# COGS 108 - Final Project

# Overview

In this notebook, we explore the shooting trends of the NBA over the last 40 years. We also dive further into the shooting trends by looking into the shots taken by each position. Furthermore, we conclude that there is a positive correlation with the three point shots and along with a corresponding negative correlation with two-point shots.

# Names


- David Palafox
- Kwok Ming Leung
- Sahil Kalra
- Jessey Villafan


# Group Members IDs


- A14752623
- A15753757
- A14066280
- A14174074


# Research Question

We aim to answer the question as to whether or not the NBA has shitfted from a more physical style of play involving to a more shooting focused league.

## Background and Prior Work

Before 1980, The National Basketball Association (NBA) only had two ways of scoring points: The 2-pointer and the Free-throw. However, a few years after the American Basketball Association (ABA) merged with the NBA, they soon after began to incorporate the 3-point shot.

A lot of teams and coaches were skeptical at first, but to this present day we believe that the 3-point shot has been progressively overcoming the 2-point shot. Previous work has shown that five years after the shot was introduced, the average three point shots taken per game was 2.4. In the present seasons, star players like James Harden shoot up to 10 per game alone. Up to the point where we believe that soon enough, it will be diminished so much that it will be eventually eradicated from the game of basketball.

As a team, we have considered the timing of this project, and given that it's in the Spring what better way to enjoy the tail end of the 2018-2019 NBA season than to look at some Basketball stats. We would like to explore trends in the shooting patterns for each player position. These explorations are important because it would be important for scouts to understand how the game is changing overall, in order to make the best player choice when looking for future players. Subsequently, we would like to see if we can make any predictions about the future of the National Basketball Association. References (include links):

• 1) https://shottracker.com/articles/the-3-point-revolution (https://shottracker.com/articles/the-3-point-revolution) • 2) https://improvehoops.com/3-point-line-distance/ (https://improvehoops.com/3-point-line-distance/)

# Hypothesis

With the rapid growth of the amount of three-point shots taken over the entire NBA, especially over the last few years, we believe that there will be a change in the game as time progresses. Furthermore, we believe that due to the increasing trends of the 3-point shot, we predict that there will be a change in the rules of the game to make it more difficult, and even a change in the type of skillsets scouts use when looking for upcoming players.

# Dataset(s)

Dataset #1

- Dataset Name: NBA Players stats since 1950
- Link to the dataset: https://www.kaggle.com/drgilermo/nba-players-stats#Seasons_Stats.csv (https://www.kaggle.com/drgilermo/nba-players-stats#Seasons_Stats.csv)
- Number of observations: 24.7k

This Data set contains the Season Statistics for each player ranging from 1950 to 2017.

Dataset #2

- Dataset Name: NBA Team Game Stats from 2014 to 2018
- Link to the dataset: https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018 (https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018)
- Number of observations: 9841

This dataset contains the Team Statistics for each game throughout 2014 through 2018.

We used the first dataset to showcase the overall trend of the NBA from 1950. We used the second dataset to see more detailed statistics from recent years. We did not combine these datasets, as we could transform the first dataset to give us the same statistics as the second one. However, due to convenience, we used the second dataset when necessary.

# Setup

In [267]:

```
# The different python libraries we may use
import numpy as np
import pandas as pd
import requests
import bs4
import matplotlib.pyplot as plt
import seaborn as sns
from numpy import loadtxt
from operator import itemgetter
from collections import defaultdict
import random
import numpy.lib.recfunctions
from collections import deque
import copy
```

# Data Cleaning

In [268]:

```
# Holds Game Statistics from 2014-2018 season
df = pd.read_csv('datasets/nba.games.stats.csv')

# Holds Player Statistics from 1950 - 2017
df_2 = pd.read_csv('datasets/nba-players-stats/Seasons_Stats.csv')
```

## Reducing Amount of Data

Because we're dealing with a lot of data, we only take data starting from 1980 onward. The majority of data before this date is incomplete and/or missing, thus does not offer us any tangible value with regards to our data science needs. Additionally, the 3 point shot was not introduced until 1979, so any yearly records before that would not include any 3 point statistic, which is the main focus of our research quesiton. We also dropped any columns that is totally incomplete, such as the 'blank' column.

In [269]:

```
df_2 = df_2[ df_2['Year'] >= 1980]
```

## Anonymizing Data

Due to the nature of our question and the format of one of the datasets, particularly its player specific stastics, we found it necessary to remove any and all statistics that can be traced back to the player. This includes their name, number of games they played, and invidual stastics such as Field Goal Attempt, Turnovers, and even what Team they played for during a given year.
Furthermore, we accumulate the data for the league as a whole by year, thus removing any possibility of tracing data back to the any individual player, resuling in consolidated scoring tendencies and stastics for the league as a whole throughout each year.

In [270]:

```
columns_drop =  ['Player', 'G', 'GS', 'MP', 'PER', 'TS%', 'FTr',
                 'ORB%', 'DRB%', 'TRB%', 'AST%', 'STL%', 'BLK%', 'USG%', 'blanl',
                 'OWS', 'DWS', 'WS', 'WS/48', 'blank2', 'OBPM', 'DBPM', 'VORP',
                 'FT', 'FTA', 'ORB', 'DRB', 'TRB', 'AST', 'STL', 'TOV', 'PF',
                 '3P%', '2P%', 'eFG%', 'FT%', 'Age',  'FG%', 'BPM', '3PAr', 'Tm',
'TOV%',
                 'BLK', df_2.columns[0] ]
df_2 = df_2.drop( columns=columns_drop )
df_2 = df_2.dropna( axis = 1, how='all')

drop = [c for c in df.columns if 'Opp' in c]
drop+=['Blocks', 'TotalFouls', 'Home', 'OffRebounds', 'TotalRebounds',  df.colum
ns[0] ]
df = df.drop( columns = drop )
```

# Dividing Data Frames

Although we did want to focus on the goal scoring trends as the league as a whole, we also found it prudent to take each position and account for their goal scoring tendencies across years. We accumulated their respective data for each year, getting more specific data for each position, while maintaining anonymity for individual players.

