Comparison Between R and Pandas

Introduction

I have been thinking what I should choose for second topic of the post- I wanted it to be relatable to my life and the courseworks I've taken. Taking Data C100 and Stats 133 at the same time, I found out that even though the goal of both classes is in analyzing data, the functions and ways to clean the data, extract graph, and make tables were different in each class. In Stats 133, we use R, and in Data C100, we use Python with libraries such as pandas, matplotlib, seaborn and numpy to analyze data. I thought it would be both interesting and helpful to students who are thinking about becoming Data Analysts or choosing path that contains both CS and Stats, and decided compare two classes and the programs that are used in each.

Step 1 : Data Description and Preparation

To begin with, I'm using RStudio Version 1.0.153 and R Version 3.4.1 for R part of this post. For the data, I will be using roster csv of GSW (Golden State Warriors) that was also used in Lab12 of Stats 133. The data is from the following page : *<Reference 1>*. It contains number, name, position, height, weight, birthday, experience, and college of plyaers who played for GSW in 2017. I consulted this page for help with loading csv file into R : *<Reference 2>*. Link to all the references can be found at the very bottom of this post.

```
# load data file
gsw <- read.csv(file="GSW-roster-2017.csv",head=TRUE,sep=",")

# display the content of gsw table
gsw
```

```
##    roster.No.     roster.Player roster.Pos roster.Ht roster.Wt
## 1           0     Avery Bradley         SG     2-Jun       180
## 2           7      Jaylen Brown         SF     7-Jun       225
## 3          99       Jae Crowder         SF     6-Jun       235
## 4          30      Gerald Green         SF     7-Jun       205
## 5          42        Al Horford          C    10-Jun       245
## 6           9 Demetrius Jackson         PG     1-Jun       201
## 7           8     Jonas Jerebko         PF    10-Jun       231
## 8          90      Amir Johnson         PF     9-Jun       240
## 9          55     Jordan Mickey         PF     8-Jun       235
## 10         41      Kelly Olynyk          C    Jul-00       238
## 11         12      Terry Rozier         PG     2-Jun       190
## 12         36      Marcus Smart         SG     4-Jun       220
## 13          4     Isaiah Thomas         PG     9-May       185
## 14         13       James Young         SG     6-Jun       215
## 15         44      Tyler Zeller          C    Jul-00       253
##    roster.Birth.Date roster.. roster.Exp                roster.College
## 1          26-Nov-90       us          6 University of Texas at Austin
## 2          24-Oct-96       us          0      University of California
## 3           6-Jul-90       us          4           Marquette University
## 4          26-Jan-86       us          9
## 5           3-Jun-86       do          9         University of Florida
## 6           7-Sep-94       us          0      University of Notre Dame
## 7           2-Mar-87       se          6
## 8           1-May-87       us         11
## 9           9-Jul-94       us          1    Louisiana State University
## 10         19-Apr-91       ca          3            Gonzaga University
## 11         17-Mar-94       us          1      University of Louisville
## 12          6-Mar-94       us          2     Oklahoma State University
## 13          7-Feb-89       us          5      University of Washington
## 14         16-Aug-95       us          2        University of Kentucky
## 15         17-Jan-90       us          4  University of North Carolina
```

Next, I loaded the data into Jupyter notebook to demonstrate how data is loaded using pandas in Data C100. The version of Python I used is 2.7. I first had to import 'pandas' and 'numpy' libraries into the notebook. I consulted following page for help in importing csv into Jupyter notebook : *<Reference 3>*. Again, link is at the bottom of this post.

First import the data we are going to use to Jupyter notebook:

In [1]:
```
import pandas as pd
import numpy as np
```

In [36]:
```
# load data into pandas dataframe
gsw = pd.read_csv('GSW-roster-2017.csv')

# display content of gsw dataframe
gsw
```

Out[36]:

	roster.No.	roster.Player	roster.Pos	roster.Ht	roster.Wt	roster.Birth.Date	roster..	roster.Exp	roster.College
0	0	Avery Bradley	SG	2-Jun	180	26-Nov-90	us	6	University of Texas at Austin
1	7	Jaylen Brown	SF	7-Jun	225	24-Oct-96	us	0	University of California
2	99	Jae Crowder	SF	6-Jun	235	6-Jul-90	us	4	Marquette University
3	30	Gerald Green	SF	7-Jun	205	26-Jan-86	us	9	NaN
4	42	Al Horford	C	10-Jun	245	3-Jun-86	do	9	University of Florida
5	9	Demetrius Jackson	PG	1-Jun	201	7-Sep-94	us	0	University of Notre Dame
6	8	Jonas Jerebko	PF	10-Jun	231	2-Mar-87	se	6	NaN
7	90	Amir Johnson	PF	9-Jun	240	1-May-87	us	11	NaN
8	55	Jordan Mickey	PF	8-Jun	235	9-Jul-94	us	1	Louisiana State University
9	41	Kelly Olynyk	C	Jul-00	238	19-Apr-91	ca	3	Gonzaga University
10	12	Terry Rozier	PG	2-Jun	190	17-Mar-94	us	1	University of Louisville
11	36	Marcus Smart	SG	4-Jun	220	6-Mar-94	us	2	Oklahoma State University
12	4	Isaiah Thomas	PG	9-May	185	7-Feb-89	us	5	University of Washington
13	13	James Young	SG	6-Jun	215	16-Aug-95	us	2	University of Kentucky
14	44	Tyler Zeller	C	Jul-00	253	17-Jan-90	us	4	University of North Carolina

