Final Report

Hongye Xu

2019/5/5

1 Introduction

1.1 Background

I have always felt that the NBA has the best data storage in the sport filed. In the beginning, I wanted to analyze the performance of the players by scrapping the data from the official NBA.stat website. However, since the NBA.stat table is in javascript format, and the official has canceled all the existing official APIs, no possible R-based crawler method has been found after the effort. Therefore, I chose an alternative, which is the basketball-reference website. This report is based on two data sources on the basketball-reference. My goal is to predict the player's salary for next season based on player performance this season.

1.2 Glossary

Abbreviation	Explanation
Pos	Position
Age	Age of Player at the start of February 1st of that season
Tm	Team
G	Games
GS	Games Started
MP	Minutes Played Per Game
FG	Field Goals Per Game
FGA	Field Goal Attempts Per Game
FG%	Field Goal Percentage
3P	3-Point Field Goals Per Game
3PA	3-Point Field Goal Attempts Per Game
3P%	FG% on 3-Pt FGAs.

2P	2-Point Field Goals Per Game
2PA	2-Point Field Goal Attempts Per Game
eFG%	Effective Field Goal Percentage
FT	Free Throws Per Game
FTA	Free Throw Attempts Per Game
FT%	Free Throw Percentage
ORB	Offensive Rebounds Per Game
DRB	Defensive Rebounds Per Game
TRB	Total Rebounds Per Game
AST	Assists Per Game
STL	Steals Per Game
BLK	Blocks Per Game
TOV	Turnovers Per Game
PF	Personal Fouls Per Game
PTS	Points Per Game

Github Link

https://github.com/szxuhongye/NBA-Player-Salary-Predicton.git (https://github.com/szxuhongye/NBA-Player-Salary-Predicton.git)

2 Preparation

2.1 Required Packages

```
library(rvest)
library(magrittr)
library(tibble)
library(dplyr)
library(stringr)
library(data.table)
library(corrplot)
library(GGally)
library(tidyverse)
library(PerformanceAnalytics)
library(plotly)
library(caret)
library(MASS)
```

2.2 Data Scraping and Cleaning

2.2.1 Players' Regular Season Data

	Player <chr>	Pos <chr>	Age <dbl>	Tm <chr>	G <dbl>	GS <dbl>	MP <dbl>	FG <dbl>	FGA <dbl>
1	Alex Abrines	SG	25	OKC	31	2	19.0	1.8	5.1
2	Quincy Acy	PF	28	PHO	10	0	12.3	0.4	1.8
3	Jaylen Adams	PG	22	ATL	34	1	12.6	1.1	3.2
4	Steven Adams	C	25	OKC	80	80	33.4	6.0	10.1
5	Bam Adebayo	C	21	MIA	82	28	23.3	3.4	5.9
6	Deng Adel	SF	21	CLE	19	3	10.2	0.6	1.9

6 rows | 1-10 of 30 columns

2.2.2 Scale

Considering that regression analysis is mainly used in this report, i try to scale some features.

	Age <dbl>	MP <dbl>	2P% <dbl>	3P% <dbl>	FT% <dbl>	TRB <dbl>
Alex Abrines	-0.2348656	-0.1678429	-0.05111731	0.1079674	1.30497350	-0.9119870
Quincy Acy	0.4772268	-0.9471834	2.01348589	-1.5619660	-0.28148044	-0.4970106
Jaylen Adams	-0.9469579	-0.9122876	-1.76955950	0.2398043	0.27342273	-0.7874941

Steven Adams	-0.2348656	1.5071577	1.13572046	-2.7309194	-1.70430908	2.4078238
Bam Adebayo	-1.1843221	0.3323309	1.03681731	-0.9730947	-0.03248542	1.4948758
Deng Adel	-1.1843221	-1.1914543	-1.47285006	-0.4369582	1.85276253	-1.1194751

6 rows | 1-8 of 13 columns

2.2.3 Players' Salaries

	Player <chr>	...	2018-19 <chr>	2019-20 <chr>	2020-21 <chr>	2021-22 <chr>	2022-23 <chr>
1	Stephen Curry	GSW	$37,457,154	$40,231,758	$43,006,362	$45,780,966	
2	Chris Paul	HOU	$35,654,150	$38,506,482	$41,358,814	$44,211,146	
3	Russell Westbrook	OKC	$35,654,150	$38,178,000	$41,006,000	$43,848,000	$46,662,000
4	LeBron James	LAL	$35,654,150	$37,436,858	$39,219,565	$41,002,273	
5	Blake Griffin	DET	$32,088,932	$34,234,964	$36,595,996	$38,957,028	
6	Gordon Hayward	BOS	$31,214,295	$32,700,690	$34,187,085		

