proj2

Jin Ah Kang

April 5, 2019

1990-2014(inclusive), Salaries, Teams

```r
library(DBI)
library(tidyverse)

## -- Attaching packages -- tidyverse 1.2.1 --

## v ggplot2 3.1.0       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(tibble)
library(ggplot2)

db <- dbConnect(RSQLite::SQLite(), "lahman2014.sqlite")
```

Using SQL, write a query to compute the total payroll and winning percentage (number of wins / number of games * 100) for each team (that is, for each teamID and yearID combination). You should include other columns that will help when performing EDA later on (e.g., franchise ids, number of wins, number of games).

P1

```sql
select teams.teamID, teams.yearID, new_salaries.lgID, new_salaries.payroll,
teams.franchID, teams.Rank, W, G, ((w * 1.0 / G) *100) AS winning_percentage
from Teams, (
  select salaries.yearID, salaries.teamID, salaries.lgID, sum(salary) as
payroll
  from salaries
  where salaries.yearID BETWEEN 1990 AND 2014
  group by salaries.teamID, salaries.yearID
) as new_salaries
where teams.w > 0 and teams.G > 0 and new_salaries.teamID = teams.teamID and
new_salaries.yearID = teams.yearID
```

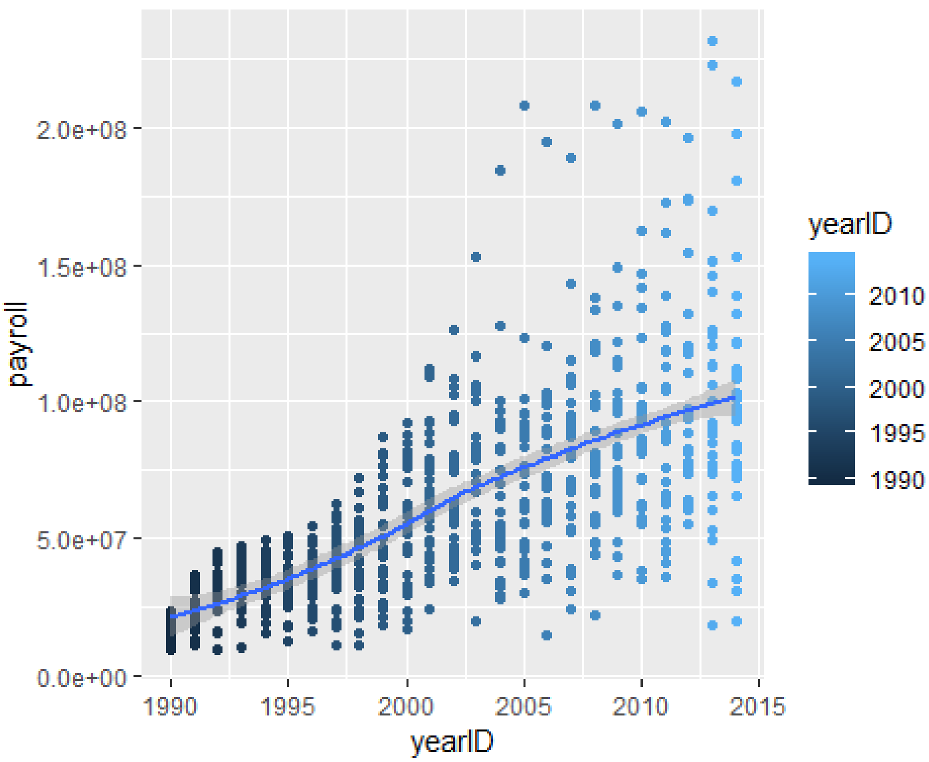
```
total_payroll %>% sample_n(10)
```

```
##    teamID yearID lgID   payroll franchID Rank  W   G winning_percentage
## 1     KCA   1999   AL  26225000      KCR    4 64 161           39.75155
## 2     SDN   2009   NL  43333700      SDP    4 75 162           46.29630
## 3     TOR   1998   AL  51376000      TOR    3 88 163           53.98773
## 4     CHA   1991   AL  16919667      CHW    2 87 162           53.70370
## 5     TEX   2012   AL 120510974      TEX    2 93 162           57.40741
## 6     SDN   1994   NL  14916333      SDP    4 47 117           40.17094
## 7     ATL   1995   NL  47235445      ATL    1 90 144           62.50000
## 8     MIN   2013   AL  75337500      MIN    4 66 162           40.74074
## 9     OAK   2003   AL  50260834      OAK    1 96 162           59.25926
## 10    NYN   1991   NL  32590001      NYM    5 77 161           47.82609
```

