The University of Texas at Arlington

Statistical Impact:

How Individual Stats Shape Counter Strike Major League Ratings

Joshua Catalan, Jair Rea, Youssef Aitbenzanzoun

CSE 4334-001

Rozsa Zaruba

6 December 2023

Abstract

Our analysis dives into the dynamics of professional CSGO player performance. Looking at the top 10 teams and player rankings, we've observed a disparity between individual player rankings and team rankings. Using a dataset from Kaggle, we applied multiple regression analysis to study the player stats that significantly impact their ratings. Surprisingly, the statistic with the highest effect is not related to kills per death or the overall kill count; the key factor is the damage a player inflicts per round. However, our findings underline the need to consider factors unrelated to statistics, like teamwork and strategy, and pushing for the removal of a one-dimensional rating which is solely dictated by statistics.

Introduction

Counter Strike is a First Person Shooter that is at the peak of world popularity throughout its lifetime. Within counter strike there are mass viewed competitions that are broadcasted throughout the world, gaining concurrent viewers between 100 thousand to 800 thousand per major match on average, reaching a peak of 3 million viewers in the last major. Adults and children around the world chant the names of their favorite teams and players while also imitating their playstyles and outfits. Being a part of the top 10 teams or players is a highly sought after position in terms of self-marketing, salary, and overall popularity worldwide. Our goal is to analyze the CSGO dataset we selected on Kaggle to conclude which player statistic affects their rating with the most significance. The reason is related to the player ranking worldwide being vastly different from the top teams in the world. When analyzing the data, we find that at most 4 to 5 players are from the top-rated teams, hence why we wanted to figure out why.

Problem Statement

Given the general overview we just discussed, a question comes to mind. Why are players from the championship teams excluded at times from the top player rankings? To answer the question, we must consider all the statistics that are recorded during a match. The data contains professional player statistics from the start of the game until the most recent major, giving us a healthy collection of samples to analyze. Going through the process will allow us to specify which statistic affects “rating” with the highest significance.

Proposed Methodology

The original approach consisted of three possible tests: a multiple regression test with a dependent variable as rating (continuous) and the rest as independent variable, the second test was a multiple regression analysis of team win percentage after the first kill and how it affects the rating, and lastly an ANOVA to find significant differences in ratings among players from different countries. We decided to go with the multiple regression test that covers all statistics and how it affects the player’s rating. The reasoning stemmed from the freedom we would have with analyzing each variable and how they each affect the rating, giving us a strong overview on all variables instead of focusing only on either categorical / continuous / standalone variable.

Analysis and Results

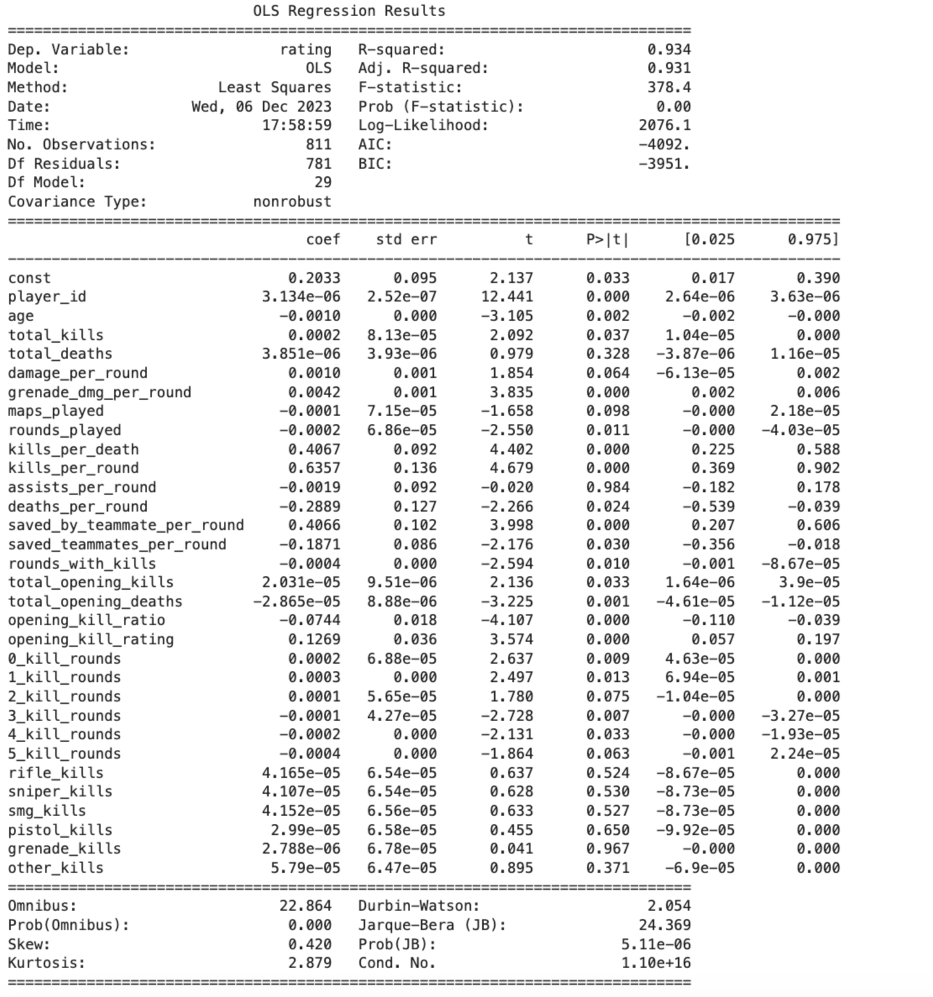
Exploratory Data Analysis (EDA):

To start off we did an EDA on the CSGO dataset, getting the general overview of the structure of the data. Our goal was to understand the factors influencing player ratings, considering in-game statistics, team affiliations, and player performance. Initial data cleaning involved removing irrelevant categorical information like player teams, names, previous affiliations, and countries. We wanted a focused analysis on variables that actually mattered in determining player ratings, with the 'rating' as our dependent variable and various player statistics as independent variables.

