MATH 6490 Final Project Report

Exploring Impactful Variables of Professional Tennis Player Success

Ryan Renken

Fall 2024

Contents

Introduction

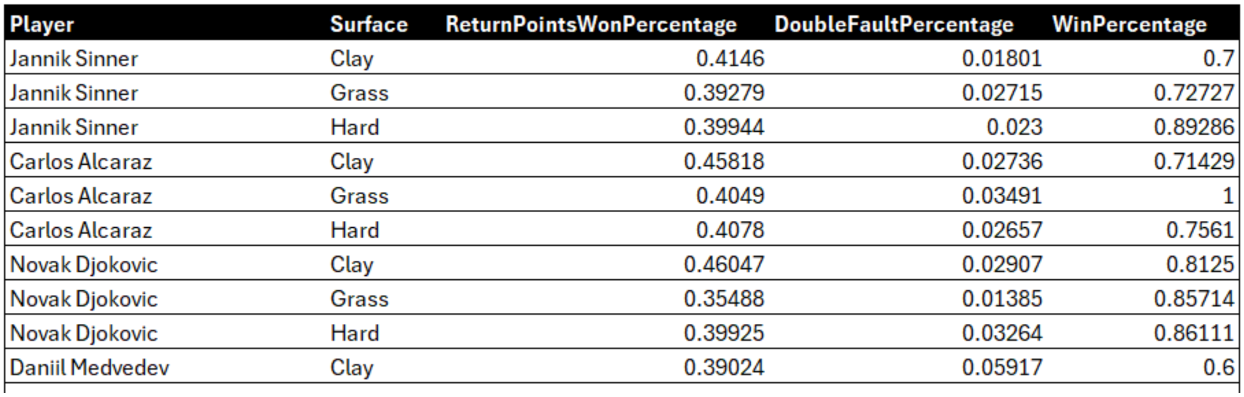
Background

Tennis is a dynamic and data-rich sport where performance is influenced by a variety of factors. From serve accuracy to the number of unforced errors, each statistic offers valuable insights into player behavior and match outcomes. Recent advances in sports analytics have shown the potential of statistical methods to uncover patterns and trends in match performance, ultimately assisting players, coaches, and analysts in making more informed decisions.

Motivation

The increasing availability of tennis match data presents a unique opportunity to explore how different player statistics impact the probability of winning a match. This study is motivated by the need to better understand these relationships through statistical analysis, providing a basis for enhancing player strategies and performance predictions.

Data



Player	Surface	ReturnPointsWonPercentage	DoubleFaultPercentage	WinPercentage
Jannik Sinner	Clay	0.4146	0.01801	0.7
Jannik Sinner	Grass	0.39279	0.02715	0.72727
Jannik Sinner	Hard	0.39944	0.023	0.89286
Carlos Alcaraz	Clay	0.45818	0.02736	0.71429
Carlos Alcaraz	Grass	0.4049	0.03491	1
Carlos Alcaraz	Hard	0.4078	0.02657	0.7561
Novak Djokovic	Clay	0.46047	0.02907	0.8125
Novak Djokovic	Grass	0.35488	0.01385	0.85714
Novak Djokovic	Hard	0.39925	0.03264	0.86111
Daniil Medvedev	Clay	0.39024	0.05917	0.6

Figure 1: Example snippet of raw data in a table.

The dataset used in this analysis was sourced from the SCORE online public sports repository and is provided in the form of a CSV file [1]. The dataset contains five key columns: Player Name, Court Surface, Return Points Won, Double Fault Percentage, and Win Percentage. These variables offer insight into various aspects of player performance, with the intention of identifying significant factors that may influence match outcomes. An example snippet of the raw data can be seen in Figure 1. Specifically, the data includes:

- **Player Name**: The name of the player.
- **Court Surface**: The type of surface the match was played on (e.g., grass, clay, hard court).
- **Return Points Won**: The percentage of points won when the opponent is serving.

- **Double Fault Percentage**: The percentage of serves that result in double faults, which surrenders a point to the opponent.
- **Win Percentage**: The overall percentage of matches won by the player over the course of a season.

This dataset is used to examine the relationships between key performance metrics and match success. While professional tennis matches typically track many additional statistics, the public repository only provided these five features for analysis. The data covers a full calendar year leading up to May 2024 and includes players who met specific thresholds: at least 10 matches overall or 5 matches on a specific surface during that time. Additionally, individual players appear multiple times in the dataset, with their performance records segmented by each of the three court surfaces.

Goals

The primary goal of this report is to identify the key factors that influence match results by applying exploratory data analysis (EDA) and regression techniques. By analyzing patterns within the dataset, we aim to uncover significant insights into tennis performance, with potential implications for players' training and competitive strategies. Specifically, this report focuses on understanding how playing surfaces and serving habits affect a player's overall win percentage over the course of a season.

The analysis begins with the application of EDA techniques to reveal trends and relationships that may not be immediately apparent from the raw data. Once the overall structure and distribution of the data are better understood, linear regression models will be fitted to explore potential linear relationships among the predictors. This approach allows for a systematic investigation into the factors that contribute most to match success.

Methodologies

Exploratory Data Analysis

This section aims to uncover the underlying story told by the dataset through exploratory data analysis (EDA). By examining the distributions, relationships, and patterns within the data, we can gain a deeper understanding of its structure and characteristics. Understanding these aspects not only provides clarity on the dataset's overall shape but also informs potential inferences and interpretations drawn from this sample. To achieve this, I will present a series of statistical graphics and accompanying insights that highlight key trends and relationships in the data.

Since win percentages are grouped by the playing surfaces on which each player competed, it is essential to examine the distribution of data points across the three surface types. Figure 2 illustrates the number of matches played on clay, grass, and hard courts. Notably, the dataset contains significantly fewer matches on grass compared to the other surfaces. However, with over 50 records for grass, the dataset does not exhibit an extreme imbalance that would raise concerns about the reliability of analyses involving this predictor. This

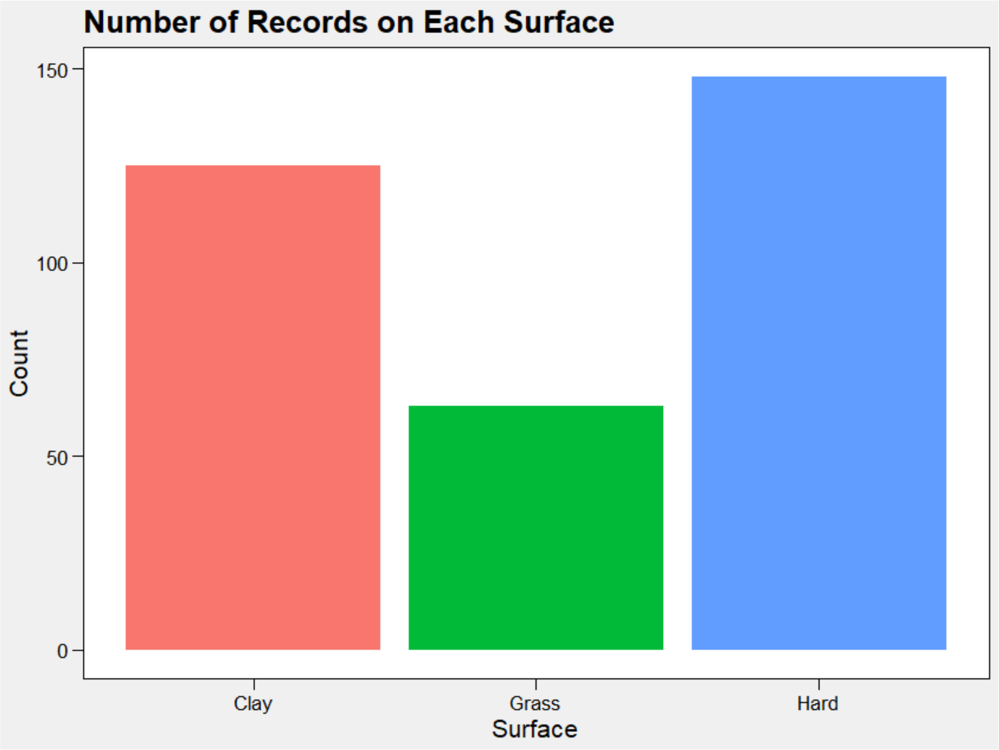


Figure 2: Total count of matches played on each of the three surfaces.

distribution highlights the differences in the frequency of play across surfaces, which may reflect the nature of the professional tennis calendar, as grass tournaments are less common than those on clay or hard courts.