6 rows | 1-9 of 11 columns

```
## Warning in function_list[[k]](value): NAs introduced by coercion
```

Player <chr>	...	2018-19 <chr> <dbl>	2019-20 <chr>	2020-21 <chr>	2021-22 <chr>	2022-23 <chr>	2023 <chr
Stephen Curry	GSW	37457154	$40,231,758	$43,006,362	$45,780,966		
Chris Paul	HOU	35654150	$38,506,482	$41,358,814	$44,211,146		
Russell Westbrook	OKC	35654150	$38,178,000	$41,006,000	$43,848,000	$46,662,000	
LeBron James	LAL	35654150	$37,436,858	$39,219,565	$41,002,273		
Blake Griffin	DET	32088932	$34,234,964	$36,595,996	$38,957,028		
Gordon Hayward	BOS	31214295	$32,700,690	$34,187,085			

6 rows | 1-8 of 10 columns

0 rows

Player	Tm	2018-19

	<chr>	<chr>	<dbl>
1	Stephen Curry	GSW	37457154
2	Chris Paul	HOU	35654150
3	Russell Westbrook	OKC	35654150
4	LeBron James	LAL	35654150
5	Blake Griffin	DET	32088932
6	Gordon Hayward	BOS	31214295

6 rows

Finally, there is no duplicate player data.

2.2.4 Merging Data

	Player <chr>	**Pos** <chr>	**Age** <dbl>	**Tm.x** <chr>	**G** <dbl>	**GS** <dbl>	**MP** <dbl>	**FG** <dbl>	**FGA** <dbl>
1	Aaron Gordon	PF	23	ORL	78	78	33.8	6.0	13.4
2	Aaron Holiday	PG	22	IND	50	0	12.9	2.1	5.2
3	Abdel Nader	SF	25	OKC	61	1	11.4	1.5	3.5
4	Al Horford	C	32	BOS	68	68	29.0	5.7	10.6
5	Al-Farouq Aminu	PF	28	POR	81	81	28.3	3.2	7.3
6	Alec Burks	SG	27	TOT	64	24	21.5	3.0	7.4

