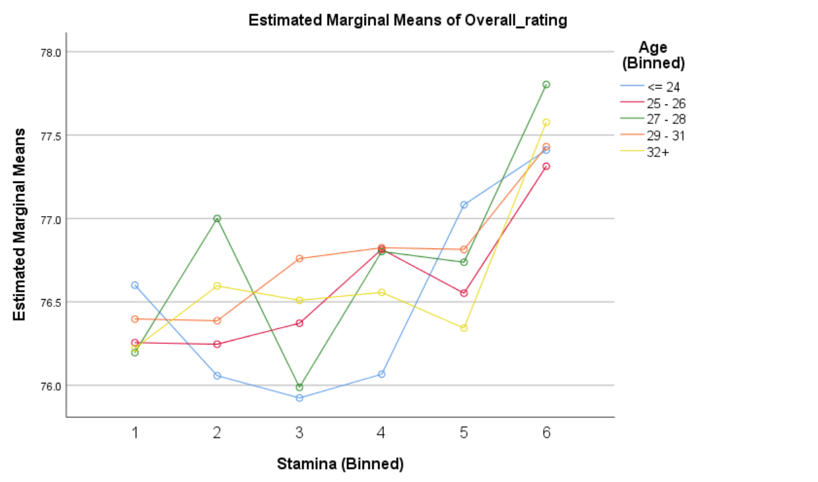
• Effect size: The effect size for the stamina is provided in the column labelled Partial Eta Squared(.011). Using Cohen’s (1988) criterion, this can be classified as small. So, although this effect reaches statistical significance, the actual difference in the mean values is very small.The difference between the groups appears to be of little practical significance.

• Post-hoc tests: Although we know that our stamina groups differ, we do not know where these differences occur: Is stamina class 1 different from class 2, is class 2 different from class 3 and so on. To investigate these questions, we need to conduct post-hoc tests. Post-hoc tests are relevant only if we have more than two levels (groups) to our independent variable. These tests systematically compare each of pairs of groups and indicate whether there is a significant difference in the means of each. The main effect for stamina and the interaction effect did not reach statistical significance. Stamina group 6 (*M*=77.41, *SD*=3.201) was significantly different from the rest of the groups ( Group1: *M*=76.60, *SD*=3.283), ( Group2: *M*=76.06, *SD*=2.620), ( Group3: *M*=7.92, *SD*=2.658), ( Group4: *M*=76.07, *SD*=2.681) and ( Group5: *77.08*, *SD*=3.084). However, the rest of the stamina groups did not differ significantly from either of the other groups.

• Multiple comparisons: The results of the post-hoc tests are provided in the table labelled Multiple Comparisons. We have requested the Tukey Honestly Significant Difference test. Stamina group 6 is significantly different from the rest of the stamina groups, i.e., stamina group 6 and 1, stamina group 6 and 2, stamina group 6 and 3, stamina group 6 and 4, stamina group 6 and 5, differ significantly from one another.





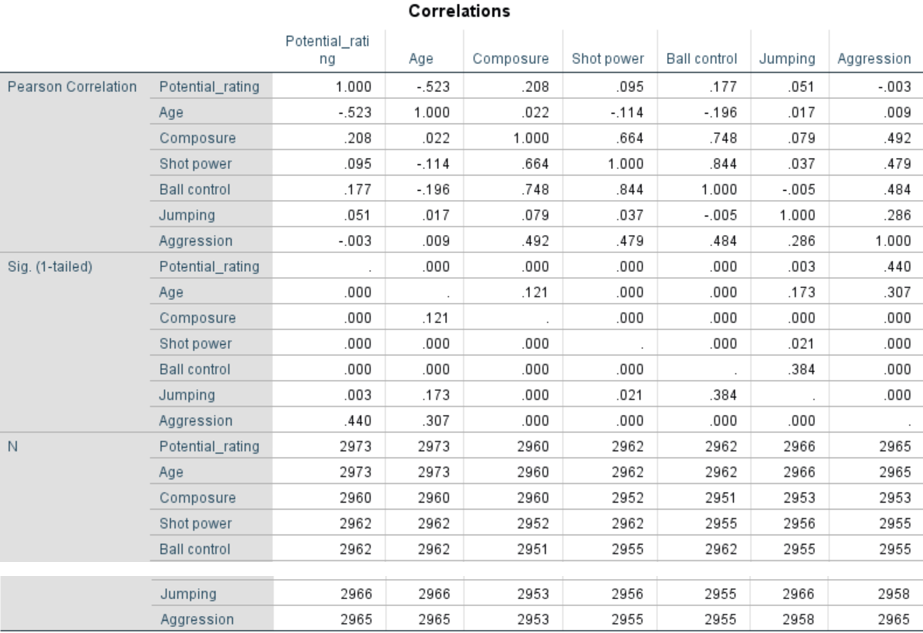
• Plots:By interpreting the output of a plot of the overall ratings, for stamina groups, across the given age groups, there appears to be small difference in stamina for the given age groups.