To explore how win percentages might be influenced by the playing surface, I began by examining the overall distribution using box plots. Figure 3 displays the spread of win percentages grouped by surface type. The box plots reveal a slight increase in win percentages for matches played on grass compared to clay and hard courts. However, this difference is subtle, with no significant separation observed among the three surface categories in terms of variability or central tendency. These findings suggest that while surface type may have some influence, it is unlikely to be a dominant factor in determining win percentages based on this dataset.

To provide a direct comparison between the predictors and their relationships, Figure 4 presents a pair plot that visualizes multiple variables simultaneously. In this visualization, each variable (excluding player names) is plotted against every other variable in a grid format, akin to the structure of a correlation matrix.

The diagonal of the pair plot displays the distributions of each variable, grouped by court surface. In this graphic, the clay surface is represented in red, grass in green, and hard court in blue. These distributions reveal the shapes of each predictor and facilitate comparisons across surfaces. Notably, the distribution of *Return Points Won %* shows an increase in return points won on clay courts compared to other surfaces. This suggests that players returning serves might gain a slight advantage on clay, potentially due to its slippery texture and slower ball bounce, which can provide more time for an effective return.

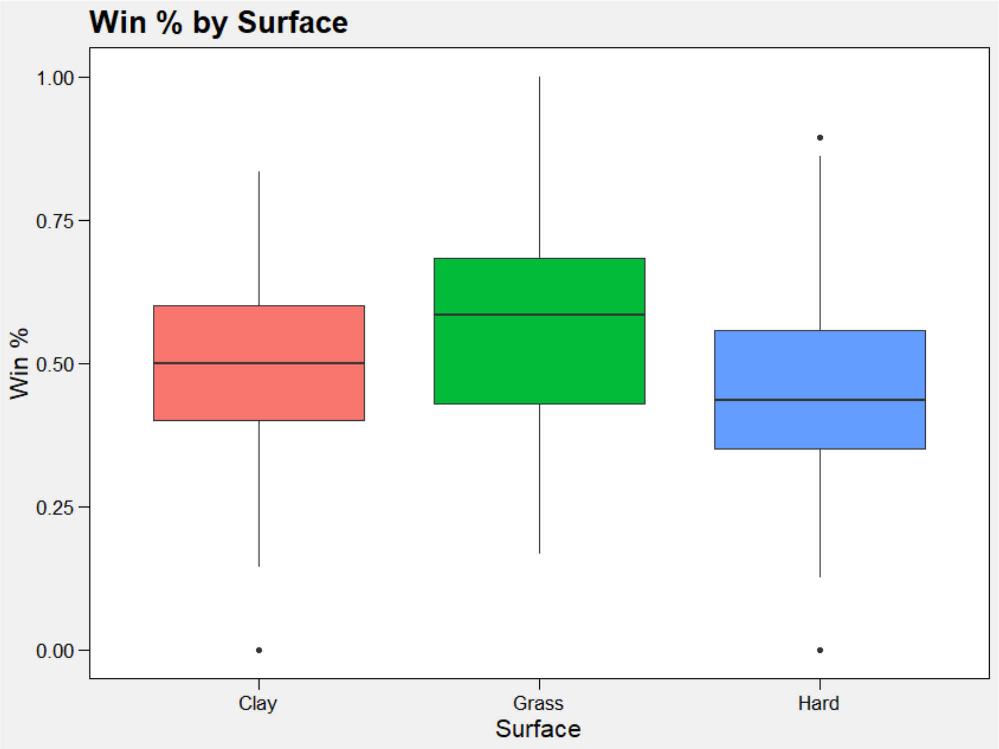


Figure 3: Side-by-side box plot of *Win %* grouped by surfaces. Overlapping interquartile ranges show that there is no distinct and significant separation between the three distributions of playing surfaces.

However, while this trend is visually apparent, the box plot corresponding to *Return Points Won %* reveals overlapping interquartile ranges across surfaces. This overlap indicates that there is no clear statistical evidence of significant separation among the surfaces, despite the observed patterns in the pair plot.

The scatter plot comparing *Return Points Won %* and *Win %* in the pairplot stands out as the most indicative of a linear relationship among the three scatter plots displayed. This visual observation is reinforced by the corresponding correlation coefficient shown on the mirrored side of the plot, which exhibits a relatively higher value compared to the other variable pairs.

Simple Linear Regression

In this section, I fit simple linear regression models to examine potential linear relationships between the predictors, *Double Fault %* and *Return Points Won %*, and the response variable, *Win %*. To better understand the model's performance, I also visualize the fitted regression line alongside the actual data points, providing a clear illustration of how well the model captures the relationship. This approach helps build intuition about the predictors' individual effects on *Win %*. In the following section, I will expand the analysis to include multiple predictors in a multiple linear regression model, where visualizing the regression line becomes impractical due to increased dimensionality.

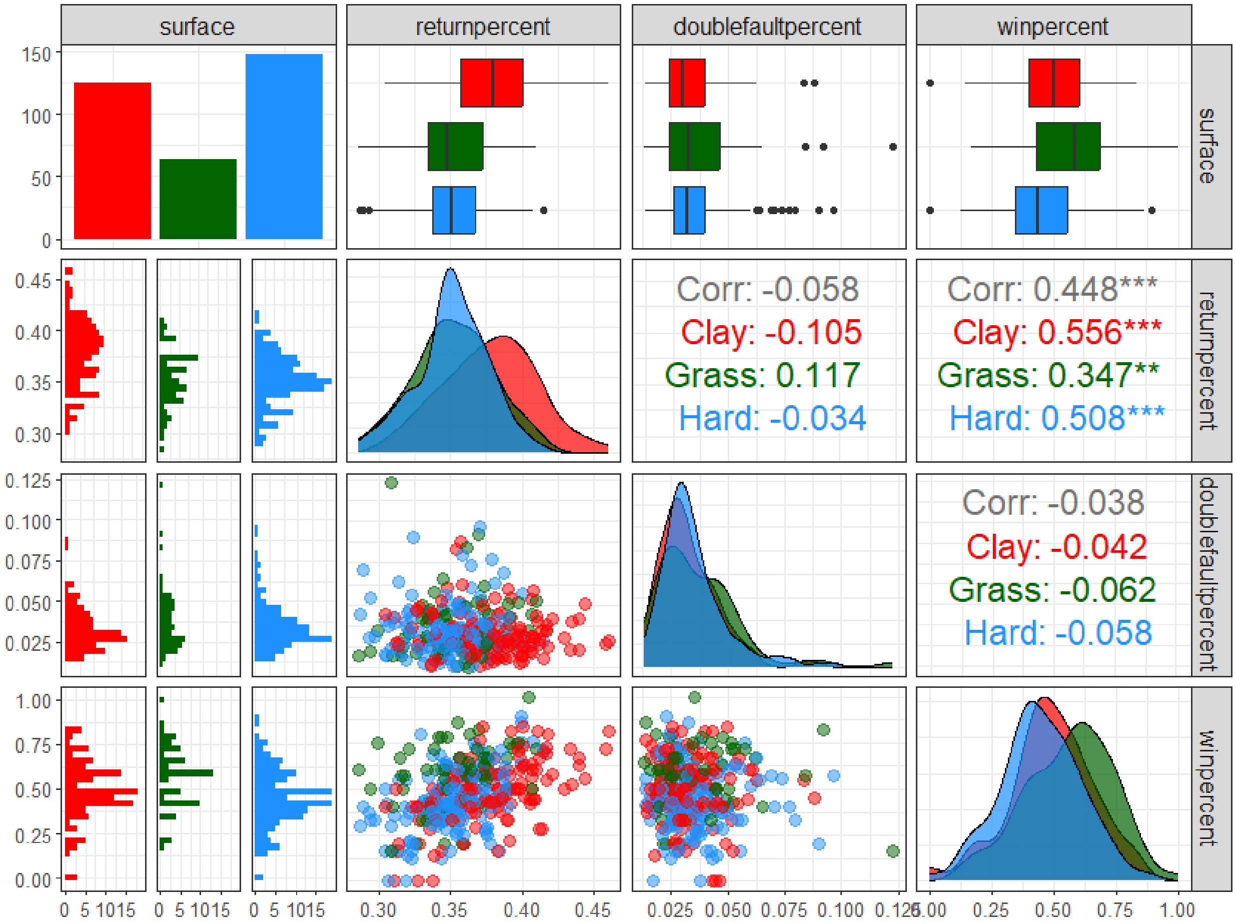
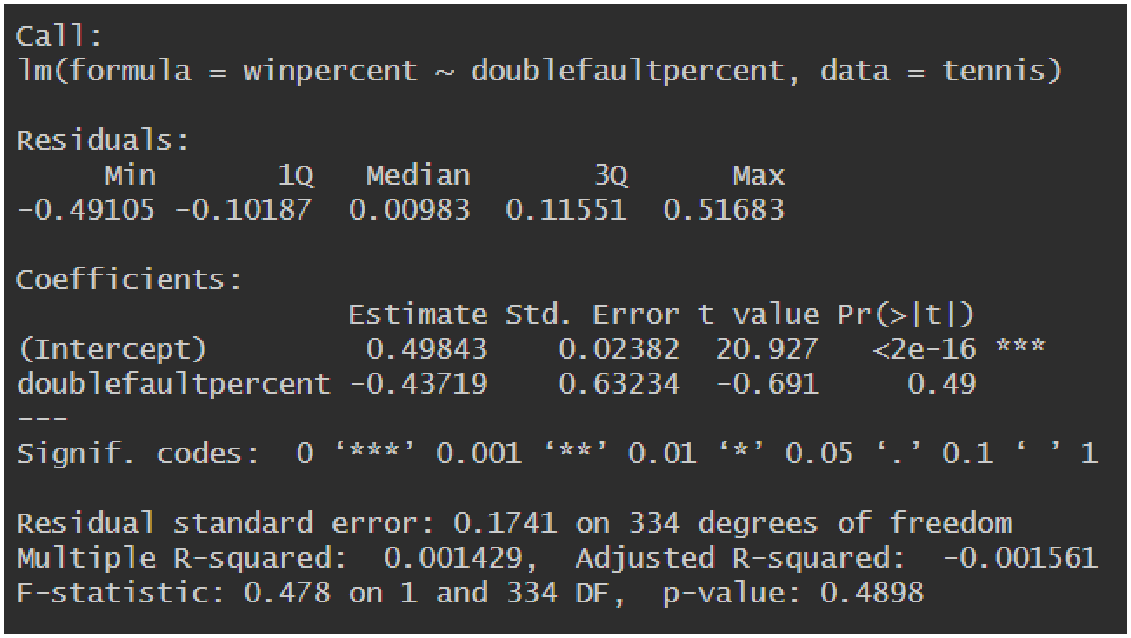