6 rows | 1-10 of 31 columns

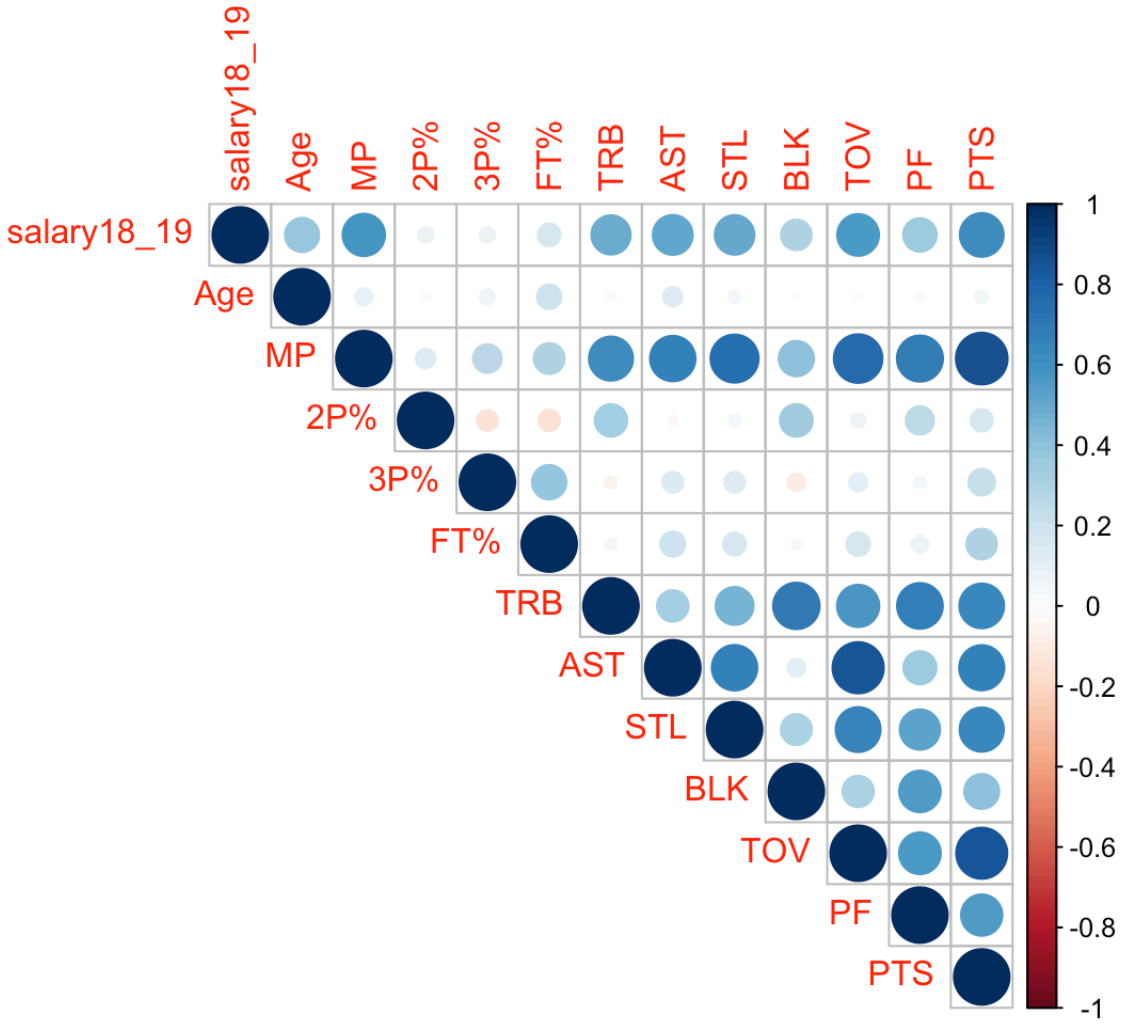
	Row.names <S3: AsIs>	**Age** <dbl>	**MP** <dbl>	**2P%** <dbl>	**3P%** <dbl>	**FT%** <dbl>	
1	Aaron Gordon	-0.7095938	1.5536855	-0.0634802	0.33648464	-0.06094200	1.536373
2	Aaron Holiday	-0.9469579	-0.8773917	-0.5579959	0.24859341	0.57221675	-0.994982
3	Abdel Nader	-0.2348656	-1.0518710	0.1096003	0.08160007	0.07422672	-0.745996
4	Al Horford	1.4266833	0.9953519	1.2346236	0.43316500	0.57933089	1.245890
5	Al-Farouq Aminu	0.4772268	0.9139283	0.1219632	0.28374990	0.90658148	1.57787
6	Alec Burks	0.2398627	0.1229558	-0.9412456	0.45953237	0.59355918	0.000960