In [271]:

```
# The columns we want to accumulate for each position
acc_cols = ['FG', 'FGA', '3P', '3PA', '2P', '2PA', 'PTS' ]

# Creating the dataframe that will hold League stastistics for each year
df_yearSum = pd.DataFrame( columns = ['Year'] + acc_cols  )
i = 0

# Populating the dataframe
for year in range( 1980, 2018 ):
    dfx = df_2[ df_2['Year'] == year ]
    df_yearSum.loc[i] = [year, dfx[acc_cols[0]].sum(), dfx[acc_cols[1]].sum(),df
x[acc_cols[2]].sum(),
                         dfx[acc_cols[3]].sum(),dfx[acc_cols[4]].sum(),dfx[acc_co
ls[5]].sum(),
                         dfx[acc_cols[6]].sum()]
    i+=1
df_yearSum


# Creating invididual data frames for each major position on a team
df_C = pd.DataFrame( columns = ['Pos', 'Year'] + acc_cols )
df_PF = pd.DataFrame( columns = df_C.columns )
df_PG = pd.DataFrame( columns = df_C.columns )
df_SG = pd.DataFrame( columns = df_C.columns )
df_SF =pd.DataFrame( columns = df_C.columns )



# Populating each dataframe with their respective accumulated stastic for each y
ear
for year in range( 1980, 2018 ):
    dfc = df_2[ (df_2['Pos'] == 'C') & (df_2['Year'] == year )]
    dfpf = df_2[ (df_2['Pos'] == 'PF')& (df_2['Year'] == year )]
    dfsg = df_2[ (df_2['Pos'] == 'SG')& (df_2['Year'] == year )]
    dfpg = df_2[ (df_2['Pos'] == 'PG')& (df_2['Year'] == year )]
    dfsf = df_2[ (df_2['Pos'] == 'SF')& (df_2['Year'] == year )]

    df_C.loc[i] = ['C', year,dfc[acc_cols[0]].sum(), dfc[acc_cols[1]].sum(), dfc
[acc_cols[2]].sum(),
                                dfc[acc_cols[3]].sum(), dfc[acc_cols[4]].sum(),
dfc[acc_cols[5]].sum(),
                                dfc[acc_cols[6]].sum()]
```

```
    df_PF.loc[i] = ['PF', year,dfpf[acc_cols[0]].sum(), dfpf[acc_cols[1]].sum(),
dfpf[acc_cols[2]].sum(),
                                dfpf[acc_cols[3]].sum(), dfpf[acc_cols[4]].sum()
, dfpf[acc_cols[5]].sum(),
                                dfpf[acc_cols[6]].sum()]

    df_PG.loc[i] = ['PG', year,dfpg[acc_cols[0]].sum(), dfpg[acc_cols[1]].sum(),
dfpg[acc_cols[2]].sum(),
                                dfpg[acc_cols[3]].sum(), dfpg[acc_cols[4]].sum()
, dfpg[acc_cols[5]].sum(),
                                dfpg[acc_cols[6]].sum()]

    df_SG.loc[i] = ['SG', year,dfsg[acc_cols[0]].sum(), dfsg[acc_cols[1]].sum(),
dfsg[acc_cols[2]].sum(),
                                dfsg[acc_cols[3]].sum(), dfsg[acc_cols[4]].sum()
, dfsg[acc_cols[5]].sum(),
                                dfsg[acc_cols[6]].sum()]

    df_SF.loc[i] = ['SF', year,dfsf[acc_cols[0]].sum(), dfsf[acc_cols[1]].sum(),
dfsf[acc_cols[2]].sum(),
                                dfsf[acc_cols[3]].sum(), dfsf[acc_cols[4]].sum()
, dfsf[acc_cols[5]].sum(),
                                dfsf[acc_cols[6]].sum()]

    i+=1


# Creating another dataframe that holds the
df_sumPos = pd.concat( [df_C, df_PF, df_PG, df_SG, df_SF] )
```

## Simplifying Data

Our second DataFrame holds statistics for each game from 2014-2018. Because of this, each game hold the date it was played in. However, it doesn't become quickly obvious what season it was played in. In order to fix this, we write a method that would take a date and output what season it was played in. We created a new column called 'Season' to hold this new feature.

In [272]:

```
def inRange( start_date, end_date, year, month, day ):
    y_start = start_date[0]
    m_start = start_date[1]
    d_start = start_date[2]
    m_end = end_date[1]
```

```
    d_end = end_date[2]

    y_end = end_date[0]
    if( year < y_end ):
        if( year > y_start):
            return True
        elif( year == y_start ):
            if( month > m_start ):
                return True
            elif( month == m_start ):
                if( day >= d_start ):
                    return True
                else:
                    return False
            else:
                return False
        else:
            return False
    elif( year == y_end ):
        if( month < m_end ):
            return True
        elif( month == m_end ):
            if( day <= d_end ):
                return True
            else:
                return False
        else:
            return False
    else:
        return False


def getSeason( date_str ):

    date = date_str.split('-')
    date[0] = int(date[0])
    date[1] = int(date[1])
    date[2] = int( date[2])
    year = date[0]
    month = date[1]
    day = date[2]
    s1_start=[2014,10,28]
    s1_end = [2015,4,15]
    s2_start=[2015,10,27]
    s2_end=[2016,4,13]
    s3_start=[2016,10,25]
    s3_end =[2017,4,12]
    s4_start = [2017,10,17]
    s4_end = [2018,4,11]

    if( inRange( s1_start, s1_end, year, month, day ) ):
        return 2014
    if( inRange(s2_start, s2_end, year, month, day ) ):
        return 2015
```

```
    if( inRange( s3_start, s3_end, year, month, day ) ):

        return 2016
    if( inRange( s4_start, s4_end, year, month, day ) ):
        return 2017

    return -1



def scatPlot( stat_str, df1, df2 ):


    colors = {'W':'g', 'L':'r'}
    fig, axes = plt.subplots( nrows=1, ncols=2, sharey=True )
    fig.set_size_inches(10,3)

    axes[0].set_xlabel('Games')
    axes[1].set_xlabel('Games')
    axes[0].set_ylabel( stat_str )
    axes[1].set_ylabel( stat_str )


    axes[0].set_title(df1.iloc[0].Team)
    axes[1].set_title(df2.iloc[0].Team )

    for i in range( len( df1['X3PointShots'])):
        axes[0].scatter(df1['Game'].iloc[i], df1[stat_str].iloc[i], color=colors
[df1['WINorLOSS'].iloc[i]] )
    #axes.scatter()

    for i in range( len( df2['X3PointShots'])):
        axes[1].scatter(df2['Game'].iloc[i], df2[stat_str].iloc[i], color=colors
[df2['WINorLOSS'].iloc[i]] )
```

