Domain: Sports

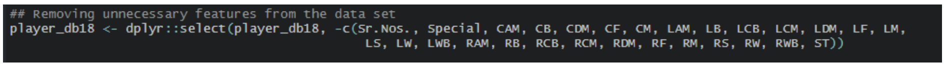
Problem Statement:

Using the FIFA 18 & FIFA 19 datasets from kaggel.com to try and understand the features / variables contributing towards the value and over all of the player & try to build a prediction model to predict the overall & value of the players on FIFA 20 dataset. As well as doing exploratory analysis by visualizing the data. And at last creating a player recommendation model.

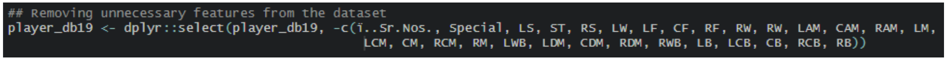
There being more than 18000 players in FIFA, it becomes very confusing to pick players for the team. There are many players who can give you similar attributes and skills, and this being a game that's all that matters. There can be a situation where you cannot afford a player or you can't choose him because he is under a contract or maybe one of your player is leaving and you need to replace him, etc. in that situation you can use this model to find similar player.

Plan of Action:

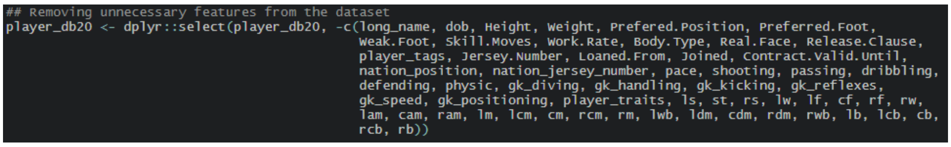
1. Initial Data Cleaning (Using Excel & R)
    a. Removing unnecessary fields (Using R)



```
## Removing unnecessary features from the data set
player_db18 <- dplyr::select(player_db18, -c(Sr.Nos., Special, CAM, CB, CDM, CF, CM, LAM, LB, LCB, LCM, LDM, LF, LM,
                                             LS, LW, LWB, RAM, RB, RCB, RCM, RDM, RF, RM, RS, RW, RWB, ST))
```



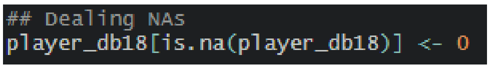
```
## Removing unnecessary features from the dataset
player_db19 <- dplyr::select(player_db19, -c(ï..Sr.Nos., Special, LS, ST, RS, LW, LF, CF, RF, RW, RW, LAM, CAM, RAM, LM,
                                             LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB))
```



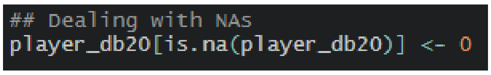
```
## Removing unnecessary features from the dataset
player_db20 <- dplyr::select(player_db20, -c(long_name, dob, Height, Weight, Prefered.Position, Preferred.Foot,
                                             Weak.Foot, Skill.Moves, Work.Rate, Body.Type, Real.Face, Release.Clause,
                                             player_tags, Jersey.Number, Loaned.From, Joined, Contract.Valid.Until,
                                             nation_position, nation_jersey_number, pace, shooting, passing, dribbling,
                                             defending, physic, gk_diving, gk_handling, gk_kicking, gk_reflexes,
                                             gk_speed, gk_positioning, player_traits, 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))
```

    b. Simplifying the position into Goal Keeper (GK), Right Back (RB), Centre Back (CB), Left Back (LB), Defensive Midfielder (DM), Centre Midfielders (CM), Wingers & Forwards. (Using Excel)

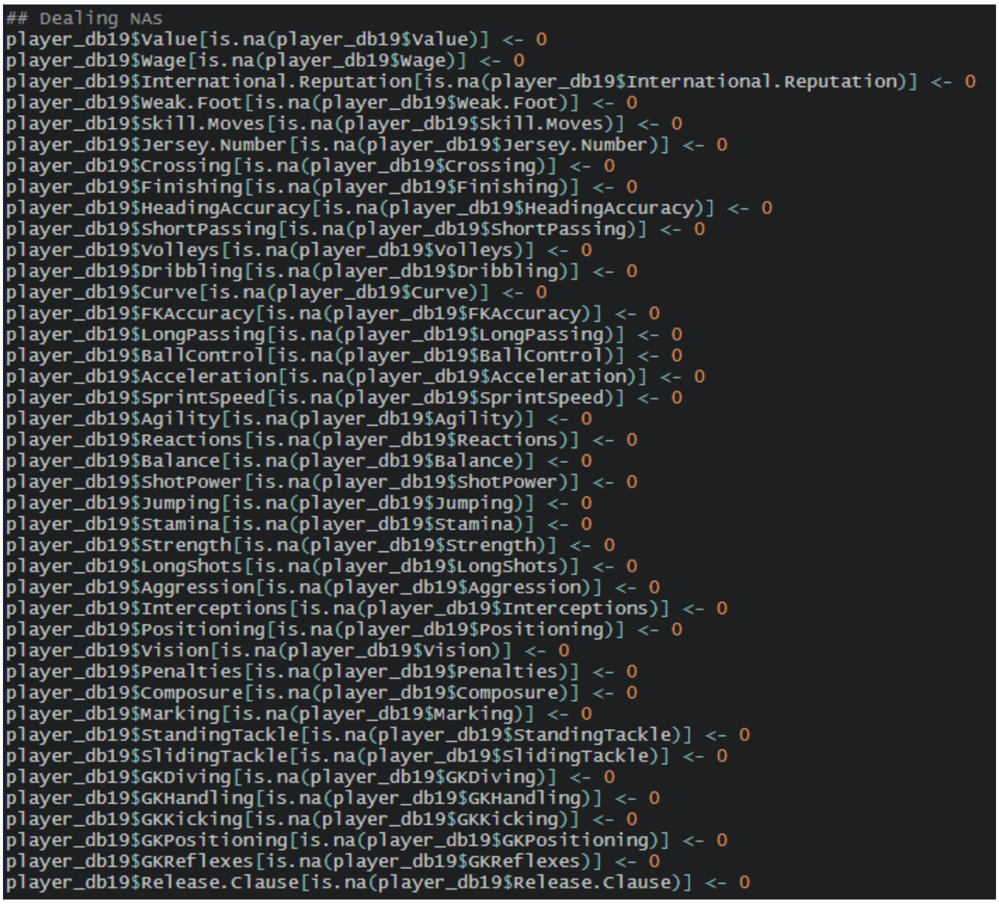
c. Dealing with NAs. (Using R)



```
## Dealing NAs
player_db18[is.na(player_db18)] <- 0
```

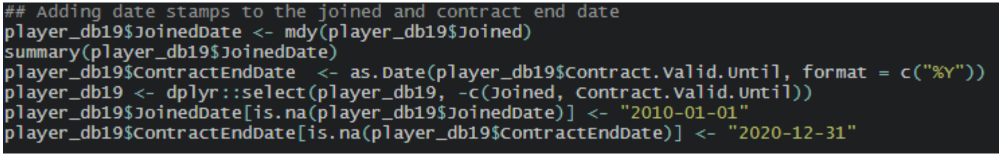


```
## Dealing with NAs
player_db20[is.na(player_db20)] <- 0
```