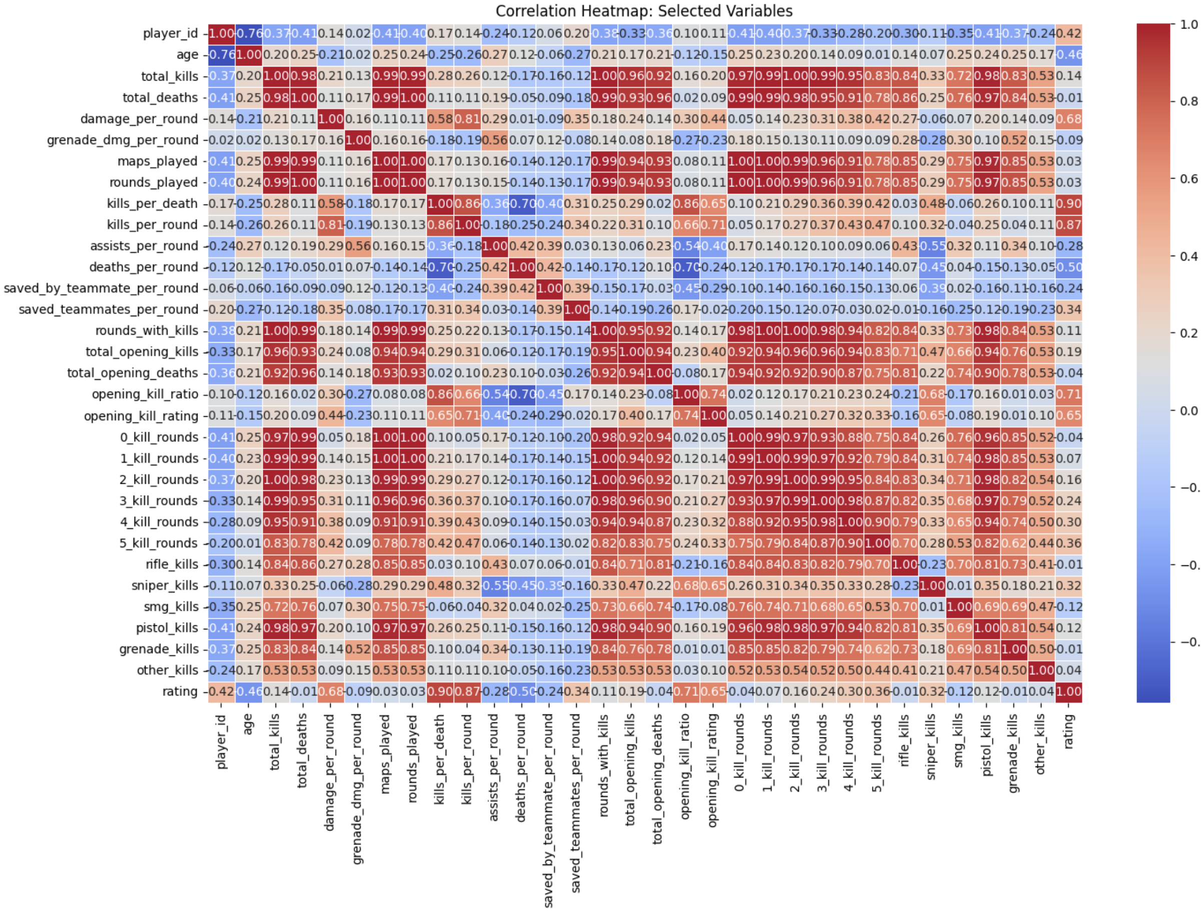
Multiple Regression Analysis:

To understand how each variable contributed towards player ratings, we decided to go with a multiple regression analysis. This allowed us to explore the impact of different factors in detail to determine each variable significance in the relation. We also set our Null Hypothesis (H_0) as all coefficients of the independent variables are equal to zero and the Alternative Hypothesis (H_1) as at least one coefficient is not equal to zero. The initial analysis showed a strong R-squared value of 93.4%, indicating that our chosen variables could explain a significant portion of the variability in player ratings. Our general assumption at the start that 'player_kills_per_death' would positively affect the rating was confirmed by the statistical significance of the coefficients and a low p-value in the Ordinary Least Squares (OLS) regression results. Unfortunately, there eigen value of the first model is 1.18e-20, indicating multicollinearity. To gain better results we utilized the VIF to create an improved model, giving us a model with an R-squared value of 0.747 and little to no multicollinearity. The methods we used for reducing multicollinearity and analyzing the VIF are explored in the future sections.