• **Results:**  Players were divided into 5 groups according to their age (Group 1: <=24, Group 2: 25-26, Group 3: 27-28, Group 4: 29-31, Group 5: 32+) and stamina levels were classified into 6 groups. Stamina classes were grouped from 1 to 6 based on the stamina scores which ranged from 15 to 94 with mean of 70.05. A significant main effect for stamina group (stamina: sig=.000) was found, but no significant main effect for age group (age: sig=.638) was reported. Post-hoc comparisons using the Tukey HSD test indicated the Stamina group 6 (*M*=77.41, *SD*=3.201) was significantly different from the rest of the groups ( Group1: *M*=76.60, *SD*=3.283), ( Group2: *M*=76.06, *SD*=2.620), (Group3: *M*=7.92, *SD*=2.658), ( Group4: *M*=76.07, *SD*=2.681) and ( Group5: *77.08*, *SD*=3.084). The rest of the stamina groups did not differ significantly with each other. Partial Eta Squaredof .011 was reported and classified as small using Cohen’s (1988) criterion.

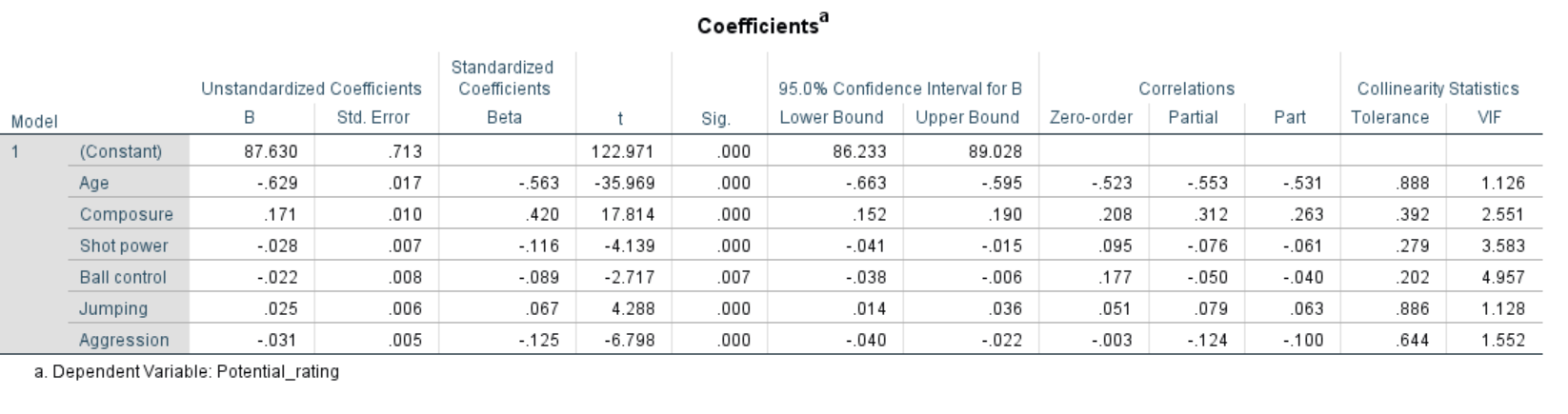
**2. For Standard Multiple Linear Regression:**

