HOMEWORK 5

a. Data gathering and integration

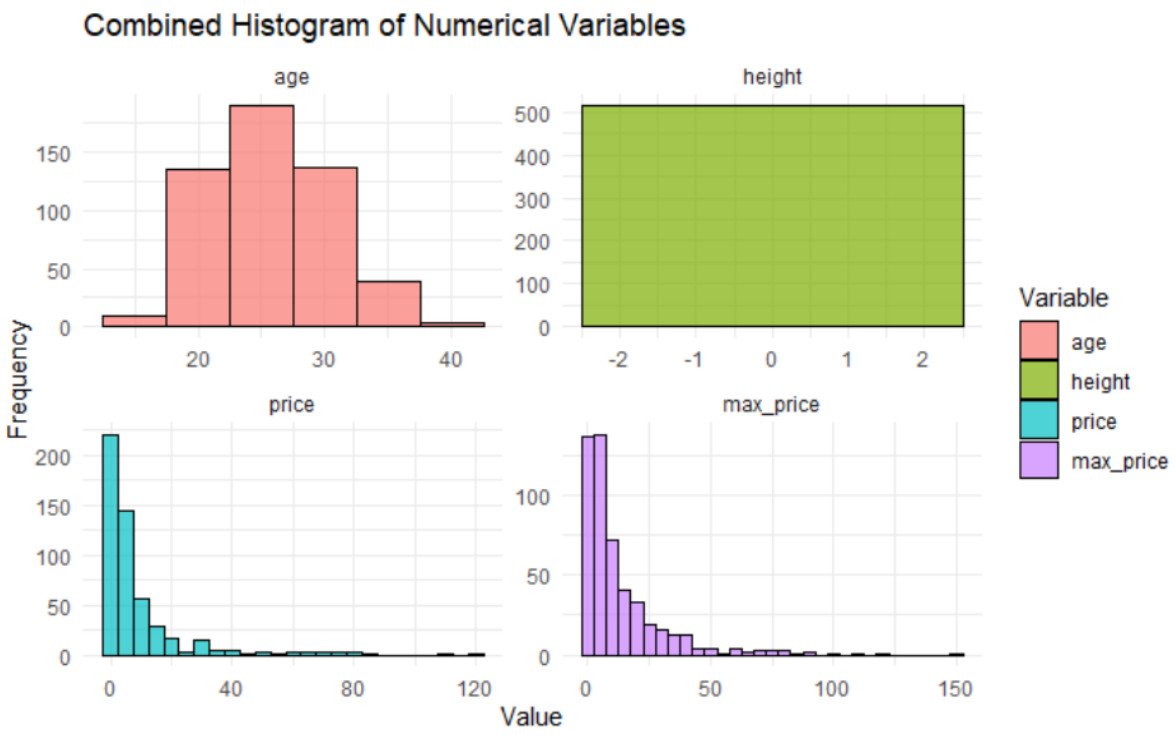
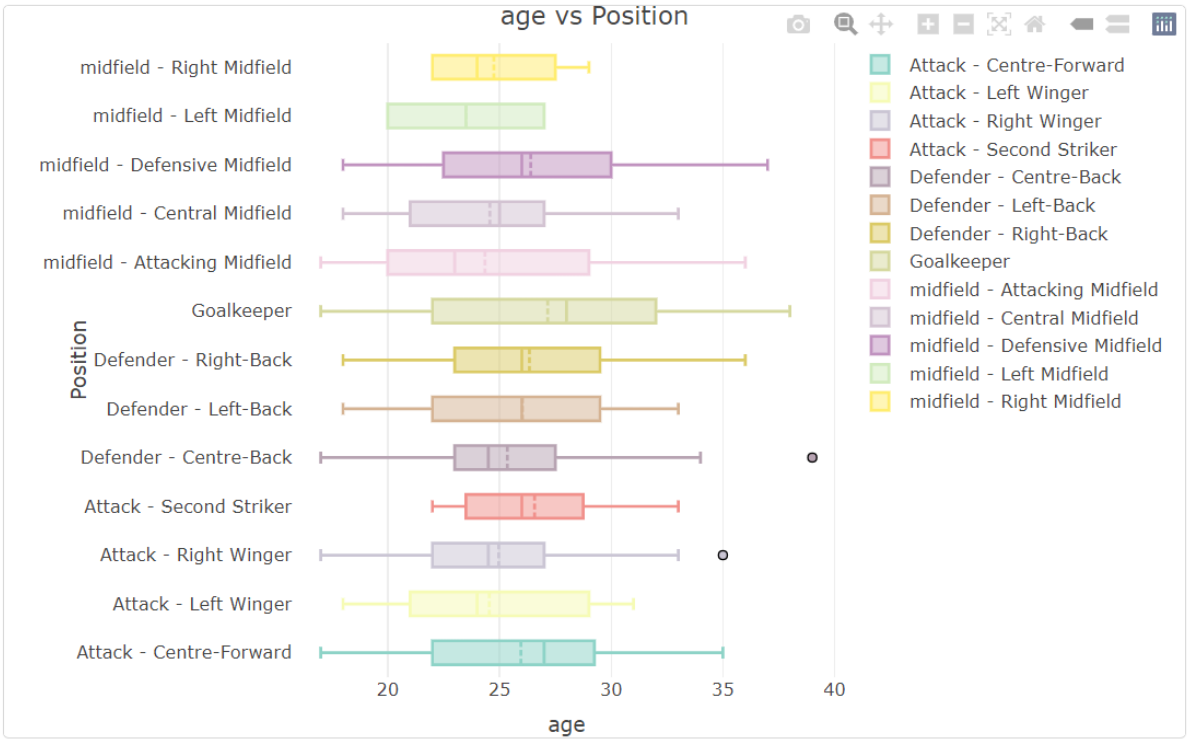
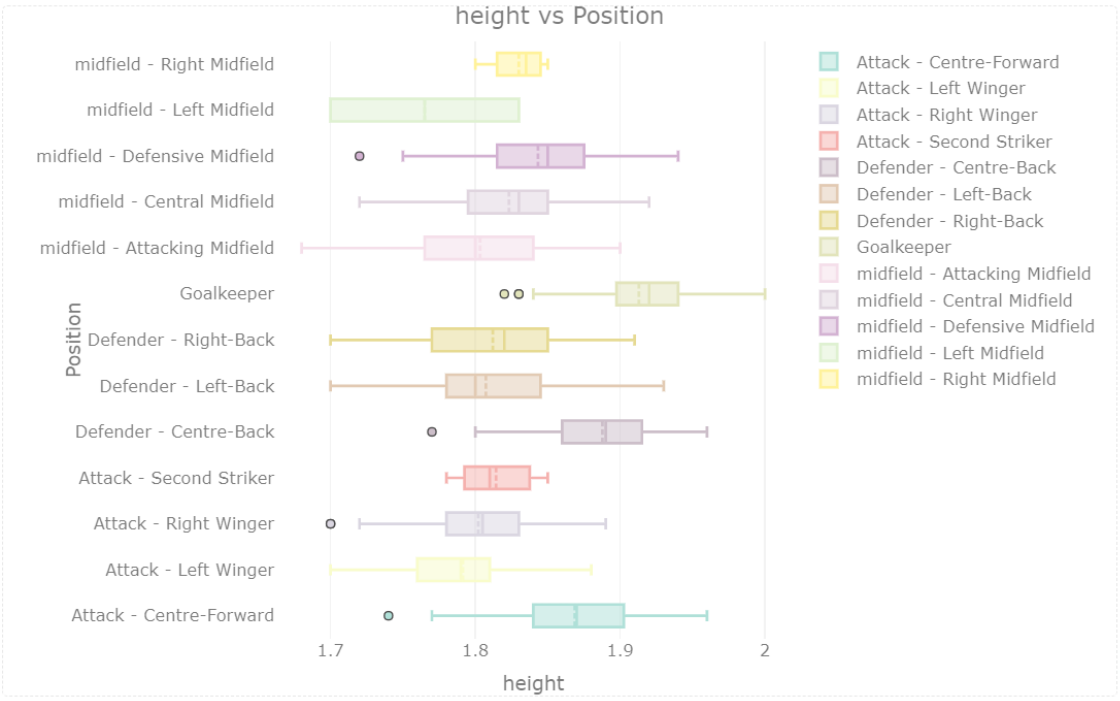
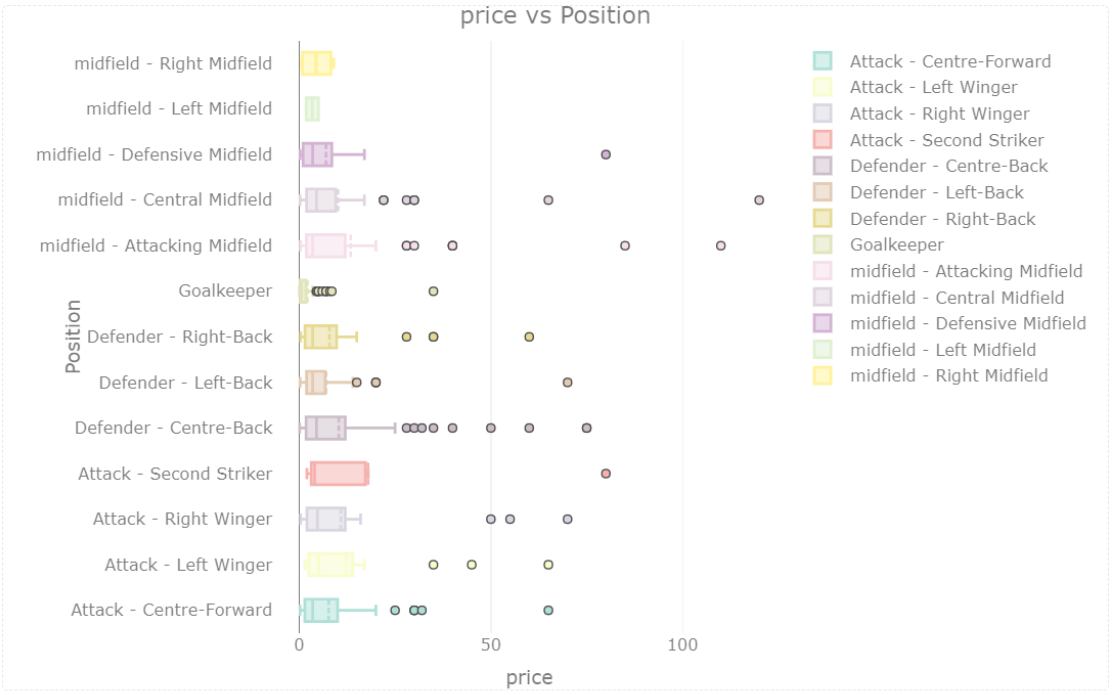
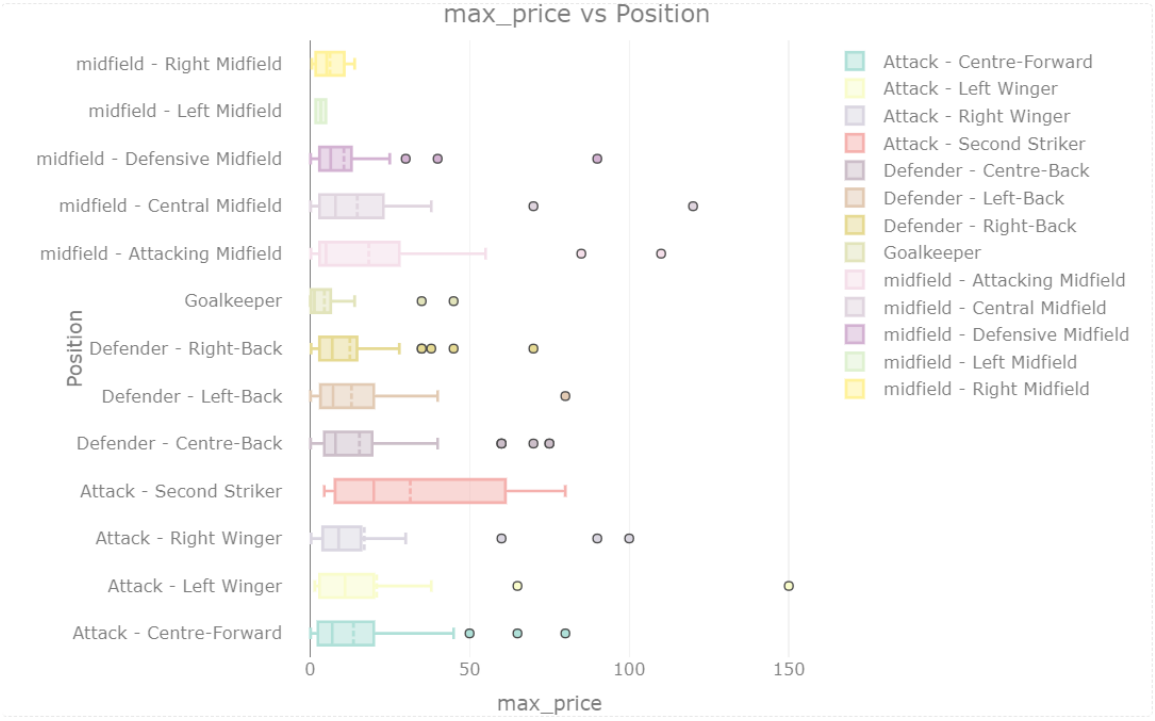
- The Database I got from Kaggle is a soccer player database from Germany league called Bundesliga.
- It has 514 Players along with the following attributed
  1. Name: Player's name.
  2. Age: Player's age.
  3. Height: Player's height in meters.
  4. Nationality: Player's nationality.
  5. Place of Birth: Player's place of birth.
  6. Price: Current market value of the player.
  7. Max Price: Maximum market value of the player.
  8. Position: Playing position of the player.
  9. Shirt Number: Player's shirt number.
  10. Foot: Preferred foot of the player.
  11. Club: Current club of the player.
  12. Contract Expires: Contract expiration date.
  13. Joined Club: Date when the player joined the club.
  14. Outfitter: Player's outfitter brand.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		name	full_name	age	height	nationalit	place_of_birt	price	max_price	position	shirt_nr	foot	club	contract_e	joined_clu	player_ag	outfitter	
2	0	Manuel N	Manuel Pe	37	1.93	Germany	Gelsenkircher	7	45	Goalkeeper	1	right	Bayern Mu	########	########	PRO Profil	adidas	
3	1	Yann Sommer		34	1.83	Switzerlan	Morges	5	13	Goalkeeper	27	right	Bayern Mu	########	########	Relatives	Puma	
4	2	Sven Ulreich		34	1.92	Germany	Schorndorf	0.9	6	Goalkeeper	26	right	Bayern Mu	########	########	BMS Sport	adidas	
5	3	Johannes Schenk		20	1.91	Germany	Schweinfurt	0.3	0.3	Goalkeeper	35		Bayern Mu	########	########	11WINS		
6	4	Matthijs de Ligt		23	1.89	Netherlan	Leiderdorp	75	75	Defender - Centre-Ba	4	right	Bayern Mu	########	########	Rafaela Pi	adidas	
7	5	Dayot Upa	Dayotchar	24	1.86	FranceÃ Â	Ã‰vreux	60	60	Defender - Centre-Ba	2	right	Bayern Mu	########	########	Unique Sp	Nike	
8	6	Lucas Her	Lucas Frar	27	1.84	FranceÃ Â	Marseille	50	70	Defender - Centre-Ba	21	left	Bayern Mu	########	########	Manuel Gi	Nike	
9	7	Alphonso	Alphonso	22	1.85	CanadaÃ /	Bududuram	70	80	Defender - Left-Back	19	left	Bayern Mu	########	########	ATG Sports		