In [273]:

```
eastern_conf= ['ATL', 'BKN', 'BOS', 'CHA', 'CHI', 'CLE', 'DET','IND','MIA', 'MIL
','NYK','ORL', 'PHI','TOR','WAS']
western_conf = ['DAL', 'DEN','GSW', 'HOU','LAC', 'LAL', 'MEM','MIN','NOP','OKC',
'PHX', 'POR', 'SAC', 'SAS','UTA']

# Creating the Season Column and applying the getSeason method
df['Season'] = df['Date']
df['Season'] = df['Season'].apply(getSeason)
df.columns
```

Out[273]:

```
Index(['Team', 'Game', 'Date', 'WINorLOSS', 'TeamPoints', 'FieldGoal
s',
       'FieldGoalsAttempted', 'FieldGoals.', 'X3PointShots',
       'X3PointShotsAttempted', 'X3PointShots.', 'FreeThrows',
       'FreeThrowsAttempted', 'FreeThrows.', 'Assists', 'Steals', 'T
urnovers',
       'Season'],
      dtype='object')
```

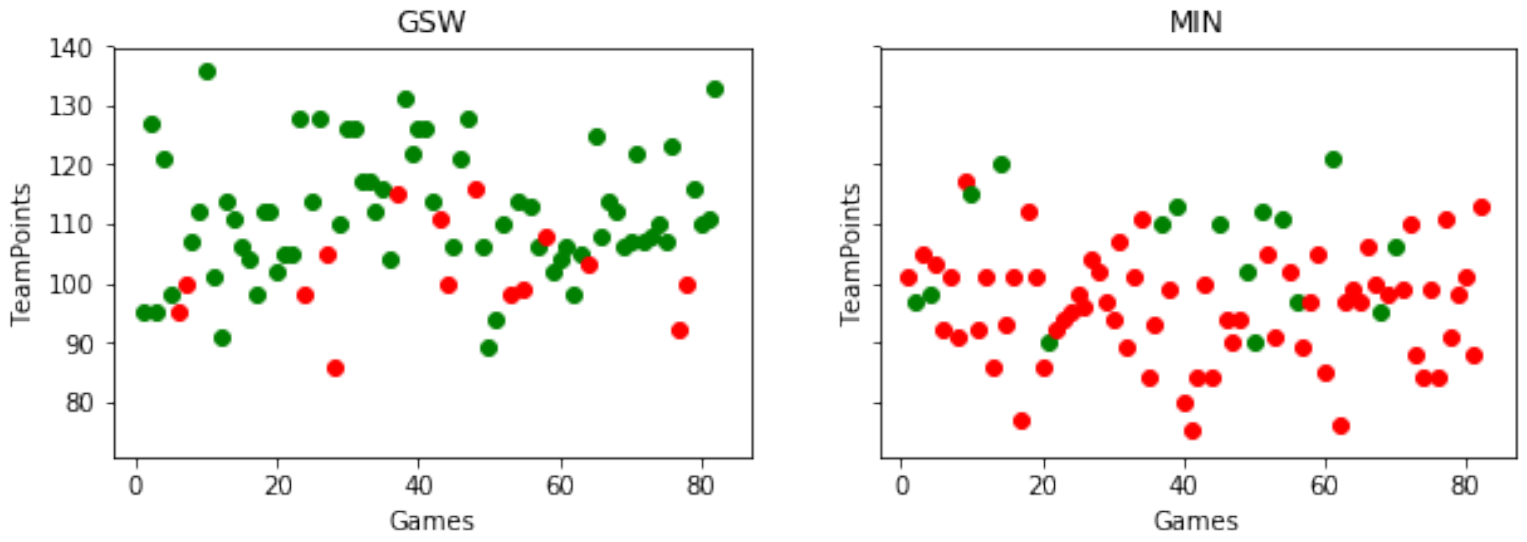
# Data Analysis & Results

Let us take a look at the first seed and last seed teams and see how they differed with regards to scoring tendencies. We can see that the first seed scores more points on average than the last seed. This isn't exactly groundbreaking. But we wondered why? Our first inclination was that maybe the last seed just wasn't shooting as much. So we explored that next.

In [274]:

```
# The first seed in the league in 2014 were the Golden State Warriors
firstSeed_2014 = df[ (df['Season'] == 2014) & (df['Team'] == 'GSW')]
# The last seed in the league in 2014 were the Minnesoata Timberwolves
lastSeed_2014 = df[ (df['Season'] == 2014) & (df['Team'] == 'MIN')]
print( "Green = Win ")
print( "Red = Loss")
scatPlot( 'TeamPoints', firstSeed_2014, lastSeed_2014 )
```

```
Green = Win
Red = Loss
```