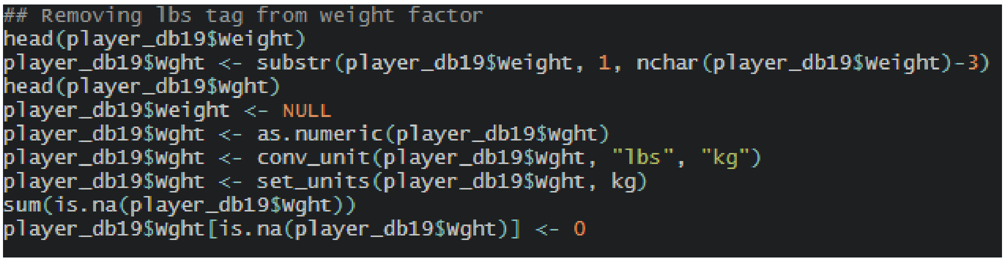
```
## Dealing NAs
player_db19$Value[is.na(player_db19$Value)] <- 0
player_db19$Wage[is.na(player_db19$Wage)] <- 0
player_db19$International.Reputation[is.na(player_db19$International.Reputation)] <- 0
player_db19$Weak.Foot[is.na(player_db19$Weak.Foot)] <- 0
player_db19$Skill.Moves[is.na(player_db19$Skill.Moves)] <- 0
player_db19$Jersey.Number[is.na(player_db19$Jersey.Number)] <- 0
player_db19$Crossing[is.na(player_db19$Crossing)] <- 0
player_db19$Finishing[is.na(player_db19$Finishing)] <- 0
player_db19$HeadingAccuracy[is.na(player_db19$HeadingAccuracy)] <- 0
player_db19$ShortPassing[is.na(player_db19$ShortPassing)] <- 0
player_db19$Volleys[is.na(player_db19$Volleys)] <- 0
player_db19$Dribbling[is.na(player_db19$Dribbling)] <- 0
player_db19$Curve[is.na(player_db19$Curve)] <- 0
player_db19$FKAccuracy[is.na(player_db19$FKAccuracy)] <- 0
player_db19$LongPassing[is.na(player_db19$LongPassing)] <- 0
player_db19$BallControl[is.na(player_db19$BallControl)] <- 0
player_db19$Acceleration[is.na(player_db19$Acceleration)] <- 0
player_db19$SprintSpeed[is.na(player_db19$SprintSpeed)] <- 0
player_db19$Agility[is.na(player_db19$Agility)] <- 0
player_db19$Reactions[is.na(player_db19$Reactions)] <- 0
player_db19$Balance[is.na(player_db19$Balance)] <- 0
player_db19$ShotPower[is.na(player_db19$ShotPower)] <- 0
player_db19$Jumping[is.na(player_db19$Jumping)] <- 0
player_db19$Stamina[is.na(player_db19$Stamina)] <- 0
player_db19$Strength[is.na(player_db19$Strength)] <- 0
player_db19$LongShots[is.na(player_db19$LongShots)] <- 0
player_db19$Aggression[is.na(player_db19$Aggression)] <- 0
player_db19$Interceptions[is.na(player_db19$Interceptions)] <- 0
player_db19$Positioning[is.na(player_db19$Positioning)] <- 0
player_db19$Vision[is.na(player_db19$Vision)] <- 0
player_db19$Penalties[is.na(player_db19$Penalties)] <- 0
player_db19$Composure[is.na(player_db19$Composure)] <- 0
player_db19$Marking[is.na(player_db19$Marking)] <- 0
player_db19$StandingTackle[is.na(player_db19$StandingTackle)] <- 0
player_db19$SlidingTackle[is.na(player_db19$SlidingTackle)] <- 0
player_db19$GKDiving[is.na(player_db19$GKDiving)] <- 0
player_db19$GKHandling[is.na(player_db19$GKHandling)] <- 0
player_db19$GKKicking[is.na(player_db19$GKKicking)] <- 0
player_db19$GKPositioning[is.na(player_db19$GKPositioning)] <- 0
player_db19$GKReflexes[is.na(player_db19$GKReflexes)] <- 0
player_db19$Release.Clause[is.na(player_db19$Release.Clause)] <- 0
```

d. Changing joined date and contract valid date to time stamps. (Using R)



```
## Adding date stamps to the joined and contract end date
player_db19$JoinedDate <- mdy(player_db19$Joined)
summary(player_db19$JoinedDate)
player_db19$ContractEndDate  <- as.Date(player_db19$Contract.Valid.Until, format = c("%Y"))
player_db19 <- dplyr::select(player_db19, -c(Joined, Contract.Valid.Until))
player_db19$JoinedDate[is.na(player_db19$JoinedDate)] <- "2010-01-01"
player_db19$ContractEndDate[is.na(player_db19$ContractEndDate)] <- "2020-12-31"
```

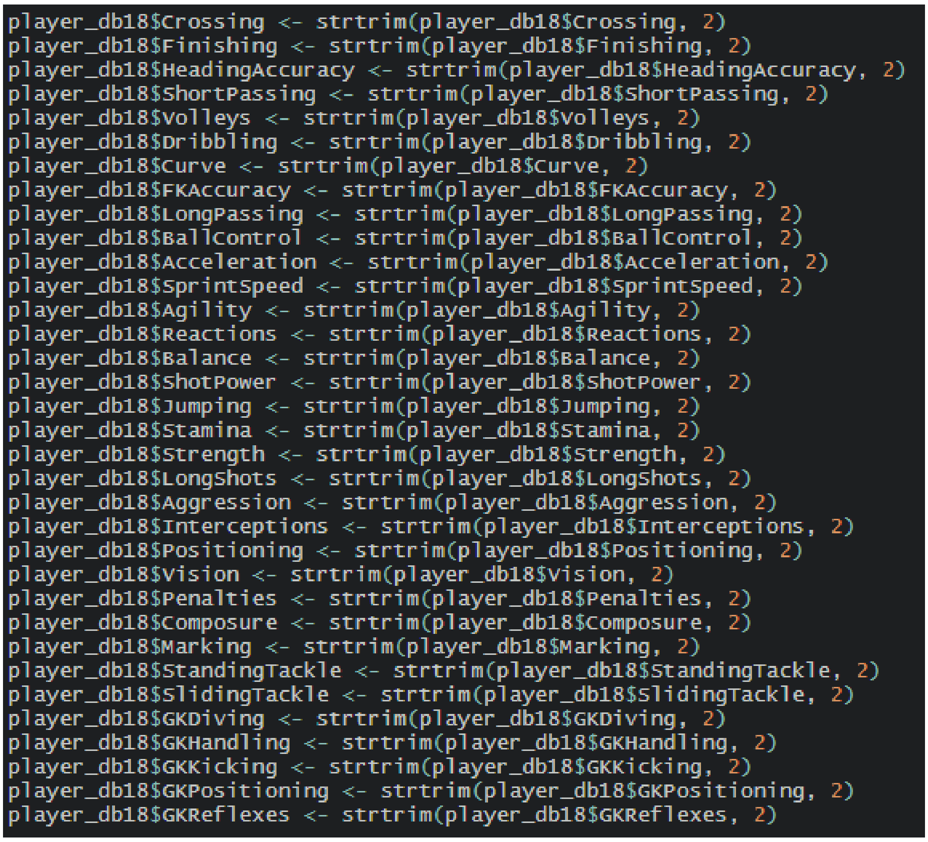
e. Removing units from weight feature. (Using R)



