# STATISTICS REPORT ON FOOTBALL TRANSFER DATA

## I. Introduction

Football is one of the most popular sports branches in the world. Football teams are ready to pay significant amounts of money to transfer each other's players. This dataset includes information about the world's most valuable 199 football players in the world extracted and arranged from Kaggle (Sayed, 2022). The data is based on four variables: market values and ages as the numerical variables; countries and positions played as the categorical variables**.** Both categorical variables are nominal data as there is no order of importance in either of them (assuming you are not a racist or in favor of a certain footballer position), and both numerical variables are ratio data as they both have an absolute zero. The aim of this report is to evaluate the dataset with the help of statistical measurements with the help of SPSS. The dataset will be regarded as two separate datasets: Categorical variables and numerical variables. The report also has two distinct parts on these two types of variables where all four variables are investigated individually and compared to the same type of variable. Using the acquired information at the end of both parts, we will evaluate what we learned from the dataset.

## II. Categorical Variables

This part includes two variables: position played by and nationality. Following, you can find pie charts, frequency tables, Pareto diagrams, contingency tables, stacked bar charts, and clustered bar charts describing the given variables.

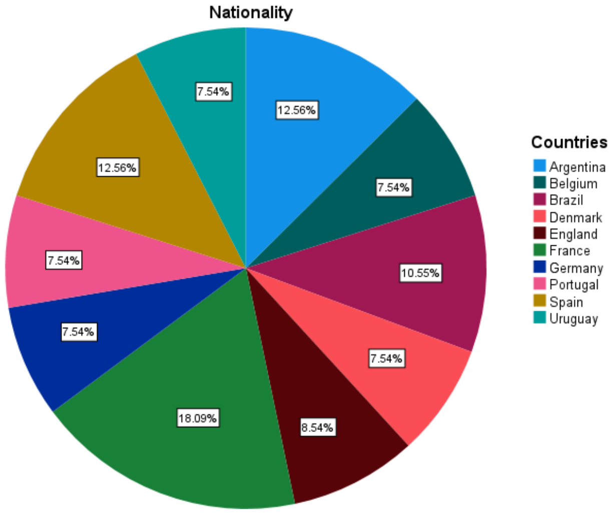


Figure 1: Pie Chart 1.

The pie chart above illustrates one of the categorical variables: nationality. While France has the highest rate, about a twentieth, five countries share the title of having the lowest proportion. Belgium, Denmark, Portugal, Germany, and Uruguay have the same proportion rate, corresponding to 15 players out of 199.

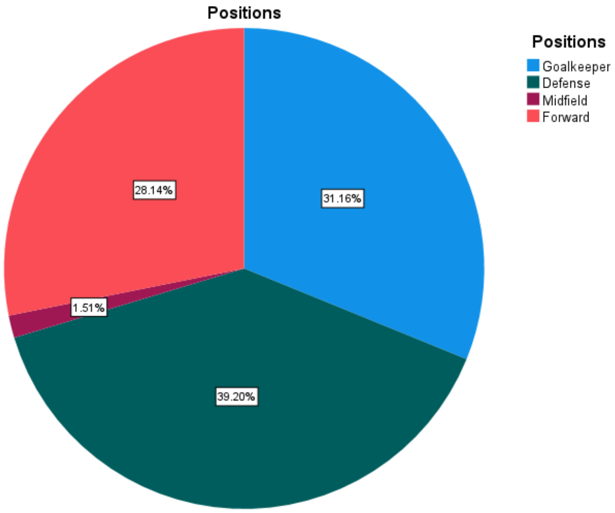
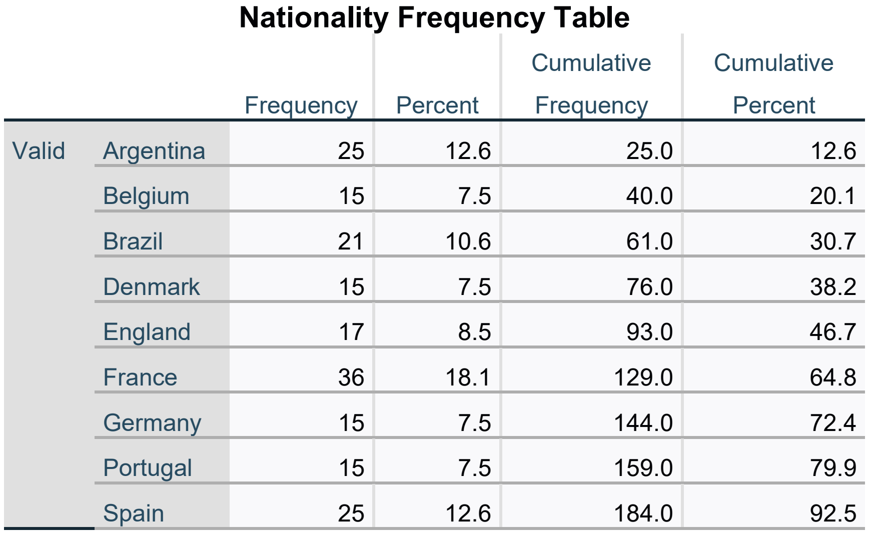


Figure 2: Pie Chart 2.

If you look at the second pie chart showing the positions played by football players, you can see that players whose positions are midfield have a negligible amount of percentage among others. Besides that, defenders have the most share reaching a percentage of almost 40%. In the frequency tables below (Table 1 and Table 2), you may view the precise values.



**Nationality Frequency Table**

| | | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Argentina | 25 | 12.6 | 25.0 | 12.6 |
| | Belgium | 15 | 7.5 | 40.0 | 20.1 |
| | Brazil | 21 | 10.6 | 61.0 | 30.7 |
| | Denmark | 15 | 7.5 | 76.0 | 38.2 |
| | England | 17 | 8.5 | 93.0 | 46.7 |
| | France | 36 | 18.1 | 129.0 | 64.8 |
| | Germany | 15 | 7.5 | 144.0 | 72.4 |
| | Portugal | 15 | 7.5 | 159.0 | 79.9 |
| | Spain | 25 | 12.6 | 184.0 | 92.5 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  | Uruguay | 15 | 7.5 | 199.0 | 100.0 |
|  | Total | 199 | 100.0 | 199.0 |  |

Table 1: Frequency Table 1.

**Position Frequency Table**

|  |  | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Goalkeeper | 62 | 31.2 | 62.0 | 31.2 |
|  | Defense | 78 | 39.2 | 140.0 | 70.4 |
|  | Midfield | 3 | 1.5 | 143.0 | 71.9 |
|  | Forward | 56 | 28.1 | 199.0 | 100.0 |
|  | Total | 199 | 100.0 | 199.0 |  |

Table 2: Frequency Table 2.

There are two frequency tables above, which indicate information based on both nationality and position of the players. In the nationality frequency table, you can see each country by the number of players they have, while the position frequency table contains details about the positions. The second columns of both tables show the percentage of the frequency in the same row. When you question how often a characteristic occurs above or below a particular value, you might look into the cumulative frequency columns. The Pareto diagrams enrich the visualization of cumulative information.

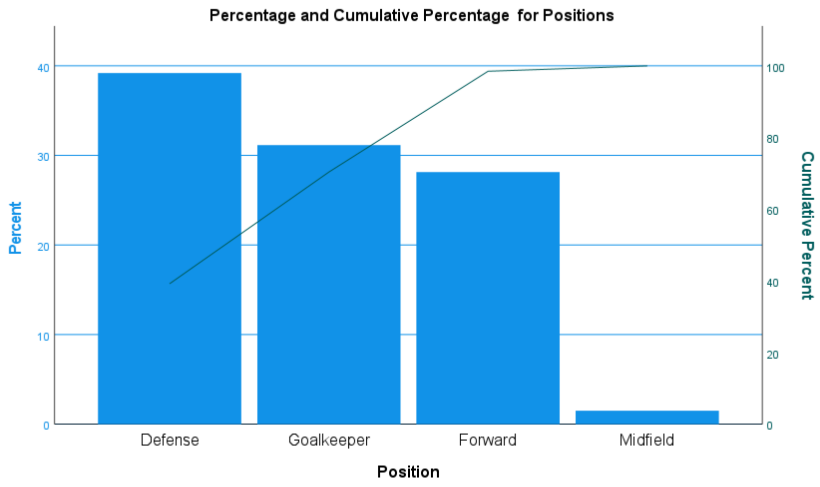


Figure 3: Pareto Diagram 1.

