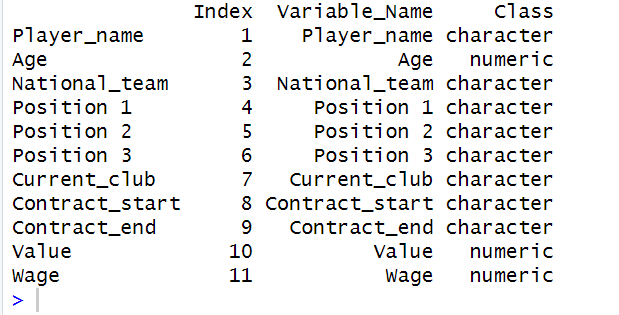
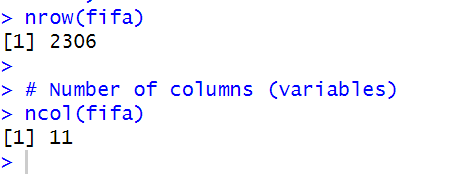
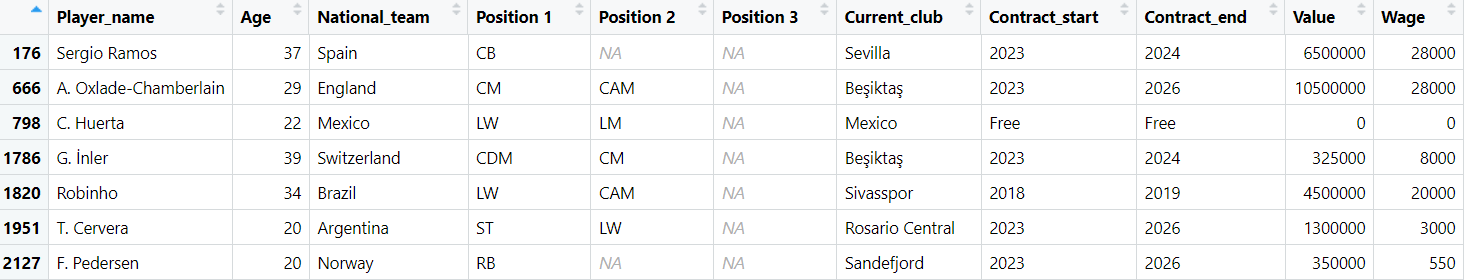
2.1a) The variables names, their index, and their R class.



2.1b) The size of the variable vectors.

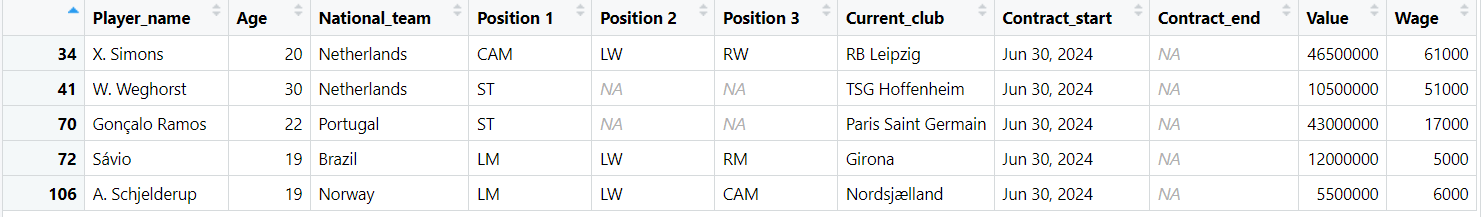
  
  
2.1c) The values in the first 6 records/rows for each variable in a table (labelled  
with row index and variable names)  


2.2a) duplicates: 7 records found



Create a variable which shows only the duplicated players, use the duplicated function to find only the duplicate players