Figure 4: Pairplot of each predictor and response variable. This visual was created using the *GGally* package in R. The diagonal boxes in the grid show the distribution of each variable. The scatter plots under these distributions show the relationships between the corresponding variables. Each figure is broken down by *Surface* where red is clay, green is grass, and blue is hard courts.

Double Fault Percentage

The *Double Fault %* predictor does not appear to be a suitable variable for linear regression in this analysis. The model summary in Figure 5 reveals a high p-value for this predictor, indicating that it is not statistically significant in predicting the response variable, *Win %*. Additionally, the model's very low R^2 value suggests that it fails to account for any meaningful portion of the variance in the target variable. These findings highlight that double fault percentage does not have a strong or consistent linear relationship with match outcomes, making it an ineffective predictor in this context.

A visual inspection of the fitted regression line in Figure 6 shows that it does not align well with the response data points, indicating a poor fit. Although the line has a slightly negative slope, the data points do not follow this trend, which is consistent with the findings from the model summary in Figure 5. These observations suggest that the double fault percentage predictor lacks a meaningful linear relationship with *Win %*.



```
Call:
lm(formula = winpercent ~ doublefaultpercent, data = tennis)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.49105 -0.10187  0.00983  0.11551  0.51683 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         0.49843    0.02382  20.927   <2e-16 ***
doublefaultpercent -0.43719    0.63234  -0.691     0.49    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1741 on 334 degrees of freedom
Multiple R-squared:  0.001429,	Adjusted R-squared:  -0.001561 
F-statistic: 0.478 on 1 and 334 DF,  p-value: 0.4898
```

Figure 5: R output of *Double Fault %* linear model summary. A p-value of 0.49 suggests that this predictor is not significant in predicting the response variable.

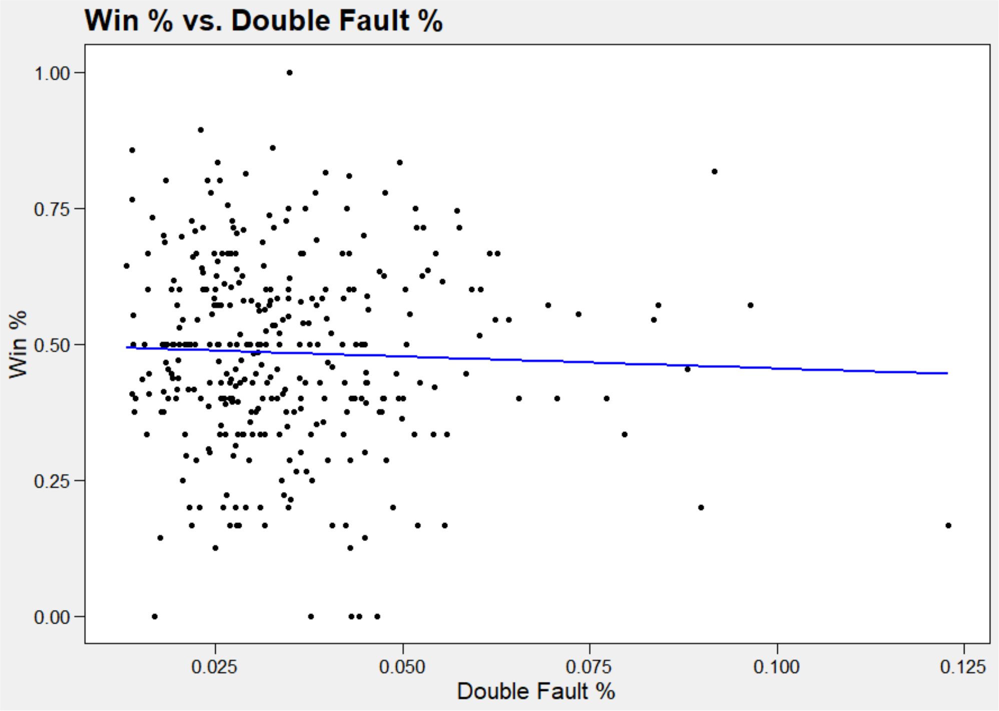


Figure 6: The blue line represents the fitted regression line predicting *Win %* as a variable of *Double Fault %*. The line suggests that the dots follow a slightly negative trend, but none of the points in the plot seem to follow this pattern.

The residual plot in Figure 7 further examines the model's performance, where we would expect the residuals to be randomly scattered without discernible patterns. While the LOWESS line in the residual plot shows some curvature, this may primarily result from the sparse number of data points in the lower value ranges. The limited data in these areas could cause the line to curve more than expected, rather than indicating a true pattern. Overall, apart from this curvature, the residuals appear relatively randomly dispersed, suggesting no strong evidence of systematic violations of the normality assumption.