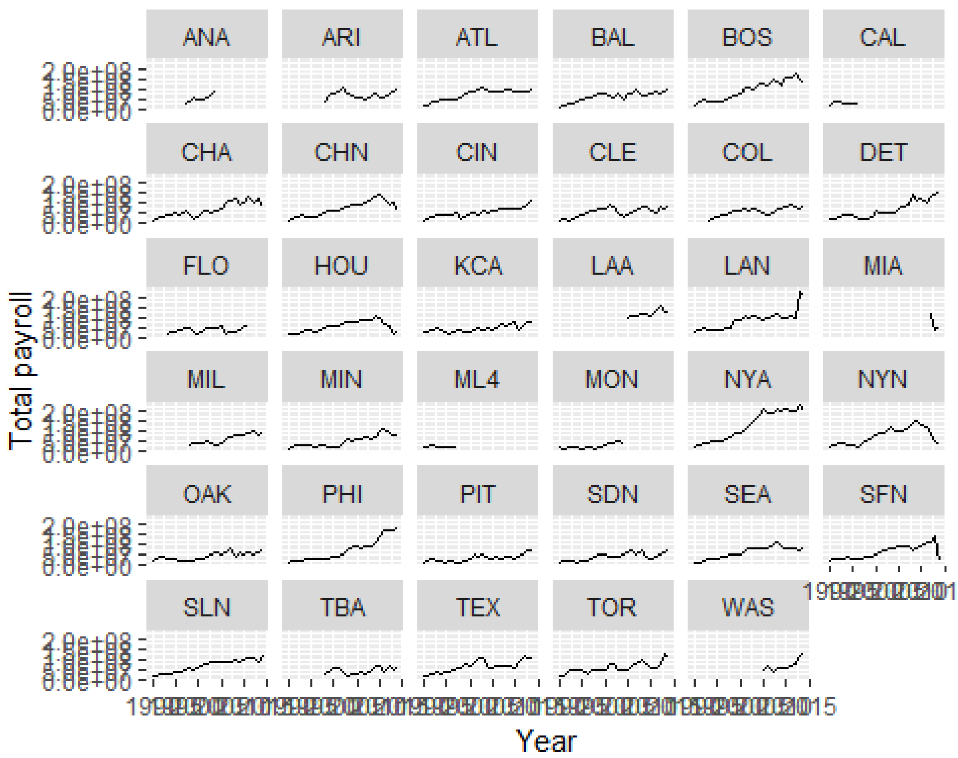
Write code to produce a plot (or plots) that shows the distribution of payrolls across teams conditioned on time (from 1990-2014). Note: you may create a single plot as long as the distributions for each year are clearly distinguishable (e.g., a single plot overlaying histograms is not OK).

P2

```
total_payroll %>%
  filter(yearID >= 1990 && yearID <= 2014) %>%
  ggplot(aes(x = yearID, y = payroll, color = yearID)) +
    geom_point() +
    geom_smooth(method = "loess")
```




```
total_payroll %>%
  filter(yearID >= 1990 && yearID <= 2014) %>%
  ggplot(aes(x = yearID, y = payroll)) +
  geom_line() +
  facet_wrap(~teamID) +
  xlab("Year") +
  ylab("Total payroll")
```



What statements can you make about the distribution of payrolls conditioned on time based on these plots? Remember you can make statements in terms of central tendency, spread, etc.

