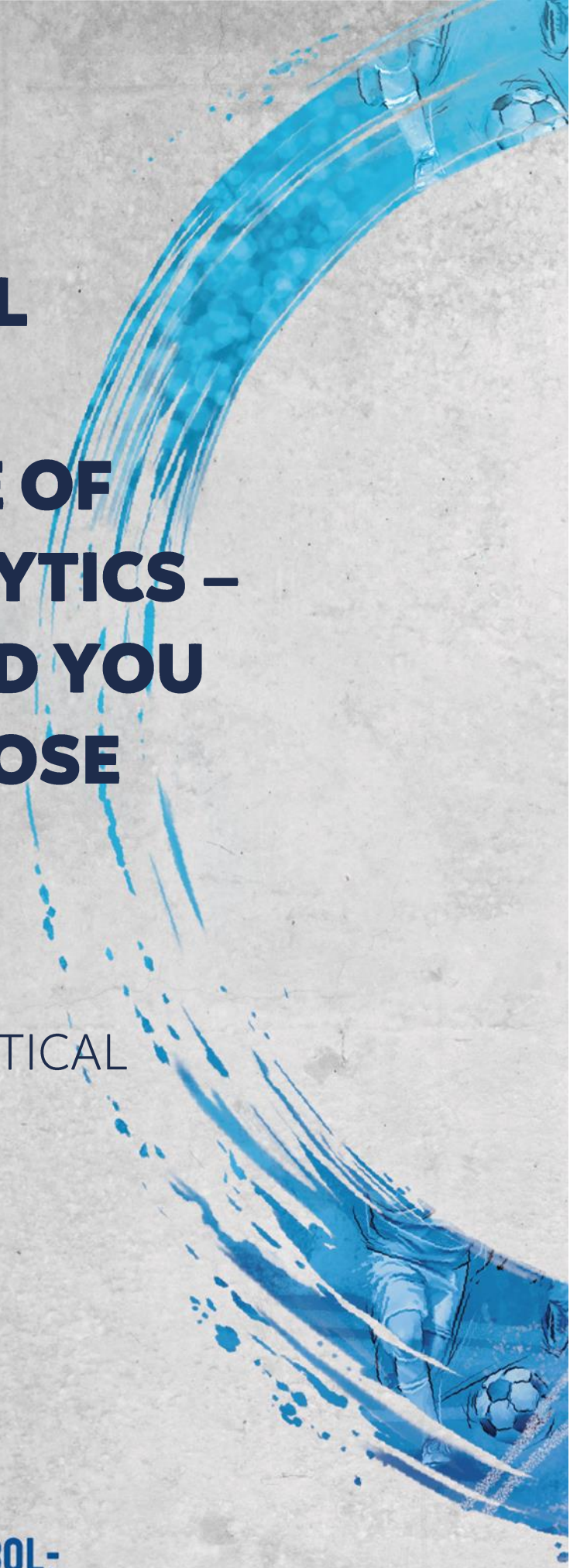
THEORETICAL STATISTICAL KNOWLEDGE OF SPORT ANALYTICS – HOW SHOULD YOU ANALYZE THOSE KPIS?

MODULE 3. STATISTICAL MODELLING IN R



-CONMEBOL-
EVOLUCIÓN

Module 3. Statistical Modelling in R

As a sports performance analyst or sports scientist, not only are you required to understand the sport, the players, and the performance data, but also how to convey this information to key stakeholders such as front office management, coaches, and players. A crucial step which lies between understanding the performance data and the decision-making process is the selection of statistical and predictive models for optimal analysis (Atkinson & Nevill, 2001; Anderson, 2015; Davenport, 2006).

Aside from generating hypothesis- or data-driven analyses, it is up to you to present the results in a meaningful and digestible way to whomever the key stakeholder is. For instance, although we may be excited about p-values, standard errors, and beta coefficients, key stakeholders typically do not desire such detailed background information. They want the bottom-line result. They want to answer questions such as: What does all this mean? And How can I apply this information to the training regimens of players?

This module is designed to guide you through some basic principles of statistical models and help provide a rationale for choosing models based on the types of variables being examined and the questions you are interested in answering. This module provides a template for statistical models which can be used to better present your data to key stakeholders, such as players, coaches, and team management (Atkinson, 2001; Andrews et al., 2011; Brown & Sethna, 2003; Slack & Parent, 2006).

There are several terms you should become well acquainted with as a sports scientist. The first important concept to understand is that there are different types of variables. Incidentally, variables were categorized in 1946 by S. S. Stevens, who declared that all measurement in science is performed using one of four scales: nominal, ordinal, interval, and ratio. Below are examples based on module 1's data types applied to sports.

Review of Data Types Applied to Sports

In tennis, for example, a nominal (categorical) variable, for instance, might be the type of tennis court surface. There are different surfaces such as red clay, hard court, grass court, and carpet. Since the order is not relevant, it can be considered a nominal variable. An example of an ordinal variable is that of the level of a tennis player. For instance, if a dataset were supplied with the categories as follows - top five professional

tennis players, top five Division I tennis players, top five Division II tennis players, and top five Division III tennis players - you would have a dataset consisting of an ordinal variable. The set has a hierarchy, which brands it an ordinal variable. An example of a variable on an interval scale would be the scoring of points in tennis if scoring "hypothetically"' followed the pattern zero (love), fifteen, thirty, forty-five, and game. However, please note that this is a theoretical example, as the true scoring of points in tennis follows the sequence zero or love, fifteen, thirty, forty, then game. This sequence of points is not a true interval variable. The comparison is provided as an example for conceptual understanding. Finally, an example of a measurement in tennis that lies on the ratio scale is the speed of serve, since it is on a continuous scale and has a value of zero as its starting point. An example of an interval variable that can be applied to all sports is temperature (Fahrenheit or Celsius). However, interval variables are hard to find in the sports world, since most measures in sports have a meaningful zero point and are ratio variables. Examples of ratio variables in tennis are the weight of a tennis racquet as well as the speed of the serve because they have a meaningful metrics and a zero point as well (Reid & Schneiker, 2008; O'Donoghue & Ingram, 2001).

In the sport of American football, an example of a nominal variable are different divisions within each conference (east, south, west, and north). Tier levels in football ordered from most to least skilled are considered an example of an ordinal variable. For instance, the National Football League (NFL), the American Football Association (AFA), National Collegiate Athletic Association (NCAA) Division I, NCAA Division II, and NCAA Division III as a whole can be considered an ordinal variable because of their hierarchical nature. An example of an interval variable is the component of time which is evenly distributed among four quarters. An example of a ratio variable in football, as in other sports, is the score because it has a starting point of zero and scores have meaningful magnitude, as the points lie on a continuous scale.

In basketball, examples of nominal variables include team divisions and player positions (centre, point guard, small forward, power forward, and shooting guard). An obvious example of an ordinal variable in basketball includes team rankings. An example of an interval variable in basketball are the four quarters that are played because they are evenly spaced throughout the game. Finally, examples of ratio variables in basketball include the number of assists, field goals, three-pointers, and free throws made, as all of these begin with zero as a starting point and are on a continuous scale. An example of an ordinal variable is the level of the league at which basketball is played, such as the NBA, FIBA, and the D-League. An example of a ratio variable in basketball is the flight time of a player slam-dunking a ball.

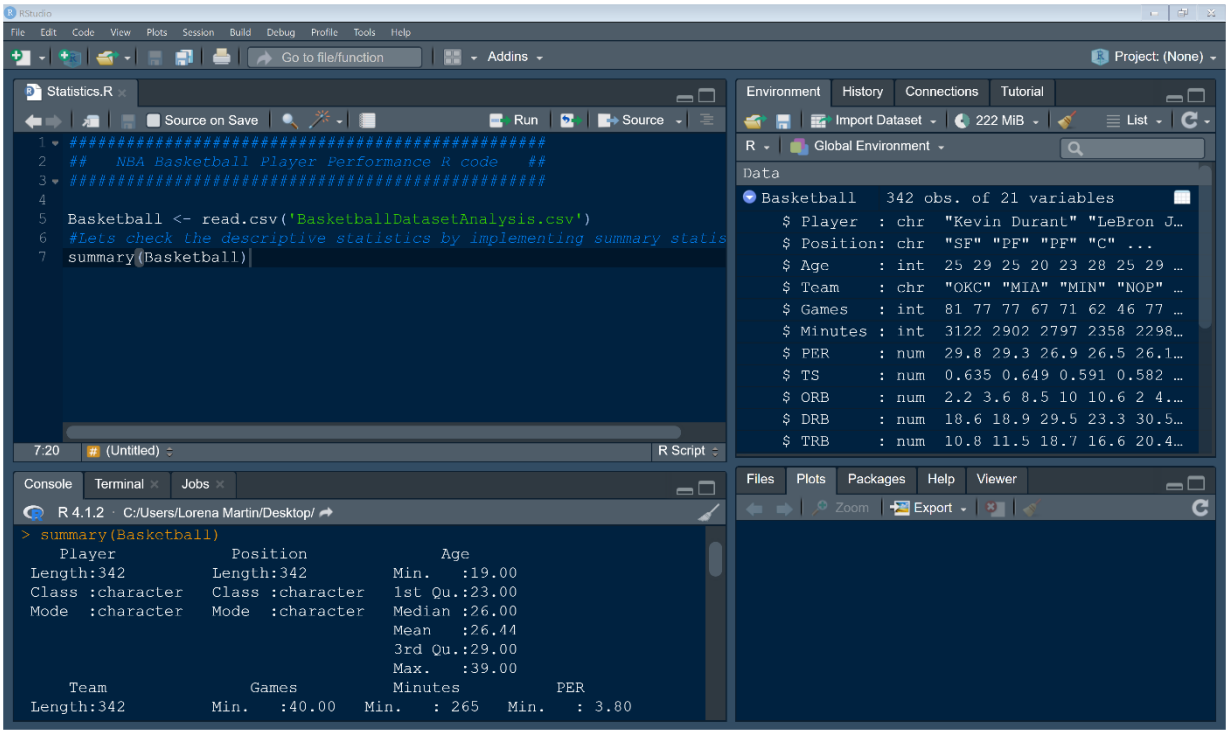
In baseball, an example of a nominal variable is player position, as there are nine different player positions. An example of an ordinal variable in baseball is ranking or tier levels: major league and minor league baseball. As for an interval variable in baseball, each inning is determined by six outs (three from each team), and therefore, can be considered an interval variable. Although baseball is a team sport, it differs from other sports in that time is not an interval variable. This is because an inning can theoretically go on indefinitely. Finally, examples of ratio variables in baseball are the number of home runs scored by a player, the number of strikeouts by a jug, or the overall score of the game. An example of a ratio variable in baseball is a baseball jug's pitching speed.