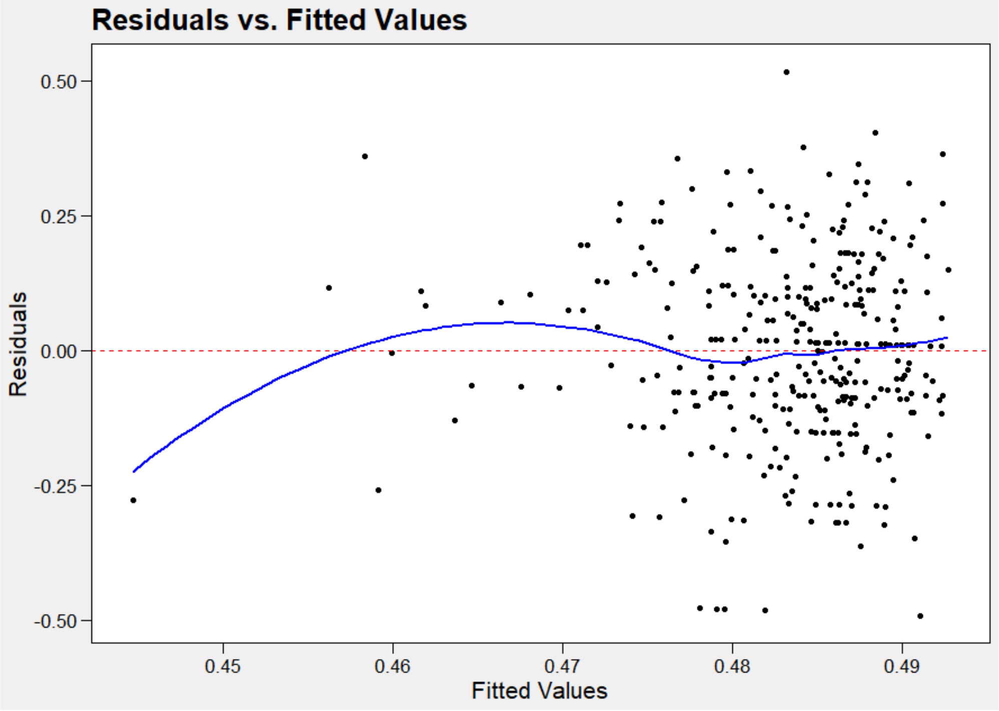


Figure 7: *Double Fault %* Residual Plot with LOWESS curve to help indicate patterns in the scatter plot.

To further evaluate the normality of the residuals, a Quantile-Quantile (QQ) plot can be used to compare the residuals against a diagonal reference line representing a normal distribution. Figure 8 demonstrates that most data points follow the normal reference line relatively closely, with slight deviations observed at the tails. To confirm these observations, a Shapiro-Wilk normality test was conducted. The resulting large p-value indicates insufficient evidence to reject the null hypothesis, suggesting that the residuals do not deviate significantly from a normal distribution.

While the visual diagnostics for the residuals, including the QQ plot and residual plot, did not provide definitive evidence of a violation of normality assumptions, the overall performance of the model suggests that *Double Fault %* is not a meaningful predictor of *Win %*. The QQ plot indicated that the residuals followed a normal distribution reasonably well, and any deviations appeared minor. However, the model summary output in R revealed a high p-value for the *Double Fault %* predictor and a very low R^2, confirming that it is not statistically significant and explains little to no variance in the response variable. Based on these findings, I will exclude *Double Fault %* from any subsequent models in my analysis, focusing instead on predictors that demonstrate stronger relationships with *Win %*.

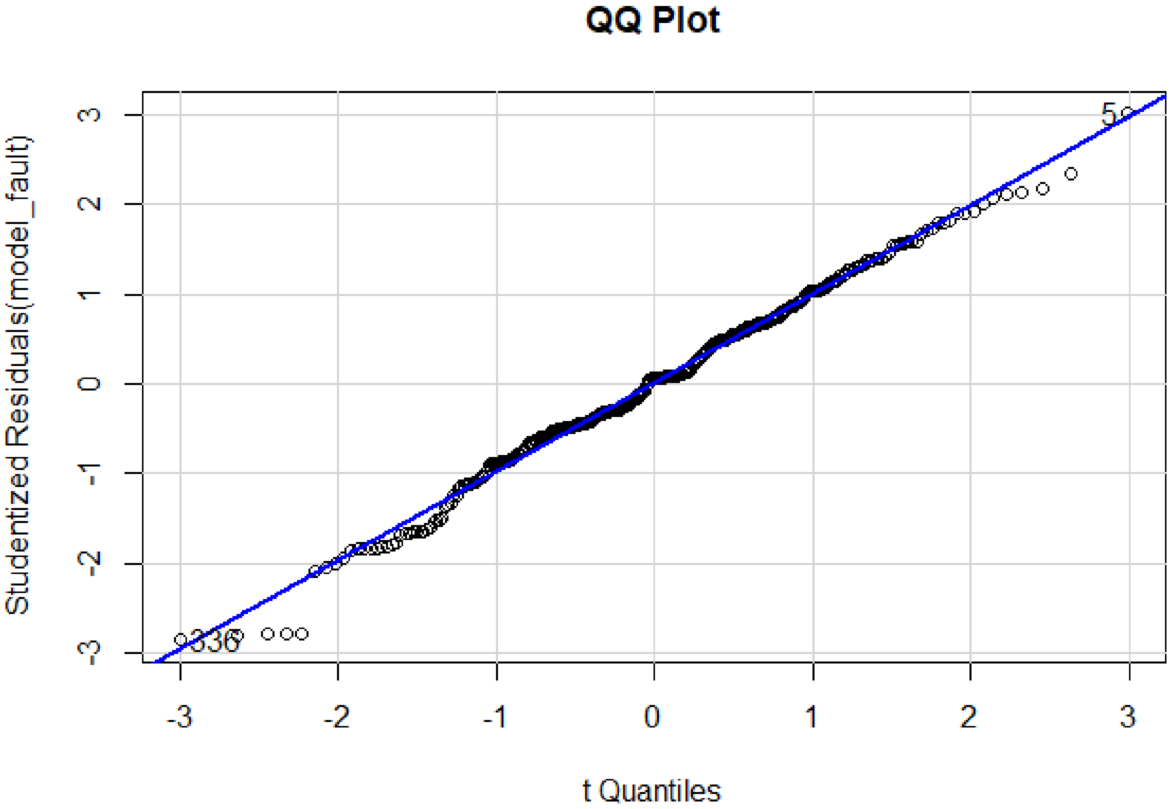


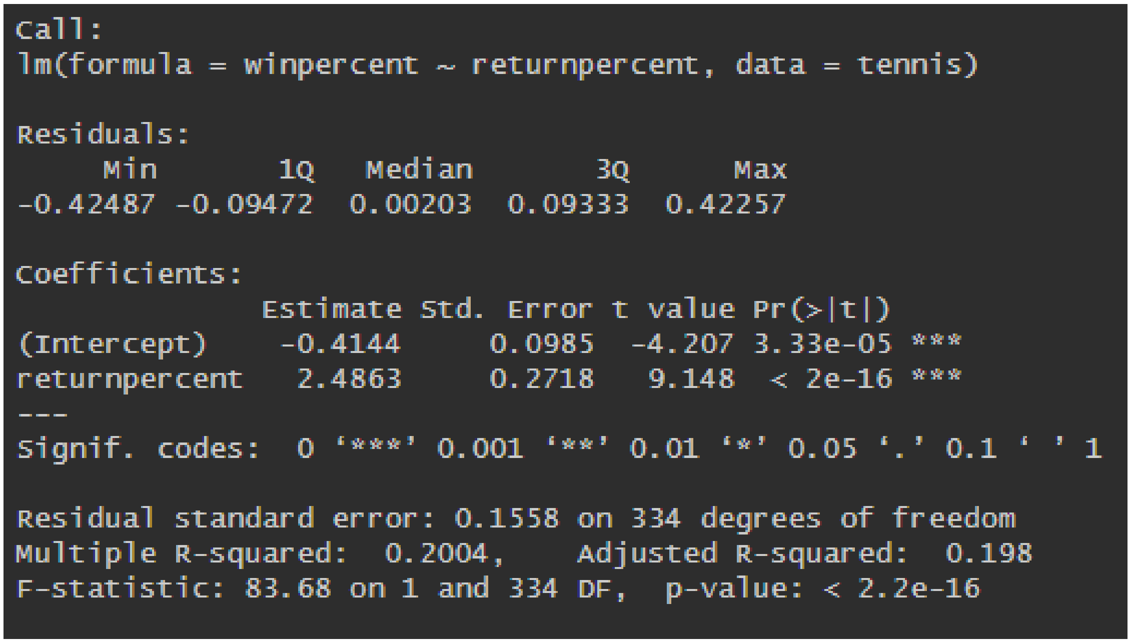
Figure 8: QQ Plot of residuals from *Double Fault %* simple linear regression model on training data. Data points that follow closely to the blue line would indicate they are likely to follow an approximately normal distribution.

Return Points Won Percentage

It is intuitive to assume that a higher percentage of return points won would correlate with greater overall success in tennis. To evaluate this relationship, I fit a simple linear regression model using *Return Points Won %* as the sole predictor of the response variable, *Win %*. The model summary, shown in Figure 9, reveals a very low p-value for the predictor, indicating that it is statistically significant in predicting *Win %*. However, the R^2 value of 0.2004 indicates that the model explains only about 20% of the variability in the response variable. While *Return Points Won %* has a significant and positive linear relationship with *Win %*, its predictive power remains relatively modest.

