hw3-Sayan-Biswas

Sayan Biswas

11 February 2019

Part A

Problem 1

Fifa 19 dataset has been selected for exploratory data analysis. The dataset includes lastest edition FIFA 2019 players attributes like Age, Nationality, Overall, Potential, Club, Value, Wage, Preferred Foot, International Reputation, Weak Foot, Skill Moves, Work Rate, Position, Jersey Number, Joined, Loaned From, Contract Valid Until, Height, Weight, LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB, Crossing, Finishing, Heading, Accuracy, ShortPassing, Volleys, Dribbling, Curve, FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, StandingTackle, SlidingTackle, GKDiving, GKHandling, GKKicking, GKPositioning, GKReflexes, and Release Clause.

The Fifa 19 dataset can be found at https://www.kaggle.com/karangadiya/fifa19.

```r
fifa19 <- read_csv("fifa.csv")
fifa19
```

```
## # A tibble: 18,207 x 89
##       X1     ID Name    Age Photo Nationality Flag  Overall Potential Club
##    <dbl>  <dbl> <chr> <dbl> <chr> <chr>       <chr>   <dbl>     <dbl> <chr>
##  1     0 158023 L. M~    31 http~ Argentina   http~      94        94 FC B~
##  2     1  20801 Cris~    33 http~ Portugal    http~      94        94 Juve~
##  3     2 190871 Neym~    26 http~ Brazil      http~      92        93 Pari~
##  4     3 193080 De G~    27 http~ Spain       http~      91        93 Manc~
##  5     4 192985 K. D~    27 http~ Belgium     http~      91        92 Manc~
##  6     5 183277 E. H~    27 http~ Belgium     http~      91        91 Chel~
##  7     6 177003 L. M~    32 http~ Croatia     http~      91        91 Real~
##  8     7 176580 L. S~    31 http~ Uruguay     http~      91        91 FC B~
##  9     8 155862 Serg~    32 http~ Spain       http~      91        91 Real~
## 10     9 200389 J. O~    25 http~ Slovenia    http~      90        93 Atlé~
## # ... with 18,197 more rows, and 79 more variables: `Club Logo` <chr>,
## #   Value <chr>, Wage <chr>, Special <dbl>, `Preferred Foot` <chr>,
## #   `International Reputation` <dbl>, `Weak Foot` <dbl>, `Skill
## #   Moves` <dbl>, `Work Rate` <chr>, `Body Type` <chr>, `Real Face` <chr>,
## #   Position <chr>, `Jersey Number` <dbl>, Joined <chr>, `Loaned
## #   From` <chr>, `Contract Valid Until` <chr>, Height <chr>, Weight <chr>,
## #   LS <chr>, ST <chr>, RS <chr>, LW <chr>, LF <chr>, CF <chr>, RF <chr>,
## #   RW <chr>, LAM <chr>, CAM <chr>, RAM <chr>, LM <chr>, LCM <chr>,
## #   CM <chr>, RCM <chr>, RM <chr>, LWB <chr>, LDM <chr>, CDM <chr>,
## #   RDM <chr>, RWB <chr>, LB <chr>, LCB <chr>, CB <chr>, RCB <chr>,
## #   RB <chr>, Crossing <dbl>, Finishing <dbl>, HeadingAccuracy <dbl>,
## #   ShortPassing <dbl>, Volleys <dbl>, Dribbling <dbl>, Curve <dbl>,
## #   FKAccuracy <dbl>, LongPassing <dbl>, BallControl <dbl>,
## #   Acceleration <dbl>, SprintSpeed <dbl>, Agility <dbl>, Reactions <dbl>,
## #   Balance <dbl>, ShotPower <dbl>, Jumping <dbl>, Stamina <dbl>,
## #   Strength <dbl>, LongShots <dbl>, Aggression <dbl>,
## #   Interceptions <dbl>, Positioning <dbl>, Vision <dbl>, Penalties <dbl>,
## #   Composure <dbl>, Marking <dbl>, StandingTackle <dbl>,
## #   SlidingTackle <dbl>, GKDiving <dbl>, GKHandling <dbl>,
## #   GKKicking <dbl>, GKPositioning <dbl>, GKReflexes <dbl>, `Release
## #   Clause` <chr>
```

The Fifa 19 dataset is already available in tidy format.

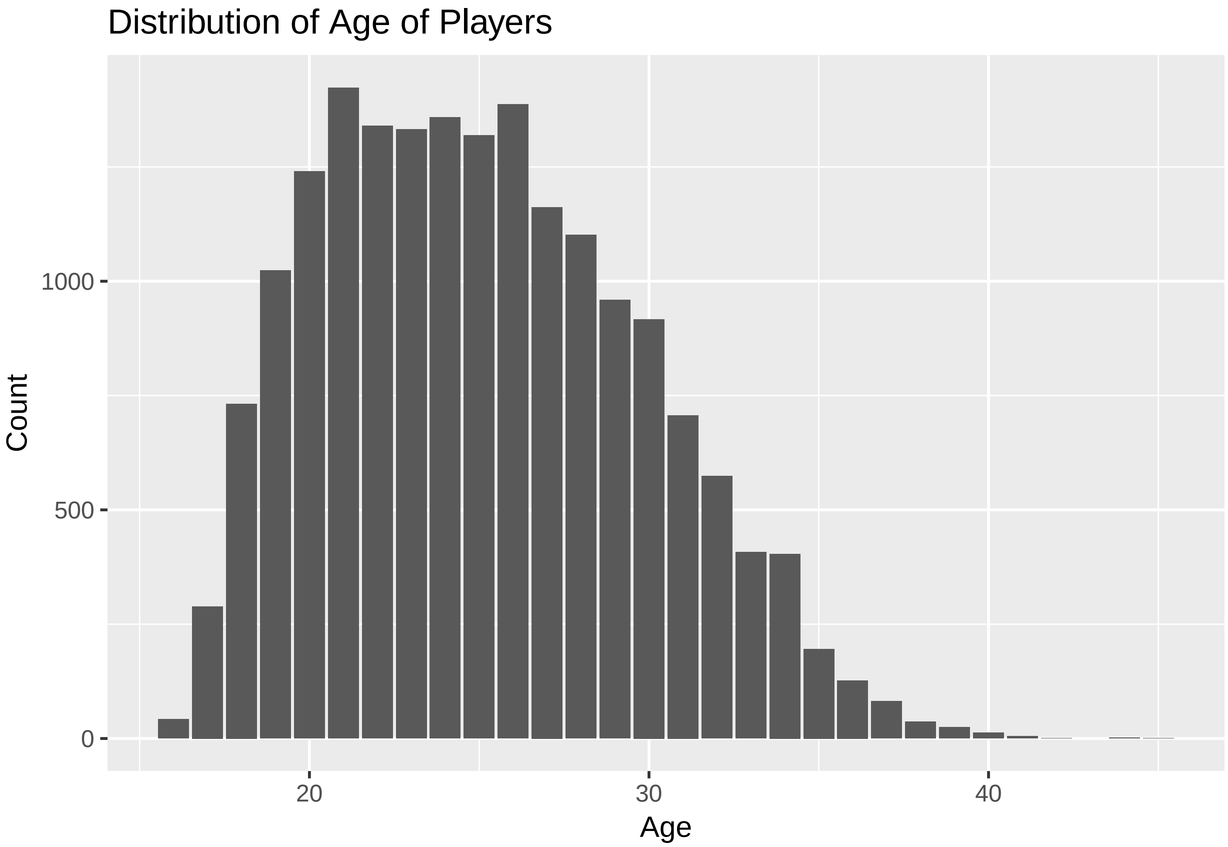
Problem 2

```
#Converting Value, Wage and Release Clause to their actual value in euros.
#Replacing the NA from position with "Unknown"

fifa19 <- fifa19%>%
  mutate(convertval=ifelse(str_detect(Value,"K"),1000,
                           ifelse(str_detect(Value,"M"),1000000,1)))%>%
  mutate(Value_inEuros = as.numeric(str_extract(Value, "[0-9]+")) * convertval) %>%
  mutate(Position = ifelse(is.na(Position), "Unknown", Position))%>%
  mutate(convertwage = ifelse(str_detect(Wage, "K"), 1000,
                              ifelse(str_detect(Wage, "M"), 1000000, 1))) %>%
  mutate(Wage_inEuros = as.numeric(str_extract(Wage, "[0-9]+")) * convertwage)%>%
  mutate(relclause = ifelse(str_detect(`Release Clause`,"K"),1000,
                            ifelse(str_detect(`Release Clause`, "M"), 1000000, 1))) %>%
  mutate(relclause_inEuros =
           as.numeric(str_extract(`Release Clause`, "[0-9]+.[0-9]+")) * relclause)
```

```
#creating groups of position
#create age groups
positions <- unique(fifa19$Position)
gk <- "GK"
defs <- positions[str_detect(positions, "B$")]
mids <- positions[str_detect(positions, "M$")]
fwds <- positions[!(positions%in%c(gk,defs,mids,"Unknown"))]

fifa19 <- fifa19 %>%
  mutate(PositionGroup =
           ifelse(Position %in% gk, "GoalKeepers",
                  ifelse(Position %in% defs, "Defenders",
                         ifelse(Position %in% mids, "Midfielders",
                                ifelse(Position %in% fwds, "Forwards", "Unknown")))))%>%
  mutate(AgeGroup=cut(Age,breaks = c(-Inf,20,25,30,35,Inf),
                    labels =c("20 and under", "21 to 25",
                              "26 to 30","31 to 35","Over 35")))
```

1. Distribution of the age of players

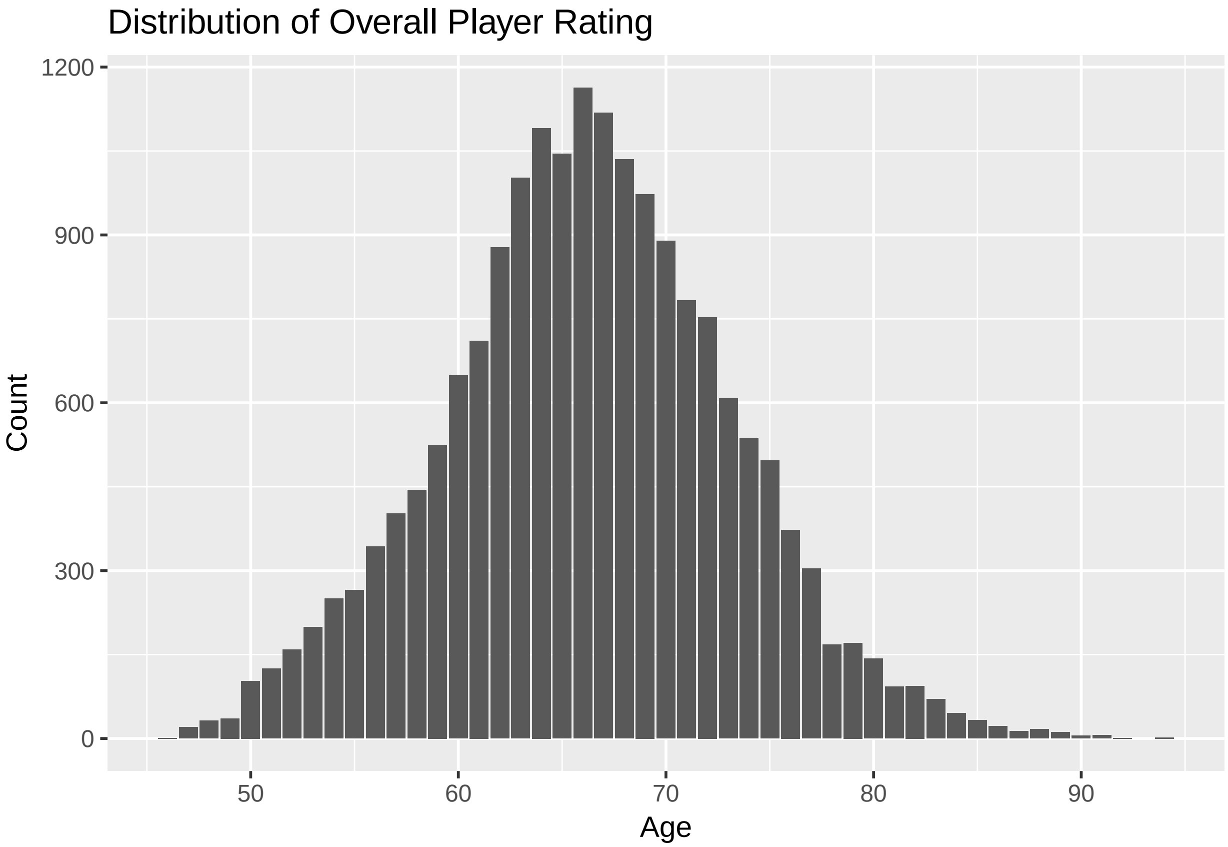
```
fifa19%>%
  ggplot()+
  geom_bar(aes(x=Age))+
  labs(title = "Distribution of Age of Players",
       x="Age",
       y="Count")
```



The plot indicates that the count of players with age 21 years is the highest. The count increases till the age of 26 years post that the count starts to decrease which would be true for any physical sport.