In soccer, an example of a nominal variable is player position. Examples of ordinal variables in soccer include team rankings and league divisions, such as FC Barcelona and the FC Barcelona B-team, which have a hierarchy and specified order to them. An example of an interval ratio in soccer is the length of time a game is played, with two forty-five minute quarters that comprise the official ninety-minute game. Finally, examples of soccer ratio variables include the number of saves caught by a goalie, the number of goals scored, the number of penalty shots taken, and the number of assists. An example of a ratio variable in soccer is the distance players run throughout an entire soccer match.

Enough of the data types recap.

Having introduced types of variables, we turn to the analysis of data, beginning with data exploration. What does it mean to explore data? One way to begin exploring data is to plot them, construct frequency distributions, and examine descriptive statistics (see the figure below).

Figure 1: RStudio implementing summary function for descriptives analytics



Source: Screenshot by author from RStudio (RStudio, 2022).

Statistics such as means, medians, standard deviations, and correlation coefficients may guide us in developing interesting questions to be examined with inferential statistics and more advanced models. See the figure below for an overview of methods and models and when to apply each.

Figure 2: Statistical methods and their application

Method or Model	Definition or Usage
Indices of Central Tendency and Variability	
Mode	The most common value
Mean	The mathematical average
Median	The center value
Variance	The spread of the distribution
Standard Deviation	How much the values deviate from the mean
Inferential Statistics Used to Examine Group Differences	
Chi-square	Compare observed frequencies with expected frequencies
t-test	Examine differences between two groups on variable of interest
ANOVA	Examine differences between two or more groups
ANCOVA	Control for another variable that may influence the dependent variable
MANOVA	Examine group differences on multiple dependent variables
MANCOVA	Control for another variable that may influence the dependent variables
Statistics and Models Used to Examine Relationships or Predict Outcomes	
Correlation	Examine the association among two variables
Simple Linear Regression	Predict outcome with a single predictor variable
Multiple Linear Regression	Predict outcome with multiple predictors
Logistic Regression	Estimate the probability of the dependent variable class as the values of independent variables change

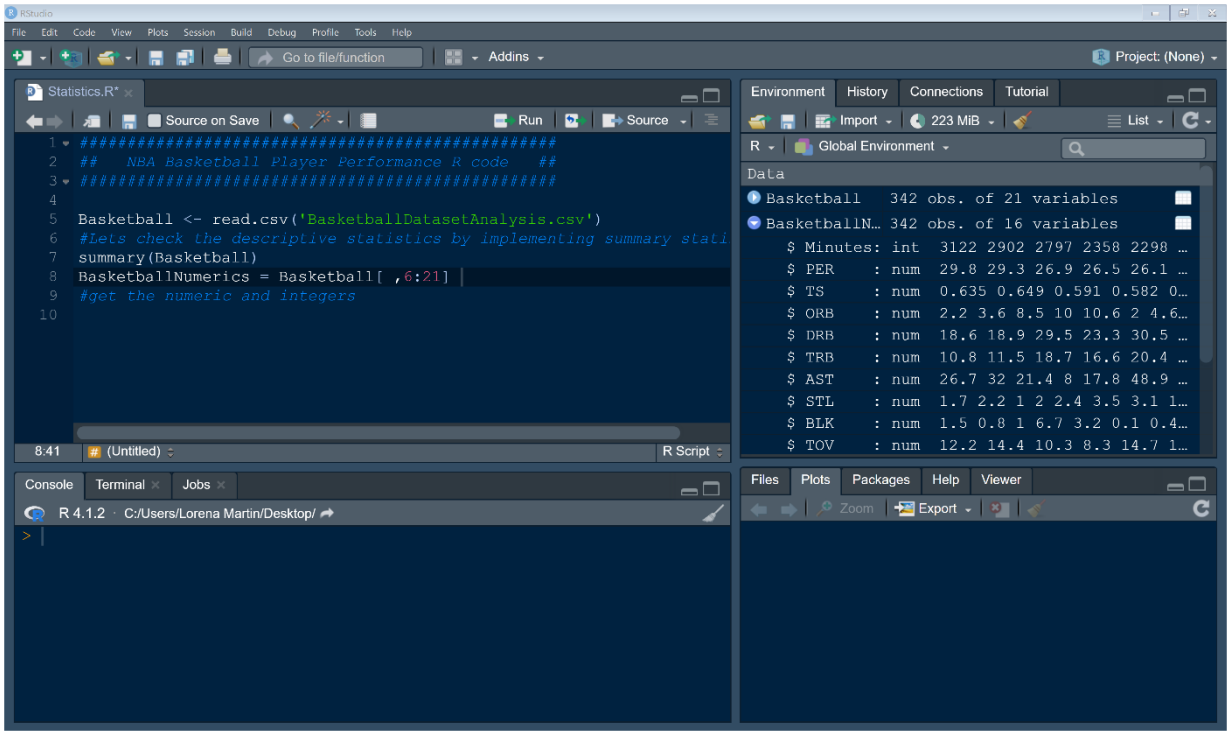
Source: Martin, 2016.

Analysing data

Let us begin by taking a look at correlational analysis. Pearson product-moment correlation quantifies the strength and direction of a linear relationship between two variables. The strength of the relationship is determined by how close or far the correlation values are to zero or one. The closer the numbers are to one, the stronger and more positive the relationship between two variables. The closer the numbers are

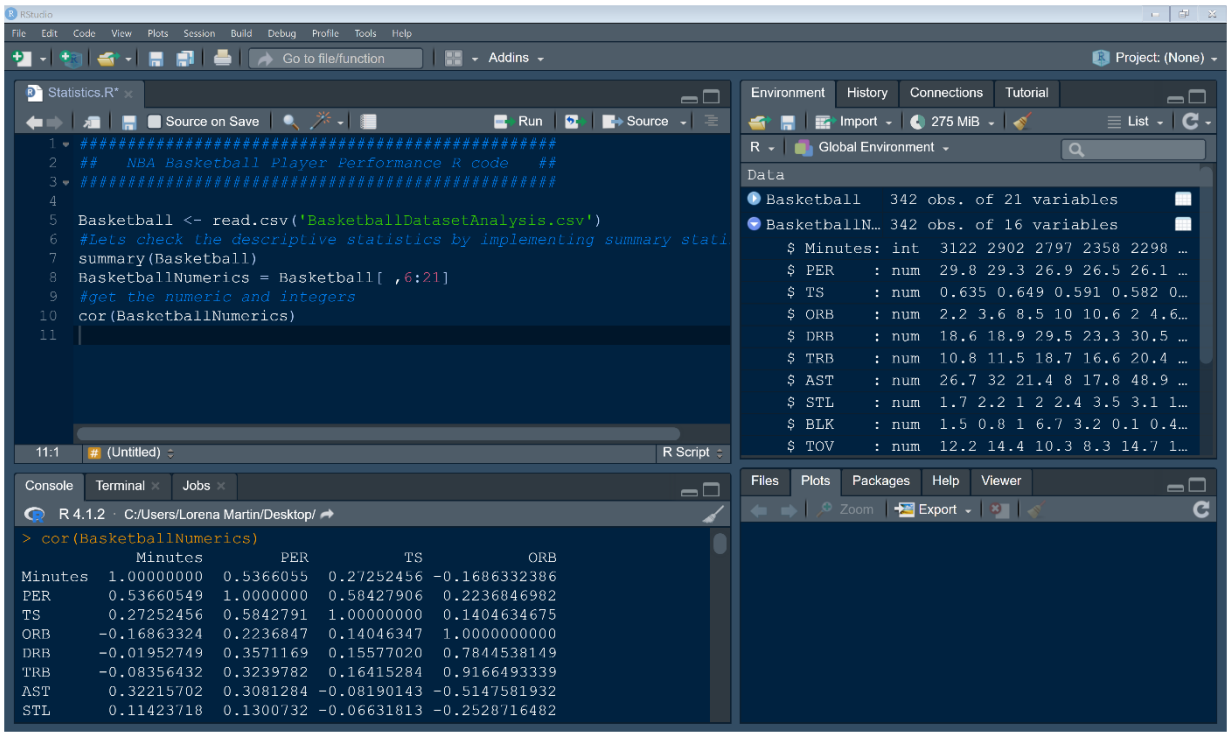
to zero, the weaker the relationship, and if close to negative one, then the association may be strong, but inversely correlated. It is important to recognise that, although a correlation between two variables may exist, it does not necessarily mean that one variable caused the other. It is strongly recommended that you get to know your data, as there are no specified units of measurement displayed in the output from statistical analyses. Pearson's product-moment correlation is intended for two normally distributed variables, otherwise a non-parametric test should be used. Spearman's rho and Kendall's tau may be used because they are not restricted in this manner - they are distribution free. Moreover, correlational analysis is based on normally distributed data, otherwise a non-parametric test should be used. If that is the case, the counterpart to Pearson's r correlation coefficient is Spearman's rho and should be utilized for assessing data that is not normally distributed. Note that many use the terms “correlation” and “association” interchangeably, and this should not be the case. In the field of data science, the term correlation is specific to the intensity and direction of the linear relationship between variables; the term association is used in a more casual manner, and does not imply direct inference from your analyses.

