Data Wrangling Final Report

Hang Qi

2020/4/24

My final project for Data Wrangling is to analyse NBA data and do some data visualization.

The first dataset is downloaded from the website https://www.kaggle.com/dansbecker/nba-shot-logs/data The table below is the first 6 rows of my dataset

```
##    GAME_ID                  MATCHUP LOCATION W FINAL_MARGIN SHOT_NUMBER
## 1 21400899 MAR 04, 2015 - CHA @ BKN        A W           24           1
## 2 21400899 MAR 04, 2015 - CHA @ BKN        A W           24           2
## 3 21400899 MAR 04, 2015 - CHA @ BKN        A W           24           3
## 4 21400899 MAR 04, 2015 - CHA @ BKN        A W           24           4
## 5 21400899 MAR 04, 2015 - CHA @ BKN        A W           24           5
## 6 21400899 MAR 04, 2015 - CHA @ BKN        A W           24           6
##   PERIOD GAME_CLOCK SHOT_CLOCK DRIBBLES TOUCH_TIME SHOT_DIST PTS_TYPE
## 1      1       1:09       10.8        2        1.9       7.7        2
## 2      1       0:14        3.4        0        0.8      28.2        3
## 3      1       0:00         NA        3        2.7      10.1        2
## 4      2      11:47       10.3        2        1.9      17.2        2
## 5      2      10:34       10.9        2        2.7       3.7        2
## 6      2       8:15        9.1        2        4.4      18.4        2
##   SHOT_RESULT  CLOSEST_DEFENDER CLOSEST_DEFENDER_PLAYER_ID CLOSE_DEF_DIST
## 1        made    Anderson, Alan                     101187            1.3
## 2      missed Bogdanovic, Bojan                     202711            6.1
## 3      missed Bogdanovic, Bojan                     202711            0.9
## 4      missed     Brown, Markel                     203900            3.4
## 5      missed   Young, Thaddeus                     201152            1.1
## 6      missed   Williams, Deron                     101114            2.6
##   FGM PTS   player_name player_id
## 1   1   2 brian roberts    203148
## 2   0   0 brian roberts    203148
## 3   0   0 brian roberts    203148
## 4   0   0 brian roberts    203148
## 5   0   0 brian roberts    203148
## 6   0   0 brian roberts    203148
```

```
## [1] 128069
```

```
## [1] 21
```

There are 21 variables and 128069 observations in this dataset. 128069 observations means that there are 128069 shots attempted in the dataset.

I will explain some variables for further analysis: LOCATION: H means home team, A means away team W: W means win, L means lose FINAL_MARGIN: final score difference at the end of game SHOT_DIST: shotting distance from basket PTS_TPYE: two-pointer or three pointer(no free throw included)

The first goal of my project is to visualize top 5 score players in the dataset

After deeply looking at this table, I decide to create a new table which is grouped by player and this can make it easier to analyse

```
## # A tibble: 6 x 4
##   player_name     shots_num made_num points
##   <fct>               <int>    <int>  <int>
## 1 aaron brooks          561      233    555
## 2 aaron gordon          104       55    119
## 3 al farouq aminu       258      111    248
## 4 al horford            715      387    783
## 5 al jefferson          800      382    766
## 6 alan anderson         337      146    352
```

```
## # A tibble: 10 x 4
##    player_name       shots_num made_num points
##    <fct>                 <int>    <int>  <int>
##  1 stephen curry           968      470   1130
##  2 james harden           1054      474   1103
##  3 klay thompson           971      449   1075
##  4 lebron james            978      478   1041
##  5 mnta ellis             1052      473   1018
##  6 kyrie irving            942      439    998
##  7 damian lillard          986      426    995
##  8 lamarcus aldridge      1050      473    971
##  9 nikola vucevic          902      480    962
## 10 chris paul              885      425    947
```