2. Distribution of Overall Player Rating

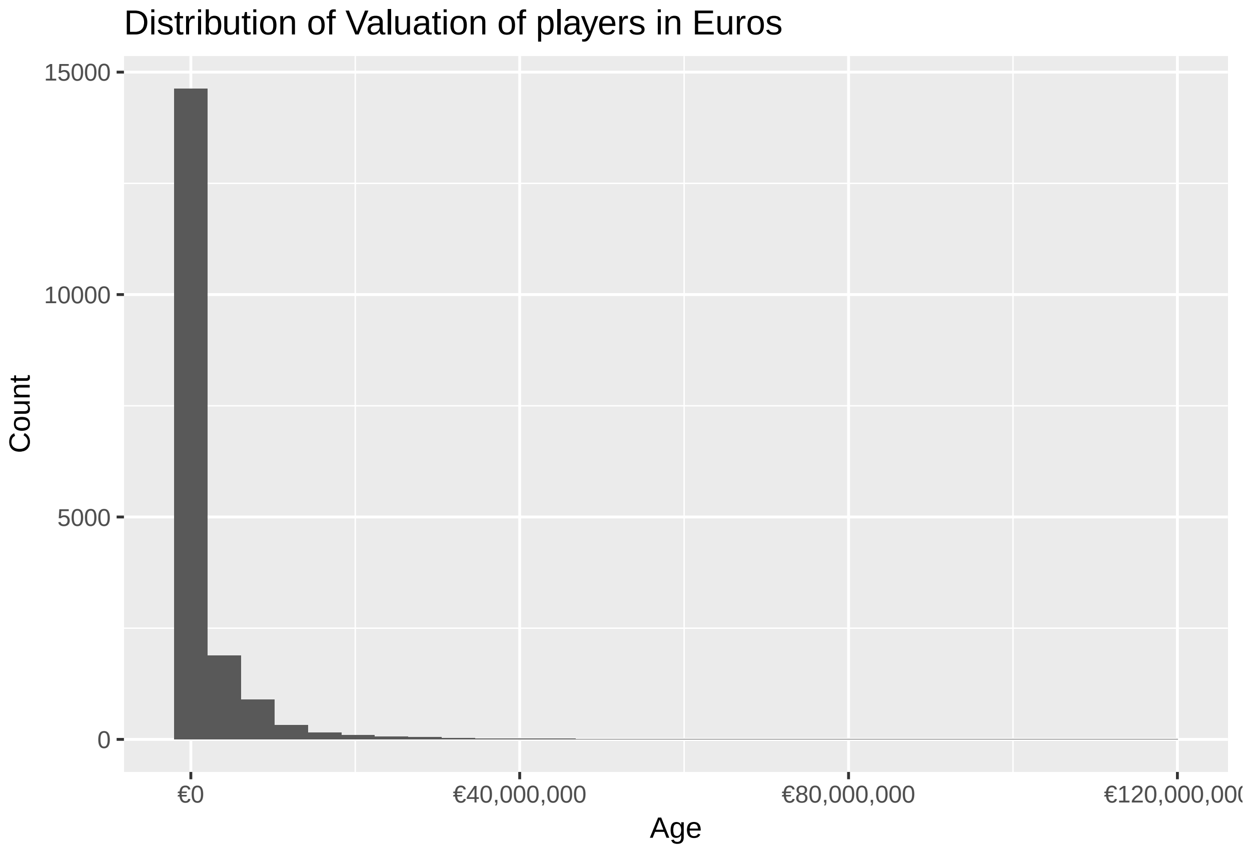
```
fifa19%>%
  ggplot()+
  geom_bar(aes(x=Overall))+
  labs(title = "Distribution of Overall Player Rating",
       x="Age",
       y="Count")
```



The Overall player rating is normally distributed with the mean rating being at 66 points.

3. Distribution of Valuation of players in euros

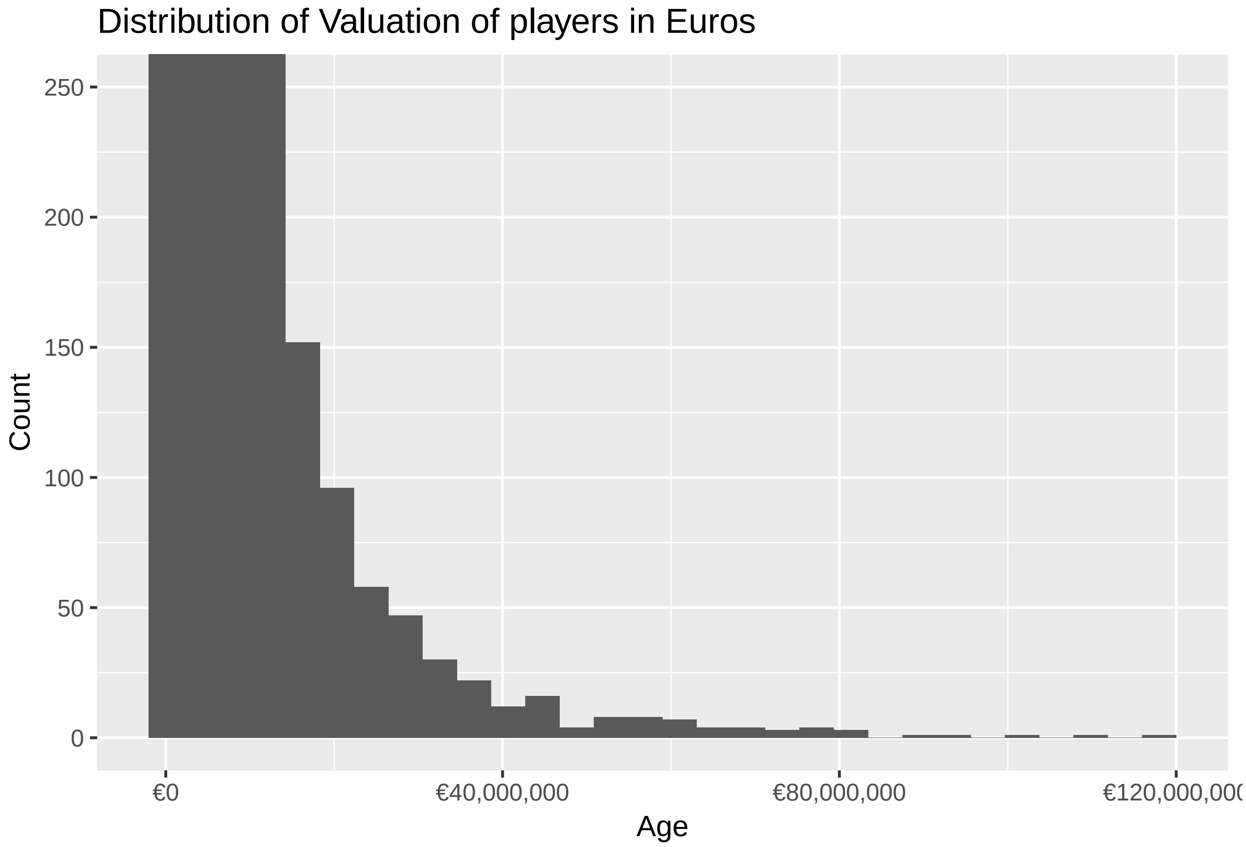
```
fifa19%>%
  ggplot(aes(x=Value_inEuros))+
  geom_histogram()+
  scale_x_continuous(labels=comma_format(prefix = "€"))+
  labs(title = "Distribution of Valuation of players in Euros",
       x="Age",
       y="Count")
```



The distribution is heavily skewed and the evidence of outliers is the unusually wide limits on the x axis. To see the outliers, we need to zoom to small values of the y-axis.

Zooming in with ylim

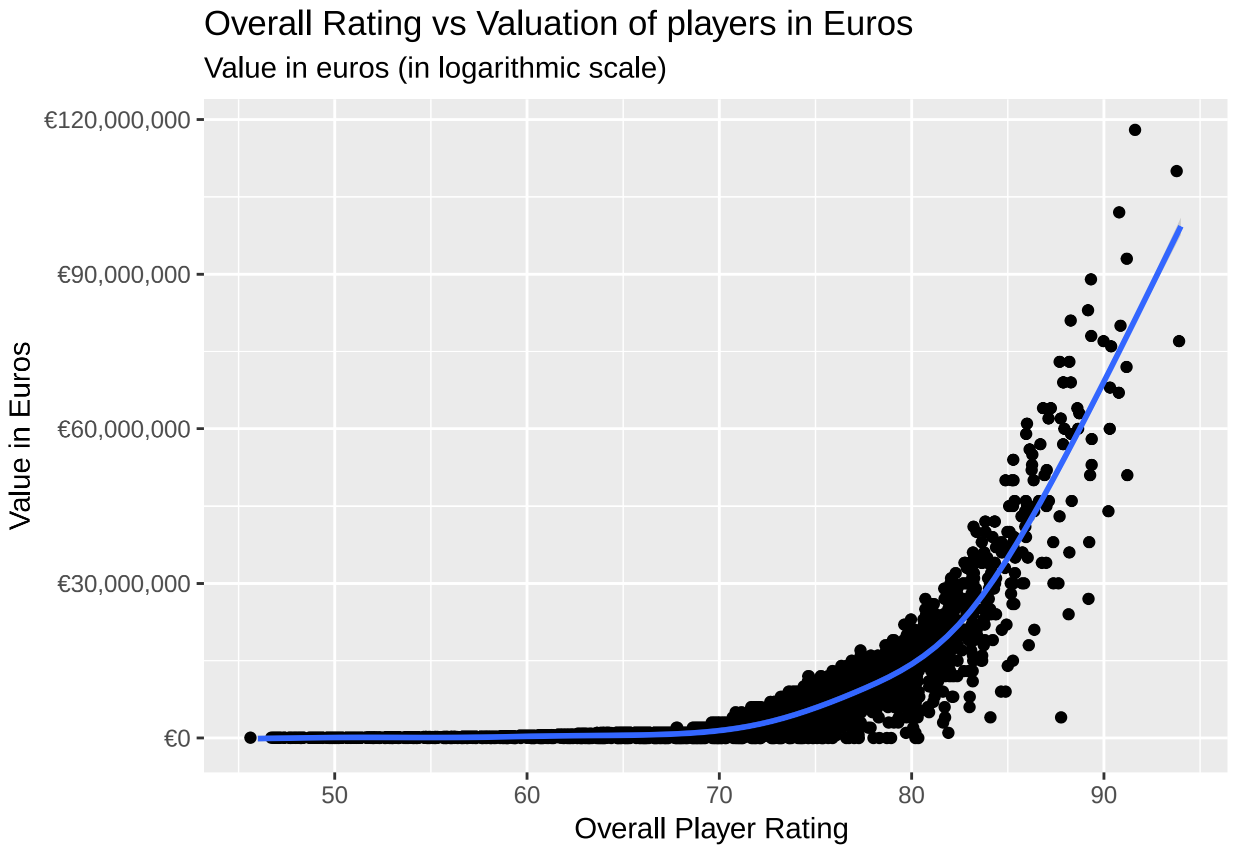
```
fifa19%>%
  ggplot(aes(x=Value_inEuros))+
  geom_histogram()+
  scale_x_continuous(labels=comma_format(prefix = "€"))+
  coord_cartesian(ylim = c(0,250))+
  labs(title = "Distribution of Valuation of players in Euros",
       x="Age",
       y="Count")
```



There are players are like Neymar,Messi, Ronaldo, De Bruyne who have high player valuations and it makes this distribution heavily skewed on the right.

4. Overall Rating vs Valuation of players in euros

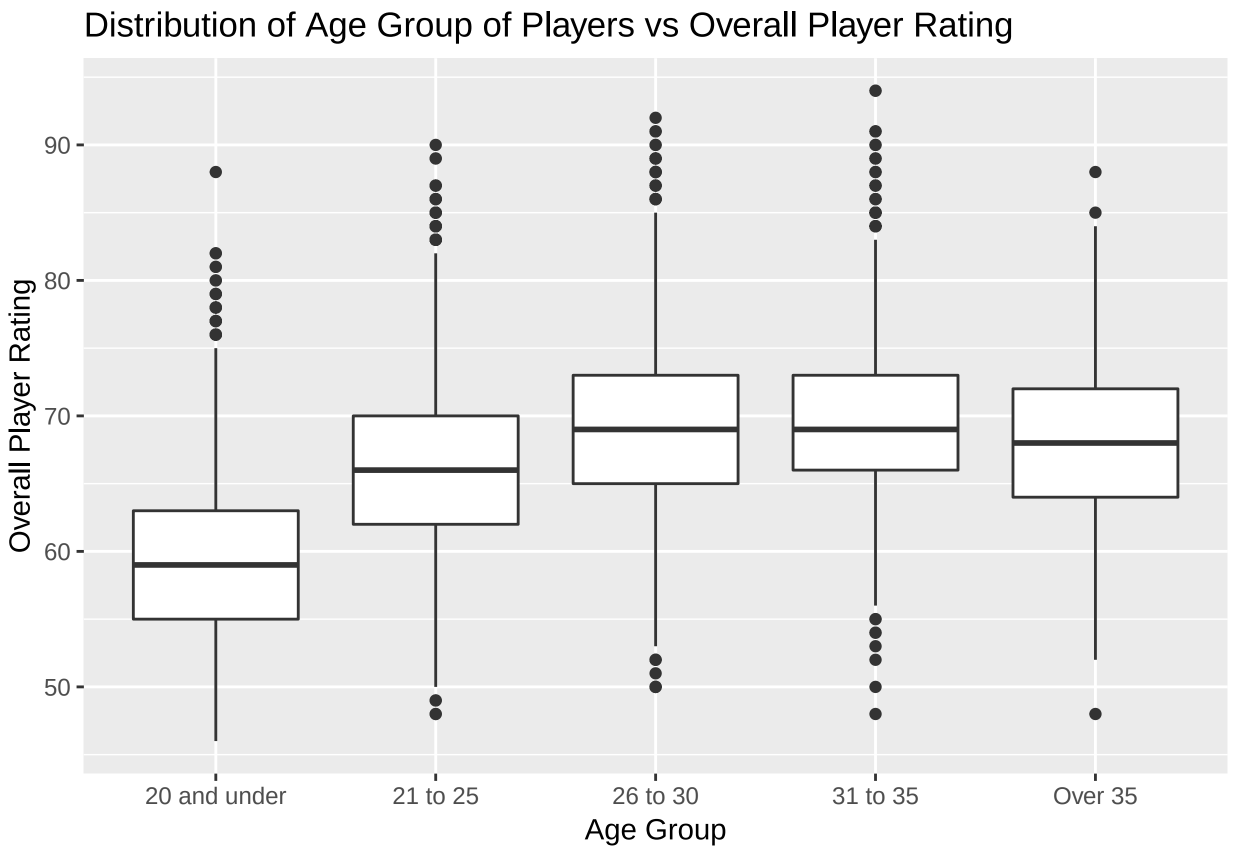
```
fifa19%>%
  ggplot(aes(x=Overall,y=Value_inEuros))+
  geom_jitter()+
  scale_y_continuous(labels=comma_format(prefix = "€"))+
  geom_smooth()+
  labs(title = "Overall Rating vs Valuation of players in Euros",
       subtitle = "Value in euros (in logarithmic scale)",
       x="Overall Player Rating ",
       y="Value in Euros"
  )
```



As expected, the value in euros increases as the overall rating increases beyond 70. The players with overall rating greater than 70 will have earn more than players with rating less than 70.