Figure 3: Sub-setting the dataset to include numeric and integer data types in R, sub-setting using square brackets [rows, columns] to examine correlations



Source: Screenshot by author from RStudio (RStudio, 2022).

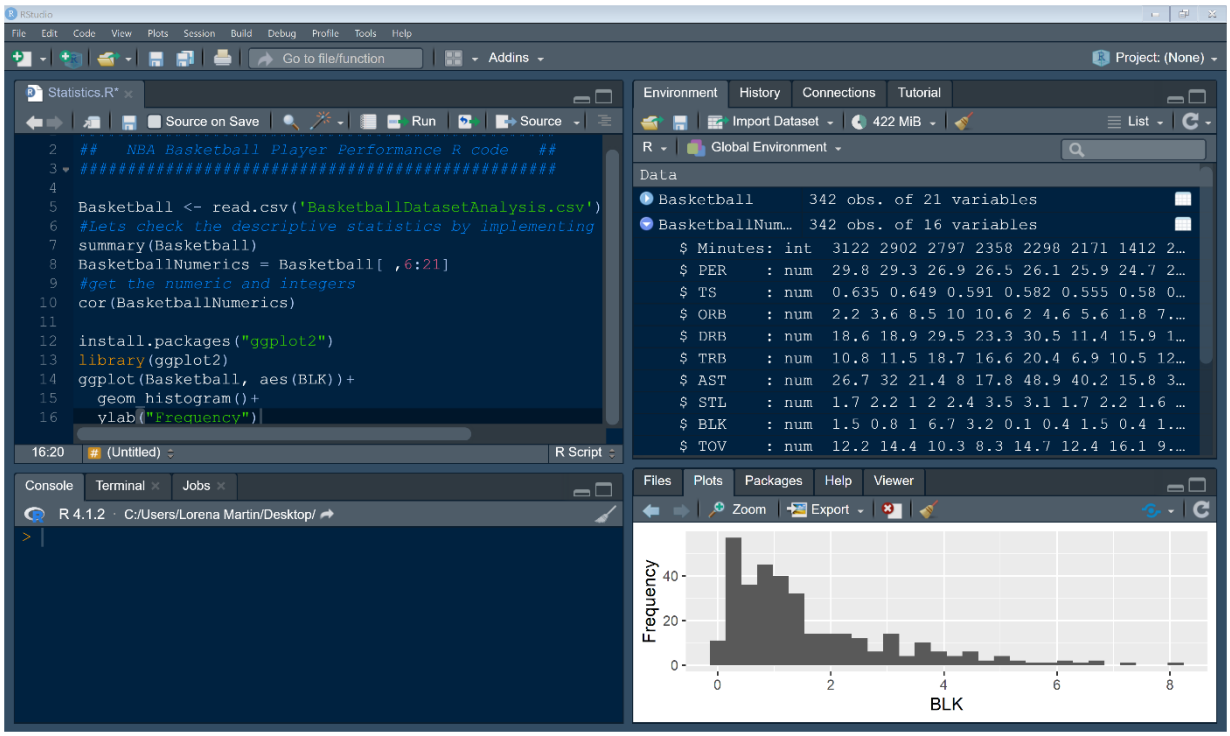
Figure 4: Correlations in R implementing the cor() function



Source: Screenshot by author from RStudio (RStudio, 2022).

It is important to know what you are trying to quantify and its application to sport. Get to know your data and examine the summary statistics. For interval and ratio variables, check histograms and look for normality, as shown in the figure below.

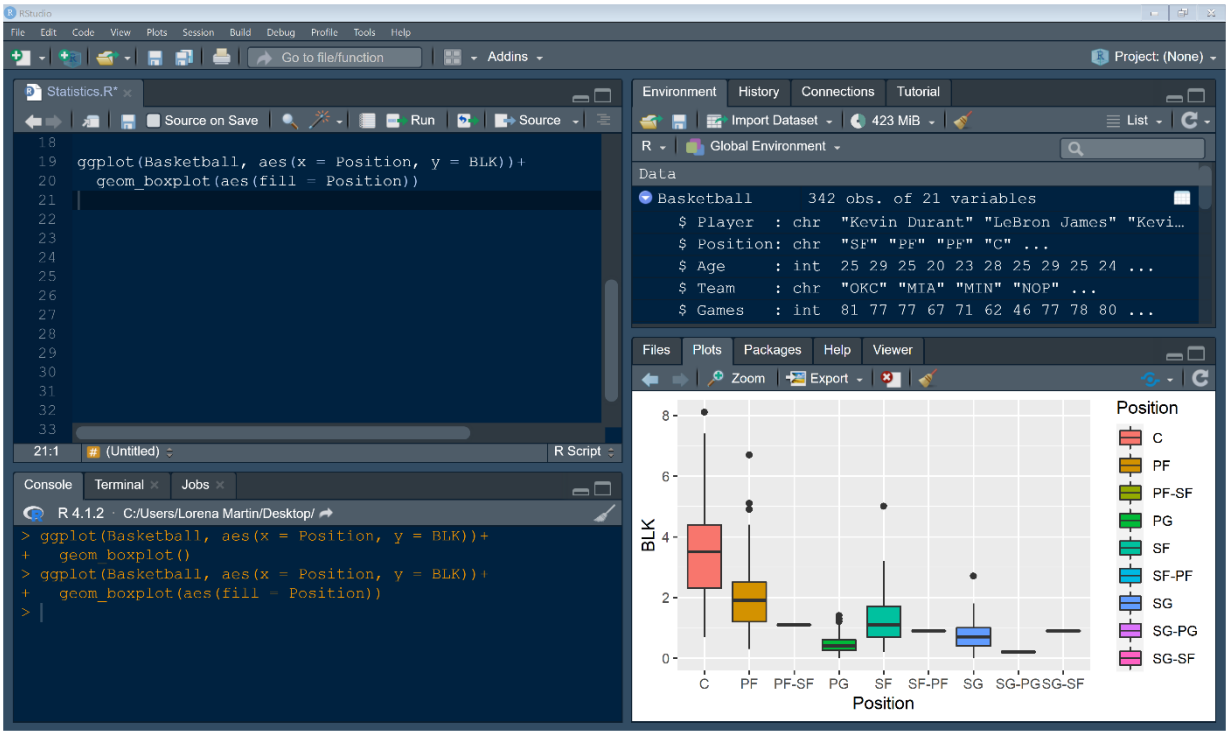
Figure 5: Histogram displaying distribution of the variable blocks (BLK)



Source: Screenshot by author from RStudio (RStudio, 2022).

Additionally, it is important to examine your data for outliers as is common in sports you will have to get creative with different types of analyses. This is important because if the sample chosen is biased or contains extreme outliers, the results of your analyses may contradict the actual values of the norm. As an example, I ran a "hypothetical model" utilising outliers only for demonstration purposes. Examining height (twenty of the shortest and tallest) of former NBA basketball players on field goal percentage and points per game made. Results showed that neither field goal percentage nor points per game differed by height. These findings are impractical as they are based on an aberrant sample. Conversely, results from running an analysis on a normally distributed sample of current NBA basketball players revealed that there is a significant difference in field goal percentage and points per game by height. Specifically, taller players have a higher field goal percentage compared to shorter players, while shorter players score significantly more points per game. This example exemplifies the importance of understanding both the sport of interest and the ability to run the appropriate statistical models.

Figure 6: BLK by player position to examine outliers denoted by the dots above the maximum value of the box plot



Source: Screenshot by author from RStudio (RStudio, 2022).

Note: Graphs and charts displayed were generated implementing the Grammar of Graphics ggplot2 package. The structure of graphing with ggplot is as follows: ggplot(dataset, aes(x variable, y variable) + geom type of graph you want, as shown in the figure ggplot(Basketball, aes(Position, BLK)) + geom_boxplot()

After determining whether your sample is normally distributed and distinguishing which types of variables make up your dataset, it is time to choose a model and check assumptions. The assumptions that must be met to use a parametric test include normally distributed data, homogeneity of variance, and independence of observations. If, however, your data do not meet the assumptions for parametric tests, then non-parametric counterparts should be implemented.

When examining associations between ordinal variables, rather than using Pearson's product-moment correlation, it is recommended to use Spearman's rho, a non-parametric statistic that is a function of ranking the data before applying Pearson's equation to it. Spearman's rho is typically used when you explore the data and observe a large sample data set for which a non-parametric test should be used. For a relatively small data set, Kendall's tau is preferred over Spearman's rho. Below, we detail some of

the commonly applied analyses and the assumptions that should be met for some of the most common parametric models of sport performance.

