NBA Game Stats Since the 1996-97 Season

Zohaib Zahid  
Project #2 – Data Exploration and Design  
Due: 2/22/2021

Data

The dataset chosen for this assignment is from Kaggle: <https://www.kaggle.com/kenhuang41/nba-basic-game-data-by-player>

This dataset features individual player data from all games they played in from start of 1996-97 season to December 31, 2020.

Loading in the library

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

Read the file

games <- read.csv(file = 'games.csv')

Check the top of the data:

head(games)

## GAME\_ID TEAM OPPT TEAM\_SCORE OPPT\_SCORE  
## 1 /boxscores/202012310UTA.html Phoenix Suns Utah Jazz 106 95  
## 2 /boxscores/202012310UTA.html Phoenix Suns Utah Jazz 106 95  
## 3 /boxscores/202012310UTA.html Phoenix Suns Utah Jazz 106 95  
## 4 /boxscores/202012310UTA.html Phoenix Suns Utah Jazz 106 95  
## 5 /boxscores/202012310UTA.html Phoenix Suns Utah Jazz 106 95  
## 6 /boxscores/202012310UTA.html Phoenix Suns Utah Jazz 106 95  
## RESULT SCORE\_DIFF PLAYER MP FG FGA FG3 FG3A FT FTA ORB DRB TRB AST STL  
## 1 W 11 Jae Crowder 37 6 11 4 9 0 0 1 5 6 1 2  
## 2 W 11 Mikal Bridges 37 6 9 4 6 0 0 1 5 6 0 2  
## 3 W 11 Devin Booker 35 10 17 1 2 4 7 0 3 3 7 0  
## 4 W 11 Chris Paul 30 4 13 2 4 1 1 1 3 4 8 0  
## 5 W 11 Deandre Ayton 28 2 8 0 0 0 0 2 7 9 3 1  
## 6 W 11 Cameron Payne 18 4 8 2 3 0 0 0 1 1 2 0  
## BLK TOV PF PLUS\_MINUS PTS TOTAL\_MINS STARTER  
## 1 1 1 3 10 16 48 Starter  
## 2 1 0 0 -2 16 48 Starter  
## 3 1 6 3 -2 25 48 Starter  
## 4 0 3 2 12 11 48 Starter  
## 5 2 1 3 -1 4 48 Starter  
## 6 0 1 3 -1 10 48 Reserve

* GAMES\_ID - Unique Game ID (Nominal)
* TEAM - Player’s team (Nominal)
* OPPT - Opponent’s team (Nominal)
* TEAM\_SCORE - Player’s Team Score (Interval)
* OPPY\_SCORE - Oppoenet’s Team Score (Interval)
* RESULT - Win/Loss (Ordinal)
* SCORE\_DIFF - Team Score - Opponent Team (Interval)
* PLAYER - Player Name (Nominal)
* MP - Minutes played (Interval)
* FG - Field Goals Made (Interval)
* FGA - Field Goals Attempted (Interval)
* FG3 - 3 Pointers Made (Interval)
* FG3A - 3 Pointers Attempted (Interval)
* FT - Free Throws Made (Interval)
* FTA - Free Throws Attempted (Interval)
* ORB - Offensive Rebounds (Interval)
* DRB - Defensive Rebounds (Interval)
* TRB - Total Rebounds (Interval)
* AST - Assists (Interval)
* STL - Steals (Interval)
* BLK - Blocks (Interval)
* TOV - Turnovers (Interval)
* PF - Personal Fouls (Interval)
* PLUS\_MINUS - Individual Plus Minus (Interval)
* PTS - Total Points (Interval)
* TOTAL\_MINS - Total Minutes (Interval)
* STARTER - Started or Not (Ordina)

Run a quick analysis on the data provided:

summary(games)