6 rows | 1-8 of 15 columns

2.2.5 Save No_scale Data Into CSV

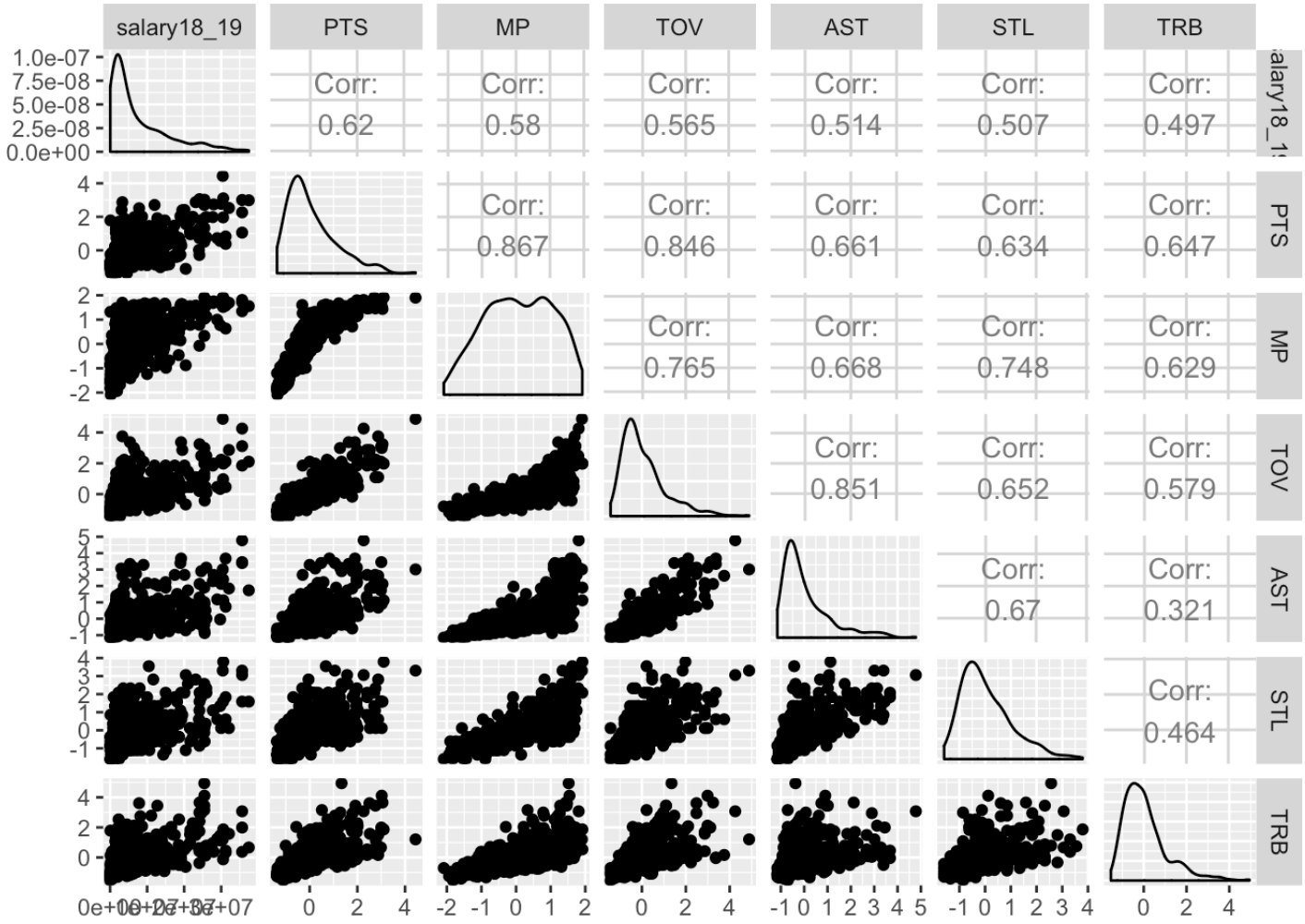
3 Correlation Check

3.1 Frist Check



The features that have strong correlation with salary are:PTS,TOV,STL,AST,TRB and MP. Besides, MP is strongly correlated with multiple features and may have multiple collinearities(This is in line with our common sense. The more time we play, the better the data will be). What I didn't expect was that the correlation between field goal and salary was not high, that is to say, the output of players influenced the salary of players more than efficiency.

3.2 Second Check




```
## salary18_19         PTS          MP         TOV         AST         STL
##   1.0000000   0.6198192   0.5803967   0.5645525   0.5142283   0.5066299
##         TRB
##   0.4972563
```

Correlation strength is: PTS > MP > TOV > AST > STL > TRB There's also one thing that surprises me: the number of players'turnivers is positively correlated with their salaries. I mean, generally speaking, assuming that a player's turnover rate is constant, the total number of turnovers will increase as his minutes played increases, and important players will have higher minutes played and higher salaries.