central trends (mean) spread (variance) skew outliers

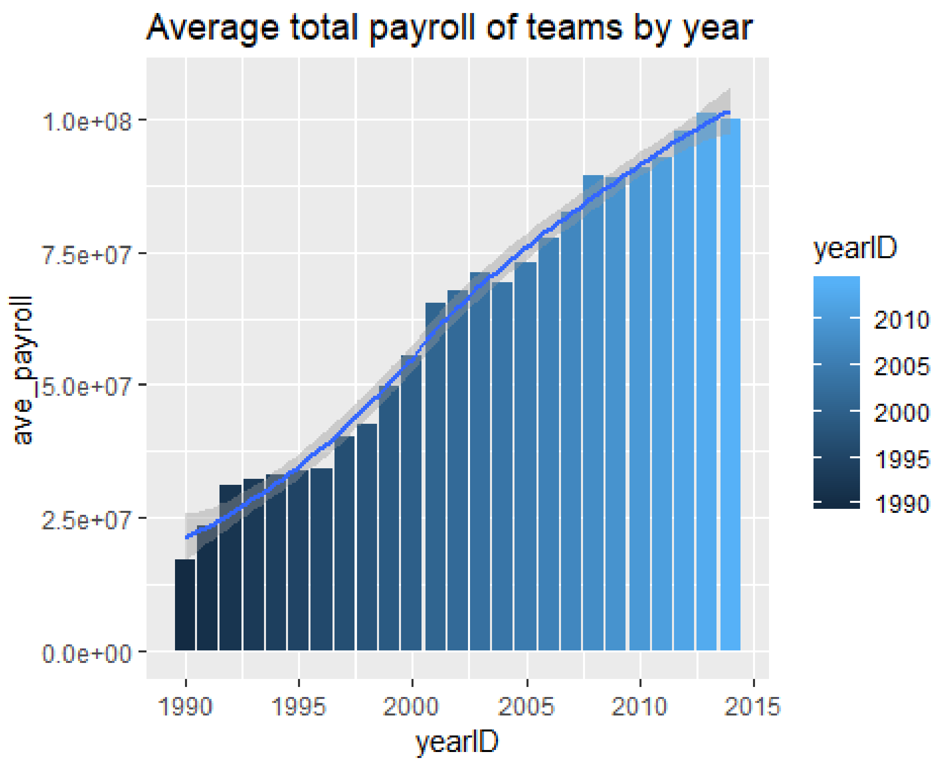
It seems that the average payrolls are increasing over time. The spread of the payroll of the teams also increases among the other teams as time passes. Some teams' payroll become much higher than the other and the gap between two are also increasing.

Write code to produce a plot (or plots) that specifically shows at least one of the statements you made in Question 1. For example, if you make a statement that there is a trend for payrolls to decrease over time, make a plot of a statistic for central tendency (e.g., mean payroll) vs. time to show that specifically.

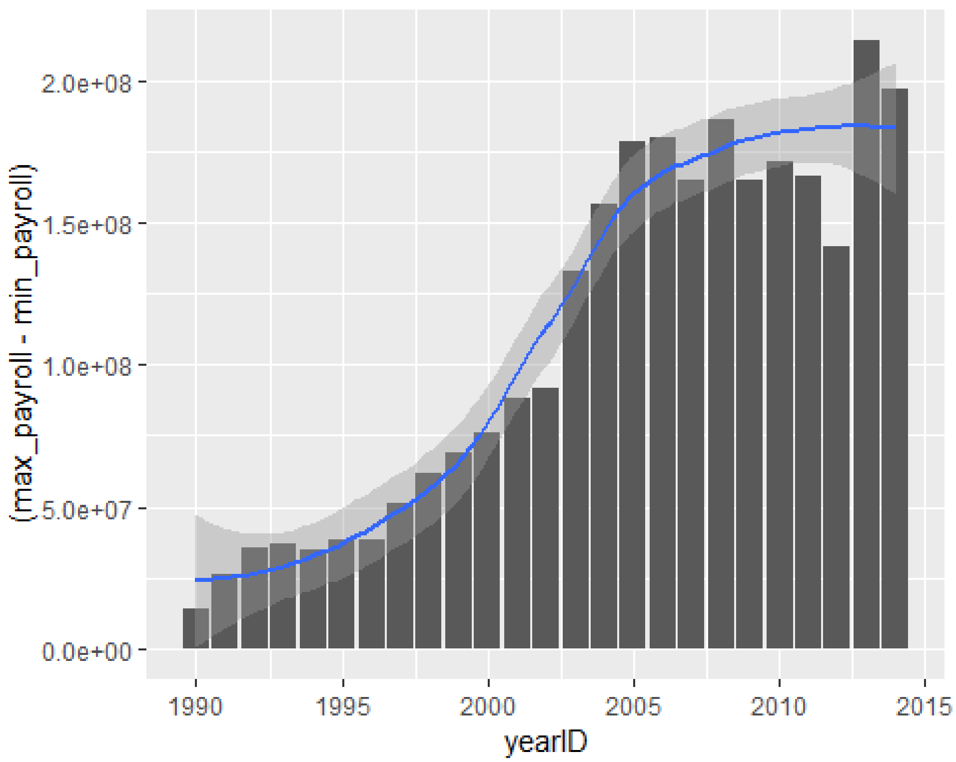
P3

```
total_payroll %>%
  group_by(yearID) %>%
  summarize(ave_payroll = mean(payroll)) %>%
    ggplot(mapping=aes(y = ave_payroll, x = yearID, fill = yearID)) +
    geom_bar(stat = "identity") +
```

```r
  ggtitle("Average total payroll of teams by year") +
  geom_smooth(method = "loess")
```




```r
total_payroll %>%
  group_by(yearID) %>%
  summarise(max_payroll = max(payroll), min_payroll = min(payroll)) %>%
    ggplot(aes(y = (max_payroll-min_payroll), x = yearID)) +
      geom_bar(stat = "identity") +
      geom_smooth(method = "loess")
```