## GAME\_ID TEAM OPPT TEAM\_SCORE   
## Length:743423 Length:743423 Length:743423 Min. : 49.00   
## Class :character Class :character Class :character 1st Qu.: 91.00   
## Mode :character Mode :character Mode :character Median : 99.00   
## Mean : 99.51   
## 3rd Qu.:108.00   
## Max. :168.00   
## OPPT\_SCORE RESULT SCORE\_DIFF PLAYER   
## Min. : 49.0 Length:743423 Min. :-65.00000 Length:743423   
## 1st Qu.: 91.0 Class :character 1st Qu.: -9.00000 Class :character   
## Median : 99.0 Mode :character Median : 1.00000 Mode :character   
## Mean : 99.5 Mean : 0.00897   
## 3rd Qu.:108.0 3rd Qu.: 9.00000   
## Max. :168.0 Max. : 65.00000   
## MP FG FGA FG3   
## Min. : 0.00 Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 7.00 1st Qu.: 0.000 1st Qu.: 1.000 1st Qu.: 0.000   
## Median :21.00 Median : 2.000 Median : 6.000 Median : 0.000   
## Mean :19.73 Mean : 3.022 Mean : 6.688 Mean : 0.566   
## 3rd Qu.:31.00 3rd Qu.: 5.000 3rd Qu.:10.000 3rd Qu.: 1.000   
## Max. :65.00 Max. :28.000 Max. :50.000 Max. :14.000   
## FG3A FT FTA ORB   
## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.0000   
## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.0000   
## Median : 0.000 Median : 0.000 Median : 0.000 Median : 0.0000   
## Mean : 1.593 Mean : 1.498 Mean : 1.985 Mean : 0.9212   
## 3rd Qu.: 3.000 3rd Qu.: 2.000 3rd Qu.: 3.000 3rd Qu.: 1.0000   
## Max. :24.000 Max. :26.000 Max. :39.000 Max. :18.0000   
## DRB TRB AST STL   
## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.0000   
## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.0000   
## Median : 2.000 Median : 3.000 Median : 1.000 Median : 0.0000   
## Mean : 2.535 Mean : 3.457 Mean : 1.781 Mean : 0.6242   
## 3rd Qu.: 4.000 3rd Qu.: 5.000 3rd Qu.: 3.000 3rd Qu.: 1.0000   
## Max. :25.000 Max. :33.000 Max. :25.000 Max. :11.0000   
## BLK TOV PF PLUS\_MINUS   
## Min. : 0.0000 Min. : 0.000 Min. :0.00 Min. :-5.7e+01   
## 1st Qu.: 0.0000 1st Qu.: 0.000 1st Qu.:0.00 1st Qu.:-5.0e+00   
## Median : 0.0000 Median : 1.000 Median :2.00 Median : 0.0e+00   
## Mean : 0.4012 Mean : 1.137 Mean :1.74 Mean :-3.0e-05   
## 3rd Qu.: 1.0000 3rd Qu.: 2.000 3rd Qu.:3.00 3rd Qu.: 5.0e+00   
## Max. :13.0000 Max. :14.000 Max. :7.00 Max. : 5.7e+01   
## PTS TOTAL\_MINS STARTER   
## Min. : 0.000 Min. :48.00 Length:743423   
## 1st Qu.: 0.000 1st Qu.:48.00 Class :character   
## Median : 6.000 Median :48.00 Mode :character   
## Mean : 8.109 Mean :48.36   
## 3rd Qu.:13.000 3rd Qu.:48.00   
## Max. :81.000 Max. :68.00

Notice that this dataset has 743423 entries. This represents every player, in every game played since 1996-97 season, upto December 31, 2020.

When taking a closer look at the score difference, it is pretty symmetrical almost with the mean being zero, because the data also include data of those players on the losing team, therefore the score difference shown for that player is negative.

Lets add a column just for year the game was played, which can be processed as a subset of the game\_ID. Although the year does not exactly correspond to the season, this will still help analyze how over the years players have changed.