Step 2 : Exploring Data

In this stage, I will play around with data and find out if there are any missing or wrong values and correct them.

```
# find attributes of gsw table
attributes(gsw)
```

```
## $names
## [1] "roster.No."        "roster.Player"     "roster.Pos"
## [4] "roster.Ht"         "roster.Wt"         "roster.Birth.Date"
## [7] "roster.."          "roster.Exp"        "roster.College"
##
## $class
## [1] "data.frame"
##
## $row.names
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
```

I consulted following page to change the names of columns of gsw for clarity : <*Reference 4*>.

```
# change name of columns
names(gsw) <- c("No.", "Name", "Pos", "Height", "Weight", "Birth", "Country", "Exp", "College")
gsw
```

```
##    No.              Name Pos Height Weight     Birth Country Exp
## 1    0     Avery Bradley  SG   2-Jun    180 26-Nov-90      us   6
## 2    7      Jaylen Brown  SF   7-Jun    225 24-Oct-96      us   0
## 3   99       Jae Crowder  SF   6-Jun    235  6-Jul-90      us   4
## 4   30      Gerald Green  SF   7-Jun    205 26-Jan-86      us   9
## 5   42        Al Horford   C  10-Jun    245  3-Jun-86      do   9
## 6    9 Demetrius Jackson  PG   1-Jun    201  7-Sep-94      us   0
## 7    8     Jonas Jerebko  PF  10-Jun    231  2-Mar-87      se   6
## 8   90      Amir Johnson  PF   9-Jun    240  1-May-87      us  11
## 9   55     Jordan Mickey  PF   8-Jun    235  9-Jul-94      us   1
## 10  41      Kelly Olynyk   C Jul-00    238 19-Apr-91      ca   3
## 11  12      Terry Rozier  PG   2-Jun    190 17-Mar-94      us   1
## 12  36      Marcus Smart  SG   4-Jun    220  6-Mar-94      us   2
## 13   4     Isaiah Thomas  PG   9-May    185  7-Feb-89      us   5
## 14  13       James Young  SG   6-Jun    215 16-Aug-95      us   2
## 15  44      Tyler Zeller   C Jul-00    253 17-Jan-90      us   4
##                        College
## 1  University of Texas at Austin
## 2       University of California
## 3           Marquette University
## 4
## 5          University of Florida
## 6       University of Notre Dame
## 7
## 8
## 9     Louisiana State University
## 10           Gonzaga University
## 11      University of Louisville
## 12     Oklahoma State University
## 13      University of Washington
## 14        University of Kentucky
## 15  University of North Carolina
```

By looking at the table again, I found out that there was something wrong with the data of height. It seems like height somehow got messed up with the date, so I fixed it.

```
# change height
gsw$Height = c("6-2", "6-7", "6-6", "6-7", "6-10", "6-1", "6-10", "6-9", "6-8", "7-0", "6-2", "6-4", "5-9", "6-6",
"7-0")
gsw
```

```
##    No.              Name Pos Height Weight     Birth Country Exp
## 1    0     Avery Bradley  SG    6-2    180 26-Nov-90      us   6
## 2    7      Jaylen Brown  SF    6-7    225 24-Oct-96      us   0
## 3   99       Jae Crowder  SF    6-6    235  6-Jul-90      us   4
## 4   30      Gerald Green  SF    6-7    205 26-Jan-86      us   9
## 5   42        Al Horford   C   6-10    245  3-Jun-86      do   9
## 6    9 Demetrius Jackson  PG    6-1    201  7-Sep-94      us   0
## 7    8     Jonas Jerebko  PF   6-10    231  2-Mar-87      se   6
## 8   90      Amir Johnson  PF    6-9    240  1-May-87      us  11
## 9   55     Jordan Mickey  PF    6-8    235  9-Jul-94      us   1
## 10  41      Kelly Olynyk   C    7-0    238 19-Apr-91      ca   3
## 11  12      Terry Rozier  PG    6-2    190 17-Mar-94      us   1
## 12  36      Marcus Smart  SG    6-4    220  6-Mar-94      us   2
## 13   4     Isaiah Thomas  PG    5-9    185  7-Feb-89      us   5
## 14  13       James Young  SG    6-6    215 16-Aug-95      us   2
## 15  44      Tyler Zeller   C    7-0    253 17-Jan-90      us   4
##                        College
## 1  University of Texas at Austin
## 2       University of California
## 3           Marquette University
## 4
## 5          University of Florida
## 6       University of Notre Dame
## 7
## 8
## 9     Louisiana State University
## 10           Gonzaga University
## 11      University of Louisville
## 12     Oklahoma State University
## 13      University of Washington
## 14        University of Kentucky
## 15  University of North Carolina
```

