Project Summary Report

Matthew Pool

matthew.pool@snhu.edu

Southern New Hampshire University

October 16, 2022

Project Summary Report

The nba_wins_data.csv file, based on the FiveThirtyEight NBA Elo historical dataset (FiveThirtyEight, 2019), will be used as the source for this statistical analysis, with the intent to predict the number of NBA (National Basketball Association) regular-season wins, based on average points and average Elo (relative skill level) of the teams, as well as the differences between them. This study focuses on data from 1995 to 2015 for all teams in the NBA. Analyses used in this study include descriptive, inferential, and prescriptive statistical methods – including simple linear regression (SLR) and multiple regression.

It should be noted that Elo includes final scores, game location (home-court or away), and outcome of the game relative to the probability of that outcome. Wins with wider margins are worth more points in this zero-sum system, in which the winning team gains points equivalent to the points the other team loses (Silver and Fischer-Baum, 2015). See Table 1 for more information on the variables being studied.

Table 1

Dataset variables that will be examined (based on regular-seasons only)

Variable	Interpretation
total_wins	Total number of wins in a season
avg_pts	Average points scored in a game
avg_elo_n	Average relative skill of team compared to others
avg_pts_differential	Difference between average points scored and points allowed
avg_elo_differential	Difference between a team's Elo and their opponent's Elo

Data visualization can be used to easily identify correlations and relationships between predictor (independent) variables and a response (dependent) variable. A regression line (trend line) can be used to find the “best-fit” relationship trend of a set of data points. The slope of this line is the change in *Y* (response variable) for every one-unit change in *X* (predictor variable).

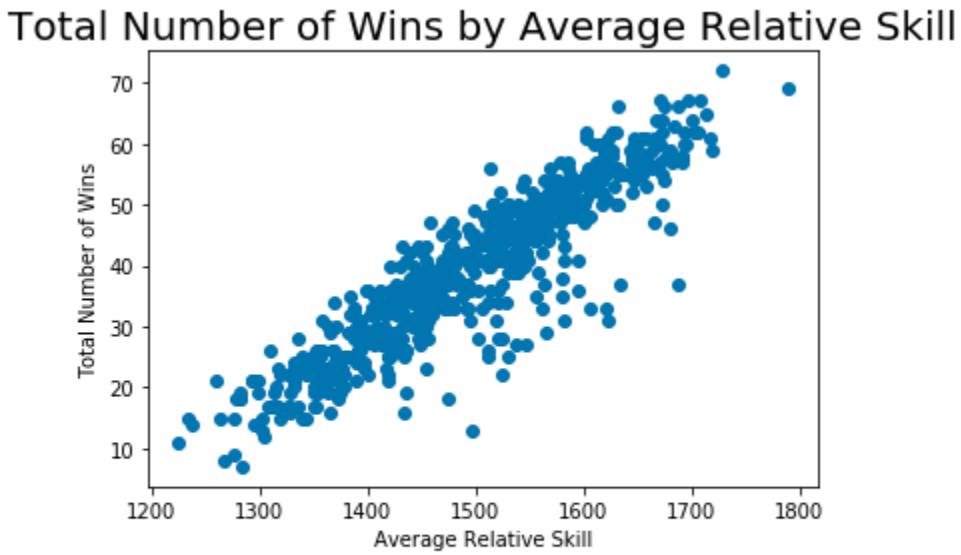
The Pearson correlation coefficient, *R*, is a measure of the strength and direction (positive or negative) of the association between two variables. This coefficient, along with the probability value (*p*-value), can determine statistical significance in the relationship. Pearson’s *R* is a value between negative one and positive one, with zero indicating no correlation.

Simple Linear Regression

One Predictor Variable

Model 1: Average Relative Skill (Elo)

Figure 1




```
Correlation between Average Relative Skill and the Total Number of Wins
Pearson Correlation Coefficient = 0.9072
P-value = 0.0
```

Looking at the scatterplot (Figure 1) created in Python (see attached HTML named *Project Three Jupyter Notebook*), there appears to be a positive correlation. In other words, as *X* increases, so does *Y*. Because the Pearson correlation coefficient is quite high ($R = .91$), a very strong (positive) correlation between average relative skill and total number of wins is confirmed. Comparing the level of significance of 1% ($\alpha = .01$) to the calculated *p*-value ($p < .001$), it is determined that this correlation is statistically significant.