10	8	Daley Blind		33	1.8	Netherlan	Amsterdam	6	25	Defender - Left-Back	23	left	Bayern Mu	########	########	SEG	adidas	
11	9	JoÃ£o Can	JoÃ£o Ped	28	1.82	Portugal	Barreiro	60	70	Defender - Right-Back	22	right	Bayern Mu	########	########	Gestifute	Nike	
12	10	Benjamin Pavard		27	1.86	France	Maubeuge	35	45	Defender - Right-Back	5	right	Bayern Mu	########	########	Carmenta	adidas	
13	11	Noussair N	Ù†ØµÙŠØ	25	1.83	MoroccoÃ	Leiderdorp	28	28	Defender - Right-Back	40	right	Bayern Mu	########	########	Rafaela Pi	adidas	
14	12	Josip Stan	Josip Stan	23	1.87	CroatiaÃ Â	MÃ¼nchen	12	12	Defender - Right-Back	44	both	Bayern Mu	########	########	BALLWERK	adidas	
15	13	Bouna Sarr		31	1.77	SenegalÃ /	Lyon	2.5	9	Defender - Right-Back	20	right	Bayern Mu	########	########	Wasserman		
16	14	Joshua Kir	Joshua Wa	28	1.77	Germany	Rottweil	80	90	midfield - Defensive M	6	right	Bayern Mu	########	########			
17	15	Leon Gore	Leon Chris	28	1.89	Germany	Bochum	65	70	midfield - Central Mid	8	right	Bayern Mu	########	########	Neubauer	13 GmbH	
18	16	Ryan Grav	Ryan Jiro (	20	1.9	Netherlan	Amsterdam	30	35	midfield - Central Mid	38	right	Bayern Mu	########	########	Team Raic	adidas	
19	17	Jamal Musiala		20	1.84	Germany/	Stuttgart	110	110	midfield - Attacking M	42	right	Bayern Mu	########	########	11WINS		
20	18	Paul Wanner		17	1.85	Germany/	Dornbirn	3	3	midfield - Attacking M	14	left	Bayern Mu	########	########	Agent is k	adidas	
21	19	Arijon Ibra	Arijon Ibra	17	1.76	Germany/	NÃ¼rnberg	1	1	midfield - Attacking M	46	right	Bayern Mu	########	########	Agent is known - Player under 18		
22	20	Kingsley C	Kingsley Ju	26	1.81	FranceÃ Â	Paris	65	65	Attack - Left Winger	11	right	Bayern Mu	########	########	CAA Base	Nike	
23	21	Sadio ManÃ©		31	1.74	Senegal	Bambaly	45	150	Attack - Left Winger	17	right	Bayern Mu	########	########	ROOF	New Balance	
24	22	Leroy San	Leroy Aziz	27	1.83	Germany/	Essen	70	100	Attack - Right Winger	10	left	Bayern Mu	########	########	LIAN Sport	Nike	
25	23	Serge Gna	Serge Dav	27	1.76	Germany/	Stuttgart	55	90	Attack - Right Winger	7	right	Bayern Mu	########	########	ROOF	adidas	

