INVESTIGATING THE RELATIONSHIP BETWEEN THE AGE OF A SOCCER PLAYER AND THE NUMBER OF MEAN GOALS SCORED PER GAME USING STATISTICAL AND CORRELATIVE ANALYSIS IN WOMEN'S SOCCER LEAGUES OF FINLAND

Internal assessment
Mathematics SL
May 2021

TABLE OF CONTENTS

1. INTRODUCTION

Soccer is the most globally popular sport in the world (Shivili, 2020) and hence has a large amount of statistical analysis performed on individual players, teams and strategies. This is useful for the teams themselves in order for them to improve their gameplay strategy, but it is also financially profitable for betters who place wagers on certain results from games. Understanding the effect of potential variables on the players allows for more reliable prediction of results. In soccer, there are many factors that may affect the amount of goals players score, for example the team they are playing against as well as agility and motor skills of the player themselves (Five attributes that make a good soccer player, 2018). The fitness level of any individual player will likely begin to drop after a certain age due to natural aging processes, however the technical and tactical skill level of the player will keep increasing as one ages and practices. It is expected that older players are more skilled, however it is not so clear whether they score more goals than their young counterparts due to differences in skill levels.

My brother has played soccer for years now and every so on mentions certain, usually attacker players in his or in opposing teams who score much more goals than your average competitive player. These players always seem to move on to more professional teams, and I can't help but wonder what happens to these players. I always imagined they would not be so highly regarded in those teams as surely everyone else there would be just as good, but even so, some players remain exceptional. I never heard from the players he mentioned again but it inspired me to investigate what may happen to players like them in terms of how their exceptional goal scoring abilities in the junior teams manifest in the more professional teams.

The data for this investigation was gathered from the Kansallinen Liiga website (Tilastot - Kansallinen Liiga, n.d.), which has data on Finnish women's teams' players. The data includes only female attacker players between the ages of 15-19, in order to limit the scope of the investigation and control variables like differences in muscle mass between men and women or between significantly younger or older players. The investigation is strictly focused on the statistical analysis of the relationship between the age of the female attacker players between the ages of 15-19 and the number of mean goals they score per game as they age. Due to such a limited sample, the findings will not and cannot be generalized to other settings or samples, but may provide insight into how increasing age along with increasing skill and difficulty as games become more professional affect the number of goals players tend to score per game. The independent variable is the age of the player and the dependent variable is the number of mean goals they score per game during that age. The research question is: *Is there a correlation between the number of mean goals scored by an attacker female soccer player and the age of the player as they age from 15-19?*

H1: There is a positive or negative correlation between the age of the player and the number of mean goals they score per game.

H0: There is no correlation between the age of the player and the number of mean goals they score per game.

2. DESCRIPTIVE STATISTICS

Central tendency refers to the measure of a single value that attempts to identify a central position in the data (Measures of Central Tendency, n.d.). This can be measured through the mean, median and mode of the data set in question. For this investigation, only mean and median were used to measure central tendency. The mean of a set of data is calculated by adding together the numbers and diving by the number of items (n), which is described by the formula (Measures of Central Tendency, n.d.):

$$\bar{x} = \frac{\sum x}{n}$$

Where $\bar{X}$ is the mean, x is the data set and n is the number of items in the data set. The mean of a data set describes the average value of that data set (Measures of Central Tendency, n.d.). It is a good way to look at the distribution of data as long as there are no significant outliers in the data set. For example, in a data set of 1, 1, 1, 1, 1, 2, 3, 3, 3, 300 the mean calculated using the formula is $\frac{1+1+1+1+1+2+3+3+3+300}{10}$ = 31.6. Outliers can be found using the formula Q_1 – (1.5 x IQR) for the low outlier and Q_3 + (1-5 x IQR) for the higher outlier, where Q_1 is the lower quartile, Q_3 is the upper quartile and IQR is the interquartile range of the data set (Statistical Language - Measures of Spread, n.d.). However, for this investigation the calculation of outliers was not necessary, as by definition outliers need to be abnormal, compared to the majority of the data. For example, if one is looking at the mean height of high-schoolers, and one of the students had dwarfism, their height in the data set would be abnormal and an outlier, which does not represent the general population. There is a specific cause for this abnormality, their disorder. However, when considering football players, if most players score 1 goal per game on average, but a singular player scores 4 goals per game on average, this should not be counted as an outlier. The reason for the player scoring so high is their high skill level, which while not the standard, is not itself abnormal or something to discard when investigating competitive football players. In the earlier example, the data point 300 skews the data and the mean does not represent what most of the data indicated. Due to this, the position of the median, when numbers in a data set are listed in ascending order, is calculated using the following formula (Measures of Central Tendency, n.d.):

$$\frac{n+1}{2} th\ item\ in\ the\ data\ set$$

For example, if there are 7 items in a data set, the position of the median would be

$$\frac{7+1}{2} = 4th\ item\ in\ the\ data\ set$$

While the mean represents the average value in a data set, the median describes the middle of the data set (Measures of Central Tendency, n.d.). The median is the value above and below which 50% of the rest of the data set is. It is a good way to avoid the influence of outliers and to represent the general population. For example when looking at incomes, the average income of citizens can be skewed by either the extremely rich or the extremely poor, hence the median which describes what the middle of these two is, is a better way of representing the general wealth of citizens. Along with this, mode is the number occurring most often in the data set

(Measures of Central Tendency, n.d.). It's most useful when looking at large quantities of data. There is no general formula for the mode. It is found by finding the item in the data set that occurs most frequently.

All the calculations for the means and medians of the data sets follow these two formulas, hence only a single example is shown for each. For the example, the player, whose stats which include the age, total goals scored and total games played can be found in the appendix, Karoliina Sydänmaa is used for the calculation of the mean, and the for the median all the whole age group's stats are used. This data can also be found in the appendix.

