Data Quality Report

Machine learning for predictive analytics |

Assignment 1

K00228585 - Sean Daly

2021

Table of Contents

[**Q1 Continuous Features:** 2](#_Toc68520033)

[**Q2 Categorical Features:** 2](#_Toc68520034)

[**Q3 Histograms of each feature** 2](#_Toc68520035)

[**Q4 Discussion of each histogram** 6](#_Toc68520036)

[**Position** 6](#_Toc68520037)

[**Height & Weight** 6](#_Toc68520038)

[**Career Stage** 6](#_Toc68520039)

[**Age** 6](#_Toc68520040)

[**Sponsorship Earnings** 6](#_Toc68520041)

[**Shoe Sponsor** 6](#_Toc68520042)

[**Q5 Identify data quality issues** 7](#_Toc68520043)

[**Q6 Relationships between each feature:** 8](#_Toc68520044)

[**Weight** 8](#_Toc68520045)

[**Age** 8](#_Toc68520046)

[**Height** 8](#_Toc68520047)

[**Sponsorship Earnings** 8](#_Toc68520048)

# **Q1 Continuous Features:**

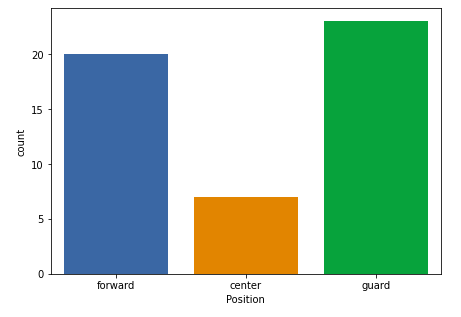
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Feature* | *Count* | *Miss* | *Card* | *1st Qrt* | *Mean* | *Median* | *3rd Qrt* | *Std* | *Min* | *Max* |
| **Height** | **50** | **4%** | **23** | **182** | **188.666** | **190** | **198** | **14.98** | **158** | **218** |
| **Weight** | **50** | **10%** | **21** | **199** | **213.11** | **215** | **223** | **19.37** | **183** | **251** |
| **Age** | **50** | **6%** | **13** | **22** | **31.234** | **27** | **29** | **42.14** | **-27** | **311** |
| **Earnings** | **50** | **0%** | **30** | **314** | **776.22** | **648** | **1347.25** | **549.79** | **12** | **1855** |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Feature* | *Count* | *Miss* | *Card* | *Mode* | *Mode Frequency* | *Mo. Per* | *2nd Mode* | *2nd Freq* | *2nd Mode %* |
| **Position** | **50** | **0%** | **3** | **Guard** | **23** | **46%** | **Forward** | **20** | **40%** |
| **Stage** | **50** | **8%** | **4** | **Mid-Career** | **19** | **38%** | **Rookie** | **18** | **36%** |
| **Sponsor** | **50** | **6%** | **3** | **No** | **25** | **50%** | **Yes** | **22** | **44%** |