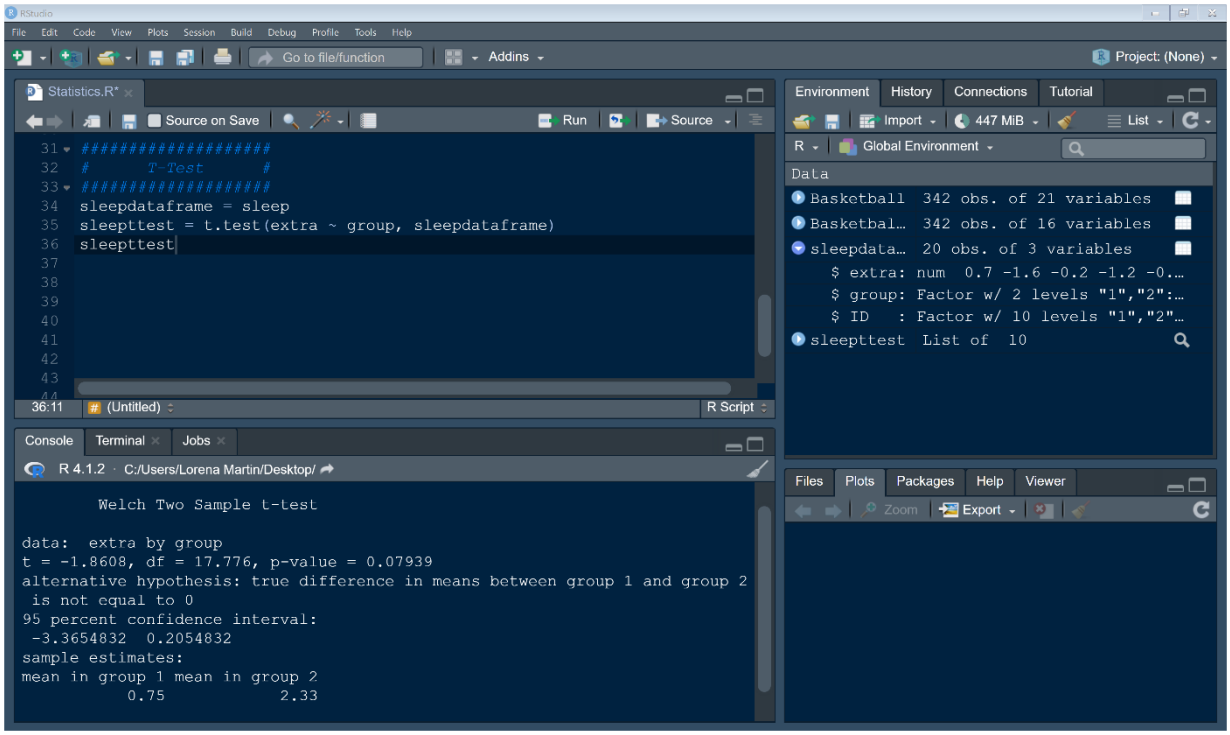
Researchers in various disciplines utilise the student's t-test, a simple statistical test used to compare group means and determine whether there is statistically a significant difference between two groups; for instance, between FC Real Madrid and FC Barcelona on number of goals scored by their top scorers. However, it should be noted that there are several types of t-tests: the one-sample t-test, independent means t-test (aka, two-sample t-test), and the matched paired t-test (aka, dependent means t-test).

A one-sample t-test should be applied when comparing the mean of the team of interest to a benchmark, for example, when comparing the number of curl ups performed to a nationally benchmarked average for a certain population or age group. Another example would be, for instance, if a tennis coach wanted to know whether the speed of students' serves exceeded the speed of serves for the world's top 20 professional tennis players' serving speed, for whom data is already available based on speed radars and Hawkeye technology, we would choose to run a relatively easy one-sample t-test. This test would look at the mean or average of speed of his students' serves, which he would compare against the average speed of the population of interest, in this case, the speed of serves of the world's top twenty tennis professionals. In summary, the group of interest is the group of tennis players the coach is training, the dependent variable is the speed of serve, and the one sample t-test will yield a p-value resulting in either significant or non-significant findings compared to a known parameter, the speed of serves of the world's top twenty professional tennis players.

This differs from the application of an independent t-test, where the situation would be set up as follows; if you wanted to compare the serve speed of two groups, female tennis players and male tennis players, then you would apply an independent t-test. In terms of assumptions that should be met for an independent t-test, the dependent variable, outcome variable, or response variable (used interchangeably; some terms may be preferred more in certain fields than in others) is numeric and continuous. Examples of continuous variables are distance run on the football field, time spent training on the tennis court, or duration of a tennis match or baseball game, hours of training in a sport, the number of unforced errors in the sport of tennis, and the number of touchdowns scored in football. In soccer, the number of goals made, the number of free throws made in basketball, and the number of runs scored in baseball are also examples of continuous variables, just to list a few. The second assumption to be met for the independent t-test is that the independent, explanatory, or predictor variable consists of two independent groups that are categorical. For instance, we could

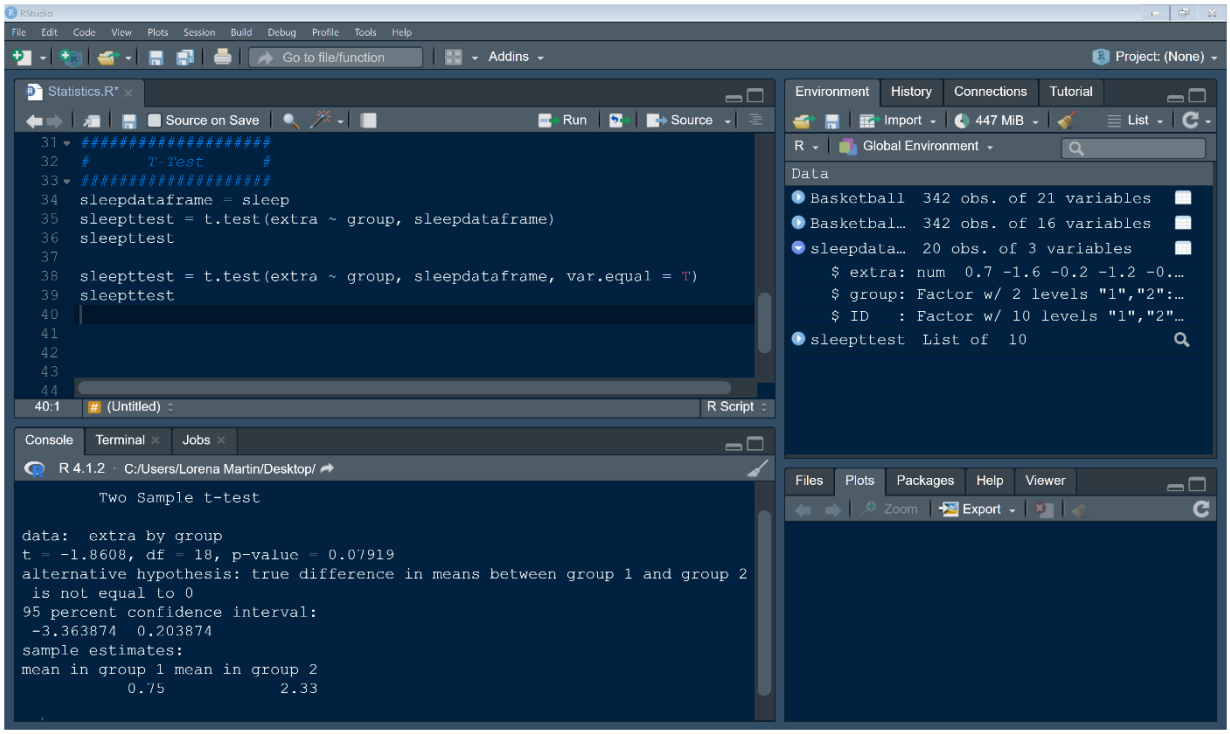
examine differences between two sports, such as baseball and football, with baseball and football representing levels of the categorical variable sport and with the study observations being independent of each other. The third assumption is that of independence of observations, which is designed to ensure that every athlete in a group be examined only in that group, and not in more than one group. Independence of observations is a critical assumption of many statistical models and tests. The fourth assumption of the t-test is that there are no extreme outliers. Including extreme outliers in your data set might tarnish your independent t-test results. If you find something interesting and would like to include the extreme outlier, you may want to use a different analysis that is more robust. Remember, you want to use the best model for the type of data you have. The fifth assumption is that your dependent variable is normally distributed for each of the independent variable levels (in an example where sport is the categorical IV with 2 levels, such as baseball and football). Tests of normality such as the Shapiro-Wilk test and the Kolmogorov-Smirnov test can be used to check that this assumption is met. Finally, the sixth assumption pertains to the criterion of homogeneity of variances, or that the variances should be equal among the groups. Levene's test is specific for assessing homogeneity of variances.