Mean goals scored at age 17 is $\frac{13}{33} = 0.39$ goals scored per game on average

Median goals scored by 19-year-old players is $\frac{15+1}{2} = 8\text{th}$

Out of the data set of median goals scored by 19 year olds, the 8^{th} term is 0.26

The following tables (tables 1-5) show the individual total goals scored, the median of the total goals scored, mean goals scored by the players and the median of the means can be seen.

Table 1 (Mean and total goals scored by 15-year old players)

AGE	MEAN GOALS SCORED PER GAME	TOTAL GOALS SCORED
15	1.107143	31
15	0.848484848	28
15	0.681818181	15
15	0.272727272	6
15	0.8076923	8
15	0.5833333	7
15	0.7333333	11
15	0.606060606	20
15	0.11111111	1
15	0	0
15	1	7
15	0.869565	20
15	0	0
15	0.105263	2
15	0.115384	3
MEDIAN:	0.606061	7

Table 2 (Mean and total goals scored by 16-year old players)

AGE	MEAN GOALS SCORED PER GAME	TOTAL GOALS SCORED
16	0.326923	17
16	0.44186	19
16	0.793103	23
16	0.44	11
16	0.56	14
16	0.2	1
16	0.45833333	11
16	0.56	14
16	0.166667	4
16	0	0
16	0.818181818	18
16	0.44	11
16	0.052631	1
16	0.074074	2
16	1.3	39
MEDIAN:	0.44	11

Table 3 (Mean and total goals scored by 17-year old players)

AGE	MEAN GOALS SCORED PER GAME	TOTAL GOALS SCORED
17	0.241379	7
17	0.3103448	9
17	0.47058823	8
17	0.173912	4
17	0.94444444	17
17	0.6842105	13
17	0.533333333	16
17	0.1	2
17	0.95238	2
17	0.4	6
17	0.31818181	7
17	0.393939394	13
17	0.23529411	4
17	0.813953	35
17	1.3103448	38
MEDIAN:	0.4	8

Table 4 (Mean and total goals scored by 18-year old players)

AGE	MEAN GOALS SCORED PER GAME	TOTAL GOALS SCORED
18	0.4	12
18	0.40625	13
18	1	5
18	0.11764705	2
18	0.275	11
18	0.25	5
18	0.3043478	7
18	0.1666666	2
18	0	0
18	0.27777777	5
18	0.384615	10
18	0.15789473	3
18	0.2692307	7
18	0.4375	14
18	0.84375	27
MEDIAN:	0.277778	7

Table 5 (Mean and total goals scored by 19-year old players)

AGE	MEAN GOALS SCORED PER GAME	TOTAL GOALS SCORED
19	0.60869565	14
19	0.26086956	6
19	0.57602307	15
19	0	0
19	0.153846153	2
19	0.1	2
19	0.045454545	1
19	0.06666666	1
19	0.26086956	6
19	0.75	21
19	0.363636364	12
19	0.4	4
19	0.3333333	9
19	0.48275862	14
19	0.227272727	5
MEDIAN:	0.260869	6

The next tables (tables 6 and 8) shows the total goals and the mean goals made by the age group per game. The mode of goals is not shown due to the sample of data being rather small (n = 15 for each age group), meaning there is not enough data to look at repetition of data points reliably. In tables X and Y as well as graphs X and Y standard deviation of the sample measures how much values vary from the sample mean. This indicates the dispersion, that is, how consistent the data is (Statistical Language - Measures of Spread, n.d.). How standard deviation and other measures of dispersion are measured is further explored in section X.

Table 6 (Total goals scored per game and standard deviation by age group)

Age	Total goals (all players)	Standard deviation
15	159	10.062661
16	185	10.417176
17	181	10.977032
18	123	6.763347
19	112	6.379282

Table 7 (Mean goals scored per game and standard deviation by age group)

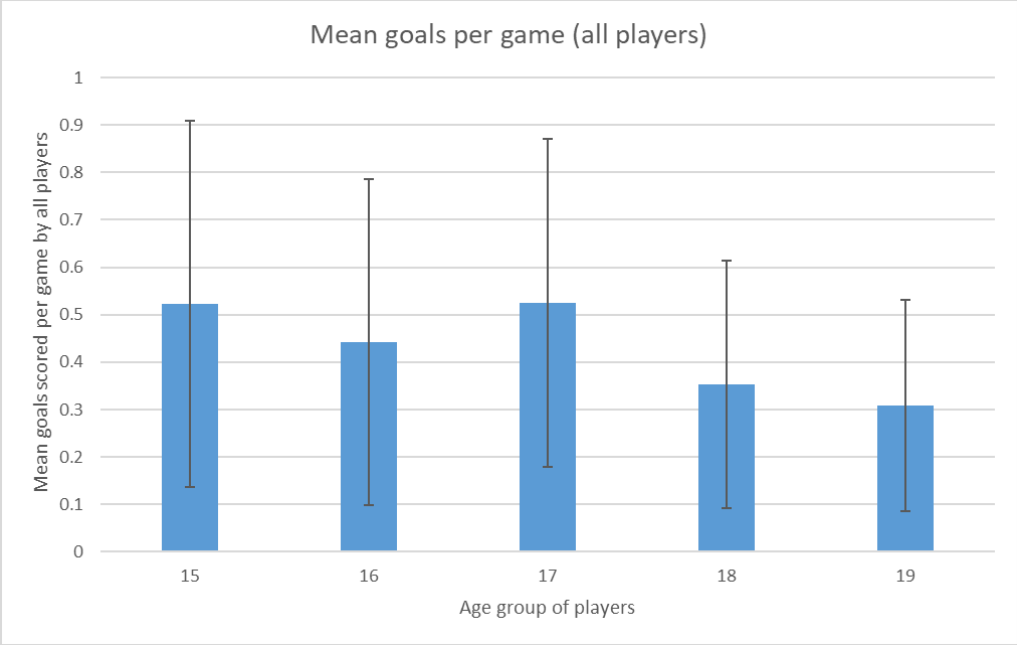
Age	Mean goals per game (all players)	Standard deviation
15	0.522794	0.385065
16	0.442118	0.344241
17	0.525487	0.346114
18	0.352712	0.261472
19	0.308628	0.223667