Then, I looked at the data to see if there are any missing values. The College column had some empty spaces, so I replaced them with 'NA'.

```
# fill empty spaces in gsw with 'NA'
gsw$College[gsw$College==""] <- "NA"
```

```
## Warning in `[<-.factor`(`*tmp*`, gsw$College == "", value =
## structure(c(12L, : invalid factor level, NA generated
```

```
gsw
```

```
##    No.              Name Pos Height Weight     Birth Country Exp
## 1    0     Avery Bradley  SG    6-2    180 26-Nov-90      us   6
## 2    7      Jaylen Brown  SF    6-7    225 24-Oct-96      us   0
## 3   99       Jae Crowder  SF    6-6    235  6-Jul-90      us   4
## 4   30      Gerald Green  SF    6-7    205 26-Jan-86      us   9
## 5   42        Al Horford   C   6-10    245  3-Jun-86      do   9
## 6    9 Demetrius Jackson  PG    6-1    201  7-Sep-94      us   0
## 7    8     Jonas Jerebko  PF   6-10    231  2-Mar-87      se   6
## 8   90      Amir Johnson  PF    6-9    240  1-May-87      us  11
## 9   55     Jordan Mickey  PF    6-8    235  9-Jul-94      us   1
## 10  41      Kelly Olynyk   C    7-0    238 19-Apr-91      ca   3
## 11  12      Terry Rozier  PG    6-2    190 17-Mar-94      us   1
## 12  36      Marcus Smart  SG    6-4    220  6-Mar-94      us   2
## 13   4     Isaiah Thomas  PG    5-9    185  7-Feb-89      us   5
## 14  13       James Young  SG    6-6    215 16-Aug-95      us   2
## 15  44      Tyler Zeller   C    7-0    253 17-Jan-90      us   4
##                         College
## 1  University of Texas at Austin
## 2       University of California
## 3           Marquette University
## 4                          <NA>
## 5          University of Florida
## 6       University of Notre Dame
## 7                          <NA>
## 8                          <NA>
## 9     Louisiana State University
## 10           Gonzaga University
## 11      University of Louisville
## 12     Oklahoma State University
## 13      University of Washington
## 14        University of Kentucky
## 15  University of North Carolina
```

For the final data manipulation, I converted column Birth into a year that the player was born. I consulted this page for help : <*Reference 5*>. I had to use 'stringr' library for this work. If this library is not installed in your computer, you should install it first by using **packages.install("stringr")**.

```
# import stringr to use
library(stringr)

# clean Birth column
vals = array("character")
for (x in seq(1, 15, 1)) {
  vals[x] = str_sub(gsw$Birth[x],-2,-1)
}
gsw$Birth = as.numeric(vals) + 1900
gsw
```

```
##    No.             Name Pos Height Weight Birth Country Exp
## 1    0     Avery Bradley  SG    6-2    180  1990      us   6
## 2    7      Jaylen Brown  SF    6-7    225  1996      us   0
## 3   99       Jae Crowder  SF    6-6    235  1990      us   4
## 4   30      Gerald Green  SF    6-7    205  1986      us   9
## 5   42        Al Horford   C   6-10    245  1986      do   9
## 6    9 Demetrius Jackson  PG    6-1    201  1994      us   0
## 7    8     Jonas Jerebko  PF   6-10    231  1987      se   6
## 8   90      Amir Johnson  PF    6-9    240  1987      us  11
## 9   55     Jordan Mickey  PF    6-8    235  1994      us   1
## 10  41      Kelly Olynyk   C    7-0    238  1991      ca   3
## 11  12      Terry Rozier  PG    6-2    190  1994      us   1
## 12  36      Marcus Smart  SG    6-4    220  1994      us   2
## 13   4     Isaiah Thomas  PG    5-9    185  1989      us   5
## 14  13       James Young  SG    6-6    215  1995      us   2
## 15  44      Tyler Zeller   C    7-0    253  1990      us   4
##                        College
## 1  University of Texas at Austin
## 2       University of California
## 3           Marquette University
## 4                         <NA>
## 5          University of Florida
## 6       University of Notre Dame
## 7                         <NA>
## 8                         <NA>
## 9     Louisiana State University
## 10           Gonzaga University
## 11     University of Louisville
## 12     Oklahoma State University
## 13      University of Washington
## 14        University of Kentucky
## 15  University of North Carolina
```

I did the same in Jupyter notebook, but using other libraries and functions. Note that in Python, the index starts from 0. For the renaming of columns, I looked at this page: *<Reference 6>*.