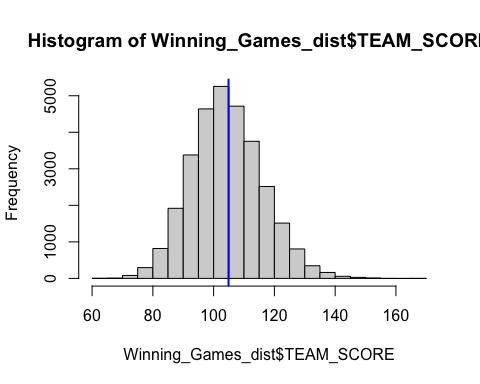
games <- mutate(games, year = substr(GAME\_ID, 12, 15))

Lets take a look at the score distribution of the winning teams:

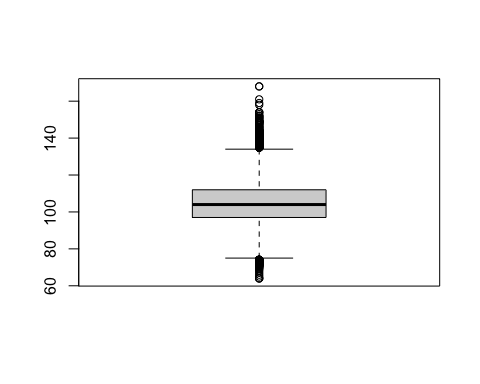
by\_game <- group\_by(games,GAME\_ID)  
by\_winning\_game\_score <- group\_by(by\_game,TEAM\_SCORE, .add=TRUE) %>%  
 filter(RESULT=="W")  
Winning\_Games\_dist <- summarise(by\_winning\_game\_score)

## `summarise()` has grouped output by 'GAME\_ID'. You can override using the `.groups` argument.

hist(Winning\_Games\_dist$TEAM\_SCORE)  
abline(v = mean(Winning\_Games\_dist$TEAM\_SCORE), col = "blue", lwd = 2)



boxplot(Winning\_Games\_dist$TEAM\_SCORE)



summary(Winning\_Games\_dist$TEAM\_SCORE)

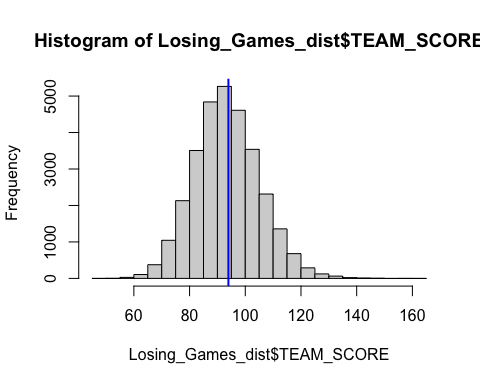
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 64.0 97.0 104.0 104.9 112.0 168.0

Lets take a look at the score distribution of the Losing teams:

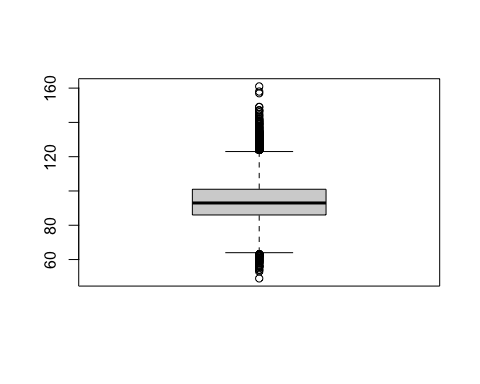
by\_game <- group\_by(games,GAME\_ID)  
by\_losing\_game\_score <- group\_by(by\_game,TEAM\_SCORE, .add=TRUE) %>%  
 filter(RESULT=="L")  
Losing\_Games\_dist <- summarise(by\_losing\_game\_score)

## `summarise()` has grouped output by 'GAME\_ID'. You can override using the `.groups` argument.

hist(Losing\_Games\_dist$TEAM\_SCORE)  
abline(v = mean(Losing\_Games\_dist$TEAM\_SCORE), col = "blue", lwd = 2)



boxplot(Losing\_Games\_dist$TEAM\_SCORE)



summary(Losing\_Games\_dist$TEAM\_SCORE)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 49.00 86.00 93.00 93.94 101.00 161.00

Analysis

Now that there is a better understanding of the data, lets take a look at five analytical questions at hand:

1. Players with the most points scored since 1996-97 season?
2. Players with the most games over 40 points?
3. Who has the best field goal ratios?
4. Player with the best overall stats?
5. Who had the best stats?