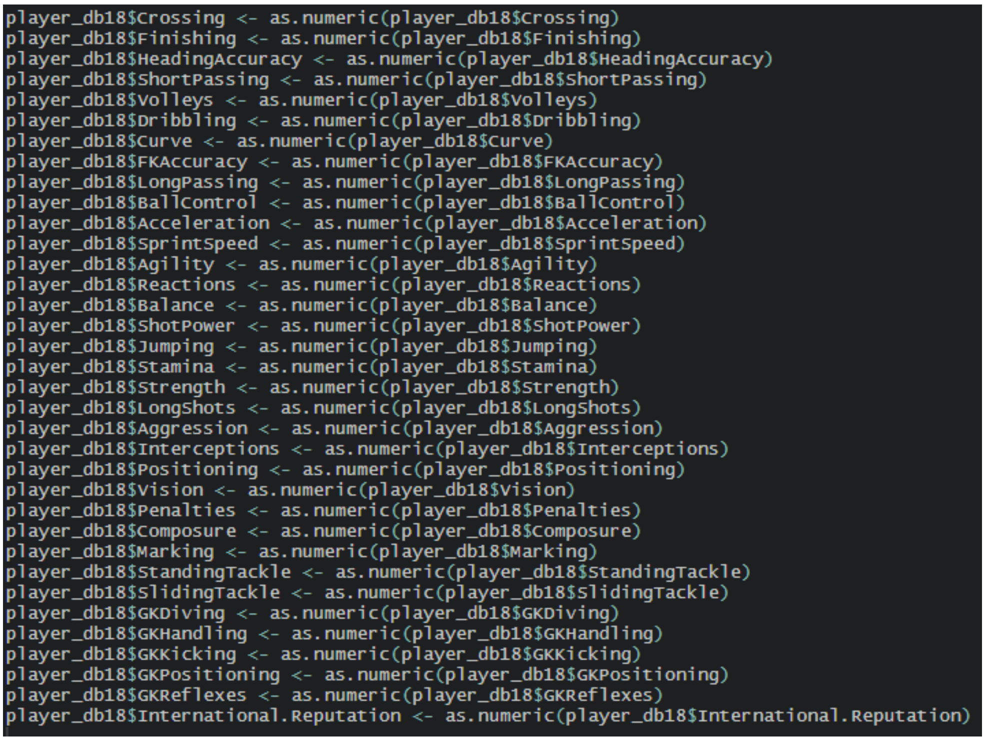
```
## Removing lbs tag from weight factor
head(player_db19$Weight)
player_db19$Wght <- substr(player_db19$Weight, 1, nchar(player_db19$Weight)-3)
head(player_db19$Wght)
player_db19$Weight <- NULL
player_db19$Wght <- as.numeric(player_db19$Wght)
player_db19$Wght <- conv_unit(player_db19$Wght, "lbs", "kg")
player_db19$Wght <- set_units(player_db19$Wght, kg)
sum(is.na(player_db19$Wght))
player_db19$Wght[is.na(player_db19$Wght)] <- 0
```

f. Changing Value, Wages and Release Clause from human form to real values. (Using Excel (LEFT and LEN function))

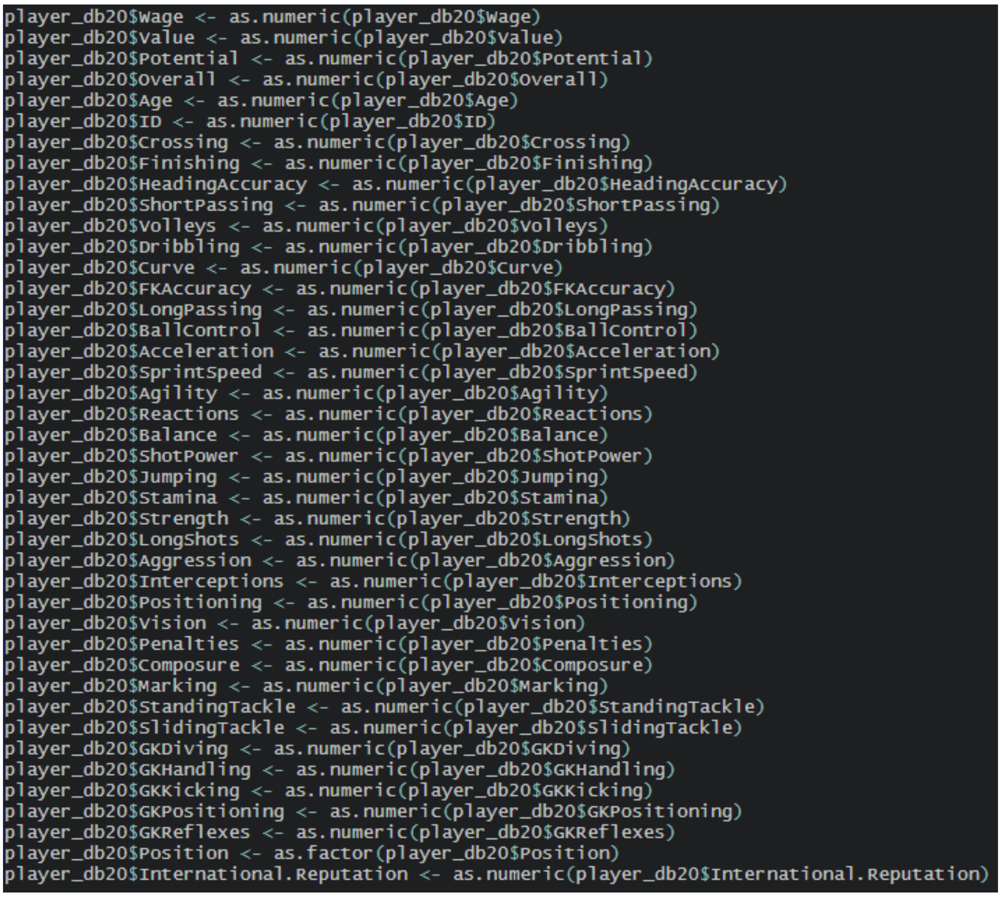
g. Trimming data & changing the data type. (Using R)



```
player_db18$Crossing <- strtrim(player_db18$Crossing, 2)
player_db18$Finishing <- strtrim(player_db18$Finishing, 2)
player_db18$HeadingAccuracy <- strtrim(player_db18$HeadingAccuracy, 2)
player_db18$ShortPassing <- strtrim(player_db18$ShortPassing, 2)
player_db18$Volleys <- strtrim(player_db18$Volleys, 2)
player_db18$Dribbling <- strtrim(player_db18$Dribbling, 2)
player_db18$Curve <- strtrim(player_db18$Curve, 2)
player_db18$FKAccuracy <- strtrim(player_db18$FKAccuracy, 2)
player_db18$LongPassing <- strtrim(player_db18$LongPassing, 2)
player_db18$BallControl <- strtrim(player_db18$BallControl, 2)
player_db18$Acceleration <- strtrim(player_db18$Acceleration, 2)
player_db18$SprintSpeed <- strtrim(player_db18$SprintSpeed, 2)
player_db18$Agility <- strtrim(player_db18$Agility, 2)
player_db18$Reactions <- strtrim(player_db18$Reactions, 2)
player_db18$Balance <- strtrim(player_db18$Balance, 2)
player_db18$ShotPower <- strtrim(player_db18$ShotPower, 2)
player_db18$Jumping <- strtrim(player_db18$Jumping, 2)
player_db18$Stamina <- strtrim(player_db18$Stamina, 2)
player_db18$Strength <- strtrim(player_db18$Strength, 2)
player_db18$LongShots <- strtrim(player_db18$LongShots, 2)
player_db18$Aggression <- strtrim(player_db18$Aggression, 2)
player_db18$Interceptions <- strtrim(player_db18$Interceptions, 2)
player_db18$Positioning <- strtrim(player_db18$Positioning, 2)
player_db18$Vision <- strtrim(player_db18$Vision, 2)
player_db18$Penalties <- strtrim(player_db18$Penalties, 2)
player_db18$Composure <- strtrim(player_db18$Composure, 2)
player_db18$Marking <- strtrim(player_db18$Marking, 2)
player_db18$StandingTackle <- strtrim(player_db18$StandingTackle, 2)
player_db18$SlidingTackle <- strtrim(player_db18$SlidingTackle, 2)
player_db18$GKDiving <- strtrim(player_db18$GKDiving, 2)
player_db18$GKHandling <- strtrim(player_db18$GKHandling, 2)
player_db18$GKKicking <- strtrim(player_db18$GKKicking, 2)
player_db18$GKPositioning <- strtrim(player_db18$GKPositioning, 2)
player_db18$GKReflexes <- strtrim(player_db18$GKReflexes, 2)
```