b. Data Exploration

- Summary of uncleaned data

```
```{r}
summary(bundesliga)
```
```

```
       X              name           full_name             age            height      nationality       
 Min.   :  0.0   Length:515         Length:515         Min.   :17.00   Min.   :1.680   Length:515        
 1st Qu.:128.5   Class :character   Class :character   1st Qu.:22.00   1st Qu.:1.800   Class :character  
 Median :257.0   Mode  :character   Mode  :character   Median :25.00   Median :1.850   Mode  :character  
 Mean   :257.0                                         Mean   :25.68   Mean   :1.848                     
 3rd Qu.:385.5                                         3rd Qu.:29.00   3rd Qu.:1.890                     
 Max.   :514.0                                         Max.   :39.00   Max.   :2.000                     

 place_of_birth         price           max_price        position           shirt_nr        foot          
 Length:515         Min.   :  0.025   Min.   :  0.10   Length:515         Min.   : 1.0   Length:515        
 Class :character   1st Qu.:  1.200   1st Qu.:  2.50   Class :character   1st Qu.: 9.0   Class :character  
 Mode  :character   Median :  3.500   Median :  7.00   Mode  :character   Median :20.0   Mode  :character  
                    Mean   :  8.483   Mean   : 13.51                      Mean   :19.8                     
                    3rd Qu.:  9.000   3rd Qu.: 16.75                      3rd Qu.:29.0                     
                    Max.   :120.000   Max.   :150.00                      Max.   :49.0                     
                    NA's   :5         NA's   :5                                                            
     club           contract_expires   joined_club        player_agent        outfitter        
 Length:515         Length:515         Length:515         Length:515         Length:515        
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
```

- Distribution of variables



- Box Plots for Numerical Variables by Position






price vs Position
midfield - Right Midfield
midfield - Left Midfield
midfield - Defensive Midfield
midfield - Central Midfield
midfield - Attacking Midfield
Goalkeeper
Defender - Right-Back
Defender - Left-Back
Defender - Centre-Back
Attack - Second Striker
Attack - Right Winger
Attack - Left Winger
Attack - Centre-Forward
Position
0
50
100
price
Attack - Centre-Forward
Attack - Left Winger
Attack - Right Winger
Attack - Second Striker
Defender - Centre-Back
Defender - Left-Back
Defender - Right-Back
Goalkeeper
midfield - Attacking Midfield
midfield - Central Midfield
midfield - Defensive Midfield
midfield - Left Midfield
midfield - Right Midfield


max_price vs Position
midfield - Right Midfield
midfield - Left Midfield
midfield - Defensive Midfield
midfield - Central Midfield
midfield - Attacking Midfield
Goalkeeper
Defender - Right-Back
Defender - Left-Back
Defender - Centre-Back
Attack - Second Striker
Attack - Right Winger
Attack - Left Winger
Attack - Centre-Forward
Position
0
50
100
150
max_price
Attack - Centre-Forward
Attack - Left Winger
Attack - Right Winger
Attack - Second Striker
Defender - Centre-Back
Defender - Left-Back
Defender - Right-Back
Goalkeeper
midfield - Attacking Midfield
midfield - Central Midfield
midfield - Defensive Midfield
midfield - Left Midfield
midfield - Right Midfield

c. Data Cleaning

- Removed unwanted characters and whitespaces in nationality and place_of_birth variables.
- Changed Date Columns to Date Format in contract_expires and joined_club.
- Removed the unnecessary columns full_name, player_agent, place_of_birth, contract_expires and joined_club.
- Removed missing values

```
##Fixing Typos and Cleaning Strings