Let's explore and clean the data so we can use it for later analysis:

In [37]:
```
gsw.rename(index=str, columns={"roster.No": "No.", "roster.Player": "Name", "roster.Pos": "Pos",
                               "roster.Ht": "Height", "roster.Wt": "Weight", "roster.Birth.Date":
                               "Birth_Date", "roster..": "Country", "roster.Exp" : "Exp",
                               "roster.College": "College"}, inplace = True)
gsw
```

Out[37]:

	roster.No.	Name	Pos	Height	Weight	Birth_Date	Country	Exp	College
0	0	Avery Bradley	SG	2-Jun	180	26-Nov-90	us	6	University of Texas at Austin
1	7	Jaylen Brown	SF	7-Jun	225	24-Oct-96	us	0	University of California
2	99	Jae Crowder	SF	6-Jun	235	6-Jul-90	us	4	Marquette University
3	30	Gerald Green	SF	7-Jun	205	26-Jan-86	us	9	NaN
4	42	Al Horford	C	10-Jun	245	3-Jun-86	do	9	University of Florida
5	9	Demetrius Jackson	PG	1-Jun	201	7-Sep-94	us	0	University of Notre Dame
6	8	Jonas Jerebko	PF	10-Jun	231	2-Mar-87	se	6	NaN
7	90	Amir Johnson	PF	9-Jun	240	1-May-87	us	11	NaN
8	55	Jordan Mickey	PF	8-Jun	235	9-Jul-94	us	1	Louisiana State University
9	41	Kelly Olynyk	C	Jul-00	238	19-Apr-91	ca	3	Gonzaga University
10	12	Terry Rozier	PG	2-Jun	190	17-Mar-94	us	1	University of Louisville
11	36	Marcus Smart	SG	4-Jun	220	6-Mar-94	us	2	Oklahoma State University
12	4	Isaiah Thomas	PG	9-May	185	7-Feb-89	us	5	University of Washington
13	13	James Young	SG	6-Jun	215	16-Aug-95	us	2	University of Kentucky
14	44	Tyler Zeller	C	Jul-00	253	17-Jan-90	us	4	University of North Carolina

Then I filled the missing NaNs with 'Unknown', following this page : *<Reference 7>*.

In [10]:
```
gsw['College'].fillna('Unknown', inplace=True)
gsw
```

Out[10]:

	roster.No.	Name	Pos	Height	Weight	Birth_Date	Country	Exp	College
0	0	Avery Bradley	SG	2-Jun	180	26-Nov-90	us	6	University of Texas at Austin
1	7	Jaylen Brown	SF	7-Jun	225	24-Oct-96	us	0	University of California
2	99	Jae Crowder	SF	6-Jun	235	6-Jul-90	us	4	Marquette University
3	30	Gerald Green	SF	7-Jun	205	26-Jan-86	us	9	Unknown
4	42	Al Horford	C	10-Jun	245	3-Jun-86	do	9	University of Florida
5	9	Demetrius Jackson	PG	1-Jun	201	7-Sep-94	us	0	University of Notre Dame
6	8	Jonas Jerebko	PF	10-Jun	231	2-Mar-87	se	6	Unknown
7	90	Amir Johnson	PF	9-Jun	240	1-May-87	us	11	Unknown
8	55	Jordan Mickey	PF	8-Jun	235	9-Jul-94	us	1	Louisiana State University
9	41	Kelly Olynyk	C	Jul-00	238	19-Apr-91	ca	3	Gonzaga University
10	12	Terry Rozier	PG	2-Jun	190	17-Mar-94	us	1	University of Louisville
11	36	Marcus Smart	SG	4-Jun	220	6-Mar-94	us	2	Oklahoma State University
12	4	Isaiah Thomas	PG	9-May	185	7-Feb-89	us	5	University of Washington
13	13	James Young	SG	6-Jun	215	16-Aug-95	us	2	University of Kentucky
14	44	Tyler Zeller	C	Jul-00	253	17-Jan-90	us	4	University of North Carolina

Then I changed the heights into right units.