Figure 7: T-test



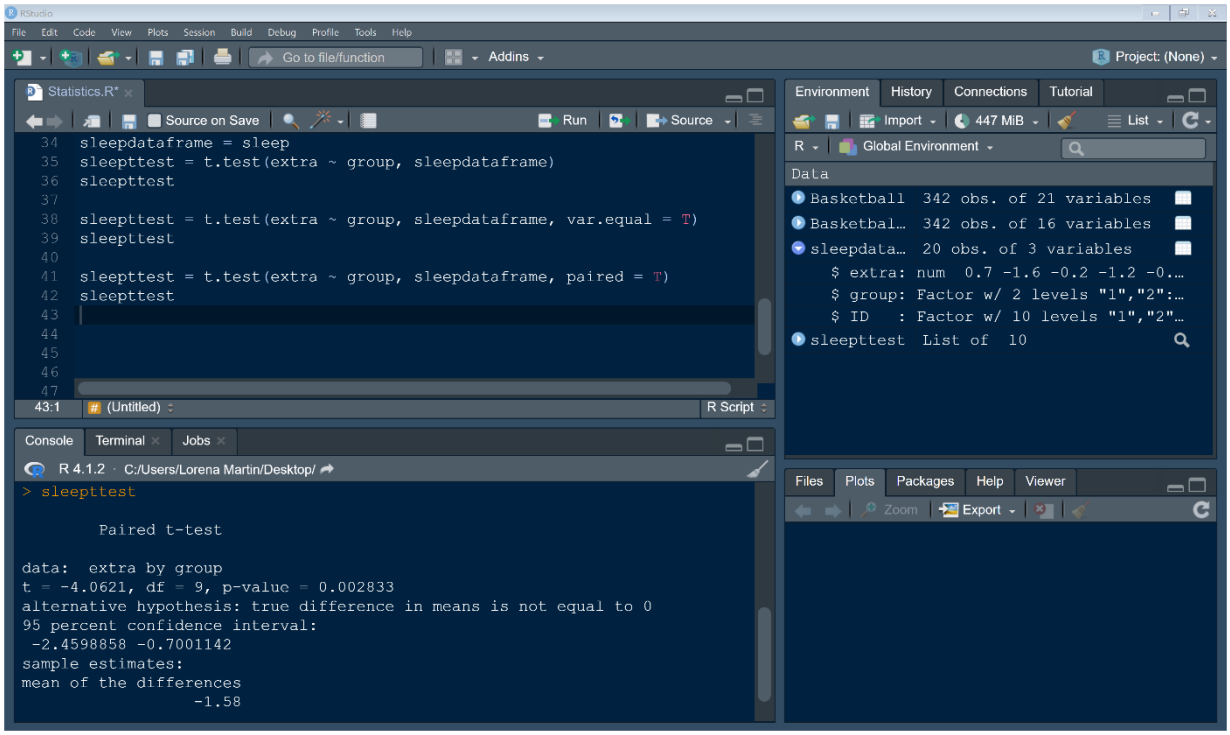
Source: Screenshot by author from RStudio (RStudio, 2022).

Figure 8: Independent t-test



Source: Screenshot by author from RStudio (RStudio, 2022).

Figure 9: Dependent t-test



Source: Screenshot by author from RStudio (RStudio, 2022).

If you are interested in comparing two time points or a before and after, or year to year performance, then you can implement a dependent t-test. There is an additional assumption of having no extreme outliers which must be met to satisfy the criteria for running this type of t-test. A detailed example of when to implement this type of t-test is if you want to answer a sport performance question, such as: Did our soccer players' speed improve from the first day of preseason conditioning compared to the last day of preseason conditioning? Two time points are assumed in this question. Thus, you will have values for the soccer players' speed at baseline and after conditioning. The dependent t-test will yield a result that indicates whether the improvement in speed was significant over those two time points. However, caution is advised, especially when using too many t-tests. They may yield false positives and increase the likelihood of a Type I error.

In the situation that you want to compare more than two groups, such as team performance across your region, etc., then you will need to implement a statistical test that can handle several combinational comparisons. For this, the analysis of variance (ANOVA) model is the right choice. When your question consists of examining the differences between three or more groups on a continuous numeric dependent variable, this is the model to apply. For instance, if we were interested in examining the differences between player positions in basketball on three-point percentage, this model would be appropriate. Our research question would be: Are there differences between centres, small forwards, and shooting guards on three-point percentage?

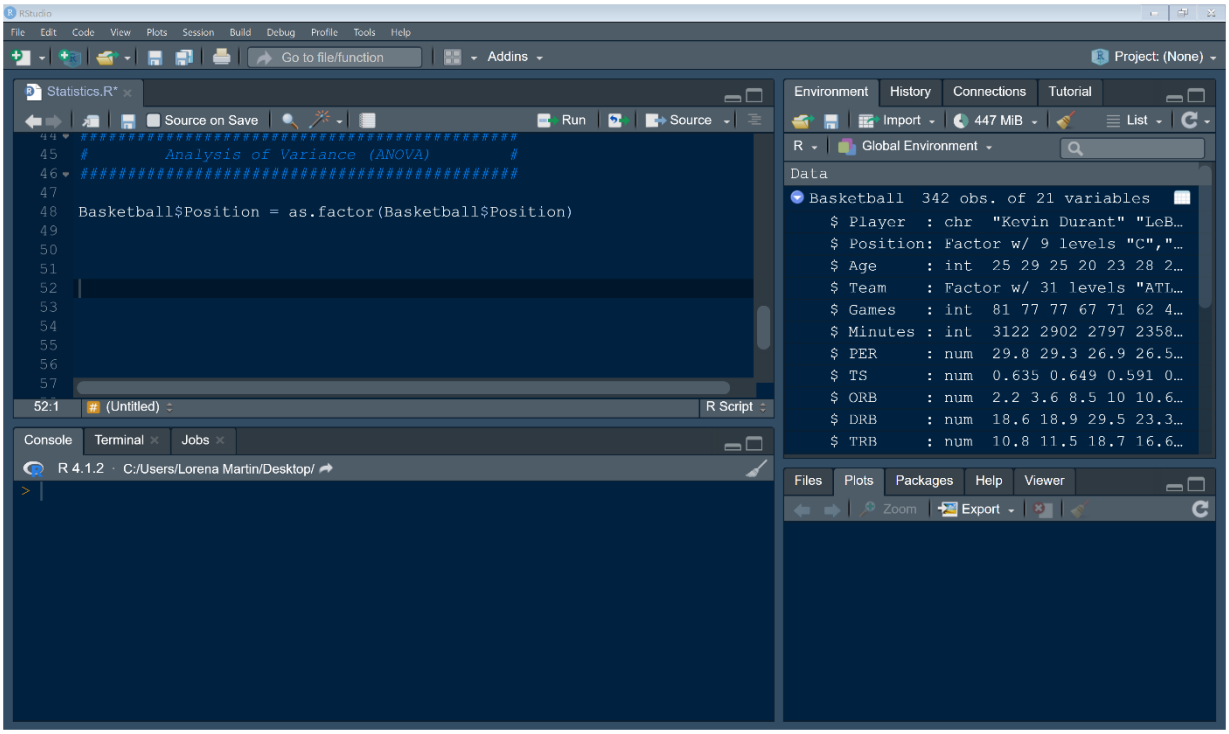
It is important to understand that the ANOVA model can determine significant differences between player positions on the three-point percentage, but it does not tell us between which player positions the significant differences are located. Thus, further investigation is needed, and post hoc analyses should be performed to determine which group pairings are statistically different from one another. There are numerous numbers of post hoc analyses, estimated 18 different types, we recommend going with Tukey HSD as it represents examining for Honestly Significant Difference (HSD).

Before deciding that ANOVA is the model you need; you must make sure to check that six assumptions are met. The first assumption is that the dependent variable is continuous. The second assumption is that your independent variable or variable of interest consists of two or more groups or categories. The third assumption of ANOVA is that there is independence of observations, meaning that no relationship between the independent variables exists. For instance, if interested in examining differences between basketball teams, this last assumption is met when we verify that players on

the Miami Heat are only in that group and do not belong to any other group or team, such as the Oklahoma City Thunder. If we were trying to compare shooting percentage across player positions, for example, the assumption would be violated if an athlete played multiple positions, such as point guard and shooting guard. If we want to use ANOVA, we must ensure that the players in each position are designated, and playing only a single position. The fourth assumption of ANOVA that should be met is that there are no extreme outliers. Typically, the rule of thumb is, if the value lies outside two standard deviations, it is considered an outlier; although in classical statistics, the data point would have to lie further out than three standard deviations above or below the mean. The fifth assumption for ANOVA is that of a normal distribution. ANOVA tends to be robust against the violation of normality, however, if the data are obviously skewed, you might be better off transforming your data or opting for a non-parametric test, such as the Kruskal-Wallis model. The sixth and final assumption that needs to be met to run an ANOVA model is homogeneity of variance, which can be verified by running Levene's test. If the assumption for homogeneity of variance is not met, there are two alternate models that can be used: Welch's test, and the Brown and Forsythe test. If all six assumptions are met, you are free to run the ANOVA model. Remember, to run post hoc analyses to identify where the significant group differences are.

Below, we are going to guide you step by step in R on how to set up your data type to be able to implement a one-way ANOVA, where the independent variable is position and the dependent variable is offensive rebounds. In other words, this would be the analysis, if the question was the following: We wonder if there is a significant difference between player positions on offensive rebounds?