5. Distribution of Overall Player Rating with each Age Group

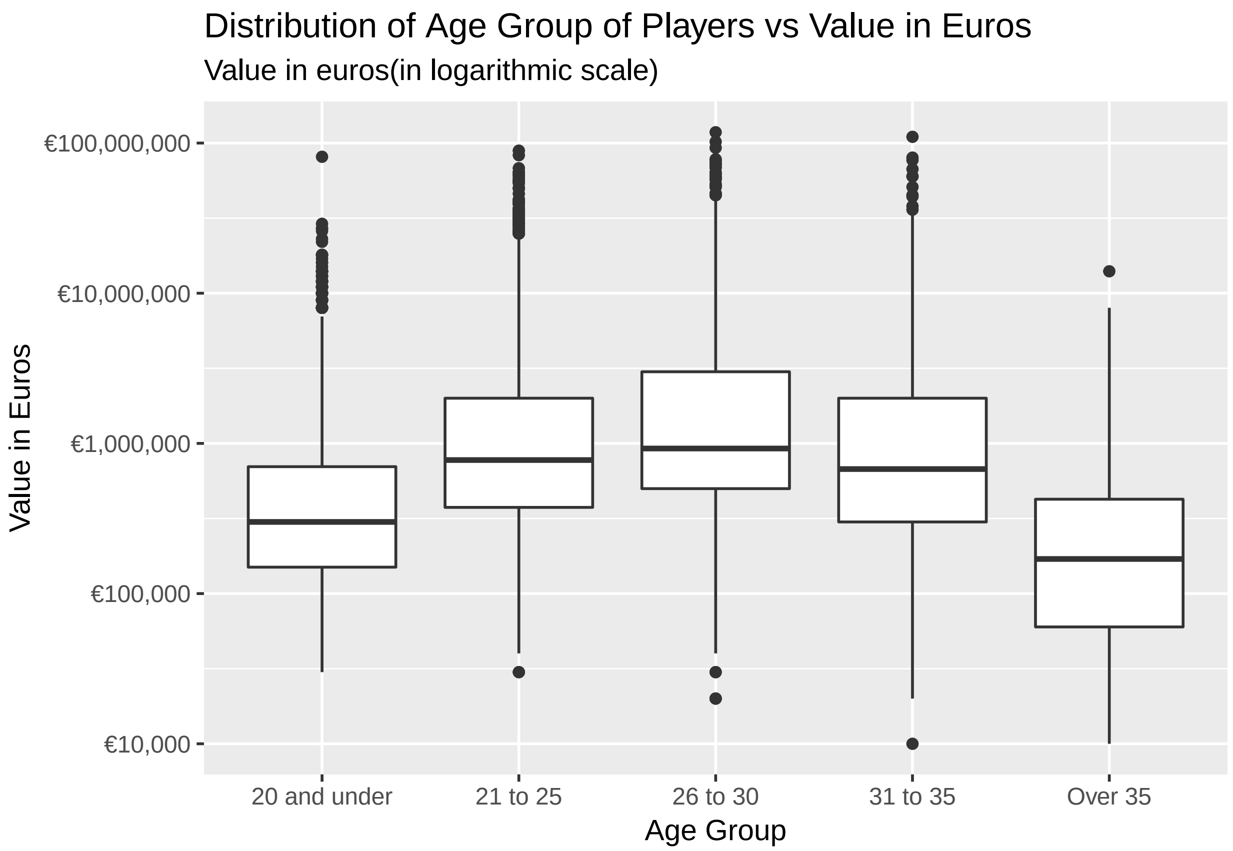
```
fifa19%>%
  ggplot(aes(x=AgeGroup,y=Overall))+
  geom_boxplot()+
  labs(title = "Distribution of Age Group of Players vs Overall Player Rating",
       x="Age Group",
       y="Overall Player Rating")
```



The median Overall Player rating is higher for the Age Groups of "25 to 30" and "31 to 35" and lowest for "20 and under".

6. Distribution of Valuations with each Age Group

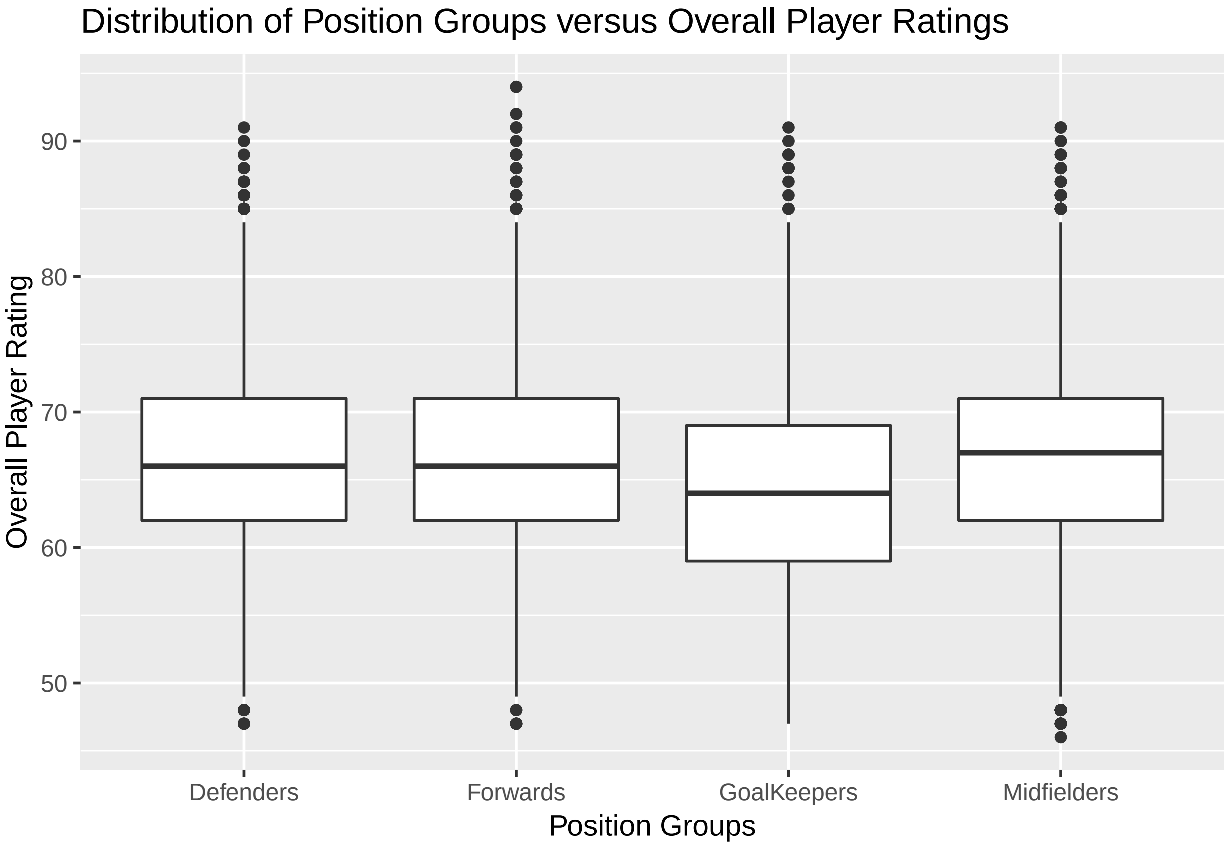
```
fifa19%>%
  ggplot(aes(x=AgeGroup,y=Value_inEuros))+
  geom_boxplot()+
  scale_y_log10(labels=dollar_format(prefix = "€"))+
  labs(title = "Distribution of Age Group of Players vs Value in Euros",
       subtitle = "Value in euros(in logarithmic scale)",
       x="Age Group",
       y="Value in Euros")
```



The valuation of the players increases as the age group increases till the age is within 30 and the valuation decreases as the age increases beyond 30 which would be true for any physical sport. The players within the Age group of 25 to 30 have the highest valuations as this is the age when most of the players are at the peak of their career.

7. Distribution of Overall Rating with each Position Groups

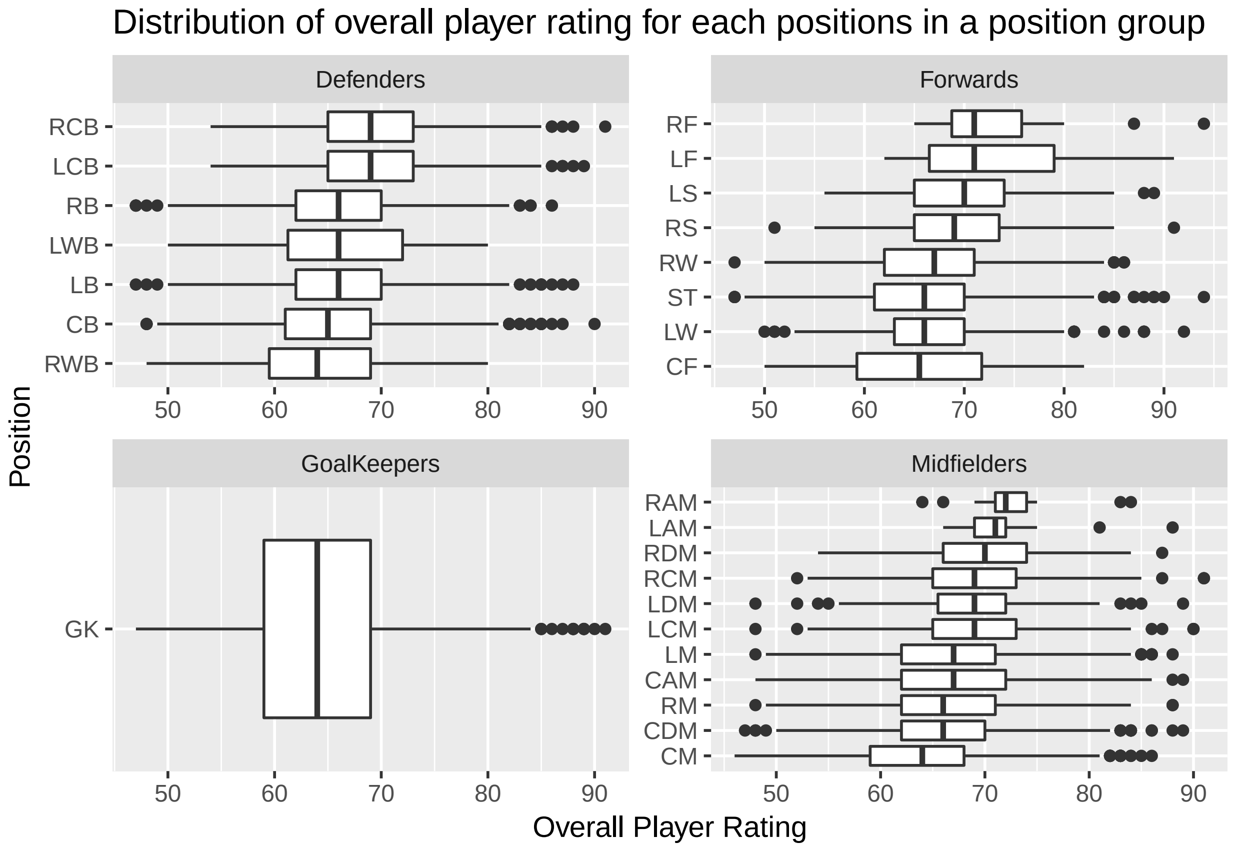
```
fifa19 %>%
  filter(PositionGroup !="Unknown")%>%
  ggplot(aes(x=PositionGroup,y=Overall))+
  geom_boxplot()+
  labs(title = "Distribution of Position Groups versus Overall Player Ratings",
       x="Position Groups",
       y="Overall Player Rating")
```



The median rating of midfielders are higher comapred to Forwards and the median rating of goalkeepers are a bit lower compared to all position groups.

8. Distribution of Overall rating for each positions within each position groups

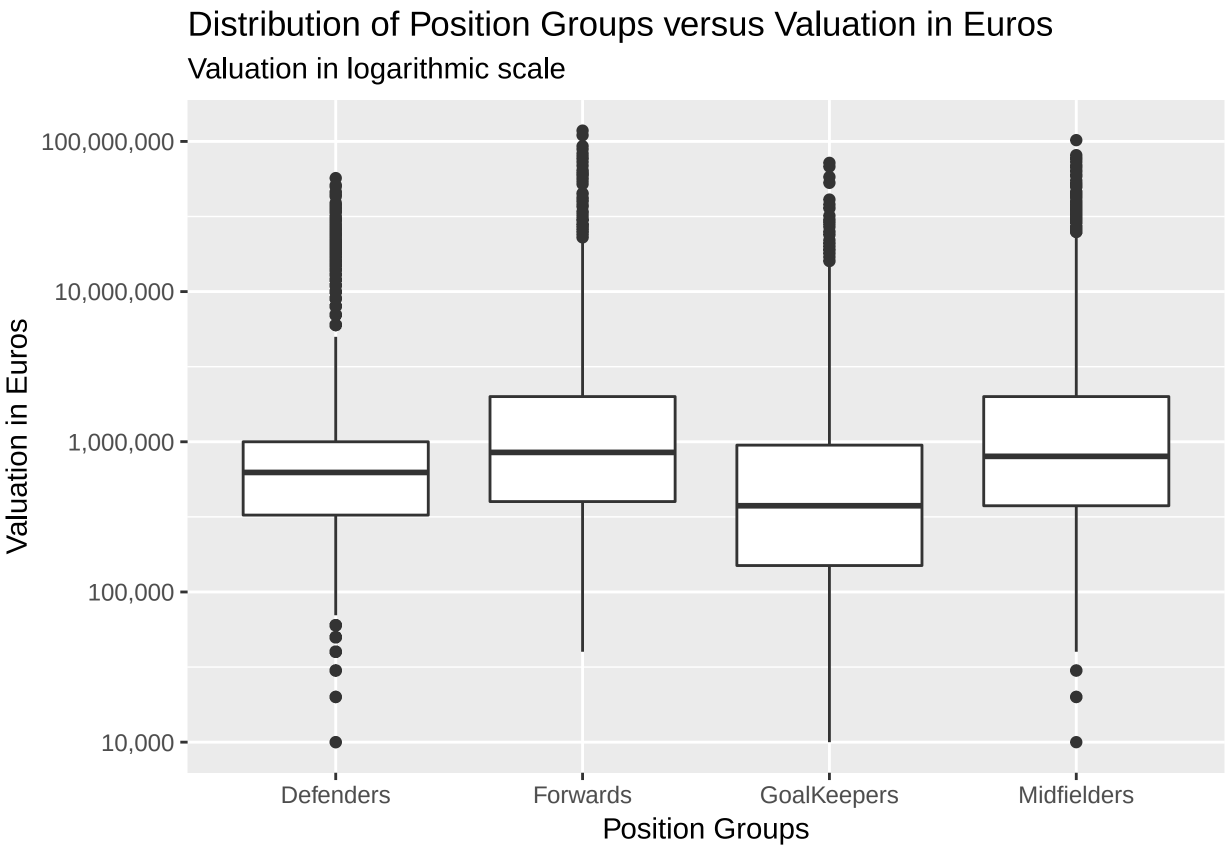
```
fifa19%>%
  filter(Position!="Unknown")%>%
  ggplot(aes(x=reorder(Position,Overall,FUN=median),y=Overall))+
  geom_boxplot()+
  coord_flip()+
  facet_wrap(~PositionGroup,scales = "free")+
  labs(title =
         "Distribution of overall player rating for each positions in a position group",
       x="Position",
       y="Overall Player Rating")
```



1. Among the defenders, Right Centre back and Left Centre back has a higher overall rating.
2. Among the forwards, the Right forward and Left forward has a higher overall rating.
3. Among the midfielders, players playing as right attacking midfield and left attacking midfield has a higher overall rating. Among all the positions Right attacking midfield has the highest median overall rating.