In [14]:
```
gsw['Height'] = ["6-2", "6-7", "6-6", "6-7", "6-10", "6-1", "6-10", "6-9", "6-8", "7-0",
                 "6-2", "6-4", "5-9", "6-6", "7-0"]
gsw
```

Out[14]:

	roster.No.	Name	Pos	Height	Weight	Birth_Date	Country	Exp	College
0	0	Avery Bradley	SG	6-2	180	26-Nov-90	us	6	University of Texas at Austin
1	7	Jaylen Brown	SF	6-7	225	24-Oct-96	us	0	University of California
2	99	Jae Crowder	SF	6-6	235	6-Jul-90	us	4	Marquette University
3	30	Gerald Green	SF	6-7	205	26-Jan-86	us	9	Unknown
4	42	Al Horford	C	6-10	245	3-Jun-86	do	9	University of Florida
5	9	Demetrius Jackson	PG	6-1	201	7-Sep-94	us	0	University of Notre Dame
6	8	Jonas Jerebko	PF	6-10	231	2-Mar-87	se	6	Unknown
7	90	Amir Johnson	PF	6-9	240	1-May-87	us	11	Unknown
8	55	Jordan Mickey	PF	6-8	235	9-Jul-94	us	1	Louisiana State University
9	41	Kelly Olynyk	C	7-0	238	19-Apr-91	ca	3	Gonzaga University
10	12	Terry Rozier	PG	6-2	190	17-Mar-94	us	1	University of Louisville
11	36	Marcus Smart	SG	6-4	220	6-Mar-94	us	2	Oklahoma State University
12	4	Isaiah Thomas	PG	5-9	185	7-Feb-89	us	5	University of Washington
13	13	James Young	SG	6-6	215	16-Aug-95	us	2	University of Kentucky
14	44	Tyler Zeller	C	7-0	253	17-Jan-90	us	4	University of North Carolina

Finally, I chagned the Birth_Date to numeric year. I used following page to add the values to column of a table : <*Reference 8*>.

In [41]:
```
gsw.Birth_Date = gsw.Birth_Date.str[-2:]
gsw.Birth_Date = pd.to_numeric(gsw.Birth_Date) + 1900
gsw
```

Out[41]:

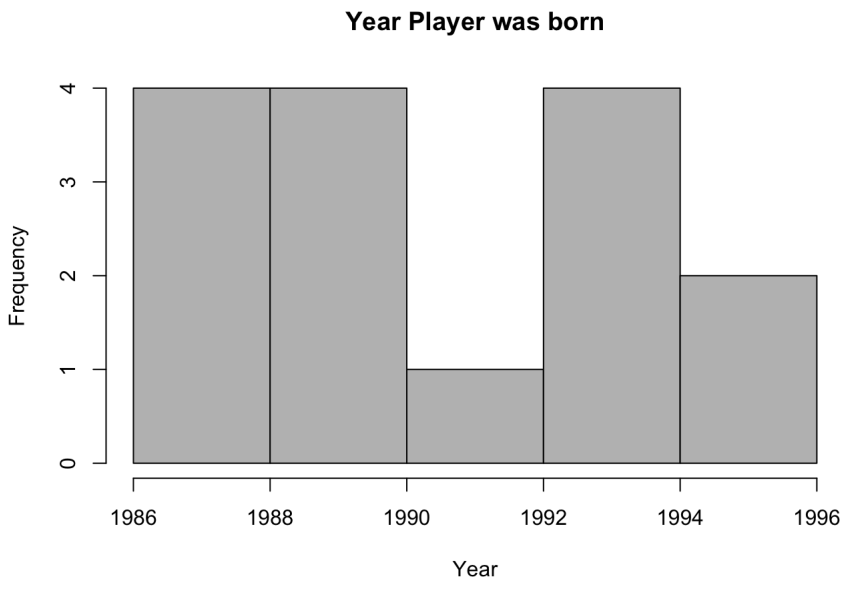
	roster.No.	Name	Pos	Height	Weight	Birth_Date	Country	Exp	College
0	0	Avery Bradley	SG	6-2	180	1990	us	6	University of Texas at Austin
1	7	Jaylen Brown	SF	6-7	225	1996	us	0	University of California
2	99	Jae Crowder	SF	6-6	235	1990	us	4	Marquette University
3	30	Gerald Green	SF	6-7	205	1986	us	9	Unknown
4	42	Al Horford	C	6-10	245	1986	do	9	University of Florida
5	9	Demetrius Jackson	PG	6-1	201	1994	us	0	University of Notre Dame
6	8	Jonas Jerebko	PF	6-10	231	1987	se	6	Unknown
7	90	Amir Johnson	PF	6-9	240	1987	us	11	Unknown
8	55	Jordan Mickey	PF	6-8	235	1994	us	1	Louisiana State University
9	41	Kelly Olynyk	C	7-0	238	1991	ca	3	Gonzaga University
10	12	Terry Rozier	PG	6-2	190	1994	us	1	University of Louisville
11	36	Marcus Smart	SG	6-4	220	1994	us	2	Oklahoma State University
12	4	Isaiah Thomas	PG	5-9	185	1989	us	5	University of Washington
13	13	James Young	SG	6-6	215	1995	us	2	University of Kentucky
14	44	Tyler Zeller	C	7-0	253	1990	us	4	University of North Carolina

Step 3 : Data Visualization

Since now we have clean data, we can use it to graph and analyze. With the data that I have, I will look for relationship between the year the player was born and year of experience he had.

