# AMS 572 Project Report

Michael Deisler, Hasan Moonam, Taejin Park, Gabrielle Vaillant
Stony Brook University
AMS 572

# Table of Contents

## I. Background Information

We opted for a tennis statistics dataset for our analysis, which focuses on the four Grand Slam tournaments of 2014, encompassing both men's and women's competitions[1]. We concentrated our analysis on the men's Australian Open data. The Australian Open 2014 data included 42 variables and 127 observations. The dataset was structured based on individual matches played in the tournament, with match statistics presented in the columns. It was sourced from UC Irvine's Machine Learning repository where it includes supplementary variable information to better understand the column names.

We decided to create a processed data set for our analysis that was based off of the original data set for the 2014 Australian Open. It should also be noted that since there are 128 unique players in this tournament, 127 matches occur to determine the winner of the tournament. Since there are player statistics for 2 players for each match, we should have 127 * 2 = 252 rows of statistics in our data. However, one of the scheduled matches was canceled before it even started due to one of the players withdrawing from the third round because of a serious back injury.[2] The original dataset we used did not include the statistics for this match, therefore we decided to exclude that match from our analyses as well (since there are no meaningful player statistics from a match that never happened). As a result, our dataset contains 252 rows of player statistics. Since the original dataset provides match by match statistics as 127 rows where each row contains data for both players, we used Excel to split each row into two rows and then assigned the match outcome for each player to their corresponding statistics. This makes more sense to analyze the performance of a player using the corresponding statistics such as first or second serve percentage, first or second serve won etc.

## II. Objectives

The objective of this project was to understand how certain variables will affect the performance of a player in grand slam tournaments. We chose to study grand slam tournaments because they are the most watched tennis tournaments in the world and they also give the players the highest prize money and the highest ranking points (2000 points for the winner). In preparation for a grand slam, players dedicate their training to specific strategies and game statistics, such as first serve in percentage and the number of unforced errors. This focused approach aims to enhance their performance in grand slam events. Analyzing data from past tournaments allows players to identify the specific statistical benchmarks required not only to qualify for the tournament but also to excel in it. Our first hypothesis focuses on the correlation between two variables. We want to see if the number of net points (a specific strategy) won is correlated to the winners earned by players in the tournament. For this hypothesis, we use two different methods (Spearman and Kendall) to analyze the linear relationship between net points won and winners earned in the population. Our second hypothesis focuses on how a specific set of variables affects the outcome of a match. In other words, we want to know if there is a

[1] Though the original data set says it's from 2013, after some research we have concluded it is actually 2014 grand slam match data.

[2] https://www.tennis.com/news/articles/pospisil-withdraws-from-australian-open-wawrinka-into-fourth-round

statistically significant relationship between the match outcome as well as the performance of a player and independent variables.

## III. Missing Values

### A. Background Information on Missing Values

There are two distinct categories of missing values that you may encounter when dealing with a data set: values that are Missing Completely at Random (MCAR) and values that are Missing Not at Random (MNAR). MCAR values are values whose probabilities are unrelated to observed data and unobserved data. These values are equally likely to be missing as any other missing values in the dataset, showing no observable pattern or relationship with the other missing values. MNAR values are the opposite case. The missingness of these values depend on the missing values themselves or depend on unobserved data.

An example of a data set containing MCAR values would be if each survey respondent decides whether to answer a specific question on a survey by rolling a die and refusing to answer if the die lands on “6”[3] . With our tennis data set, these MCAR values might exist if a random event such as someone forgetting or missing to record a certain tennis statistic for a match occurs. Another scenario, may be if a match was played on a court that did not have the resources to record a certain variable. As long as the probability of the value being missing is not influenced by any other tennis statistics during the tournament, but rather by random events that are not related to the tennis statistics themselves. On the other hand, MNAR are related to the missing values themselves or on unobserved data. For example, MNAR values depending on missing values themselves can occur when people with higher earnings are less likely to reveal them in a survey[4]. An example of this occurring for our data set could be if a player was embarrassed about a low statistic for a match and they refused to report the statistic to the data set.

Therefore, in MNAR scenarios, the probability of an observation being missing is related to the unobserved value. This means that certain patterns of missingness are systematic and not random. Multiple imputation assumes that the missingness is random, and when this assumption is violated, imputations will not accurately reflect the underlying distribution of the missing values. It relies on creating plausible imputed datasets to account for uncertainty. In the presence of MNAR data, the imputations may not fully capture the variability and patterns of the missing values. As a result, the standard errors of the estimates may be underestimated, leading to overly optimistic confidence intervals.

To handle missing values, either MCAR and MNAR, we used the Multivariate Imputation by Chained Equations (MICE) package in R. The MICE algorithm predicts missing values for a variable using a model in which that variable serves as the outcome, and the remaining variables act as predictors. Once these predictions are made, the algorithm proceeds to the next variable, predicting its missing values based on the values of all other variables. These regression equations are interconnected in a chain, thus giving rise to the algorithm's name.

---

[3] www.stat.columbia.edu/~gelman/arm/missing.pdf.

[4] www.stat.columbia.edu/~gelman/arm/missing.pdf.

### B. Imputing MCAR values

Our original data set did not have any missing values, therefore we created our own. We first set the probability of missingness to be 0.2 for all missing values. In our case, it's very easy to see which values are MCAR because we created them ourselves, but in data sets that naturally have missing values, it can be difficult to distinguish between MCAR and MNAR.

We first created MCAR values by using a function called "delete_MCAR" from the missMethods library. This function allows us to specify the probability of missing values for the entire dataset. For our project, we only generated missing values for the columns we were directly working with for our two hypotheses. These columns are named "WNR", "NPW" in the processed data set. Those two columns give the number of winners and number of net points won for each player during each match of the tournament. The "NPW" variable was interesting to study because going to the net during a point is a strategy that some players may use to win the point. The "WNR" variable is also interesting to study because we can examine the correlation between the number of winners and the number of net points won.

After generating the missing values, our next step was to figure out how to deal with the missing values. This step was very important, because we wanted to pick a method that did not disrupt the distribution of the variable drastically. First, we check the original distribution of the variable with the missing values ("WNR") by visualizing the data using the "ggplot" library in R. We then tried some basic imputation methods such as mean, median and zero imputation for the missing values and compared their distributions to the original (Figure 1). These methods are ways to retain the full sample size which can be helpful for bias and precision but can create different kinds of bias.