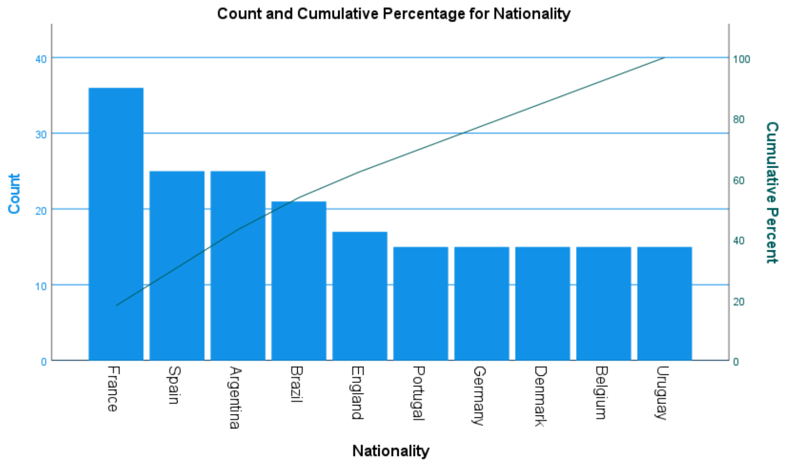
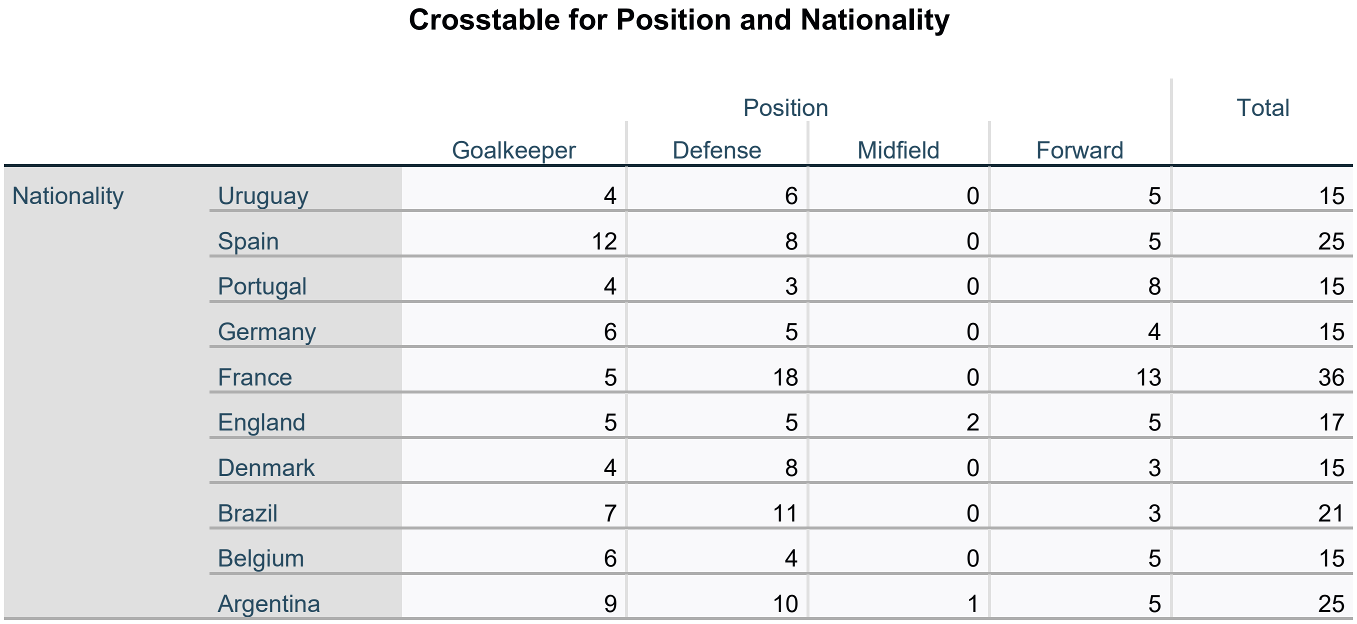


Figure 4: Pareto Diagram 2.

Since the Pareto diagrams are in order from the highest value to the lowest, it is obvious how few midfielders there are, compared to the others in Figure 3. In the same way, countries can make comments by looking at Figure 4, seeing and comparing the countries in front of and behind them, and they can take action accordingly. Thus, it can cause people responsible for football to self-criticize or appreciate themselves. The crosstable below gives a numerical approach to such comparisons.



**Crosstable for Position and Nationality**

| | | Position | | | | Total |
|---|---|---|---|---|---|---|
| | | Goalkeeper | Defense | Midfield | Forward | |
| Nationality | Uruguay | 4 | 6 | 0 | 5 | 15 |
| | Spain | 12 | 8 | 0 | 5 | 25 |
| | Portugal | 4 | 3 | 0 | 8 | 15 |
| | Germany | 6 | 5 | 0 | 4 | 15 |
| | France | 5 | 18 | 0 | 13 | 36 |
| | England | 5 | 5 | 2 | 5 | 17 |
| | Denmark | 4 | 8 | 0 | 3 | 15 |
| | Brazil | 7 | 11 | 0 | 3 | 21 |
| | Belgium | 6 | 4 | 0 | 5 | 15 |
| | Argentina | 9 | 10 | 1 | 5 | 25 |

| | | | | | |
|---|---|---|---|---|---|
| Total | 62 | 78 | 3 | 56 | 199 |

Table 3: Contingency Table.

As seen on the contingency table, we can answer questions such as how many players there are from each country and in which position. What is more, by looking at the table, it is possible to say that the most goalkeepers came from Spain by far. We may claim the same about France for the defensive and forward positions.

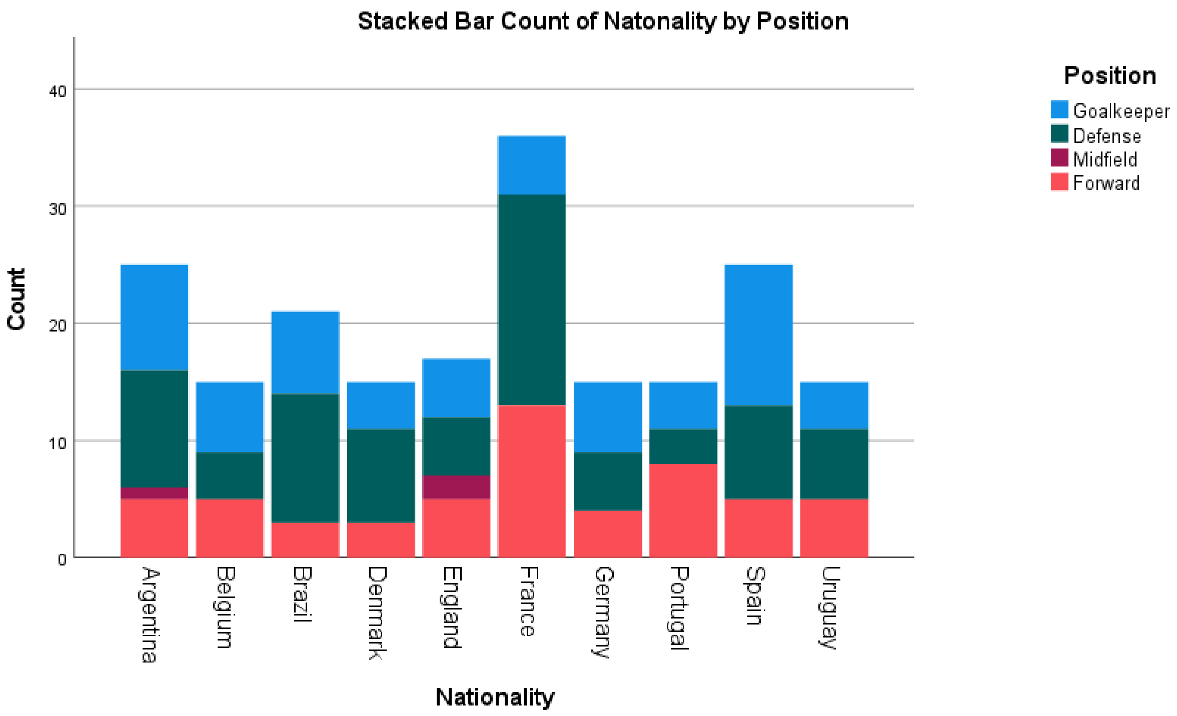


Figure 5: Stacked Bar.