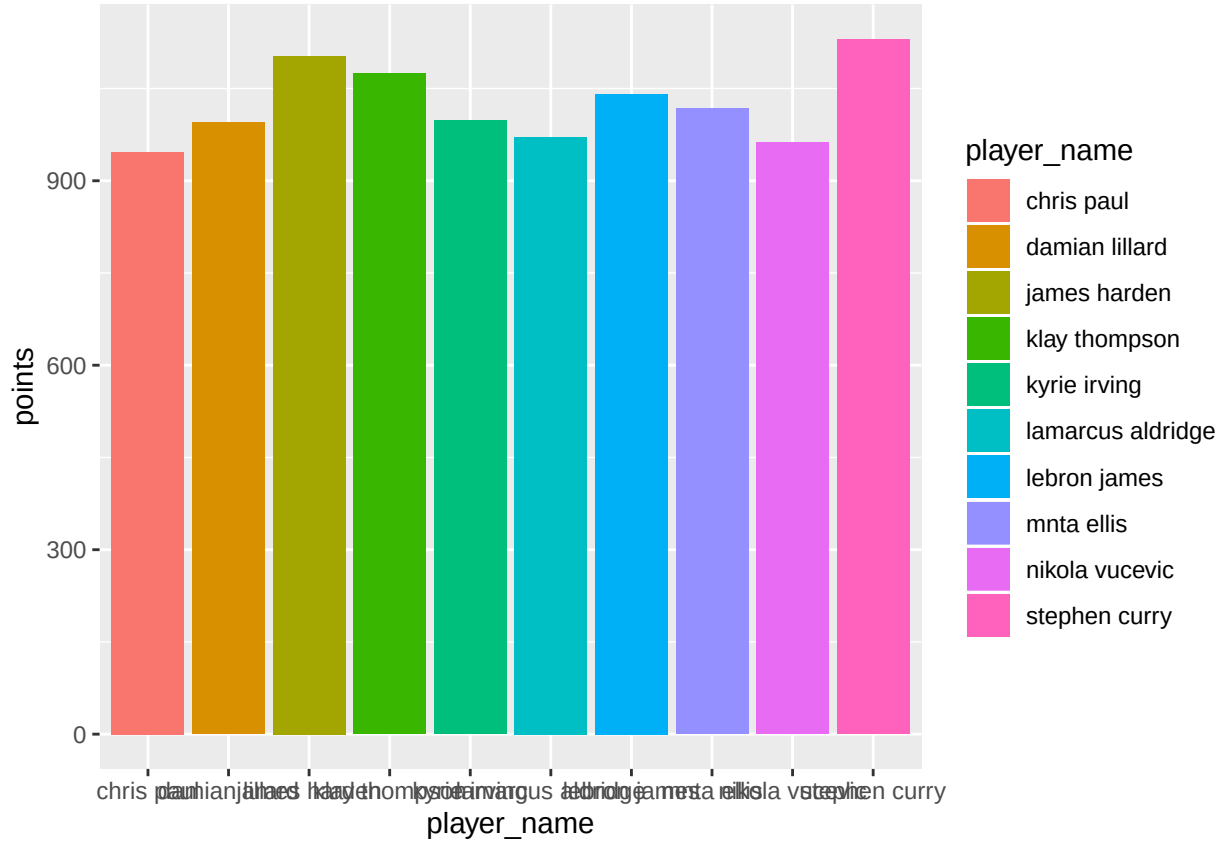
The players table above is for top 5 score in the dataset I found that the highest total score is 1130 which is from player Stephen Curry, and I realized that the data from kaggle dataset is not from a whole season. To test my thoughts, I calculated the number of unique game_ID.

```
## [1] 904
```

As we know that there is 30 teams in NBA and they all need to play 82 games for each regular season,and I calculated the total number of games for the whole season.

```
## [1] 1230
```

So the data does not contain the whole season.