Figure 10 displays the scatter plot of *Return Points Won %* versus *Win %*, with the fitted linear regression line overlaid to visually assess the model's fit. Unlike the previous model in Figure 6, this plot clearly shows a closer alignment between the data points and the fitted line, suggesting that a linear model is more appropriate for this relationship.

To evaluate the assumptions of linear regression, I first examined the residual plot shown in Figure 11. The plot does not reveal any obvious patterns around the red reference line, indicating that the constant variance assumption holds, and there is no evidence of heteroscedasticity.



```
Call:
lm(formula = winpercent ~ returnpercent, data = tennis)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.42487 -0.09472  0.00203  0.09333  0.42257 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -0.4144     0.0985  -4.207 3.33e-05 ***
returnpercent   2.4863     0.2718   9.148  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1558 on 334 degrees of freedom
Multiple R-squared:  0.2004,	Adjusted R-squared:  0.198 
F-statistic: 83.68 on 1 and 334 DF,  p-value: < 2.2e-16
```

Figure 9: R output of *Return Points Won %* linear model summary.

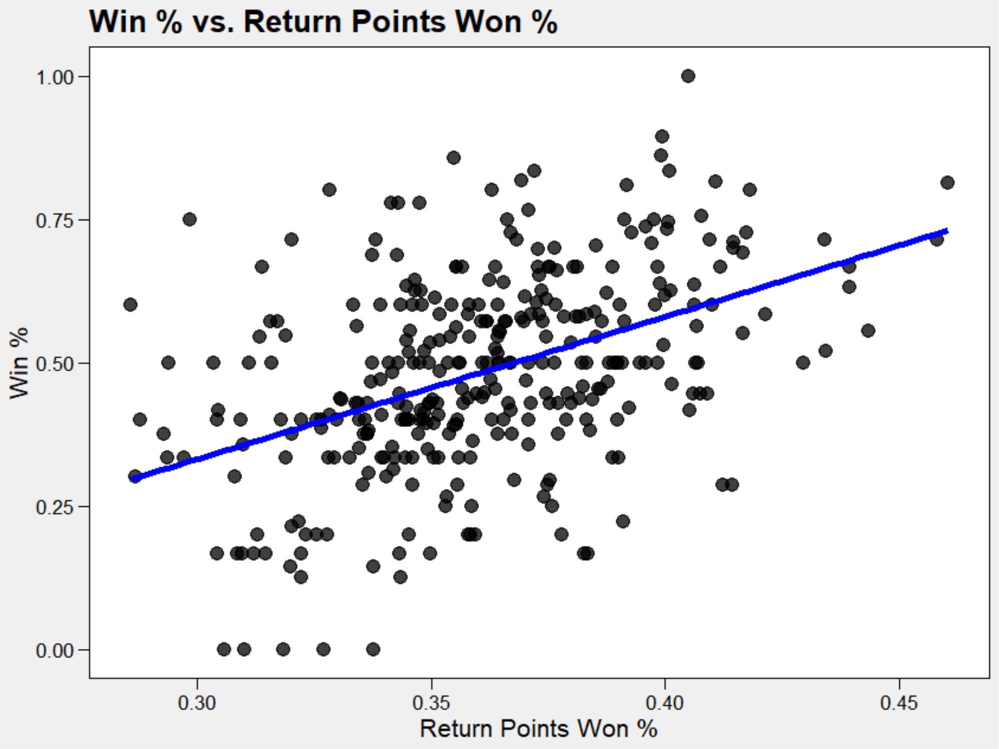


Figure 10: Fitted regression line predicting *Win %* as a variable of *Return Points Won %*

Next, I used a QQ plot to assess whether the residuals follow a normal distribution. Figure 12 shows that the residuals closely follow the normal reference line, suggesting that the normality assumption is not violated. This conclusion is further supported by the Shapiro-Wilk normality test, which yields a large p-value, indicating no significant deviation from normality.

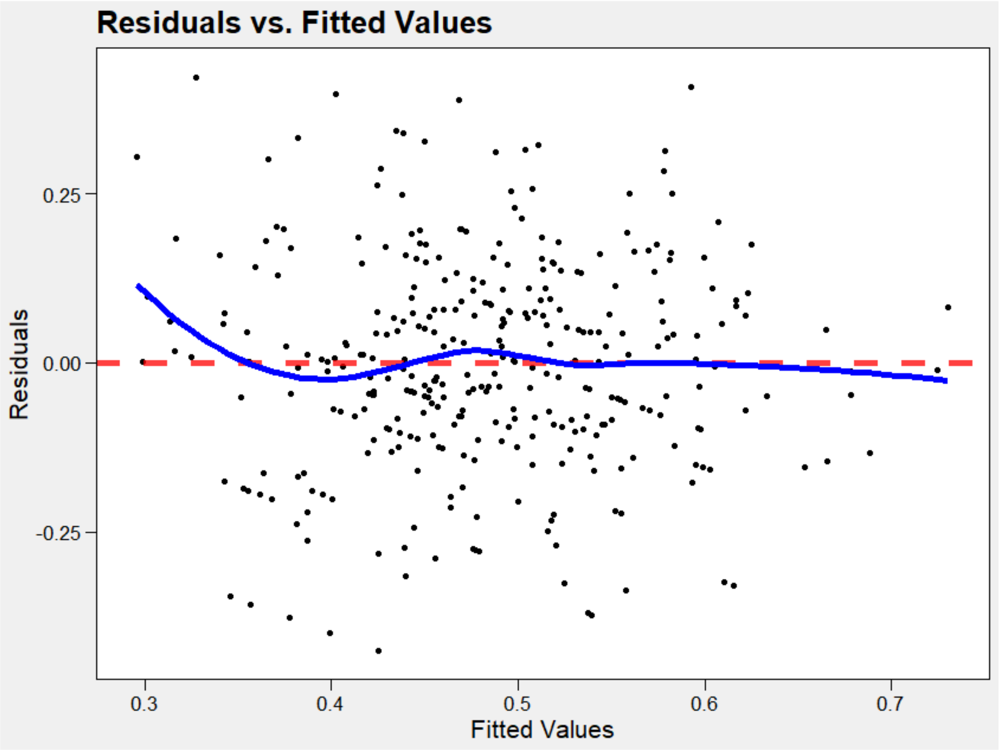


Figure 11: *Return Points Won %* Residual Plot with LOWESS curve to help indicate patterns in the scatter plot.

Overall, these diagnostic checks suggest that *Return Points Won %* has a positive linear relationship with *Win %* in this dataset. However, given the relatively low R^2 value, which indicates that only about 20% of the variance in *Win %* is explained by the predictor, this model would not be suitable for situations requiring precise predictions.

Multiple Linear Regression

In this section, I extend the simple linear regression model from the previous analysis to incorporate the categorical variable *Surfaces*. Because categorical variables cannot be directly interpreted by a regression model, they must first be converted into numeric representations using indicator (or ”dummy”) variables. This approach ensures that the model does not incorrectly interpret text labels as having a meaningful order or magnitude. For a variable with three categories, such as *Surfaces*, we create two new binary variables to represent the first two categories, while the third category serves as the reference (baseline) group. When both indicator variables are set to zero, the record implicitly belongs to the baseline group. Fortunately, R simplifies this process when categorical variables are properly preprocessed