When we examine the stacked bar, we can see how far France is ahead of other countries. Thanks to the stacked bar, countries can compare the distribution of their players by positions in the dataset. It is even possible to compare the number of goalkeepers, defenders, midfielders, and forwards in each country, as each position has its own color. A similar approach may be followed on a clustered bar.

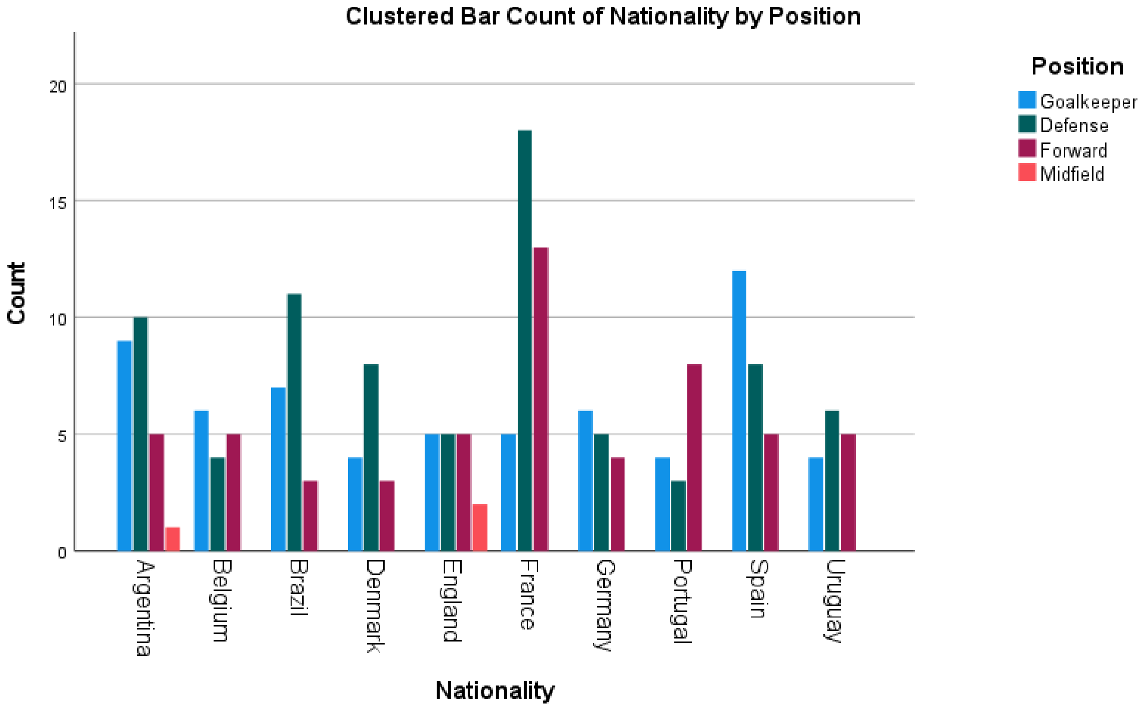


Figure 6: Clustered Bar.

Unlike a stacked bar, the clustered bar is not built on top of each other, but side by side. We can guess the number of players thanks to the lines starting from the left vertical part. For instance, England has five goalkeepers, five defenders, and five strikers. In addition, it is possible to say that there is a big difference between the number of French goalkeepers and defenders. Overall, no other country has a midfielder besides England and Argentina.

## III. Numerical Variables

This part describes the numerical variables - market value and age - by using frequency tables, histograms, ogives, stem and leaf displays, Box-and-Whisker plots, Chebyshev's Theorem, and a scatter plot. To further describe the dataset, five-number-summary, range, IQR, variance, standard deviation, coefficient of variation, covariance, and correlation coefficient are computed. In the end, by using acquired tables and figures, the variables are interpreted with respect to the shape of the distribution and the dispersion of the data.

Since we have numerical data, it is convenient to use binned formats. As the row number of the dataset is between 100-500, classes of 8-10 are applicable. Following, you can find frequency tables and histograms for both binned and unbinned formats of both variables.

**Age**

| | | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 18 | 1 | .5 | 1.0 | .5 |
| | 19 | 4 | 2.0 | 5.0 | 2.5 |
| | 20 | 10 | 5.0 | 15.0 | 7.5 |
| | 21 | 9 | 4.5 | 24.0 | 12.1 |
| | 22 | 24 | 12.1 | 48.0 | 24.1 |
| | 23 | 14 | 7.0 | 62.0 | 31.2 |
| | 24 | 27 | 13.6 | 89.0 | 44.7 |
| | 25 | 31 | 15.6 | 120.0 | 60.3 |
| | 26 | 26 | 13.1 | 146.0 | 73.4 |
| | 27 | 14 | 7.0 | 160.0 | 80.4 |
| | 28 | 12 | 6.0 | 172.0 | 86.4 |
| | 29 | 15 | 7.5 | 187.0 | 94.0 |
| | 30 | 6 | 3.0 | 193.0 | 97.0 |
| | 31 | 4 | 2.0 | 197.0 | 99.0 |
| | 33 | 1 | .5 | 198.0 | 99.5 |
| | 34 | 1 | .5 | 199.0 | 100.0 |
| | Total | 199 | 100.0 | 199.0 | |

Table 4: Frequency Table 3.

**Market Value (10^6 €)**

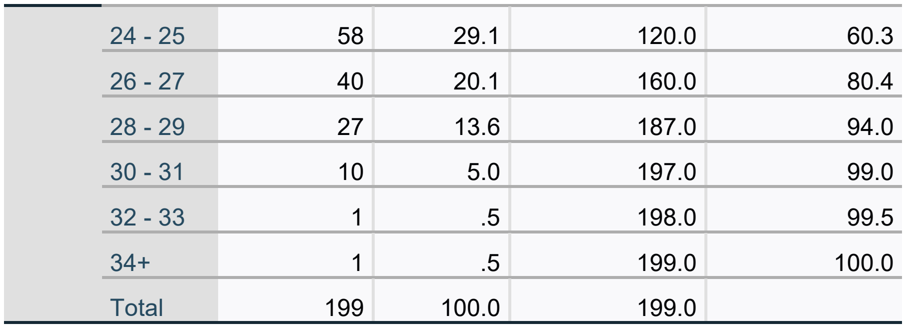
| | | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 14 | 7 | 3.5 | 7.0 | 3.5 |
| | 15 | 1 | .5 | 8.0 | 4.0 |
| | 15 | 27 | 13.6 | 35.0 | 17.6 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  | 16 | 15 | 7.5 | 50.0 | 25.1 |
|  | 17 | 16 | 8.0 | 66.0 | 33.2 |
|  | 18 | 24 | 12.1 | 90.0 | 45.2 |
|  | 19 | 1 | .5 | 91.0 | 45.7 |
|  | 20 | 20 | 10.1 | 111.0 | 55.8 |
|  | 22 | 9 | 4.5 | 120.0 | 60.3 |
|  | 23 | 1 | .5 | 121.0 | 60.8 |
|  | 24 | 1 | .5 | 122.0 | 61.3 |
|  | 25 | 18 | 9.0 | 140.0 | 70.4 |
|  | 27 | 3 | 1.5 | 143.0 | 71.9 |
|  | 28 | 5 | 2.5 | 148.0 | 74.4 |
|  | 30 | 10 | 5.0 | 158.0 | 79.4 |
|  | 32 | 1 | .5 | 159.0 | 79.9 |
|  | 33 | 1 | .5 | 160.0 | 80.4 |
|  | 35 | 13 | 6.5 | 173.0 | 86.9 |
|  | 40 | 6 | 3.0 | 179.0 | 89.9 |
|  | 42 | 1 | .5 | 180.0 | 90.5 |
|  | 45 | 3 | 1.5 | 183.0 | 92.0 |
|  | 48 | 3 | 1.5 | 186.0 | 93.5 |
|  | 50 | 3 | 1.5 | 189.0 | 95.0 |
|  | 55 | 2 | 1.0 | 191.0 | 96.0 |
|  | 60 | 2 | 1.0 | 193.0 | 97.0 |
|  | 65 | 1 | .5 | 194.0 | 97.5 |
|  | 70 | 4 | 2.0 | 198.0 | 99.5 |
|  | 150 | 1 | .5 | 199.0 | 100.0 |
|  | Total | 199 | 100.0 | 199.0 |  |

Table 5: Frequency Table 4.