A simple linear regression (SLR) model estimates the relationship between two quantitative variables, using a straight line as the best representation of that relationship. This allows estimation of how a response variable changes, as the independent variable changes and

shows the strength of that relationship. Plugging in the intercept (β_0) and the independent variable values into the equation for this line, the "best-guess" value of the response variable can be easily determined.

Table 2

Simple linear regression results

OLS Regression Results

Dep. Variable:	total_wins	R-squared:	0.823
Model:	OLS	Adj. R-squared:	0.823
Method:	Least Squares	F-statistic:	2865.
Date:	Tue, 11 Oct 2022	Prob (F-statistic):	8.06e-234
Time:	18:54:40	Log-Likelihood:	-1930.3
No. Observations:	618	AIC:	3865.
Df Residuals:	616	BIC:	3873.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-128.2475	3.149	-40.731	0.000	-134.431	-122.064
avg_elo_n	0.1121	0.002	53.523	0.000	0.108	0.116

Omnibus:	152.822	Durbin-Watson:	1.098
Prob(Omnibus):	0.000	Jarque-Bera (JB):	393.223
Skew:	-1.247	Prob(JB):	4.10e-86
Kurtosis:	6.009	Cond. No.	2.14e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.14e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Using the ordinary least squares (OLS) regression results (Table 2), the equation of the line can be found to be as follows:

$$\hat{Y} = \beta_0 + \beta_1 X_1$$

$$\hat{Y} = -128.2475 + 0.1121X$$

$$\hat{Y} = -128.2475 + 0.1121(avg\_elo\_n) \quad (1)$$

Hypothesis Test

Is there evidence to support the claim that the number of wins an NBA team has during a regular-season positively correlates with mean relative skill level (Elo)?

Null Hypothesis

$$H_0: \beta_1 = 0$$

The regression coefficient (β_1) in this model is equal to zero, and the predictor variable (X_1), the team relative skill level, has no statistically significant relationship with the response variable (Y), the total number of regular-season game wins.

Alternative Hypothesis

$$H_a: \beta_1 \neq 0$$

The regression coefficient (β_1) in this model is not equal to zero, and the predictor variable (X_1), the team relative skill level, has a statistically significant relationship with the response variable (Y), the total number of regular-season game wins.

Level of Significance

$$\alpha = .01$$

Table 3

SLR model test statistic and p-value

Statistic	Value
F-statistic	2865.00 *(rounded to 2 decimal places)*
P(*F*-statistic)	8.0600e-234 *(rounded to 4 decimal places)*

Statistical Significance

$$p - value\text{: } (F = 2865, p < .001)$$

Since the *p*-value (probability value) of 8.0600e-234 is lower than the level of significance ($\alpha = .01$), the null hypothesis can be rejected, in favor of the alternative hypothesis. The regression coefficient, or slope coefficient, (β_1) in the SLR model is not equal to zero, and a statistically significant positive linear relationship exists between the response variable and the predictor variable and is not likely due to chance.

The coefficient of determination, R^2, is a value between zero and one and is known as the "goodness of fit" and explains how much variability of one factor is caused by its relationship to another. This model's coefficient of determination ($R^2 = .82$), indicating a very strong positive correlation between the predictor and response variables. This means approximately 82% of the data fits the regression model and can be used to predict the total number of wins in a regular-season.

Example 1

If average relative skill is 1550:

$$\hat{Y} = -128.2475 + 0.1121X$$

$$\hat{Y} = -128.2475 + 0.1121(1550)$$

$$\hat{Y} = 45.5075 \cong 45$$

Example 2

If average relative skill is 1450:

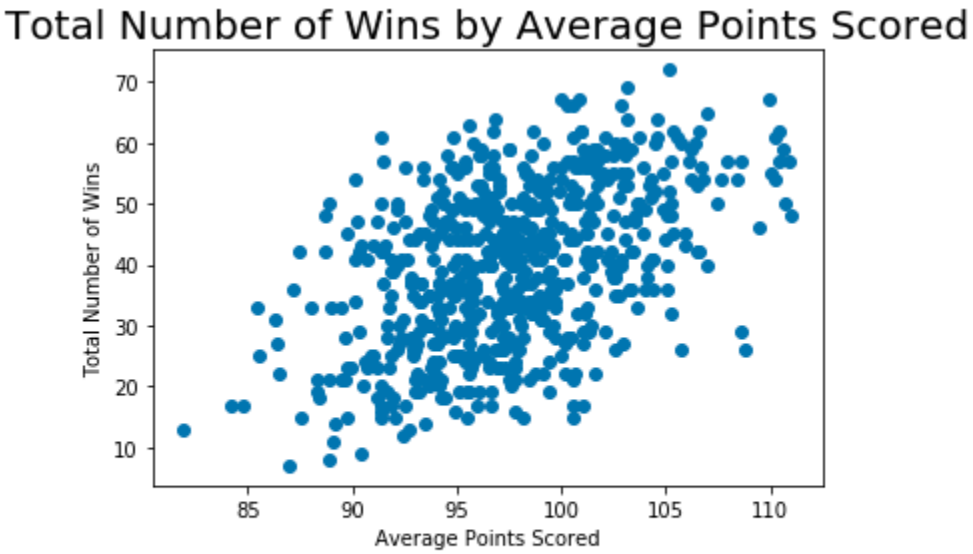
$$\hat{Y} = -128.2475 + 0.1121X$$

$$\hat{Y} = -128.2475 + 0.1121(1450)$$

$$\hat{Y} = 34.2975 \cong 34$$

Model 2: Average Total Points

Figure 2




```
Correlation between Average Points Scored and the Total Number of Wins
Pearson Correlation Coefficient = 0.4777
P-value = 0.0
```

Based on the scatterplot, there doesn't appear to be any pattern or correlation between the predictor variable (average points scored) and the response variable (total number of wins). The Pearson correlation coefficient ($R = .48$) further verifies this assumption, indicating a low correlation between the variables. Comparing the *p*-value ($p < .001$) to the level of significance ($\alpha = .01$), it is determined that this result is statistically significant.

Multiple Regression

A multiple regression model estimates the relationship between multiple quantitative (predictor) variables and a single quantitative response variable. This allows estimation of how a response variable changes, as the multiple independent variables change, as well as the strength of the relationships in between those variables. Plugging in the intercept (β_0) and the independent variable values $\{X_0, X_1, \dots, X_n\}$ into a multiple regression equation will result in the best estimation of the response variable (Y).

Model 3: Two Predictor Variables

Table 4

Multiple regression results

OLS Regression Results

Dep. Variable:	total_wins	R-squared:	0.837
Model:	OLS	Adj. R-squared:	0.837
Method:	Least Squares	F-statistic:	1580.
Date:	Tue, 11 Oct 2022	Prob (F-statistic):	4.41e-243
Time:	18:59:19	Log-Likelihood:	-1904.6
No. Observations:	618	AIC:	3815.
Df Residuals:	615	BIC:	3829.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-152.5736	4.500	-33.903	0.000	-161.411	-143.736
avg_pts	0.3497	0.048	7.297	0.000	0.256	0.444
avg_elo_n	0.1055	0.002	47.952	0.000	0.101	0.110

Omnibus:	89.087	Durbin-Watson:	1.203
Prob(Omnibus):	0.000	Jarque-Bera (JB):	160.540
Skew:	-0.869	Prob(JB):	1.38e-35
Kurtosis:	4.793	Cond. No.	3.19e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.19e+04. This might indicate that there are strong multicollinearity or other numerical problems.

In this model, the total number of wins is used as the response variable (Y), and the average points scored, as well as the average relative skill (Elo), will both be used as the predictor variables, X_1 and X_2 (respectively). The equation for this model is as follows:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\hat{Y} = -152.5736 + 0.3497 X_1 + 0.1055 X_2$$

$$\hat{Y} = (-152.5736) + (0.3497)(avg\_pts) + (0.1055)(avg\_elo\_n) \quad (2)$$

Is there evidence to support the claim that that the number of wins an NBA team has during a regular-season positively correlates with mean relative skill level (Elo) and average points scored?