9. Distribution of Valuation of players within each position group

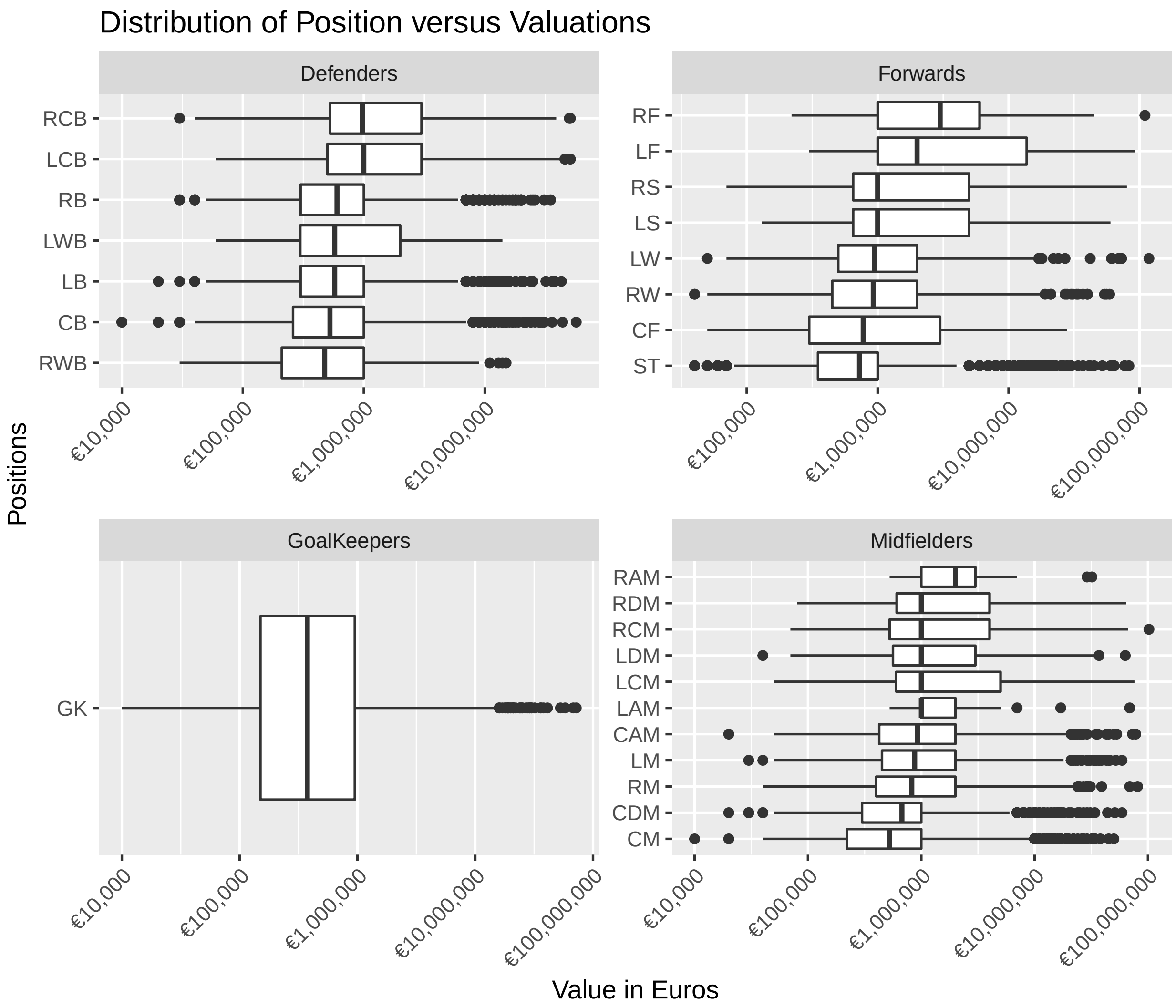
```
fifa19 %>%
  filter(PositionGroup !="Unknown")%>%
  ggplot(aes(x=PositionGroup,y=Value_inEuros))+
  geom_boxplot()+
  scale_y_log10(labels=comma)+
  labs(title = "Distribution of Position Groups versus Valuation in Euros",
       subtitle = "Valuation in logarithmic scale",
       x="Position Groups",
       y="Valuation in Euros")
```



Overall the Forwards and the midfielders have a higher valuation as compared to defenders and golakeepers.

10. Distribution of valuation for each positions within each position group

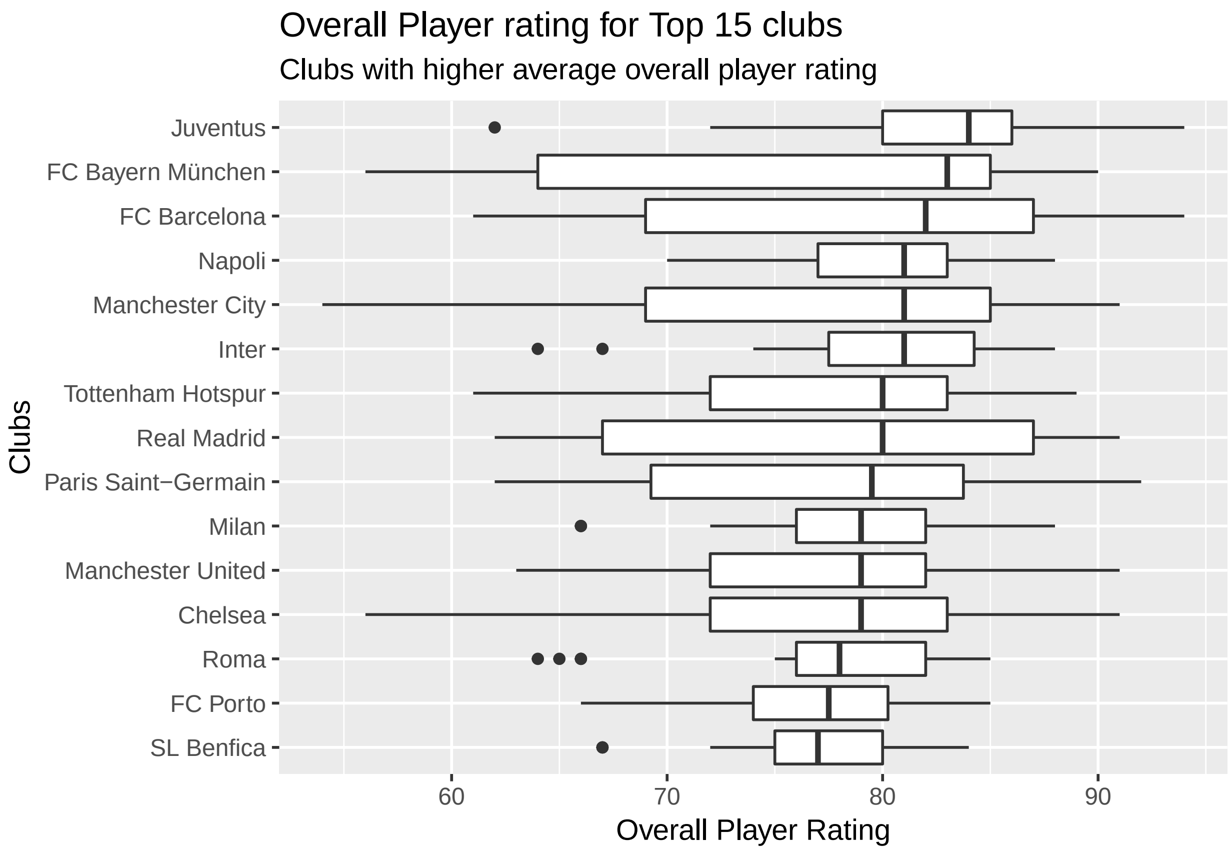
```
fifa19%>%
filter(PositionGroup !="Unknown")%>%
  ggplot(aes(x=reorder(Position,Value_inEuros,FUN=median),y=Value_inEuros))+
  geom_boxplot()+
  scale_y_log10(labels = dollar_format(prefix = "€"))+
  coord_flip()+
  facet_wrap(~PositionGroup,scales = "free")+
  labs(title = "Distribution of Position versus Valuations",
       x="Positions",
       y="Value in Euros")+
  theme(axis.text.x = element_text(angle=45, hjust = 1))
```



1. The Right Centre Back and the Left Centre Back has a higher median valuation among the defenders.
2. The Right Forward has the highest median value among all the forward players.
3. The Right attacking midfield player has the highest median value among all the midfield players.
4. The trend also indicates a positive co-relation between overall rating and valuation as the position with higher overall rating fetch more values.

11. Overall player rating of the top 15 clubs

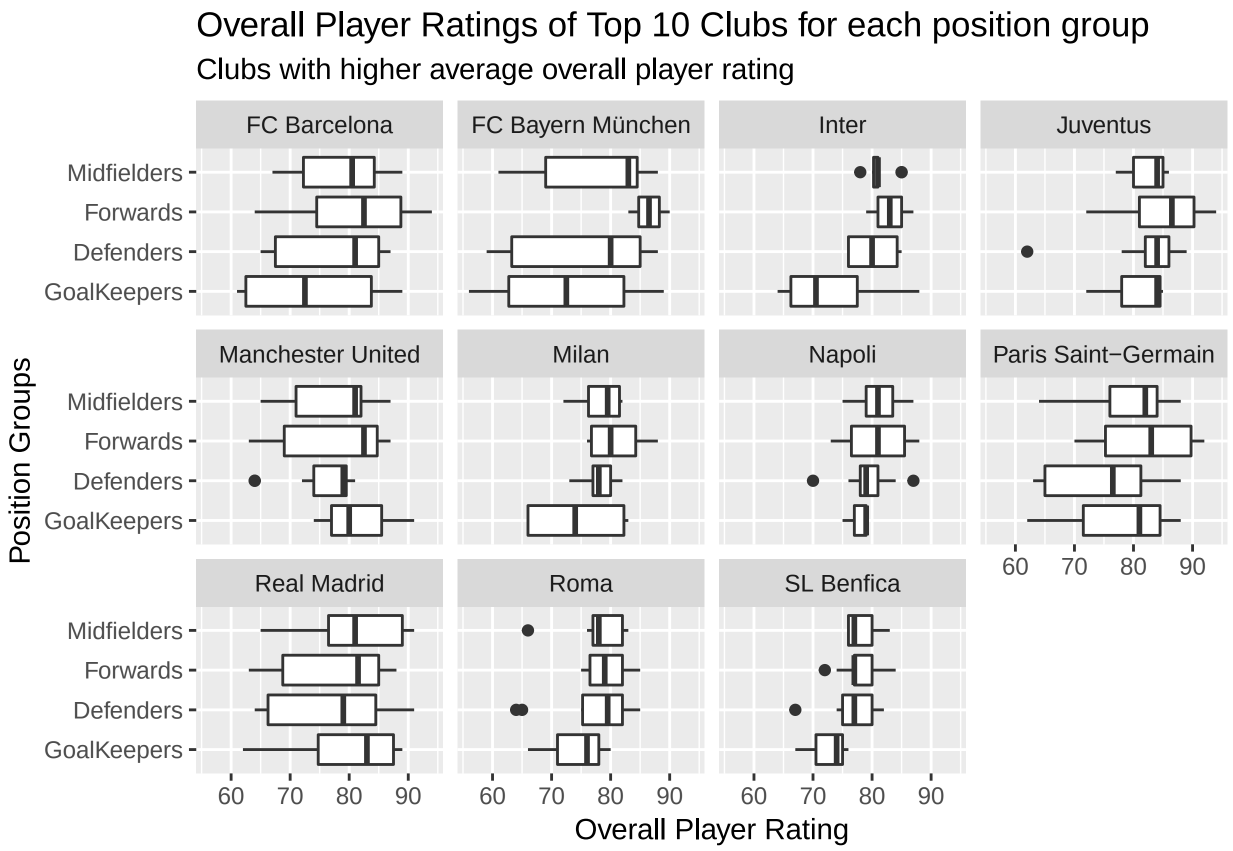
```
top15 <- fifa19%>%
  group_by(Club)%>%
  summarise(mean=mean(Overall))%>%
  arrange(desc(mean))%>%
  top_n(15)

fifa19%>%
  semi_join(top15,by="Club")%>%
  ggplot(aes(x=reorder(Club,Overall,FUN=median),y=Overall))+
  geom_boxplot()+
  coord_flip()+
  labs(title = "Overall Player rating for Top 15 clubs",
       subtitle = "Clubs with higher average overall player rating",
       x="Clubs",
       y="Overall Player Rating")
```



The top 15 teams have been selected based on their average overall player rating. The plot shows Juventus as team as a higher overall player rating, followed by FC Bayern Munich and FC Barcelona. SL Benfica, FC Porto and Roma are in the bottom three in the top 15 clubs with overall player rating.