2.2b) Records with non-year contract dates (start, end): 248 records

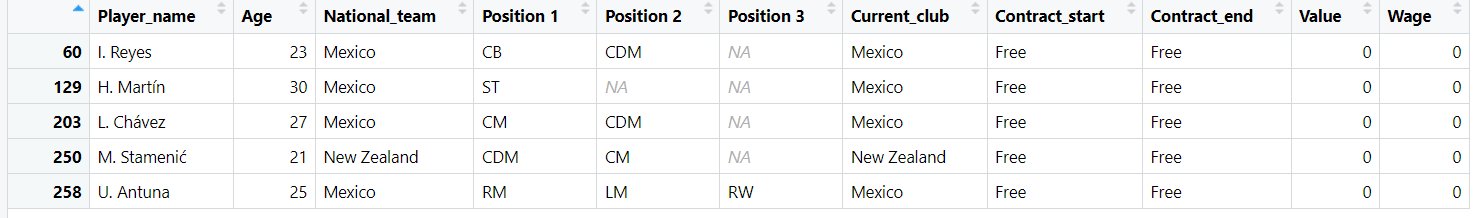


Filtering process:

Create a variable which shows only the players where the contract start and contract end are not 4 digits, an actual contract will have 4 digits eg 2029

2.2c) zero value salary: 58 records found

First 5 records:

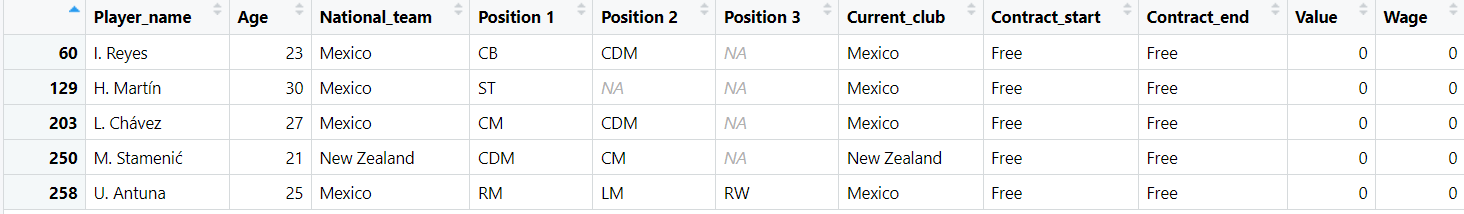


Filtering process:

Create a variable that shows only the players from the fifa file where wage == 0

2.2d) zero value value: 62 records found

First 5 records:



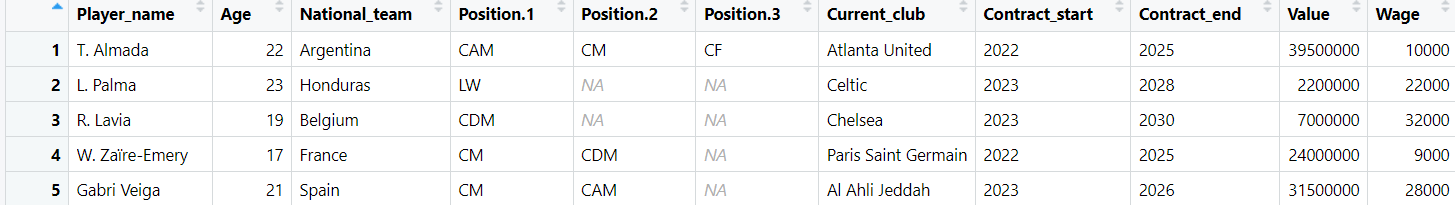
Create a variable that shows only the players from the fifa file where value == 0

2.2e) fake players:



2.4a) All players – 1992 records found

Sample of 5 records:



Filtering Process – since we cleaned out all the wrong players, the new file would just contain all the players left so we create a variable <- clean\_fifa