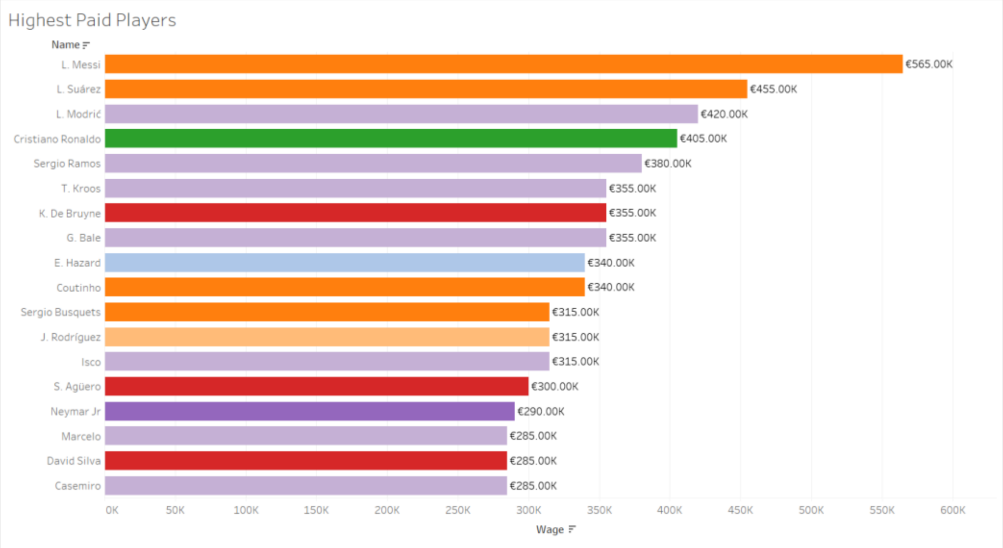


```
player_db18$Crossing <- as.numeric(player_db18$Crossing)
player_db18$Finishing <- as.numeric(player_db18$Finishing)
player_db18$HeadingAccuracy <- as.numeric(player_db18$HeadingAccuracy)
player_db18$ShortPassing <- as.numeric(player_db18$ShortPassing)
player_db18$Volleys <- as.numeric(player_db18$Volleys)
player_db18$Dribbling <- as.numeric(player_db18$Dribbling)
player_db18$Curve <- as.numeric(player_db18$Curve)
player_db18$FKAccuracy <- as.numeric(player_db18$FKAccuracy)
player_db18$LongPassing <- as.numeric(player_db18$LongPassing)
player_db18$BallControl <- as.numeric(player_db18$BallControl)
player_db18$Acceleration <- as.numeric(player_db18$Acceleration)
player_db18$SprintSpeed <- as.numeric(player_db18$SprintSpeed)
player_db18$Agility <- as.numeric(player_db18$Agility)
player_db18$Reactions <- as.numeric(player_db18$Reactions)
player_db18$Balance <- as.numeric(player_db18$Balance)
player_db18$ShotPower <- as.numeric(player_db18$ShotPower)
player_db18$Jumping <- as.numeric(player_db18$Jumping)
player_db18$Stamina <- as.numeric(player_db18$Stamina)
player_db18$Strength <- as.numeric(player_db18$Strength)
player_db18$LongShots <- as.numeric(player_db18$LongShots)
player_db18$Aggression <- as.numeric(player_db18$Aggression)
player_db18$Interceptions <- as.numeric(player_db18$Interceptions)
player_db18$Positioning <- as.numeric(player_db18$Positioning)
player_db18$Vision <- as.numeric(player_db18$Vision)
player_db18$Penalties <- as.numeric(player_db18$Penalties)
player_db18$Composure <- as.numeric(player_db18$Composure)
player_db18$Marking <- as.numeric(player_db18$Marking)
player_db18$StandingTackle <- as.numeric(player_db18$StandingTackle)
player_db18$SlidingTackle <- as.numeric(player_db18$SlidingTackle)
player_db18$GKDiving <- as.numeric(player_db18$GKDiving)
player_db18$GKHandling <- as.numeric(player_db18$GKHandling)
player_db18$GKKicking <- as.numeric(player_db18$GKKicking)
player_db18$GKPositioning <- as.numeric(player_db18$GKPositioning)
player_db18$GKReflexes <- as.numeric(player_db18$GKReflexes)
player_db18$International.Reputation <- as.numeric(player_db18$International.Reputation)
```

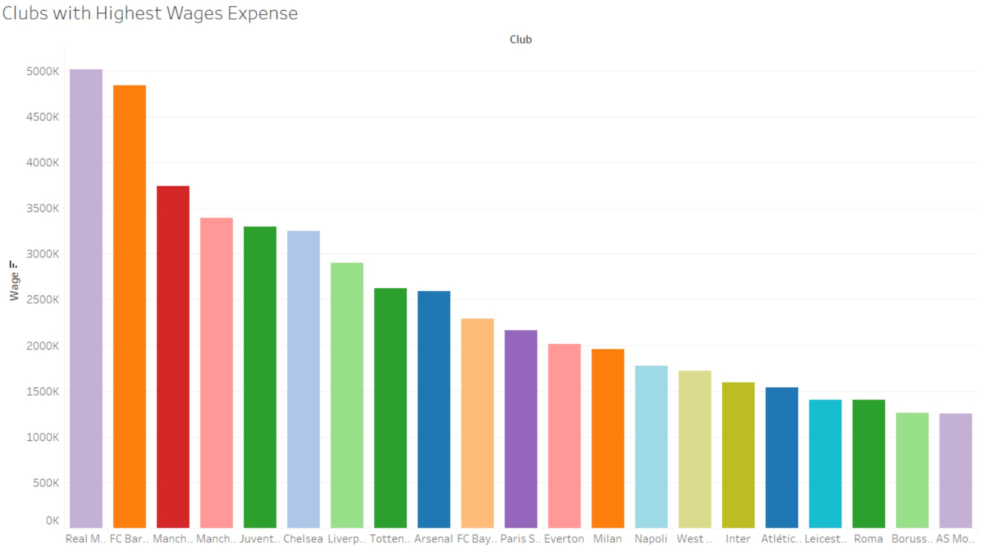


```
player_db20$Wage <- as.numeric(player_db20$Wage)
player_db20$Value <- as.numeric(player_db20$Value)
player_db20$Potential <- as.numeric(player_db20$Potential)
player_db20$Overall <- as.numeric(player_db20$Overall)
player_db20$Age <- as.numeric(player_db20$Age)
player_db20$ID <- as.numeric(player_db20$ID)
player_db20$Crossing <- as.numeric(player_db20$Crossing)
player_db20$Finishing <- as.numeric(player_db20$Finishing)
player_db20$HeadingAccuracy <- as.numeric(player_db20$HeadingAccuracy)
player_db20$ShortPassing <- as.numeric(player_db20$ShortPassing)
player_db20$Volleys <- as.numeric(player_db20$Volleys)
player_db20$Dribbling <- as.numeric(player_db20$Dribbling)
player_db20$Curve <- as.numeric(player_db20$Curve)
player_db20$FKAccuracy <- as.numeric(player_db20$FKAccuracy)
player_db20$LongPassing <- as.numeric(player_db20$LongPassing)
player_db20$BallControl <- as.numeric(player_db20$BallControl)
player_db20$Acceleration <- as.numeric(player_db20$Acceleration)
player_db20$SprintSpeed <- as.numeric(player_db20$SprintSpeed)
player_db20$Agility <- as.numeric(player_db20$Agility)
player_db20$Reactions <- as.numeric(player_db20$Reactions)
player_db20$Balance <- as.numeric(player_db20$Balance)
player_db20$ShotPower <- as.numeric(player_db20$ShotPower)
player_db20$Jumping <- as.numeric(player_db20$Jumping)
player_db20$Stamina <- as.numeric(player_db20$Stamina)
player_db20$Strength <- as.numeric(player_db20$Strength)
player_db20$LongShots <- as.numeric(player_db20$LongShots)
player_db20$Aggression <- as.numeric(player_db20$Aggression)
player_db20$Interceptions <- as.numeric(player_db20$Interceptions)
player_db20$Positioning <- as.numeric(player_db20$Positioning)
player_db20$Vision <- as.numeric(player_db20$Vision)
player_db20$Penalties <- as.numeric(player_db20$Penalties)
player_db20$Composure <- as.numeric(player_db20$Composure)
player_db20$Marking <- as.numeric(player_db20$Marking)
player_db20$StandingTackle <- as.numeric(player_db20$StandingTackle)
player_db20$SlidingTackle <- as.numeric(player_db20$SlidingTackle)
player_db20$GKDiving <- as.numeric(player_db20$GKDiving)
player_db20$GKHandling <- as.numeric(player_db20$GKHandling)
player_db20$GKKicking <- as.numeric(player_db20$GKKicking)
player_db20$GKPositioning <- as.numeric(player_db20$GKPositioning)
player_db20$GKReflexes <- as.numeric(player_db20$GKReflexes)
player_db20$Position <- as.factor(player_db20$Position)
player_db20$International.Reputation <- as.numeric(player_db20$International.Reputation)
```