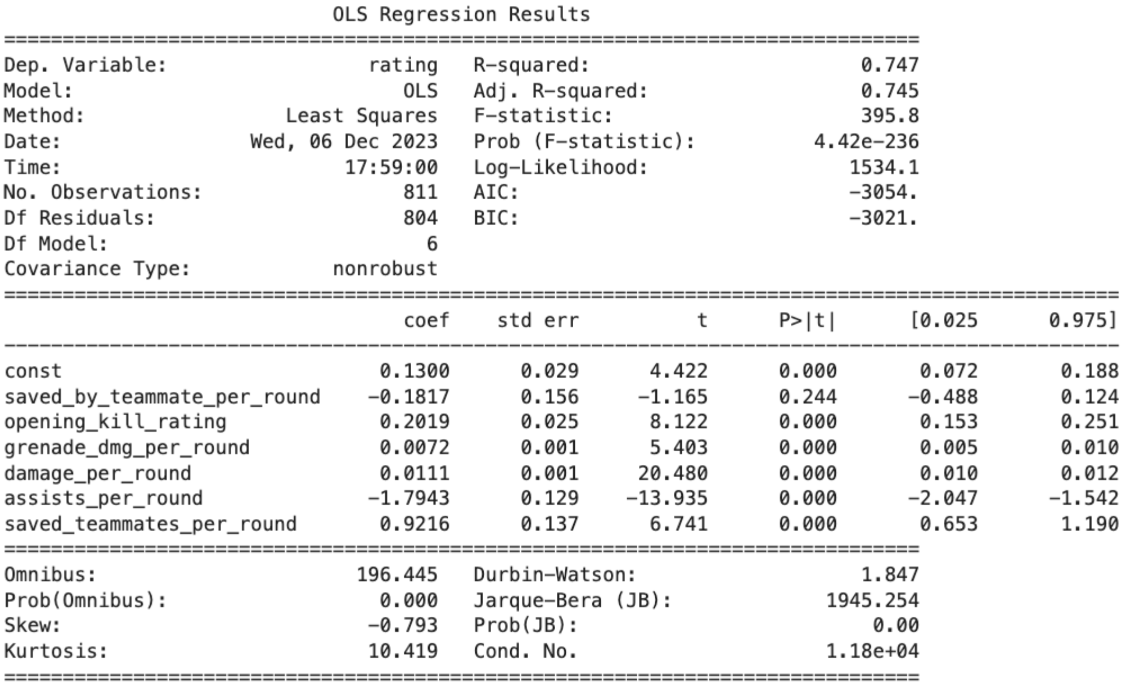
Before:



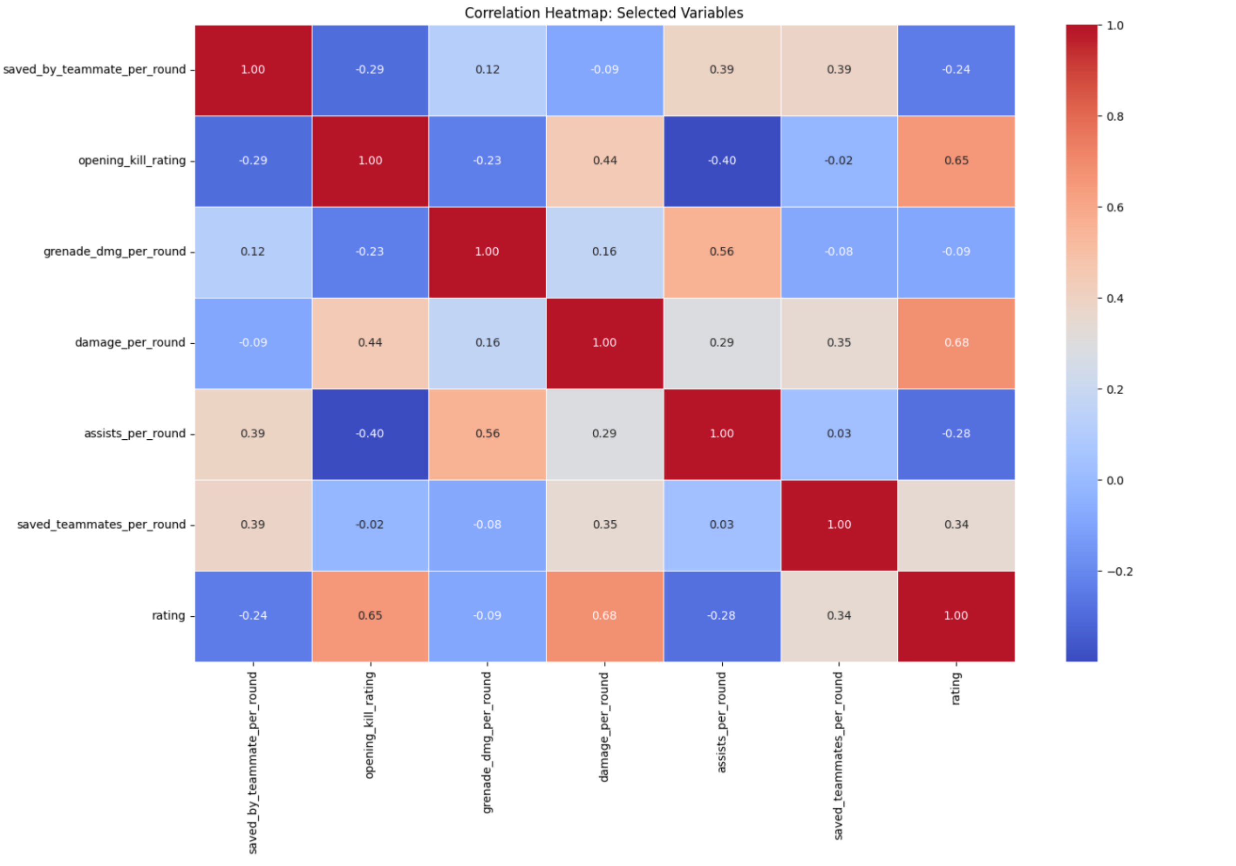
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rating   R-squared:                       0.934
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                     378.4
Date:                Wed, 06 Dec 2023   Prob (F-statistic):               0.00
Time:                        17:58:59   Log-Likelihood:                 2076.1
No. Observations:                 811   AIC:                            -4092.
Df Residuals:                     781   BIC:                            -3951.
Df Model:                          29
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                           0.2033      0.095      2.137      0.033       0.017       0.390
player_id                    3.134e-06   2.52e-07     12.441      0.000    2.64e-06    3.63e-06
age                            -0.0010      0.000     -3.105      0.002      -0.002      -0.000
total_kills                     0.0002   8.13e-05      2.092      0.037    1.04e-05       0.000
total_deaths                 3.851e-06   3.93e-06      0.979      0.328   -3.87e-06    1.16e-05
damage_per_round                0.0010      0.001      1.854      0.064   -6.13e-05       0.002
grenade_dmg_per_round           0.0042      0.001      3.835      0.000       0.002       0.006
maps_played                    -0.0001   7.15e-05     -1.658      0.098      -0.000    2.18e-05
rounds_played                  -0.0002   6.86e-05     -2.550      0.011      -0.000   -4.03e-05
kills_per_death                 0.4067      0.092      4.402      0.000       0.225       0.588
kills_per_round                 0.6357      0.136      4.679      0.000       0.369       0.902
assists_per_round              -0.0019      0.092     -0.020      0.984      -0.182       0.178
deaths_per_round               -0.2889      0.127     -2.266      0.024      -0.539      -0.039
saved_by_teammate_per_round     0.4066      0.102      3.998      0.000       0.207       0.606
saved_teammates_per_round      -0.1871      0.086     -2.176      0.030      -0.356      -0.018
rounds_with_kills              -0.0004      0.000     -2.594      0.010      -0.001   -8.67e-05
total_opening_kills          2.031e-05   9.51e-06      2.136      0.033    1.64e-06     3.9e-05
total_opening_deaths        -2.865e-05   8.88e-06     -3.225      0.001   -4.61e-05   -1.12e-05
opening_kill_ratio             -0.0744      0.018     -4.107      0.000      -0.110      -0.039
opening_kill_rating             0.1269      0.036      3.574      0.000       0.057       0.197
0_kill_rounds                   0.0002   6.88e-05      2.637      0.009    4.63e-05       0.000
1_kill_rounds                   0.0003      0.000      2.497      0.013    6.94e-05       0.001
2_kill_rounds                   0.0001   5.65e-05      1.780      0.075   -1.04e-05       0.000
3_kill_rounds                  -0.0001   4.27e-05     -2.728      0.007      -0.000   -3.27e-05
4_kill_rounds                  -0.0002      0.000     -2.131      0.033      -0.000   -1.93e-05
5_kill_rounds                  -0.0004      0.000     -1.864      0.063      -0.001    2.24e-05
rifle_kills                  4.165e-05   6.54e-05      0.637      0.524   -8.67e-05       0.000
sniper_kills                 4.107e-05   6.54e-05      0.628      0.530   -8.73e-05       0.000
smg_kills                    4.152e-05   6.56e-05      0.633      0.527   -8.73e-05       0.000
pistol_kills                  2.99e-05   6.58e-05      0.455      0.650   -9.92e-05       0.000
grenade_kills                2.788e-06   6.78e-05      0.041      0.967      -0.000       0.000
other_kills                   5.79e-05   6.47e-05      0.895      0.371    -6.9e-05       0.000
==============================================================================
Omnibus:                       22.864   Durbin-Watson:                   2.054
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               24.369
Skew:                           0.420   Prob(JB):                     5.11e-06
Kurtosis:                       2.879   Cond. No.                     1.10e+16
==============================================================================
```



After:

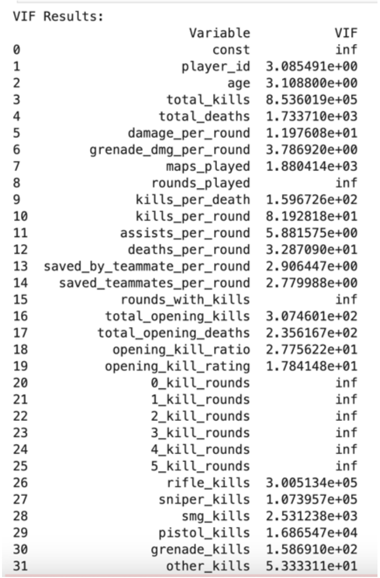


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rating   R-squared:                       0.747
Model:                            OLS   Adj. R-squared:                  0.745
Method:                 Least Squares   F-statistic:                     395.8
Date:                Wed, 06 Dec 2023   Prob (F-statistic):         4.42e-236
Time:                        17:59:00   Log-Likelihood:                 1534.1
No. Observations:                 811   AIC:                            -3054.
Df Residuals:                     804   BIC:                            -3021.
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                          0.1300      0.029      4.422      0.000       0.072       0.188
saved_by_teammate_per_round   -0.1817      0.156     -1.165      0.244      -0.488       0.124
opening_kill_rating            0.2019      0.025      8.122      0.000       0.153       0.251
grenade_dmg_per_round          0.0072      0.001      5.403      0.000       0.005       0.010
damage_per_round               0.0111      0.001     20.480      0.000       0.010       0.012
assists_per_round             -1.7943      0.129    -13.935      0.000      -2.047      -1.542
saved_teammates_per_round      0.9216      0.137      6.741      0.000       0.653       1.190
==============================================================================
Omnibus:                      196.445   Durbin-Watson:                   1.847
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1945.254
Skew:                          -0.793   Prob(JB):                         0.00
Kurtosis:                      10.419   Cond. No.                     1.18e+04
==============================================================================
```