4 Data Visualization

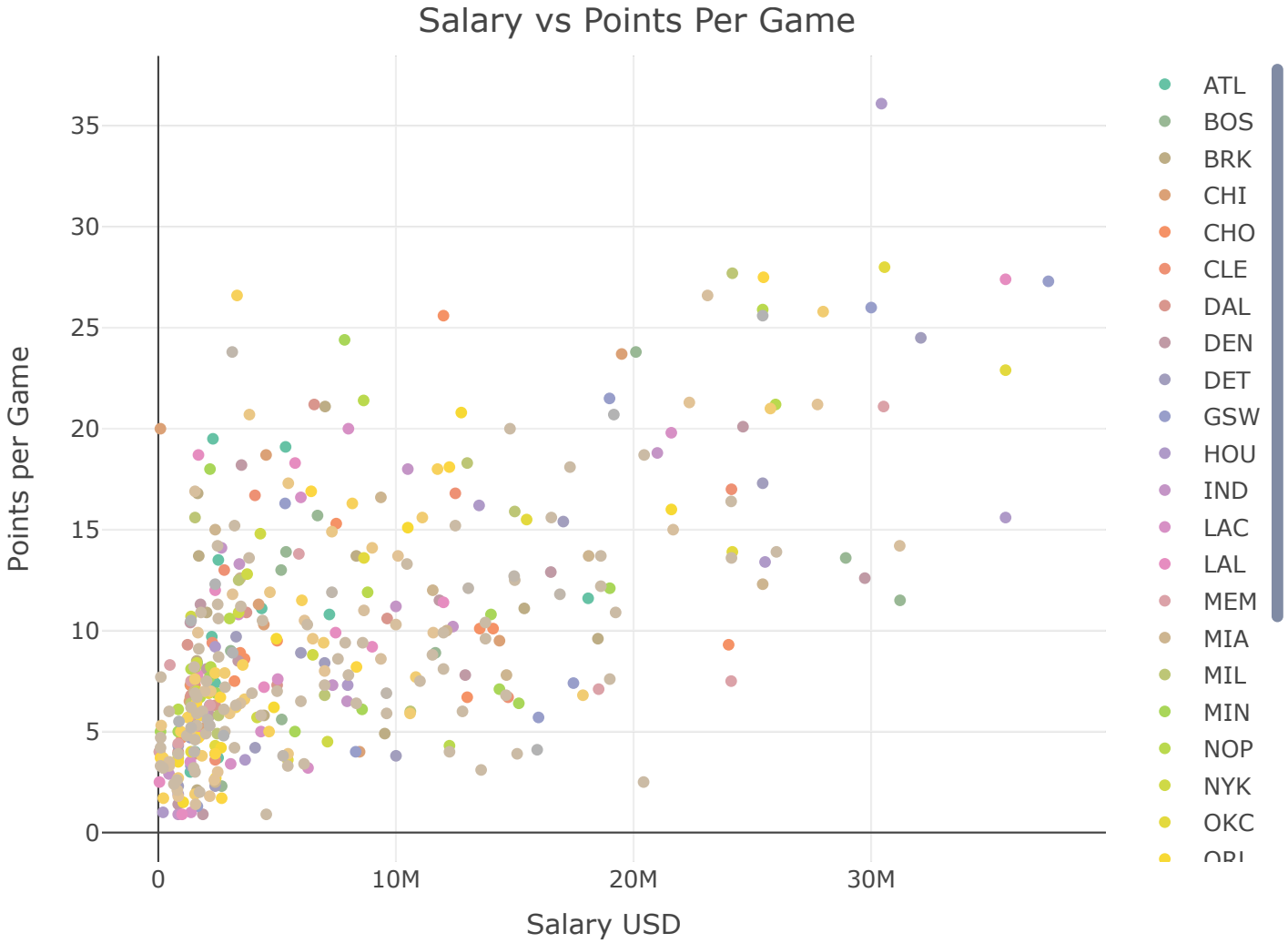
4.1 Interactive Plot

```
## No trace type specified:
##   Based on info supplied, a 'scatter' trace seems appropriate.
##   Read more about this trace type -> https://plot.ly/r/reference/#scatter
```