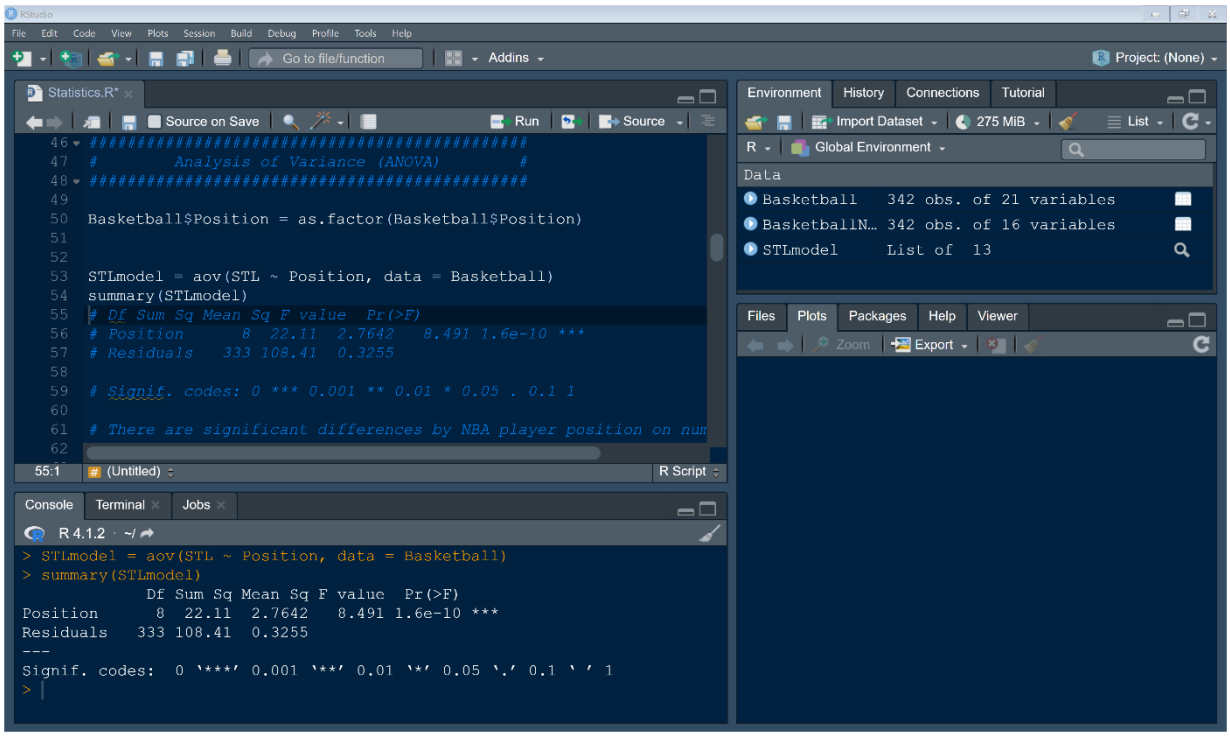
Figure 10: Setting up position variable to factor data type for implementation of ANOVA



Source: Screenshot by author from RStudio (RStudio, 2022).

After you ensure that the independent variable of interest is of data type factor, then you can implement the ANOVA as shown in the figure below.

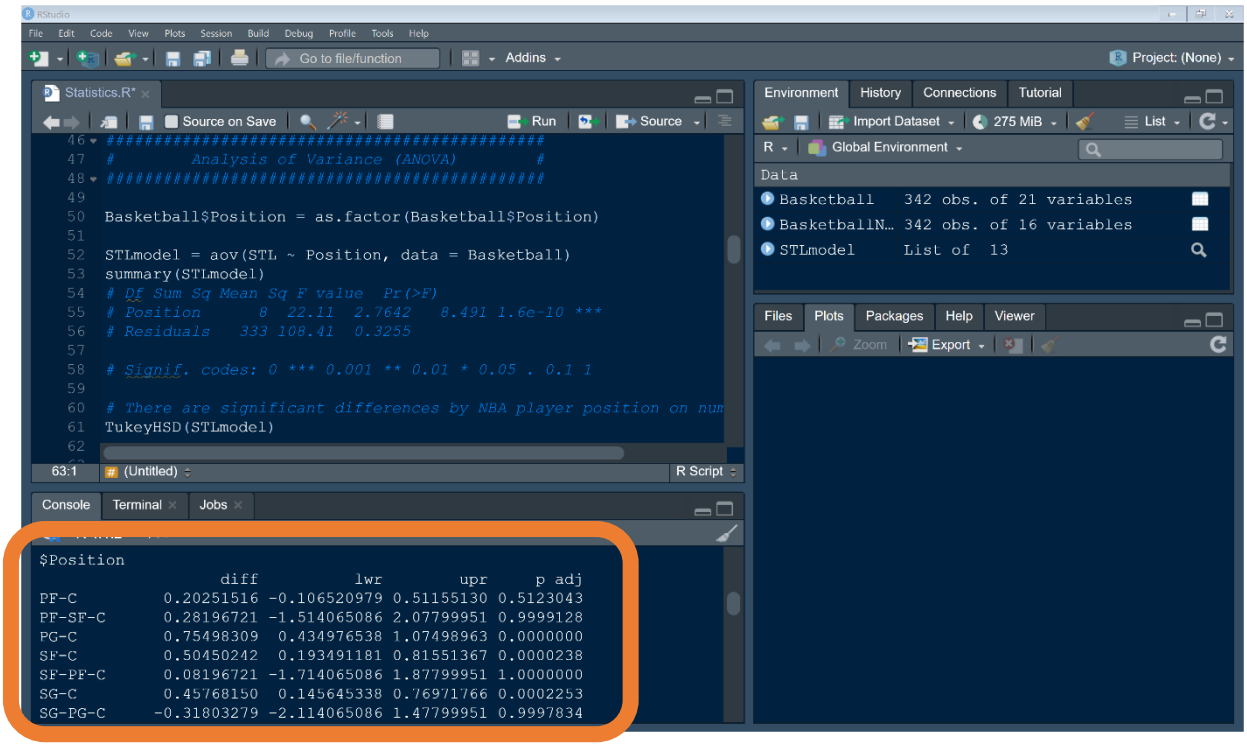
Figure 11: ANOVA, steals by player position



Source: Screenshot by author from RStudio (RStudio, 2022).

Then, after examining the ANOVA results, if there are statistically significant findings, as indicated by the triple asterisks *** in the R console (statistically significant at an alpha 0.001 level), then implement the follow-up post hoc analyses as displayed in the figure below.

Figure 12: Post hoc analyses TukeyHSD to examine group comparisons



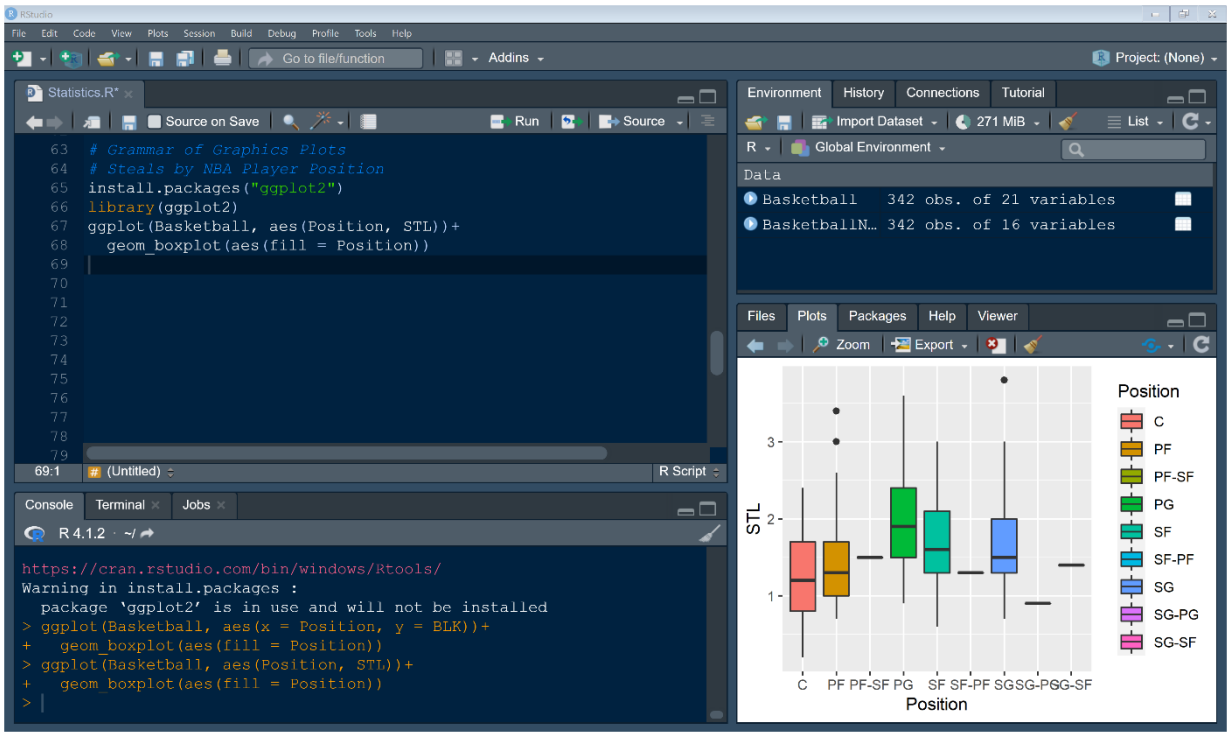
Source: Screenshot by author from RStudio (RStudio, 2022).

As portrayed in the figure above, we can easily see that there were significant differences between the power forward position and the centre position on steals, as displayed in the R console. See the encircled highlighted orange section that details: the position followed by the difference, followed by lower and upper range, then followed by the adjusted p-value. If the p-value is smaller than the specified alpha (which is typically either set at 0.05 or defaulted to 0.05, then you could state that there is a statistically significant difference between the groups.

Recall the following: Alpha is the threshold chosen by the sports scientist or data scientist as the acceptable ability to make a Type I error (a false positive mistake), also known as 1 – Confidence Interval.