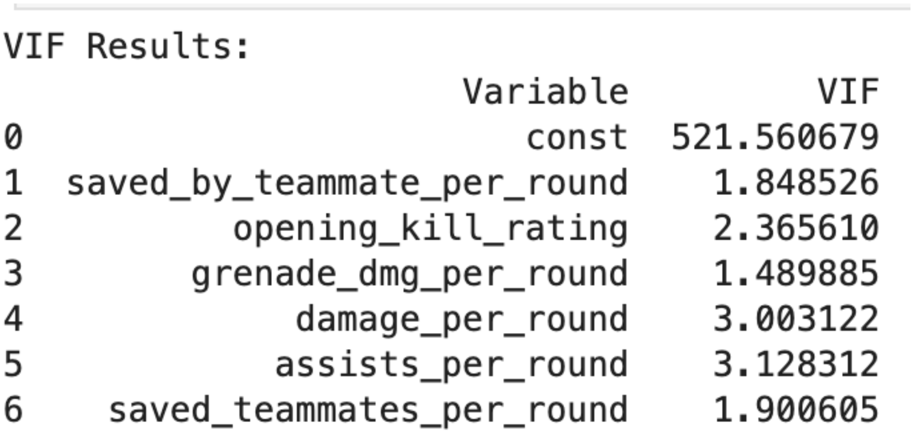
By looking at the OLS Regression results we find a low P-value allowing us to reject the Null hypothesis (the F-statistic), as for the variables that we selected displayed in the OLS regression results we can see that all the variables are significant except for saved_by_teammate_per_round which is greater than 0.05 at 0.244. To solidify the variables further we conducted a VIF test which we limited to less than a VIF value of 10, we then removed variables that exceeded 10 to achieve acceptable levels of multicollinearity. To study the multicollinearity of our multiple regression analysis, we utilized the eigenvalues from the OLS regression results. We then cross verified the assessment by calculating the VIF output through trial and error to achieve a more acceptable multicollinearity level for our variables. Through various attempts we were finally able to achieve between minimal to no multicollinearity cases.

Before:



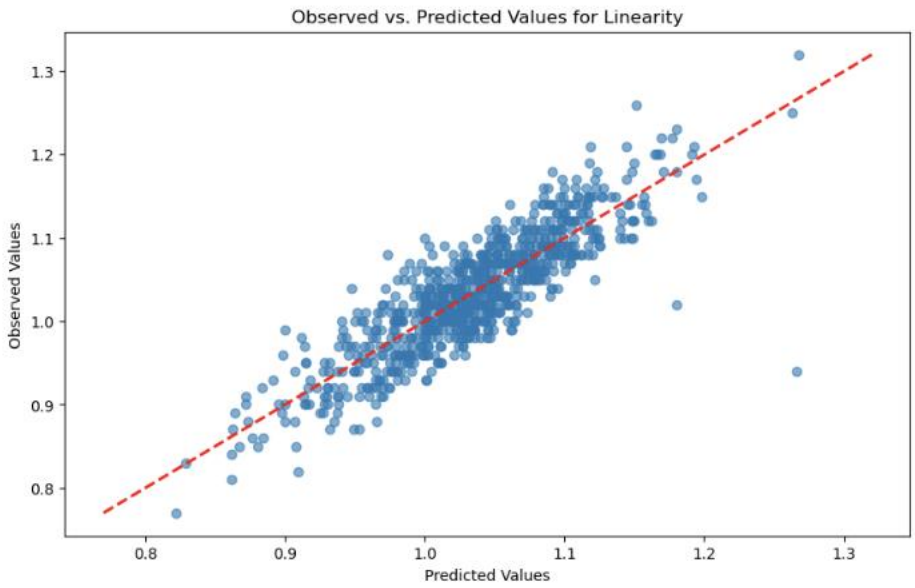
```
VIF Results:
                        Variable           VIF
0                          const           inf
1                      player_id  3.085491e+00
2                            age  3.108800e+00
3                    total_kills  8.536019e+05
4                   total_deaths  1.733710e+03
5               damage_per_round  1.197608e+01
6           grenade_dmg_per_round  3.786920e+00
7                    maps_played  1.880414e+03
8                  rounds_played           inf
9                 kills_per_death  1.596726e+02
10                kills_per_round  8.192818e+01
11              assists_per_round  5.881575e+00
12               deaths_per_round  3.287090e+01
13   saved_by_teammate_per_round  2.906447e+00
14     saved_teammates_per_round  2.779988e+00
15              rounds_with_kills           inf
16           total_opening_kills  3.074601e+02
17          total_opening_deaths  2.356167e+02
18            opening_kill_ratio  2.775622e+01
19           opening_kill_rating  1.784148e+01
20                 0_kill_rounds           inf
21                 1_kill_rounds           inf
22                 2_kill_rounds           inf
23                 3_kill_rounds           inf
24                 4_kill_rounds           inf
25                 5_kill_rounds           inf
26                   rifle_kills  3.005134e+05
27                  sniper_kills  1.073957e+05
28                     smg_kills  2.531238e+03
29                  pistol_kills  1.686547e+04
30                 grenade_kills  1.586910e+02
31                   other_kills  5.333311e+01
```