Table 8 (Median goals scored per game by age group)

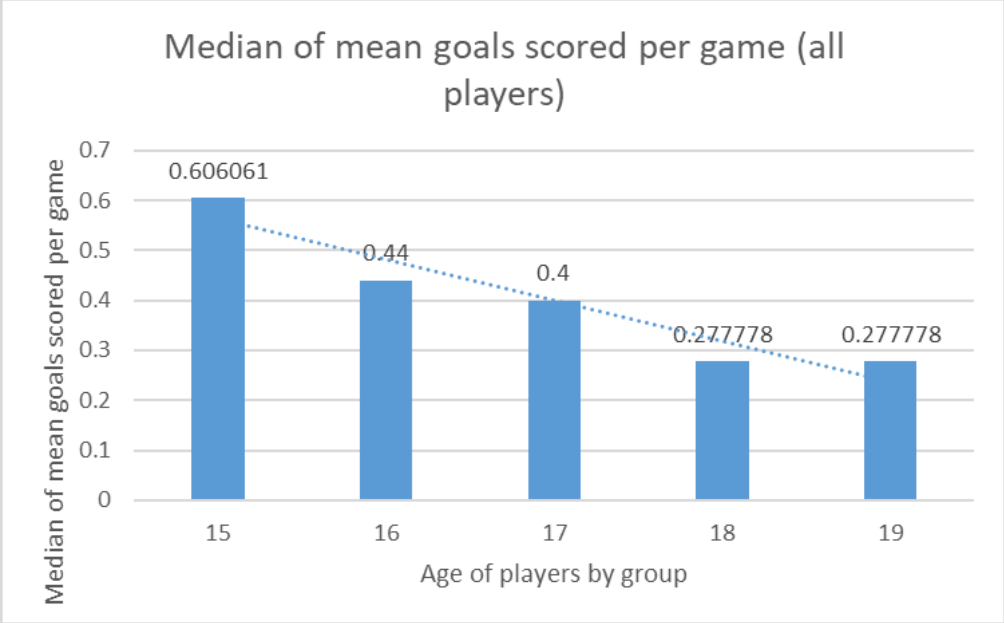
Age	Median of mean goals scored per game (all players)
15	0.606061
16	0.44
17	0.4
18	0.277778
19	0.277778

The trends in data can best be represented using diagrams. Since the investigation follows the effect of age on the mean goals scored per game, the diagrams only illustrate the mean goals per game of all players in diagram 1 and the median of the mean goals in diagram 2.

Graph 1 (Mean goals scored per game by age group)



Graph 2 (Median of mean goals scored per game by age group)



There trends in data are that the median of the means as well as the mean goals themselves are clearly the lowest during the ages 18-19. The median of the mean goals is lower at 15 and higher at 16, even though the mean goals are higher at 15 and lower at 16. This is because in the 15-year old players, the dispersion of the data was much greater; there were more players scoring very high but also more players scoring very low, raising the average goals scored, but when looking at the median, in the 16-year old group the dispersion of the scores is much more even. This means that most were scoring similar amounts of goals consistently. This trend of less dispersion is examined in more detail later in the dispersion section. Regardless, looking at the trend in data, it seems that the amount mean goals, be that median of the means or simply the mean goals scored by the players in a certain age groups, goes down as the players age, and there is a considerable drop between the ages of 17 and 18 in both cases, however this is not the case between the ages 18 and 19. The reason for the drop in mean goals scored after age 17 is suggested to be the transference of the players, who previously were in junior league as minors, to playing in the adult teams with older, more experienced players. Due to the increased difficulty of the games and the higher skillset of the opposing players, the amount of goals one scores goes down, and due to the same reason, even the players who used to be shining stars in the junior teams experience a considerable drop in the amount of goals they generally score. It is comparable to gifted children in high school, who work hard and outshine the other students, only to get into university and realize everyone else is working just as hard at a higher difficulty, and realize they are no longer exceptional. This is also suggested to be the reason there is not a huge difference between 18 and 19-year-old players, since the difficulty remains largely the same.

2.1 FREQUENCY DISTRIBUTION TABLE AND HISTOGRAM

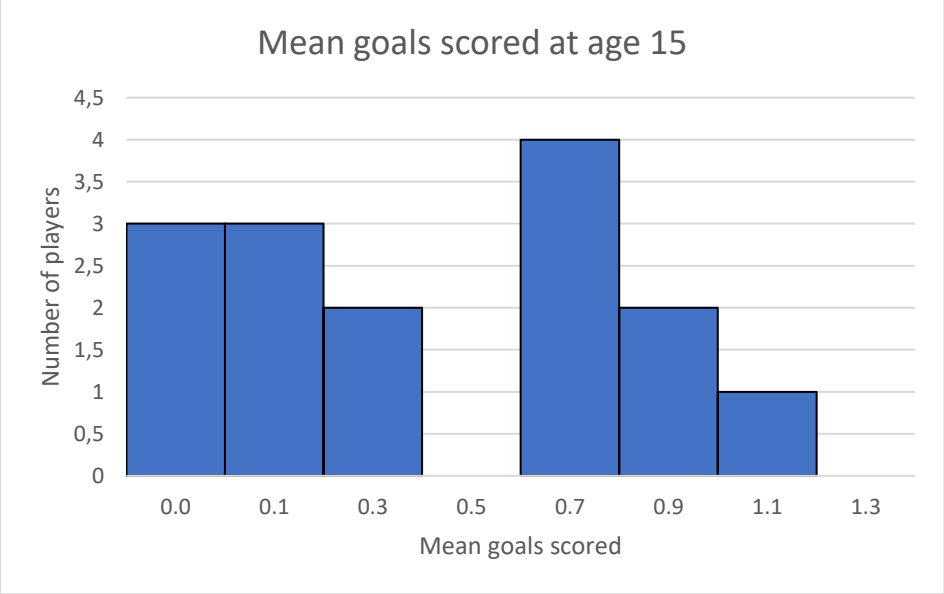
A frequency table allows a person to get a generalized look at the data, and a diagram allows visual interpretation of central tendency (Frequency and Frequency Tables, n.d.). For the table, the goals made per game are categorized as continuous variables, since they can exist as any

values in a finite interval. A frequency tables (table 9) and histograms were created using the data of mean goals per game from tables 1-6. For table 9 the grouping is made in intervals of 0.2.