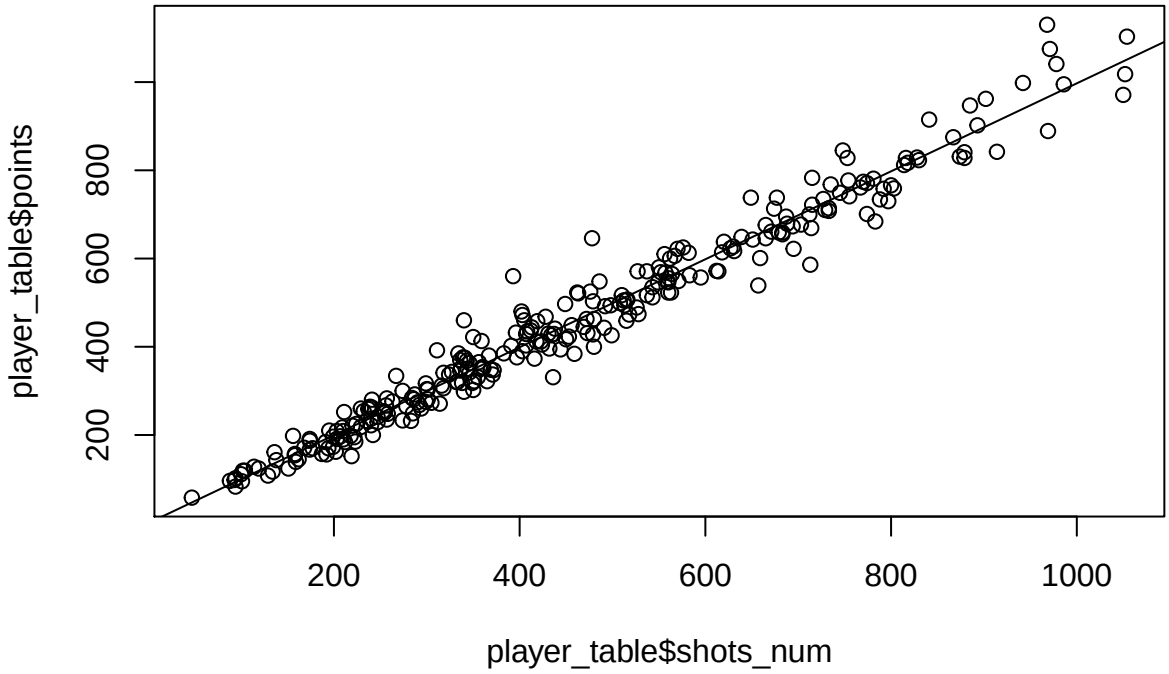
We can conclude that these 10 players' scoring abilities are similar in season 2014. In these 10 players, only Lebron James aldridge and vucevic are not guards, and all the other 7 players are guards. We can then conclude that small players are more likely to get higher score among excellent players.

My second goal is to analyse that whether there is a strong linear relationship between total score and number of shots. So we can set shots_num as the predictor and set points as the response

```
##
## Call:
## lm(formula = player_table$points ~ player_table$shots_num, data = player_table)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -125.019  -22.961   -3.369   17.396  169.290
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.11594    5.41983   0.021    0.983
## player_table$shots_num  0.99706    0.01062  93.876   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.86 on 279 degrees of freedom
## Multiple R-squared:  0.9693, Adjusted R-squared:  0.9692
## F-statistic:  8813 on 1 and 279 DF,  p-value: < 2.2e-16
```