2.4b) players that can play CAM – 318 records found

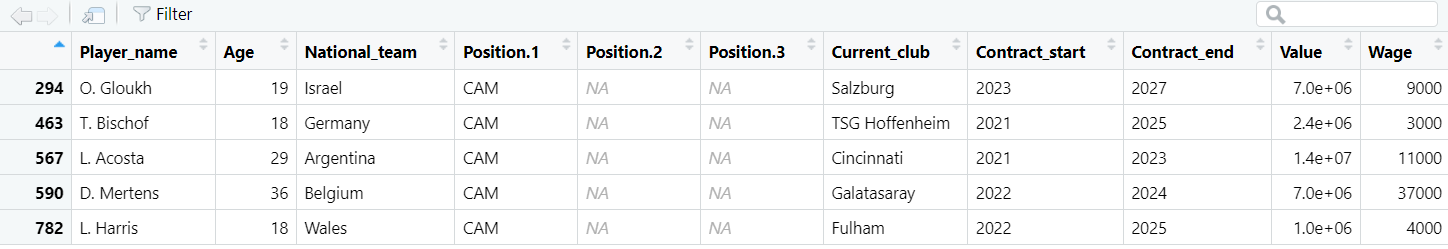
Sample of 5 records:



Filtering process: only included players with cam, cam could be in position1,2 or 3

2.4c) CAM only - 14 records found

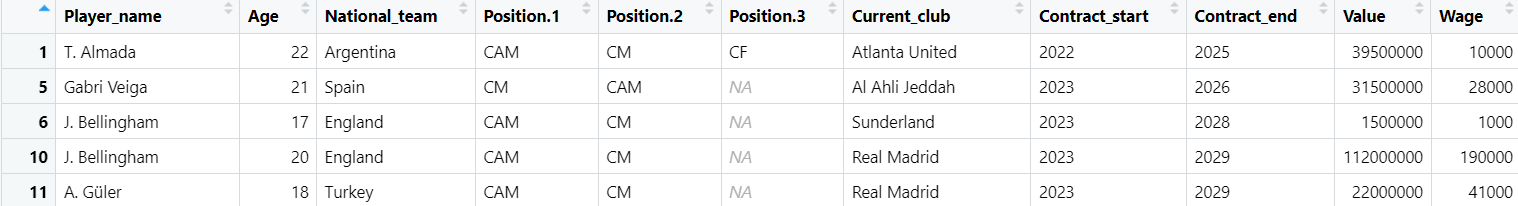
Sample of 5 records:



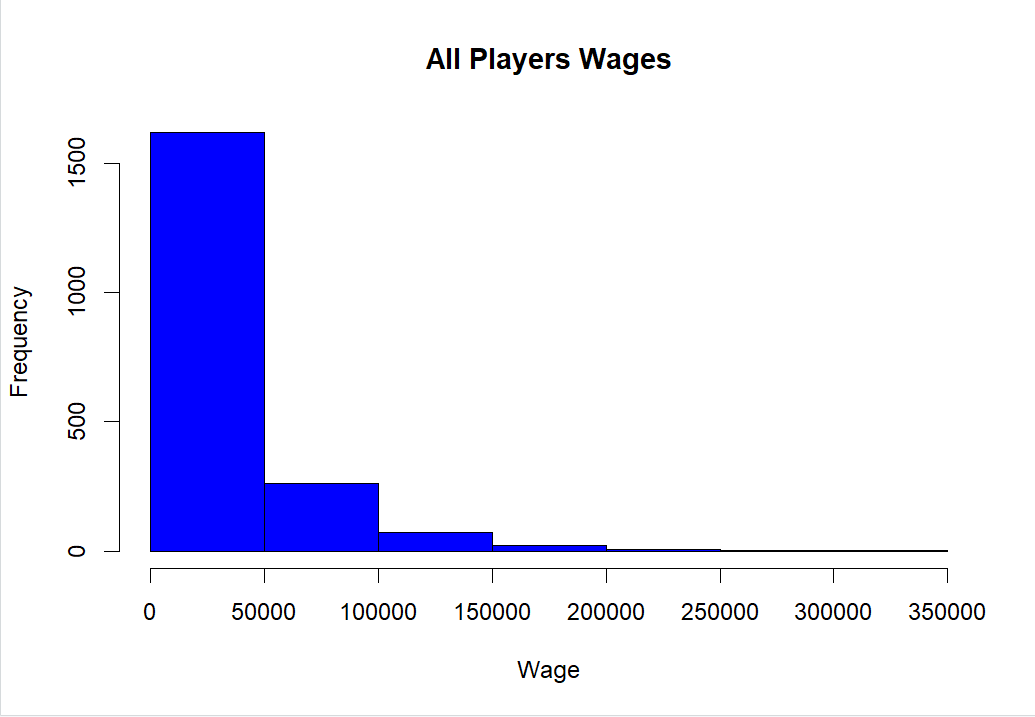
Filtering process – only include players with position1 = cam, position 2 and 3 = NA