Table 9 (Frequency distribution table of mean goals scored by age group)

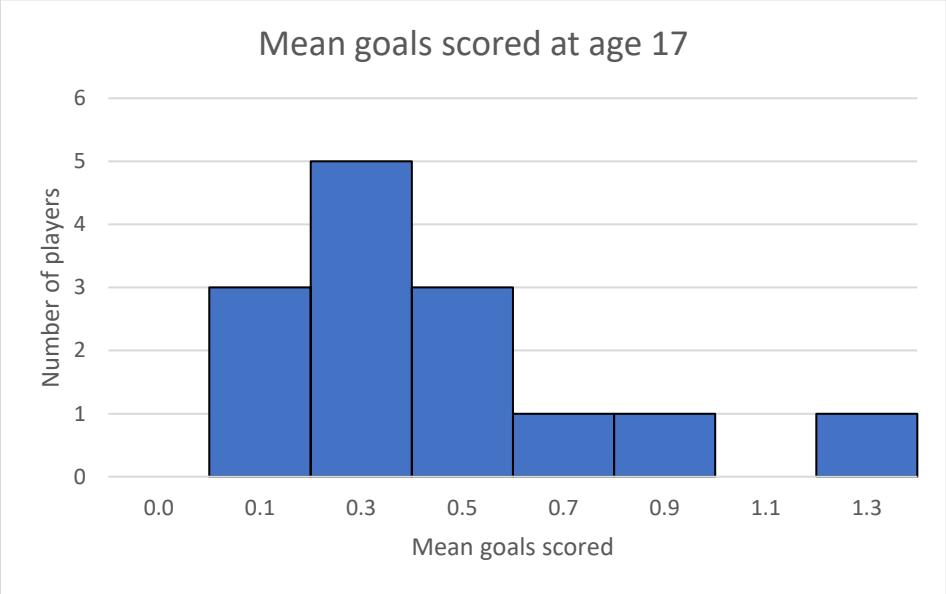
			Age		
Mean goals scored	Age 15	Age 16	Age 17	Age 18	Age 19
x = 0	3	1	0	1	1
0 < x ≤ 0.2	3	4	3	4	4
0.2 < x ≤ 0.4	2	1	5	5	6
0.4 < x ≤ 0.6	0	6	3	3	2
0.6 < x ≤ 0.8	4	1	1	0	2
0.8 < x ≤ 1	2	1	1	2	0
1 < x ≤ 1.2	1	0	0	0	0
1.2 < x ≤ 1.4	0	1	1	0	0

Graph 3 (Frequency distribution of mean goals scored by -year old players)

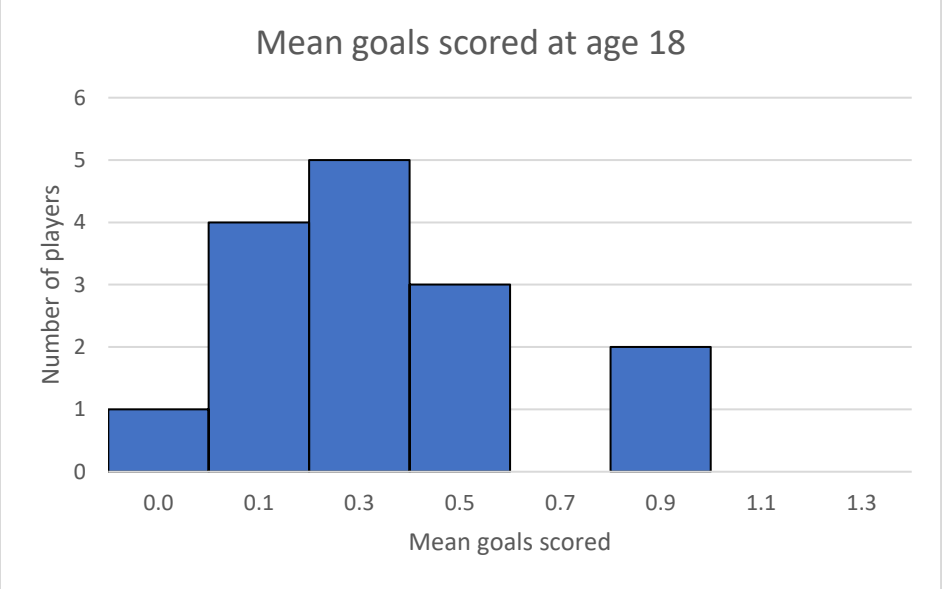


Graph 4 (Frequency distribution of mean goals scored by 16-year old players)

Graph 5 (Frequency distribution of mean goals scored by 17-year old players)



Graph 6 (Frequency distribution of mean goals scored by 18-year old players)



Graph 7 (Frequency distribution of mean goals scored by 19-year old players)

3. DISPERSION AND VARIANCE

While central tendency tells one about the mid-value of a data set (Measures of Central Tendency, n.d.), and how frequency is distributed, it tells very little about the so called range of the data set. The range in mathematics simply refers to the difference between the extreme values of the data set, and it's measured by subtracting the lowest value from the highest value: Range = Highest value – smallest value (Statistical Language - Measures of Spread, n.d.). For example. the range of the mean goals made by 18 year old players is 1.00 – 0.00 = 1. The ranges for the age groups can be found in the table below.