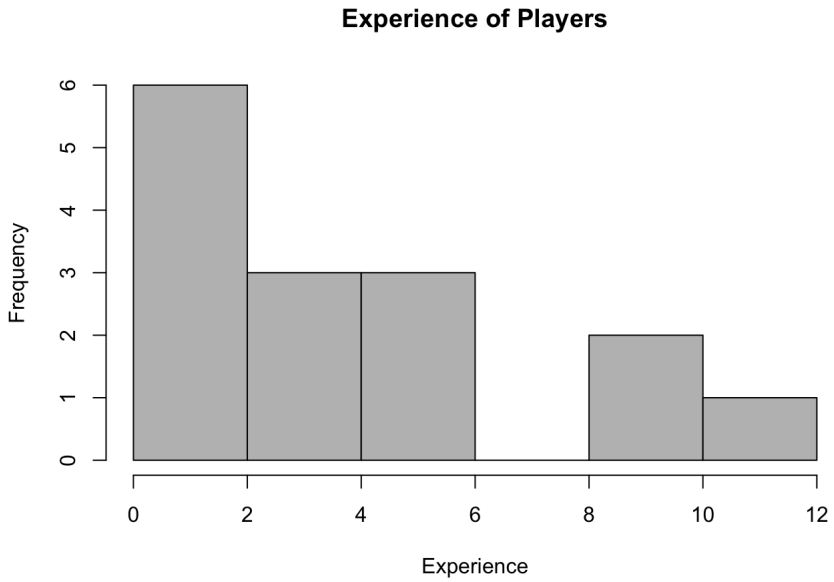
Let's first look at the year that the players were born.

```
# histogram of the years players were born
hist(gsw$Birth, main = "Year Player was born", col = "grey", xlab="Year")
```



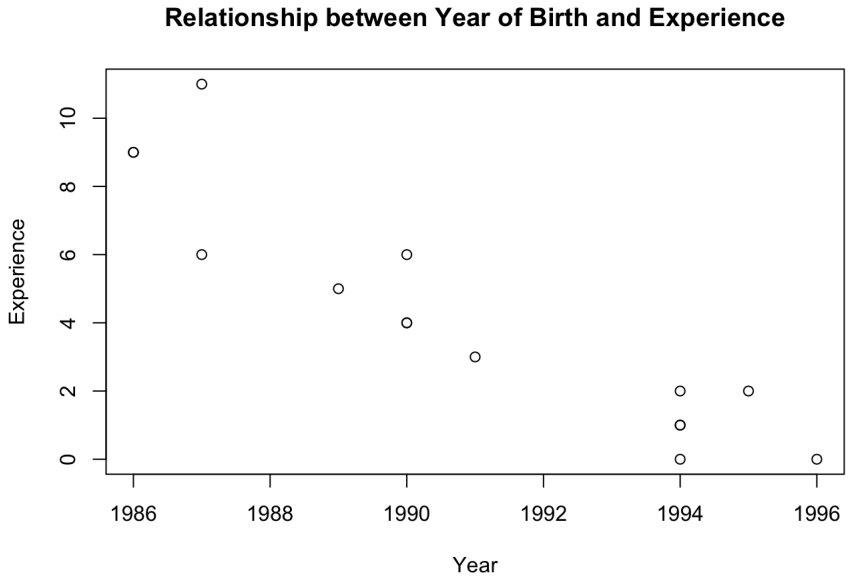
Now let's look at the experience players had.

```
# histogram of experiences of players
hist(gsw$Exp, main = "Experience of Players", col = "grey", xlab="Experience")
```



Finally, let's find the relationship between these two variables by graphing scatter plot.

```
# scatter plot between birth year and experience
plot(x = gsw$Birth, y = gsw$Exp, xlab="Year", ylab="Experience", main="Relationship between Year of Birth and Expe
rience")
```



Then I moved onto Jupyter notebook to plot graphs in Python. I consulted the following page for visualization of data : <*Reference 9*>.