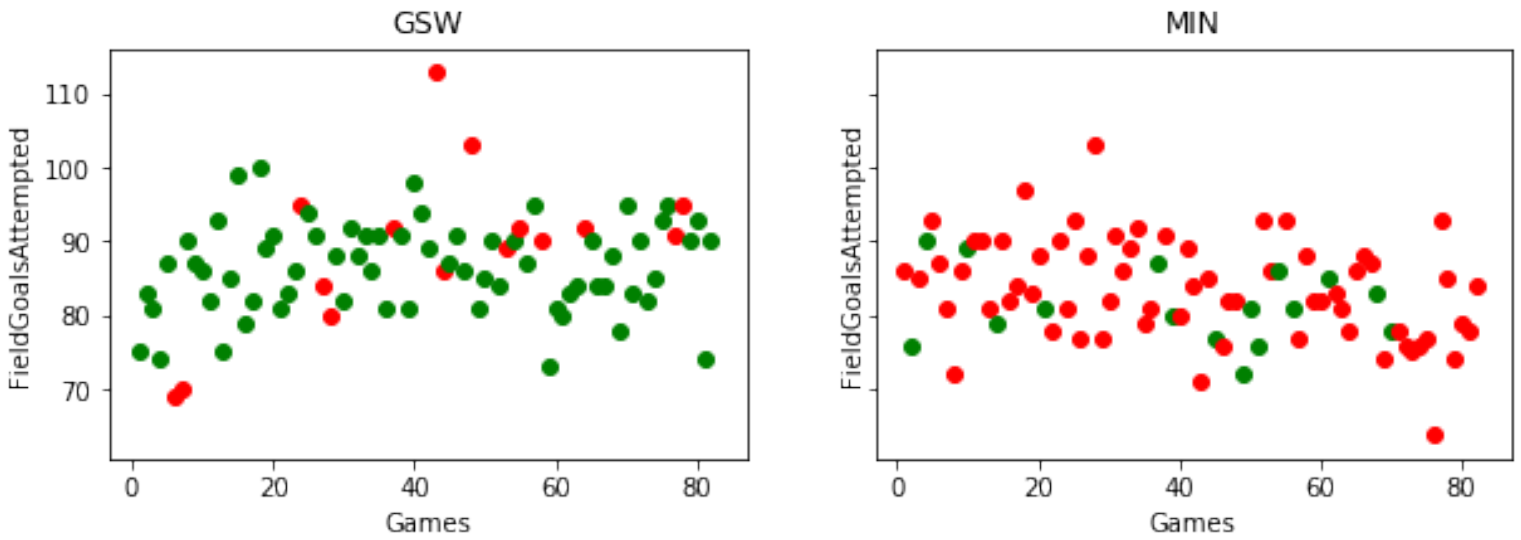


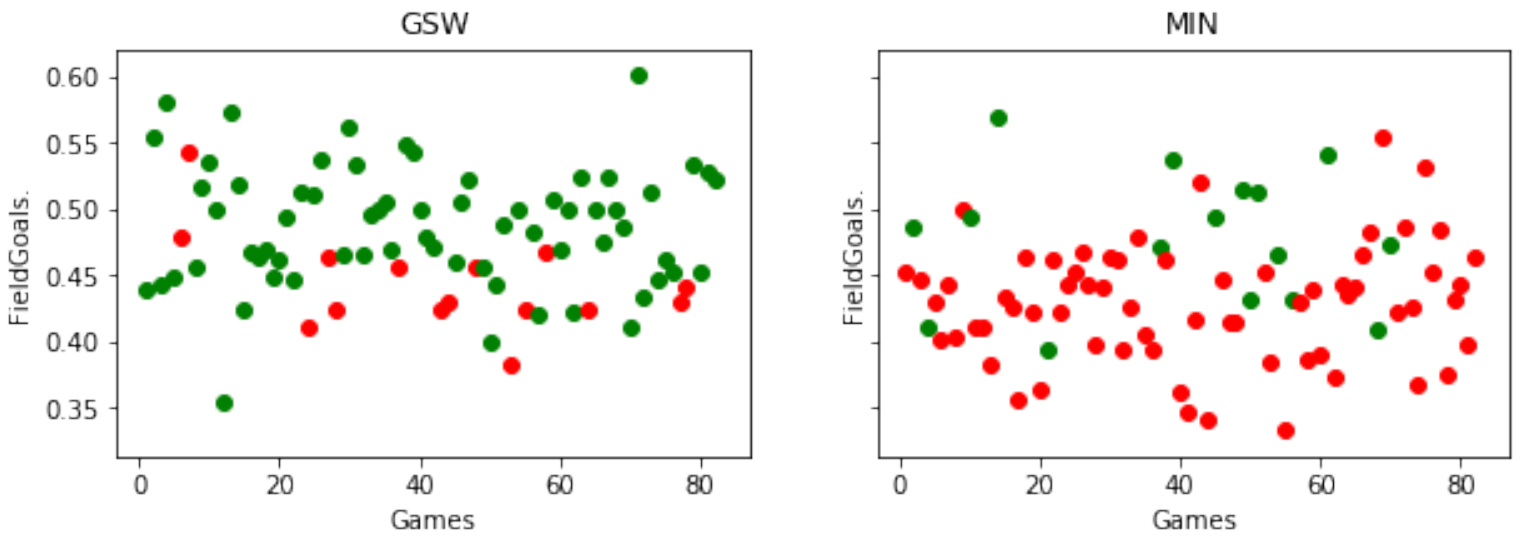

Taking a look at the Field Goals Attempted by each team, we don't see such a drastic difference. Both teams appear to be taking shots at about the same frequency. Perhaps the last seed just doesn't make as many shots despite taking about the same number of shots. Here, we did find a difference. However, it was not as drastic as we thought it would be. Let's look at the type of shots attempted and made.

In [275]:

```
print( "Green = Win ")
print( "Red = Loss")
scatPlot( 'FieldGoalsAttempted', firstSeed_2014, lastSeed_2014 )
scatPlot( 'FieldGoals.', firstSeed_2014, lastSeed_2014 )
```

```
Green = Win
Red = Loss
```