Checking assumptions of linear regression provided us the following information:

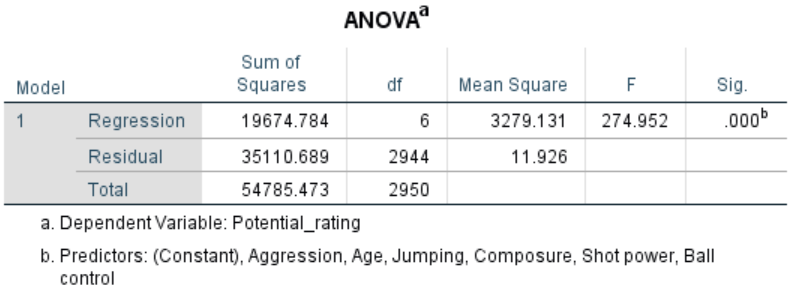
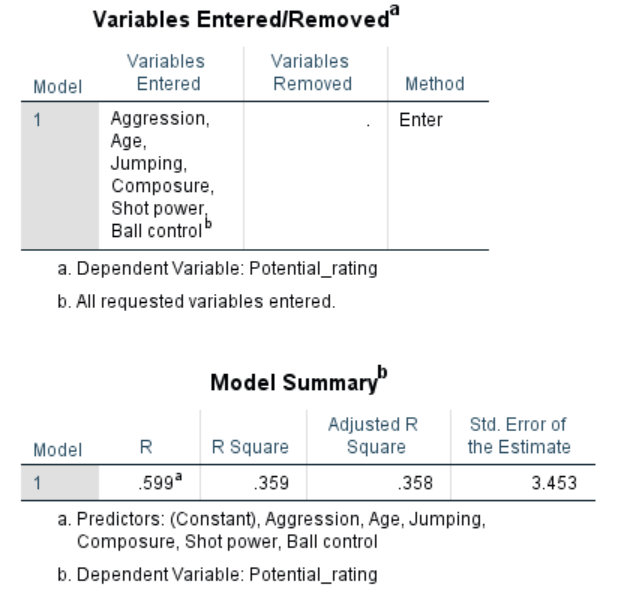
• Multicollinearity: The correlations between potential rating and age, composure, shot power, ball control, jumping, aggression in the given model are -.523, .208, .095, .177, .051, -0.03. Though the remaining variables have little correlation with the dependent variable (potential rating), age indicated high positive correlation with potential rating. Since the dependent variables have bivariate correlations less than .7, we retained all the independent variables. In this case, the independent variables are not multicollinear.



• Tolerance and VIF: Tolerance for age, composure, shot power, ball control, jumping, aggression are .888, .392, .279, .202, .886, .644 which are higher than the tolerance of .20. In addition, VIF are below 10 for all given independent variables. Hence there is no indication of possibility of multicollinearity in the given model.