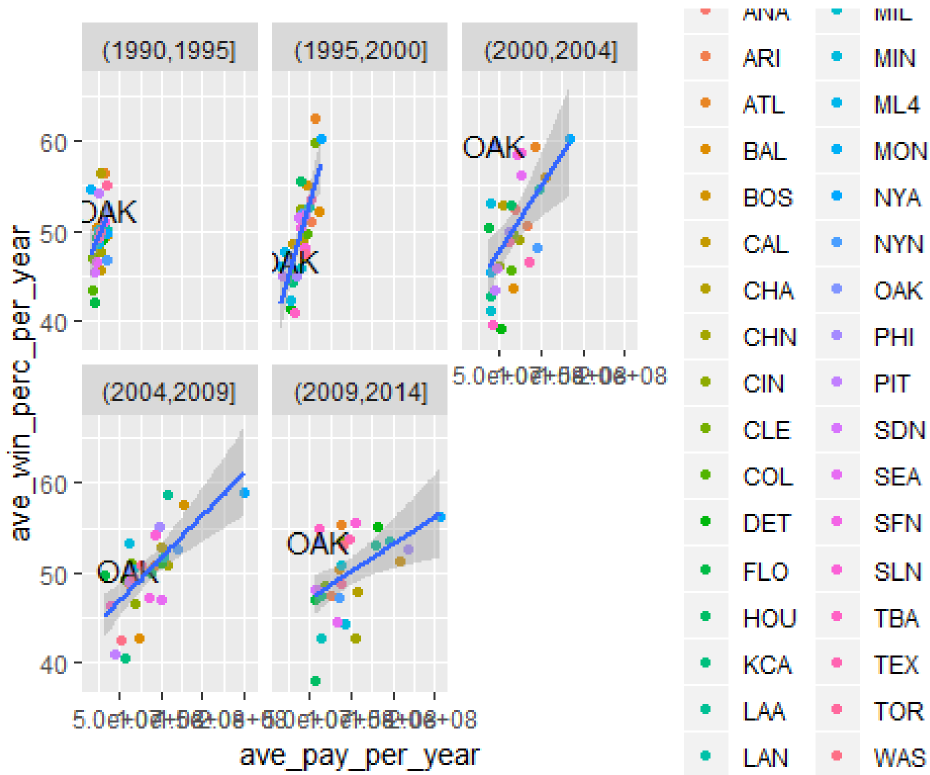


Write code to discretize year into five time periods (using the cut function with parameter breaks=5) and then make a scatterplot showing mean winning percentage (y-axis) vs. mean payroll (x-axis) for each of the five time periods. You could add a regression line (using geom_smooth(method=lm)) in each scatter plot to ease interpretation. Note: look at the discussion on faceting in the visualization EDA lecture notes. P4

```
total_payroll$year_range <- cut(total_payroll$yearID, breaks = 5)

ave_per_year <- total_payroll %>%
  group_by(year_range, teamID) %>%
  summarise(ave_pay_per_year = mean(payroll), ave_win_perc_per_year =
mean(winning_percentage, na.rm = TRUE))

ave_per_year %>% sample_n(5)

## # A tibble: 25 x 4
## # Groups:   year_range [5]
##    year_range  teamID ave_pay_per_year ave_win_perc_per_year
##    <fct>       <chr>             <dbl>                 <dbl>
##  1 (1990,1995] SEA           22670033.                  46.5
##  2 (1990,1995] MON           16227678.                  54.6
##  3 (1990,1995] ATL           31721853.                  56.5
##  4 (1990,1995] CHA           27090400.                  56.4
##  5 (1990,1995] DET           29670214.                  49.1
##  6 (1995,2000] COL           46062938.                  49.6
##  7 (1995,2000] MIN           26357000                   42.3
##  8 (1995,2000] BOS           47732454.                  55.0
```

```
## 9 (1995,2000] SLN        45455610                    48.0
## 10 (1995,2000] SDN       37744771.                   51.6
## # ... with 15 more rows
```

```
ave_per_year %>%
  ggplot(aes(y = ave_win_perc_per_year, x = ave_pay_per_year)) +
  geom_point(aes(color = teamID)) +
  geom_text(data = subset(ave_per_year, teamID == "OAK"), aes(label =
teamID)) +
  facet_wrap(~year_range) +
  geom_smooth(method = 'lm')
```