```{r}
bundesliga$nationality <- gsub("Â·Â·", "", bundesliga$nationality)
bundesliga$place_of_birth <- gsub("Â·Â·", "", bundesliga$place_of_birth)

```

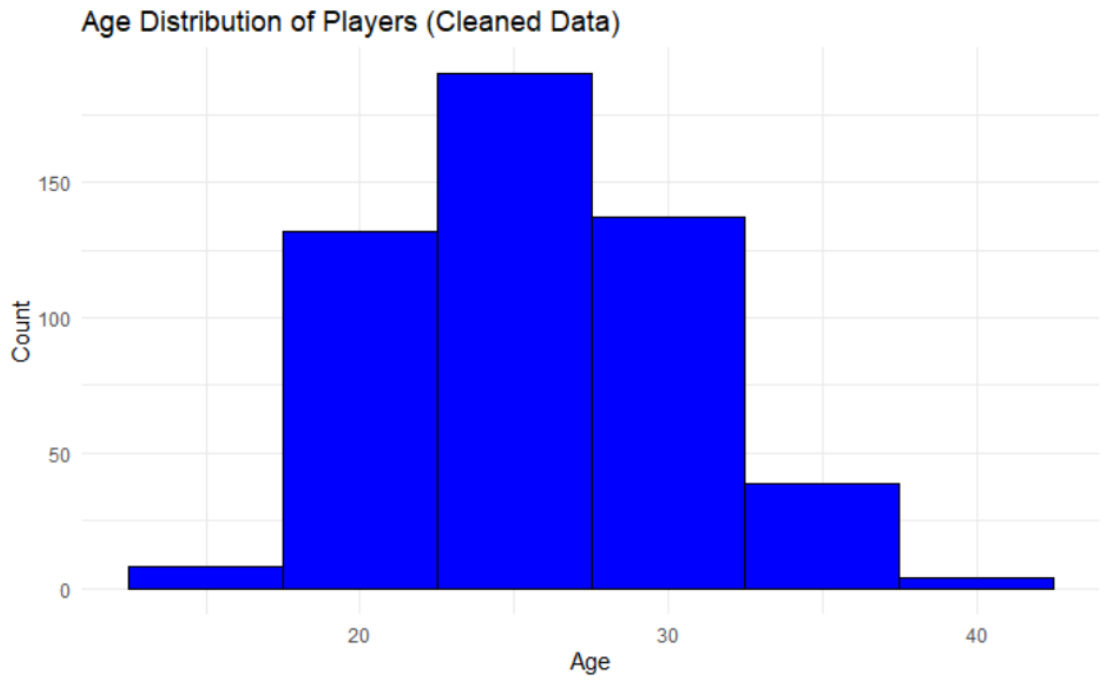
##Changing Date Columns to Date Format

```{r}

bundesliga$contract_expires <- as.Date(bundesliga$contract_expires, format="%d/%m/%Y")
bundesliga$joined_club <- as.Date(bundesliga$joined_club, format="%d/%m/%Y")

```

##Removing Unnecessary Columns

```{r}
bundesliga <- bundesliga %>% select(-full_name, -player_agent,-place_of_birth, -contract_expires, -joined_club)
```


##Removing Missing Values

```{r}

bundesliga <- na.omit(bundesliga)

```
```

- Summary after cleaning

```
```{r}
summary(bundesliga)

```
```

```
       X             name                age            height      nationality            price        
 Min.   :  0.0   Length:510         Min.   :17.00   Min.   :1.680   Length:510         Min.   :  0.025  
 1st Qu.:128.2   Class :character   1st Qu.:22.00   1st Qu.:1.800   Class :character   1st Qu.:  1.200  
 Median :257.5   Mode  :character   Median :25.00   Median :1.850   Mode  :character   Median :  3.500  
 Mean   :257.5                      Mean   :25.76   Mean   :1.847                      Mean   :  8.483  
 3rd Qu.:386.8                      3rd Qu.:29.00   3rd Qu.:1.890                      3rd Qu.:  9.000  
 Max.   :514.0                      Max.   :39.00   Max.   :2.000                      Max.   :120.000  
   max_price        position           shirt_nr        foot               club            outfitter        
 Min.   :  0.10   Length:510         Min.   : 1.0   Length:510         Length:510         Length:510        
 1st Qu.:  2.50   Class :character   1st Qu.: 9.0   Class :character   Class :character   Class :character  
 Median :  7.00   Mode  :character   Median :19.5   Mode  :character   Mode  :character   Mode  :character  
 Mean   : 13.51                      Mean   :19.7                                                           
 3rd Qu.: 16.75                      3rd Qu.:29.0                                                           
 Max.   :150.00                      Max.   :49.0                                                           
```

- Visualization of clean data



d. Data Preprocessing

- Normalized numerical variables
- Created dummy variables for categorical columns
- Binned age into 3 categories young, mid and old.

```
##Normalization

```{r}
bundesliga[numerical_vars] <- scale(bundesliga[numerical_vars])

```

##Creating Dummy Variables for Categorical Columns

```{r}
categorical_cols <- c('position', 'nationality', 'club')

bundesliga <- bundesliga %>%
  mutate(across(all_of(categorical_cols), as.factor)) %>%
  model.matrix(~.-1, data=.) %>%
  as.data.frame()

```

##Binning 'Age' into Categories: 'young', 'mid', 'old'

```{r}
# Binning 'age' into categories: 'young', 'mid', 'old'
bundesliga$age_group <- cut(bundesliga$age, breaks = 3, labels = c("young", "mid", "old"))