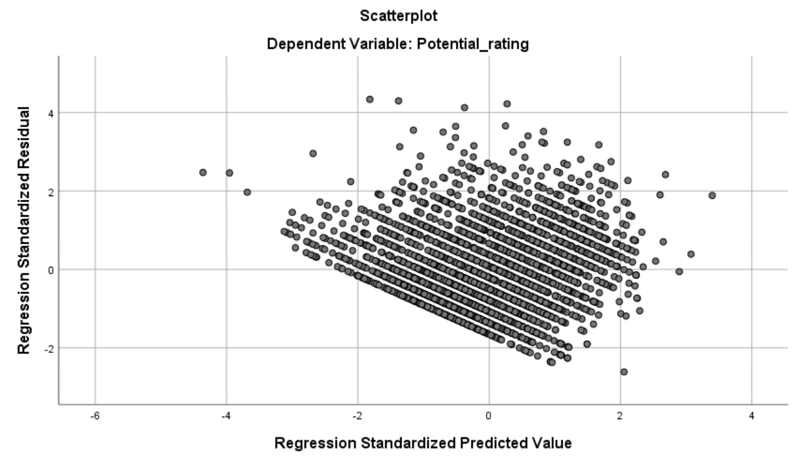


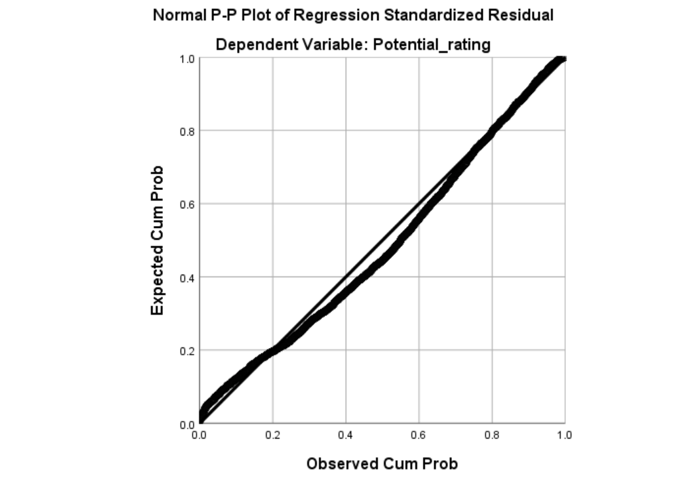
• Statistical Significance of the model: From the ANOVA table, we can infer that the model reaches statistical significance (Sig = .000, which means p<.0005).



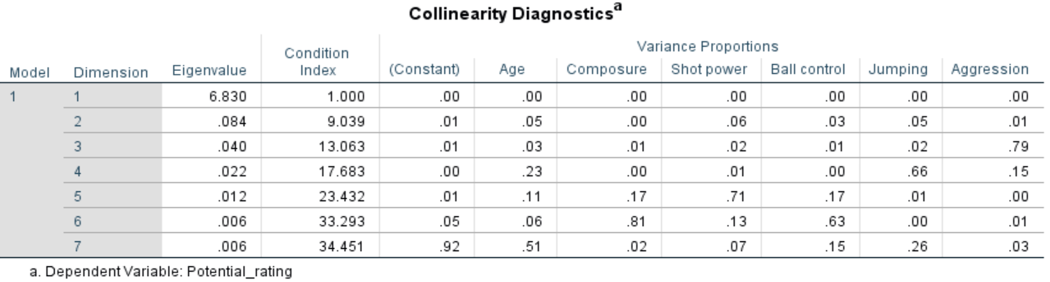
• Evaluating the model: By inferring Model Summarybox, our model (which includes age, composure, shot power, ball control, jumping, aggression) explains 35.8 percent (based on adjusted R square) of the variance in potential rating. This is a quite moderate result.

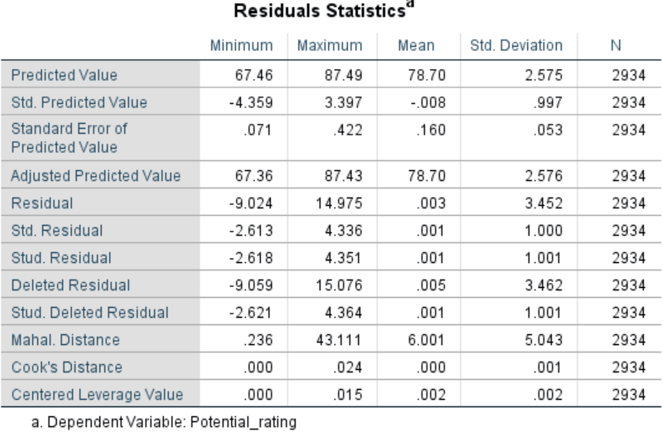
• Outliers, Normality, Linearity, Homoscedasticity, Independence of Residuals: In the Normal Probability Plot, we see the points lie in a reasonably straight diagonal line from bottom left to top right. This suggests no major deviations from normality. In the Scatterplot, the standardized residuals are roughly rectangularly distributed, with most of the scores concentrated in the center (along the 0 point). No significant outliers are seen in the scatterplot.