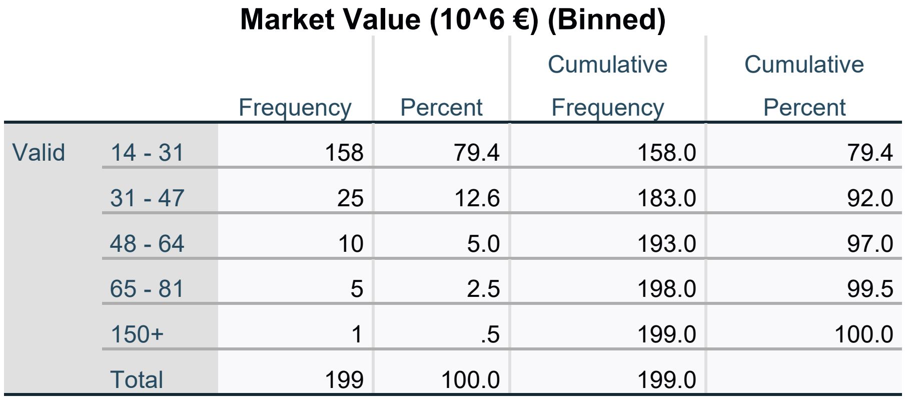
### Age (Binned)

|  |  | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 18 - 20 | 5 | 2.5 | 5.0 | 2.5 |
|  | 20 - 21 | 19 | 9.5 | 24.0 | 12.1 |
|  | 22 - 23 | 38 | 19.1 | 62.0 | 31.2 |



| | | | | | |
|---|---|---|---|---|---|
| | 24 - 25 | 58 | 29.1 | 120.0 | 60.3 |
| | 26 - 27 | 40 | 20.1 | 160.0 | 80.4 |
| | 28 - 29 | 27 | 13.6 | 187.0 | 94.0 |
| | 30 - 31 | 10 | 5.0 | 197.0 | 99.0 |
| | 32 - 33 | 1 | .5 | 198.0 | 99.5 |
| | 34+ | 1 | .5 | 199.0 | 100.0 |
| | Total | 199 | 100.0 | 199.0 | |

Table 6: Frequency Table 5.



**Market Value (10^6 €) (Binned)**

| | | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 14 - 31 | 158 | 79.4 | 158.0 | 79.4 |
| | 31 - 47 | 25 | 12.6 | 183.0 | 92.0 |
| | 48 - 64 | 10 | 5.0 | 193.0 | 97.0 |
| | 65 - 81 | 5 | 2.5 | 198.0 | 99.5 |
| | 150+ | 1 | .5 | 199.0 | 100.0 |
| | Total | 199 | 100.0 | 199.0 | |

Table 7: Frequency Table 6

These tables provide simple yet efficient information about the variables. By looking at the tables, we can directly observe which values are the most frequently occurring ones. For Market Value, it is the group between 14-31 – the lowest value; for Age, it is 24-25, which is towards the middle. These should give us a hint about what we expect to see in the histogram graphs below.

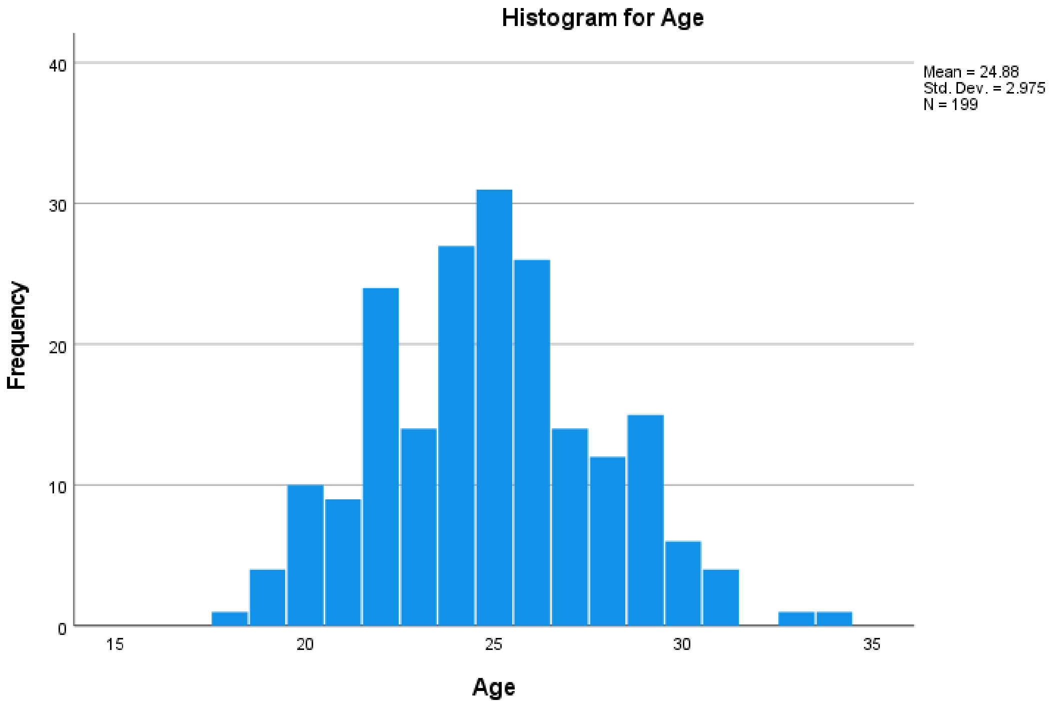


Figure 7: Histogram 1.

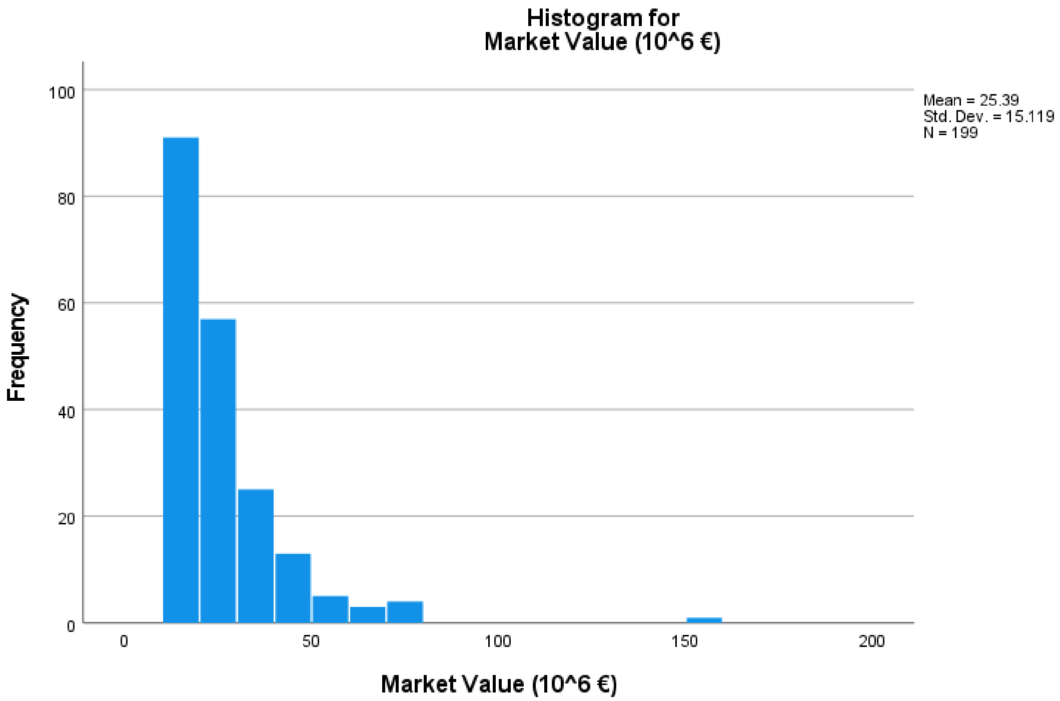


Figure 8: Histogram 2.