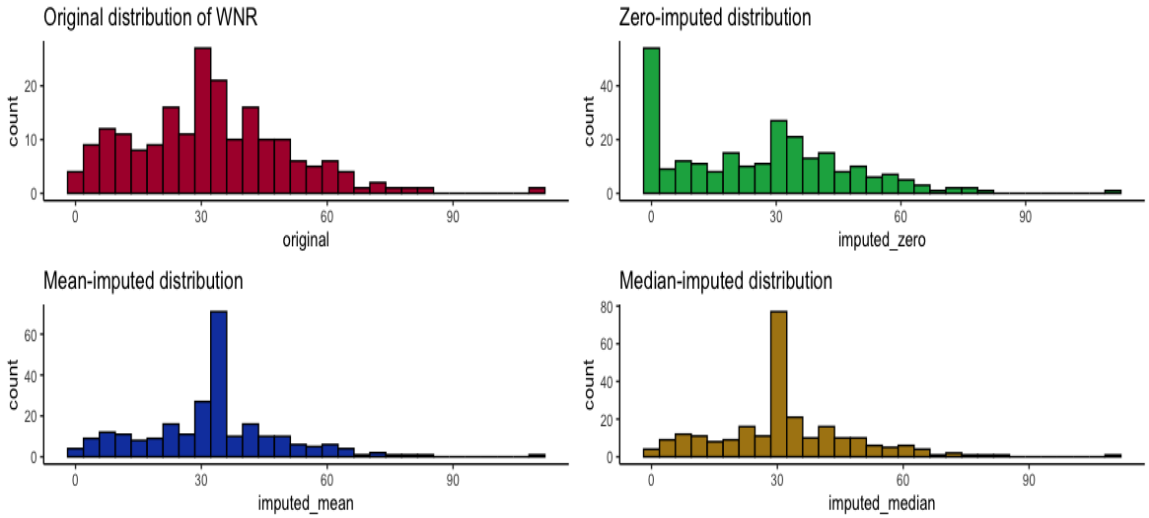


Figure 1: Comparing the original distribution with the resulting distributions from value imputed imputation methods

Figure 1 shows how none of the three imputation methods would be good to use for our analysis. The mean imputed method can distort the distribution of the variable which could lead to errors in the summary measures such as underestimating the standard deviation and can

also distort the relationship between two variables[5]. This is something we especially do not want because this can pull the correlation of the two variables closer to 0 and in our first hypothesis we are specifically looking at the correlation of variables.

Now we look for a different method to handle our MCAR values. One of the most widely used packages in R to handle missing values is the MICE package. First, we wanted to focus on a certain set of variables in our data set. We then created a new value (“ds_num”) that only selects a certain set of columns which we will use during the imputation. Using the “md.pattern” function from MICE, we check the missing data pattern of the selected columns. The output of the md.pattern function is illustrated in Figure 2, which is tailored to data featuring 20% missing values that follow the Missing Completely at Random (MCAR) pattern in the "WNR" and "NPW" columns. The lower right-hand corner of the figure indicates the total number of missing values in the dataset, revealing that both the "WNR" and "NPW" columns have 50 missing values each.


Figure 2: Matrix explaining the missing data patterns in the data set.

The md.pattern function is a good first step to check if the data set you are dealing with even has missing values. If our data set had thousands of rows and columns, this function is a good approach to understand which columns have and how much missing data.

We create a new data frame in R that will hold 3 new values. It will include the original distribution of the data (including the MCAR values) and three other distributions of the same variable but with 3 imputation methods accessed through the MICE package. The first method, “pmm,” stands for predictive mean matching. Predictive mean matching is a method that imputes missing values by building a predictive model using other observed values and then matches the predicted values to the observed values based on their mean. The matched values would then replace the missing values in the data set. The second method, “cart,” stands for Classification and Regression Trees (CART) and the last method used in our analysis is the “norm” method. The norm method in MICE uses Bayesian linear regression to impute the missing values.

[5] www.stat.columbia.edu/~gelman/arm/missing.pdf.

After we use the specific methods in MICE to impute the missing data, we now look at the distributions of the WNR variable after the three imputation methods were applied and compare them to the original distribution. We examined the distributions of the variable and selected the method that most accurately replicates the original distribution for WNR. Based on these graphs, we chose the cart method to impute the missing values.

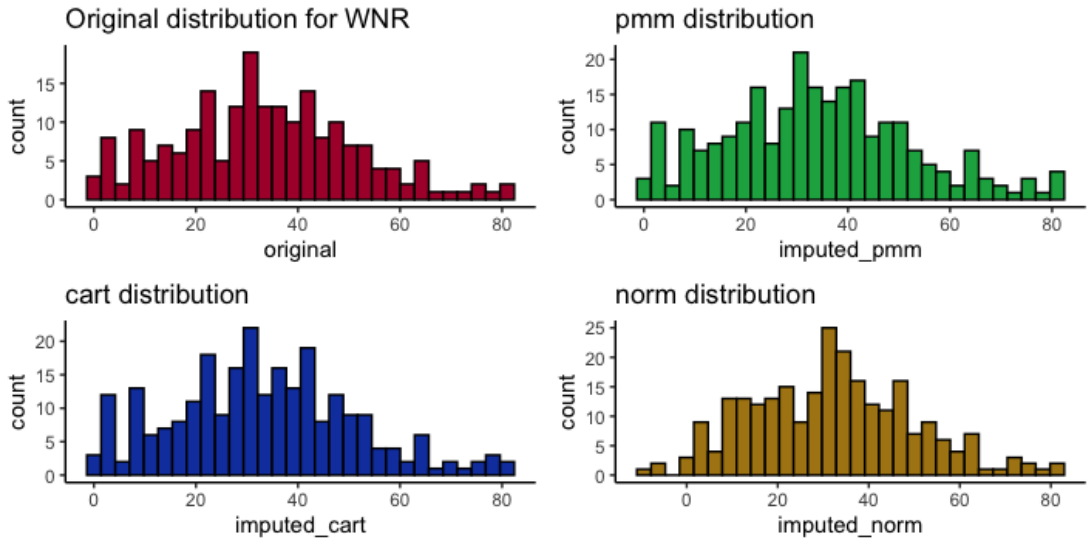


Figure 3: Comparing the original distribution with the resulting distributions from MICE imputed imputation methods for the “WNR” variable.

We want to double check to make sure we do not have any more missing values in the “WNR” column since we have finished the imputation for that column. Using the md.pattern function again, we can check this (Found in Figure 4).

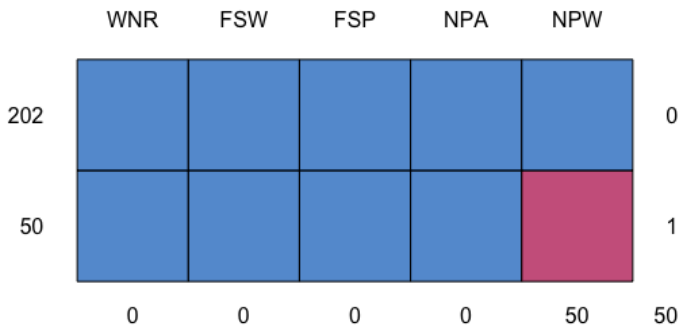


Figure 4: Matrix explaining the missing data patterns in the data set after dealing with the MCAR values for the “WNR” variable.