Null Hypothesis

$$H_0: \beta_1 = \beta_2 = 0$$

The regression coefficients, β_1 and β_2, in this model are equal to zero, and the predictor variables, X_1 and X_2 (team relative skill level and average points scored, respectively), have no statistically significant relationship with the response variable (*Y*), the total number of regular-season wins.

Alternative Hypothesis

$$H_a: \text{at least one } \beta_i \neq 0, \text{for } i = 1, 2$$

At least one of the regression coefficients, β_1 and β_2, in this model are not equal to zero, and the predictor variables, X_1 and X_2 (team average relative skill level and average points scored, respectively), have a statistically significant relationship with the response variable (*Y*), the total number of regular-season wins.

Level of Significance

$$\alpha = .01$$

Table 5

Hypothesis test for the Overall F-Test

Statistic	Value
F-statistic	1580.00 *(Rounded to 2 decimal places)*
P(*F*-statistic)	4.4100e-234 *(Rounded to 4 decimal places)*

Statistical Significance

$$p - value: (F = 1580, p < .001)$$

Since the *p*-value of the Overall *F*-test of 4.4100e-234 is much lower than the level of significance ($\alpha = .01$), the null hypothesis can be rejected, in favor of the alternative hypothesis. At least one of the regression coefficients (β_1 and β_2) in the multiple regression model is not

equal to zero, and a statistically significant positive linear relationship exists between the response variable and the predictor variables and is not likely due to chance.

Table 6

Hypothesis tests for the Student's t-test

	Average Points	**Average Elo**
t-statistic	7.297	47.952
p-value	0.000	0.000

Both predictor variables (X_1 and X_2) with regression coefficients of 0.3497 and 0.1055 (respectively) are statistically significant ($p < .001$) at a 1% significance level ($\alpha = .01$). The regression coefficients indicate that for every one unit change in average relative skill level or average points scored, the estimated number of total wins increases by 0.3497 and 0.1055 (respectively), as long as all other independent variables remain constant.

The coefficient of determination ($R^2 = .84$) for this multiple regression model indicates a very strong positive correlation between the variance of the response variable and the predictor variables.

Example 1:

If average total points is 75 and relative skill level is 1350:

$$\hat{Y} = (-152.5736) + 0.3497X_1 + 0.1055X_2$$

$$\hat{Y} = (-152.5736) + 0.3497(75) + 0.1055(1350)$$

$$\hat{Y} = 16.0789 \cong 16$$

Example 2:

If average total points is 75 and relative skill level is 1350:

$$\hat{Y} = (-152.5736) + 0.3497X_1 + 0.1055X_2$$

$$\hat{Y} = (-152.5736) + 0.3497(100) + 0.1055(1600)$$

$$\hat{Y} = 51.1964 \cong 51$$

Model 4: Four Predictor Variables

Table 7

OLS Regression Results

Dep. Variable:	total_wins	R-squared:	0.878
Model:	OLS	Adj. R-squared:	0.877
Method:	Least Squares	F-statistic:	1102.
Date:	Tue, 11 Oct 2022	Prob (F-statistic):	3.07e-278
Time:	19:01:36	Log-Likelihood:	-1815.5
No. Observations:	618	AIC:	3641.
Df Residuals:	613	BIC:	3663.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	34.5753	25.867	1.337	0.182	-16.223	85.373
avg_pts	0.2597	0.043	6.070	0.000	0.176	0.344
avg_elo_n	-0.0134	0.017	-0.769	0.442	-0.048	0.021
avg_pts_differential	1.6206	0.135	12.024	0.000	1.356	1.885
avg_elo_differential	0.0525	0.018	2.915	0.004	0.017	0.088

Omnibus:	193.608	Durbin-Watson:	0.979
Prob(Omnibus):	0.000	Jarque-Bera (JB):	598.416
Skew:	-1.503	Prob(JB):	1.14e-130
Kurtosis:	6.769	Cond. No.	2.11e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.11e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

In this multiple regression model (Table 7), total number of wins is again the response variable. The independent variables include average points scored, Elo, average points differential, and the Elo differential. The equation for this model is as follows:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

$$\hat{Y} = 34.5753 + 0.2597 X_1 - 0.0134 X_2 + 1.6206 X_3 + 0.0525 X_4$$

$$\hat{Y} = 34.5753 + 0.2597 avg\_pts - 0.0134 avg\_elo\_n + 1.6206 avg\_pts\_differential + 0.0525 avg\_elo\_differential \quad (3)$$

Is there evidence to support the claim that the number of wins an NBA team has during a regular-season positively correlates with mean points scored, Elo, average points differential, and Elo differential?