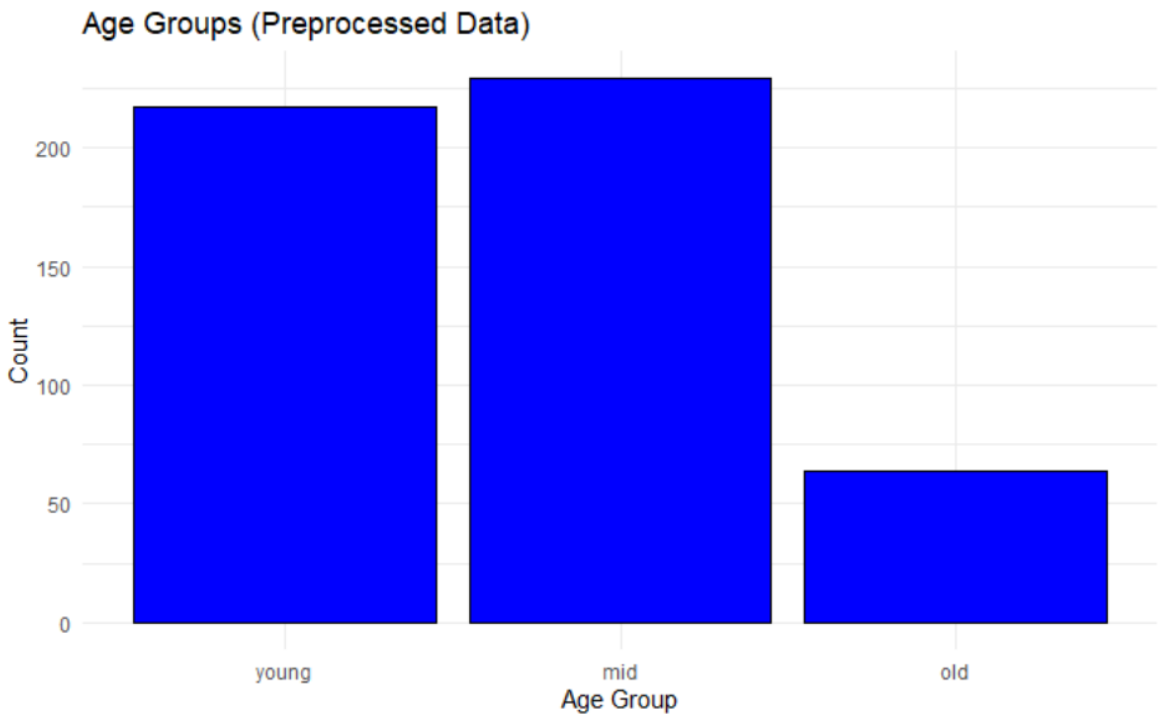
```
```

```{r}
head(bundesliga)
```

Description: df [6 × 712]

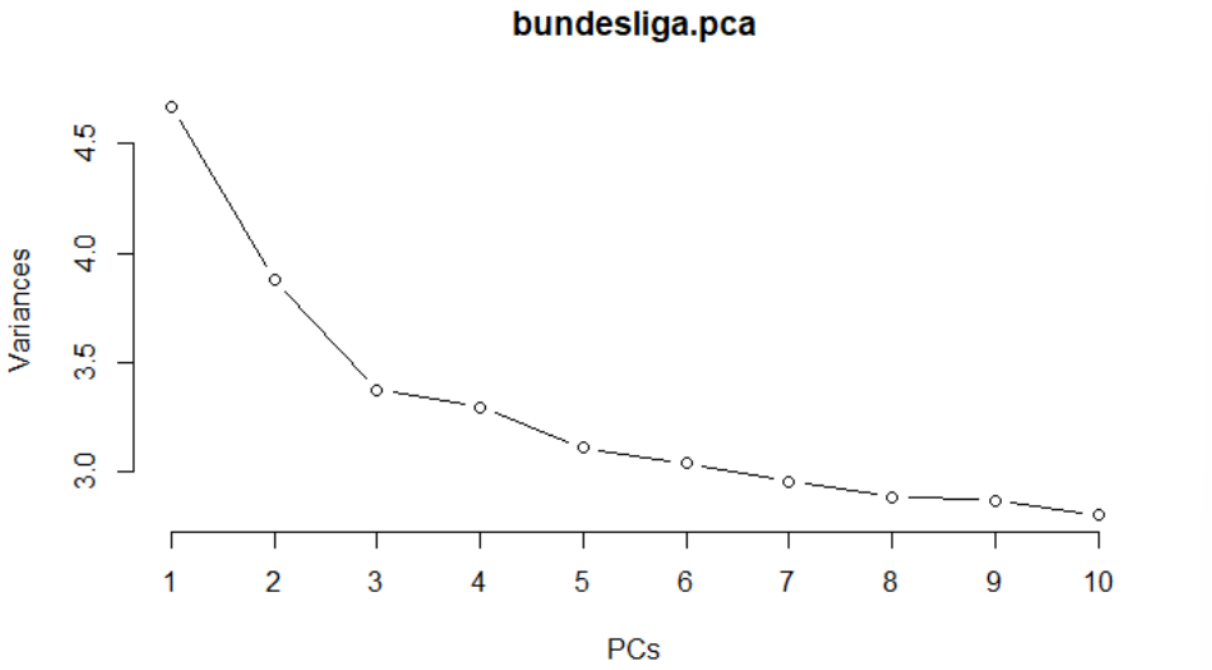
	X <dbl>	nameAarón Martín <dbl>	nameAaron Zehnter <dbl>	nameAbdou Diallo <dbl>	nameAbdoulaye Kamara <dbl>	nameAdam Hlozek <dbl>	nameAgustín Rogel <dbl>
1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0
3	2	0	0	0	0	0	0
4	3	0	0	0	0	0	0
5	4	0	0	0	0	0	0
6	5	0	0	0	0	0	0