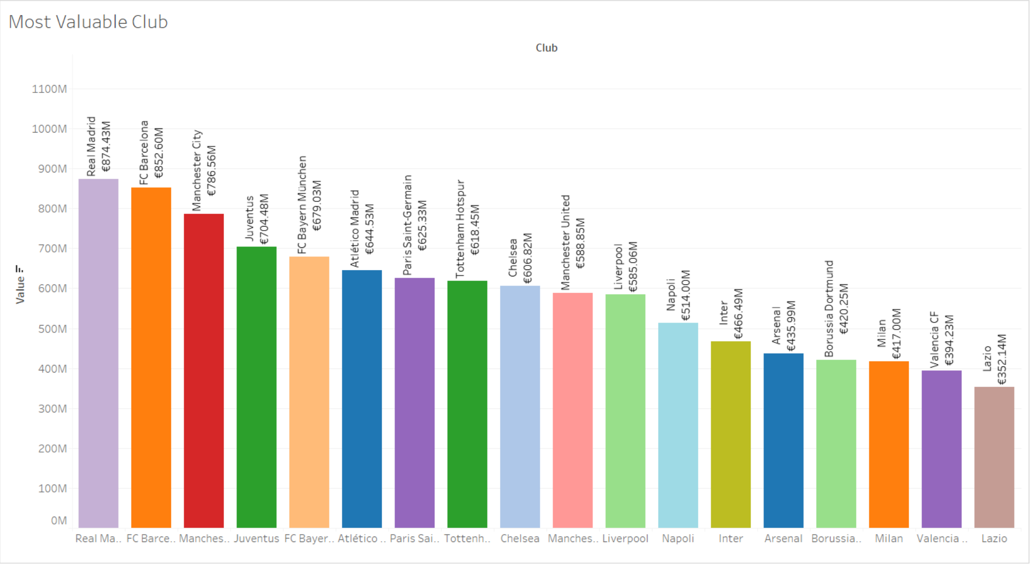
2. Conducting Exploratory Data Analysis, answering various questions to learn the characteristics of the population using cleaned FIFA 19 dataset. (Using Tableau)
    a. Highest Paid Players



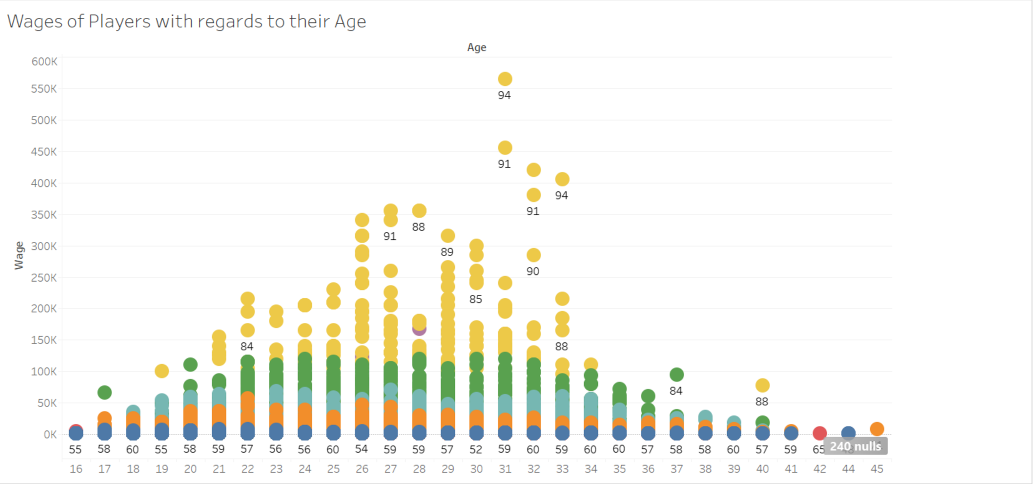
b. Clubs with Highest Wages Expense



c. Most Valuable Club



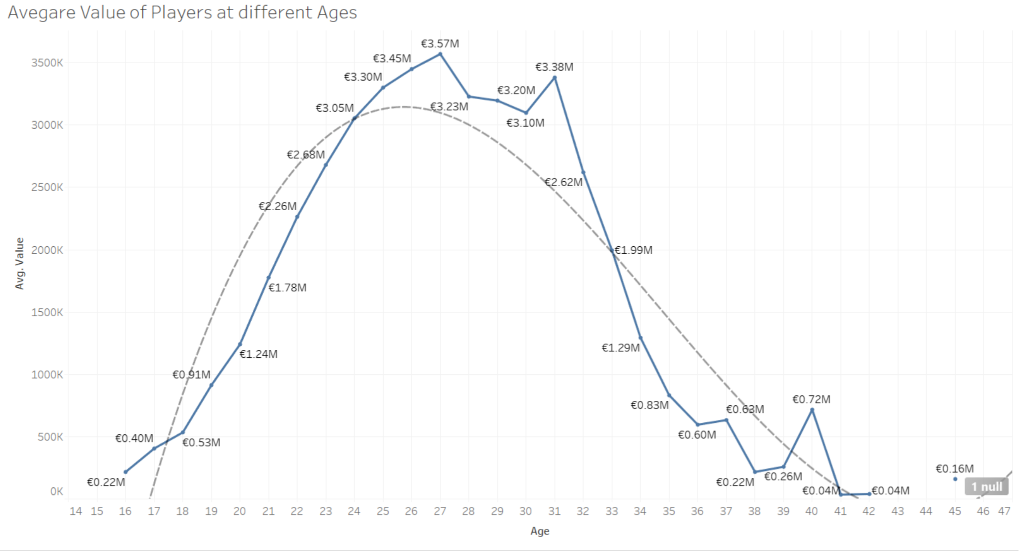
d. Wages of Players with regards to their Age