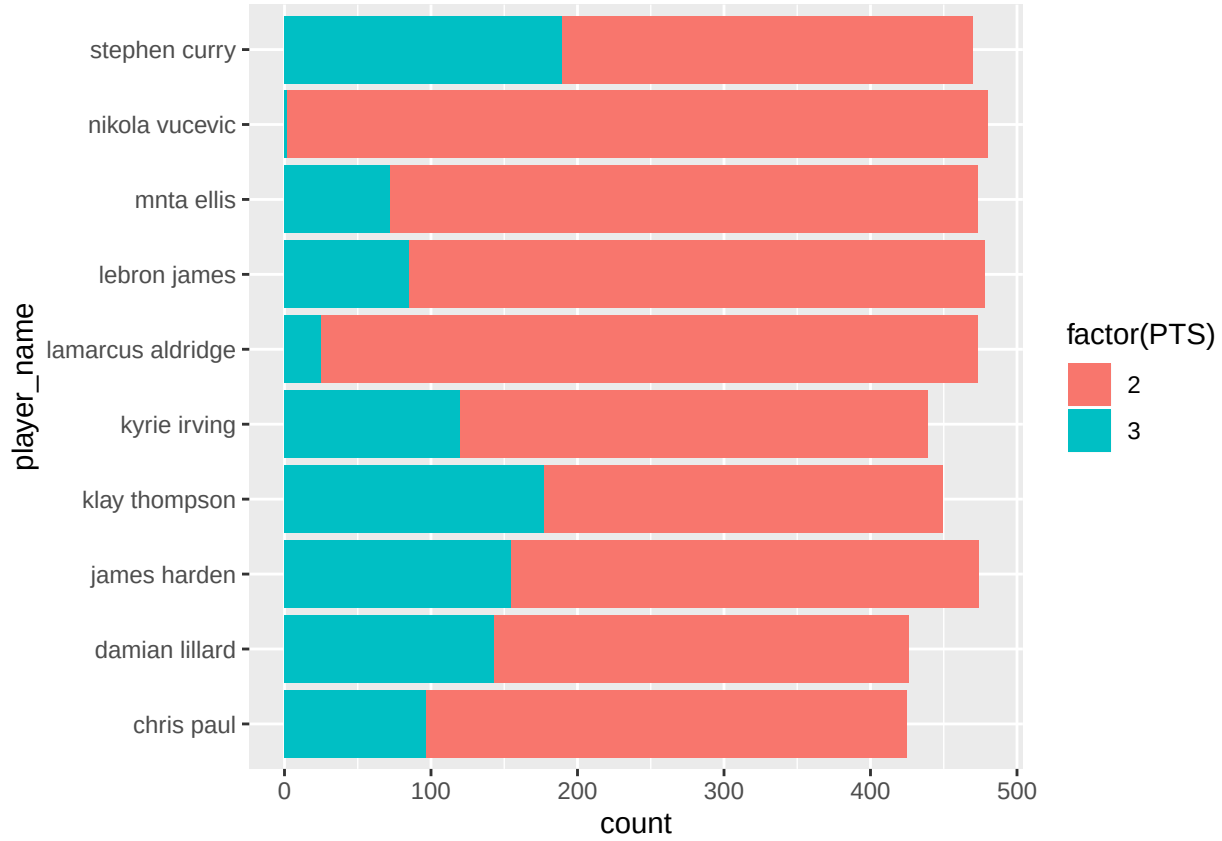
As we can see that the R-squared is 0.9693, which means the model can explain 96.93% data.

So we can conclude that there is a strong linear relationship between number of shots and score.



We can also get the same conclusion from this graph, as all the observations are close to the linear model.

Next,I want to know the ratio of 3-pointer and 2-pointer for the top10 players mentioned before



We can see that the ratio of made 2-pointer/3pointer for all the top 10 players are more than 1. Stephen Curry and Klay Thompson are more balanced between 2-pointer and 3-pointer

My second data source is web-scraped from the website https://www.espn.com/nba/player.

The table below contains the statistics about player "Lebron James" for his current season (first row) and his regular season career total (second row).