Null Hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

The regression coefficients ($\beta_1, \beta_2, \beta_3,$ and β_4) in this model are equal to zero, and the predictor variables, $X_1, X_2, X_3,$ and X_4 (team average Elo, average points, average Elo differential, and average points differential, respectively), have no statistically significant relationship with the response variable (*Y*), the total number of regular-season wins.

Alternative Hypothesis

$$H_a: \text{at least one } \beta_i \neq 0, \text{for } i = 1, 2, 3, 4$$

At least one of the regression coefficients ($\beta_1, \beta_2, \beta_3, \text{and } \beta_4$) in this model are not equal to zero, and the predictor variables, $X_1, X_2, X_3, \text{and } X_4$ (team average Elo, average points, average Elo differential, and average points differential, respectively), have a statistically significant relationship with the response variable (*Y*), the total number of regular-season wins.

Level of Significance

$$\alpha = .01$$

Table 8

Hypothesis test for the Overall F-Test

Statistic	Value
F-statistic	1102.00 *(Rounded to 2 decimal places)*
P(*F*-statistic)	3.0700e-278 *(Rounded to 4 decimal places)*

Statistical Significance

$$p - value: (F = 1102, p < .001)$$

Since the *p*-value of the Overall *F*-test of 3.07e-278 is much lower than the level of significance ($\alpha = .01$), the null hypothesis can be rejected, in favor of the alternative hypothesis. At least one of the regression coefficients ($\beta_1, \beta_2, \beta_3, \text{and } \beta_4$) in the multiple regression model is

not equal to zero, and a statistically significant positive relationship exists between the response variable and the predictor variables and is not likely due to chance.

Table 9

Hypothesis test for the Student's t-test

	Average Points	**Average Elo**	**Points Differential**	**Elo Differential**
t-statistic	6.070	-0.769	12.024	2.915
p-value	0.000	0.442	0.000	0.004

The predictor variables ($X_1, X_3,$ and X_4) with regression coefficients of 0.2597, 1.6206, and 0.0525 (respectively) are statistically significant ($p < .001$) at the significance level ($\alpha = .01$). These coefficients indicate that for every one unit added to average Elo, average points differential, or average Elo differential, the estimated number of total wins increases by 0.2597, 1.6206, or 0.0525 (respectively), as long as all other independent variables remain constant.

The regression equation for this model appears to show that for every unit added, the predictor variable, X_2, will decrease the predicted number of wins by 0.0134 ($\beta_2 = -0.0134$). However, this is not statistically significant at this *p*-value ($p = .442$) at the level of significance ($\alpha = .01$). Thus, this correlation should be considered redundant and should be ignored.

The coefficient of determination ($R^2_{\text{adj}} = .88$) for this multiple regression model indicates a very strong positive correlation between the variance of the response variable and (at least one of) the predictor variables.

Example 1:

If average total points is 75, Elo is 1350, point differential is -5, and Elo differential is -30:

$$\hat{Y} = (34.5753) + 0.2597X_1 + (-0.0134)X_2 + 1.6206X_3 + 0.0525X_4$$

$$\hat{Y} = 34.5753 + 0.2597(75) - 0.0134(1350) + 1.6206(-5) + 0.0525(-30)$$

$$\hat{Y} = 26.2848 \cong 26$$

Example 2:

If average total points is 100, Elo is 1600, point differential is +5, and Elo differential is +95:

$$\hat{Y} = (34.5753) + 0.2597X_1 + (-0.0134)X_2 + 1.6206X_3 + 0.0525X_4$$

$$\hat{Y} = 34.5753 + 0.2597(100) - 0.0134(1600) + 1.6206(5) + 0.0525(95)$$

$$\hat{Y} = 52.1958 \cong 52$$

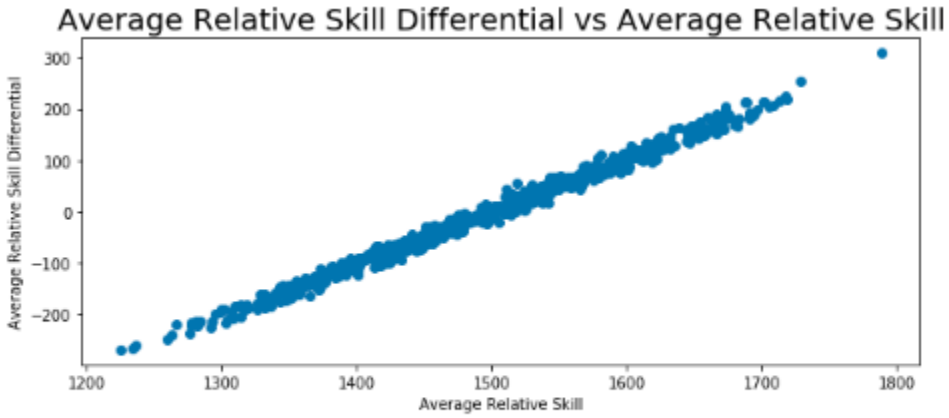
Further Analysis

A major concern with using so many independent variables in a single model is multicollinearity, which is when (some of) the independent variables have a very strong correlation between them. This creates redundancy and other problems like inflated standard errors and unreliable regression parameters. “Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of your regression model. You might not be able to trust the *p*-values to identify independent variables that are statistically significant” (Frost, 2022). Since it would be beneficial to understand the role of each independent variable, multicollinearity should be reduced as much as possible.

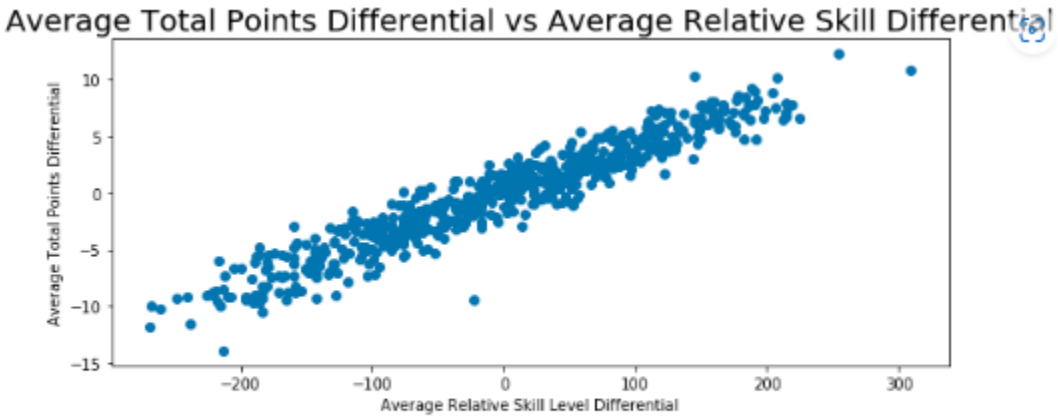
To combat this potential issue, I created scatterplots and (Python) calculated the Pearson correlation coefficient and *p*-value for these independent-to-independent relationships. (See attached HTML file named *Project Three – Further* Analysis.) Figure 5 (below) shows very high correlation between average relative skill (Elo) and relative skill, as well as between average points differential and average Elo differential, and also between average points differential and Elo.

Figure 5

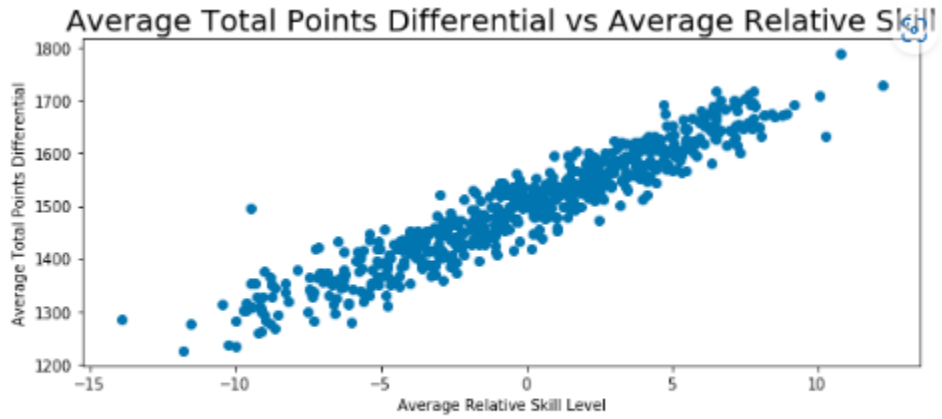
Testing for multicollinearity: high correlation




```
Correlation between Average Relative Skill and the Average Relative Skill Differential
Pearson Correlation Coefficient = 0.9949
P-value = 0.0
```




```
Correlation between Average Relative Skill Differential and the Average Total Points
Pearson Correlation Coefficient = 0.9525
P-value = 0.0
```