Q2 What can you say about team payrolls across these periods? Are there any teams that standout as being particularly good at paying for wins across these time periods? What can you say about the Oakland A's spending efficiency across these time periods (labeling points in the scatterplot can help interpretation).

The spread of the average payroll increases as more teams are paying their players more and more over time. The regression lines changes from vetical to diagonal that the more money the team pay to players, the more winnings they have. NYA

Write dplyr code to create a new variable in your dataset that standardizes payroll conditioned on year. So, this column for team i in year j should equal

P5

```
std_payroll <- total_payroll %>%
  group_by(yearID) %>%
  summarize(ave_payroll_per_year = mean(payroll), sd_payroll_per_year =
sd(payroll))

#join std_payroll to the original data: total_payroll by yearID
total_payroll <- total_payroll %>%
  inner_join(std_payroll, by = c("yearID"))

# new variable: std payroll for each team on each year
total_payroll <- total_payroll %>%
  mutate(std_payroll_conditioned_on_year = (payroll - ave_payroll_per_year) /
sd_payroll_per_year)

total_payroll %>% select(teamID, yearID, ave_payroll_per_year,
sd_payroll_per_year, std_payroll_conditioned_on_year) %>%
  sample_n(10)
```

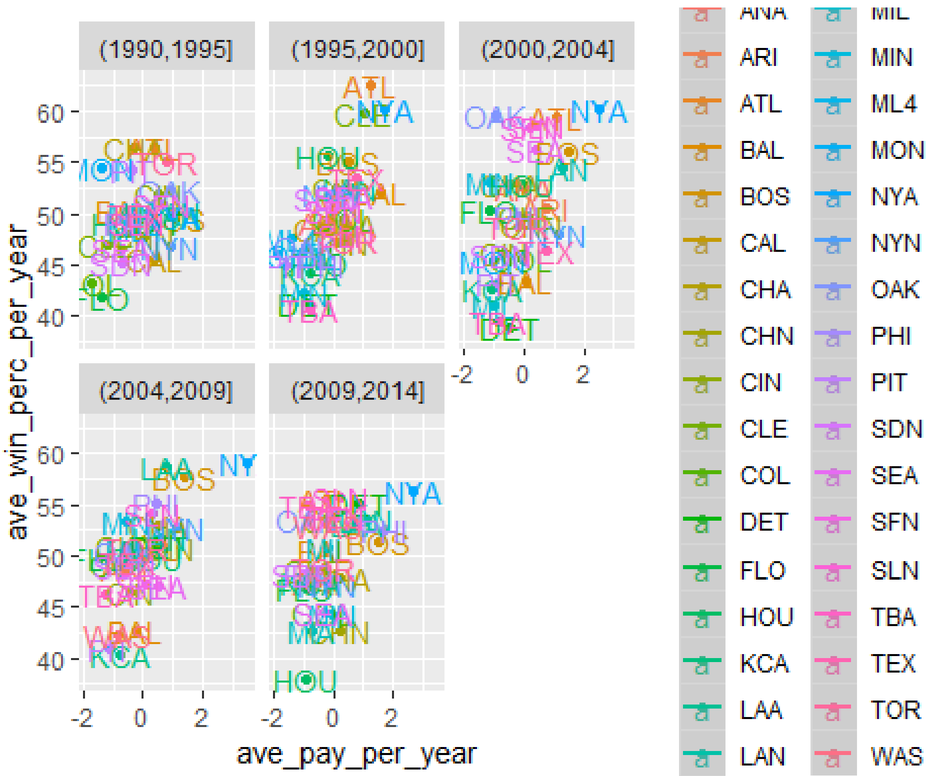
```
##    teamID yearID ave_payroll_per_year sd_payroll_per_year
## 1     TEX   1997             40260210            13060728
## 2     MIN   2005             72957113            34174781
## 3     SFN   1991             23578785             6894669
## 4     CLE   2005             72957113            34174781
## 5     LAA   2014             99800016            45705053
## 6     OAK   2013            101150855            48830287
## 7     CHN   1995             33981049             9447998
## 8     NYA   2011             92816843            40811974
## 9     DET   2002             67469251            24692193
## 10    OAK   2003             70942071            28011963
##    std_payroll_conditioned_on_year
## 1                        1.0097927
## 2                       -0.4907453
## 3                        1.0716803
## 4                       -0.9204042
## 5                        0.4854657
## 6                       -0.8400187
## 7                       -0.4736680
## 8                        2.6820115
## 9                       -0.5030437
## 10                      -0.7383002
```