6 rows | 1-8 of 712 columns

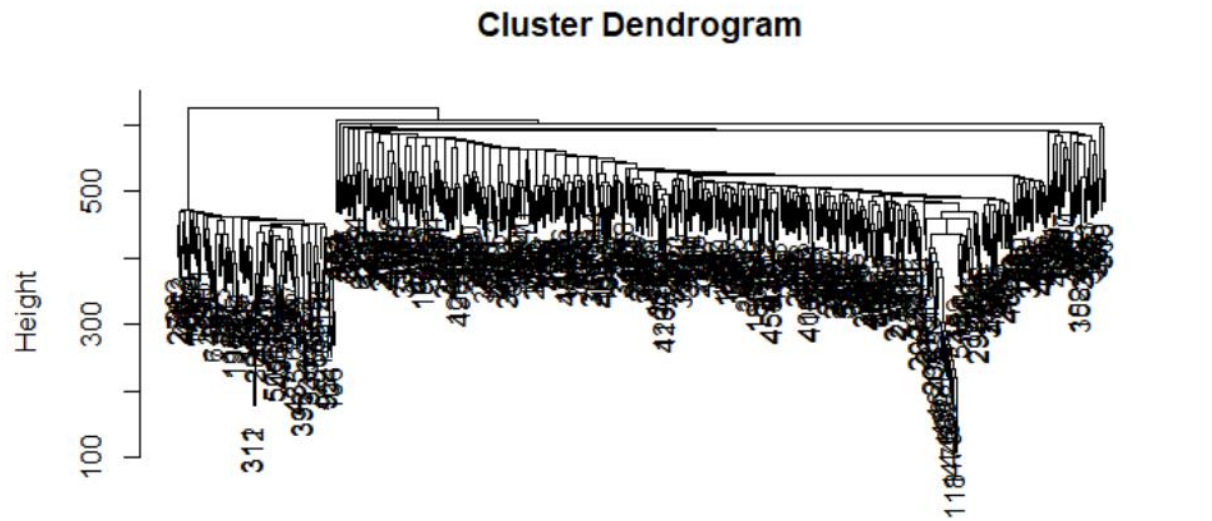


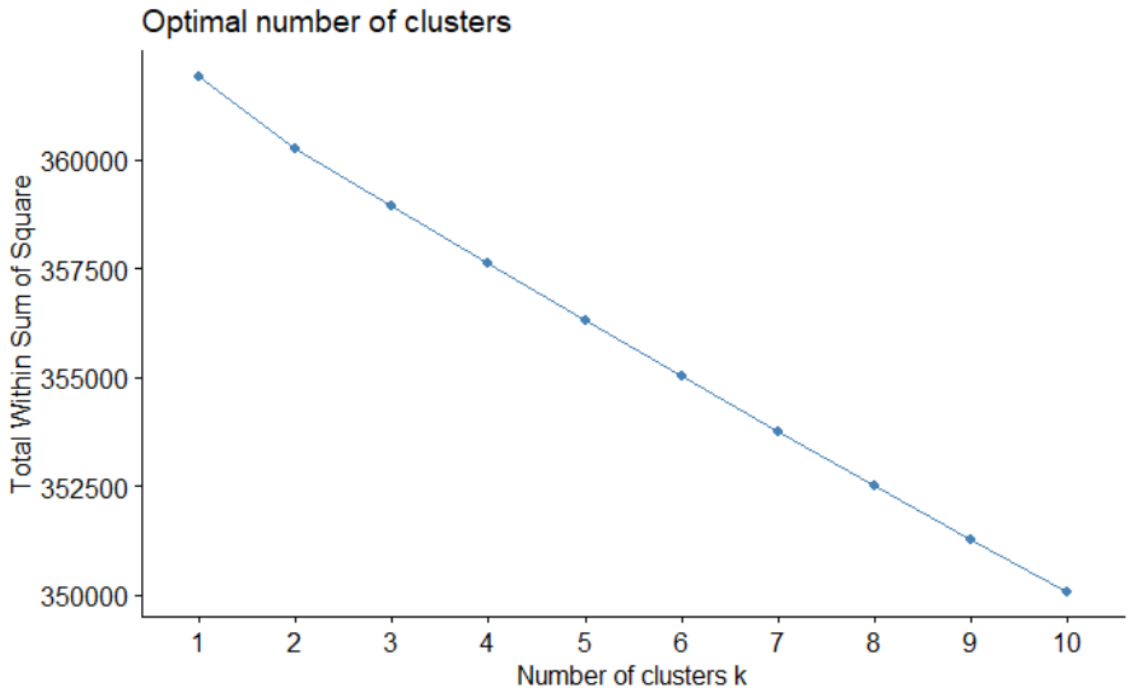
e. Clustering

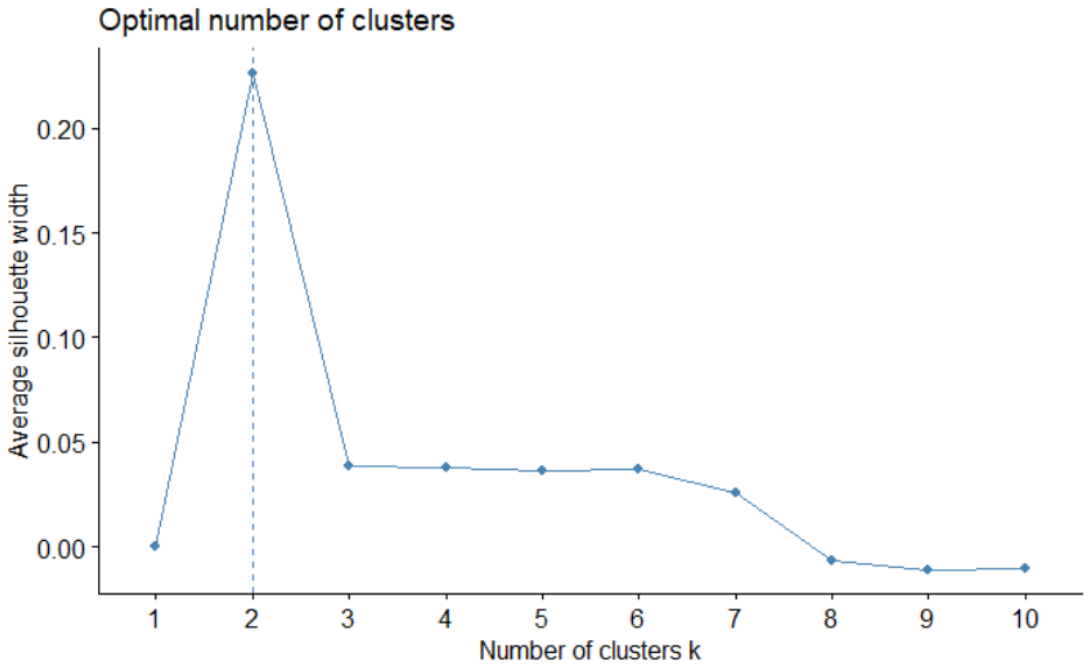
- PCA was used to reduce the dimensionality of the dataset, making it easier to visualize and analyze clusters.

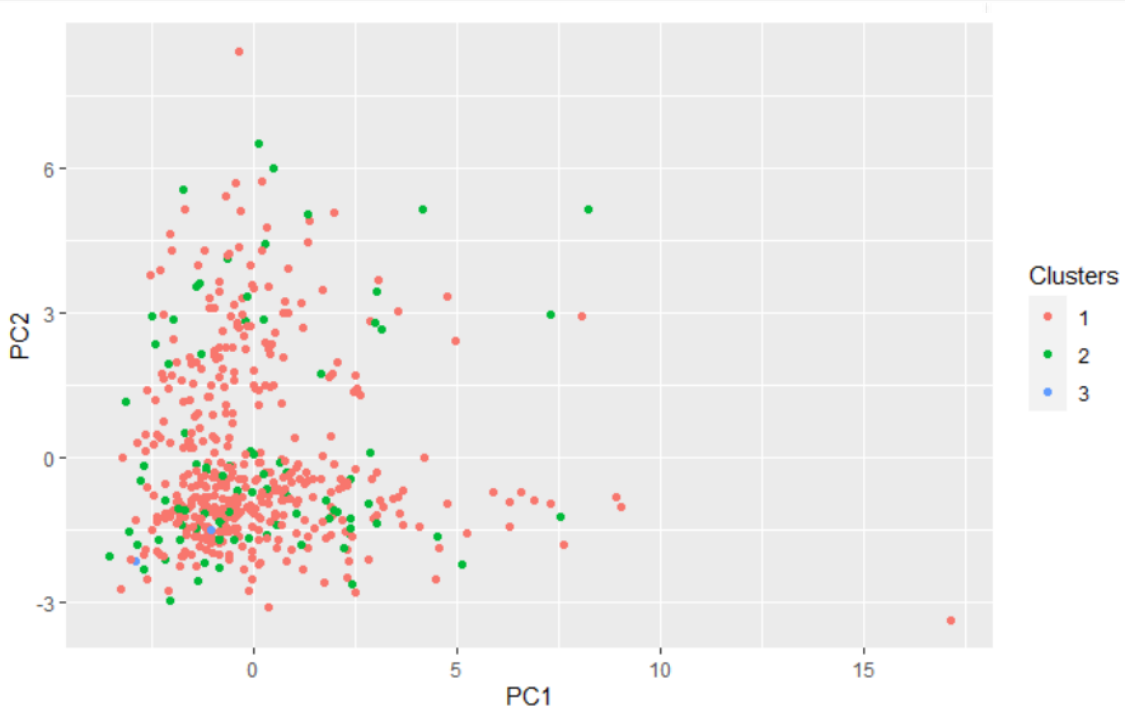


- Hierarchical clustering with 3 clusters was determined to be appropriate based on silhouette scores.