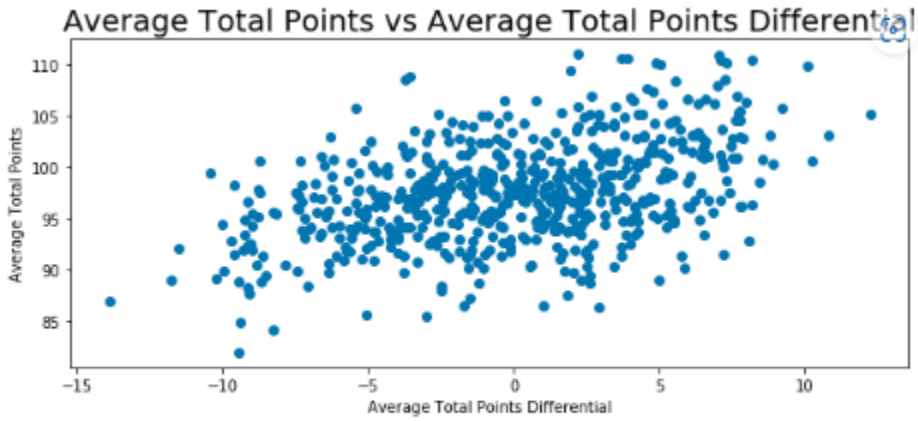



```
Correlation between Average Relative Skill and the Average Total Points Differential
Pearson Correlation Coefficient = 0.9463
P-value = 0.0
```

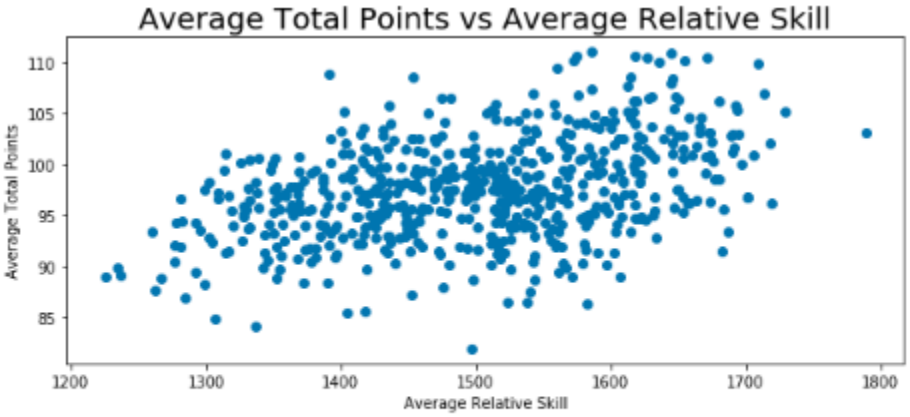
The relationships between average total points and points differential and between average points and Elo each have a moderate correlation, as seen in Figure 6 (below). This makes sense, since Elo is a value calculated from data that includes total points scored in each game.