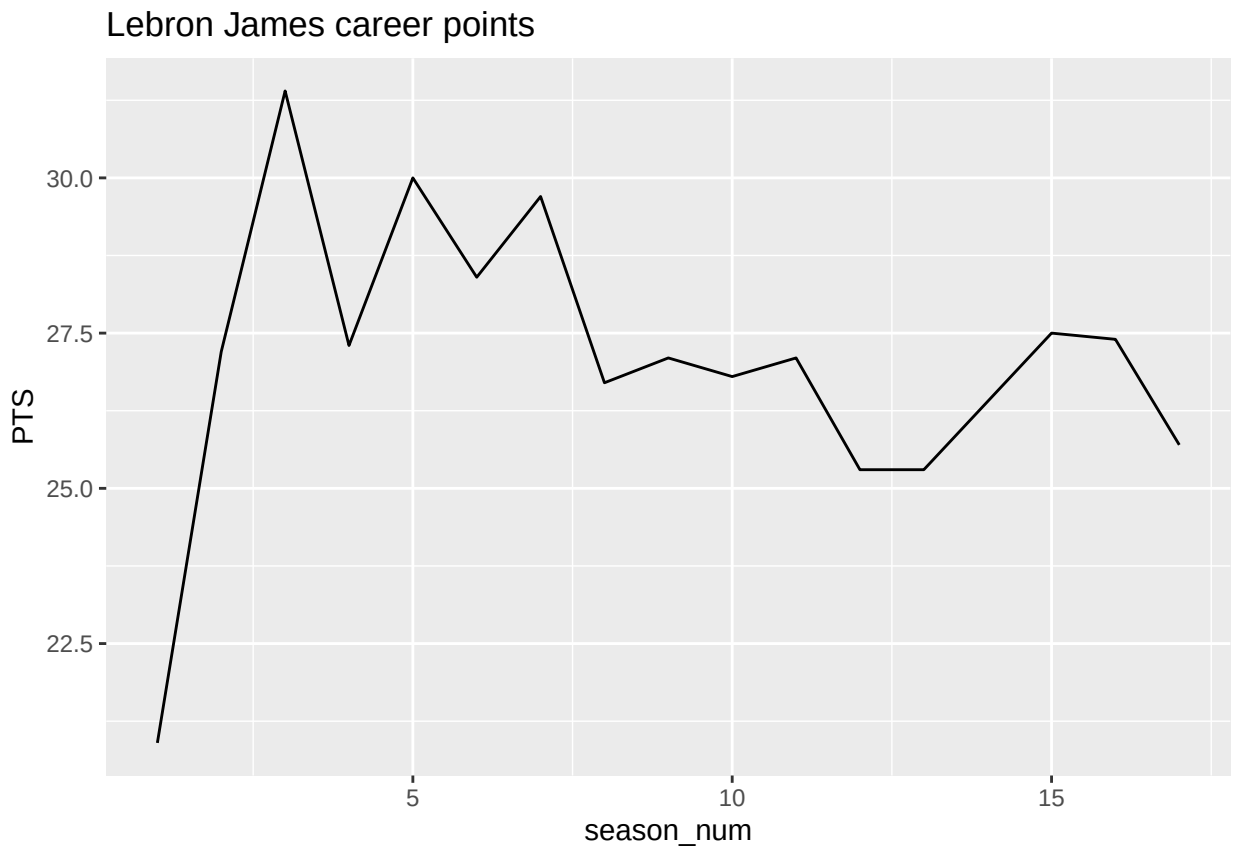
```
##      GP  MIN  FG%  3P%  FT% REB  AST BLK STL  PF  TO  PTS
## 1   60 34.9 49.8 34.9 69.7 7.9 10.6 0.5 1.2 1.8 4.0 25.7
## 2 1258 38.4 50.4 34.4 73.5 7.4  7.4 0.8 1.6 1.8 3.5 27.1
```