According to Figure 4, the imputation for the “WNR” column has been successful. There still remain missing values for the “NPW" column. Now we perform the same operations as we did

for imputing missing values for the “WNR” column, but now we do it for the “NPW” column. Figure 5 shows the distributions for the “NPW” variable. Again, we use the “cart” method to impute the missing values in the “NPW” because the distributions look similar. It is important to point out how the norm distribution now includes values that are less than 0. This does not make sense for our data set, so it is clear that we should not pick that method for the imputation.

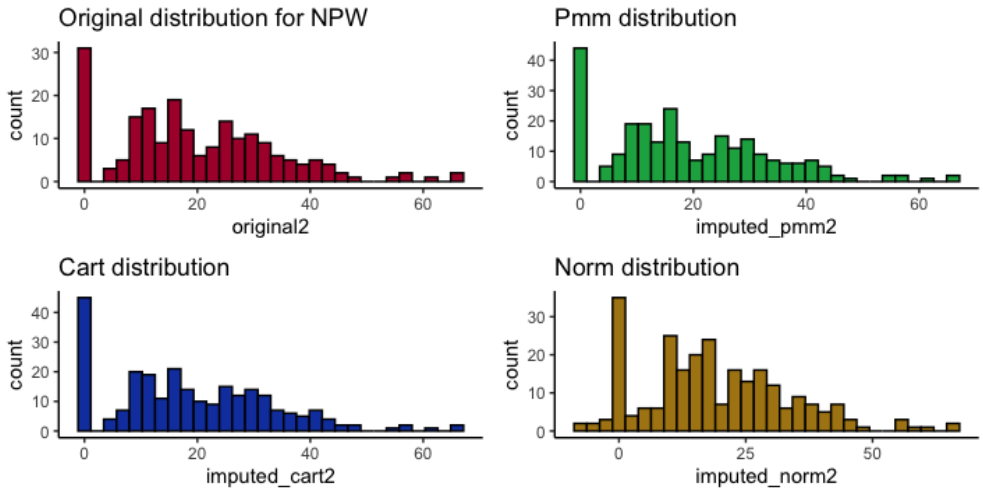


Figure 5: Comparing the original distribution with the resulting distributions from MICE imputed imputation methods for the “NPW” variable.

After dealing with all the MCAR values in our data set, we will lastly double check that there are no more missing values. We perform the md.pattern again to check this. Figure 6 shows that there are no missing values in our data set since there is no pinkish color in the output and there are all zero missing values for all 252 observations. This figure is difficult to read because we have so many columns that it could not format properly. This is also a reason why we came up with the “ds_num” value so that the output is easier to read. The console should also print out a statement explicitly telling you there are no missing values. This output is found in Figure 7. We are officially done dealing with MCAR values and can now run the two hypotheses on the complete data set.

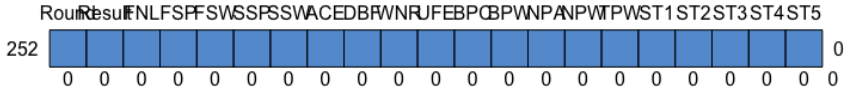

Figure 6: Matrix explaining the missing data patterns in the data set after dealing with the MCAR values for both variables. There are no more missing values.

```
> md.pattern(ds_complete)
 /\     /\
{  `---'  }
{  0   0  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'

    Round Result FNL FSP FSW SSP SSW ACE DBF WNR UFE BPC BPW NPA NPW TPW ST1 ST2 ST3 ST4
252     1      1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
        0      0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    ST5
252   1 0
      0 0
```

Figure 7: R console output notifying the user that there are no more missing values in the data set.

### C. Imputing MNAR Values

For MNAR values, the probability of missing data depends on the unobserved (missing) data itself, and missingness is related to the missing value of the variable. We set the probability of missingness for the missing values to 0.2, to keep consistency with MCAR data. For the MNAR data, we first created MNAR values by using a function called “delete_MNAR_censoring” from the “missMethods” library, and used to create missing values in WNR and NPW variables.

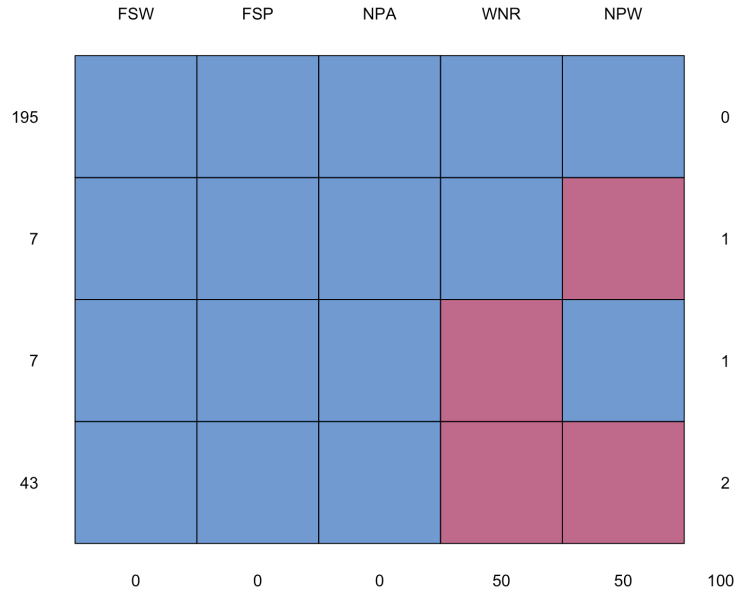


Figure 8: Matrix explaining the missing patterns in the data set.

We can see that the columns “WNR” and “NPW” contains 50 missing values each.

Similar to what we did in MCAR values, we use the MICE package to perform multiple imputation to our MNAR values, with three methods.

The code imputes the data using “pmm”, “cart”, “rf” methods of multiple imputation, and outputs visualized data of the original distribution of “WNR”, and results of those methods. The “pmm” and “cart” are the same as mentioned above, but the new “rf” method uses multiple decision trees to make predictions for missing values, and the final imputation is a summary measure based on the predictions of these trees.

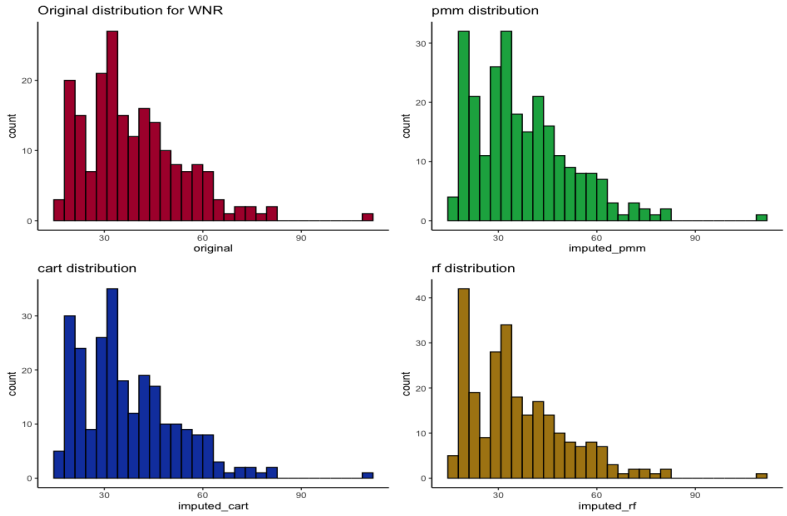


Figure 9: Comparing the original distribution with the resulting distributions from MICE imputed imputation methods for the “WNR” variable.