Figure 6

Relationships with moderate correlation




```
Correlation between Average Relative Skill and the Average Total Points
Pearson Correlation Coefficient = 0.4395
P-value = 0.0
```

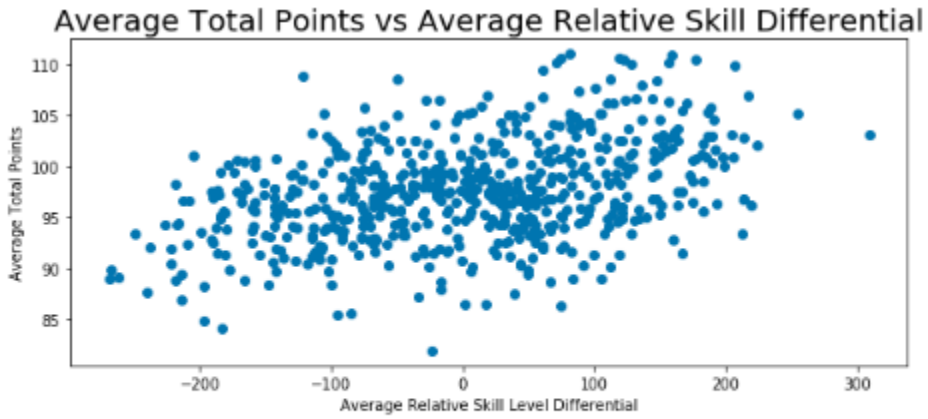



```
Correlation between Average Relative Skill and the Average Total Points
Pearson Correlation Coefficient = 0.407
P-value = 0.0
```

Finally, as seen in Figure 7 (below), average total points and average relative skill differential have a low correlation and should not cause any issues based on multicollinearity.

Figure 7

Relationships with low correlation




```
Correlation between Average Relative Skill Differential and the Average Total Points
Pearson Correlation Coefficient = 0.3976
P-value = 0.0
```

Table 10 (below) shows a summary of the four different models that were analyzed and the associated calculated values for each. Model 2 has the lowest *R* value ($R = .48$) and coefficient of determination ($R^2 = .23$) and should not be considered as a practical model. The other three models, however, show very strong correlation between (at least some of) the independent variable(s) and the response variable (total number of wins).

Table 10

Relevant statistics for the four models that were evaluated

Required Models	Predictor X[i]	R	R-sq.[adj]	S B[0]	p-value B[0]	Slope B[i] (regression coefficient)	S B[i]	p-value B[i]
MODEL 1	elo	0.9072	0.823	3.149	0.000	0.1121	0.002	0.000
MODEL 2	pts	0.4777	0.227	9.305	0.000	1.2849	0.095	0.000
MODEL 3	pts	0.9149	0.837	4.500	0.000	0.3497	0.048	0.000
	elo					0.1055	0.002	0.000
MODEL 4	pts	0.9370	0.877	25.867	0.182	0.2597	0.043	0.000
	elo					-0.013	0.017	0.442
	pts diff					1.6206	0.135	0.000
	elo diff					0.0525	0.018	0.004

Note. Red values indicate (relevant) minimums for each column. The R value and p-value were the only considerations needed to analyze Model 2.

The independent variable for average Elo, in Model 4, has a negative regression coefficient ($\beta = -0.013$) and a high probability value ($p = .442$) and should be considered statistically insignificant. Also, as noted earlier, there is multicollinearity present in this model. This model should not be considered as a valid prediction model.

That leaves Model 1 and 3 for consideration for the best prediction model (of the four models analyzed). Because average total points scored shows to have a weak correlation with the response variable ($R^2 = .23$), Model 3 should also be disregarded.

In the end, I would recommend Model 1 as the best predictor out of these four models. Using only one variable can keep things simple and keep our coefficients and *p*-values unaffected by other variables. Other models should be considered in the future, such as simple regression with only average Elo differential as the independent variable, for example. Perhaps even a different multiple regression model should be evaluated, as even with multicollinearity

present, the actual prediction power of the model will not be influenced by such an issue. For understanding individual key metrics based on the regression coefficients though, multicollinearity should be avoided.

References

Bhandari, P. (2022). *Reporting Statistics in APA Style | Guidelines & Examples.* Scribbr. Retrieved October 15, 2022, from https://www.scribbr.com/apa-style/numbers-and-statistics/

FiveThirtyEight. (2019). *FiveThirtyEight NBA Elo dataset*. Kaggle. Retrieved October 14, 2022, from https://www.kaggle.com/fivethirtyeight/fivethirtyeight-nba-elo-dataset/

Frost, J. (2022). *Do I Have to Fix Multicollinearity.* Statistics By Jim. Retrieved October 15, 2022, from https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Purdue University. (2019). *APA Sample Paper*. Purdue Writing Lab. Retrieved October 13, 2022, from https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/apa_sample_paper.html