12. Overall Player Ratings of Top 10 Clubs for each position group

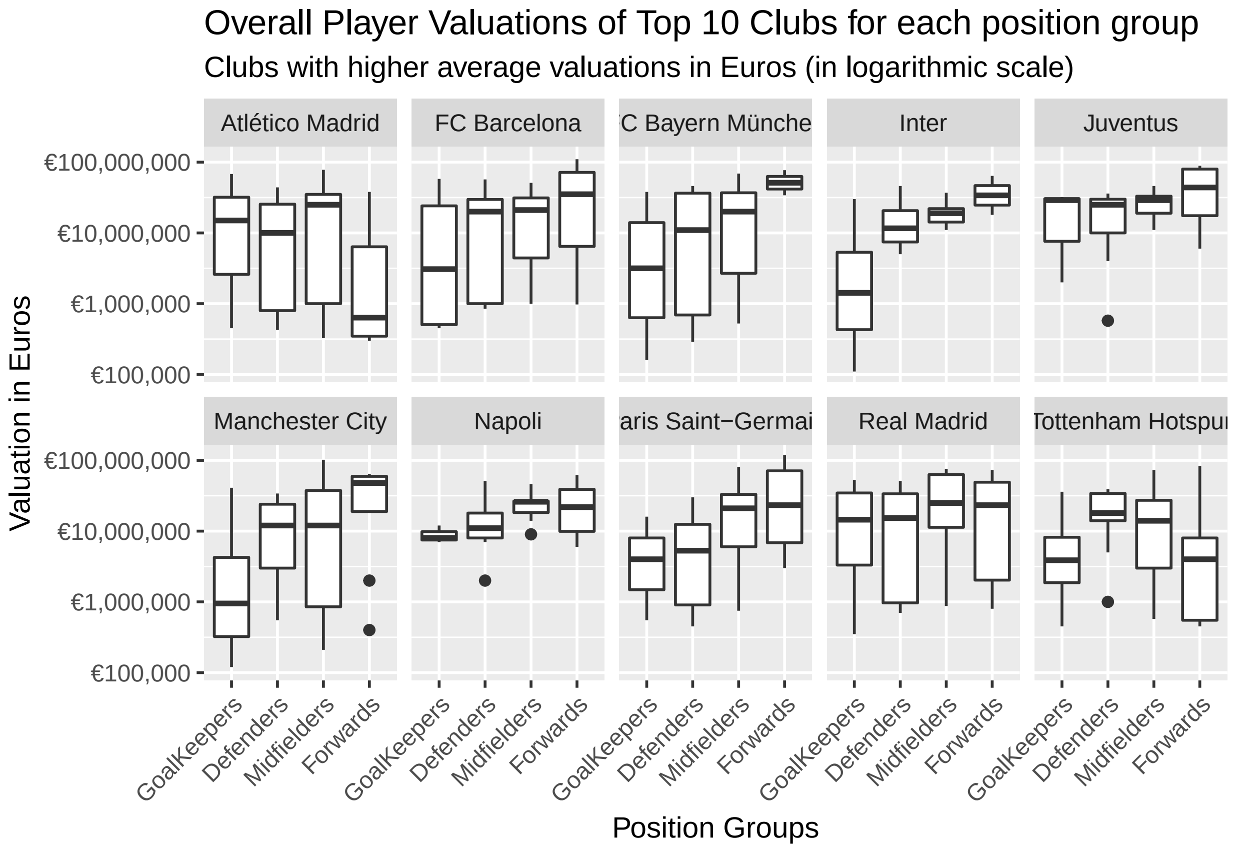
```
top10 <- fifa19%>%
  group_by(Club)%>%
  summarise(op=mean(Overall,na.rm = TRUE))%>%
  arrange(desc(op))%>%
  top_n(10,op)

fifa19%>%
  semi_join(top10,by="Club")%>%
  ggplot(aes(x=reorder(PositionGroup,Overall,FUN=median),y=Overall))+
  geom_boxplot()+
  coord_flip()+
  facet_wrap(~Club,nrow = 3)+
  labs(title = 'Overall Player Ratings of Top 10 Clubs for each position group',
       subtitle = 'Clubs with higher average overall player rating',
       x='Position Groups',
       y='Overall Player Rating')
```



Juventus as a team has a higher median overall rating in all the position groups among the top 10 teams. Median Overall rating of goalkeepers for Bracelona, Bayern Munich and Inter are lower compared to other teams in the TOp 10. This plot will help the teams to understand their weaknesses and will help them to understand where they stand against other teams and invest smartly in players.

13. Overall Player Valuations of Top 10 Clubs for each position group

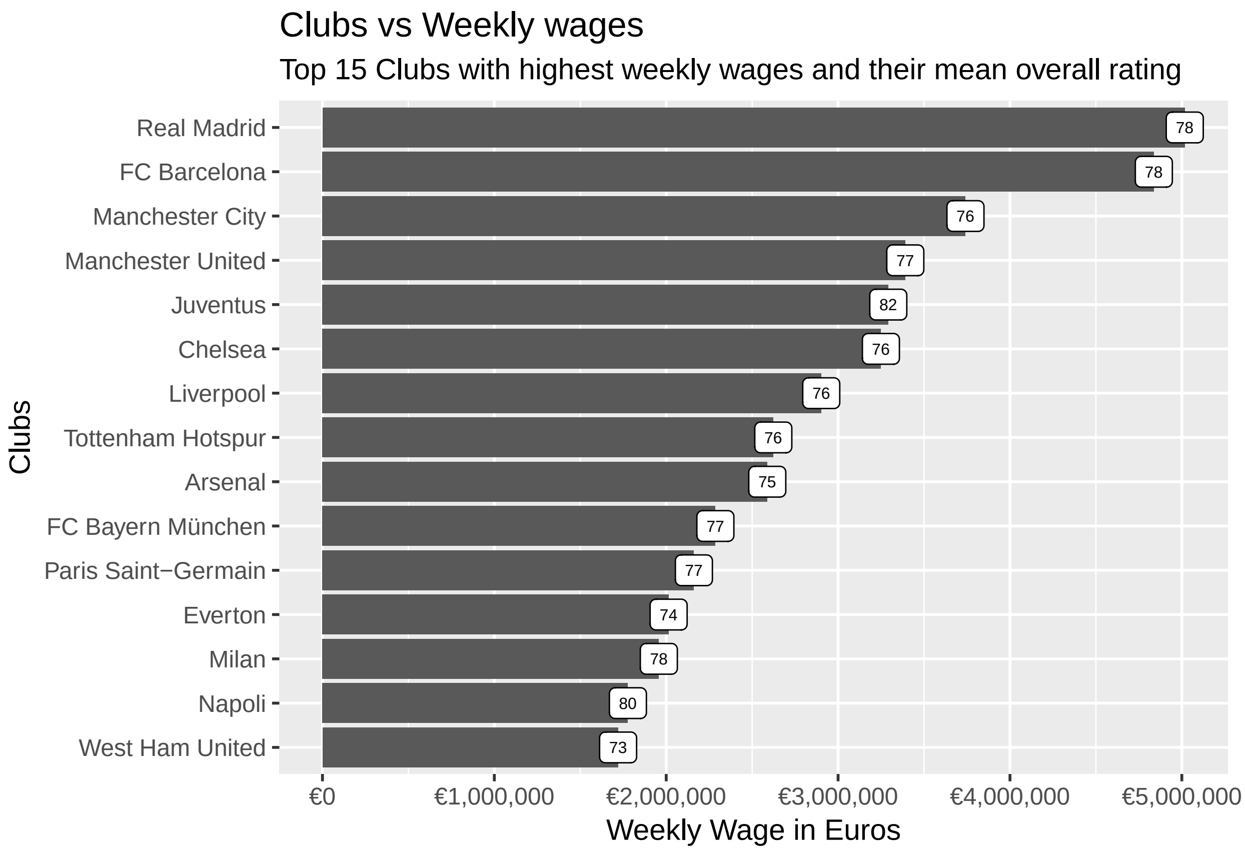
```
top10v <- fifa19%>%
  group_by(Club)%>%
  summarise(op=mean(Value_inEuros,na.rm = TRUE))%>%
  arrange(desc(op))%>%
  top_n(10,op)

fifa19%>%
  semi_join(top10v,by="Club")%>%
  ggplot(aes(x=reorder(PositionGroup,Value_inEuros,FUN=median),y=Value_inEuros))+
  geom_boxplot()+
  facet_wrap(~Club,nrow = 2)+
  scale_y_log10(labels = dollar_format(prefix = "€"))+
  labs(title = 'Overall Player Valuations of Top 10 Clubs for each position group',
       subtitle = 'Clubs with higher average valuations in Euros (in logarithmic scale)',
       x='Position Groups',
       y='Valuation in Euros')+
  theme(axis.text.x = element_text(angle=45, hjust = 1))
```



For the top 10 clubs with a higher average player valuations, teams like FC Barcelona, FC Bayern Munich, Inter, Juventus, Manchester City, PSG has a trend of Forwards has a higher median valuation follwed by midfielders, defenders and goalkeepers. However, for Athletico Madrid the forwards have a lower median valuation compared to even golkeepers and midfielders have a higher valuation. For Tottenham Hotspur, the defenders have a higher valuation compared to the other position groups.

14. Clubs vs Weekly wages

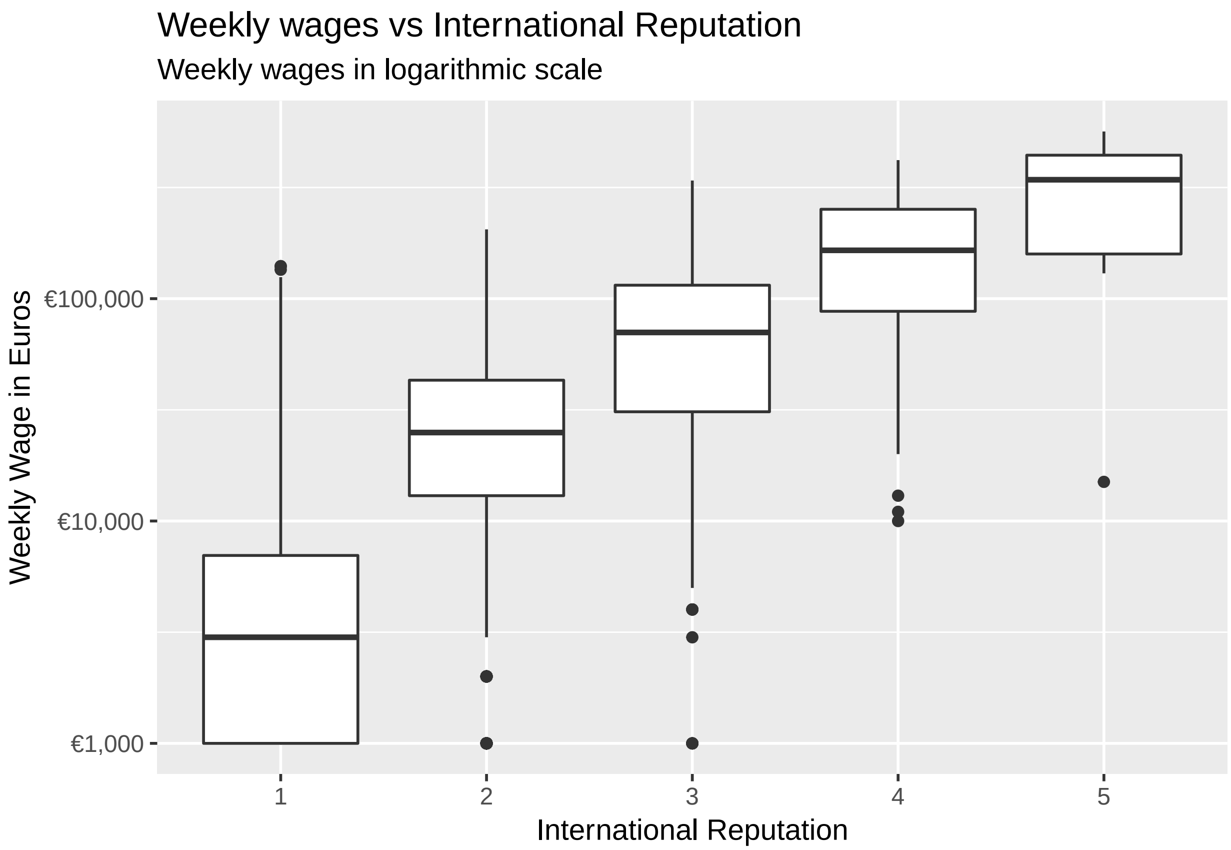
```
fifa19%>%
    group_by(Club)%>%
    summarise(TotalWage=sum(Wage_inEuros),
              mo=mean(Overall))%>%
    arrange(desc(TotalWage))%>%
    top_n(15,TotalWage)%>%
    ggplot(aes(x=reorder(Club,TotalWage),y=TotalWage))+
    geom_col()+
    coord_flip()+
    scale_y_continuous(labels = comma_format(prefix = "€"))+
    labs(title = "Clubs vs Weekly wages",
         subtitle = "Top 15 Clubs with highest weekly wages and their mean overall rating",
         x="Clubs",
         y="Weekly Wage in Euros")+
    geom_label(aes(label=floor(mo)), size = 2,nudge_y= 0.5)
```



Real Madrid has the highest weekly wage among all other teams followed by Barcelona and Manchester City. Clubs like Everton and West Ham United despite having lower mean overall player rating spends high in weekly wages.