Below, is an example of the graphical representation of group player position differences on steals using the Grammar of Graphics package in R. The template structure is the following; ggplot(dataset, aes(x variable, y variable)) + geom_boxplot()

Figure 13: Steals (STL) by player position, emulating the findings from ANOVA

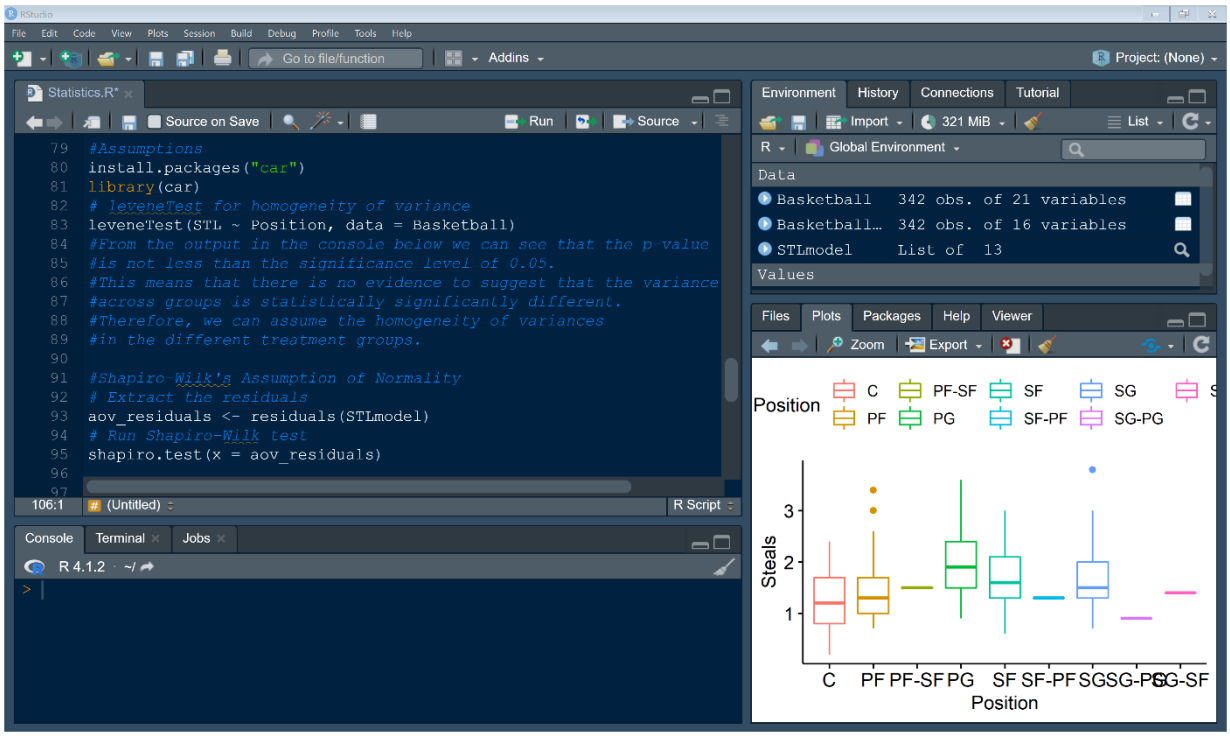


Source: Screenshot by author from RStudio (RStudio, 2022).

Also, if you want to display a graph that is suited specifically for displaying significant differences, you will want to install a package called ggpubr and call the library as displayed in the figure below.

A final check is recommended where assumptions are examined and final determination on whether to implement the ANOVA or a non-parametric version such as the Kruskal Wallis test is to be considered. See the figure below.

Figure 14: Assumption quality control of ANOVA

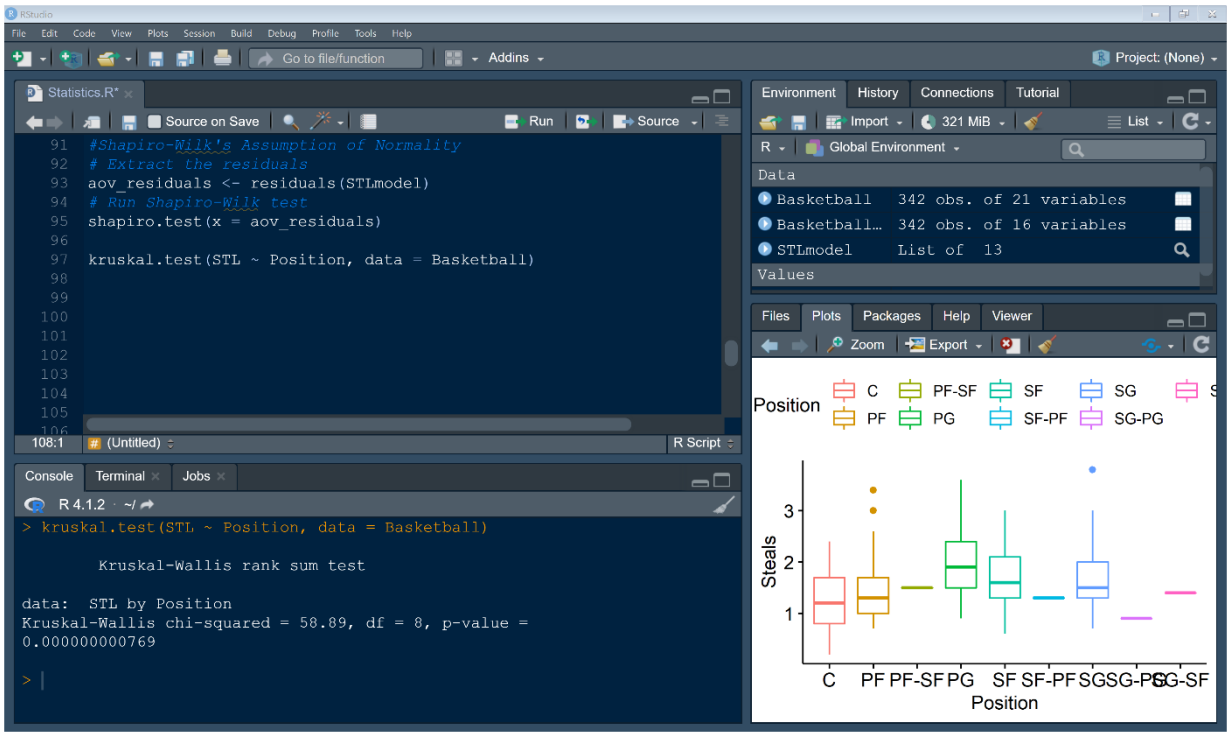


Source: Screenshot by author from RStudio (RStudio, 2022).

As we see above, the assumption of homogeneity of variance is verified by implementing Levene's test with the leveneTest() function and it is yielding a p-value that is not significant, thereby indicating that there are no statistically significant differences between groups in terms of homogeneity of variance.

We then examine for the assumption of normality by implementing a Shapiro Wilk's test, which is done by extracting the residuals of your ANOVA model and then implementing the Shapiro.test() function on the residuals as displayed in the figure above. To recap in assumption testing, we do not want p-values that are less than alpha, they need to be greater than alpha to continue with the ANOVA model. If any of the assumptions fail, then we can implement a non-parametric assessment. The non-parametric version of the ANOVA is the Kruskal Wallis test as implemented with the Kruskal.test() function in R displayed in the figure below.

Figure 15: Assumption of normality

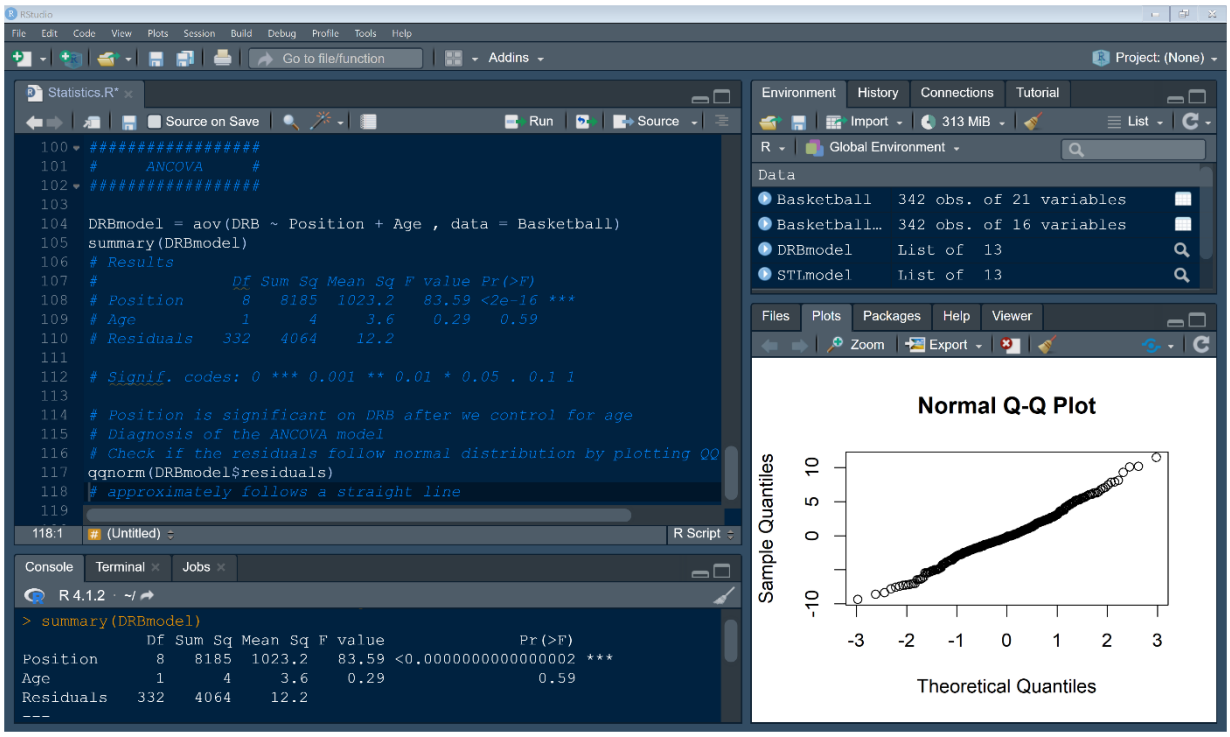


Source: Screenshot by author from RStudio (RStudio, 2022).