As this season is paused due to coronavirus, He played 60 games this season

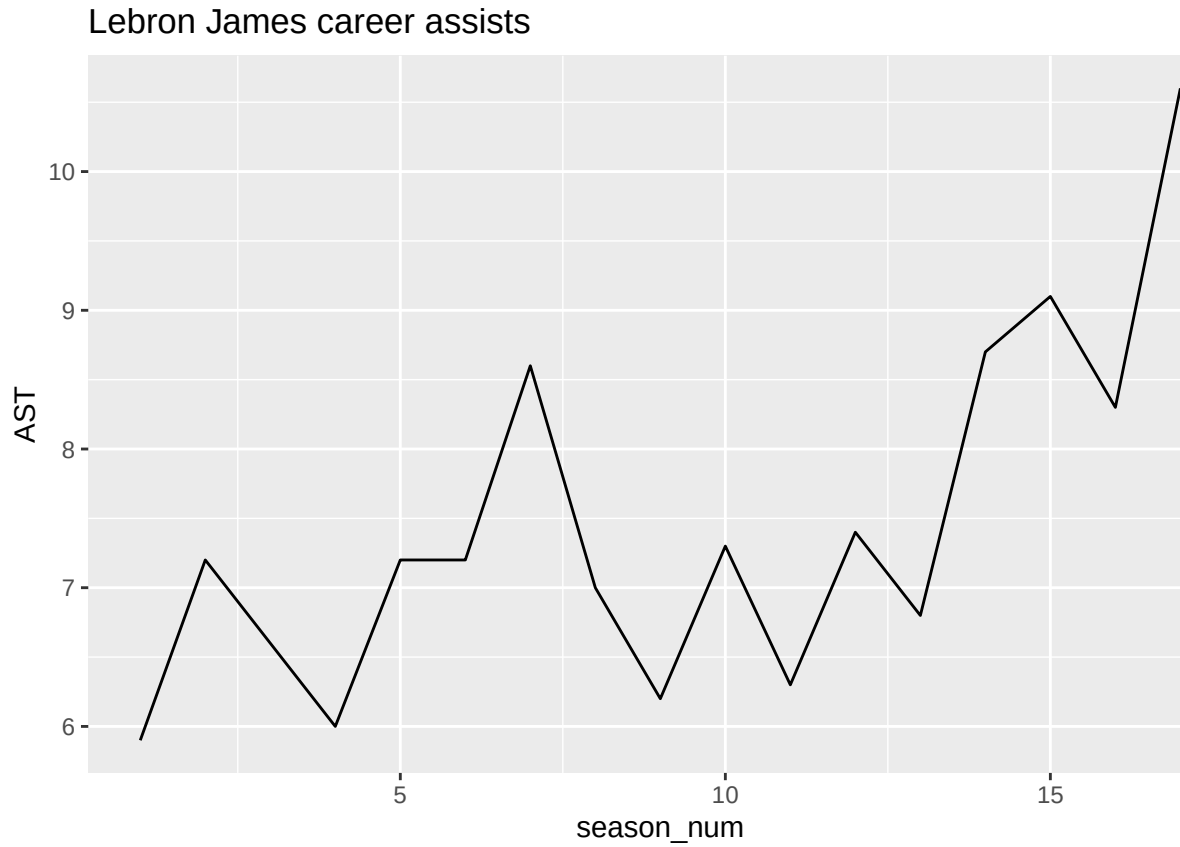
The table below is the first 5 rows of detail information for Lebron James for his 17 seasons, and the first row is for season 2003-2004.

```
##   GP GS  MIN        FG  FG%     3PT  3P%       FT  FT%  OR  DR REB AST BLK
## 1 79 79 39.5  7.9-18.9 41.7 0.8-2.7 29.0  4.4-5.8 75.4 1.3 4.2 5.5 5.9 0.7
## 2 80 80 42.4  9.9-21.1 47.2 1.4-3.9 35.1  6.0-8.0 75.0 1.4 6.0 7.4 7.2 0.7
## 3 79 79 42.5 11.1-23.1 48.0 1.6-4.8 33.5 7.6-10.3 73.8 0.9 6.1 7.0 6.6 0.8
## 4 78 78 40.9  9.9-20.8 47.6 1.3-4.0 31.9  6.3-9.0 69.8 1.1 5.7 6.7 6.0 0.7
## 5 75 74 40.4 10.6-21.9 48.4 1.5-4.8 31.5 7.3-10.3 71.2 1.8 6.1 7.9 7.2 1.1
##   STL  PF  TO  PTS
## 1 1.6 1.9 3.5 20.9
## 2 2.2 1.8 3.3 27.2
## 3 1.6 2.3 3.3 31.4
```

```
## 4 1.6 2.2 3.2 27.3
## 5 1.8 2.2 3.4 30.0
```



We can see that Lebron James's scoring ability is significantly increasing in his first 5 seasons, and declines a little bit for next 5 years and be stable until now. It is very amazing to see a 35 year-old man's average score is



higher than 25 points.

There is an interesting thing that Lebron James's assists is higher and higher as he becomes old. As he's average score being stable(conclusion from last graph), his ability to help teammates is even higher for a 35-year-old man!

I like to compare some statistics between Lebron James and Stephen Curry, so I also scraped the data for Curry.