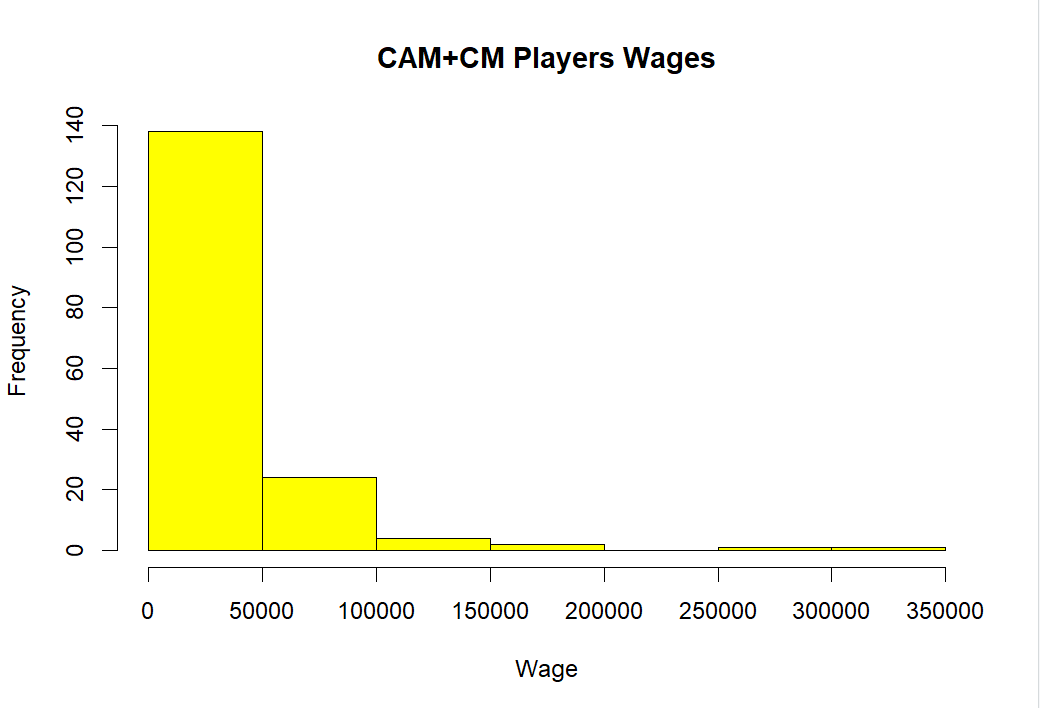
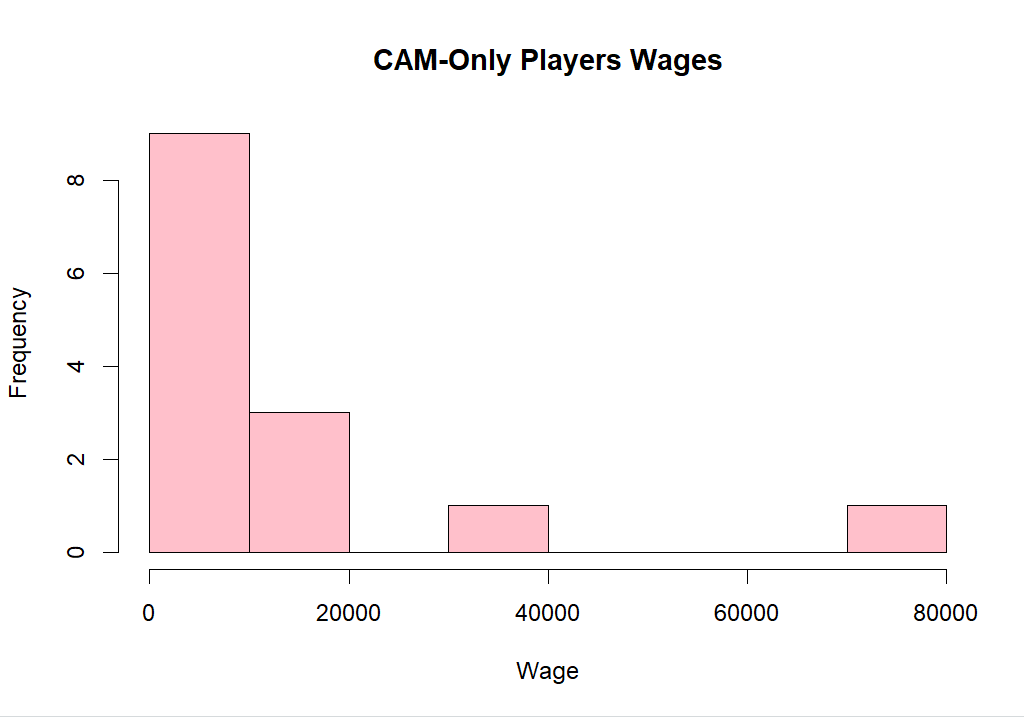