After:

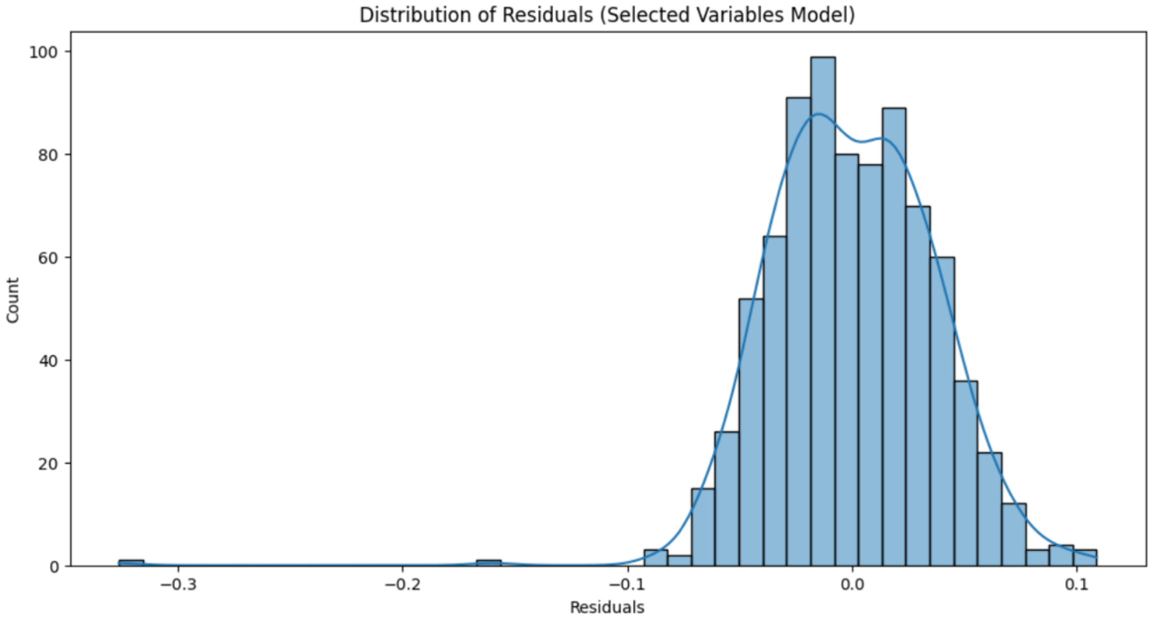


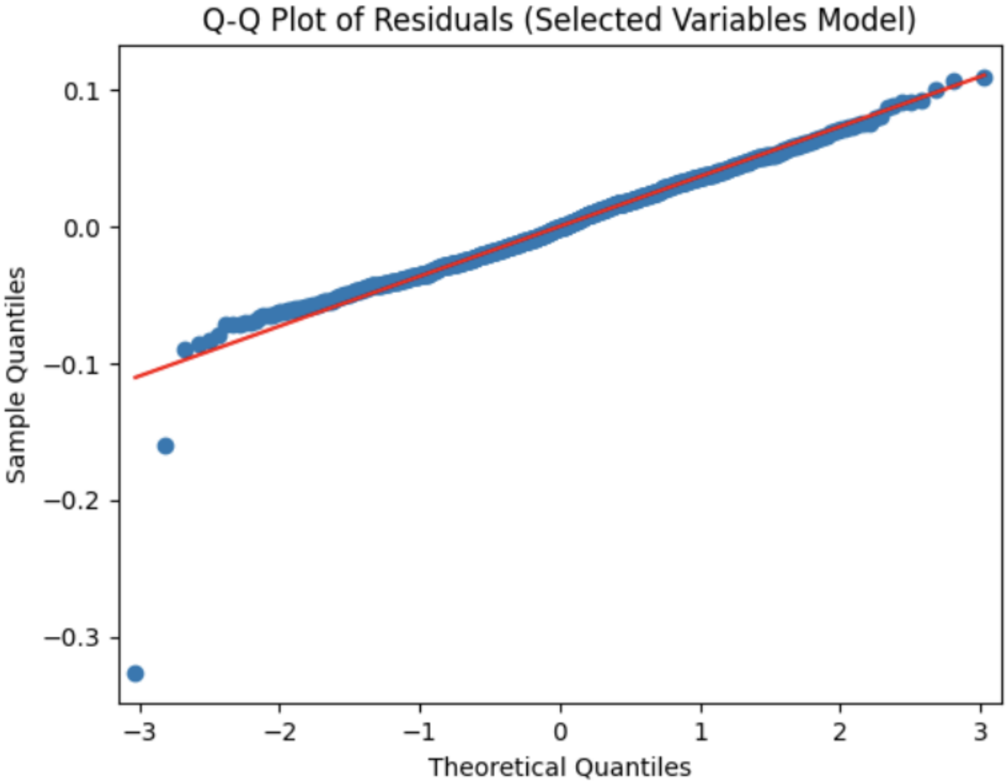
```
VIF Results:
                        Variable         VIF
0                          const  521.560679
1    saved_by_teammate_per_round    1.848526
2            opening_kill_rating    2.365610
3          grenade_dmg_per_round    1.489885
4               damage_per_round    3.003122
5              assists_per_round    3.128312
6      saved_teammates_per_round    1.900605
```

For the linearity assessment on the relationship between the rating and the player statistics, we observed and predicted the values which show a possible positive trend in the relation. As shown, between the observed and predicted values in the figure that will be shown below:



As for evaluating normality we examined whether the distribution of difference between observed and predicted values followed a normal distribution, the figure used was a histogram and a Q-Q plot of the residuals with values that should be symmetrical. As for homoscedasticity the points were consistently spread above and below the mid region with similar variability of the residuals across the whole region, thus further solidifying the presence of homoscedasticity.