Table 10 (Range of mean goals per game by age group)

Age	Range of mean goals made per game
15	1
16	1.3
17	1.22
18	1
19	0.75

In addition, three other quartiles can be measured. The first quartile, Q_1 refers to the first 25% of the data set in ascending numerical order, the third quartile Q_3 refers to the first 75% of the data set and the interquartile range, IQR refers to the middle space between Q_1 and Q_3 (Statistical Language - Measures of Spread, n.d.). The quartiles can be calculated using the following formulas (Statistical Language - Measures of Spread, n.d.).

$$Q_1 = \frac{1}{4}(n+1)$$

$$Q_3 = \frac{3}{4}(n+1)$$

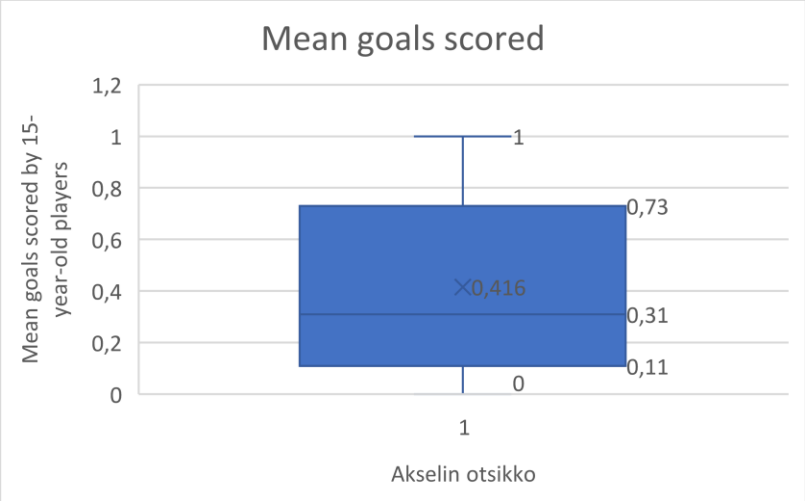
$$IQR = Q_3 - Q_1$$

The ranges for the data in each age group can be seen in the table below along with an example calculation. The values for this are obtained from table 1 in section 1.

Range of the mean of mean goals made by 15 year old players: 1.00 – 0.00 = 1
$Q_1 = \frac{1}{4}(15+1) = 4$, in the data set this value is 0.11
$Q_3 = \frac{3}{4}(15+1) = 12$, in the data set this value is 0.73
IQR = 0.73-9.11 = 0.62

The quartiles can be shown as a box-and-whisker plot, which is a visual summary of the data:

Diagram 1 (Box-and-whisker plot of quartile ranges of mean goals scored)



The quartiles for the mean goals scored per game for the age groups can be found in the table below.

Table 11 (Quartiles and interquartile ranges of mean goals scored by age group)

Age	Q1	Q3	IQR
15	0.11	0.73	0.62
16	0.1	0.56	0.46
17	0.24	0.68	0.44
18	0.17	0.4	0.23
19	0.1	0.48	0.38

The ranges and, the upper quartiles and the interquartile ranges generally seem to become smaller as the players age. There are a few exceptions, like the increase of Q3 at age 19 and the increase of Q1 at ages 16 and 17 compared to age 15. Regardless, there is a general trend, which is suggested to be due to the stagnation of skill observed in all sports. As a player reaches a certain skill level, improving becomes much more difficult and is a much slower process.

Measuring dispersion in this way tells us about how the distribution is ‘’stretched’’ when looking at it in the form of a diagram and is useful in data interpretation (Sample Variance, n.d.). It is useful for the identification of outliers and whether that is mild or strong. In this plot, we see that the data set used had no strong or mild outliers.

Another way to measure dispersion in this investigation is to measure standard deviation and variance of the means. Standard deviation refers to the measure of the dispersion of a set of data relative to the mean of the set of data for a sample, not population, meaning the interpretations and conclusion are not generalizable to other contexts/populations, is calculated using the following formula (Sample Variance, n.d.):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{n-1}}$$

Where

σ = sample variance — x_i = Value of the i[th] element

μ = sample mean — n = sample size

Variance, specifically sample variance for this investigation on the other hand refers to the measure of the spread of numbers and is measured using the following formula (Sample Variance, n.d.):

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{n-1}$$

Where

σ= sample variance — x_i = Value of the i[th] element

μ= sample mean — n = sample size

The practical difference between standard deviation and variance in relation to this investigation is that while standard deviation looks at how spread out the data is from the mean of the data, the variance measures the mean degree to which the points differ from the mean. From analysis of variance we see the degree to which the mean goals scored per game differ from the overall mean, and standard deviation shows us how spread out the data as a whole is relative to the mean. The example calculations for standard deviation and variance of the mean goals per game can be found below along with a data summary table of these values for each age group.

To calculate an example for the standard deviation for 16 year old players, we first have to find the sample mean. We begin this by addition of all the mean goals per game. The data for this is obtained from table 2.

$$0.326923 + 0.441860 + 0.793103 + 0.440000 + 0.560000 + 0.2000000 + 0.458333 + 0.560000 + 0.166667 + 0 + 0.818182 + 0.440000 + 0.052631 + 0.074074 + 1.300000 =$$

Next the sum is divided by the sample size of 15.

$$\frac{6.631773}{15} \approx 0.4421182$$

The next step is to subtract the sample mean from each individual data point and square it. Then the products are added together.

$$(0.326923 - 0.4421182)^2 + (0.441860 - 0.4421182)^2 + (0.793103 - 0.4421182)^2 + (0.440000 - 0.4421182)^2 + (0.560000 - 0.4421182)^2 + (0.2000000 - 0.4421182)^2 + (0.458333 - 0.4421182)^2 + (0.560000 - 0.4421182)^2 + (0.166667 - 0.4421182)^2 + (0.00 - 0.4421182)^2 + (0.818182 - 0.4421182)^2 + (0.440000 - 0.4421182)^2 + (0.052631 - 0.4421182)^2 + (0.074074 - 0.4421182)^2 + (1.300000 - 0.4421182)^2 \approx 1.65903$$

We can now plug in this number to the formula for standard deviation. For the formula, n, the sample size is 15 as mentioned earlier.

$$\sqrt{\frac{1.65903}{15 - 1}} \approx 0.344241$$

To calculate an example for variance, the same group and data is used. The mean of the sample is 0.4421182 as calculated earlier. To begin, the sum of all individual data points subtracted by the mean of the sample is calculated. This is the same process as the standard deviation and yields the same result of 1.65903. This number is then divided by n-1, that is, 15-1.