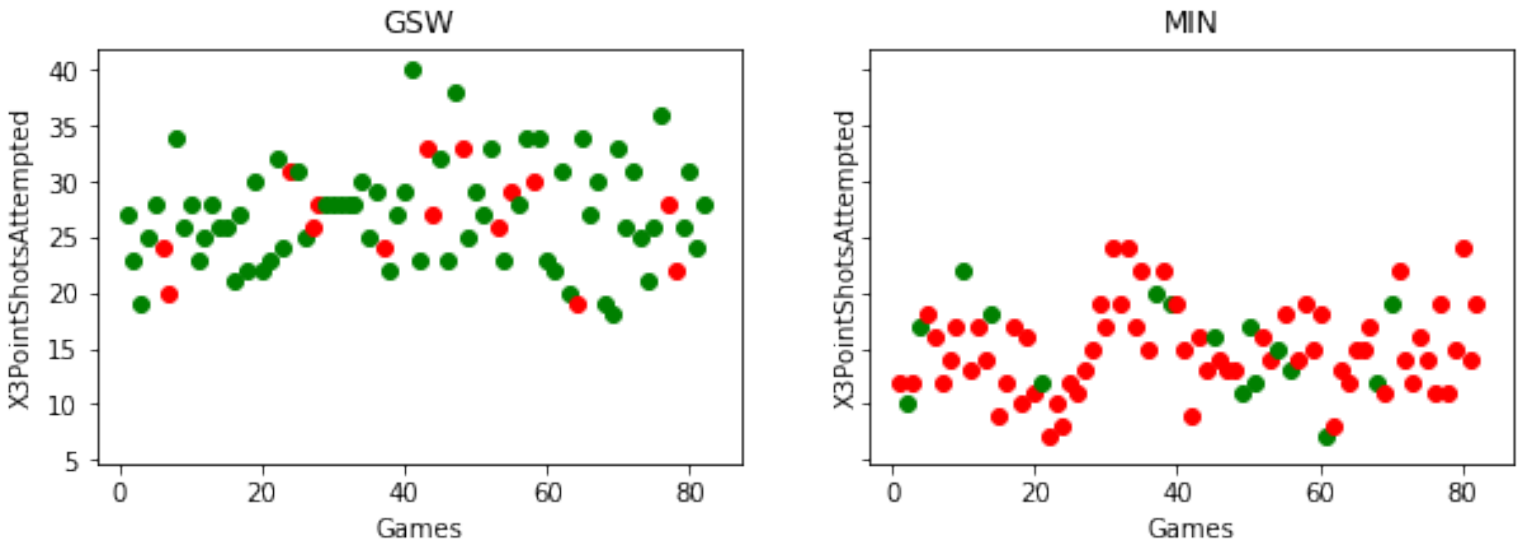



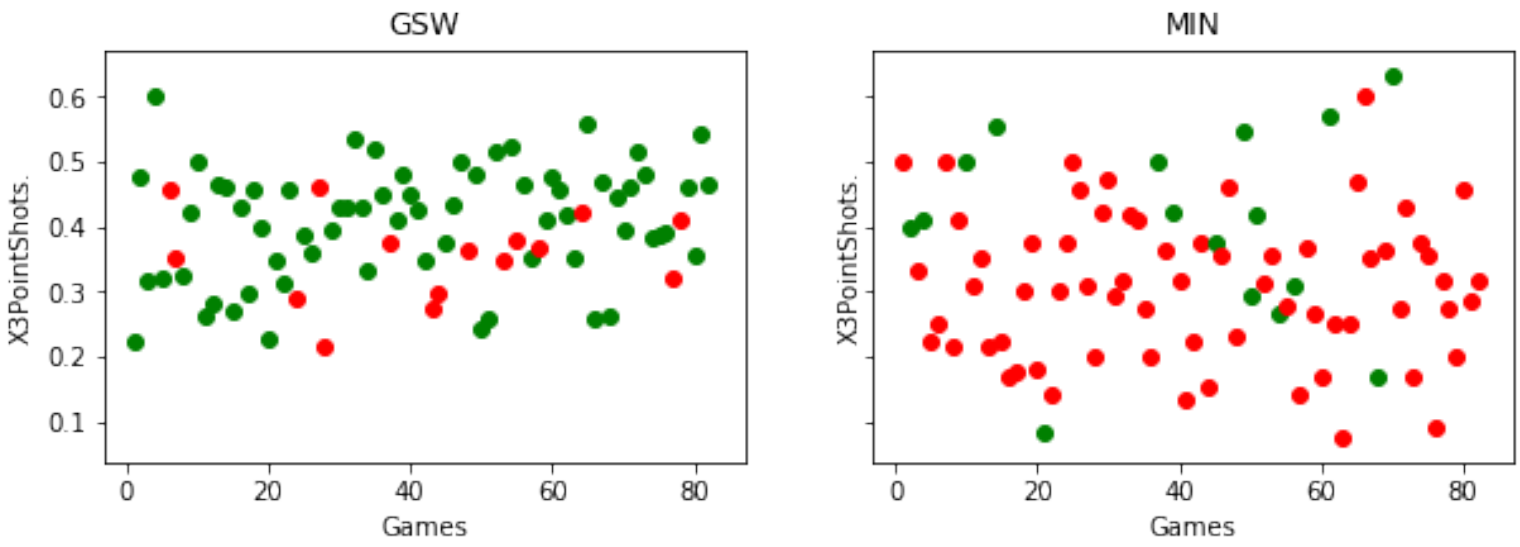

Let us now compare the type of shots during this season. We see that the overall 3 Point Shots attempted by the first seed team is consistently higher than the last seed team. Furthermore, The first seed had a higher and more consistent 3 point shot making rate than the last seed. This falls in line with our hypothesis. Let's look at past years and see the trend of the 3 point shot.

In [276]:

```
print( "Green = Win ")
print( "Red = Loss")
scatPlot( 'X3PointShotsAttempted', firstSeed_2014, lastSeed_2014 )
scatPlot( 'X3PointShots.', firstSeed_2014, lastSeed_2014 )
```

```
Green = Win
Red = Loss
```




## The 3 point shot over the years

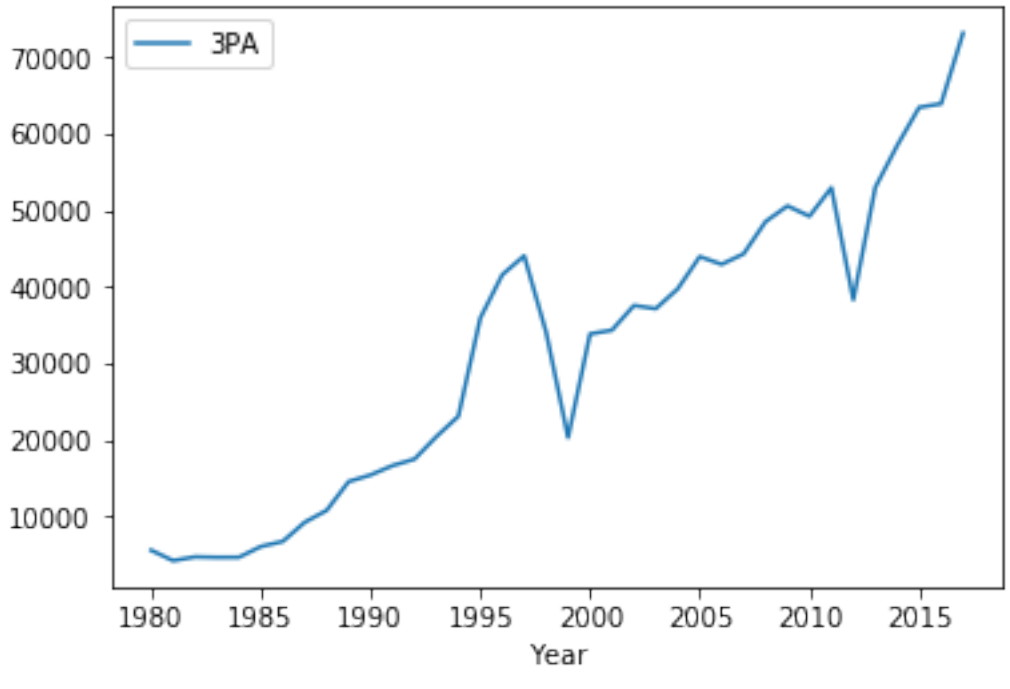
We plotted the amount of 3 point attempts as well as the number of 3 points made throughout the NBA from 1980. It quickly becomes evident that the 3 point shot has gained popularity as the League got older and older. We see a few spikes between 1995 and 2000, but variants are to be expected. However, the over trend is consistent and is positively correlated with the year. We see the maximum 3 Point Attempts and 3 Point shots made in 2018, which is the latest data we have.