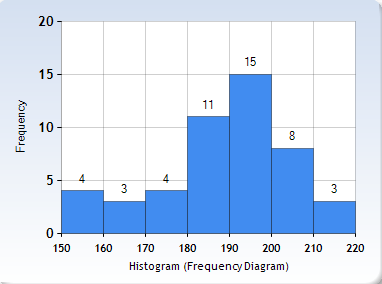
## **Q2 Categorical Features:**

## **Q3 Histograms of each feature**

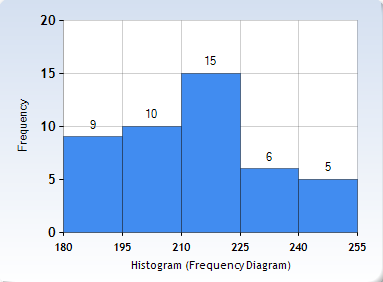
**Position**



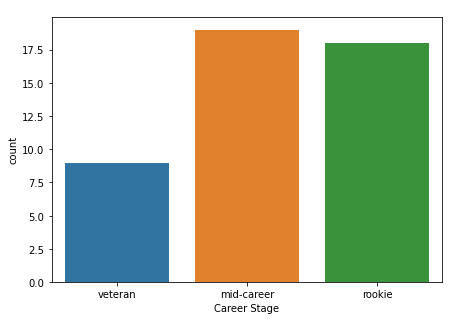
**Height**



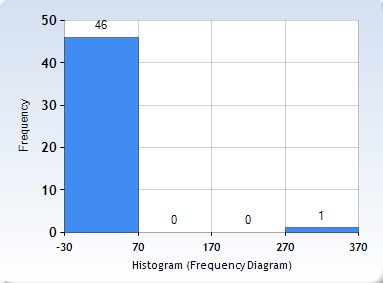
**Weight**



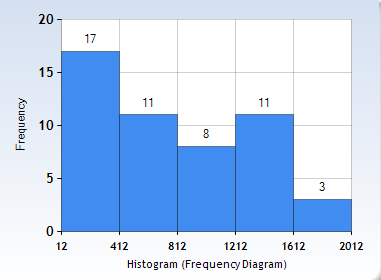
**Career Stage**



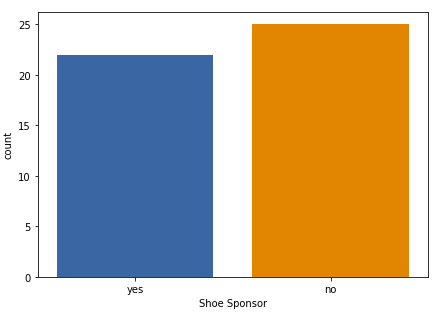
**Age**



**Sponsorship Earnings**



**Shoe Sponsor**



## **Q4 Discussion of each histogram**

**Position** – There are three positions: forward, center position and guard in the data with different counts for each result. As you can see the count for forward is 20. The count for the center position is around 7.5 and the guard is 23. This bar graph displays these results very neatly.

**Height & Weight** – The height and weight range from 158 to 218, and 183 to 251. The concentration is around a height of 190 and a weight of 215.

**Career Stage** – Most of the players are ‘rookies’ or ‘mid-career’. It is worth mentioning they are missing 4 values. These 4 missing values impacted the results.

**Age** – The values are between 19 & 35. There are two outliers: one negative value and one very large value. On the left, it is the histogram plot for the Age after removing the outliers.

**Sponsorship Earnings** - The sponsorship earnings vary between 12 and 1855. This is a large gap between the lowest and the highest value.

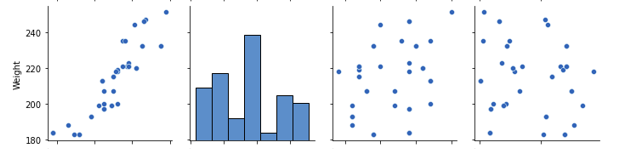
**Shoe Sponsor** – There are 22 players that have a shoe sponsor, 25 do not. There are 3 rows without values that do not appear on the histogram.

## **Q5 Identify data quality issues**

|  |  |  |
| --- | --- | --- |
| ***Feature*** | ***Data quality issue*** | ***Potential handling strategies*** |
| ***Age*** | Outliers: one negative value and one very big value | * Because both are irregular data for ages, the specified issue is an outlier and falls under the category of invalid outliers. * The best way to solve this is to remove all outliers because they aren't a good representation of the results. |
| ***Age*** | 3 Missing values | - impute with the mode  - Delete the row |
| ***Height*** | 2 Missing values / 4%  Cardinality is 7 | * Binning could help with both issues; missing data isn't necessarily a bad thing, but the cardinality could be reduced if the data is binned into broader ranges, resulting in a more uniform distribution. The issue with binning is that it may remove some data. * Range normalization may be preferable because it can help with missing data. * Equal-width binning could also be a viable solution in this case. |
| ***Weight*** | 5 Missing values / 10% | Missing data is common in weight categories because it may be purposefully missing. |
| ***Career Stage*** | 4 Missing values / 8% | This should be noted, but the data may still be accurate and do not need further processing. |
| ***Shoe Sponsor*** | 3 Missing values / 6% | This should be noted, but the data may still be accurate and do not need further processing. |

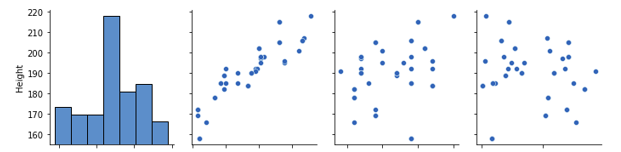
## **Q6 Relationships between each feature:**

## **Weight**

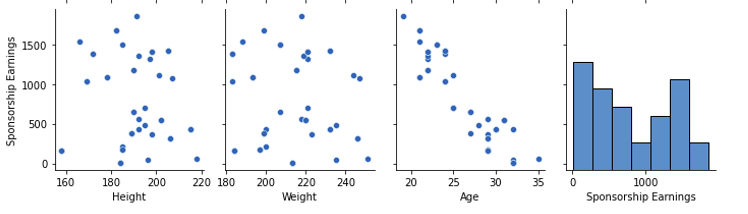


### **Age**

### **Height**



### **Sponsorship Earnings**



In this diagram you can see the linear relationship between these pairs of variables:

* Age and Sponsorship earning
* Height and Weight

The scatter plot of the other pairs of variables do not give us much information.