15. Weekly wages vs International Reputation

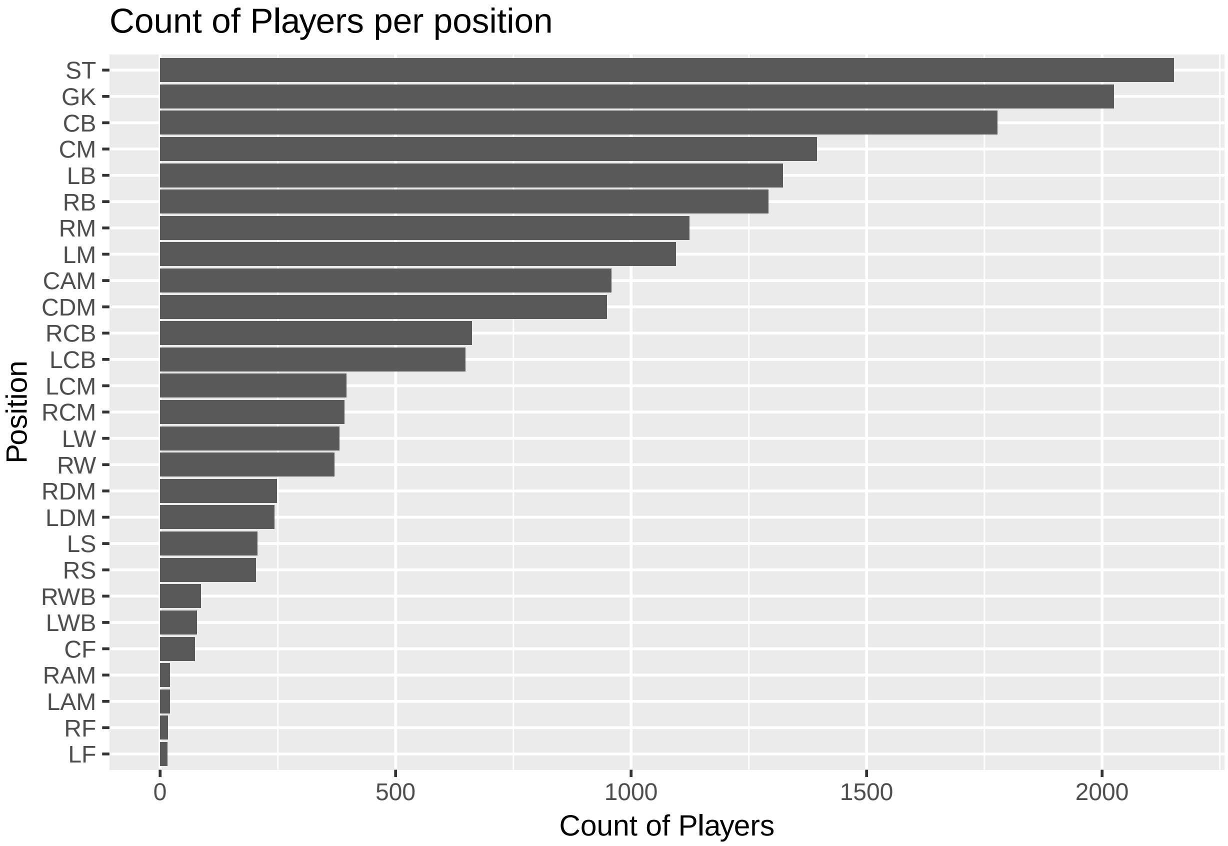
```
fifa19%>%
  mutate(Reputation=`International Reputation`)%>%
  mutate(Reputation=ifelse(is.na(Reputation),"Unknown",Reputation))%>%
  filter(Reputation!="Unknown")%>%
  ggplot(aes(x=as.factor(Reputation),y=Wage_inEuros))+
  geom_boxplot()+
  scale_y_log10(labels=comma_format(prefix = "€"))+
  labs(title = "Weekly wages vs International Reputation",
       subtitle = "Weekly wages in logarithmic scale",
       x="International Reputation",
       y="Weekly Wage in Euros")
```



Clearly the wages increase as the reputation increases. 5 points being the highest for international reputation in the data set has a higher median for weekly wages.

16. Count of Players per position

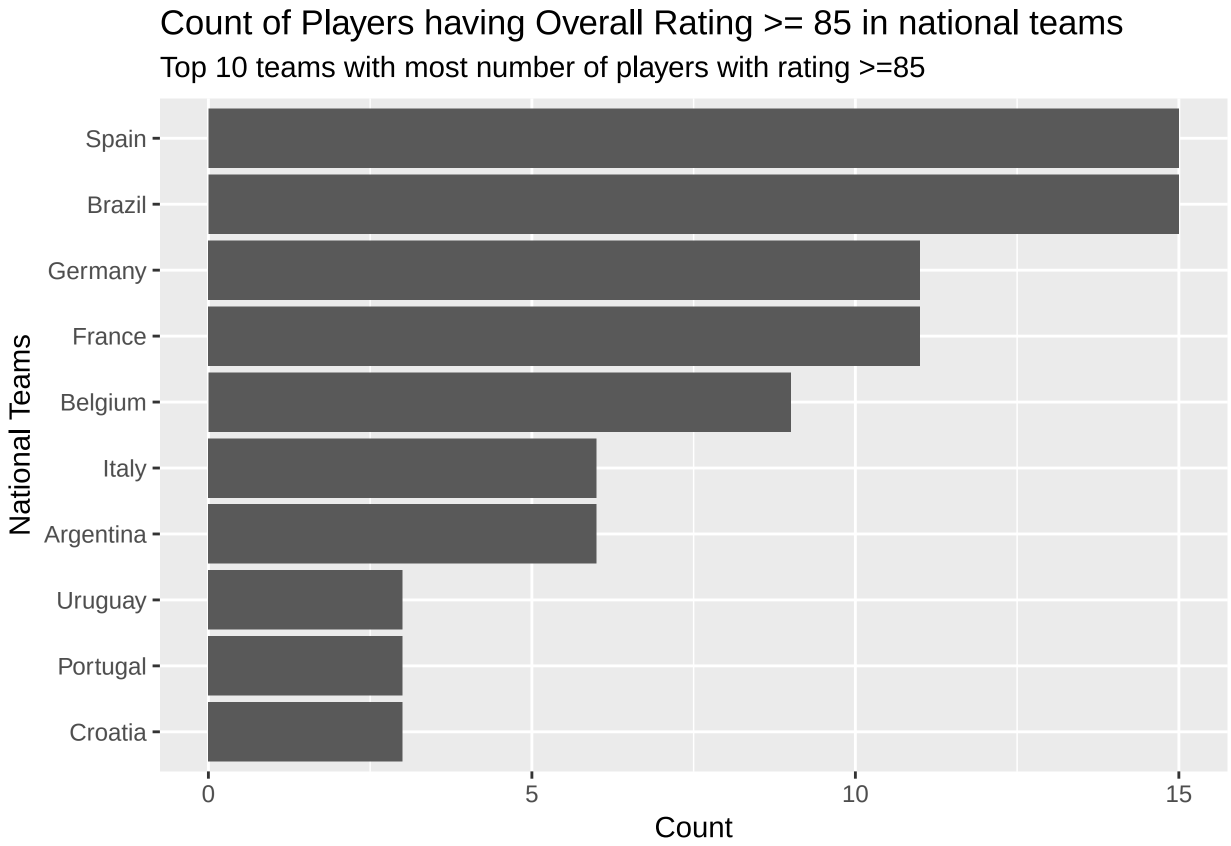
```
fifa19%>%
  filter(Position!="Unknown")%>%
  count(Position)%>%
  ggplot(aes(x=reorder(Position,n),y=n))+
  geom_col()+
  coord_flip()+
  labs(title = "Count of Players per position",
       x="Position",
       y="Count of Players")
```



Strikers are in large numbers followed by Goalkeepers and Centre Backs. Very few players for the Left attacking midfield, Right Forward and Left forward, maybe because teams generally don't prefer playing with formations that would require this positions.

17. Count of Players having Overall Rating >= 85 in National teams

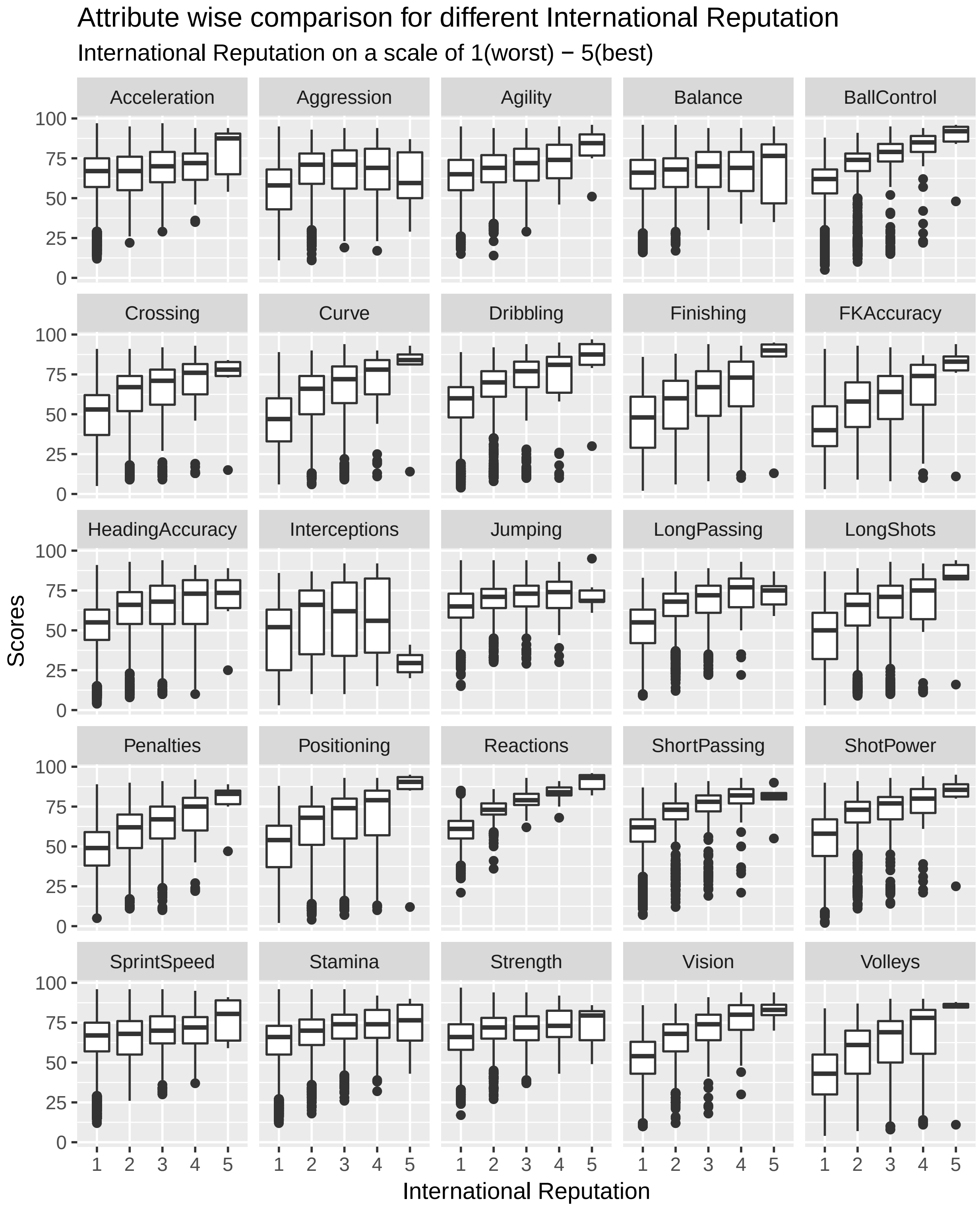
```
fifa19%>%
  filter(Overall>=85)%>%
  count(Nationality)%>%
  arrange(desc(n))%>%
  top_n(10)%>%
  ggplot(aes(x=reorder(Nationality,n),y=n))+
  geom_col()+
  coord_flip()+
  labs(title = "Count of Players having Overall Rating >= 85 in national teams",
       subtitle = "Top 10 teams with most number of players with rating >=85",
       x="National Teams",
       y="Count")
```



The count of players with overall rating greater than or equal to 85 is higher for Spain and Brazil followed by Germany and France.

18.Attribute wise comparison for different International Reputation

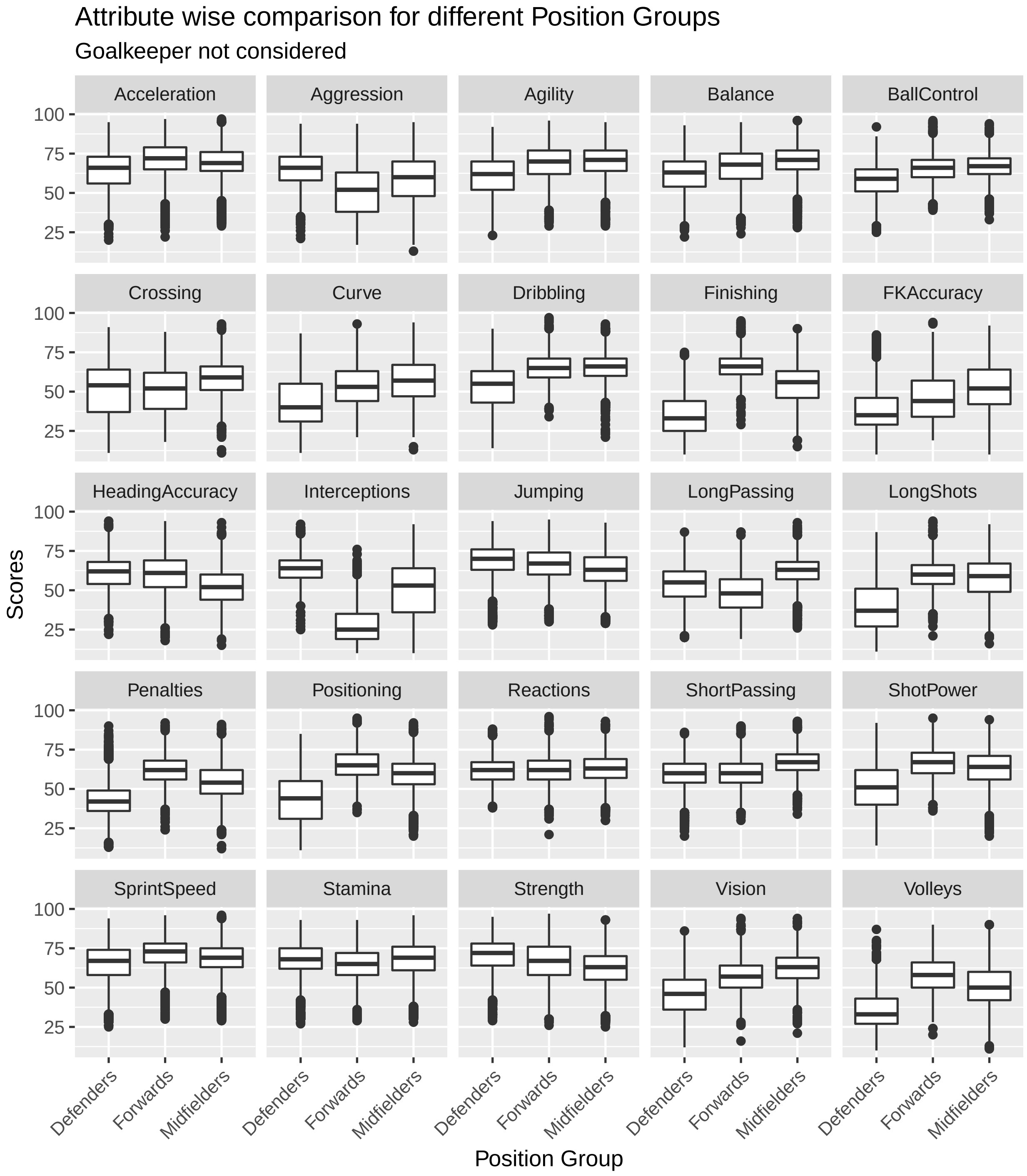
```r
fifa19%>%
  mutate(Reputation=`International Reputation`)%>%
  mutate(Reputation=ifelse(is.na(Reputation),"Unknown",Reputation))%>%
  filter(Reputation!="Unknown")%>%
  gather(Crossing,Finishing,HeadingAccuracy,ShortPassing,
         Volleys,Dribbling,Curve,FKAccuracy,LongPassing,
         BallControl,Acceleration,SprintSpeed,Agility,
         Reactions,Balance,ShotPower,Jumping,Stamina,Strength,LongShots,
         Aggression,Interceptions,Positioning,Vision,Penalties,
         key = "Feature",value="Scores")%>%
  ggplot(aes(x=as.factor(`International Reputation`),y=Scores))+
           geom_boxplot()+
           facet_wrap(~Feature)+
  labs(title = "Attribute wise comparison for different International Reputation",
       subtitle = "International Reputation on a scale of 1(worst) - 5(best)",
       x="International Reputation",
       y="Scores")
```



The players with best international reputation has better Agility, Ball Control, Crossing, Curve, Dribbling, Finishing, First kick accuracy, Heading Accuracy, Long Shots, Shot Power ,Penalties among other attributes. But the players with best reputation does not have better aggression, jumping, interceptions, stamina and strength, which helps to understand that the players should focus on the skills mentioned in the first paragraph and improve them to gain a better reputation and value.