Looking at Figure 9, it seems the “pmm” method has a better representation of the original data set of WNR than other methods. So, we can update the dataset with the new imputed “WNR” values with the “pmm” method.

Then we move on to the new missing values in “NPW”. Like the procedure we did for the “WNR”, we create a plot to compare the imputed results between three methods. The codes impute missing values in the "WNR" column using three different methods: “pmm”, “cart”, and “rf”.

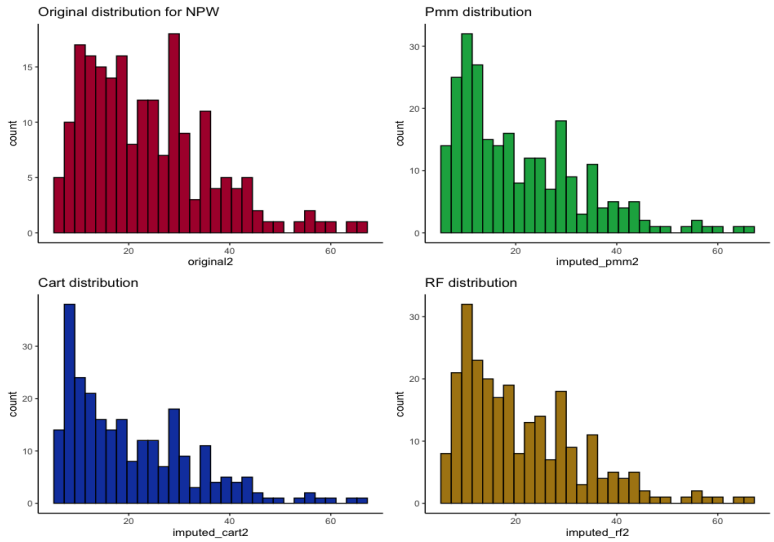


Figure 10: Comparing the original distribution with the resulting distributions from MICE imputed imputation methods for the “NPW” variable.

Looking at Figure 10, it seems that the “pmm” method has a better representation of the original data set of “NPW” as well, so we select the “pmm” method to impute the missing values in our “NPW” column. We update our data set “ds_miss_MNAR” with the new imputed column of “NPW”, and finalize our data set with MNAR data. To check the completeness of our final data set whether we have data that are still missing, we used the function md.pattern to visualize our final data set. Our final data set looks complete, and we can confirm that we are done dealing with our MNAR values in our data set.

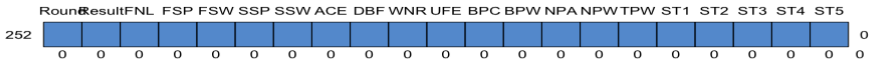


Figure 11: Matrix explaining the missing data patterns in the data set after dealing with the MCAR values for both variables. There are no more missing values.

## IV. First Hypothesis: Correlation

For the first hypothesis, we are interested in finding out if any of the sample correlation coefficients had any statistical significance in the population. Since our dataset contains the match statistics for each player in the 2014 men’s singles Australian Open, the population in this case is the match statistics for each player across all of the 4 different tournaments. To do this analysis, we performed correlation tests. The variables that we were interested in were the winners earned by players and the net points won by players in the tournament. In tennis, a winner is when a player hits the ball to the opponent’s side and the ball was not reached or hit back by the opposing player, thus granting the player who made the shot the point.[6] Net points are points that are won or lost by players that approach the net on the tennis court, instead of earning a point by hitting the ball from the baseline of the court.[7] We performed four different correlation hypothesis tests split into two different groups: the first group of tests were conducted with missing values created from MCAR and MNAR; the second group of tests were conducted with the missing values imputed from the MCAR and MNAR data sets. We will look at the results of the first group of hypothesis tests.

When analyzing the tennis match dataset with missing values in the “NPW” and “WNR” variables using MCAR and MNAR, we first checked the assumptions required for the correlation

[6] https://en.wikipedia.org/wiki/Glossary_of_tennis_terms
[7] https://en.wikipedia.org/wiki/Glossary_of_tennis_terms

tests. One key assumption when performing a correlation test (particularly for the Pearson correlation) is whether the data we have is normally distributed. To check this assumption, we looked at the Q-Q plots for the “NPW” and “WNR” variables to see if the data appeared linear. See Figure 12 and Figure 13.

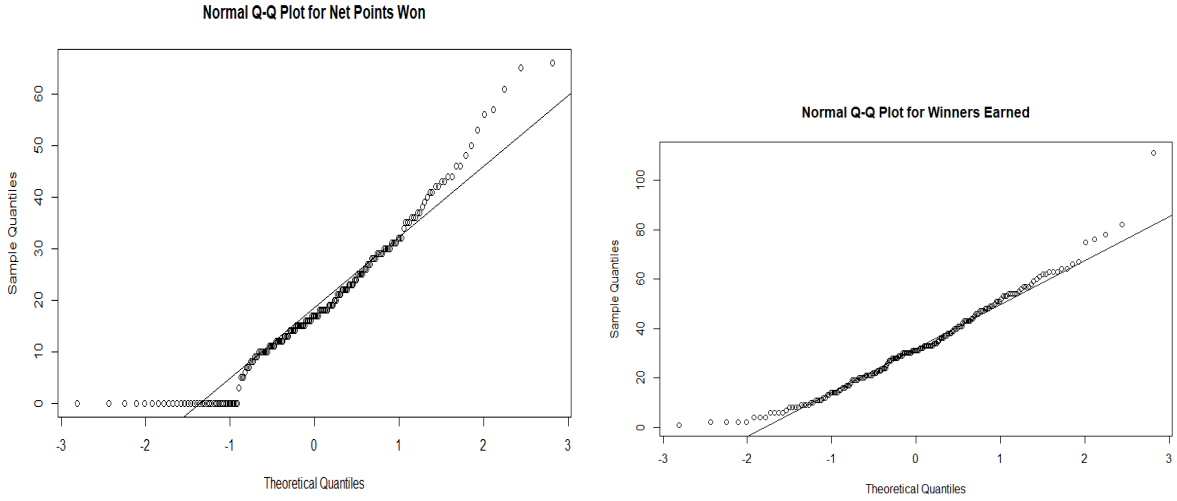


Figure 12 Normal Q-Q plots of “NPW” and “WNR” for the dataset with MCAR missing values.