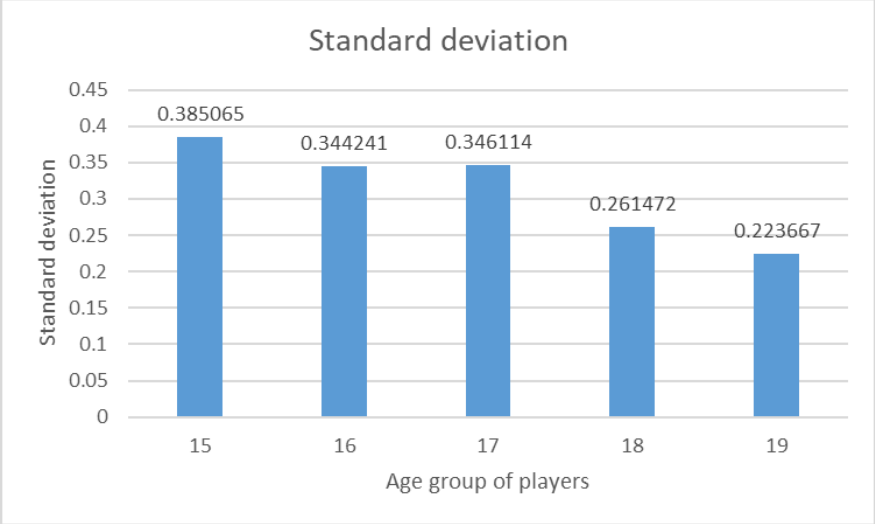
$$\frac{1.65903}{15 - 1} \approx 0.118502$$

The variance of the mean goals made per game for the 16-year-old players is 0.12. The variance and standard deviation for all age groups for mean goals scored per game can bee seen in table Y and graphs X and Z below.

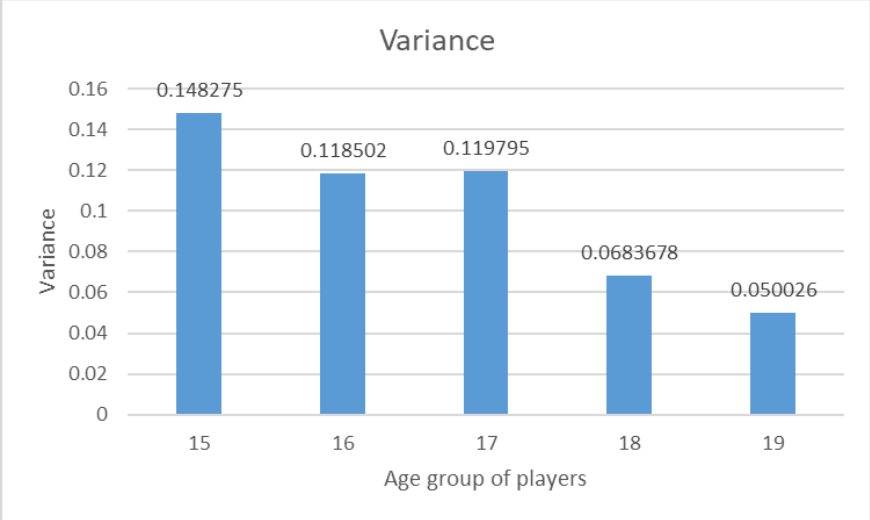
Table 12 (Standard deviation and variance of mean goals scored per game by age group)

Age	Standard deviation	Variance
15	0.385065	0.148275
16	0.344241	0.118502
17	0.346114	0.119795
18	0.261472	0.0683678
19	0.223667	0.050026

Graph 8 (Standard deviation of mean goals scored per game by age group)



Graph 9 (Variance of mean goals scored per game by age group)



Both the standard deviation and the variance data show the same trend as diagrams 1 and 2 in the descriptive statistics section. There is a notable drop in both standard deviation and variance between the age groups 17 and 18, while this is not the case between age groups 18 and 19. There is also a clear trend of standard deviation and variance both decreasing as the players get older. This can reasonably be attributed to two main causes. First, there is an increase in difficulty when the 17 year old players are transferred to adult teams, meaning the high scoring players tend to experience a drop in the amount of goals they make, while the players who were previously scoring less high do not tend to experience this at the same intensity, hence there is an evening out of the goals being made and the range of values tends to decrease; this was already demonstrated earlier in this section. Second, the skill level of the players in the teams even out as the players become professionals.

4. CORRELATION

Pearson's correlation coefficient is a mathematical tool used to measure the relationship between two variables (Correlation Coefficient, n.d.). The correlation can be positive or negative and it can range from strong to null. The values range from -1 to 1, -1 being perfect negative correlation, 1 being perfect positive correlation and 0 being no correlation (Correlation Coefficient, n.d.). A limitation of the Pearson correlation is that it does not measure the direction of the correlation, for example, there could be a 0.8 strong correlation between the amount of time spent playing video games and aggressive outbursts at school. However, nothing can definitively be concluded about the relationship between the variables, since while increased video game time may lead to more aggressive outbursts, it may also be that people who tend to have more aggressive outbursts tend to play more video games. In relation to this experiment, this weakness isn't considerable. The age of a player can affect the mean goals made per game, but the mean goals made per game cannot affect the age of the player.

The formula for the Pearson correlation is the following (Correlation Coefficient, n.d.):

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (x)^2][n\sum y^2 - (y)^{2]}}}$$

The values for x, y, xy, x^2 and y^2 for the following calculations can be found in the appendix.