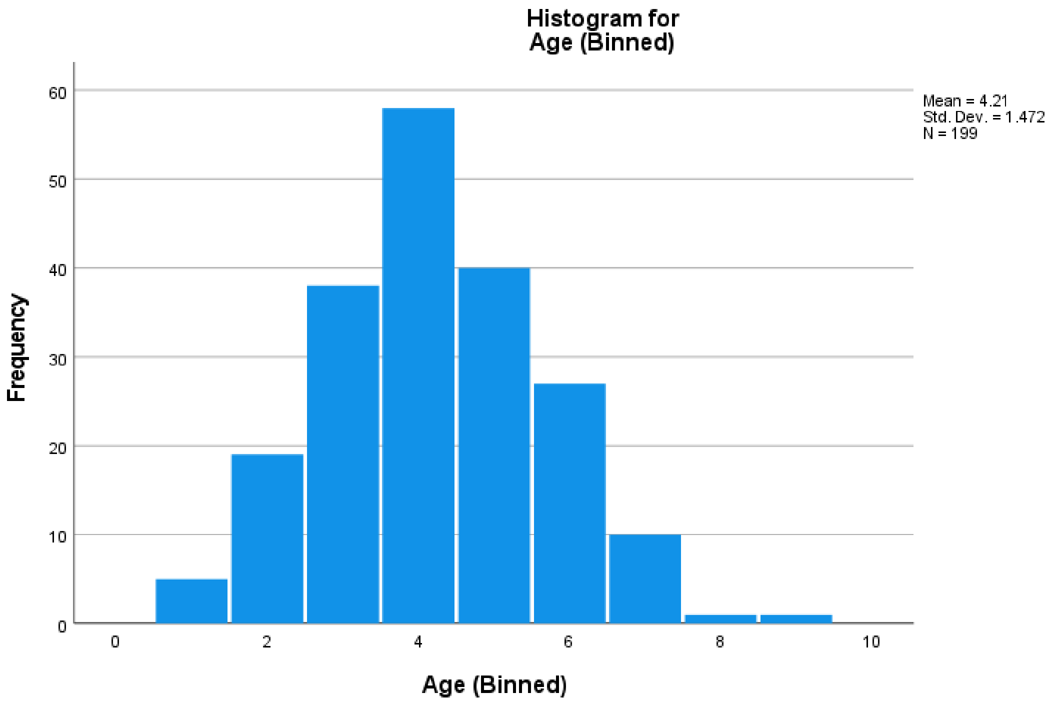


Figure 9: Histogram 3.

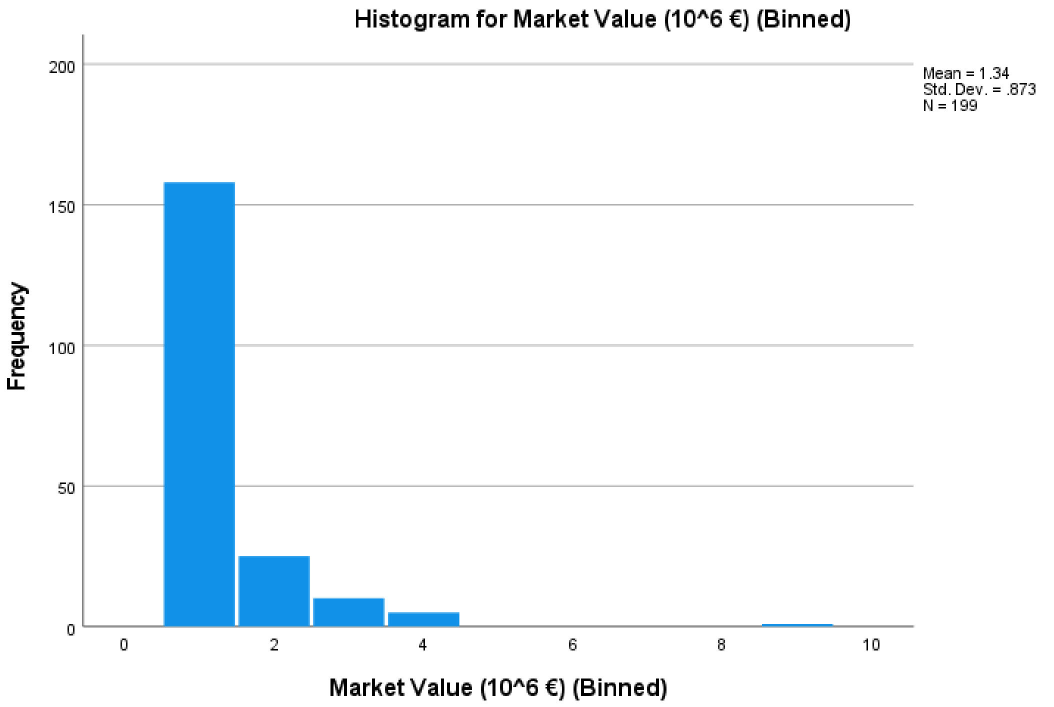


Figure 10: Histogram 4.

The histogram graphs show the frequency of the variables. As seen – and as would have been guessed by the frequency tables – the graphs for Age (Figure 7 and Figure 9) shows a distribution similar to symmetrical, whereas the graphs for Market Value (Figure 8 and Figure 10) are right-skewed. Below, you can find ogive graphs for cumulative percentage frequencies.

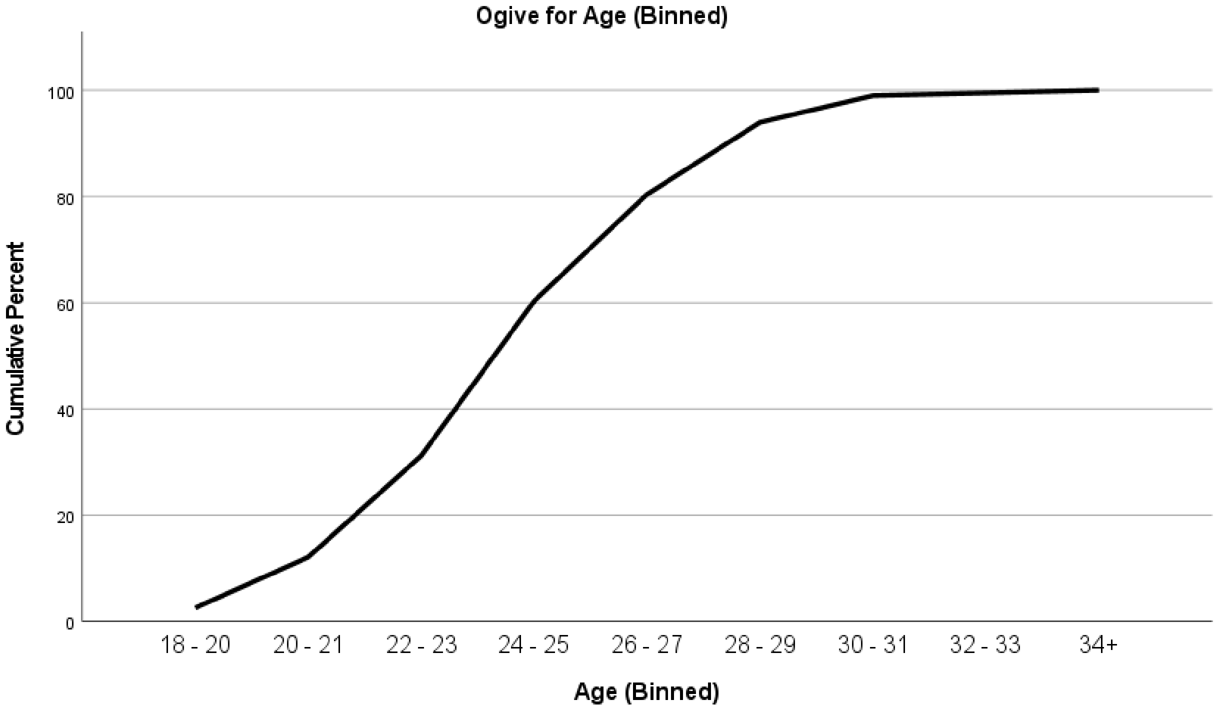


Figure 11: Ogive 1.