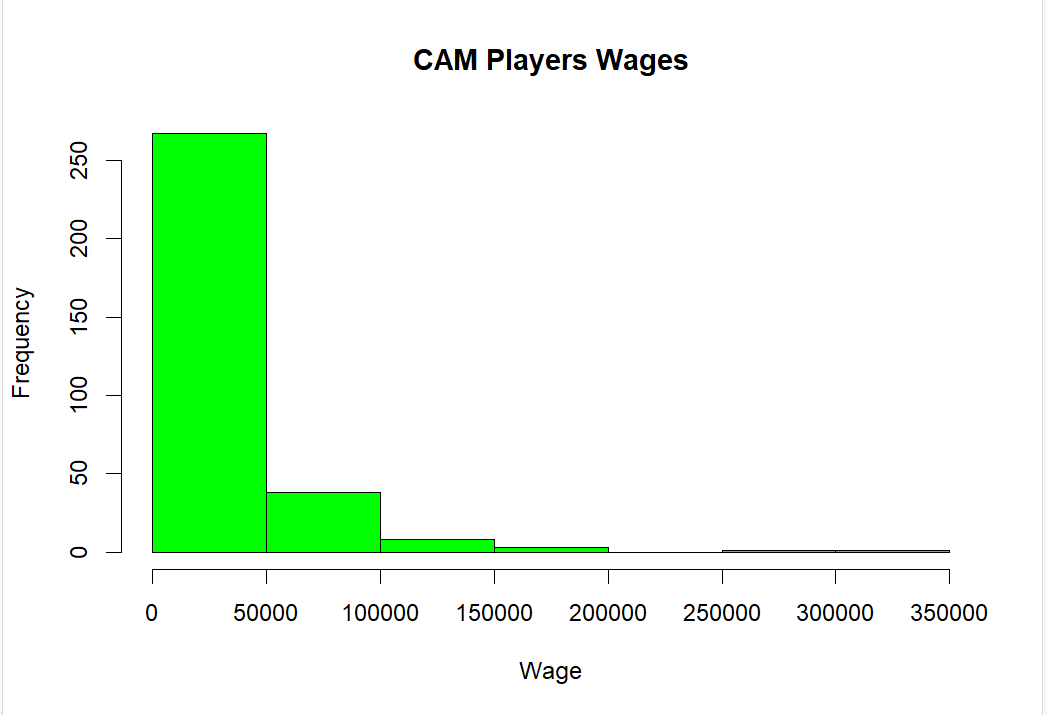
2.4d) CAM and CM – 170 records found