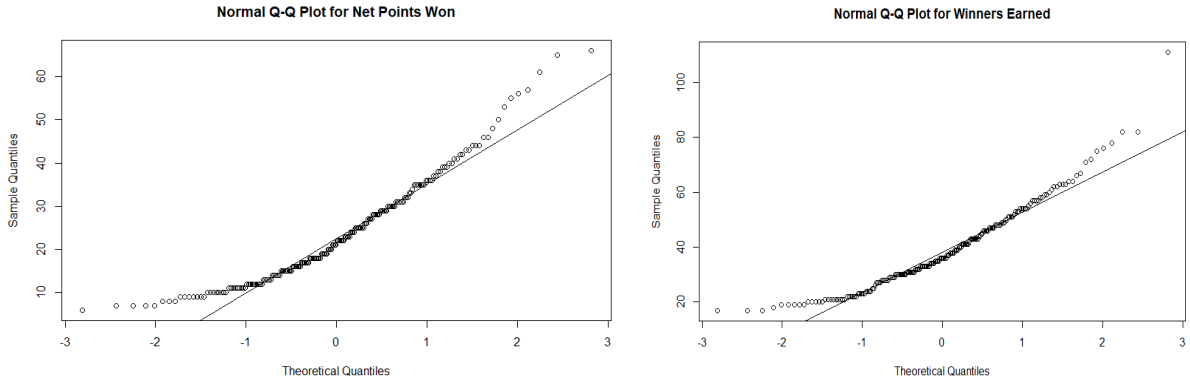


Figure 13 Normal Q-Q plots of “NPW” and “WNR” for the dataset with MNAR missing values.

Based on the normal Q-Q plots for the datasets with MCAR and MNAR missing values, we can see from the curvature of the data that these random variables do not appear very linear, which indicates that the normal distribution assumption is violated. These plots are also very similar to the Q-Q plots for “NPW” and “WNR” where we imputed the missing values for MCAR and MNAR. Since we do not want to rely on the Q-Q plots alone, we also looked at the results of the Shapiro-Wilk normality test. These tests were done on lines 231-232 and 316-317 for “NPW” and “WNR.” The p-values from the Shapiro test for the MCAR missing values dataset are 1.509e-07 and 0.0002459 for “NPW” and “WNR,” respectively. For the MNAR missing values dataset, the p-values are 2.627e-08 and 5.166e-08 for “NPW” and “WNR,” respectively.

In addition, when looking at the Shapiro tests for “NPW” and “WNR” in the MCAR and MNAR datasets where we imputed missing values, all of the resulting p-values are similarly very small. Since each of these p-values from the Shapiro tests are quite low (below $\alpha = 0.01$), we reject the null hypothesis for the Shapiro tests at the 0.01 significance level, meaning that there is sufficient evidence to assume that the “NPW” and “WNR” random variables are not normally distributed.

After checking this assumption, there are a few other assumptions that must be verified.[8] One is that our data must be interval or ratio. Ratio data is data that has a defined zero point, and interval data is data that lacks a defined zero point.[9] Since our “winners earned” and “net points won” variables cannot go below 0, we know that our data is ratio data, which satisfies this assumption. Another assumption is that the data must be linear or approximately linear. This assumption is checked later on in the analysis where we look at the correlation of “NPW” and “WNR” using scatter plots. One additional assumption is that the data should not contain too many outliers, since outliers can heavily impact the results of the sample correlation coefficients and the correlation tests. There are a few outliers in the MCAR and MNAR missing values datasets (<10 for both datasets), but these outliers shouldn’t make a significant impact on the correlation tests and sample correlation coefficients.

We proceeded with calculating the sample correlation coefficients for “NPW” and “WNR.” We used three different methods to calculate the sample correlation coefficients: Pearson, Spearman, and Kendall. The resulting sample correlation coefficients for the MCAR missing values dataset are $r = 0.687$, $r = 0.759$, and $r = 0.5755$ using Pearson, Spearman, and Kendall, respectively. Similarly, with the MNAR missing values dataset, the r values are $r = 0.453$, $r = 0.471$, and $r = 0.3325$ using Pearson, Spearman, and Kendall, respectively. These sample correlation coefficients are very similar to the coefficients calculated when the missing values are imputed as well. It is interesting to note that generating MNAR missing values made quite a noticeable impact on the sample correlation coefficients, going from a moderately positive linear relationship in MCAR to a weak positive linear relationship in MNAR. It is also important to determine which sample correlation coefficient is the most accurate for our dataset. Since the Pearson correlation method requires the assumption that the data is normally distributed, we did not use this method for the correlation hypothesis test. The Kendall method is best suited for data that is smaller and has many tied ranks from the Spearman method.[10] Since our data is not small and did not have too many tied ranks, we used the Spearman method to calculate the sample correlation coefficient and proceed with the correlation test.

In order to visualize the strength of the correlation in the sample, we looked at the scatter plots between “NPW” and “WNR” for the MNAR and MCAR datasets. See Figure 14, which shows the MCAR dataset scatter plot (first plot) and the MNAR dataset scatter plot (second plot).

[8] https://ademos.people.uic.edu/Chapter22.html

[9] https://www.cvent.com/en/blog/events/data-types-interval-ratio-data#:~:text=Ratio%20data%20has%20a%20defined,a%20reference%20point%20for%20measurement.

[10] https://towardsdatascience.com/kendall-rank-correlation-explained-dee01d99c535