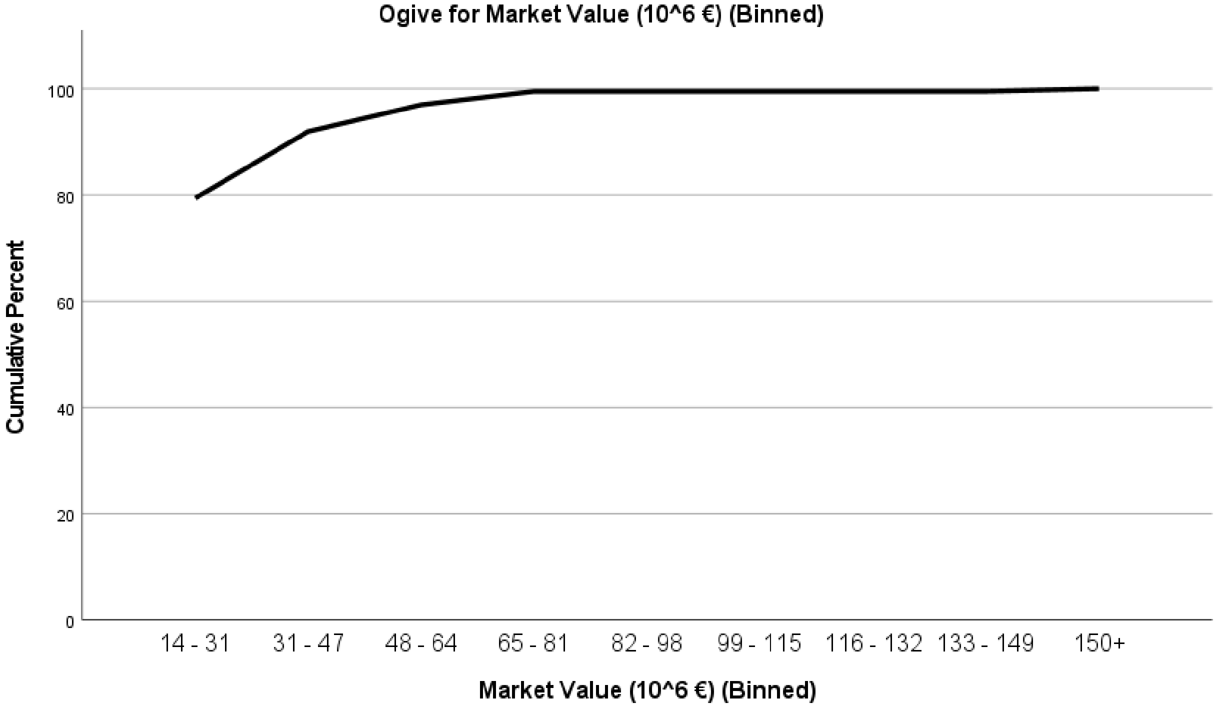
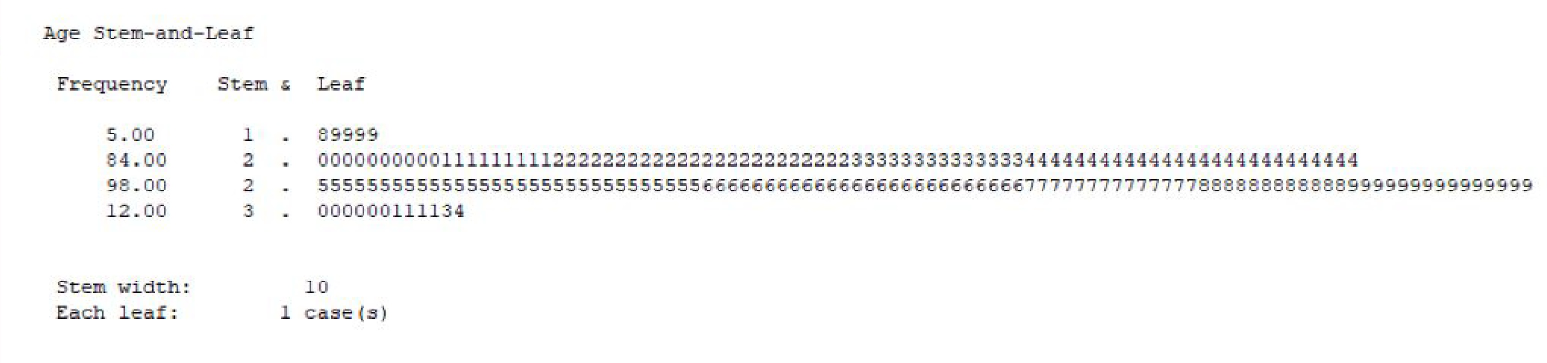


Figure 12: Ogive 2.

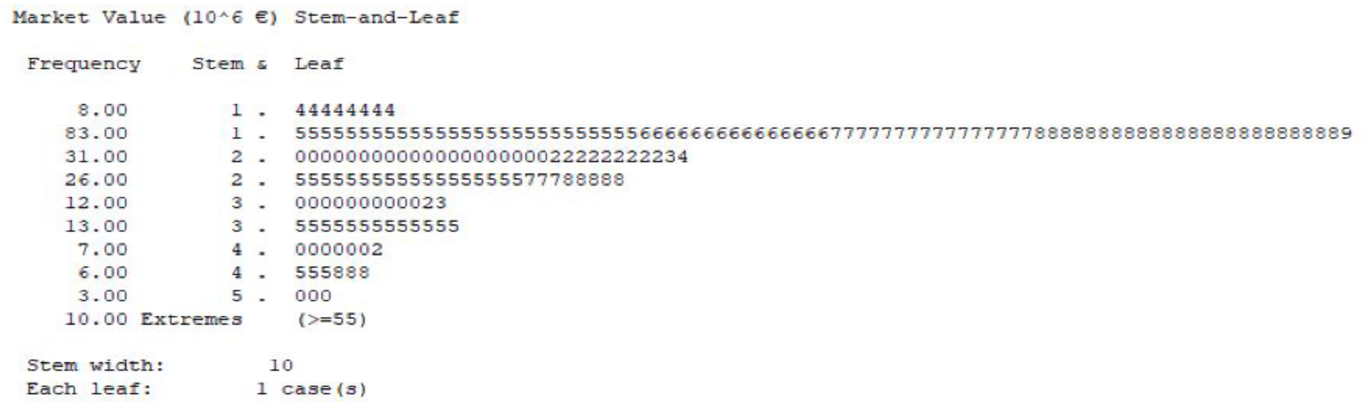
As ogive is a cumulative graph, it allows the observation to be made on all values. In Figure 11, we can spot that the line is steeper for younger ages which leads us to conclude that the dataset focuses more on younger footballers. On the other hand, Figure 12 moves almost as a straight line after the third group which tells us that there are not many footballer transfers made for costs higher than the third group values.

So far, the tables and graphs we have seen do not give us the corresponding values of the data. To overcome any problems that may be caused, we can use stem-and-leaf display to classify the values, but still see them all.



```
 Age Stem-and-Leaf

  Frequency    Stem &  Leaf

      5.00        1 .  89999
     84.00        2 .  000000000011111111122222222222222222222222233333333333333444444444444444444444444444
     98.00        2 .  55555555555555555555555555555556666666666666666666666666667777777777777788888888888889999999999999999
     12.00        3 .  000000111134


  Stem width:         10
  Each leaf:        1 case(s)
```

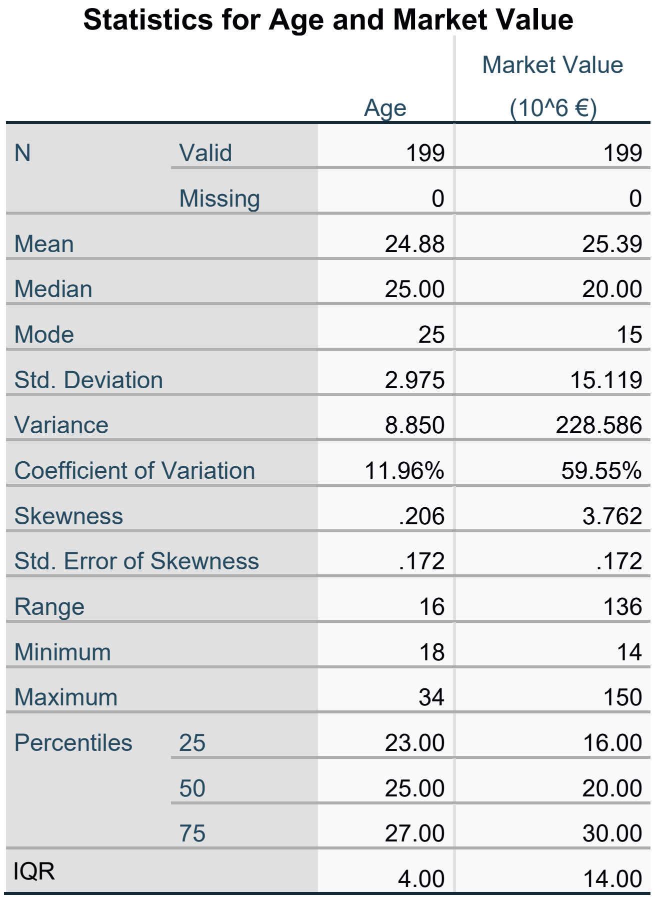
Figure 13: Stem-and-Leaf 1.



```
 Market Value (10^6 €) Stem-and-Leaf

  Frequency    Stem &  Leaf

      8.00        1 .  44444444
     83.00        1 .  55555555555555555555555555566666666666666677777777777777778888888888888888888888889
     31.00        2 .  0000000000000000000022222222234
     26.00        2 .  55555555555555555577788888
     12.00        3 .  000000000023
     13.00        3 .  5555555555555
      7.00        4 .  0000002
      6.00        4 .  555888
      3.00        5 .  000
     10.00 Extremes    (>=55)

  Stem width:        10
  Each leaf:       1 case(s)
```

Figure 14: Stem-and-Leaf 2.