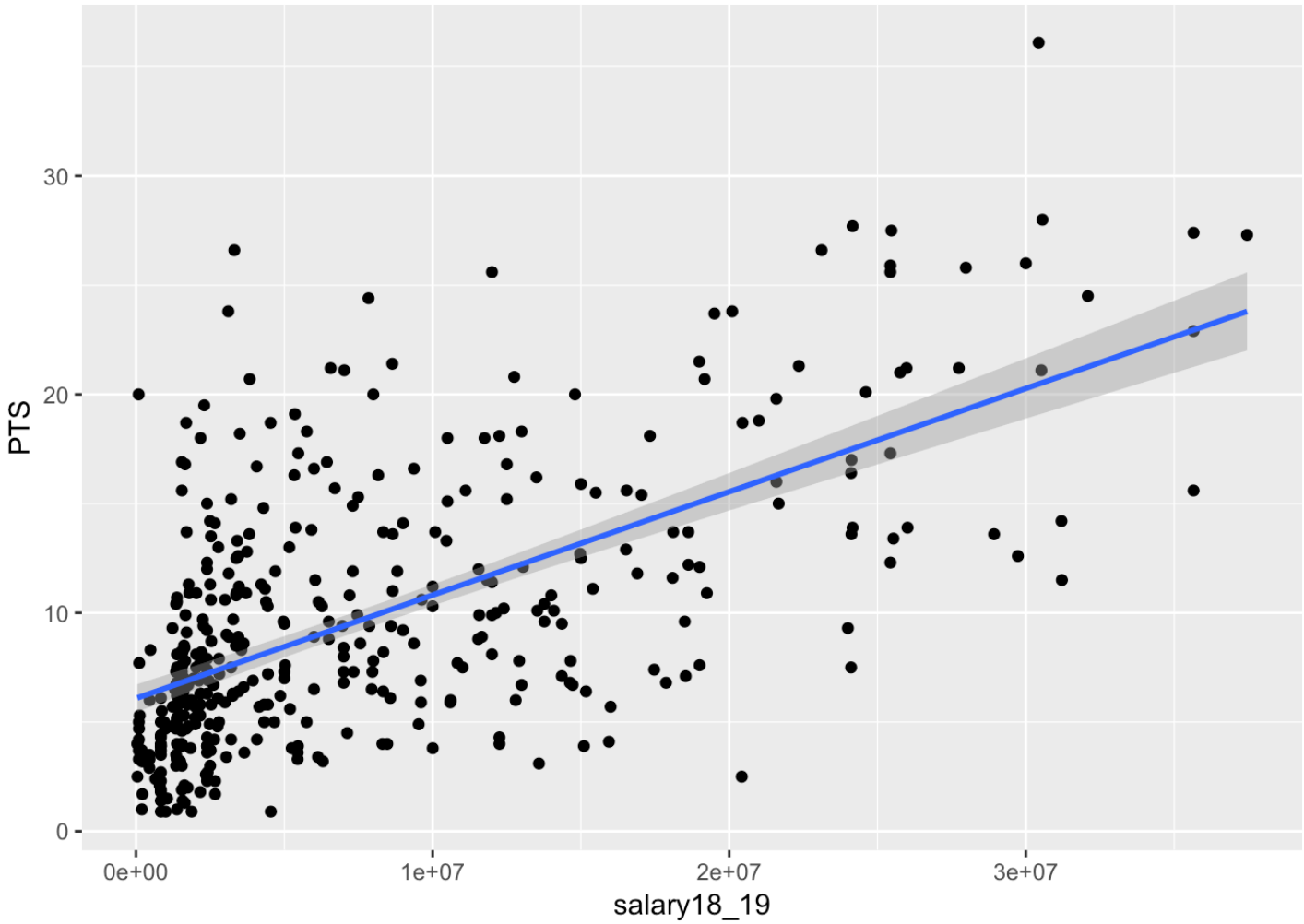
```
## No scatter mode specifed:
##   Setting the mode to markers
##   Read more about this attribute -> https://plot.ly/r/reference/#scatter-mode
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): n too large, allowed maximum for p
alette Set2 is 8
## Returning the palette you asked for with that many colors

## Warning in RColorBrewer::brewer.pal(N, "Set2"): n too large, allowed maximum for p
alette Set2 is 8
## Returning the palette you asked for with that many colors
```



4.2 Scatter Plot With Regression Line



Under the simple linear model, we can understand that the fitted curve represents the average level of the league, and the player below the curve performs worse than the expected performance corresponding to the salary. We can check their name by hovering on the points in the interactive plot. It includes a lot of All-Star players, such as Chris Paul, Kyle Lowry, Al Horford, Gordon Haywood(The Celtics are unlucky) and etc. However, it only considers the scoring feature, and does not fully reflect the players'influence on the field.

5 Multiple Regression

```
##
## Call:
## lm(formula = salary18_19 ~ PTS + MP + TOV + AST + STL + TRB,
##     data = regression)
##
## Coefficients:
## (Intercept)          PTS           MP          TOV          AST
##     7114340      3626578      -546186     -2022044      2571555
##         STL          TRB
##      833962      1888579
```

From here, we can see that points per game is the most significant feature of positive impact, while turnovers per game is the most significant feature of negative impact. However, simple multiple regression also has some problems, that is, there are multiple collinearities.