Repeat the same plots as Problem 4, but use this new standardized payroll variable.

P6

```
total_payroll %>%
  group_by(teamID, year_range) %>%
  summarize(ave_pay_per_year = mean(std_payroll_conditioned_on_year),
ave_win_perc_per_year = mean(winning_percentage, na.rm = TRUE)) %>%
```

```r
  ggplot(aes(y = ave_win_perc_per_year, x = ave_pay_per_year, label = teamID,
color = teamID)) +
  geom_point() +
  geom_text() +
  facet_wrap(~year_range) +
  geom_smooth(method = 'lm')
```

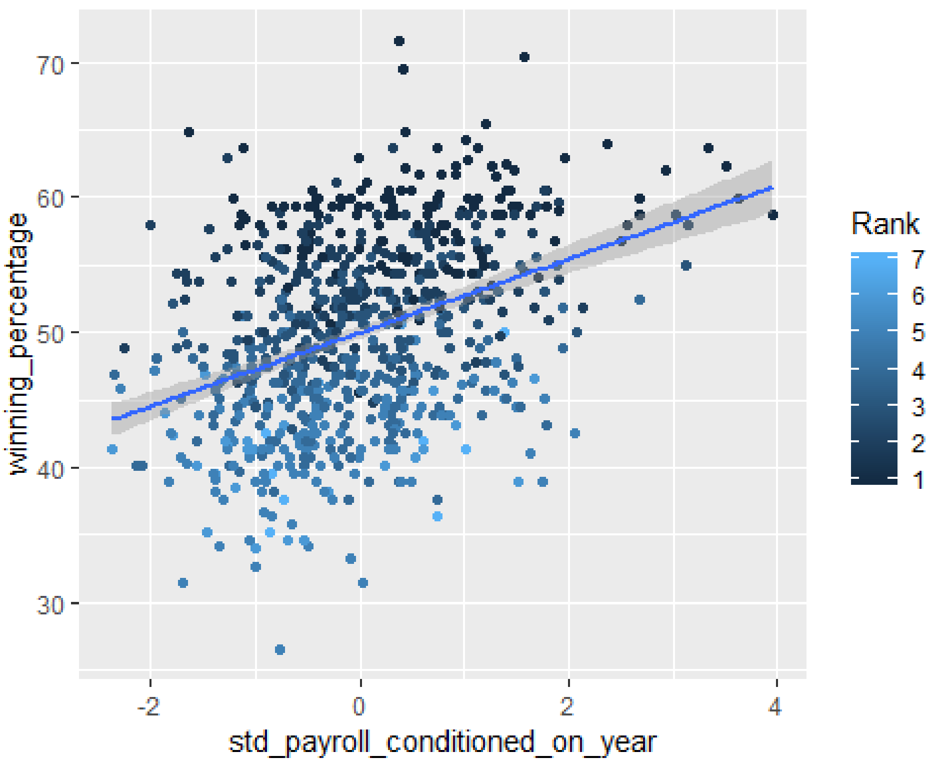


Q3 Discuss how the plots from Problem 4 and Problem 6 reflect the transformation you did on the payroll variable. Consider data range, center and spread along with observed correlation in your discussion. Some of these change after transformation, others don't. The new plot is the representation of the transformation since it is clear that each data point is relative to each other on a standard scale.