The figures and tables provided above give enough information to briefly describe the data in hand, yet it is not enough. We need more statistical information to comprehend how the data is distributed. Table 8 gives this information for both variables.



**Statistics for Age and Market Value**

| | | Age | Market Value (10^6 €) |
|---|---|---|---|
| N | Valid | 199 | 199 |
| | Missing | 0 | 0 |
| Mean | | 24.88 | 25.39 |
| Median | | 25.00 | 20.00 |
| Mode | | 25 | 15 |
| Std. Deviation | | 2.975 | 15.119 |
| Variance | | 8.850 | 228.586 |
| Coefficient of Variation | | 11.96% | 59.55% |
| Skewness | | .206 | 3.762 |
| Std. Error of Skewness | | .172 | .172 |
| Range | | 16 | 136 |
| Minimum | | 18 | 14 |
| Maximum | | 34 | 150 |
| Percentiles | 25 | 23.00 | 16.00 |
| | 50 | 25.00 | 20.00 |
| | 75 | 27.00 | 30.00 |
| IQR | | 4.00 | 14.00 |

Table 8: Statistics.

Table 8 gives us the necessary information for numerical descriptions including both central tendency measures and variation. By using Table 8, we can form a five-number-summary for each variable.

For Age,

18.00 < 23.00 < 24.88 < 27.00 < 34.00

and for Market Value (10^6 €),

14 < 16.00 < 25.39 < 30.00 < 150

Observing the five-number-summary, it is easy to see that Market Value variable has a bump to 150. To investigate this view, we can check the range for Market Value as the range is sensitive to outliers. The range for Market Value is 136 (Table 8) which supports the idea of a bump. Since IQR is 14 (Table 8) and is not near the value of the range, we may interpret that Market Value has outliers.

Moving to the five-number-summary of Age, the numbers are much closer to each other and seem precise. As the range and IQR (Table 8) do not suggest a highly imprecise situation, we may conclude that Age has more consistent values than Market Value. A comparison of standard deviation, variance, and coefficient of variation from Table 8 would support this conclusion. This may take us back to the observation we made while scanning the histogram graphs. As the values of the variable Age are gathered closer to the median - which is equal to the mean in 2 significant figures- and its skewness is close to 0, the data for Age is almost symmetric. On the other hand, the median of Market Value is further than its mean, and the values accumulate closer to the minimum value. Hence, our observation of right-skewness for the histogram graph of Market Value is justified by the numerical data.

Figure 15 and Figure 16 visualize five-number-summary as a Box-and-Whisker plot which exhibits the presence of outliers and the point where the data accumulates.

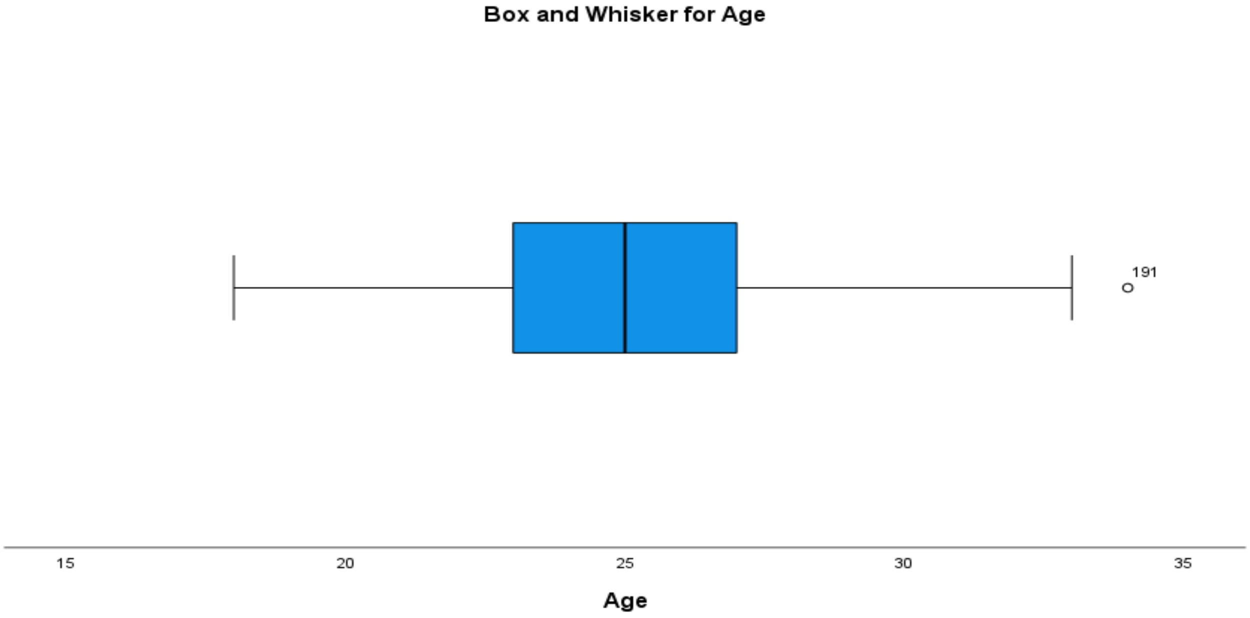


Figure 15: Box and Whisker 1.

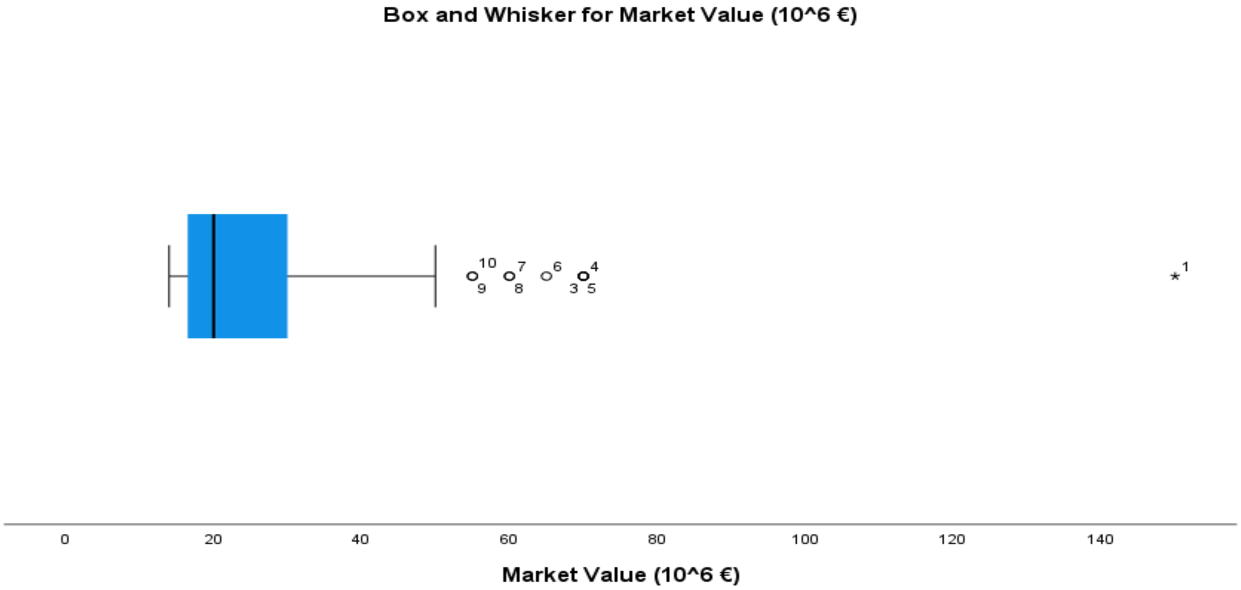
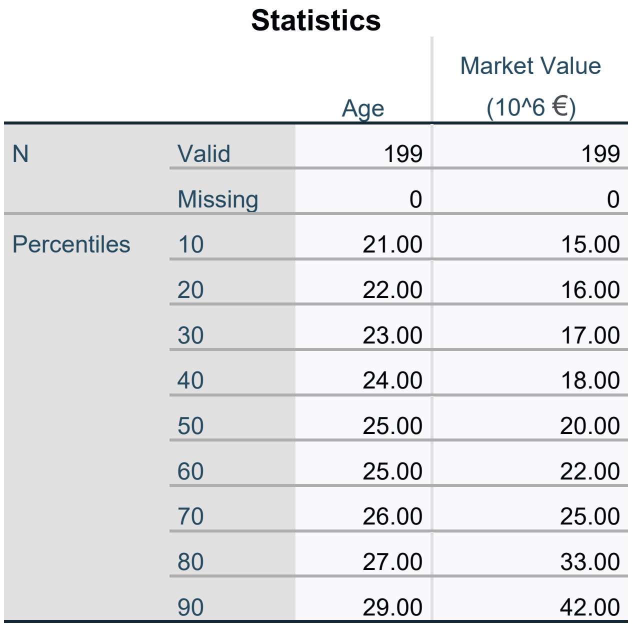


Figure 16: Box and Whisker 2.