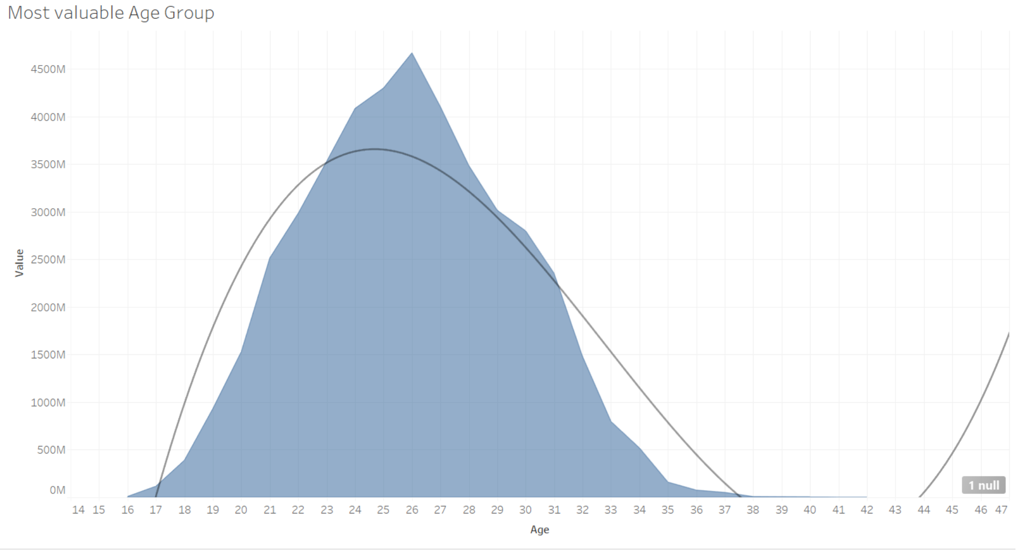
Caption

Sum of Wage for each Age. Color shows Overall as an attribute. The marks are labeled by Overall as an attribute. Details are shown for Player.
From lookin at the plot it is apparent that if we don't take the extraordinary players from each age into consideration the Wages of the players with regards to their Age fits under a Bell Curve. With the players earning peak wages earned between 26 - 29 years of age.

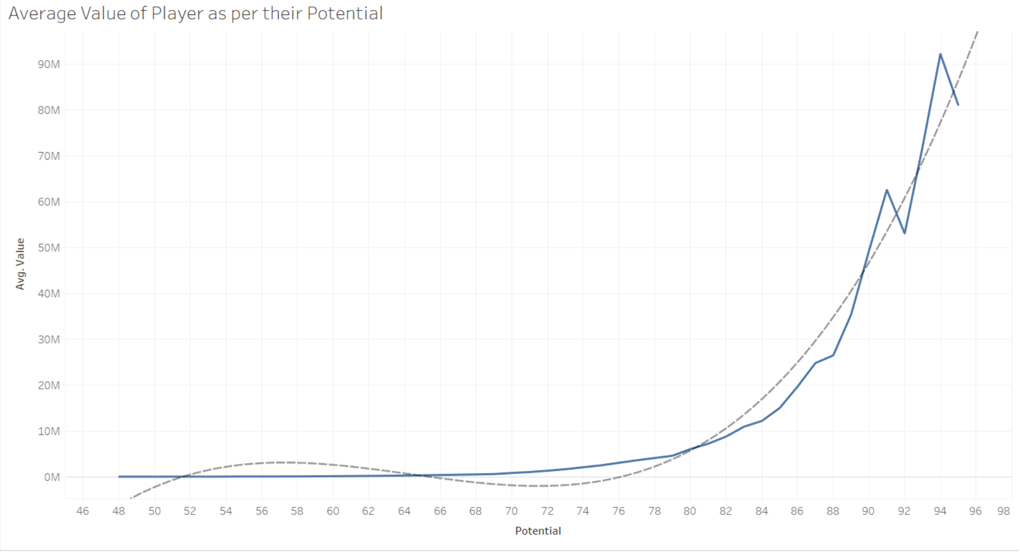
e. Average Value of Players at different Ages



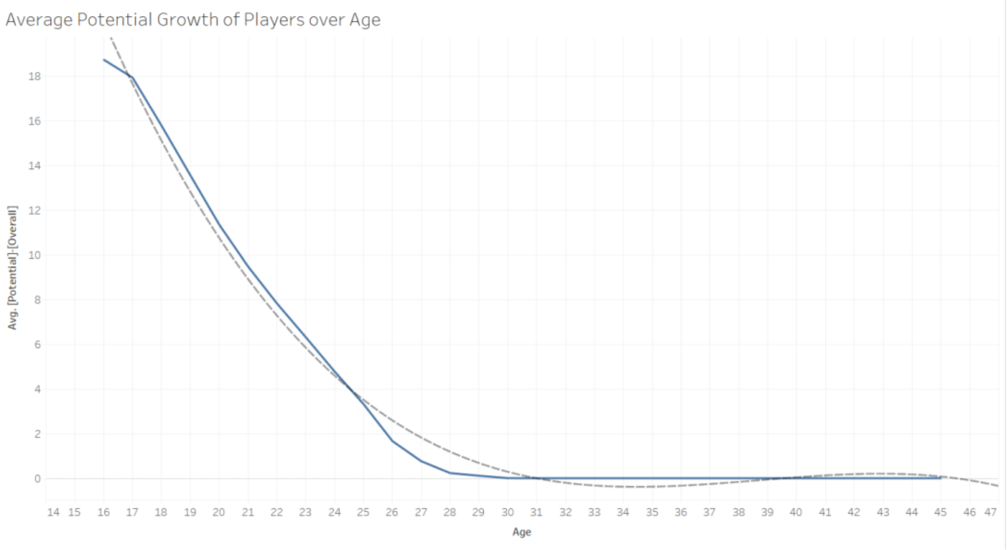
f. Most valuable Age Group



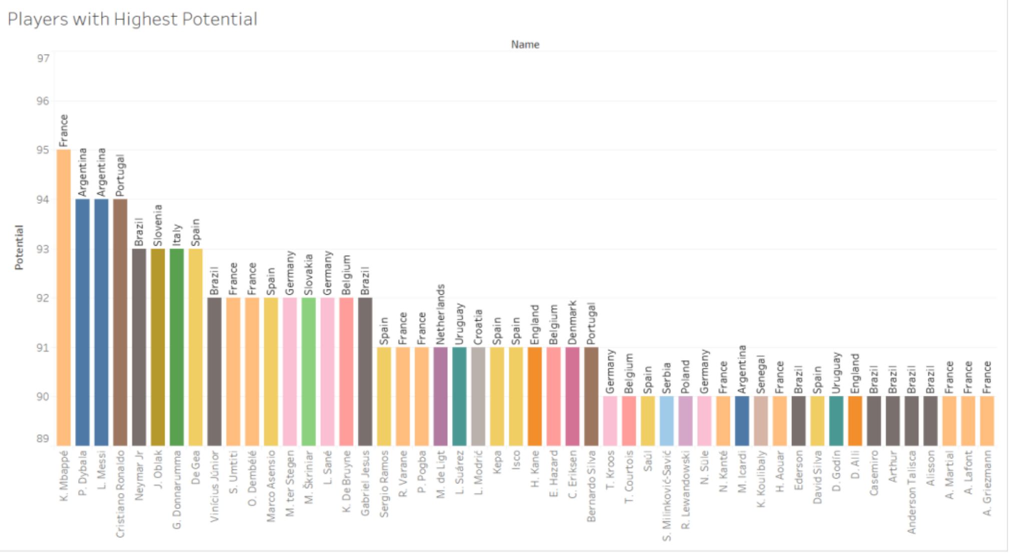
g. Average Value of Player as per their Potential