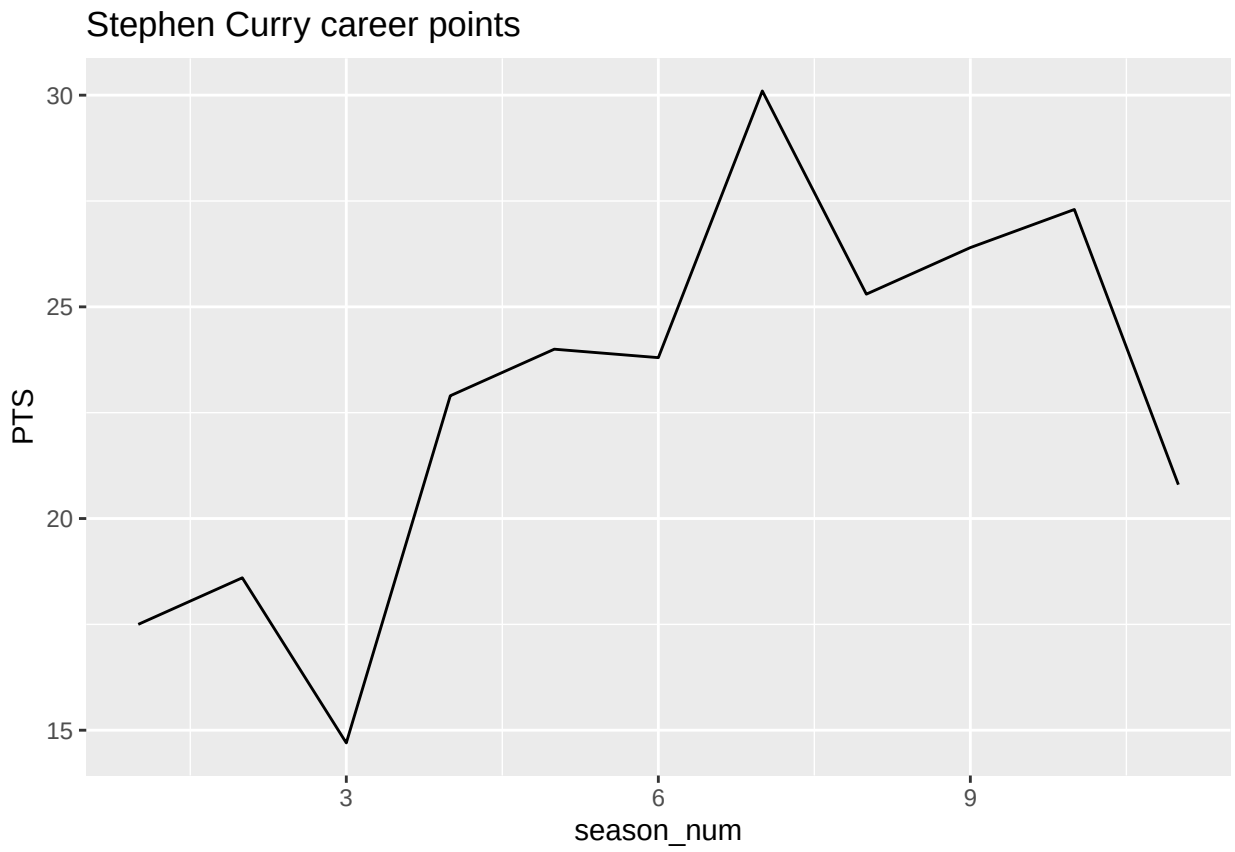
```
##    GP  MIN  FG%  3P%   FT% REB AST BLK STL  PF  TO  PTS
## 1   5 27.8 40.2 24.5 100.0 5.2 6.6 0.4 1.0 2.2 3.2 20.8
## 2 699 34.3 47.6 43.5  90.6 4.5 6.6 0.2 1.7 2.5 3.1 23.5
```

Due to injury, he only played 5 games this season. So it is not reasonable to compare the data for current season.