First the numerator is calculated. The sum of all the xy values is multiplied by n, the sample size which is 75.

$$75(540.927633) = 40\ 569.572475$$

Next, the sum of all the x values and y values is calculated, and they are multiplied by each other.

$$(1275)(32.28) = 41\ 157$$

We can then plug these values into the formula in the numerator.

$$40\ 569.572475 - 41\ 157 = -587.427525$$

Then the denominator is calculated. For the first set of brackets, the sum of all the x-squared values are multiplied by the sample size.

$$75(21825) = 1\ 636\ 875$$

Next, the sum of all x values is squared.

$$(1275)^2 = 1\ 625\ 625$$

These values can then be plugged into the formula.

$$1\,636\,875 - 1\,625\,625 = 11\,250$$

This process is repeated for the second set of brackets with y and y-squared values.

$$75(21.538187) - (32.28)^2 = 573.365625$$

These values for the brackets are then multiplied together, and the value is then plugged into the formula.

$$11\,250 \times 573.365625 = 6\,450\,363.28125$$

$$\frac{-587.427525}{\sqrt{6\,450\,363.28125}} = -0.231$$

The r-value obtained is -0.231. This indicates a weak negative correlation (Correlation Coefficient, n.d.). To find out if this is significant, the critical r value is calculated using the formula N-2, so 15-2 = 13. From a table obtained from the UCONN website (r Critical Value Table, n.d.), the critical r-value can be found. If r>critical r value, the correlation is significant, and the null hypothesis can be rejected. The critical value is 0.514, which means that since 0.231 < 0.514 and -0.231 > -0.514, the correlation in this case is not significant.

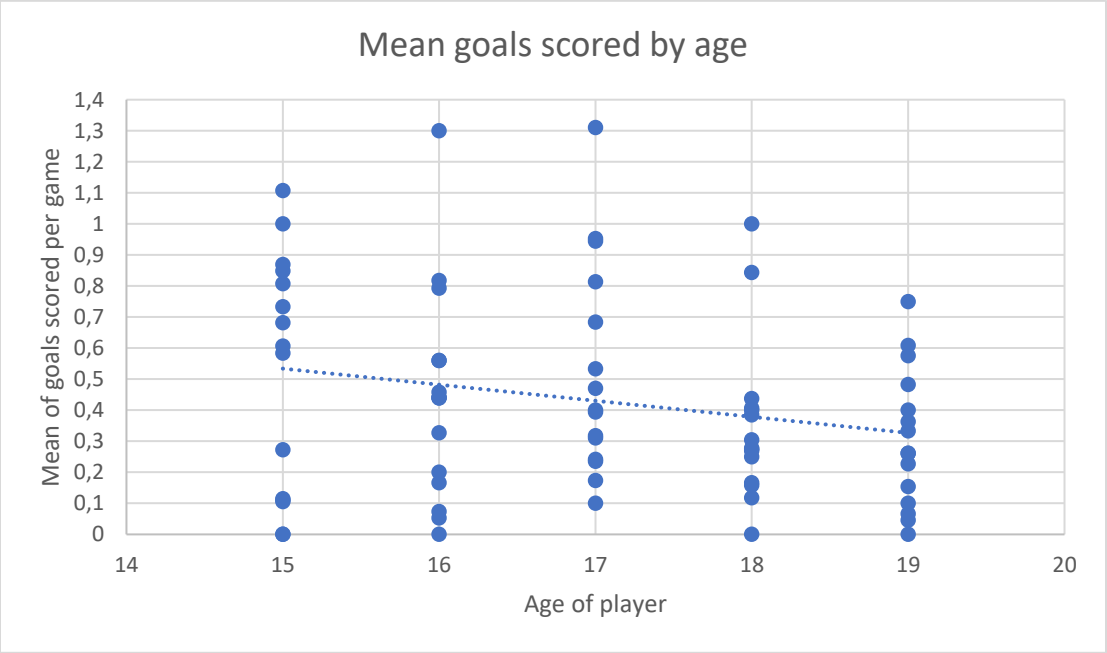
4.1 COEFFICIENT DETERMINATION

In statistical analysis, coefficient of determination, r^2 is the proportion of the variance in the dependent variable that is predictable from the independent variable (Bloomenthal, 2020). It shows how much of the variance can be accounted for due to the independent variable, and is calculated in percentages. Since the absolute value of r is 0.231, the r^2 value is 0.0534. This means that only 0.0534% of the variance can reasonably be attributed to the independent variable, and the remaining 99.9466% are attributed to other variables (Bloomenthal, 2020).

4.2 SCATTER DIAGRAM

The correlation between the age of a player and the mean goals scored per game is shown in the diagram below.

Diagram 2 (Scatter diagram of mean goals scored per game by age)



5. CONCLUSION

There is no significant correlation between the age of a player and the mean goals scored per game. This conclusion is supported by the low r-value obtained from Pearson's correlation and the very small value of r^2 obtained as the correlation coefficient, which indicates that over 99% of the variance in mean goals scored is due to variables other than the age of the player (Bloomenthal, 2020). Realistically this makes sense and is for example supported by the conclusions made in the dispersion and variance section of the investigation. As the players age and move on to more challenging games and professional teams, specifically after age 17, the dispersion of data and the variance in the number of mean goals scored as well as the degree of the variation become much smaller. This is suggested to be due to the stagnation in improvement in skill as well as the players now playing against players of similar or higher skill levels. This effect is observed the most between ages 17 and 18, however there is a constant decline in dispersion and variance of the data from age 15 to 19, and this is suggested to be due to the skill levels of the players becoming similar as they age. One can reach a certain, though individual, skill level at a sport until they begin stagnating in their skill and the improvement becomes much slower or seems to stop altogether. Age correlates with this, but not necessarily with the amount of goals scored, which is correlated more with difficulty of games and individual skill levels. The conclusion is also supported by the frequency distribution of the goals made. Whereas in the 15-year old group the mean goals per game are distributed more, with some players scoring no goals and some scoring 1 goal per game, in the 19-year old group, over 65% of the mean goals fall in two categories in the table. While the correlation between the age of the player and the mean goals scored per game is low and insignificant, it is clear that the age of the player does affect what kind of players they are and in what ways they score their goals, that is, more consistently, though less frequently on average. A considerable limitation of this investigation is the vast amount of uncontrollable environmental variables like the type of coaching received, the type and skill level of the team the player is in, the amount of transfers between teams, injuries, the play strategy of the teams and many more, not to mention personal differences like height,

weight and time spent playing on the field. For similar future investigations, a much larger sample size is suggested, preferably with players of similar socioeconomic and health profiles, as these are things that can majorly affect the skill level and playtime of a player.