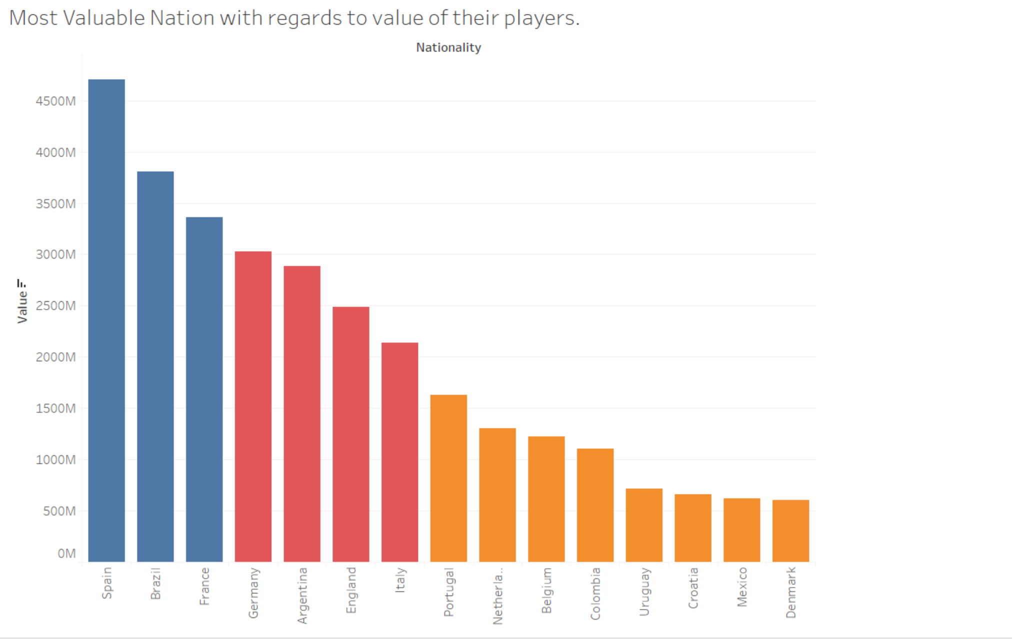
h. Average Potential Growth of Players over Age



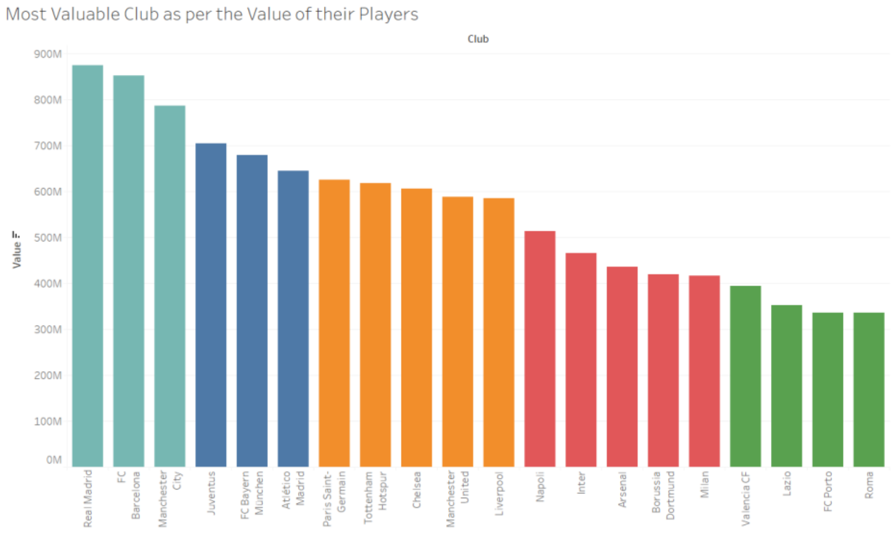
i. Players with Highest Potential



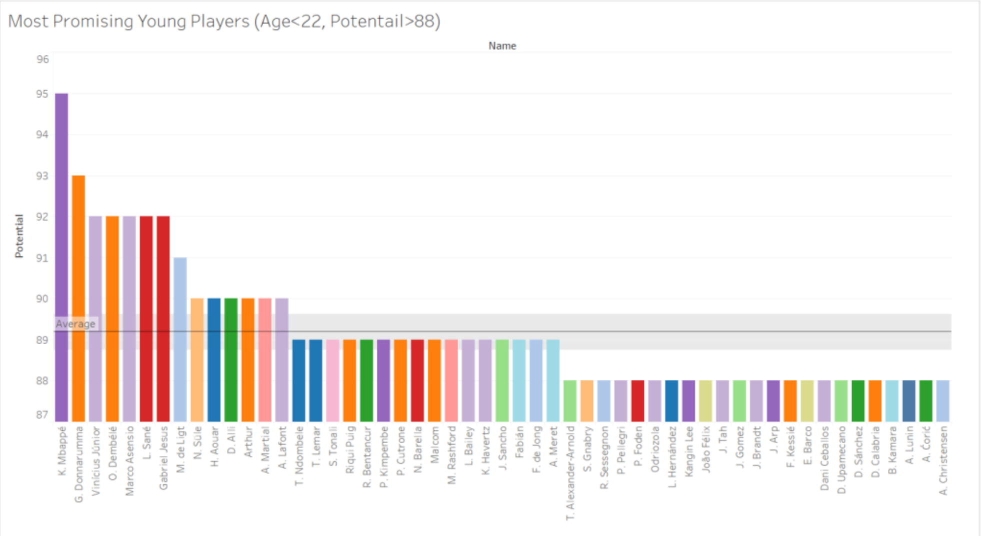
j. Most Valuable Nation with regards to value of their players



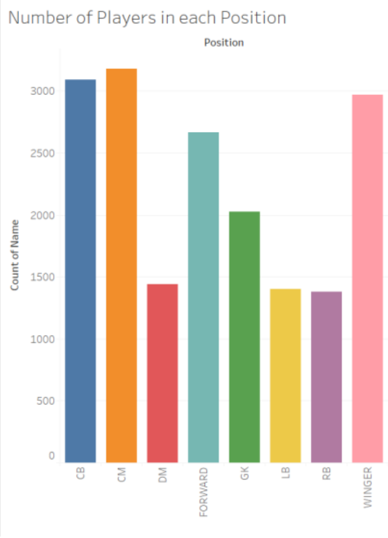
k. Most Valuable Club as per the Value of their Players



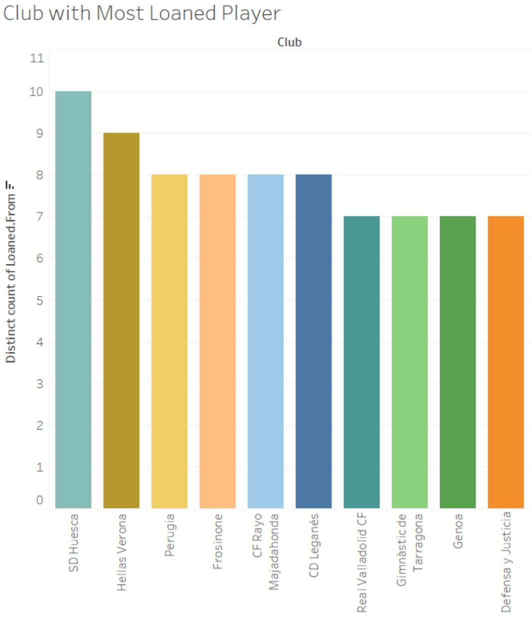
l. Most Promising Young Players (Age<22, Potential>88)