Question 1 - Players with the most points scored since 1996-97 season?

by\_player <- group\_by(games,PLAYER)  
  
most\_points\_scored <- summarise(by\_player, points = sum(PTS)) %>%  
 arrange(desc(points)) %>%  
 top\_n(5)

## Selecting by points

most\_points\_scored

## # A tibble: 5 x 2  
## PLAYER points  
## <chr> <int>  
## 1 LeBron James 41849  
## 2 Kobe Bryant 39283  
## 3 Dirk Nowitzki 35223  
## 4 Tim Duncan 31668  
## 5 Paul Pierce 29577

Question 2 - Players with the most games over 40 points?

most\_games\_over\_40pts <- group\_by(games,PLAYER) %>%  
 filter(PTS>40) %>%  
 count(PLAYER) %>%  
 arrange(desc(n)) %>%  
 ungroup(PLAYER) %>%  
 top\_n(5)

## Selecting by n

most\_games\_over\_40pts

## # A tibble: 5 x 2  
## PLAYER n  
## <chr> <int>  
## 1 Kobe Bryant 105  
## 2 James Harden 88  
## 3 LeBron James 79  
## 4 Allen Iverson 72  
## 5 Kevin Durant 48

Now that we have a few of the best players in the NBA, according to total points scored and most games with atleast 40 pts. Lets analyze a few stats from these individuals to try and help narrow down and determine who is the best of these player in the NBA since 1996-97 season.

Question 3 - Who has the best field goal ratios in their career?

fgr <- mutate(games, FGR = FG/FGA, FGR3PT = FG3/FG3A, FGRFT = FT/FTA) %>%  
 filter(PLAYER == most\_games\_over\_40pts$PLAYER | PLAYER == most\_points\_scored$PLAYER)

## Warning in PLAYER == most\_games\_over\_40pts$PLAYER: longer object length is not a  
## multiple of shorter object length

## Warning in PLAYER == most\_points\_scored$PLAYER: longer object length is not a  
## multiple of shorter object length

best\_fgr <- group\_by(fgr,PLAYER) %>%  
 summarise(avgfgr = mean(FGR, na.rm = TRUE)) %>%  
 arrange(desc(avgfgr))  
  
best\_fgr

## # A tibble: 8 x 2  
## PLAYER avgfgr  
## <chr> <dbl>  
## 1 Tim Duncan 0.514  
## 2 LeBron James 0.504  
## 3 Kevin Durant 0.498  
## 4 Dirk Nowitzki 0.472  
## 5 Kobe Bryant 0.450  
## 6 James Harden 0.438  
## 7 Paul Pierce 0.428  
## 8 Allen Iverson 0.405

best\_3ptr <- group\_by(fgr,PLAYER) %>%  
 summarise(avg3ptr = mean(FGR3PT, na.rm = TRUE)) %>%  
 arrange(desc(avg3ptr))  
  
best\_3ptr

## # A tibble: 8 x 2  
## PLAYER avg3ptr  
## <chr> <dbl>  
## 1 Kevin Durant 0.397  
## 2 Dirk Nowitzki 0.379  
## 3 James Harden 0.357  
## 4 Paul Pierce 0.335  
## 5 LeBron James 0.307  
## 6 Kobe Bryant 0.297  
## 7 Allen Iverson 0.275  
## 8 Tim Duncan 0.167

best\_ftr <- group\_by(fgr,PLAYER) %>%  
 summarise(avgftr = mean(FGRFT, na.rm = TRUE)) %>%  
 arrange(desc(avgftr))  
  
best\_ftr

## # A tibble: 8 x 2  
## PLAYER avgftr  
## <chr> <dbl>  
## 1 Kevin Durant 0.895  
## 2 Dirk Nowitzki 0.888  
## 3 James Harden 0.839  
## 4 Kobe Bryant 0.830  
## 5 Paul Pierce 0.801  
## 6 Allen Iverson 0.766  
## 7 LeBron James 0.730  
## 8 Tim Duncan 0.678

merged\_fgr <- merge(best\_fgr,best\_3ptr,by.x="PLAYER") %>%  
 merge(best\_ftr,by.x="PLAYER")  
merged\_fgr