Sample of 5 records:

Filtering process: include players that have cam in position1,2 or 3 AND have cm in position 1,2 or 3

2.5) most football players earn between 0-50,000. High wage outliers are found in CAM players and CAM only. Most CAM only players don't earn as much as the highest is 80,000 which is an outlier. All players and CAM players are similar in wage and shape, but the CAM group has a few high earners. All players, CAM, CAM+CM show a slightly skewed distribution whilst CAM only is a strong right skew. In CAM +CM there are also a few outliers earning high in wages.





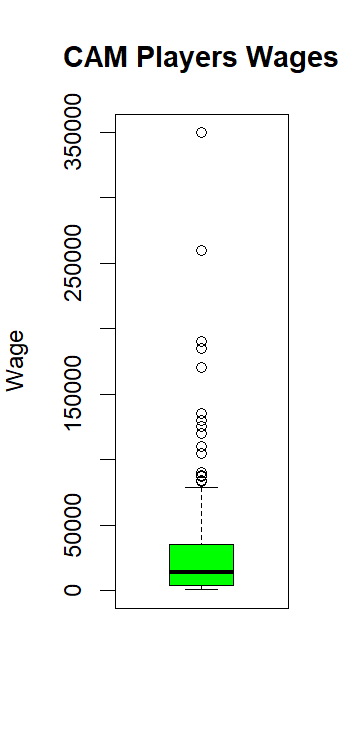
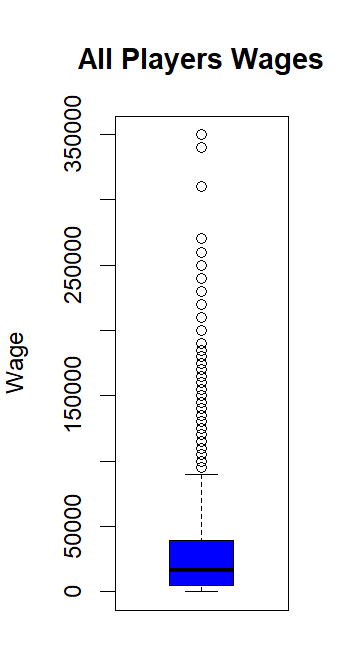
2.6) boxplots:

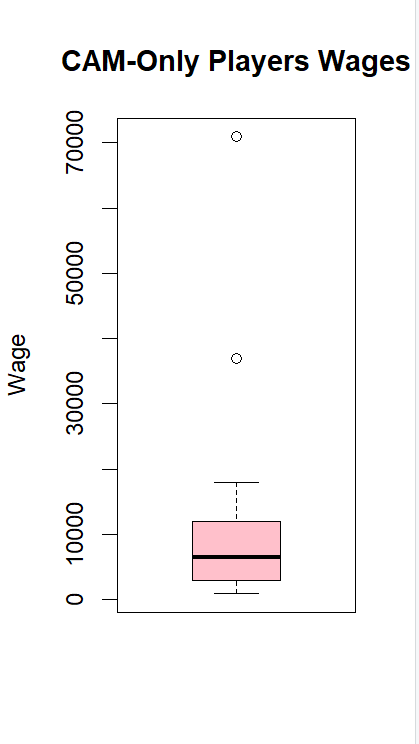
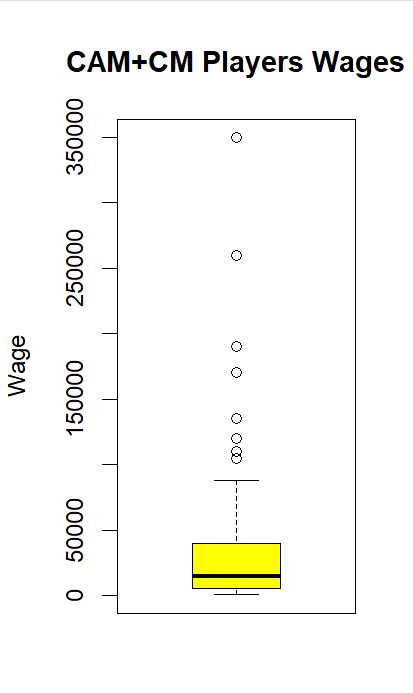
For all players – the median is around 20,000, right skewed meaning most players earn high wages, there are many extreme outliers