Make a single scatter plot of winning percentage (y-axis) vs. standardized payroll (x-axis). Add a regression line to highlight the relationship (again using geom_smooth(method=lm)).

p7

```r
total_payroll %>%
  ggplot(aes(y = winning_percentage, x = std_payroll_conditioned_on_year,
label = teamID)) +
geom_point(aes(color = Rank)) +
  geom_smooth(method = 'lm')
```



The regression line gives you expected winning percentage as a function of standardized payroll. Looking at the regression line, it looks like teams that spend roughly the average payroll in a given year will win 50% of their games (i.e. win_pct is 50 when standardized_payroll is 0), and teams increase 5% wins for every 2 standard units of payroll (i.e., win_pct is 55 when standardized_payroll is 2). We will see how this is done in general using linear regression later in the course.

From these observations we can calculate an expected win percentage for team i in year j as

expected_win_pct(ij) = 50 + 2.5 × standardized_payroll(ij)

Write dplyr code to calculate spending efficiency for each team

efficiency(ij) = win_pct(ij) ??? expected_win_pct(ij)

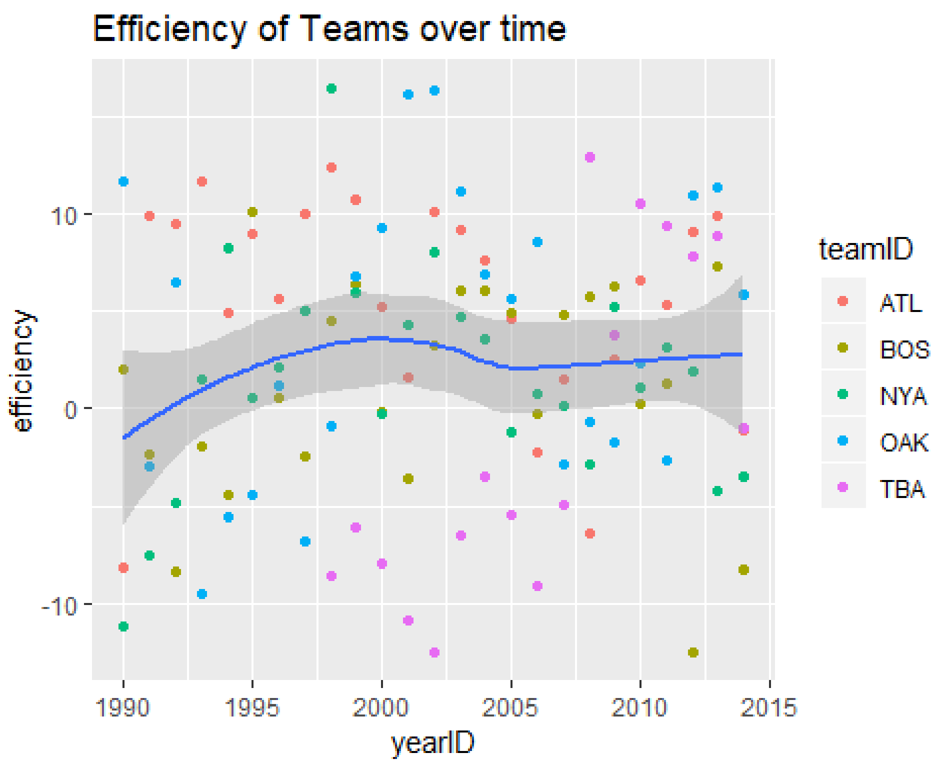
for team i in year j, where expected_win_pct is given above.

Make a line plot with year on the x-axis and efficiency on the y-axis. A good set of teams to plot are Oakland, the New York Yankees, Boston, Atlanta and Tampa Bay (teamIDs OAK, BOS, NYA, ATL, TBA). That plot can be hard to read since there is so much year to year variation for each team. One way to improve it is to use geom_smooth instead of geom_line.