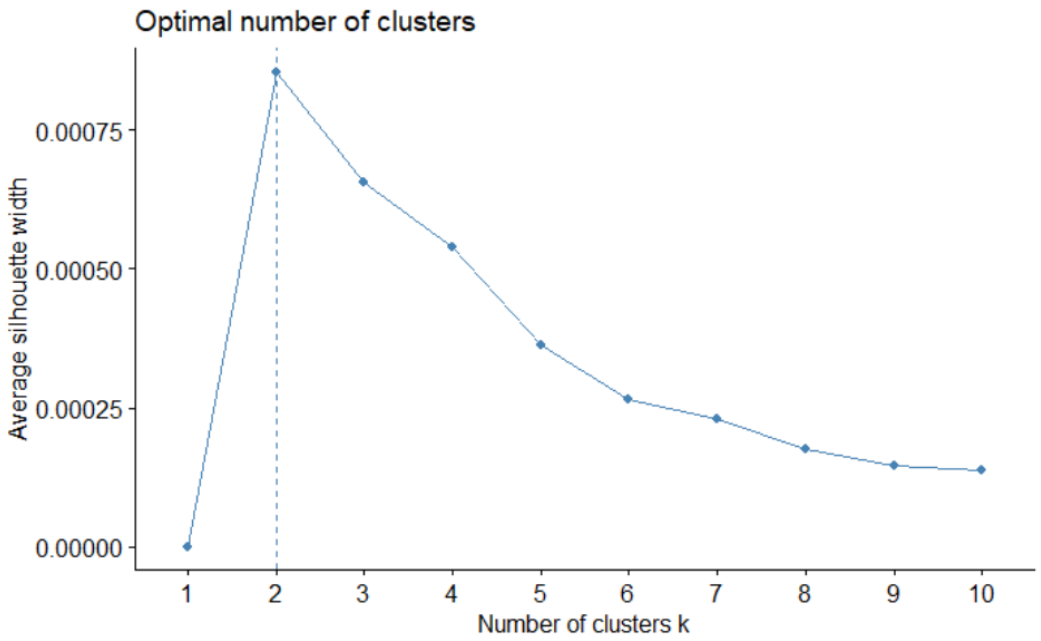



Optimal number of clusters
Total Within Sum of Square
360000
357500
355000
352500
350000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

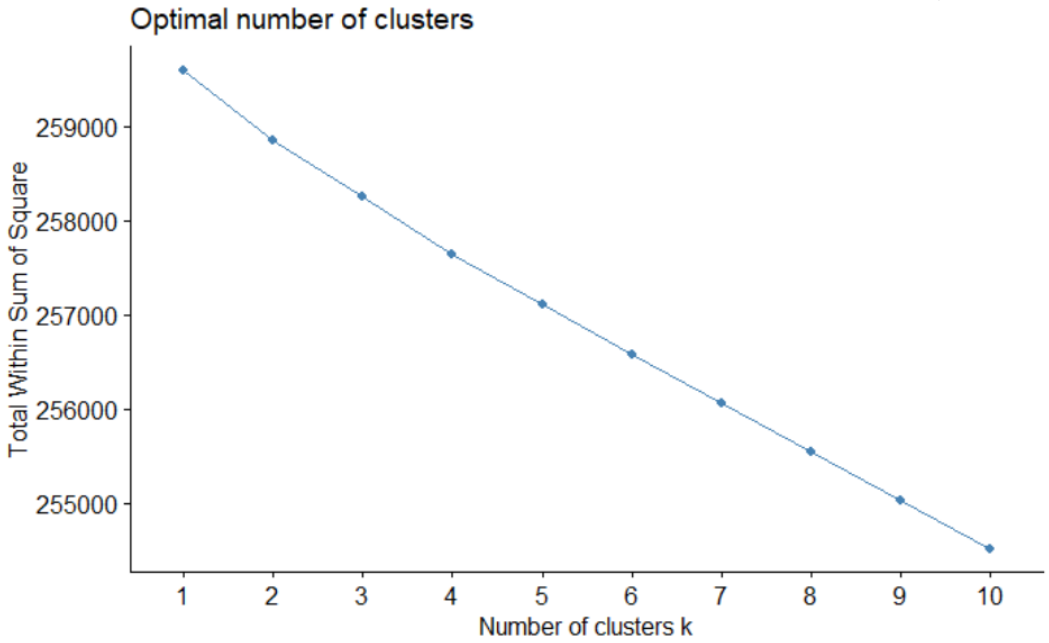

Optimal number of clusters
Average silhouette width
0.20
0.15
0.10
0.05
0.00
1
2
3
4
5
6
7
8
9
10
Number of clusters k

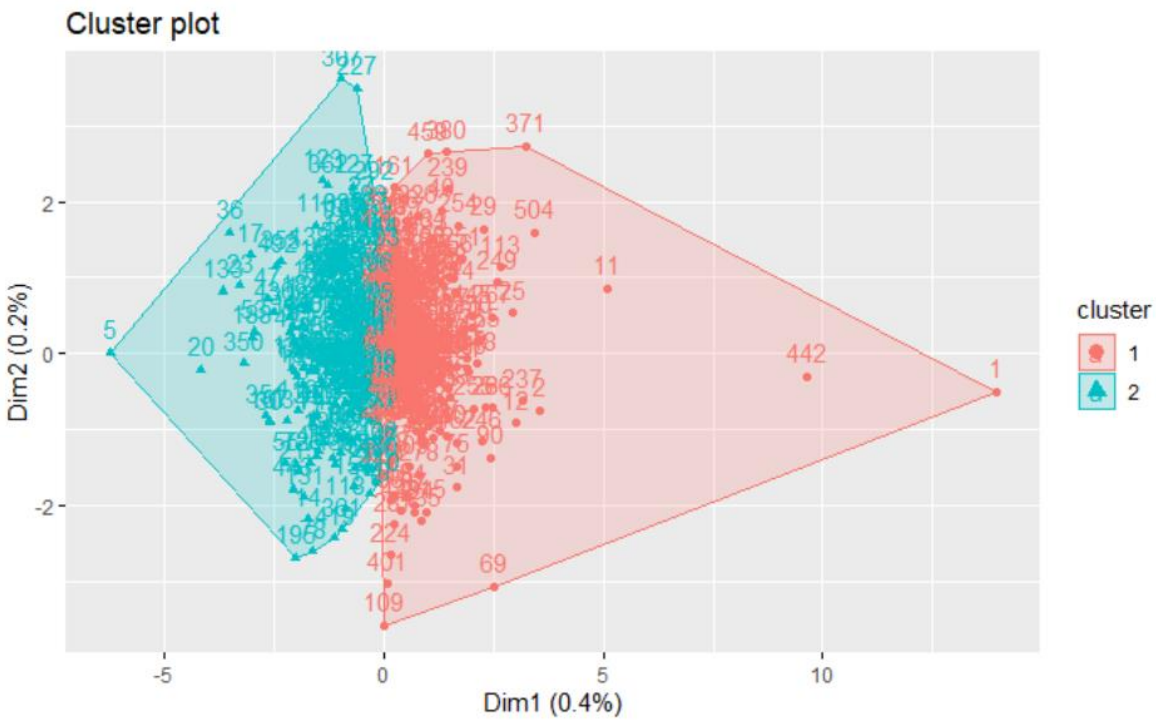


- K-means clustering was also performed, and the optimal number of clusters was found to be 3.