For CAM players – the median is around 40,000 which is higher than the all-players median, wider IQR meaning more variability, there are outliers but less than all-players

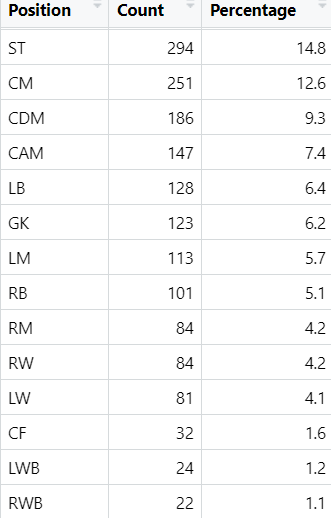
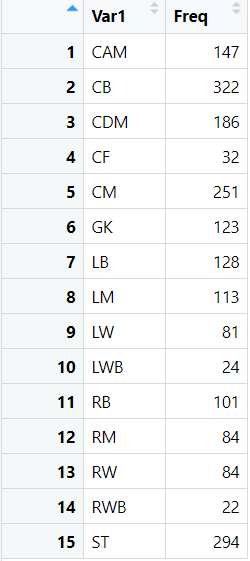
For CAM only – median is similar to CAM, whiskers are shorter, very few outliers

For CAM+CM – median is the highest, symmetrical IQR, longer upper whisker so some earn much more, very few outliers

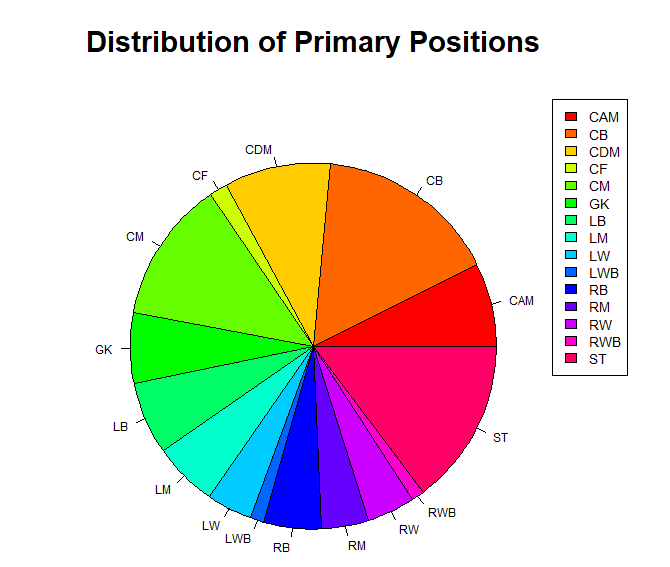




2.7a) players per primary position 2.7b) Results in frequency table



2.7c) pie chart



2.7d) the highest count for a position is ST with 294 counts which is 14.8% of all players whilst the lowest count for a position is RWB with 22 counts which us 1.1% of all players