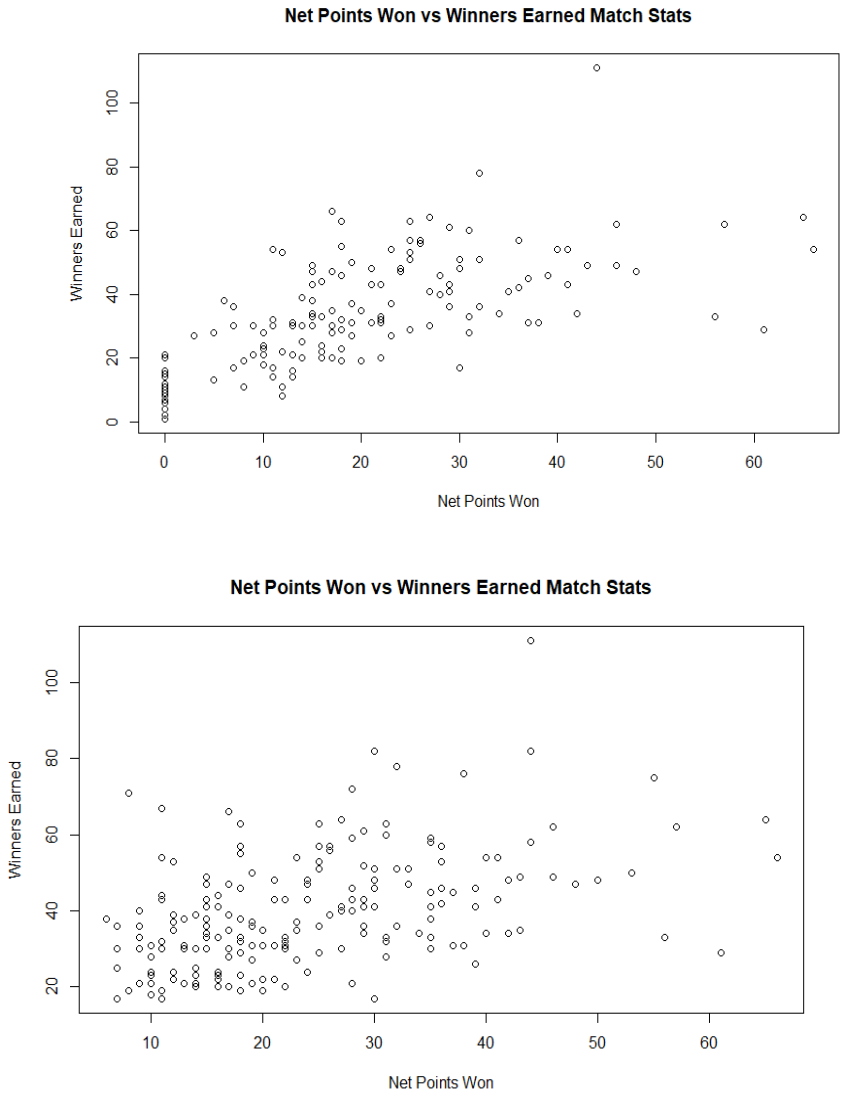


Figure 14 Scatter plots of "NPW" and "WNR" for the datasets with MCAR & MNAR missing values.

We also looked at the scatter plots for the MCAR and MNAR datasets where the missing values were imputed. See Figure 15, in which the first plot shows the relationship for the imputed MCAR values and the second plot shows the relationship for the imputed MNAR values.

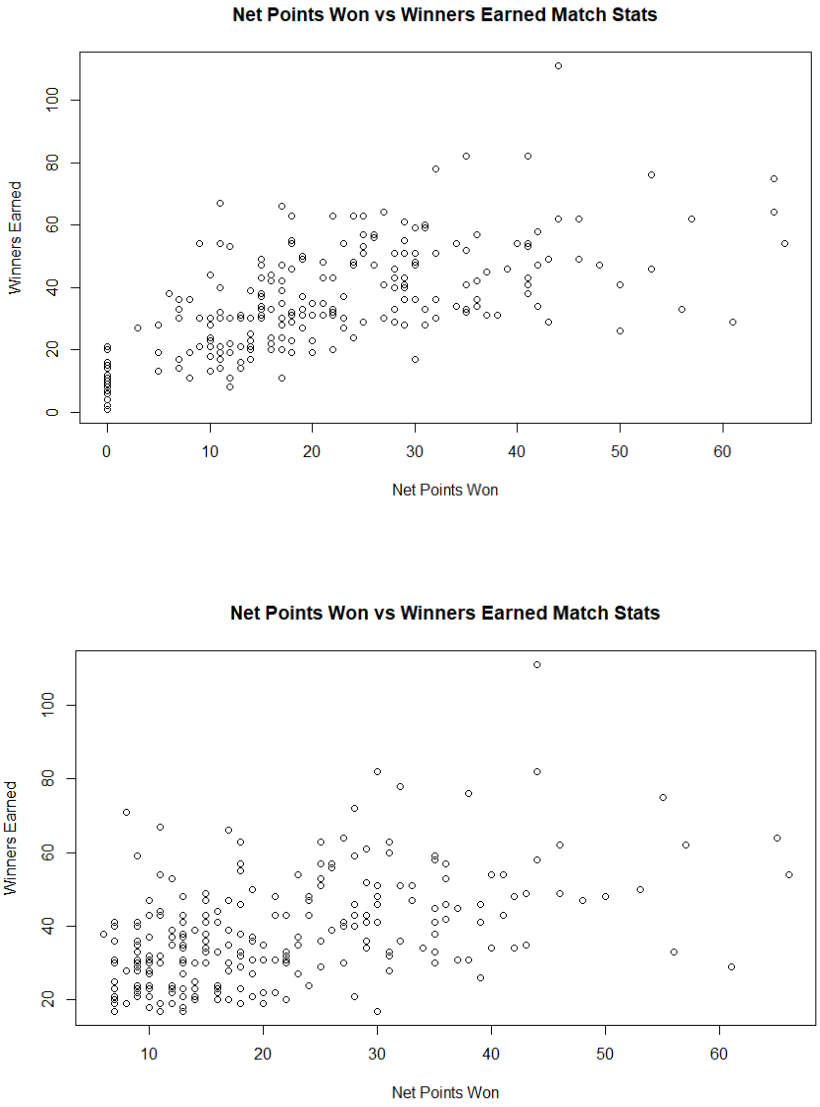


Figure 15 Scatter plots of “NPW” and “WNR” for the datasets with MCAR & MNAR imputed missing values.

We can see from the scatterplots that the linear relationship between “NPW” and “WNR” makes sense with the sample correlation coefficients we have calculated. The scatter plots with MCAR data show that there is a stronger positive linear relationship between net points won and winners earned in tennis matches, whereas the scatter plots with MNAR data show that the positive linear relationship between the two variables is weaker.

Now that we have checked our assumptions, determined the sample correlation coefficient that fit best with our dataset, and looked at the scatter plots to visualize the correlation, we then performed correlation tests using the Spearman method. We also performed correlation tests using the Kendall method to compare the results with the Spearman

method, since the Kendall method can be used when the dataset is not normally distributed. For the Spearman correlation hypothesis test, the hypotheses are:

$$H_0: \rho = 0 \; vs \; H_1: \rho \neq 0$$

For the Kendall correlation hypothesis test, the hypotheses are:

$$H_0: \tau = 0 \; vs \; H_1: \tau \neq 0$$

The results are listed below:

*Missing values (not yet imputed) for MNAR dataset*

```
Kendall's rank correlation tau

data:  ds_miss_MNAR$NPW and ds_miss_MNAR$WNR
z = 6.7557, p-value = 1.421e-11
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.3324868

Spearman's rank correlation rho

data:  ds_miss_MNAR$NPW and ds_miss_MNAR$WNR
S = 653437, p-value = 3.581e-12
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.471235
```

*Missing values (not yet imputed) for MCAR dataset*

```
Kendall's rank correlation tau

data:  ds_miss$NPW and ds_miss$WNR
z = 10.367, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.5602847

Spearman's rank correlation rho

data:  ds_miss$NPW and ds_miss$WNR
S = 192747, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7378055
```

*Missing values imputed for MNAR dataset*

```
Kendall's rank correlation tau
```

```
data:  ds_complete_MNAR$NPW and ds_complete_MNAR$WNR
z = 7.9146, p-value = 2.481e-15
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.3445739

Spearman's rank correlation rho

data:  ds_complete_MNAR$NPW and ds_complete_MNAR$WNR
S = 1359419, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4903057
```

*Missing values imputed for MCAR dataset*

```
Kendall's rank correlation tau

data:  ds_complete$NPW and ds_complete$WNR
z = 12.833, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.5588389

Spearman's rank correlation rho

data:  ds_complete$NPW and ds_complete$WNR
S = 700946, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7371905
```

Using the p-values for the Spearman correlation tests where the missing values are imputed, we have that the p-values for both MNAR and MCAR are p-value < 2.2e-16. Since these p-values are both less than $\alpha = 0.01$, we reject the null hypothesis at the 0.01 significance level. This means that we have sufficient evidence that the population correlation coefficient for winners earned and net points won is statistically significant (meaning that the population correlation coefficient is different from 0). It is important to note that the missing values, regardless of whether they were imputed or not, had very little effect on the decision for our correlation tests. In both cases, we reject the null hypothesis since the p-values from the Spearman test are very small.