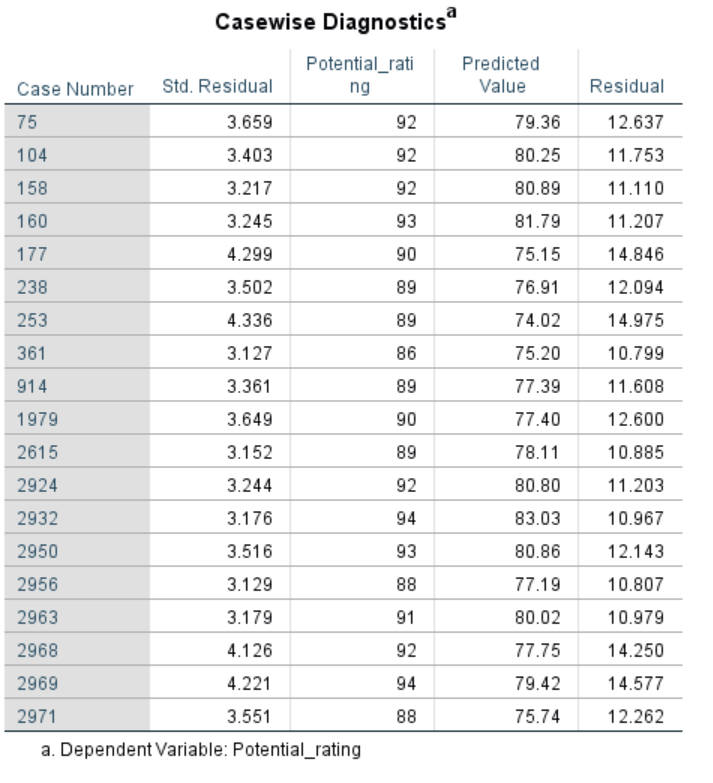




• Evaluating each of the independent variables: In the given model, the largest beta coefficient is –.563, which is for Age. Hence, Age makes the strongest unique contribution to explaining the potential rating when the variance explained by all other variables in the model is controlled for. The Beta value for composure was slightly lower (.420), indicating that it made less of a contribution. Similarly, shot power, ball control, jumping, and aggression had Beta values of -.116, -.089, .067 and -.125 respectively.

The significance values for all the independent variables in the model are less than .05 (.000 and .007), hence all these variables are making a significant unique contribution to the prediction of potential rating. Age, composure, shot power, ball control, jumping, aggression made a unique, and statistically significant contribution to the prediction of potential rating. 





**Results:** Standard multiple regression was performed. Dependent variable for the regression model is Standardized (beta) for Age, Composure, Shot-power, Ball control, Jumping and Aggression are -.563,.420, -.116,.0.89,.067 and -.125 respectively. and Unstandardized (B) coefficients for Age, Composure, Shot-power, Ball control, Jumping and Aggression are -.629, .171, -.028, -.022, .025 and -.031 respectively. Standard errors (SE) are .017, .010, .007, .008, .006 and .005 respectively.

***iv) Discussion:***

Compared to recent studies investigating the relevance of talent predictors in soccer, this study assessed an optimum sample over a single period consisting of most of the ages relevant in soccer players and an impact of stamina on ratings in their career. Dealing with a bivariate approach for overall ratings, this work also addressed predictors of potential ratings. Predictors in the regression model are multidimensional, for e.g., psychomotor (ball control) and physical (jumping) predictors. Besides the well-established procedures analyzing manifest variables based on the general linear model (e.g., ANOVAs), this study also used standard linear regression model as a further multivariate approach allowing for the consideration of latent performance factors as well as their interdependent relations. Stamina levels discriminated between the overall ratings within the study sample whereas age did not,

at least in the relevant soccer league, which is UEFA in the period of 2018.