Optimal number of clusters
Average silhouette width
0.00075
0.00050
0.00025
0.00000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

-


Optimal number of clusters
Total Within Sum of Square
259000
258000
257000
256000
255000
1
2
3
4
5
6
7
8
9
10
Number of clusters k



Cluster plot
Dim2 (0.2%)
Dim1 (0.4%)
-5
0
5
10
2
0
-2
cluster
1
2

-

f. Classification

- I used k-NN model and Decision Tree model
- The k-NN model achieved an accuracy of approximately 52%, indicating room for improvement.

```
k-Nearest Neighbors

358 samples
711 predictors
  3 classes: 'young', 'mid', 'old'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 322, 322, 322, 321, 323, 322, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.4632518  0.06926053
   7  0.4800815  0.08423095
   9  0.5075547  0.12794725
  11  0.5161218  0.14377039
  13  0.4962849  0.10413494
  15  0.4969820  0.10331296
  17  0.5219863  0.14801505
  19  0.5161175  0.13499342
  21  0.4965144  0.09920748
  23  0.4967439  0.10049915
  25  0.5080888  0.11999935
  27  0.4941248  0.09504042
  29  0.4769863  0.06443056
  31  0.4743629  0.06014139
  33  0.4661840  0.04594871
  35  0.4797726  0.06971625
  37  0.4936572  0.09467006
  39  0.4545174  0.02474169
  41  0.4602402  0.03367573
  43  0.4632518  0.03873998

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 17.
```

-

- The Decision Tree model performed perfectly with 100% accuracy

```
CART

358 samples
711 predictors
  3 classes: 'young', 'mid', 'old'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 322, 323, 322, 321, 322, 323, ...
Resampling results across tuning parameters:

  cp         Accuracy   Kappa
  0.0000000  1.0000000  1.0000000
  0.2284264  0.9313063  0.8789116
  0.7715736  0.6613857  0.3789116

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.
```

-

g. Evaluation

- Since we had 3 classes rebuild the model with 2 classes "young" and "not young"
- 2x2 Confusion Matrix

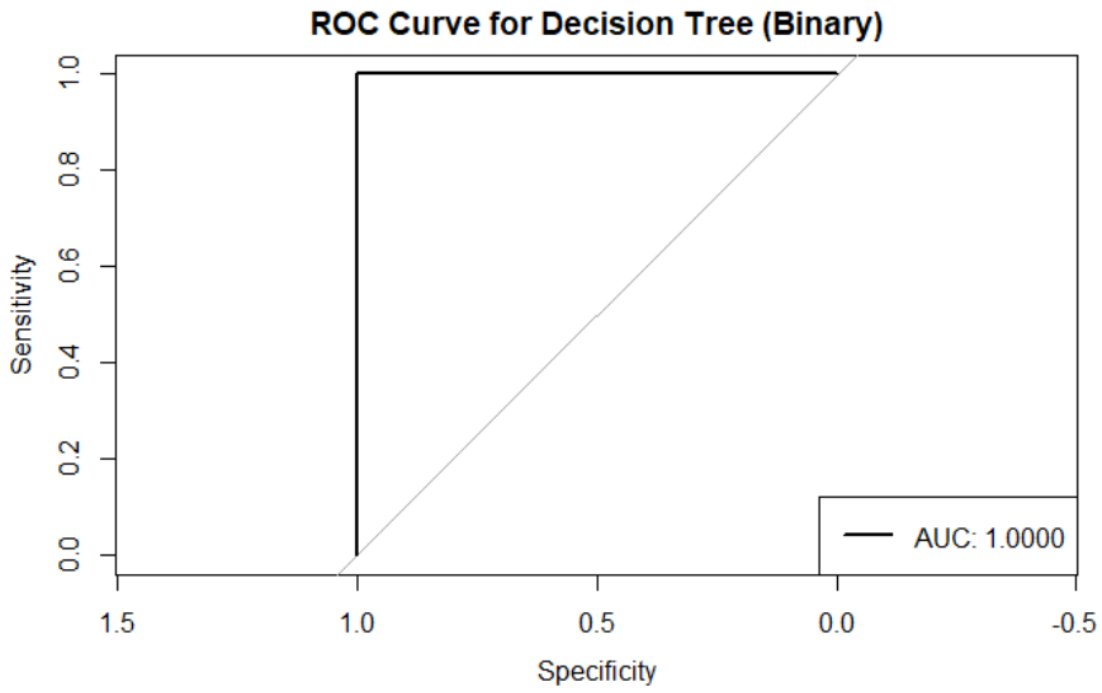
```
Confusion Matrix and Statistics

           Reference
Prediction  young not_young
  young        65         0
  not_young     0        87

               Accuracy : 1
                 95% CI : (0.976, 1)
    No Information Rate : 0.5724
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.4276
         Detection Rate : 0.4276
   Detection Prevalence : 0.4276
      Balanced Accuracy : 1.0000

       'Positive' Class : young
```

- Calculated the Precision and Recall manually Both precision and recall are 1.0000, indicating perfect precision and recall, which means the model does not make any false positives or false negatives.

```
328 ##Precision and Recal
329
330
331 ```{r}
332 |
333 cm_values <- as.numeric(tree_conf_matrix_binary$table)
334 true_negative <- cm_values[1]
335 false_positive <- cm_values[2]
336 false_negative <- cm_values[3]
337 true_positive <- cm_values[4]
338
339 precision <- true_positive / (true_positive + false_positive)
340 recall <- true_positive / (true_positive + false_negative)
341
342 precision
343 recall
344
345 ```
```

```
[1] 1
[1] 1
```

- The ROC curve confirms the model's perfect classification performance, with an AUC of 1.0000.



Description: df [6 × 2]

	young <dbl>	not_young <dbl>
5	1	0
6	1	0
7	0	1
11	0	1
14	0	1
15	0	1

6 rows

h. Report

Interesting analysis:

- The Decision Tree model performed perfectly when age groups were simplified into two categories. This means the model can accurately distinguish between "young" and "not_young" players without making any mistakes.
- Cluster analysis revealed different groupings of players based on their attributes, which helps in identifying player profiles and understanding the variety of player traits in the Bundesliga.
- The age distribution analysis showed a high number of players in their mid-20s, which is typical for professional athletes. Dividing the age into "young," "mid," and "old" categories gave a clearer picture of the age distribution and allowed for a more focused analysis.

i. Reflection

During the FDS course, I learned a lot about how to handle and analyze data. I now understand the steps needed to clean and prepare data, how to use visualizations to explore data, and how to apply different machine learning methods like k-NN, SVM, Decision Trees and Random Forest. Working on real data and using tools to measure how well my models perform, such as confusion matrices and ROC curves, has been very helpful. This course has boosted my skills and given me the confidence to work on more advanced data science projects in the future for my career.