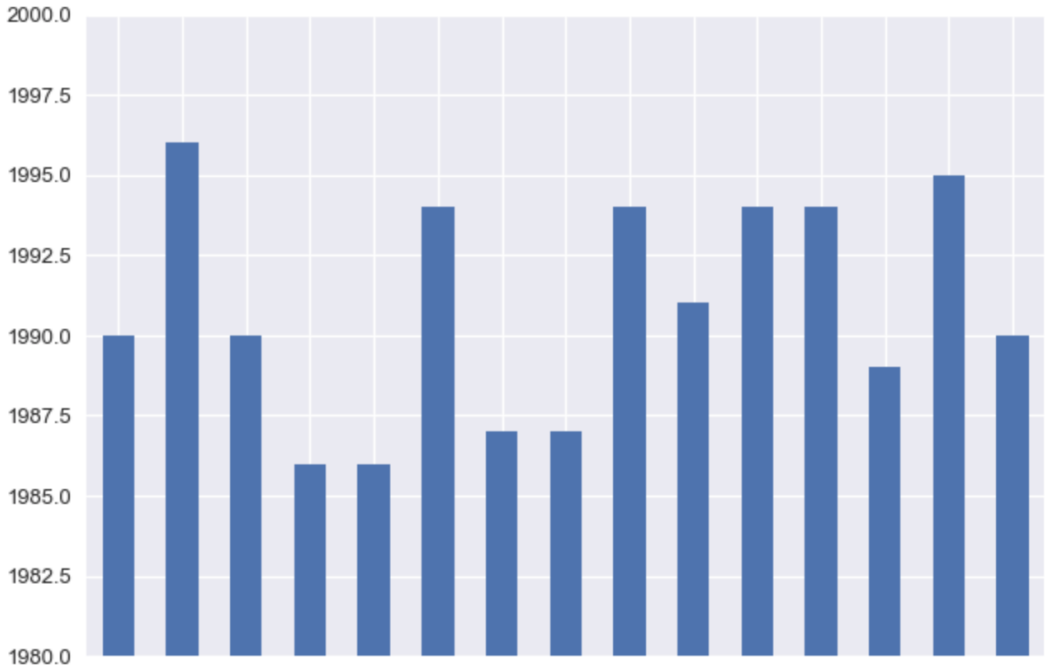
I first created bar plot of birth years of players. To do this, I had to import more libraries, such as matplotlib, seaborn and used IPython.display to show the result of plot.

```
In [46]: import matplotlib.pyplot as plt
         import seaborn as sns
         %matplotlib inline
         sns.set()
         sns.set_context("talk")

         from IPython.display import display, Latex, Markdown, HTML, Javascript
```

```
In [51]: gsw['Birth_Date'].plot(kind='bar', ylim = (1980, 2000))
```

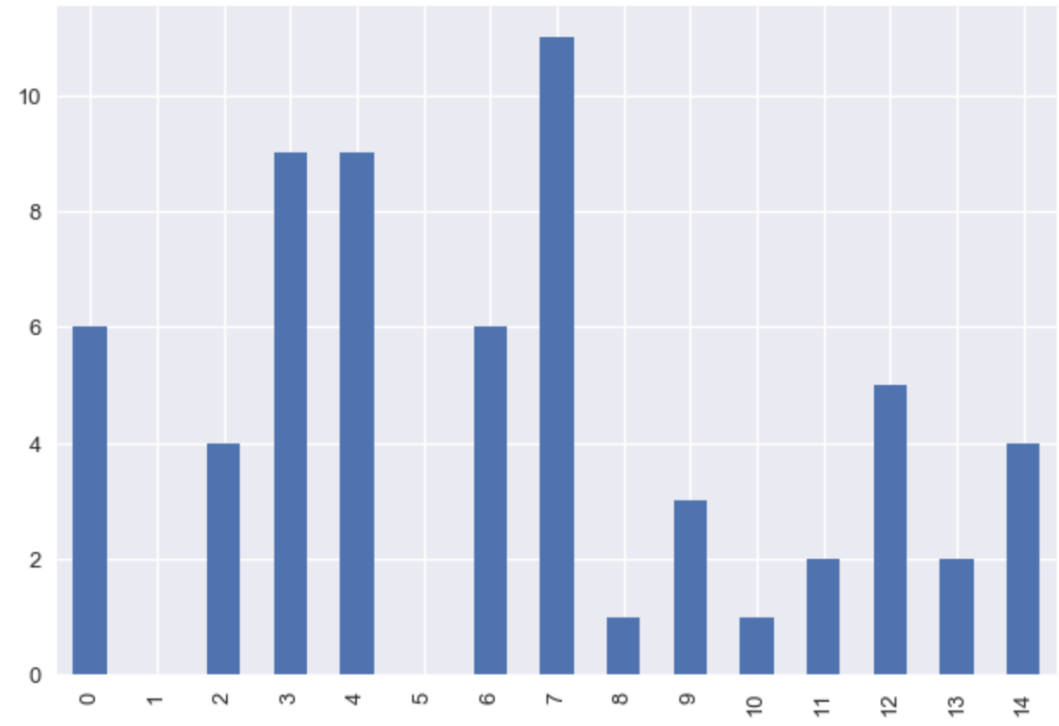
```
Out[51]: <matplotlib.axes._subplots.AxesSubplot at 0x110efaa90>
```



Then I plotted bar chart of players' experiences.