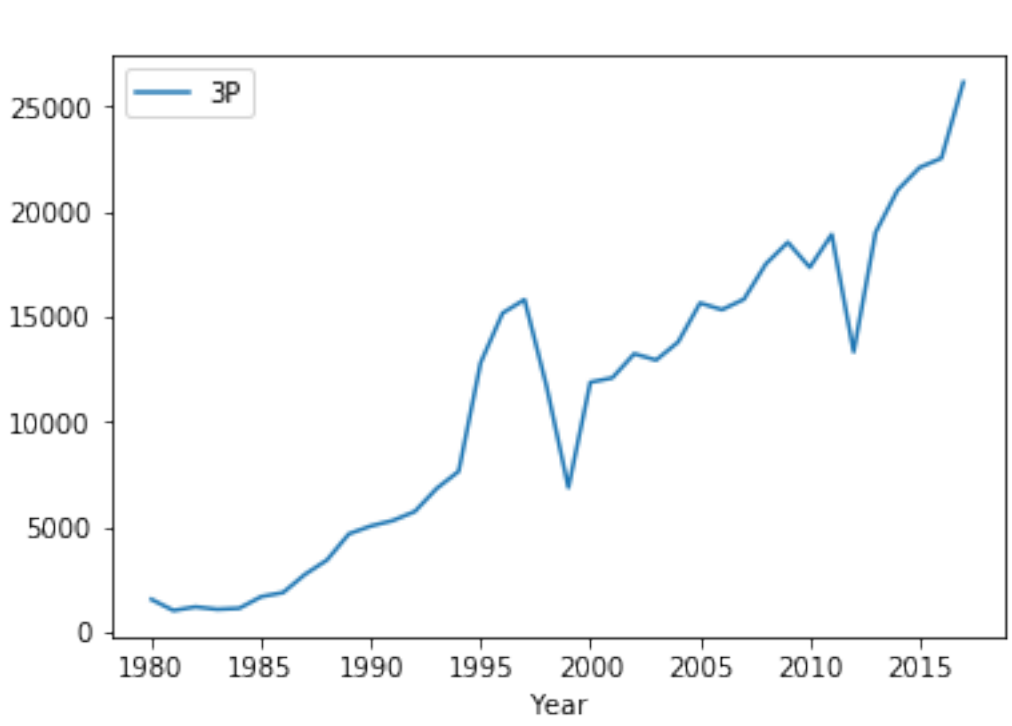
In [277]:

```
df_yearSum.set_index('Year')
df_yearSum.plot( x = 'Year', y='3PA')
df_yearSum.plot( x = 'Year', y='3P')
```

Out[277]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a21ee9ba8>
```




## 3 Point Shot respective to position

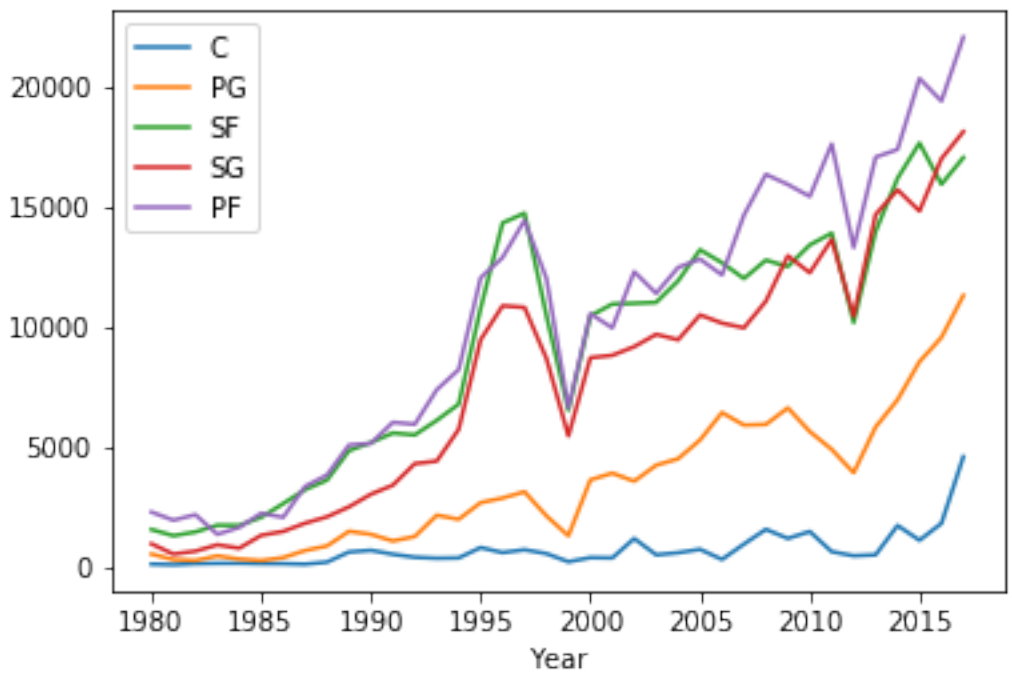
Seeing as the 3 point shot is becoming more and more dominant throughout the league, we were intersted if perhaps certain position had adopted the 3 point shot more quickly or if certain positions had not adopted the 3 point shot and opted to stick with past styles of scoring tendencies. We found that Power Forwards seemed to be adopting the 3 point shot. It's a bit suprising that they have adopted the 3 point shot faster and more frequently than point guards. Centers have more or less stuck with their style of play, but have become more and more versatile with the 3 point shot, especially over the last few years.

In [278]:

```
fig, ax = plt.subplots()
df_sumPos.groupby('Pos').plot(x='Year', y='3PA', ax=ax)
ax.legend(['C', 'PG', 'SF', 'SG', 'PF'] )
```

Out[278]:

```
<matplotlib.legend.Legend at 0x1a239916d8>
```



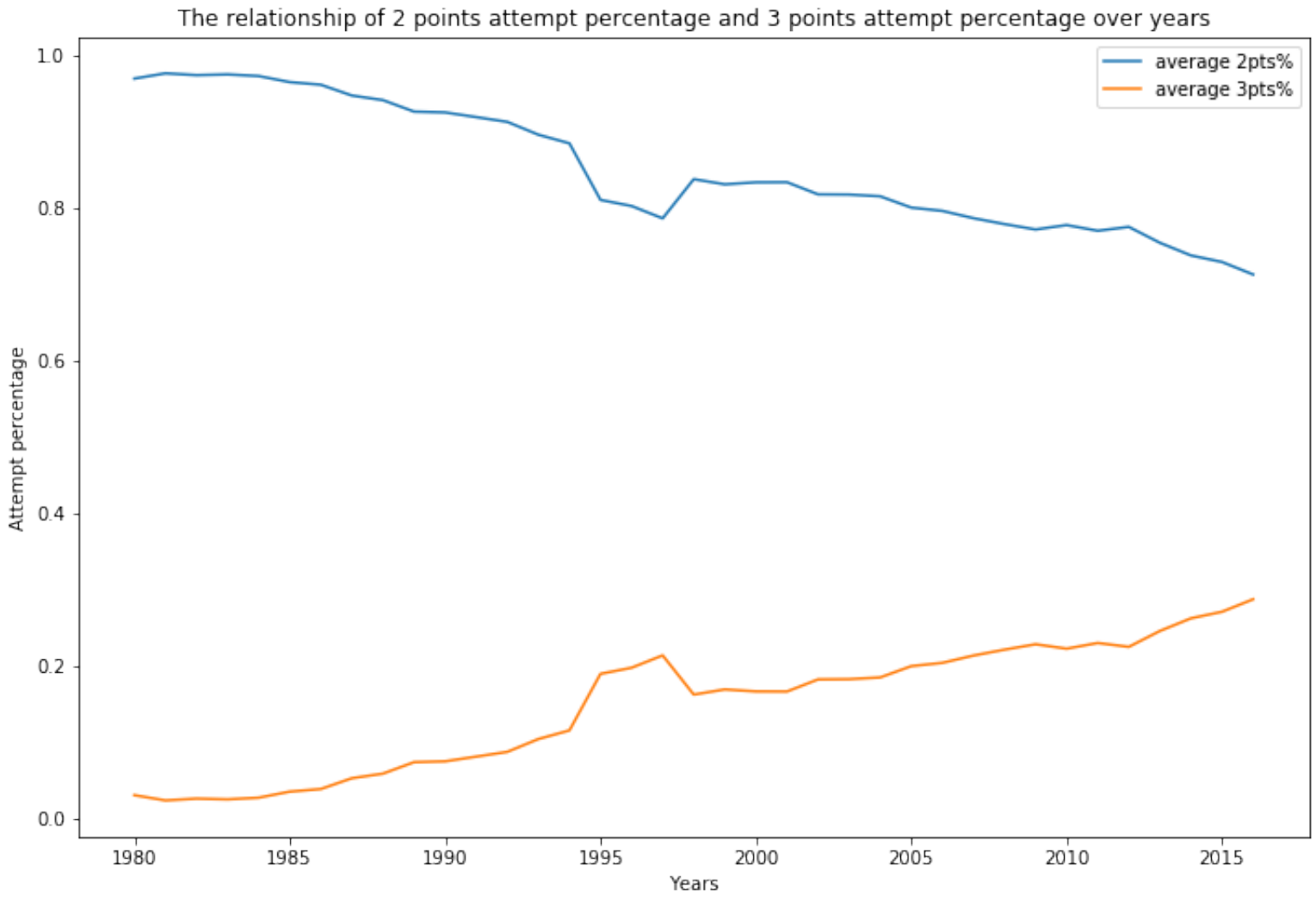

It seems 3 point shot is more agressive than in the past, whereas 2 point shot become more reserve. Therefore, we want to see the change of 2 point shot rate and 3 point shot rate over the years. In below, we show the graph of 3 points percentage and 2 points percentage VS year.

In [279]:

```
# here, we want to show the graph of 3 points percentage and 2 points percentage
VS year
# add two new column: '3PAr', '2PAr', which represents 3 points attempt percenta
ge and 2 points attempt percentage

# create two empty lists
x_2PAr = []
x_3PAr = []
df_2['Year'] = df_2['Year'].astype(np.int64)

# loop through each year and compute the 2 and 3 point shot attempt percentage
for i in range(1980, 2017):
    total_2PA = sum(df_2.loc[df_2['Year'] == i]['2PA'])
    total_3PA = sum(df_2.loc[df_2['Year'] == i]['3PA'])
    total_FGA = sum(df_2.loc[df_2['Year'] == i]['FGA'])
    x_2PAr.append(total_2PA / total_FGA)
    x_3PAr.append(total_3PA / total_FGA)