## PLAYER avgfgr avg3ptr avgftr  
## 1 Allen Iverson 0.4049864 0.2749247 0.7655203  
## 2 Dirk Nowitzki 0.4723505 0.3792939 0.8876295  
## 3 James Harden 0.4379330 0.3567053 0.8391360  
## 4 Kevin Durant 0.4982336 0.3971548 0.8951663  
## 5 Kobe Bryant 0.4501163 0.2972883 0.8297172  
## 6 LeBron James 0.5040459 0.3070461 0.7298760  
## 7 Paul Pierce 0.4275667 0.3354740 0.8007743  
## 8 Tim Duncan 0.5138060 0.1666667 0.6781307

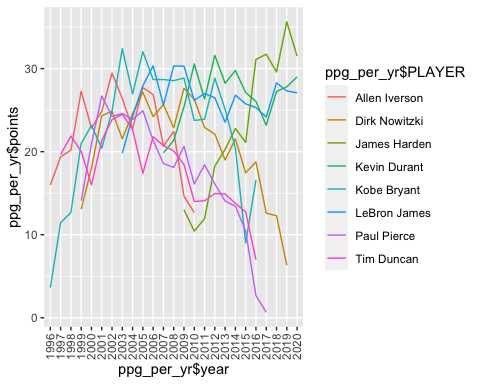
by\_player <- filter(fgr) %>%  
 group\_by(PLAYER, year)  
  
ppg\_per\_yr <- summarise(by\_player, points = mean(PTS, na.rm = TRUE)) %>%  
 arrange(year)

## `summarise()` has grouped output by 'PLAYER'. You can override using the `.groups` argument.

ppg\_per\_yr

## # A tibble: 140 x 3  
## # Groups: PLAYER [8]  
## PLAYER year points  
## <chr> <chr> <dbl>  
## 1 Allen Iverson 1996 16   
## 2 Kobe Bryant 1996 3.64  
## 3 Allen Iverson 1997 19.4   
## 4 Kobe Bryant 1997 11.4   
## 5 Tim Duncan 1997 19.7   
## 6 Allen Iverson 1998 20.2   
## 7 Kobe Bryant 1998 12.7   
## 8 Tim Duncan 1998 21.9   
## 9 Allen Iverson 1999 27.3   
## 10 Dirk Nowitzki 1999 13.1   
## # … with 130 more rows

# plot  
ggplot(data = ppg\_per\_yr, aes(x=ppg\_per\_yr$year, y=ppg\_per\_yr$points, group = ppg\_per\_yr$PLAYER, color =ppg\_per\_yr$PLAYER)) + geom\_line() + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))



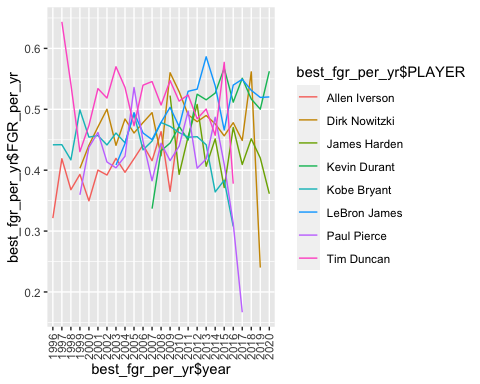
by\_player <- filter(fgr) %>%  
 group\_by(PLAYER, year)  
  
best\_fgr\_per\_yr <- summarise(by\_player, FGR\_per\_yr = mean(FGR, na.rm = TRUE)) %>%  
 arrange(year)