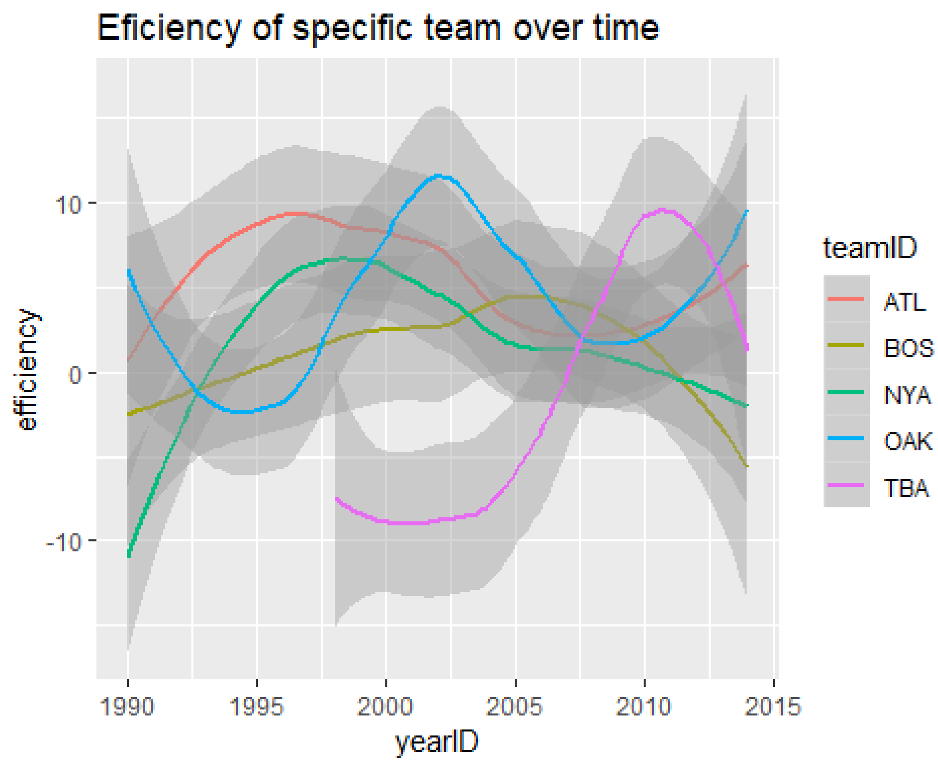
P8

```
#expected_win_pct(ij) = 50 + 2.5 × standardized_payroll(ij)
total_payroll <- total_payroll %>%
  mutate(expected_win_pct = (50 + 2.5 * std_payroll_conditioned_on_year))

#efficiency(ij) = win_pct(ij) ??? expected_win_pct(ij)
total_payroll <- total_payroll %>%
  mutate(efficiency = winning_percentage - expected_win_pct)

total_payroll %>% select(teamID, yearID, winning_percentage,
expected_win_pct, efficiency) %>% sample_n(10)

##    teamID yearID winning_percentage expected_win_pct  efficiency
## 1     NYN   1997           54.32099         49.91199   4.4090015
## 2     DET   2013           57.40741         52.29564   5.1117704
## 3     BOS   1998           56.79012         52.29955   4.4905745
## 4     BAL   2001           38.88889         50.22706 -11.3381755
## 5     DET   1990           48.76543         50.34525  -1.5798140
## 6     LAN   1997           54.32099         50.98006   3.3409325
## 7     TOR   2012           45.06173         48.45530  -3.3935703
## 8     CHN   2005           48.76543         51.02969  -2.2642614
## 9     DET   1993           52.46914         51.60985   0.8592874
## 10    TBA   2008           59.87654         46.97935  12.8971963

#(teamIDs OAK, BOS, NYA, ATL, TBA)
total_payroll %>%
  filter(teamID %in% c("OAK", "BOS", "NYA", "ATL", "TBA")) %>%
```

```r
  ggplot(aes(y = efficiency, x = yearID)) +
  geom_point(aes(color = teamID)) +
  geom_smooth(method = 'loess') +
  ggtitle("Efficiency of Teams over time")
```




```r
#(teamIDs OAK, BOS, NYA, ATL, TBA)
total_payroll %>%
  filter(teamID %in% c("OAK", "BOS", "NYA", "ATL", "TBA")) %>%
  ggplot(aes(y = efficiency, x = yearID, color = teamID)) +
  geom_smooth(method = 'loess') +
  ggtitle("Eficiency of specific team over time")
```



Q4 What can you learn from this plot compared to the set of plots you looked at in Question 2 and 3? How good was Oakland's efficiency during the Moneyball period? From the graph, efficiency of team over time, we can see that winning efficienct of teams over time is increased to an all time high in 2000 and In question 2 and 3, we observed that money has a high degree of influence on how well a team would do. Over time, the regression line of payroll and winning percentage emmerged; a team win more than 50% of their games if the team spend more than average amount of payroll on the team. Oakland is an outlier of the trend. During the Moneyball peroid, Oakland was more efficent than any other team from 2000 to 2005. In other words, Oakland was winning a lot more games than we could expected (by looking at how much they were spending on the team)