After analyzing the results and removing variables that was above the 10 VIF range and outside the significance range, we kept the following variables: opening_kill_rating, grenade_dmg_per_round, damage_per_round, assists_per_round, and saved_teammates_per_round. Unfortunately, our assumption of kills_per_death is not included in this general pool list. Looking at the selected variable we notice a trend that all of them center around Damage_per_round which was honestly to be expected and short sighted from our end. It makes sense since the objective of the point system will most likely be passed as the average damage per round or ADR for short.

Conclusion

In conclusion, the Damage_per_round has the most influence on the player rating in both positive and negative increases (has the highest weight in determining rating). After finalizing the results and discussing between ourselves we found that the results are a bit one dimensional and does not focus on the various aspect of the game that are not shown in statistics. Intelligent players implement various tactics, plans, and communication to attain a victory over the opposing team. Damage is not the end all be all and the board should consider the various strategies that teams implement to become the best in the world. These issues are what cause the vast discrepancy between the top player's world ranking and the team world rankings, not even half of the top 10 players is signed with the top 10 teams in the world.

Another thing to discuss includes the perfect plot fits that we have shown during our analysis. Our assumption for such positive results relates to the number of variable outputs that can be accumulated for each player throughout the round and the absence of NULL and unrelated data. In counter strike each team has a 100 Health with 5 players per team. Let's take for instance that players can invest in Kevlar which will provide 100 extra protections to the mid-section only and purchase a helmet to protect the head while adding 30 more points, bringing the total Health to a value of 230. The highest possible damage ceiling in a round would be between (230 hit points * 10 players) which is 2300 and the minimum would be 0 (in the case of a surrender or a 0 dmg round which is very rare). Unlike soccer or other sports, counter strike has an actual consistent value range for each variable throughout every round, minimizing outliers and making it an easier data set to analyze and predict with. Utilizing the data with proper damage / statistical parameters we were able to come to the conclusions that were discussed above, providing us with clear and consistent results in our normality, linearity, and Homoscedasticity.

Lessons Learned

While diving into Exploratory Data Analysis (EDA) and Multiple Regression Analysis of the CSGO dataset, we picked up some valuable insights. In the EDA phase, we quickly grasped the importance of having clean data and honing in on the right variables. Getting rid of unnecessary categorical info allowed us to laser-focus on factors affecting player ratings. Moving on to Multiple Regression Analysis, we realized how crucial it is to craft precise hypotheses and fine-tune the model. Identifying multicollinearity issues led us to bring in the Variance Inflation

Factor (VIF) to boost model accuracy, underlining the significance of tackling statistical assumptions.

Checking for normality and homoscedasticity gave us a peek into the model's robustness. The whole back-and-forth of variable selection and refinement underscored the need for flexibility in statistical modeling. Also, discovering that certain variables, like damage_per_round, held a pivotal role in player ratings drove home the importance of truly understanding the domain under scrutiny. These experiences don't just refine our analytical skills; they deepen our understanding of the complex factors influencing CSGO player ratings.

Bibliography

Arif, N. (2022, August 2). *CSGO: Professional players dataset*. Kaggle. https://www.kaggle.com/datasets/naumanaarif/csgo-pro-players-dataset

Kaggle. (2022). CS:GO Pro Players Dataset. *Kaggle*. https://www.kaggle.com/datasets/naumanaarif/csgo-pro-players-dataset

Martinos, Nikolas (2021) *Functional and Non-Functional Requirements for a Performance Dashboard in CS:GO.*

Zhang, W., Muric, G., & Ferrara, E. (2022a, May 19). *Individual and collective performance deteriorate in a new team: A case study of cs:go tournaments*. arXiv.org. https://arxiv.org/abs/2205.09693

Huang, W. X., Wang, J., & Xu, Y. (2022, May 24). *Predicting round result in Counter-Strike: Global ... - IEEE xplore*. IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9778597

Regensburg, D. H. U. of, Halbhuber, D., Regensburg, U. of, Regensburg, J. H. U. of, Höpfinger, J., Valentin Schwind Frankfurt University of Applied Sciences, Schwind, V., Sciences, F. U. of A., Regensburg, N. H. U. of, Henze, N., Leuven, K., York, U. of, Bremen, U. of, Maryland, U. of, University, U., & Metrics, O. M. A. (2022, November 1). *A dataset to investigate first-person shooter players: Extended abstracts of the 2022 annual symposium on computer-human interaction in play*. ACM Conferences. https://dl.acm.org/doi/abs/10.1145/3505270.3558331

Dehpanah, A., Ghori, M. F., Gemmell, J., & Mobasher, B. (2022, July 1). *Behavioral player rating in competitive online shooter games*. arXiv.org. https://arxiv.org/abs/2207.00528