Objectives: For interpretation of therelevance, one must consider that design features strongly

affect the estimates of predictors' relevance. The duration of the study period and its relative phase play an important role, because discriminating factors among players vary between the different periods of time. In addition, the sample size (determining the statistical test power) and the performance levels at which the predictors and criteria were assessed leading to different ranges of ratings are further potential moderators of the effect size. Because of the strong influence of design features, the current results are best comparable with findings from Vattier. That study also analyzed the relevance of age and ratings analysis. The present results confirm the former results of an impact of age on overall ratings in soccer. Multiple variables were assessed to determine the predictive value of age and other factors in soccer. Regarding the relevance ofrelated predictors, this study revealed slightly moderate sizes for potential ratings compared to other prospective studies dealing with players at the given age. Still, it remains unanswered whether differences in the variables are caused by related biases in the selection (e.g., players with more stamina are more often selected) or whether these variables have to be considered as performance factors prior to selection leading to no impact on overall and potential ratings.

However, despite the explained variance, the quite relevant sizes derived by the multiple group comparisons are a hint that the tests obtained prognostic relevance. The results of the analysis of variance along with post hoc tests confirm these general findings of the traditional ratings approaches, but also go beyond these findings with regard to the overall explained variance, role of the covariates and consideration of other underlying factors and attributes of the measurement model.

The linear regression model led to a considerably higher magnitude of explained variance (*R2* = 35.8%). To my knowledge, no study in soccer exists using the variables approach for the estimation of age, composure, shot power, ball control, jumping, aggression for the relevance of potential rating. Thus, the results for linear regression cannot be compared to other studies. It should be noted that the attributes of the measurement model is influenced by low levels of correlation between the predictors and potential ratings. Different models were tried to find out variables that would have higher correlations with potential ratings. Besides age, all the rest of the variables had correlation coefficient less than .3. However, these variables had significant p-values and effective R2 values. Hence, they were incorporated appropriately in the regression model. Further regression models can be generated and tested by controlling for several independent variables and introducing intervening variables. Post- treatment and pre-treatment approaches can be tried to further determine the validity, reliability and accuracy of the statistical results generated. Nevertheless, main reasons for the increase in explained variance in the latent variable model could be that within the regression model, regression coefficients are unbiased and error free. According to the results for analysis of variance in age and stamina, the relevance of the two factors was quite similar. However,

the insights into the impact of potential ratings characteristics are limited in this study, as its focus laid on overall ratings. For a more concrete clarification of the impact of age and stamina, repeated measurements and a theoretical model including core assumptions of ratings in soccer would be necessary.

Therefore, the difference in the effects should be interpreted cautiously.

***v. Conclusion***

The study provides reliable empirical knowledge on the relevance related soccer ratings.

The results demonstrated validity for 2018 player ratings for the associated examination of variances accounted by players demonstrating different classes of stamina levels. Apart from this, a simple linear regression model explained variances in potential ratings caused due to several physical or psychomotor factors in soccer players. Together with emerging reports studies, these results raise concern towards the sensitivity of tests and the levels to apply them as exclusive and reliable information for selection decisions. Thus, identifying talent in soccer remains a very complex (practical and theoretical) problem, and multidimensional approaches are recommended as they may prove more successful in talent identification by having a look at the overall ratings and potential ratings of soccer players. This study applied analysis of variance and determined the significant impact of age as well as the underlying performance factors, providing further insights for talent identification. The regression model has a greater predictive power despite portraying low correlations with the dependent variable. A promising prospect and likewise a huge challenge for future talent research lies in the application and integration of different multivariate approaches in order to go

beyond the ªtraditional predictive validity and to

be able to work on more theoretically founded questions.

**Acknowledgments**

This study is part of the research project in the Applied Data Analysis class, Koppelman School of Business, Brooklyn College. I would like to thank Dr. William Hampton Sosa for the approval of project proposal, data and productive discussions in several meetings.

**References**

1**.**Hughes M. Reviews: Talent identification in soccer: The role of maturity status on physical, physiological and technical characteristics. International Journal of Sports Science &

Coaching. 2010;

2.Rowe P. Talent in sports. Some reflections about the search for future champions.

Movement and Sport Sciences. 2015;

3.The Premier League. Elite performance. 2011. https://www.premierleague.com/youth/eliteperformance.

4. Gilbert C. Using physiological data to predict future career progression in 14- to 17-year-old

Austrian soccer academy players. Journal of Sports Science.