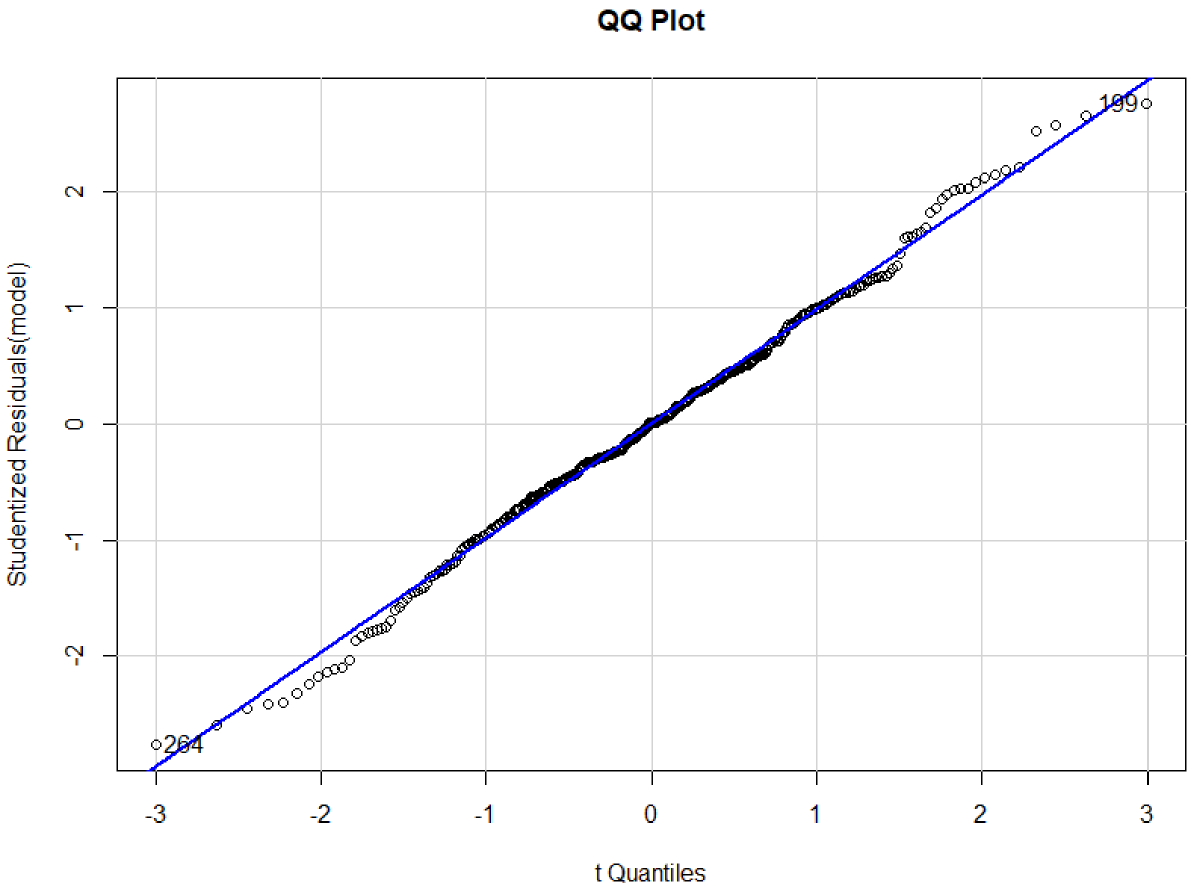


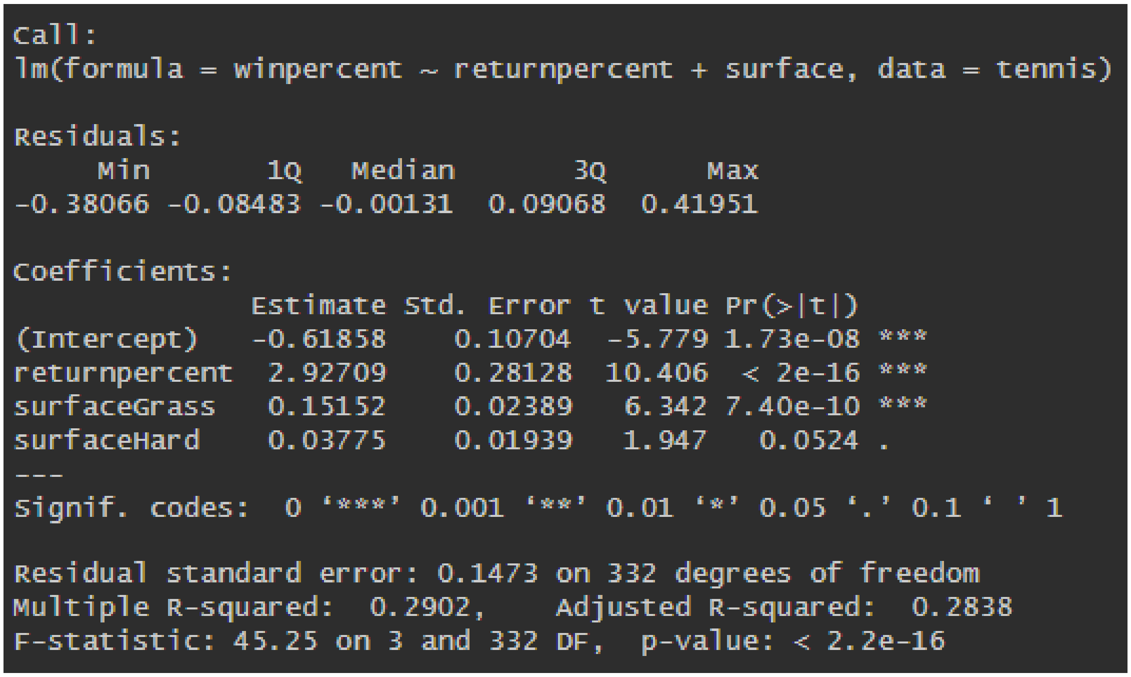
Figure 12: QQ Plot of the *Return Points Won %* simple linear regression model residuals with diagonal normal reference line. Data points that follow closely to the blue line would indicate they are likely to follow an approximately normal distribution.

as factors. The regression model will know to automatically assigns one category as the baseline and generates the corresponding dummy variables for the other categories, making it easy to include the *Surfaces* variable in the regression model.

In Figure 13, we observe how the inclusion of the *surface* variable is factored into the model. The regression output introduces two new predictors, *surfaceGrass* and *surfaceHard*, each with unique coefficient estimates. This indicates that the clay surface is treated as the baseline category for the *surface* variable in the model.

Examining the p-values of the predictors, both *returnpercent* and *surface* are statistically significant in predicting *Win %* in this dataset. Furthermore, the slight improvement in the Adjusted R^2 value suggests that incorporating the *surface* variable enhances the model's ability to explain variability in the response variable.

Again, we must next check the linear model assumptions to justify the model's validity. First, looking at the residual plot in Figure 14, we don't see any distinct patterns and the variance looks to be approximately constant among the data points. Also, the additional LOWESS line fit to the data in this plot seems to follow very closely to the red reference line further confirming that the linearity assumption is reasonably satisfied.



```
Call:
lm(formula = winpercent ~ returnpercent + surface, data = tennis)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.38066 -0.08483 -0.00131  0.09068  0.41951 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -0.61858    0.10704  -5.779 1.73e-08 ***
returnpercent  2.92709    0.28128  10.406  < 2e-16 ***
surfaceGrass   0.15152    0.02389   6.342 7.40e-10 ***
surfaceHard    0.03775    0.01939   1.947   0.0524 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1473 on 332 degrees of freedom
Multiple R-squared:  0.2902,	Adjusted R-squared:  0.2838 
F-statistic: 45.25 on 3 and 332 DF,  p-value: < 2.2e-16
```

Figure 13: R output of multiple linear regression model summary with predictors *Return Points Won %* and *surface* and response variable *Win %*.

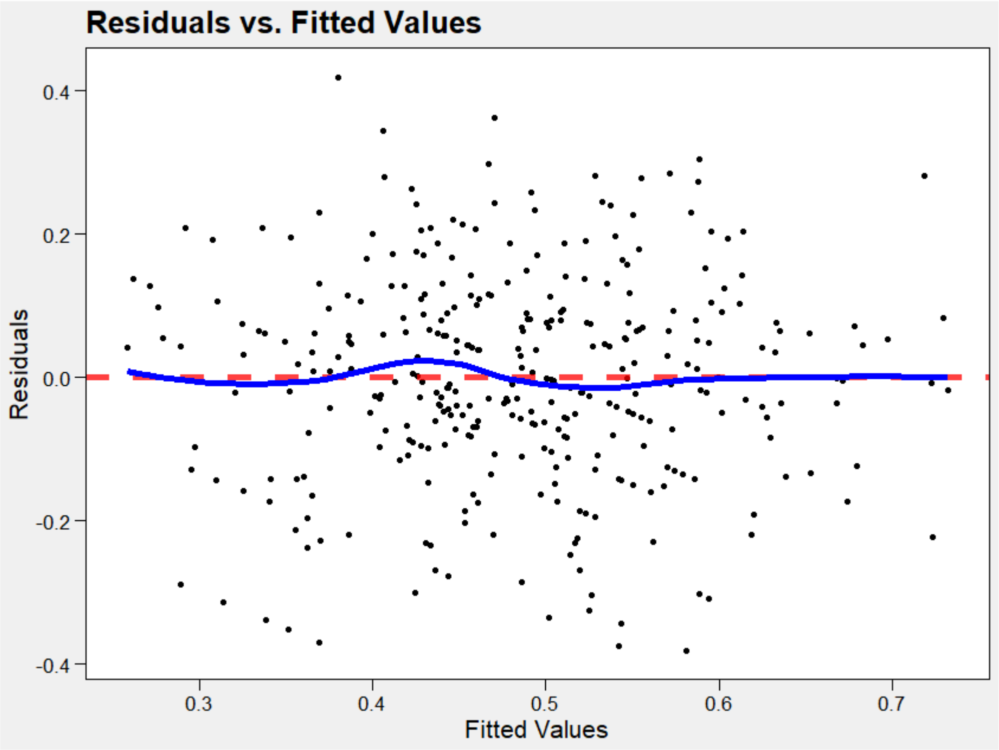


Figure 14: Final multiple linear regression model residual plot with fitted LOWESS curve to help indicate patterns in the scatter plot.