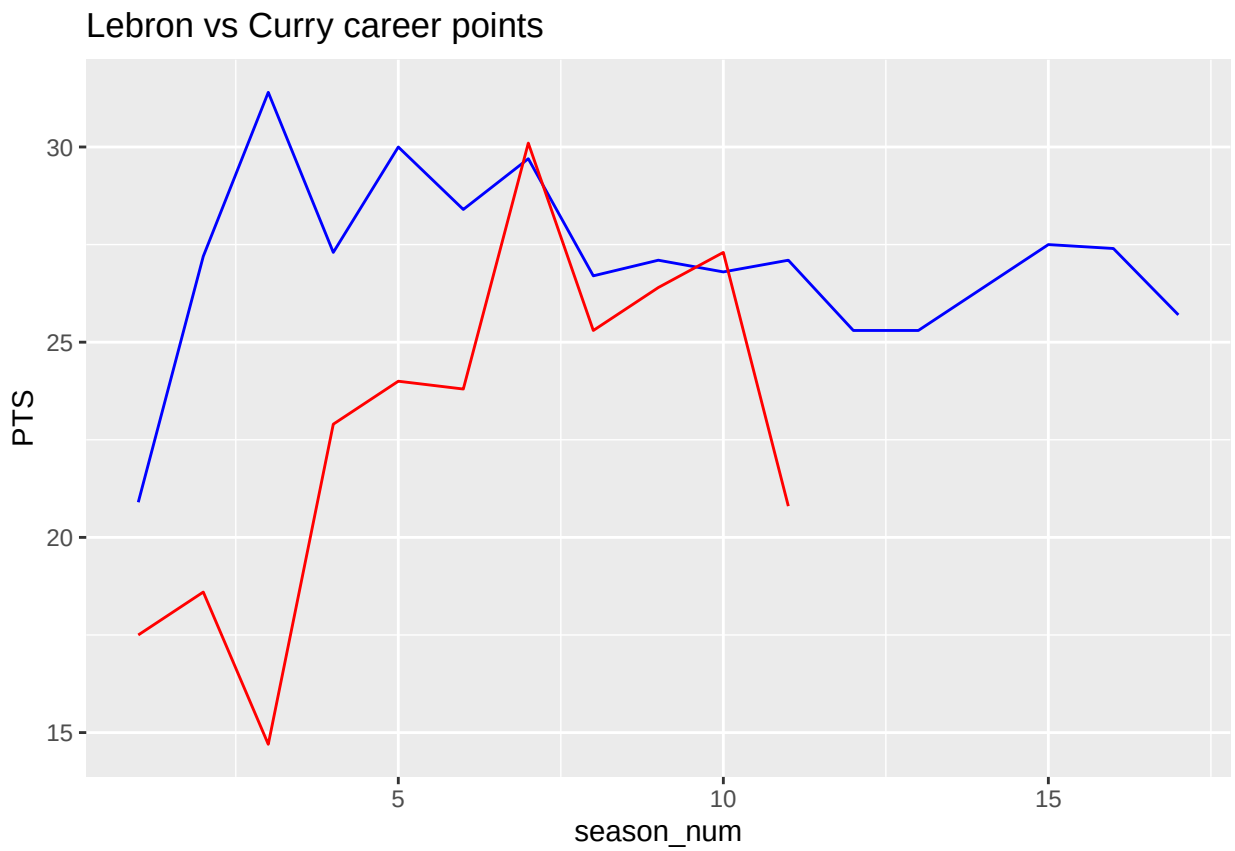
The table below is the first 5 rows of detail information for Curry for his 11 seasons, and the first row is for season 2009-2010.

```
##   GP GS  MIN       FG  FG%     3PT  3P%      FT  FT%  OR  DR REB AST BLK
## 1 80 77 36.2 6.6-14.3 46.2 2.1-4.8 43.7 2.2-2.5 88.5 0.6 3.9 4.5 5.9 0.2
## 2 74 74 33.6 6.8-14.2 48.0 2.0-4.6 44.2 2.9-3.1 93.4 0.7 3.2 3.9 5.8 0.3
## 3 26 23 28.2 5.6-11.4 49.0 2.1-4.7 45.5 1.5-1.8 80.9 0.6 2.8 3.4 5.3 0.3
## 4 78 78 38.2 8.0-17.8 45.1 3.5-7.7 45.3 3.4-3.7 90.0 0.8 3.3 4.0 6.9 0.2
## 5 78 78 36.5 8.4-17.7 47.1 3.3-7.9 42.4 3.9-4.5 88.5 0.6 3.7 4.3 8.5 0.2
##   STL  PF  TO  PTS
## 1 1.9 3.2 3.1 17.5
## 2 1.5 3.1 3.1 18.6
## 3 1.5 2.4 2.5 14.7
```

```
## 4 1.6 2.5 3.1 22.9
## 5 1.6 2.5 3.8 24.0
```



We can see that Curry's points is decreasing for the recent years, and this is because of his injury.People usually regard him as a historical scorer, however, half of his points data is below 25 points. And lebron James's points data are all above 25 points except his rookie season.



We can see that after Curry's 6th season, his scoring ability is almost same as Lebron corespondingly. However,the effect of his injury is severe to his points data. I hope he can recover as soon as possible and I hope the NBA will restart as soon as possible when we overcome the coronavirus.