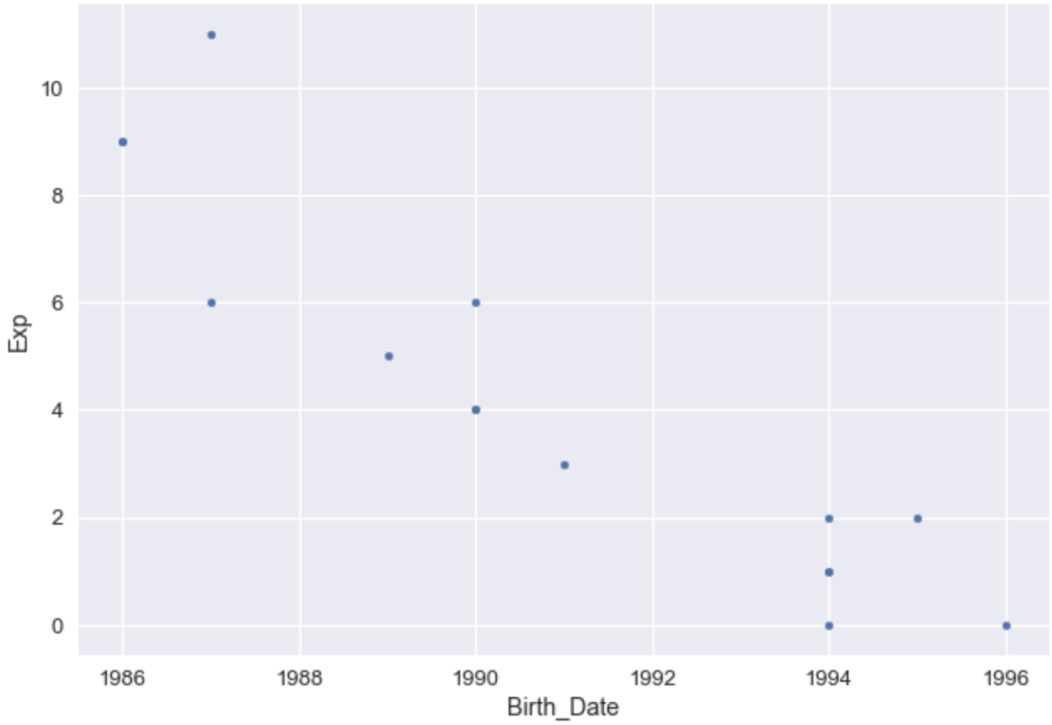
```
In [52]: gsw['Exp'].plot(kind='bar')
```

```
Out[52]: <matplotlib.axes._subplots.AxesSubplot at 0x110efa7f0>
```



Finally, I made scatterplot with these two data.

```
In [55]: gsw.plot.scatter('Birth_Date', 'Exp')
Out[55]: <matplotlib.axes._subplots.AxesSubplot at 0x110f00ac8>
```



Step 4 : Analysis

From this short data cleaning and visualization, I could find out that usually older the player is, more experienced he is, by looking at the scatter plot between birth year and experience.

Key Points and Some Lessons

One thing I found out while working on these two environments is that essence of coding is the same - only functions and libraries that we use differ for each of the programs. If you clearly know what you want to do, then there will be always a way to implement it (Google is your best friend)! You just need a bit of understanding on how the coding works, such as how to use functions, how to import libraries/packages, and how to manipulate data using different methods (for loop, while loop, indexing, etc). While working on this post, I thought that using Python (especially pandas) in Jupyter notebook was easier than using R in analyzing the data. I feel like R has many restrictions… Also, the UI and outcome of the visualizations differ a lot, and I prefer the graphs I have in Jupyter notebook. Like this, if you can choose a program to work on your project, it might be best to know the goods and bads of each programs before you begin your project and find out which one is the most comfortable and easy-to-use for you.

References

- *Reference 1* : https://www.basketball-reference.com/teams/GSW/2018.html#all_roster
- *Reference 2* : http://www.cyclismo.org/tutorial/R/input.html
- *Reference 3* : https://chrisalbon.com/python/pandas_dataframe_importing_csv.html
- *Reference 4* : http://rprogramming.net/rename-columns-in-r/
- *Reference 5* : https://stackoverflow.com/questions/7963898/extracting-the-last-n-characters-from-a-string-in-r
- *Reference 6* : https://pandas.pydata.org/pandas-docs/version/0.21/generated/pandas.DataFrame.rename.html
- *Reference 7* : https://pandas.pydata.org/pandas-docs/stable/missing_data.html
- *Reference 8* : https://stackoverflow.com/questions/30794525/adding-one-to-all-the-values-in-a-dataframe
- *Reference 9* : https://pandas.pydata.org/pandas-docs/stable/visualization.html