## `summarise()` has grouped output by 'PLAYER'. You can override using the `.groups` argument.

best\_fgr\_per\_yr

## # A tibble: 140 x 3  
## # Groups: PLAYER [8]  
## PLAYER year FGR\_per\_yr  
## <chr> <chr> <dbl>  
## 1 Allen Iverson 1996 0.321  
## 2 Kobe Bryant 1996 0.442  
## 3 Allen Iverson 1997 0.419  
## 4 Kobe Bryant 1997 0.442  
## 5 Tim Duncan 1997 0.644  
## 6 Allen Iverson 1998 0.368  
## 7 Kobe Bryant 1998 0.417  
## 8 Tim Duncan 1998 0.541  
## 9 Allen Iverson 1999 0.393  
## 10 Dirk Nowitzki 1999 0.403  
## # … with 130 more rows

# plot  
ggplot(data = best\_fgr\_per\_yr, aes(x=best\_fgr\_per\_yr$year, y=best\_fgr\_per\_yr$FGR\_per\_yr, group = best\_fgr\_per\_yr$PLAYER, color =best\_fgr\_per\_yr$PLAYER)) + geom\_line() + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))



Question 4 - Player with the best overall stats in their careers?

stats <- filter(fgr)  
  
best\_stats <- group\_by(stats,PLAYER) %>%  
 summarise(avgTRB = mean(TRB, na.rm = TRUE), avgAST = mean(AST, na.rm = TRUE), avgSTL = mean(STL, na.rm = TRUE), avgBLK = mean(BLK, na.rm = TRUE)) %>%  
 mutate(total = avgTRB+avgAST+avgSTL+avgBLK) %>%  
 arrange(desc(total))  
  
best\_stats

## # A tibble: 8 x 6  
## PLAYER avgTRB avgAST avgSTL avgBLK total  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 LeBron James 7.56 7.24 1.63 0.781 17.2   
## 2 Tim Duncan 10.4 2.85 0.622 2.12 16.0   
## 3 Kevin Durant 6.98 4.20 1.13 1.16 13.5   
## 4 James Harden 4.89 5.98 1.54 0.583 13.0   
## 5 Dirk Nowitzki 8.01 2.37 0.826 0.780 12.0   
## 6 Allen Iverson 3.19 5.91 2.12 0.190 11.4   
## 7 Kobe Bryant 4.96 4.42 1.33 0.480 11.2   
## 8 Paul Pierce 5.16 3.17 1.11 0.522 9.96

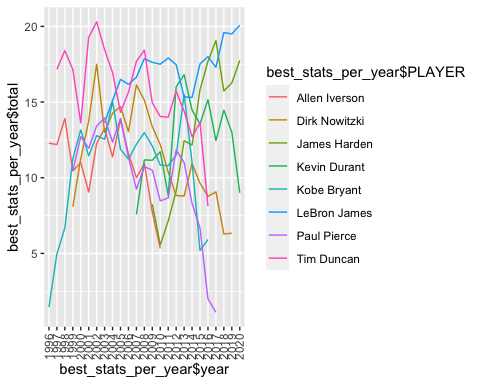
by\_player <- filter(fgr) %>%  
 group\_by(PLAYER, year)  
  
best\_stats\_per\_year <- summarise(by\_player, avgTRB = mean(TRB, na.rm = TRUE), avgAST = mean(AST, na.rm = TRUE), avgSTL = mean(STL, na.rm = TRUE), avgBLK = mean(BLK, na.rm = TRUE)) %>%  
 mutate(total = avgTRB+avgAST+avgSTL+avgBLK) %>%  
 arrange(desc(total))

## `summarise()` has grouped output by 'PLAYER'. You can override using the `.groups` argument.

best\_stats\_per\_year

## # A tibble: 140 x 7  
## # Groups: PLAYER [8]  
## PLAYER year avgTRB avgAST avgSTL avgBLK total  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Tim Duncan 2002 12.6 3.8 0.5 3.4 20.3  
## 2 LeBron James 2020 8.69 9.81 0.885 0.692 20.1  
## 3 LeBron James 2018 9.08 8.26 1.44 0.821 19.6  
## 4 LeBron James 2019 7.64 10.1 1.29 0.5 19.5  
## 5 Tim Duncan 2001 11.7 4 0.643 2.93 19.3  
## 6 James Harden 2017 7.47 9.27 1.53 0.8 19.1  
## 7 Tim Duncan 2003 12.3 3.38 0.417 2.38 18.5  
## 8 Tim Duncan 2008 12.9 3.19 0.571 1.76 18.4  
## 9 Tim Duncan 1998 12.9 2.87 0.933 1.67 18.4  
## 10 LeBron James 2016 8.05 7.58 1.71 0.658 18   
## # … with 130 more rows

ggplot(data = best\_stats\_per\_year, aes(x=best\_stats\_per\_year$year, y=best\_stats\_per\_year$total, group = best\_stats\_per\_year$PLAYER, color =best\_stats\_per\_year$PLAYER)) + geom\_line() + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))