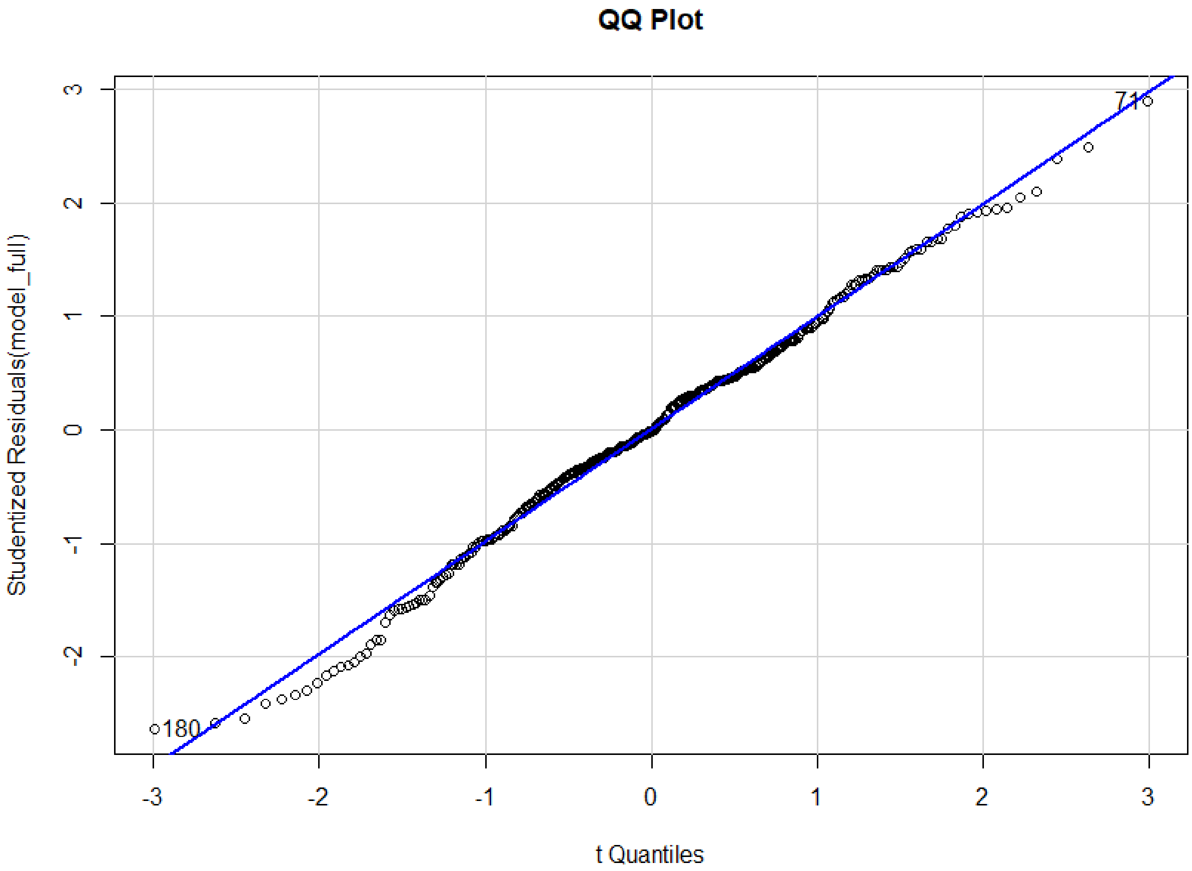


Figure 15: QQ Plot of full model residuals with diagonal normal reference line. Data points that follow closely to the blue line would indicate they are likely to follow an approximately normal distribution.

The QQ plot for the residuals, shown in Figure 15, indicates that the data points closely align with the normal reference line, suggesting that the residuals are approximately normally distributed. While there is minor deviation at the tails, this separation is not substantial enough to raise significant concerns. Additionally, the Shapiro-Wilk normality test yielded a p-value that supports failing to reject the null hypothesis, further affirming that the residuals follow a normal distribution.

With the model showing improvement and no violations of linear regression assumptions, I conducted an influence analysis to identify any data points that might significantly impact the model's fit. Highly influential points, if present, may need to be addressed or removed to ensure the model accurately represents the rest of the data.

Figure 16 displays a plot of Cook's Distance, a metric that combines leverage and residual information to quantify the influence of individual data points. Cook's Distance helps identify observations that are extreme in either the independent or dependent variables—or both. Typically, values in the range of 0.3 to 0.5 or higher are considered potentially influential. In this case, none of the data points exceed a Cook's Distance of 0.1, indicating no cause for concern regarding influential points in the dataset.

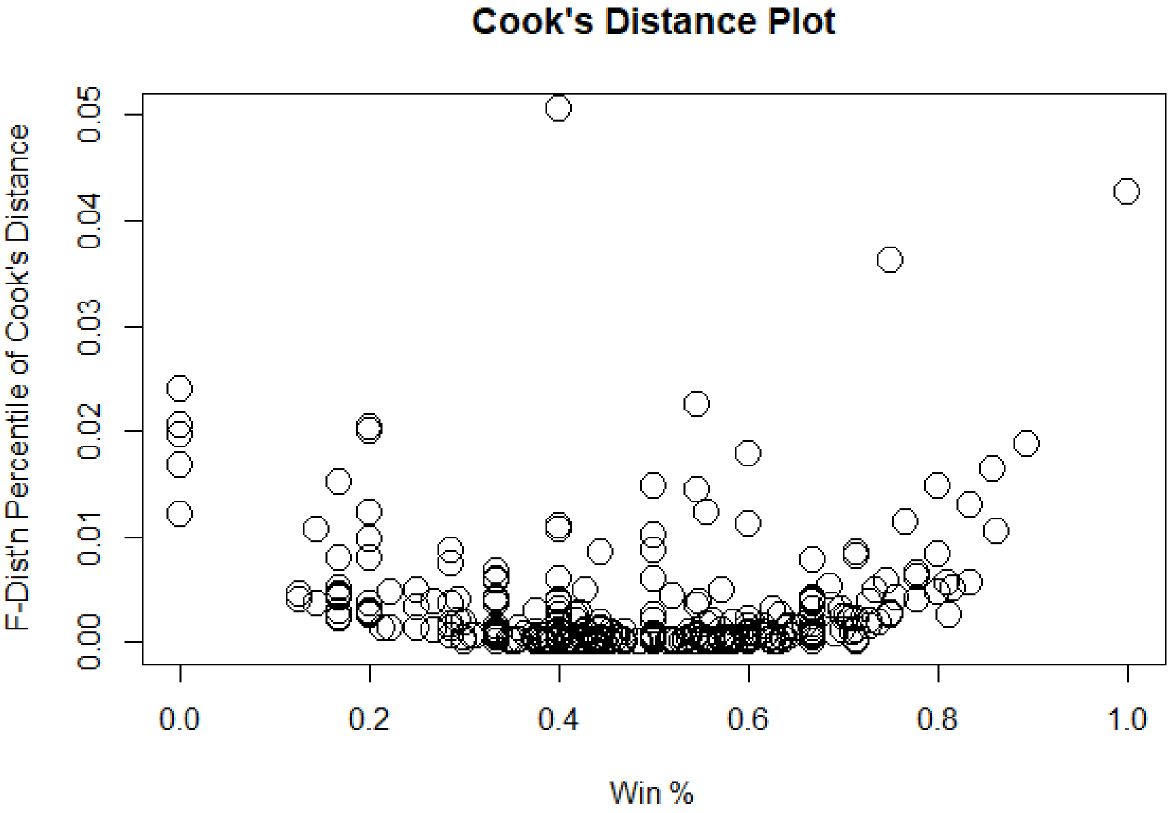


Figure 16: Cooks Distance Plot for the full final multiple linear regression model. The y-axis is scaled down and only increase in increments of 0.01.