## V. Second Hypothesis: Logistic Regression

A useful application of a regression model could be the use of the significance of coefficients to understand the relationship between dependent and independent variables.

Logistic regression models are generally used for binary responses including classification. This model computes the probability of one of the two outcomes as a function of predictor variables[11]. For example, outcome of a tennis match for a player may depend on several factors such as First Serve Percentage (FSP), First Serve Won (FSW), Second Serve Percentage (SSP), Second Serve Won (SSW), Aces won (ACE), Double Faults committed (DBF), Winners earned (WNR), Unforced Errors committed (UFE), Break Points Created (BPC), Break Points Won (BPW), Net Points Attempted (NPA), Net Points Won (NPW) by the player, and so on. Using a logistic regression model, we can test whether all the independent variables available in the dataset are significant or not. We formulated our second test hypothesis as follows:

$$H_0: All\ \beta = 0\ vs\ H_a: Not\ H_0$$

Our dataset supports the most important assumption of a binary logistic regression which is – each match has two outcomes: either player 1 or player 2 wins the match. Therefore, each match statistic was divided into two and treated as two samples each one associated with the corresponding player. The performance (win/lose) of the player is the dependent variable and other player specific statistics are independent variables. Such data structure makes it suitable for a logistic regression model. However, an important requirement for a logistic regression model is to choose only the meaningful independent variables. Although we fitted all the variables into the model at the beginning, we removed some variables based on further analysis. For example, Second Serve Percentage (SSP) is omitted due to multicollinearity issues. This variable has a perfect inverse correlation (-1.0) with the First Serve Percentage (FSP) variable (see figure below).

[11] Tamhane, Ajit, and Dunlop Dunlop. "Statistics and data analysis: from elementary to intermediate." (2000).

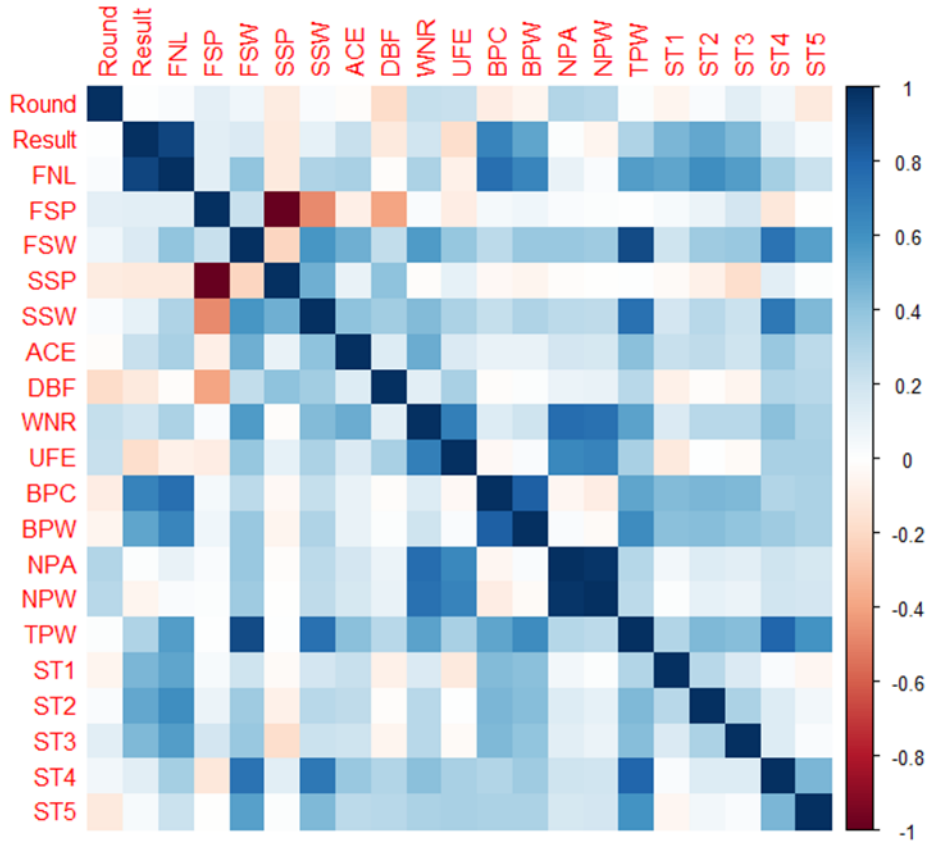


Figure 14: Correlation among all the variables

Multicollinearity can cause serious numerical and statistical difficulties in fitting the regression model unless “extra” predictor variables are deleted. Hence, we looked at the correlation matrix closely and treated the multicollinearity issue. In the previous section, we discussed that the data sometimes deviate from the normality assumption, hence, we utilized Spearman's rank correlation coefficient test to obtain the correlation matrix.

We also removed two variables – Final Number of Games Won by the player (FNL) for nearly perfect correlation, and Round of the tournament at which game is played (Round) for almost no correlation with the dependent variables. When one independent variable has nearly perfect correlation with the dependent variable it causes other independent variables to become independent and thus, it leads to a poor modeling. In addition, set specific results (e.g., set 1, set 2, etc.) are also ignored for two reasons – some of them are insignificant and not all these variables apply to all the games. Thus, we carefully selected important and meaningful variables and avoided any severe multicollinearity issues in our logistic regression model. According to Tamhane & Dunlop, for multiple covariates, $x_1, x_2, \ldots\ldots\ldots, x_k$, the logistic regression model is –

$$\ln\left\{\frac{P(Y=1|x_1,x_2,\ldots\ldots\ldots,x_k)}{P(Y=0|x_1,x_2,\ldots\ldots\ldots,x_k)}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Results from the model using data with missing values and imputed/simulated values are not very different. The following table lists all the independent variables used in the models and their corresponding significance (Y or N):

Table: Significance (Y or N) summary of the independent variables across different models

| Variables | Data w/MCAR | MCAR Data Simulated | Data w/MNAR | MNAR Data Simulated |
|---|---|---|---|---|
| FSP | Y | Y | Y | Y |
| FSW | N | N | N | N |
| SSW | Y | Y | Y | Y |
| ACE | Y | Y | Y | Y |
| DBF | N | N | N | N |
| WNR | Y | Y | N | Y |
| UFE | Y | Y | N | Y |
| BPC | Y | Y | Y | Y |
| BPW | N | N | N | N |
| NPA | Y | Y | Y | Y |
| NPW | N | Y | Y | Y |
| TPW | Y | Y | Y | Y |