During this investigation I've learned that seemingly connected variables may not always exhibit the expected correlation. My very first initial thought on the hypothesis was that of course age and mean goals scored would be positively correlated; the skill level increases, so why not the amount of goals? This was of course shortsighted, and I soon realized my error in judgement as I began to examine closer which factors are at play when scoring goals. I learned to take a deeper look into seemingly simple issues, and to keep my expectations rather neutral and accept the evidence I am confronted with, and form my conclusions based on that rather than intuitive thoughts.

6. APPENDICES

Name	**Birthyear**	**Team**
Nora Lehto	2001	TiPS
Age	**Total goals scored**	**Total games played**
15	37	40
16	24	66
17	0	5
18	15	30
19	16	23

Name	**Birthyear**	**Team**
Jenna Topra	2001	TiPS
Age	**Total goals scored**	**Total games played**
15	46	48
16	19	43
17	8	20
18	14	43
19	11	28

Name	**Birthyear**	**Team**
Aino Kröger	1998	KuPS
Age	**Total goals scored**	**Total games played**
15	15	22
16	23	29
17	8	17
18	5	5
19	15	26

Name	Birthyear	Team
Aada Törrönen	1998	KuPS
Age	**Total goals scored**	**Total games played**
15	7	12
16	2	6
17	13	19
18	5	20
19	2	20

Name	Birthyear	Team
Aino Vuorinen	2001	FC HONKA
Age	**Total goals scored**	**Total games played**
15	14	38
16	30	30
17	25	18
18	11	40
19	2	14

Name	Birthyear	Team
Tiia Savolainen	1997	FC HONKA
Age	**Total goals scored**	**Total games played**
15	0	15
16	1	19
17	4	17
18	7	26
19	9	27

Name	Birthyear	Team
Jasmin Leppioja	2000	FC HONKA
Age	**Total goals scored**	**Total games played**
15	8	16
16	24	22
17	17	22
18	10	27
19	0	16

Name	Birthyear	Team
Olivia Kåhre	1999	PK-35 HKI

Age	Total goals scored	Total games played
15	0	1
16	0	17
17	10	19
18	4	19
19	11	21

Name	Birthyear	Team
Marie Mäkinen	2001	PK-35 HKI
Age	**Total goals scored**	**Total games played**
15	20	34
16	15	26
17	2	20
18	2	12
19	1	15

Name	Birthyear	Team
Henna Tenkanen	1999	FC HONKA
Age	**Total goals scored**	**Total games played**
15	0	1
16	6	22
17	11	25
18	4	25
19	2	17

Name	Birthyear	Team
Anna Olmala	1999	PK-35 VANTAA
Age	**Total goals scored**	**Total games played**
15	7	7
16	18	22
17	7	22
18	10	26
19	11	33

Name	Birthyear	Team
Siiri Koivula	2000	PK-35 VANTAA
Age	**Total goals scored**	**Total games played**
15	13	19
16	11	24

17	16	30
18	7	24
19	1	22

Name	**Birthyear**	**Team**
Karoliina Autio	2000	PK-35 VANTAA
Age	**Total goals scored**	**Total games played**
15	2	19
16	2	27
17	35	43
18	14	32
19	14	29

Name	**Birthyear**	**Team**
Moona Talka	1999	ILVES
Age	**Total goals scored**	**Total games played**
15	0	4
16	0	4
17	6	15
18	5	18
19	18	28

Name	**Birthyear**	**Team**
Karoliina Sydänmaa	2001	JyPK
Age	**Total goals scored**	**Total games played**
15	21	23
16	11	25
17	13	33
18	3	19
19	4	9

7. REFERENCES

Bloomenthal, A., 2020. How the Coefficient of Determination Works. [online] Investopedia. Available at: <https://www.investopedia.com/terms/c/coefficient-of-determination.asp> [Accessed 3 March 2021].

Statistics How To. n.d. Correlation Coefficient. [online] Available at: <https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/> [Accessed 3 March 2021].

soccer.com. 2018. Five attributes that make a good soccer player. [online] Available at: <https://www.soccer.com/guide/five-attributes-that-make-a-good-soccer-player> [Accessed 3 March 2021].

Mathsteacher.com.au. n.d. Frequency and Frequency Tables. [online] Available at: <https://www.mathsteacher.com.au/year8/ch17_stat/03_freq/freq.htm> [Accessed 3 March 2021].

Statistics.laerd.com. n.d. Measures of Central Tendency. [online] Available at: <https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php> [Accessed 3 March 2021].

Researchbasics.education.uconn.edu. n.d. r Critical Value Table. [online] Available at: <https://researchbasics.education.uconn.edu/r_critical_value_table/> [Accessed 3 March 2021].

Statistics How To. n.d. Sample Variance. [online] Available at: <https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/sample-variance/> [Accessed 3 March 2021].

Shivili, J., 2020. The Most Popular Sports In The World. [online] WorldAtlas. Available at: <https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html> [Accessed 3 March 2021].

Abs.gov.au. n.d. Statistical Language - Measures of Spread. [online] Available at: <https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language+-+measures+of+spread> [Accessed 3 March 2021].

Kansallinen Liiga. n.d. Tilastot - Kansallinen Liiga. [online] Available at: <https://www.kansallinenliiga.fi/tilastot/> [Accessed 3 March 2021].