5.1 Player ’s Importance And Incautiousness

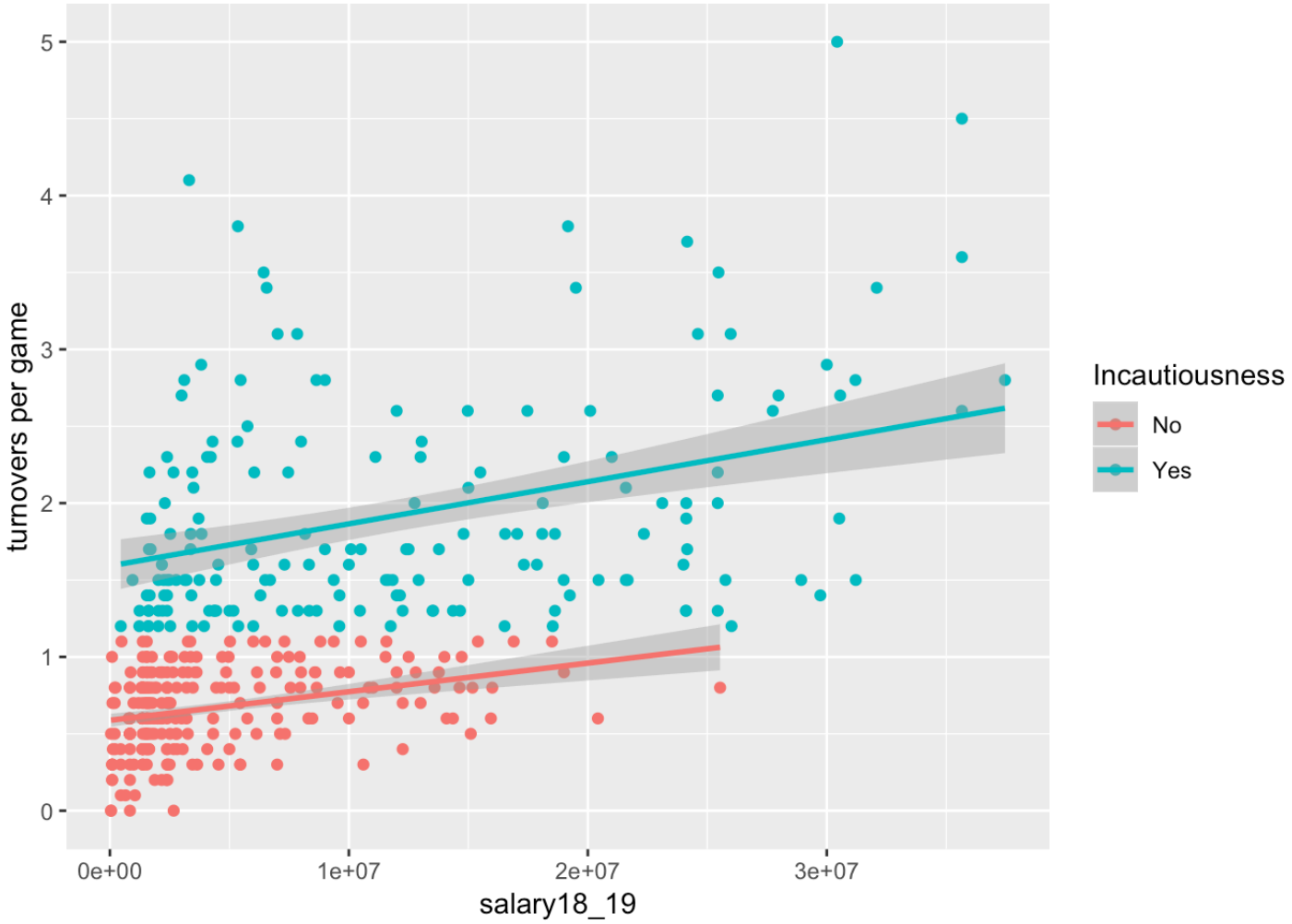
Here we make two definitions that a player is “important” if his minutes played is above average and is “incautious” if his turnover per game is above average.

	salary18_19 <dbl>	PTS <dbl>	MP <dbl>	TOV <dbl>	AST <dbl>	STL <dbl>	TR <dbl
1	21590909	1.12354560	1.5536855	1.2222684	0.90172692	0.1196179	1.536373474
2	1911960	-0.54308294	-0.8773917	-0.4188509	-0.20733649	-0.6151041	-0.994982230
3	1378242	-0.85660712	-1.0518710	-0.9238106	-0.98368087	-0.8600115	-0.745996423
4	28928710	0.72751506	0.9953519	0.4648288	1.17899277	0.6094326	1.245890033
5	6957105	0.03446161	0.9139283	-0.2926109	-0.42914917	0.3645252	1.577871109
6	11536515	-0.06454603	0.1229558	-0.1663710	-0.04097697	-0.1252894	0.000960997

6 rows | 1-9 of 10 columns

5.2 Prallel Slope Model

5.2.1 Incautiousness Comparision



It's true that players with higher salaries make more turnovers. But the tendency is weak.So in fact, turnovers don't have much impact on salaries.

```
##
## Call:
## lm(formula = salary18_19 ~ Importance * Incautiousness, data = regression)
##
## Coefficients:
##                       (Intercept)                      ImportanceYes
##                           3275995                            3754147
##                IncautiousnessYes  ImportanceYes:IncautiousnessYes
##                           3048670                            3017297
```

This shows that when a player is important, he is paid more with fewer turnovers. But the impact is limited.