The scenario we simulated was 20% missing data for both MCAR & MNAR cases. Interestingly, only Missing Not At Random (MNAR) data results vary slightly where WNR, UFE variables become not statistically significant at $\alpha=0.01$. Most of the variables are significant at $\alpha=0.01$ level. Also, the p-value for the overall Chi-Square statistic of the model turns out to be 0 for all scenarios. Therefore we reject the null hypothesis ($H_0$:All $\beta=0$). The R output results are shown below:

```
Call:
glm(formula = Result ~ FSP + FSW + SSW + ACE + DBF + WNR + UFE +
    BPC + BPW + NPA + NPW + TPW, family = "binomial", data = ds_miss_logit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.657564   3.750204  -2.575 0.010018 *
FSP          0.119775   0.058714   2.040 0.041353 *
FSW          0.025619   0.054021   0.474 0.635330
SSW          0.157418   0.068743   2.290 0.022024 *
ACE          0.054431   0.044708   1.217 0.223415
DBF          0.021341   0.070003   0.305 0.760477
WNR          0.089931   0.024838   3.621 0.000294 ***
UFE         -0.070583   0.016447  -4.291 1.77e-05 ***
BPC          1.179213   0.198222   5.949 2.70e-09 ***
BPW          0.089283   0.079272   1.126 0.260046
NPA         -0.006268   0.040749  -0.154 0.877745
NPW          0.007816   0.019764   0.395 0.692498
TPW         -0.077548   0.036307  -2.136 0.032689 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 349.35  on 251  degrees of freedom
Residual deviance: 162.48  on 239  degrees of freedom
AIC: 188.48

Number of Fisher Scoring iterations: 6

> 1-pchisq(349.35-162.48, 251-239)
[1] 0
```

Figure 15: Results from dataset with missing completely at random values.

```
Call:
glm(formula = Result ~ FSP + FSW + SSW + ACE + DBF + WNR + UFE +
    BPC + BPW + NPA + NPW + TPW, family = "binomial", data = ds_complete_logit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.091382   3.776331  -2.407 0.016064 *
FSP          0.116613   0.059508   1.960 0.050040 .
FSW          0.056211   0.057577   0.976 0.328925
SSW          0.190258   0.070344   2.705 0.006837 **
ACE          0.073446   0.047066   1.561 0.118640
DBF          0.009614   0.071441   0.135 0.892952
WNR          0.092589   0.025612   3.615 0.000300 ***
UFE         -0.059811   0.017152  -3.487 0.000488 ***
BPC          1.282934   0.216535   5.925 3.13e-09 ***
BPW          0.093904   0.080958   1.160 0.246088
NPA          0.290459   0.096866   2.999 0.002712 **
NPW         -0.209999   0.067671  -3.103 0.001914 **
TPW         -0.105921   0.039277  -2.697 0.007001 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 349.35  on 251  degrees of freedom
Residual deviance: 150.79  on 239  degrees of freedom
AIC: 176.79

Number of Fisher Scoring iterations: 6

> 1-pchisq(349.35-150.79, 251-239)
[1] 0
```

Figure 16: Results from dataset with projected data for missing completely at random values.

```
Call:
glm(formula = Result ~ FSP + FSW + SSW + ACE + DBF + WNR + UFE +
    BPC + BPW + NPA + NPW + TPW, family = "binomial", data = ds_miss_logit_MNAR)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.395953   3.842655  -3.226  0.00126 **
FSP           0.156077   0.060283   2.589  0.00962 **
FSW           0.042949   0.057057   0.753  0.45160
SSW           0.204361   0.073449   2.782  0.00540 **
ACE           0.136200   0.046988   2.899  0.00375 **
DBF          -0.009851   0.070834  -0.139  0.88939
WNR           0.045344   0.025496   1.779  0.07532 .
UFE          -0.030522   0.015918  -1.917  0.05518 .
BPC           1.295692   0.215917   6.001 1.96e-09 ***
BPW           0.099423   0.079534   1.250  0.21128
NPA           0.464273   0.106657   4.353 1.34e-05 ***
NPW          -0.294275   0.072847  -4.040 5.35e-05 ***
TPW          -0.103062   0.038704  -2.663  0.00775 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 349.35  on 251  degrees of freedom
Residual deviance: 147.86  on 239  degrees of freedom
AIC: 173.86

Number of Fisher Scoring iterations: 6

> 1-pchisq(349.35-147.86, 251-239)
[1] 0
```

Figure 17: Results from dataset with missing not at random values.

```
Call:
glm(formula = Result ~ FSP + FSW + SSW + ACE + DBF + WNR + UFE +
    BPC + BPW + NPA + NPW + TPW, family = "binomial", data = ds_complete_logit_MNAR)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.95097    3.79092  -2.625  0.00867 **
FSP          0.13491    0.06003   2.247  0.02462 *
FSW          0.04325    0.05952   0.727  0.46745
SSW          0.18636    0.07218   2.582  0.00982 **
ACE          0.12107    0.04607   2.628  0.00859 **
DBF          0.06214    0.07463   0.833  0.40502
WNR          0.05384    0.02551   2.111  0.03481 *
UFE         -0.06788    0.01666  -4.074 4.62e-05 ***
BPC          1.23366    0.21363   5.775 7.70e-09 ***
BPW          0.08534    0.08300   1.028  0.30386
NPA          0.38954    0.08413   4.630 3.65e-06 ***
NPW         -0.27004    0.06262  -4.312 1.61e-05 ***
TPW         -0.09072    0.04016  -2.259  0.02388 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 349.35  on 251  degrees of freedom
Residual deviance: 144.64  on 239  degrees of freedom
AIC: 170.64

Number of Fisher Scoring iterations: 6

> 1-pchisq(349.35-144.64, 251-239)
[1] 0
```

Figure 18: Results from dataset with projected data for missing not at random values.

## VI. Conclusions

For the first hypothesis, we have that the p-value is less than the 0.01 significance level, so we reject the null hypothesis. So, we have sufficient evidence to claim that the winners earned and net points won for all 4 of the tournaments (the population) have a significant positive correlation or linear relationship. There are two important notes pertaining to this conclusion. The first is that correlation is not causation. It is a common mistake to say that when two variables are significantly correlated, one variable causes the other variable to change. This is not necessarily true. The second note is that even though our analysis found that winners earned and net points won have a significant correlation in the population, there could be other factors or “lurking variables” that were not considered in our analysis that impact this correlation. For instance, each tennis tournament has different tennis court surfaces, such as grass, clay, and hard court, which could affect the correlation for the two variables. For the second hypothesis, since this p-value is less than .05, we reject the null hypothesis. In other

words, there is a statistically significant relationship between the match result i.e. performance of the player and the selected independent variables.