Question 5 - Which team did LeBron play the best on?

by\_player <- filter(games, PLAYER=="LeBron James") %>%  
 group\_by(TEAM,year)  
  
Lebron\_most\_points\_per\_year <- summarise(by\_player, points = sum(PTS)) %>%  
 arrange(year)

## `summarise()` has grouped output by 'TEAM'. You can override using the `.groups` argument.

Lebron\_most\_points\_per\_year

## # A tibble: 21 x 3  
## # Groups: TEAM [3]  
## TEAM year points  
## <chr> <chr> <int>  
## 1 Cleveland Cavaliers 2003 647  
## 2 Cleveland Cavaliers 2004 1715  
## 3 Cleveland Cavaliers 2005 2317  
## 4 Cleveland Cavaliers 2006 2821  
## 5 Cleveland Cavaliers 2007 2589  
## 6 Cleveland Cavaliers 2008 2730  
## 7 Cleveland Cavaliers 2009 2917  
## 8 Cleveland Cavaliers 2010 1596  
## 9 Miami Heat 2010 830  
## 10 Miami Heat 2011 1910  
## # … with 11 more rows

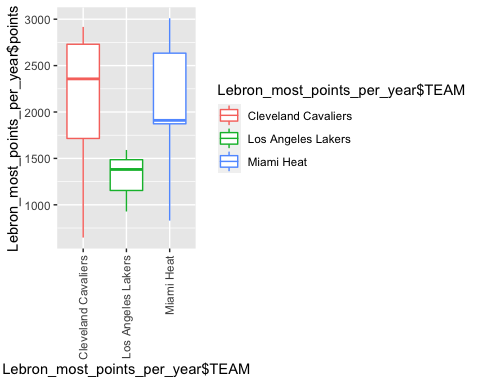
Lebron\_avg\_points\_per\_year <-summarise(by\_player, points = mean(PTS)) %>%  
 arrange(year)

## `summarise()` has grouped output by 'TEAM'. You can override using the `.groups` argument.

Lebron\_avg\_points\_per\_year

## # A tibble: 21 x 3  
## # Groups: TEAM [3]  
## TEAM year points  
## <chr> <chr> <dbl>  
## 1 Cleveland Cavaliers 2003 20.2  
## 2 Cleveland Cavaliers 2004 22.6  
## 3 Cleveland Cavaliers 2005 29.0  
## 4 Cleveland Cavaliers 2006 30.3  
## 5 Cleveland Cavaliers 2007 25.9  
## 6 Cleveland Cavaliers 2008 29.4  
## 7 Cleveland Cavaliers 2009 29.5  
## 8 Cleveland Cavaliers 2010 29.6  
## 9 Miami Heat 2010 24.4  
## 10 Miami Heat 2011 26.5  
## # … with 11 more rows

# plot  
ggplot(data = Lebron\_most\_points\_per\_year, aes(x=Lebron\_most\_points\_per\_year$TEAM, y=Lebron\_most\_points\_per\_year$points, group = Lebron\_most\_points\_per\_year$TEAM, color=Lebron\_most\_points\_per\_year$TEAM)) + geom\_boxplot() + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))



# plot  
ggplot(data = Lebron\_avg\_points\_per\_year, aes(x=Lebron\_avg\_points\_per\_year$year, y=Lebron\_avg\_points\_per\_year$points, group = Lebron\_avg\_points\_per\_year$TEAM, color=Lebron\_avg\_points\_per\_year$TEAM)) + geom\_point() + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