To further examine influence, I used an influence plot, shown in Figure 17, which visualizes Cook's Distance alongside leverage and residual information. The x-axis represents hat values, or leverage, while the y-axis reflects standardized residuals. To enhance clarity, the tick marks for leverage have been scaled down from the usual increments of 0.1. In this model, all data points are clustered below a leverage of 0.04, indicating that no observations exhibit high leverage or disproportionately affect the model.

On the y-axis, data points are plotted based on their Studentized Residuals, indicating how far each observation deviates from the predictions made by the full model. Points falling beyond the horizontal dotted reference lines may warrant closer scrutiny, as they represent predictions with relatively large deviations.

The color and size of each circle reflect the corresponding data point's Cook's Distance. The legend above the plot illustrates that darker shades of blue and larger circle sizes represent higher Cook's Distance values, while lighter colors and smaller circles correspond to lower values. Since the highest Cook's Distance in Figure 16 does not exceed 0.06, the coloring scale in Figure 17 has been adjusted accordingly. As a result, although some circles appear darker, they remain within an acceptable range, and no points exhibit Cook's Distance values that would raise concerns about undesired influence on the model.

This influence analysis confirms that the final full model, which includes *Return Points*

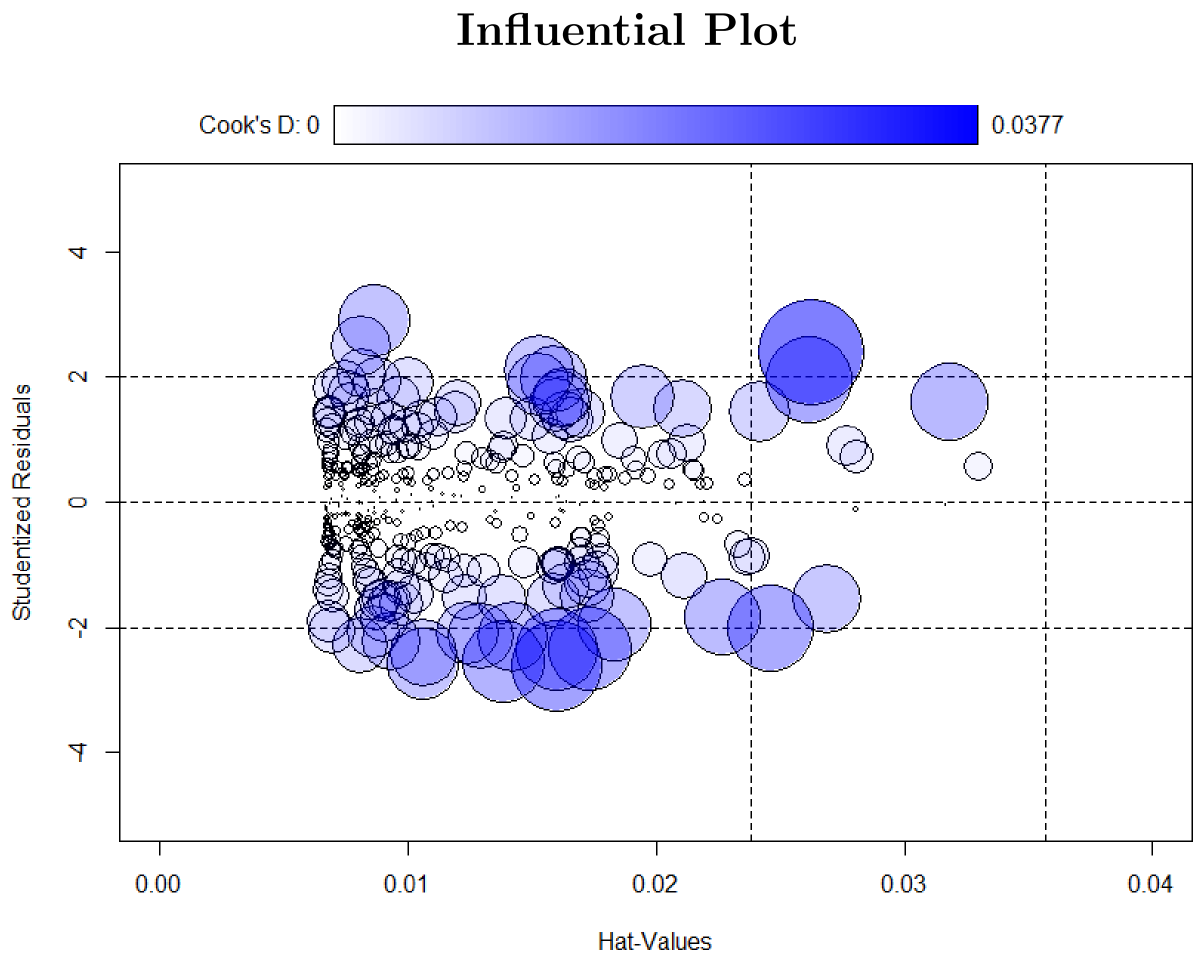


Figure 17: Influence Plot of the final multiple linear regression model. The x-axis represents the magnitude of leverage for each data point, while the y-axis reflects standardized residuals. Points that are larger in shape and darker in color indicate higher Cook's Distance values.

Won % and *Surface* as predictors of *Win %*, does not contain any data points with extreme influence or leverage that would distort the model's fit. Both the Cook's Distance plot and the influence plot indicate that all observations fall within acceptable ranges, with no outliers or highly influential points identified. As no records were removed from the dataset, this model represents the best fit identified in this study and is retained as the final model.

Conclusion

Among the predictor variables analyzed in this dataset, two were identified as statistically significant in predicting the response variable using the linear regression models developed in this study. The *Double Fault %* predictor, on the other hand, had a high p-value in the regression summary output and violated the normality assumption for residuals. These findings suggest that *Double Fault %* does not exhibit a linear relationship with player success during the analyzed season. It is important to note, however, that this conclusion applies specifically to linear models. More flexible modeling approaches might reveal non-linear relationships that were beyond the scope of this analysis.

The predictors *Return Points Won %* and *Surface* were both found to be statistically significant, with low p-values supporting their importance in predicting player success. Despite this, the final multiple linear regression model, which incorporated these predictors, produced an R^2 value no higher than 0.20, indicating that the model explains only a small proportion of the variance in the response variable. This highlights the model's limited predictive capability and suggests that linear regression may not fully capture the complexities of this dataset. Nevertheless, the significant predictors reveal meaningful relationships, such as a positive association between player win percentages and return-serve performance across different surfaces. These findings provide a foundation for further exploration using more sophisticated or non-linear modeling techniques.

Future Works

The primary objective of this project was to fit linear models to the dataset and utilize statistical graphics to draw inferences about the relationship between various predictors and player success. However, the results of this study indicate that linear regression models are not well-suited for accurately predicting professional player success based on the available predictors. As an extension to this research, exploring more flexible or nonparametric models may yield better results for prediction tasks, as they can capture more complex relationships between variables that linear models fail to account for. Prediction accuracy is particularly valuable for players aiming to forecast future performance in order to strategically prepare for upcoming matches and tournaments.

This analysis was limited by the number of predictor variables included in the dataset, which is relatively small compared to the vast range of metrics typically tracked during a tennis season. Incorporating additional player metrics could potentially improve the predictive power of the model. Therefore, future studies may benefit from broadening the scope of the dataset by including more variables, allowing for the application of linear regression models while retaining their simplicity and interpretability. This would provide users with more accessible tools for prediction without necessarily relying on more complex models that may not offer meaningful insights.

Furthermore, the introduction of additional variables could reveal interesting interactions between predictors that might enhance the model's performance. While all possible interaction terms were tested in this study, none provided significant improvements. However, in sports, the factors contributing to player success are often intertwined in complex ways. It is plausible that adding more variables could uncover synergistic relationships that not only improve model fit but also provide deeper insights into the drivers of player performance.

References

[1] E. Seltzer, “Win percentages by surface in professional tennis,” 2024. [Online]. Available: https://data.scorenetwork.org/tennis/atp_player-stats.html#references