19. Attribute wise comparison for different Position Groups

```r
fifa19%>%
  filter(PositionGroup!="GoalKeepers",PositionGroup!="Unknown")%>%
  gather(Crossing,Finishing,HeadingAccuracy,ShortPassing,
         Volleys,Dribbling,Curve,FKAccuracy,LongPassing,
         BallControl,Acceleration,SprintSpeed,Agility,
         Reactions,Balance,ShotPower,Jumping,Stamina,Strength,LongShots,
         Aggression,Interceptions,Positioning,Vision,Penalties,
         key = "Feature",value="Scores")%>%
  ggplot(aes(x=PositionGroup,y=Scores))+
  geom_boxplot()+
  facet_wrap(~Feature)+
  labs(title = "Attribute wise comparison for different Position Groups",
       subtitle = "Goalkeeper not considered",
       x="Position Group",
       y="Scores")+
  theme(axis.text.x = element_text(angle=45, hjust = 1))
```



Forwards have better acceleration, finishing, long shots,volleys, penalty taking and positioning skills

Midfielders have better agility, balance, ball control, first kick accuracy, dribbling, curve, crossing skills.

Defenders have better aggression, jumping, interception, heading accuracy, strength and stamina.

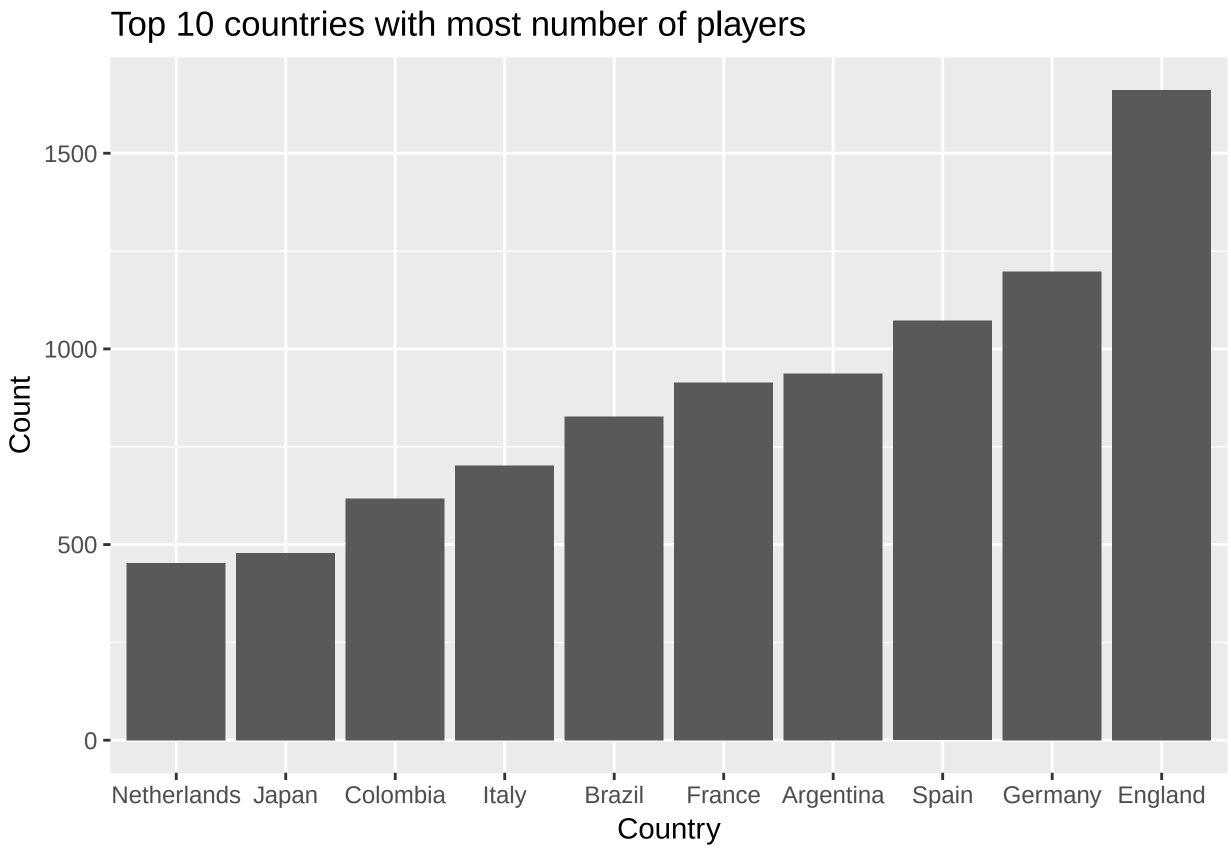
20. Best player for each position

```
fifa19%>%
  filter(Position!="Unknown")%>%
  group_by(Position)%>%
  arrange(desc(Overall))%>%
  top_n(1,Overall)%>%
  select(Position,Name)%>%
  distinct(Position,.keep_all = TRUE)%>%
  print(n=28)
```

```
## # A tibble: 27 x 2
## # Groups:   Position [27]
##    Position Name
##    <chr>    <chr>
##  1 RF       L. Messi
##  2 ST       Cristiano Ronaldo
##  3 LW       Neymar Jr
##  4 GK       De Gea
##  5 RCM      K. De Bruyne
##  6 LF       E. Hazard
##  7 RS       L. Suárez
##  8 RCB      Sergio Ramos
##  9 LCM      T. Kroos
## 10 CB       D. Godín
## 11 LDM      N. Kanté
## 12 CAM      A. Griezmann
## 13 CDM      Sergio Busquets
## 14 LS       E. Cavani
## 15 LCB      G. Chiellini
## 16 RM       K. Mbappé
## 17 LAM      J. Rodríguez
## 18 LM       P. Aubameyang
## 19 LB       Marcelo
## 20 RDM      P. Pogba
## 21 RW       Bernardo Silva
## 22 CM       Thiago
## 23 RB       Azpilicueta
## 24 RAM      J. Cuadrado
## 25 CF       Luis Alberto
## 26 RWB      M. Ginter
## 27 LWB      N. Schulz
```

21. Count of players for each country

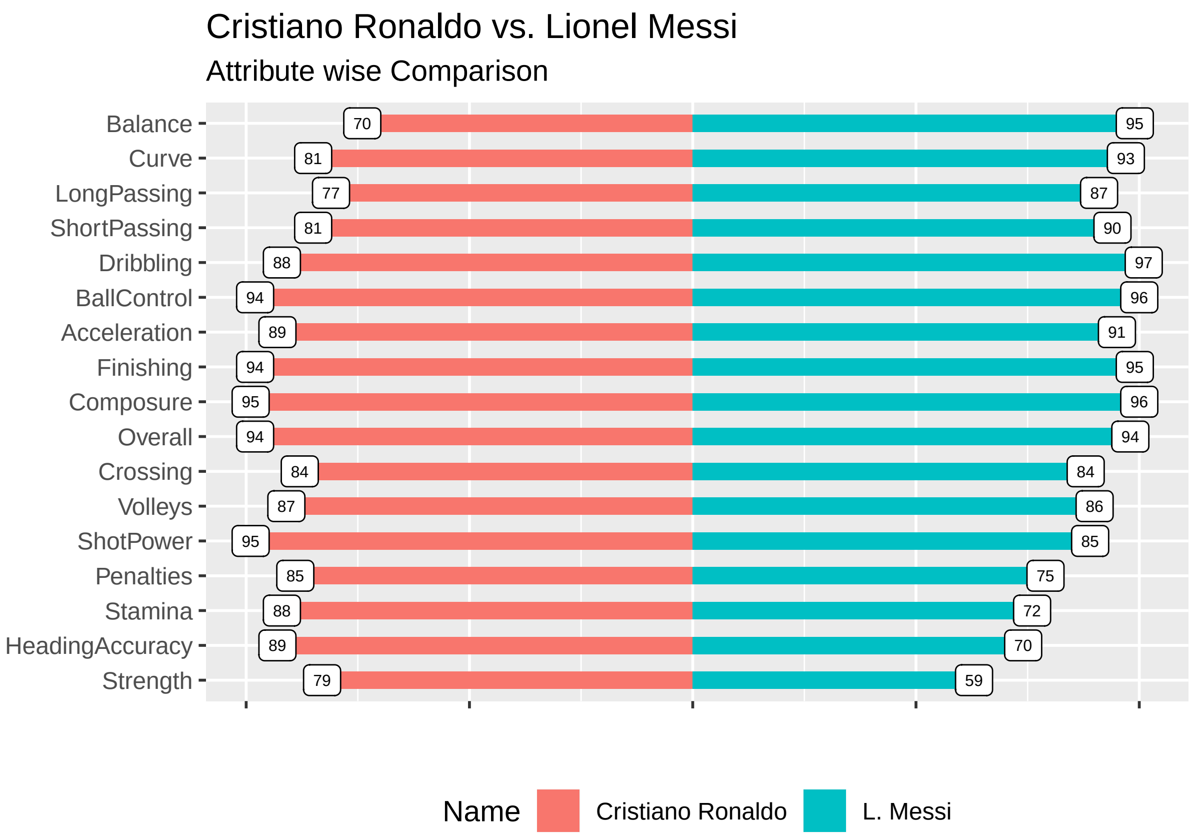
```r
fifa19%>%
  count(Nationality)%>%
  arrange(desc(n))%>%
  top_n(10,n)%>%
  ggplot(aes(x=reorder(Nationality,n),y=n))+
  geom_col()+
  labs(title = "Top 10 countries with most number of players",
       x="Country",
       y="Count")
```



England has the highest number of players followed by Germany and Spain among the top 10 nations with most number of players.

22. Cristiano Ronaldo vs. Lionel Messi

```r
M_R <- fifa19%>%
  filter(Name=="L. Messi"|Name=="Cristiano Ronaldo")%>%
  select(Name,Overall,Crossing,Finishing,HeadingAccuracy,
         ShortPassing,Volleys,Dribbling,Curve,LongPassing,
         BallControl,Acceleration,Balance,ShotPower,Stamina,
         Strength,Penalties,Composure)%>%
  gather(Overall,Crossing,Finishing,HeadingAccuracy,
         ShortPassing,Volleys,Dribbling,Curve,LongPassing,
         BallControl,Acceleration,Balance,ShotPower,Stamina,
         Strength,Penalties,Composure,key = "Features",value = "Scores")%>%
  mutate(Scores=ifelse(Name=="Cristiano Ronaldo",Scores*-1,Scores))%>%
  arrange(Scores)%>%
  mutate(nudge=ifelse(Scores>0,4,-4))


ggplot(data=M_R,aes(x=reorder(Features,Scores),y=Scores))+
  geom_bar(stat = "identity",aes(fill=Name),width=.5)+
  labs(title = 'Cristiano Ronaldo vs. Lionel Messi',
       subtitle = 'Attribute wise Comparison',
       x='',
       y='')+
  geom_label(aes(label=abs(Scores)), size = 2,nudge_y =M_R$nudge)+
  theme(axis.text.x=element_blank(),
       legend.position = 'bottom')+
  coord_flip()
```



Messi clearly has better Balance, First Kick Accuracy, Curve, Long and Short Passing, dribbling compared to Ronaldo and have similar attributes while in terms of Ball Control, Acceleration, Long Shots, Finishing and Composure, but Ronaldo is better than Messi when it comes to Volleys, Sprint Speed, Shot Power, Penalties, Stamina, Heading Accuracy and strength. The plot is there for you to decide who is the best.