Due to the distribution of the data, it might be better to use the mean as a measure of central tendency for Age, yet the mode or the median for Market Value as mean is sensitive to outliers.

Before finishing off with numerical measurements, let us use Chebyshev's Theorem to test the data one more time. Let's use k = 2 and calculate the percentage of observations that fall within the range of Chebyshev's Theorem.

Using $\mu \pm 2\sigma$ for the interval and $100(1-\left(\frac{1}{2^2}\right))$ which equals to 75% for the percentage of observations within the interval, we obtain $24.88 \pm 5.95 = [18.93, 30.83]$ for Age interval and $25.39 \pm 30.24 = [-4.85, 55.63]$ for Market Value interval. Table 9 shows that both intervals contain more than 75% of values, hence Chebyshev's Theorem holds.



**Statistics**

| | | Age | Market Value (10^6 €) |
|---|---|---|---|
| N | Valid | 199 | 199 |
| | Missing | 0 | 0 |
| Percentiles | 10 | 21.00 | 15.00 |
| | 20 | 22.00 | 16.00 |
| | 30 | 23.00 | 17.00 |
| | 40 | 24.00 | 18.00 |
| | 50 | 25.00 | 20.00 |
| | 60 | 25.00 | 22.00 |
| | 70 | 26.00 | 25.00 |
| | 80 | 27.00 | 33.00 |
| | 90 | 29.00 | 42.00 |

Table 9: Percentile statistics.

So far, we have investigated the variables independent of each other. Now let's make a scatter plot and see if and how Age and Market Value relate.

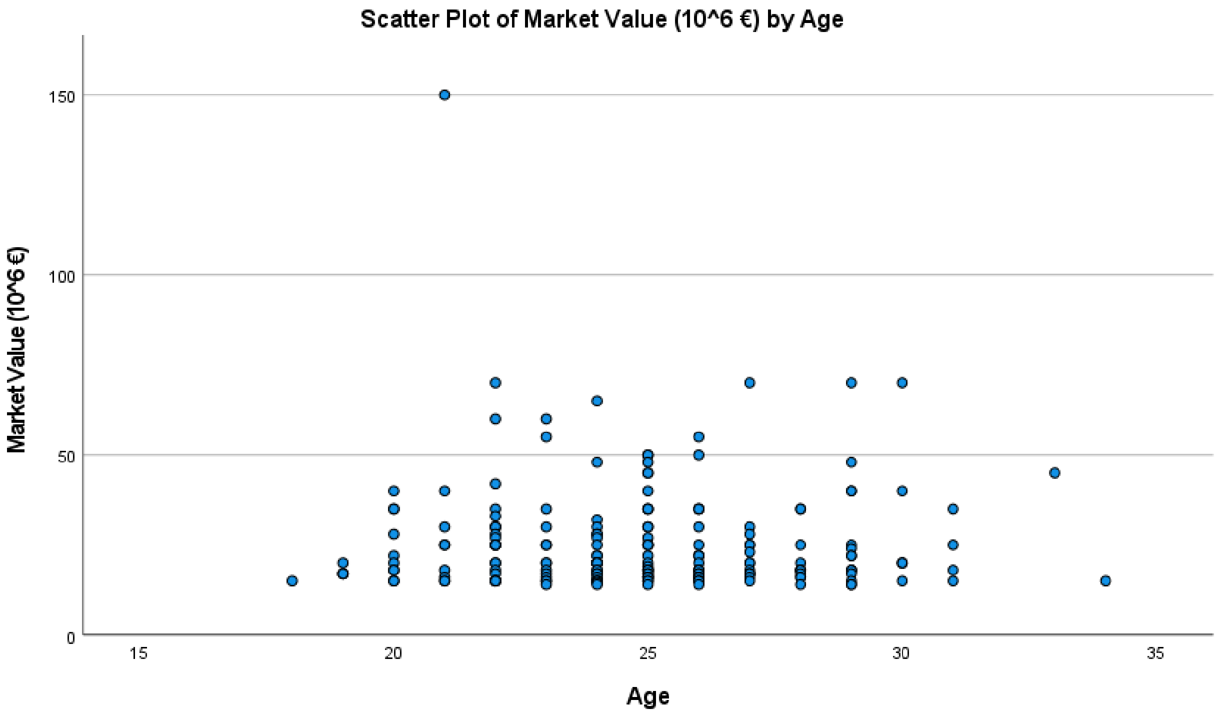


Figure 17: Scatter Plot

Figure 17 shows the relationship between Market Value and Age. If we aimed to draw a best-fit-line, it would be close to a straight line, and we would not be able to attribute an apparent relationship to the graph. Table 10 below gives us numerical data that describes this non-apparent relationship between the two variables.

**Correlations**

| | | Age | Market Value (10^6 €) |
|---|---|---|---|
| Age | Pearson Correlation | 1 | -.021 |
| | Sig. (2-tailed) | | .767 |
| | Sum of Squares and Cross-products | 1752.342 | -188.543 |
| | Covariance | 8.850 | -.952 |
| | N | 199 | 199 |
| Market Value (10^6 €) | Pearson Correlation | -.021 | 1 |
| | Sig. (2-tailed) | .767 | |
| | Sum of Squares and Cross-products | -188.543 | 45260.068 |
| | Covariance | -.952 | 228.586 |
| | N | 199 | 199 |

Table 10: Correlations.

According to Table 10, the correlation coefficient is -0.021 which supports the idea of a non-apparent relationship. Since this number is near 0, the relationship is very weak. Both covariance and correlation coefficient are negative numbers which is a result of a negative relationship (although weak) and may be verbally expressed as follows: As Age increases, Market Value decreases.

In this part, we have analyzed the two numerical variables of the dataset. We have found out that Age has a close-to-symmetrical distribution, and Market Value is right-skewed. We have also observed that Market Value was rich in outliers, and mean would be a wrong choice to measure its central tendency, but it would be a good choice for Age. Lastly, we have investigated the relationship between Age and Market Value and found a very weak negative relationship that is close to no-relationship.

## IV. Conclusion

In this report, we have evaluated the dataset on transferred footballers in two parts based on the types of variables. The first part on categorical variables has exhibited graphs and tables on the distribution of the nationalities and positions of the players. The second part has laid out the age and market value of the players and compared them. Using the first part, countries that want to develop in the football market may decide whether they should invest more in a position or not. The second part has shown that there is no correlation between the age and market value of footballers, however, it may still be used to generalize the fact that not many footballers over the age of 30 are in the market or are transferred. The interpretation of the data will differ by the question asked yet the amount of data in the dataset may not be sufficient to justify any judgments that are based on the dataset. As mentioned above, this dataset is an extraction and arrangement of a much larger dataset of about 40,000 rows. Our sample dataset includes only 199 rows of this large set. Hence, if we need to make a better judgment on footballer transfers, it would be appropriate to use the whole dataset of population instead of this small sample. Yet, a sample may still infer information about its population. Thus, the small size of the sample dataset we have should not create ambiguity for the conclusions we have made. Another restriction in our report is the lack of comparison between the categorical variables and numerical variables which can be overcome by another part that investigates the dataset intertype. The absence of the improvements discussed above does not discredit the conclusions made based solely on the sample dataset but suggest a way to ameliorate. Hence, overall, the aim of this report has been accomplished as we have used the sample dataset and explained it statistically as mentioned in the introduction.

## Bibliography

- Sayed, M. (2022, October). *Football Transfer Window from July to September*. Retrieved November 2022, from Kaggle: https://www.kaggle.com/datasets/mohamedsiika/football-transfer-window-from-july-to-september