# plot and label the graph
plt.figure(figsize=(12, 8))
plt.plot(range(1980, 2017), x_2PAr, label="average 2pts%")
plt.plot(range(1980, 2017), x_3PAr, label="average 3pts%")
plt.xlabel('Years')
plt.ylabel('Attempt percentage')
plt.title('The relationship of 2 points attempt percentage and 3 points attempt
percentage over years')
plt.legend()
```

Out[279]:

<matplotlib.legend.Legend at 0x1a21945b70>



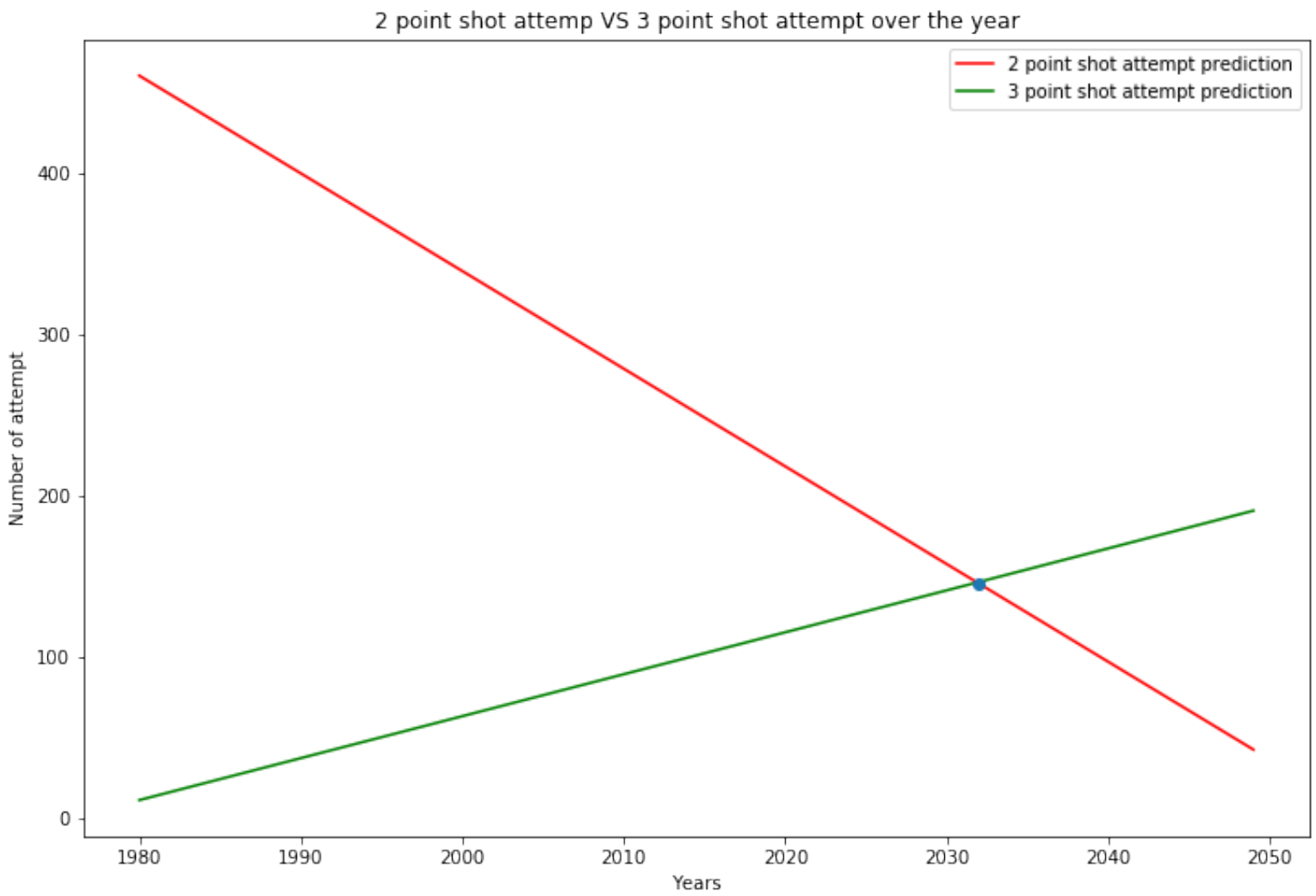
In [280]:

```
# specfy the graph
plt.figure(figsize=(12, 8))

# predict the 2 point shot in the future
a1, b1 = np.polyfit(df_2['Year'], df_2['2PA'], 1)
pred_2P = a1 * np.arange(1980, 2050) + b1
plt.plot(np.arange(1980, 2050), pred_2P, color='red', label="2 point shot attemp
t prediction")

# predict the 3 point shot in the future
a2, b2 = np.polyfit(df_2['Year'], df_2['3PA'], 1)
pred_3P = a2 * np.arange(1980, 2050) + b2
plt.plot(np.arange(1980, 2050), pred_3P, color='green', label="3 point shot atte
mpt prediction")

# now, loop through each years to find the value of intersection point, save the
x value of the intersection
# to variable i
for i in range(1980, 2050):
    if((a2 * i + b2) >= (a1 * i + b1)):
        print(i)
        break

# plot the cross point
plt.plot(i, (a1 * i + b1), 'o')

# label the graph
plt.xlabel('Years')
plt.ylabel('Number of attempt')
plt.title('2 point shot attemp VS 3 point shot attempt over the year')
plt.legend()
```

2032

Out[280]:

<matplotlib.legend.Legend at 0x1a1afb4c88>




# Ethics & Privacy

Since the datasets we used were publicly accessible and were just stats of the NBA players there wouldn't be much of a reason to get rid of many personal identifiers. But this was some sort of high school players' stats or anything else that was not publicly available there would be a lot of things that would remain anonymous. For example the way we considered ethics and privacy in our datasets was that we removed every players' name from the dataset. Also, the players' height, weight, and hometowns were removed to protect their identity as well.

As for the rest of the data such as points, blocks, steals we felt like were not data that needed to get removed for privacy reasons. The only reason that points, blocks, steals etc. are not shown in our cleaned version of our dataset is because we only focused on players' two point and three point percentages and attempted shots. That is why positions was not removed because we need to get the percentages and shots over the five different positions which were point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C).

# Conclusion & Discussion

As you can see from our data visualization, from the introduction of the three-point line in 1979 till now the three-point shot has been emphasized a lot more at an increasing rate especially over the past four years with Stephen Curry and the Warriors changing the way basketball is usually played. In the first graph, you can see that the percentage of three point shots taken in the league over the years has been increasing and the percentage of two point shots has been decreasing. This fits our hypothesis that the three-point shot is being emphasized more than the two-point shot. We also looked at the Golden State Warriors and the last seeded teams three point shot attempted where you could see the amount of wins from each team. While the difference in shot attempts was not that much Golden State Warriors had a lot more wins than the last seeded team which would mean that they probably made a higher percentage of their threes tha

We also checked the positions and the three point shots attempted over the years. Usually basketball early on was more physical where the power forward (PF) and center (C) positions were dominant since they would score a lot of mid-range shots. This meant that not a lot of three pointers would be taken by these positions. Mostly three point shots usually came from the point guard (PG) and shooting guard (SG) positions. When you look at the chart where it graphs the three-point shot by positions, you can see that along with all the other positions, the power forward and center positions have had a sharp increase in the three-point shot over the past couple of years. This follows out hypothesis because positions that were mostly known for being dominant in the mid-range and post now also are shooting three point shots more.

Also, we came up with a prediction of three-point percentage vs. two-point percentage in the future and our model predicts that the cross point will happen in 2032. This brings up an important discussion that has been brought in the recent years because of the sharp rise in three point shots. The NBA has been thinking of either introducing a four-point line that would be farther away from the basket or just extending the three-point line farther away. The mere fact that they are having this discussion follows our hypothesis that the three-point shot is getting more emphasized and maybe becoming too oversaturated for the audience that is watching the game. Another thing is that since the three-point shot is so relevant now, point guards, shooting guards, and small forwards are more dominant in the game while the power forwards and centers have taken a step back. Now power forwards and centers are expected to shoot threes whereas before they were only required to be good in the mid-range and post-game.