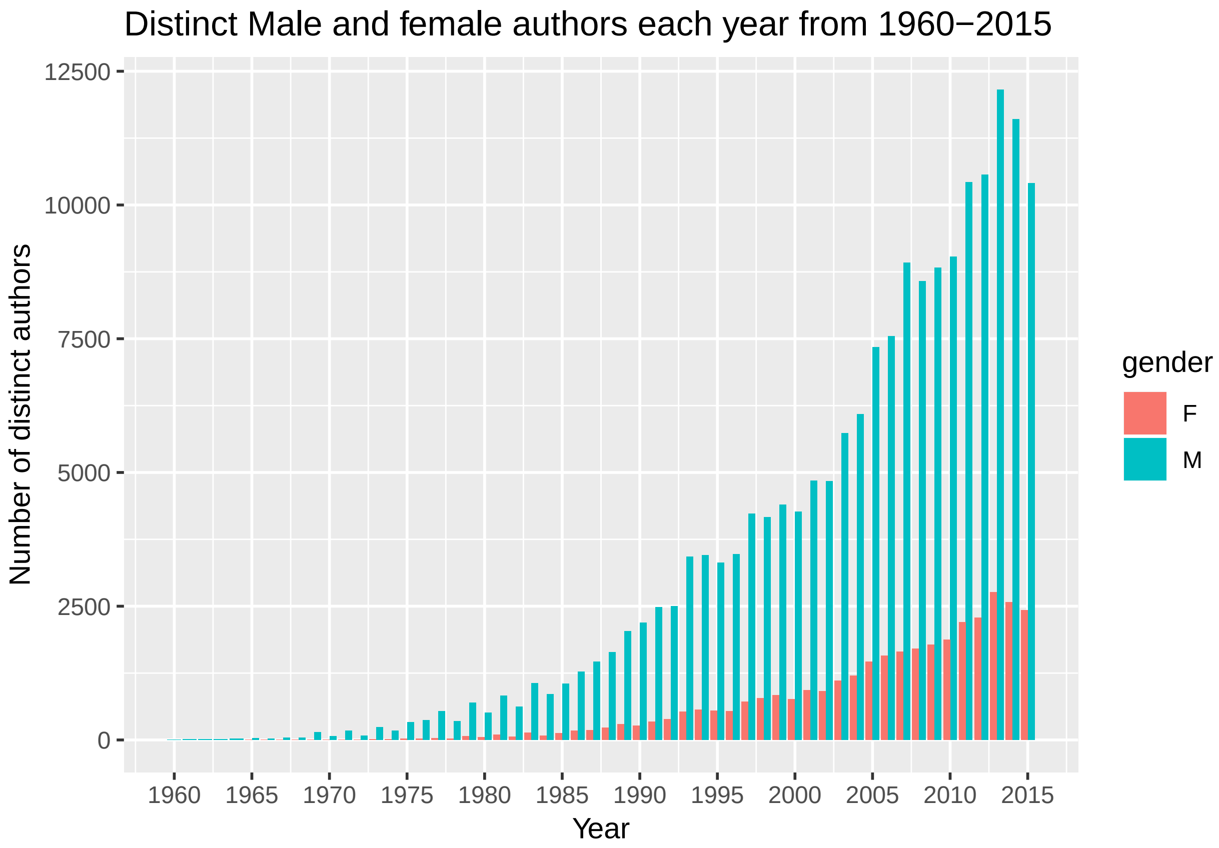
Part B

** Creating DB connection **

```r
con <- dbConnect(RMySQL::MySQL(), host = "localhost",
                 user = "root", password = "root",dbname="dblp")
dblp_authors <- tbl(con,"authors")
dblp_general <- tbl(con,"general")
```

Problem 3

```
dblp_authors %>%
  left_join(dblp_general) %>%
  filter(prob>=0.95 & gender %in% c("M","F"))%>%
  group_by(year,gender) %>%
  summarise(num_authors = n_distinct(name)) %>%
  ggplot() +
  geom_col(aes(x=year, y=num_authors,fill=gender),position = "dodge")+
  scale_x_continuous(breaks=
                       c(1960,1965,1970,1975,1980,1985,1990,1995,2000,2005,2010,2015))+
  labs(title = "Distinct Male and female authors each year from 1960-2015",
       x="Year",
       y="Number of distinct authors")
```

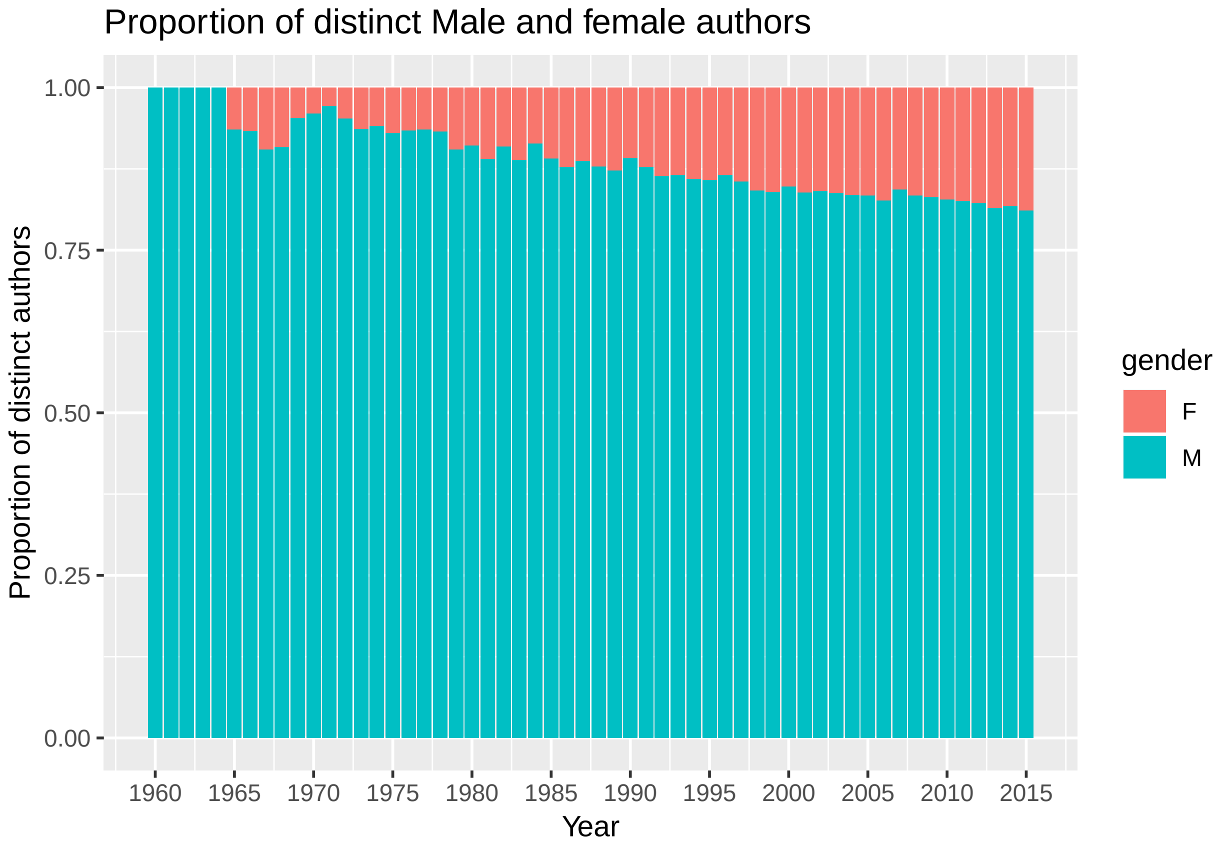


The number of distinct male authors have gradually increased over the years and it was highest in the year 2013, however the numbers started slightly coming down after 2013.

The number of distinct female authors are comparatively less than the male authors. The number of distinct female authors have also gradually increased, however at a smaller rate compared to male authors. The year 2013 witnessed maximum distinct female author and post that the number slightly came down for the female authors also.

Problem 4

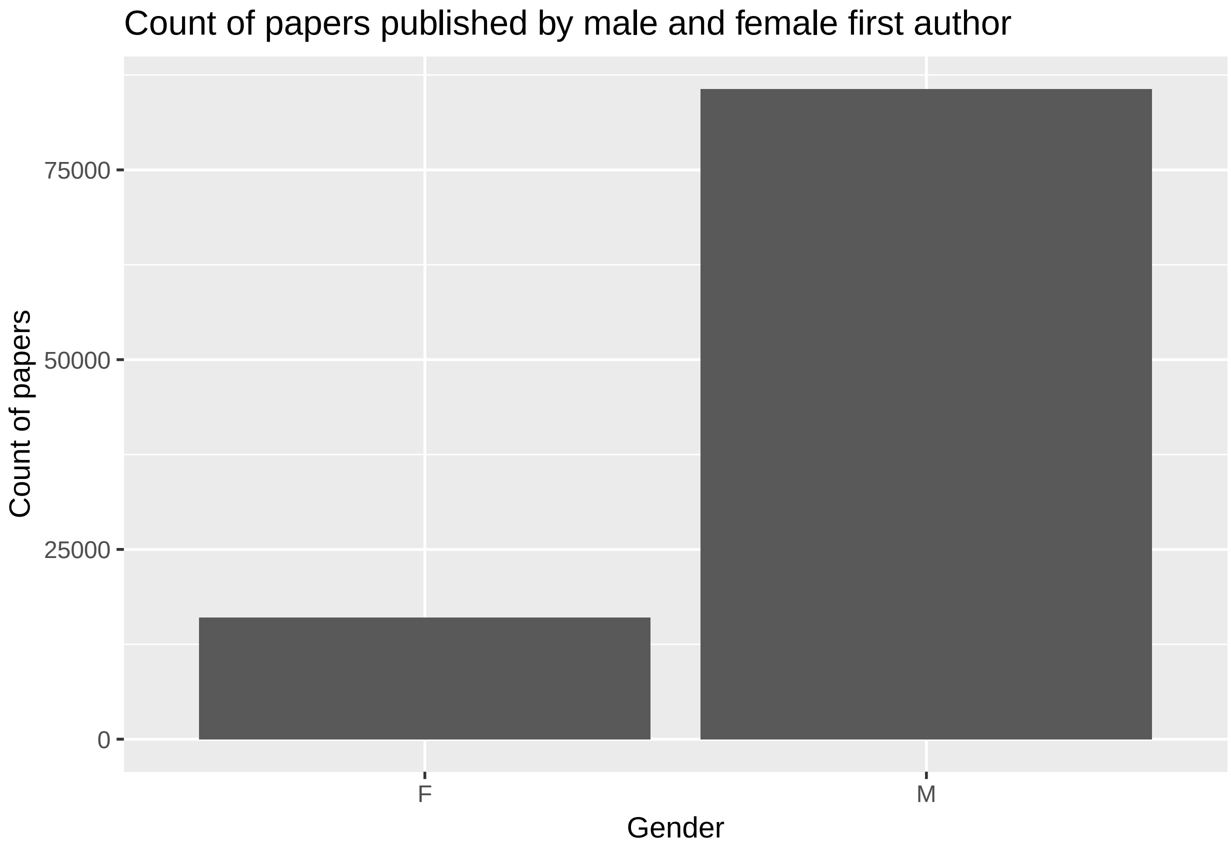
```
dblp_authors %>%
  left_join(dblp_general) %>%
  filter(prob>=0.95 & gender %in% c("M","F"))%>%
  group_by(year,gender) %>%
  summarise(num_authors = n_distinct(name)) %>%
  ggplot() +
  geom_col(aes(x=year, y=num_authors,fill=gender),position = "fill")+
  scale_x_continuous(breaks=
                       c(1960,1965,1970,1975,1980,1985,1990,1995,2000,2005,2010,2015))+
  labs(title = "Proportion of distinct Male and female authors",
       x="Year",
       y="Proportion of distinct authors")
```



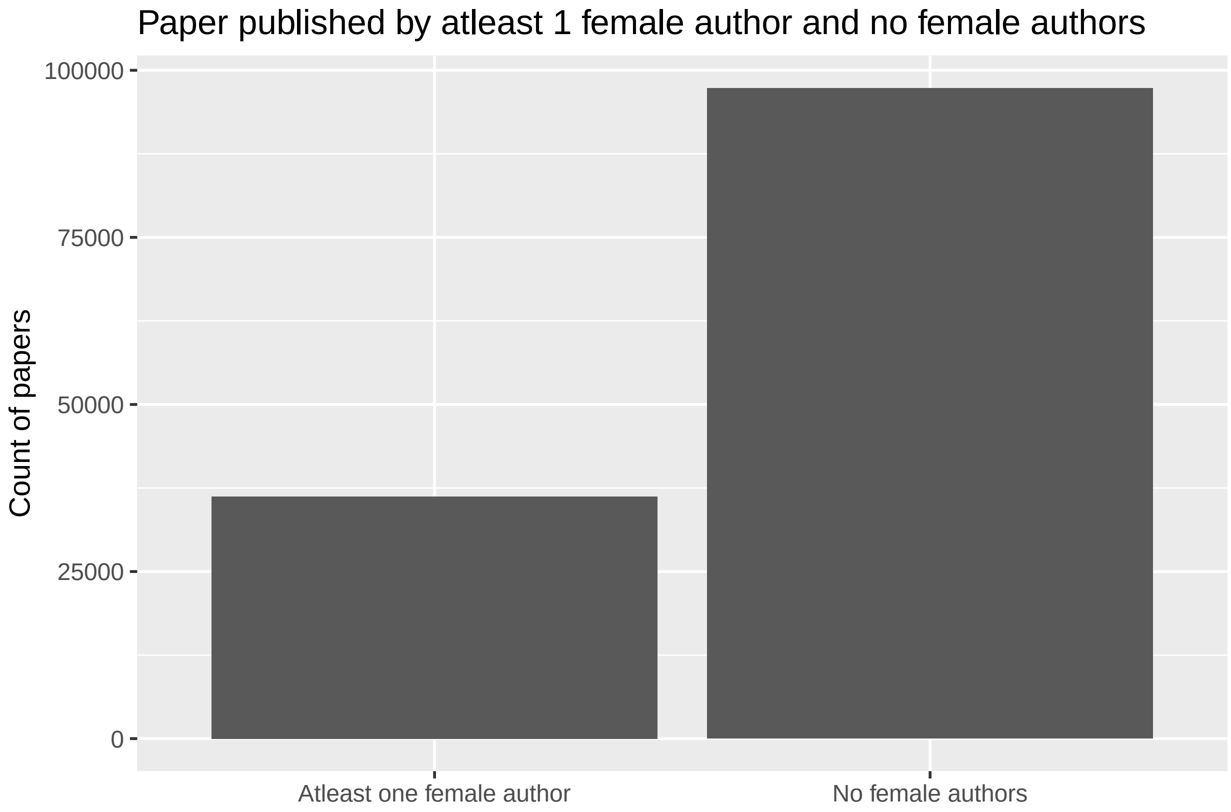
The plot indicates that there were no female authors for the years 1960-1964, as the proportion of distinct male authors is 1. The graph shows that the propotion of the distinct female authors is slowly increasing over the years. However, the proportion of distinct female authors is less compared to the distinct male authors.

Problem 5

```
dblp_authors%>%
  left_join(dblp_general,by="k")%>%
  filter(prob>=0.95 & gender %in% c("M","F"))%>%
  filter(pos==0)%>%
  collect%>%
  ggplot()+
  geom_bar(aes(x=gender))+
  labs(title="Count of papers published by male and female first author",
       x = "Gender",
       y ="Count of papers")
```



```
dblp_authors%>%
    left_join(dblp_general,by="k")%>%
    filter(prob>=0.95 & gender %in% c("M","F"))%>%
    collect%>%
    gather(key="papers", value="is_published",
           cs, de, se, th) %>%
    filter(is_published==1)%>%
    select(-is_published)%>%
    group_by(k)%>%
    summarise(F_Authors=sum(gender=="F"))%>%
    mutate(F_Authors=ifelse(F_Authors==0,
                            "No female authors","Atleast one female author"))%>%
    ggplot(aes(x=F_Authors))+
    geom_bar()+
    labs(title="Paper published by atleast 1 female author and no female authors",
       x="",
       y="Count of papers")
```



The count of papers published by male first authors are very high as compared to the papers published by female first authors. The trend is similar in the second plot as well where the number of papers published by no female authors which is actually the papers published by only male authors is higher compared to the papers published by atleast one female author.

However, the number of papers published by atleast one female author has increased compared to the papers published by female first author which indicates that the female authors are more involved as the second, third or subsequent authors.