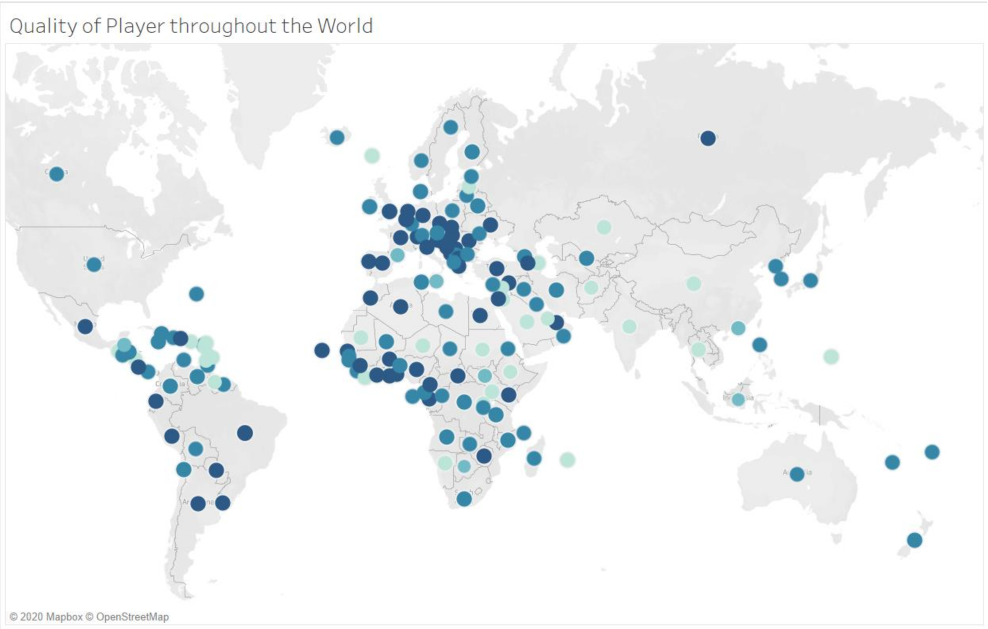
m. Number of Players in each Position



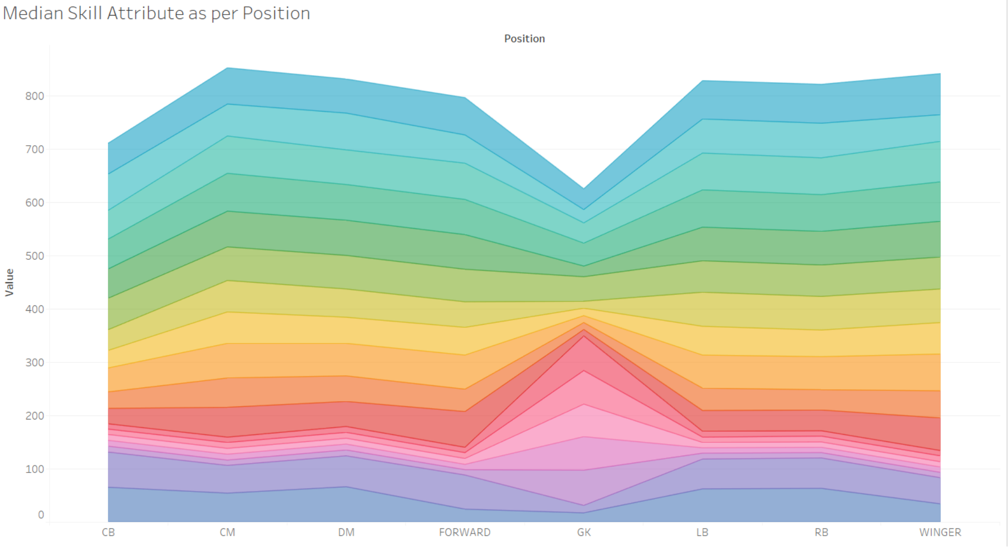
n. Club with Most Loaned Player



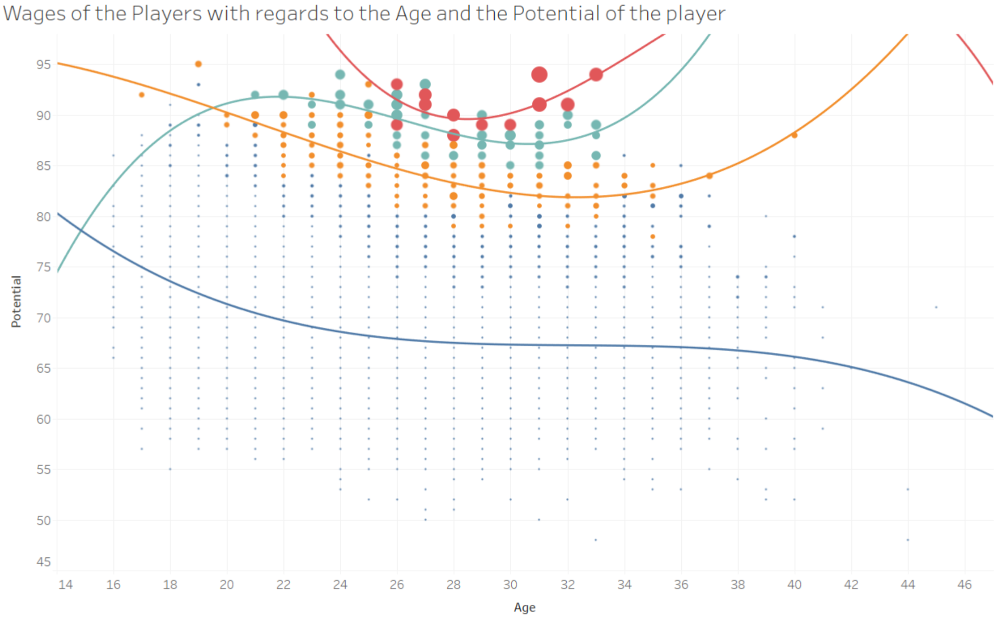
o. Quality of Player throughout the World


Quality of Player throughout the World
© 2020 Mapbox © OpenStreetMap

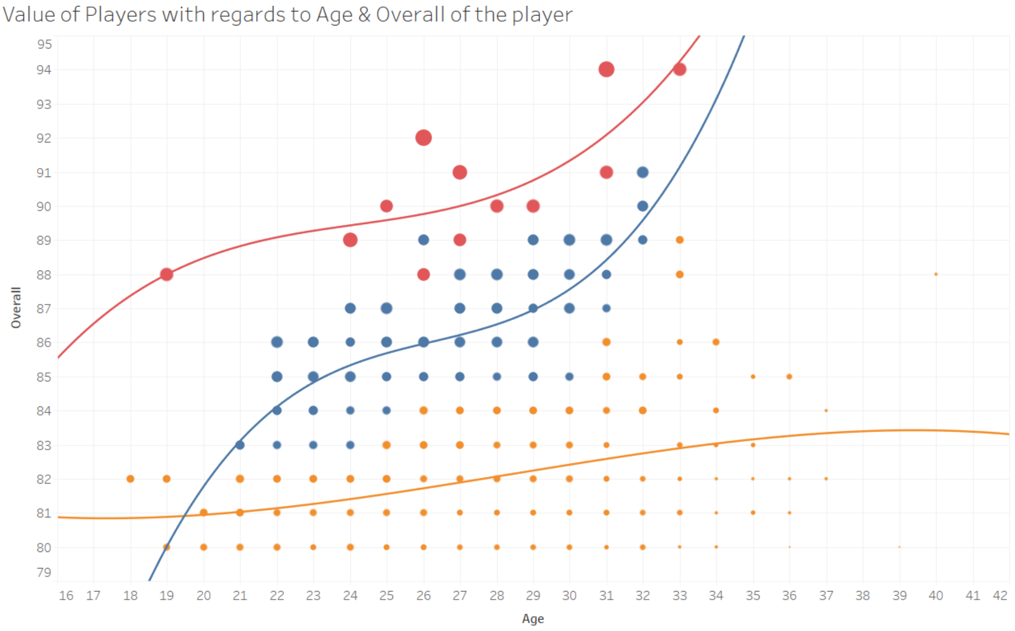
p. Median Skill Attribute as per Position



q. Wages of the Players with regards to the Age and the Potential of the player



r. Value of Players with regards to Age & Overall of the player



3. Creating various ML Model to predict the Overall & Value of the players. (Using R) [Will be done by Monday]
   a. Building models to predict the Overall of the Player
      - Multi-Linear Regression Model
      - Regression Tree Model
      - Random Forrest Model
      - Artificial Neural Network Model
   b. Building models to predict the Value of the Player
      - Multi-Linear Regression Model
      - Regression Tree Model
      - Random Forrest Model
      - Artificial Neural Network Model

4. Creating a Recommendation Model. (Using Python) [Will be done by Friday 5th June 2020]
    a. Using KNN Model to create the Model.

Source:

1. FIFA 18 dataset: https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset
2. FIFA 19 dataset: https://www.kaggle.com/karangadiya/fifa19
3. FIFA 20 dataset: https://www.kaggle.com/stefanoleone992/fifa-20-complete-player- dataset