University, S. M. T., Marshall, S., University, T., University, P. M.-B. T., Mavromoustakos-Blom, P., University, P. S. T., Spronck, P., & Metrics, O. M. A. (2022, September 1). *Enabling real-time prediction of in-game deaths through telemetry in counter-strike: Global offensive: Proceedings of the 17th International Conference on the foundations of Digital Games*. ACM Other conferences. https://dl.acm.org/doi/abs/10.1145/3555858.3555859

Klagenfurt, M. L. A.-A.-U., Lux, M., Klagenfurt, A.-A.-U., SimulaMet, P. H., Halvorsen, P., SimulaMet, Bergen, D.-T. D.-N. U. of, Dang-Nguyen, D.-T., Bergen, U. of, Laboratory, H. S. S. R., Stensland, H., Laboratory, S. R., University, M. K. D. C., Kesavulu, M., University, D. C., Leipzig, M. P. U., Potthast, M., Leipzig, U., SimulaMet, M. R., … Metrics, O. M. A. (2019, June 1). *Summarizing e-sports matches and tournaments: Proceedings of the 11th ACM Workshop on immersive mixed and Virtual Environment Systems*. ACM Conferences. https://dl.acm.org/doi/abs/10.1145/3304113.3326116

Appendix A: CS:GO Dataset Overview

This portion will provide details on the dataset and analysis used.

1. **CS:GO Pro Players Dataset Information**

   The dataset utilized in this analysis is sourced from Kaggle and focuses on professional CS:GO player performance. The dataset includes statistics spanning from the inception of the game to the most recent major, providing a comprehensive collection of samples for analysis.

   - Source: Kaggle
   - Dataset Title: CS:GO Pro Players Dataset
   - URL: CS:GO Pro Players Dataset on Kaggle

2. **Data Cleaning Process**

   During the exploratory data analysis (EDA), the dataset underwent initial cleaning. Irrelevant categorical information, such as player teams, names, previous affiliations, and countries, was removed to focus on variables directly influencing player ratings.

Appendix B: Supplementary Figures

This portion will provide figures/models relating to exploratory data and multiple regression analysis that has been conducted on the CS:GO Pro Players Dataset.

1. **EDA Figures**

2. **Multiple Regression Analysis Figures**

Appendix C: Statistical Analysis Output

This portion provides a detailed look into the statistical output from the multiple regression analysis including Ordinary Least Squares regression results, coefficients, p-values, and the variance inflation factor. The analysis is organized to highlight the significance of each variable and their contribution to the explanation of rating among players.

1. **OLS Regression Results**

2. **VIF Analysis Results**

Appendix D: Additional Analysis and Results

This portion will provide and present supplementary information on the statistical analysis, and additional tests performed.

1. **Additional Statistical Tests and Assessments**
    - Multicollinearity Assessment:
        - Explanation of the VIF Test: A detailed overview of how the Variance Inflation Factor (VIF) test was applied to identify and mitigate multicollinearity issues.
        - Eigenvalues Analysis: Examination of eigenvalues from the Ordinary Least Squares (OLS) regression results to assess multicollinearity.
    - Linearity Assessment:
        - Scatterplot Matrix: Visualization of the relationship between the rating and selected player statistics, indicating a potential positive trend
    - Normality and Homoscedasticity
        - Histogram and Q-Q Plot: Display of histograms and Q-Q plots to evaluate the normality of the distribution of differences between observed and predicted values.
        - Homoscedasticity Verification: Explanation and visualization confirming the consistency of residuals across the dataset.
2. **Variable Selection and Refinement**
    - Explanation of the iterative process of variable selection and refinement based on the VIF test results.
    - Identification of the final set of variables: opening_kill_rating, grenade_dmg_per_round, damage_per_round, assists_per_round, and saved_teammates_per_round.

Appendix E: Conclusion and Reflection

This portion provides insight into the conclusions drawn from the analysis, and reflections from the result the team has achieved.

1. **Influence of Damage_per_round on Player Rating:**
    - Detailed discussion on how Damage_per_round emerged as the most influential factor in determining player ratings, both positively and negatively.
2. **Considerations Beyond Statistics:**
    - Exploration of the limitations of a purely statistical approach and the importance of considering teamwork, strategy, and other non-statistical elements in player and team rankings.
3. **Discussion on Discrepancies in Rankings:**
    - Examination of the discrepancies between individual player rankings and team rankings, highlighting the need for a holistic assessment of player performance.
4. **Insights from Perfect Plot Fits:**
    - Analysis of the positive results obtained in the exploratory data analysis, attributing success to the consistent and defined value range of variables in CS:GO

Appendix F: Lessons Learned

This section outlines the important lessons and takeaways gained during the research process, and exploration of the CS:GO Pro players dataset.

1. **Importance of Clean Data in EDA:**
   - Reflection on the significance of data cleanliness during the Exploratory Data Analysis (EDA) phase and its impact on focusing on relevant variables.
2. **Precision in Model Crafting:**
   - Insights into the importance of crafting precise hypotheses and refining the model during Multiple Regression Analysis, with a focus on tackling multicollinearity issues.
3. **Robustness Checks for Statistical Assumptions:**
   - Discussion on the robustness checks performed for normality, linearity, and homoscedasticity, underscoring their role in ensuring the accuracy of the model.
4. **Understanding Domain-Specific Variables:**
   - Reflection on the importance of understanding domain-specific variables, exemplified by the pivotal role of damage_per_round in predicting player ratings.