5.3 Stepwise Regression

```
##
## Call:
## lm(formula = salary18_19 ~ Age + `2P%` + `FT%` + TRB + AST +
##     STL + PF + PTS, data = stepdata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -16726226  -3415620   -329804   2885318  19994129
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7029186     265808  26.445  < 2e-16 ***
## Age          2699715     263103  10.261  < 2e-16 ***
## `2P%`        -498000     300816  -1.655  0.09859 .
## `FT%`        -598141     317671  -1.883  0.06042 .
## TRB          1906249     421719   4.520 8.09e-06 ***
## AST           781799     400336   1.953  0.05151 .
## STL          1151355     403226   2.855  0.00452 **
## PF           -889339     396582  -2.243  0.02546 *
## PTS          3003649     473869   6.339 6.09e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5423000 on 412 degrees of freedom
## Multiple R-squared:  0.5536, Adjusted R-squared:  0.545
## F-statistic: 63.88 on 8 and 412 DF,  p-value: < 2.2e-16
```

nvmax <int>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	RsquaredSD <dbl>	MAESD <dbl>
1	6477031	0.3656265	4841275	740066.8	0.1181143	489025.3
2	6394271	0.3803728	4800648	640458.4	0.1150337	384294.5
3	6113242	0.4295432	4634334	621821.6	0.1299230	442098.5
4	6082951	0.4340008	4636211	662992.5	0.1335307	493985.8
5	6157695	0.4198971	4707174	644375.1	0.1258431	498597.2
6	6142967	0.4211164	4689635	668283.2	0.1298739	503369.9
7	6118851	0.4262623	4685282	678915.2	0.1321880	539508.8
8	6116129	0.4267740	4683816	673996.1	0.1312854	541582.6
9	6116129	0.4267740	4683816	673996.1	0.1312854	541582.6

9 rows

From the result we can see that three-variable model's RMSE is the smallest.

```
## Subset selection object
## 8 Variables  (and intercept)
##                   Forced in Forced out
## PTS                   FALSE      FALSE
## MP                    FALSE      FALSE
## TOV                   FALSE      FALSE
## AST                   FALSE      FALSE
## STL                   FALSE      FALSE
## TRB                   FALSE      FALSE
## ImportanceYes         FALSE      FALSE
## IncautiousnessYes     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: backward
##          PTS MP  TOV AST STL TRB ImportanceYes IncautiousnessYes
## 1  ( 1 ) "*" " " " " " " " " " " "            " "
## 2  ( 1 ) "*" " " " " "*" " " " " "            " "
## 3  ( 1 ) "*" " " " " "*" " " "*" "            " "
## 4  ( 1 ) "*" " " "*" "*" " " "*" "            " "
```

```
## (Intercept)         PTS         AST         TRB
##     7095995     2678780     1764352     1642389
```

The best model is salary18_19 ~ PTS + AST + TRB

6 Conclusion

6.1 What i want to predict

As Lebron James fan, I am concerned about the new contract for the Lakers who may stay next season. Let's find them first.

Player <chr>
Kentavious Caldwell-Pope
Rajon Rondo
Mike Muscala
Lance Stephenson
Reggie Bullock

JaVale McGee
Andre Ingram
Scott Machado

8 rows

Row.names <S3: AsIs>	Age <dbl>	MP <dbl>	2P% <dbl>	3P% <dbl>	FT% <dbl>	
Kentavious Caldwell-Pope	-0.2348656	0.5068101	0.4681242	0.3189064	0.9065815	-0.331

1 row | 1-8 of 14 columns

6.2 Analysis conclusion

```
## [1] "Expected Salary: $6,771,542"
```

```
## [1] "Expected Salary: $7,275,901"
```

```
## [1] "Expected Salary: $5,902,181"
```

So Salaries for Pope, Bullock and Stephenson for next season are $6,771,542, $7,275,901 and $5,902,181