Moving to a more complex model, analysis of covariance (ANCOVA) is recommended when the independent variables are categorical and the dependent variable is a continuous variable, but now you want to control for a possible confounding variable. A confounding variable is a factor that may be contributing variance to the independent variable. Therefore, to obtain a true measure of the variance accounted for by the independent variable on the dependent variable, you should control for this third variable. For example, if you wanted to examine the differences between basketball teams' training regimens and performance, it is important to control for the budgets invested in player training (covariate). This is of interest because the amount of money invested in training and equipment can make a difference in coaching, equipment, physical therapy, and strength and conditioning of the players, all of which indirectly influence training regimen and performance. To run the ANCOVA model, the six assumptions for an ANOVA model are to be satisfied. Additionally, the following three assumptions also have to be met: The covariate (third variable) should be linearly related to the outcome for each independent variable group, then there is the assumption of homoscedasticity, and the assumption of homogeneity of regression slopes that also have to be met. What does all this actually mean? What are we checking for? The assumption of the covariate being linearly associated with the dependent variable can be verified with a simple scatter plot. If you find that the relationship is non-linear, ANCOVA is not the optimal choice to analyse your data. The assumption of

homoscedasticity checks that the error term in the relationship between the independent variable and the dependent variable is equal across all levels of the independent variables. Finally, the assumption of homogeneity of regression slopes is meant to determine whether the slopes, although they differ, run parallel to each other as this indicates that there is no evidence of interaction between the covariate and the independent variable if slopes run parallel to each other. This is a great model for comparing differences across groups such as teams or player positions on a particular outcome, while simultaneously controlling for an intervening variable. See the figure below displaying an example, whereby we are examining defensive rebounds by player position with age as covariate. In the figure below, we also checked for normality with the qqnorm() function, if circles follow the straight line, then it is normal.

Figure 16: ANCOVA, defensive rebounds (DRB) by player position with age as covariate



Source: Screenshot by author from RStudio (RStudio, 2022).

Multivariate models are specifically designed to assess multiple variables. Some in the field call models with several independent variables multivariate, but others refer to it when there are multiple dependent variables that are correlated. For instance, suppose you are now interested in examining differences among three positions in football (quarterback, running back, and defensive lineman) on the 40-yard dash, shuttle run, and 5-10-5 agility drill. A multivariate analysis is preferred because the dependent variables are all somewhat correlated in this example; however, there seems to be some

correlation as they are all measures of anaerobic power (some mixed with linear speed and others with change of direction). In this case, a multivariate analysis of variance (MANOVA) is the optimal choice. On the other hand, suppose you wanted to analyse the differences in these same player positions on the 40-yard dash, 225-pound bench press, and the Wonderlic cognitive ability test. The MANOVA is not the correct model to use. Why is this? Because the dependent variables are not correlated. The 40-yard dash is a measure of anaerobic power and speed, while the 225-pound bench press is an assessment of upper body muscular strength, and the Wonderlic test is an assessment of a completely unrelated aptitude. Again, results obtained using this model will tell you that there are significant differences between groups, but not exactly between which pairs of groups the differences lie. Additional analyses are required to determine more specific information. If you decide that the MANOVA model is appropriate, based on the type of data and the number of dependent variables and the correlation among several variables, a total of nine assumptions must be met. The independence of observations assumption is standard. A critical assumption when using multivariate analyses is the minimum sample size required for sufficient power to analyse the data with this particular model.

Large sample sizes are preferred. It is usually preferable to have as large a sample size as possible. Using MANOVA or MANCOVA requires that there be more subjects in each independent variable group than the total number of dependent variables.
Another assumption is that there are no univariate or multivariate outliers. “Univariate outliers” is a term used interchangeably with “outliers” because it is indicative of outliers within each group of the independent variables, compared to multivariate outliers which refers to those in the dependent variables. You should assess for outliers (univariate) using boxplots, and assess for multivariate outliers using Mahalanobis distance. Another assumption is that of multivariate normality, which is checked using the Shapiro-Wilk test of normality. Additionally, MANOVA requires the assumption of a linear relationship among the dependent variables for all independent variables, which usually can be verified by a simple scatter plot matrix. Furthermore, there is the assumption of homogeneity of variance-covariance matrices that is checked using Box's M test of equality of covariance. The final assumption necessary to run these types of multivariate analyses calls for an absence of multicollinearity, meaning that there should not be too strong of a correlation between the dependent variables. It may sound counterintuitive, but to run MANOVAs or MANCOVAs you should have multiple dependent variables which are moderately correlated. If the correlation is too low, it is better to assess the dependent variables individually using ANOVA. And if the correlation is too strong, multicollinearity may be an issue.

MANOVAs can also be used to assess a particular variable at several different time points. This model is called a repeated measures MANOVA. An example of when it is appropriate to use this model is when assessing muscular power pre-season, during season, and post-season because there are at least three or more time points. When using a repeated measures MANOVA, it is important to refer to Wilks' Lambda and the significance of the overall multivariate tests. If you do not find significance, your analysis is done. However, if you find significance, you should follow up using univariate ANOVAs and determine if the tests of between-subjects' effects (across the independent variables) are significant. It is protocol to run a Bonferroni correction after finding significance in the multivariate and between-subjects tests, to correct for the number of ANOVAs conducted.

An additional test that is similar to correlations is the chi-square test of independence, also termed Pearson's chi-square. The chi-square test differs from the well-known Pearson's product-moment correlation in that it is used to examine the relationships between two categorical (not continuous) variables. This test requires that only two assumptions be met: both variables are categorical, and there are at least two independent groups. Suppose you wanted to explore the relationship between the two soccer teams Real Madrid and Barcelona (considered one categorical variable consisting of two groups) on penalty shots made versus penalty shots missed during a season, this would be the analysis of choice. The chi-square test is appropriate for exploring this type of data, especially because the dependent variable is of a dichotomous (penalty shots made/penalty shots missed) nature. Typically, your output, depending on the software being used to analyse the data, will yield a cross tabs section and the chi-square test results.

The figure below displays statistical models, data types and variables, and the type of sports performance questions that can be answered with each type of model.

Figure 17: Statistical models, data types and variables, and sports performance questions

Statistical Model	Data and Variables	Questions Answered by the Statistical Model
Chi-square	One or more categorical variables	Are basketball players more susceptible to injuries than baseball players? (Are two categorical variables related?)
t-test	Dichotomous independent variable for groups, one continuous dependent variable variable	Are there differences between the New England Patriots and the Miami Dolphins on touchdowns scored? (Do differences exist between two groups on a dependent variable?)
ANOVA	One or more categorical independent variables, one continuous dependent variable	Are there differences between the sports of basketball, tennis, and soccer on athletes' salaries? (Do differences exist between two or more groups on one continuous dependent variable?)
ANCOVA	One or more categorical independent variables, one continuous dependent variable, and one or more control variables	Are there differences between the sports of basketball, tennis, and soccer on athletes' salaries after controlling for ticket sales? (Do differences exist between two or more groups after controlling for a covariate on one dependent variable?)
MANOVA	One or more categorical independent variables, two or more continuous dependent variables	Are there differences between basketball player positions; center, point guard, and power forward on field goals, rebounds, and assists? (Do differences exist between two or more groups on multiple dependent variables?)
MANOVA with Repeated Measures	One or more categorical independent variables, two or more continuous dependent variables, with the dependent variables being repeated measures of the same attribute	Are there differences between basketball player positions; center, point guard, and power forward on field goals, rebounds, and assists at at pre season, during the season, and post season? (Do differences exist between two or more groups on multiple dependent variables over different time points?)
MANCOVA	One or more categorical independent variables, two or more continuous dependent variables, and one or more control variables	Are there differences between basketball player positions; center, point guard, and power forward on field goals, rebounds, and assists after controlling for minutes played? (Do differences exist between two or more groups after controlling for a covariate on multiple dependent variables?)

Source: Martin, 2016

Keep in mind the R scripts and datasets accompany these modules.

References

Allaire, J. J. (2022). R 4.2.1 [Computer Software]. RStudio, Inc. https://cran.r-project.org/index.html

Anderson, R. (2015). Modeling niches and distributions: it's not just "click, click, click". Biogeografía, 8, 4-27.

Andrews, F. T., Croke, B. F., & Jakeman, A. J. (2011). An open software environment for hydrological model assessment and development. Environmental Modelling & Software, 26(10), 1171-1185.

Atkinson, G., & Nevill, A. M. (2001). Selected issues in the design and analysis of sport performance research. Journal of sports sciences, 19(10), 811–827. https://doi.org/10.1080/026404101317015447

Brown, K. S., & Sethna, J. P. (2003). Statistical mechanical approaches to models with many poorly known parameters. Physical review E, 68(2), 021904.

Davenport, T. H. (2006). Competing on analytics. Harvard Business Review, 84(1), 98–134.

Martin, L. (2016). Sports performance measurement and analytics: The science of assessing performance, predicting future outcomes, interpreting statistical models, and evaluating the market value of athletes. FT Press.

O' Donoghue, P., & Ingram, B. (2001). A notational analysis of elite tennis strategy. Journal of sports sciences, 19(2), 107–115. https://doi.org/10.1080/026404101300036299

Reid, M., & Schneiker, K. (2008). Strength and conditioning in tennis: current research and practice. Journal of Science and medicine in Sport, 11(3), 248-256. https://doi.org/10.1016/j.jsams.2007.05.002

Slack, T., & Parent, M. M. (2006). Understanding sport organizations: The application of organization theory. Human Kinetics.