Report

Data Wrangling and Exploratory Data Analysis

Setting up the connection

```
library(tidyverse)
```

```
##   Attaching packages                     tidyverse 1.3.0

##   ggplot2 3.2.1      purrr   0.3.3
##   tibble  2.1.3      dplyr   0.8.3
##   tidyr   1.0.0      stringr 1.4.0
##   readr   1.3.1      forcats 0.4.0

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

```
library(RSQLite)
db <- DBI::dbConnect(RSQLite::SQLite(), "lahman2016.sqlite")
```

Problem 1 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).

```
select Salaries.yearID as year,
  Teams.teamID as team,
  (cast(sum(Teams.W) as float) / sum(Teams.G))*100 as win_percentage,
  franchName,
  sum(Salaries.salary) as payroll
from Salaries
join Teams
on Salaries.teamID = Teams.teamID
and Salaries.yearID = Teams.yearID
join TeamsFranchises
on Teams.franchID = TeamsFranchises.franchID
where year between 1990 and 2014
group by year, team
```

Exploratory data analysis

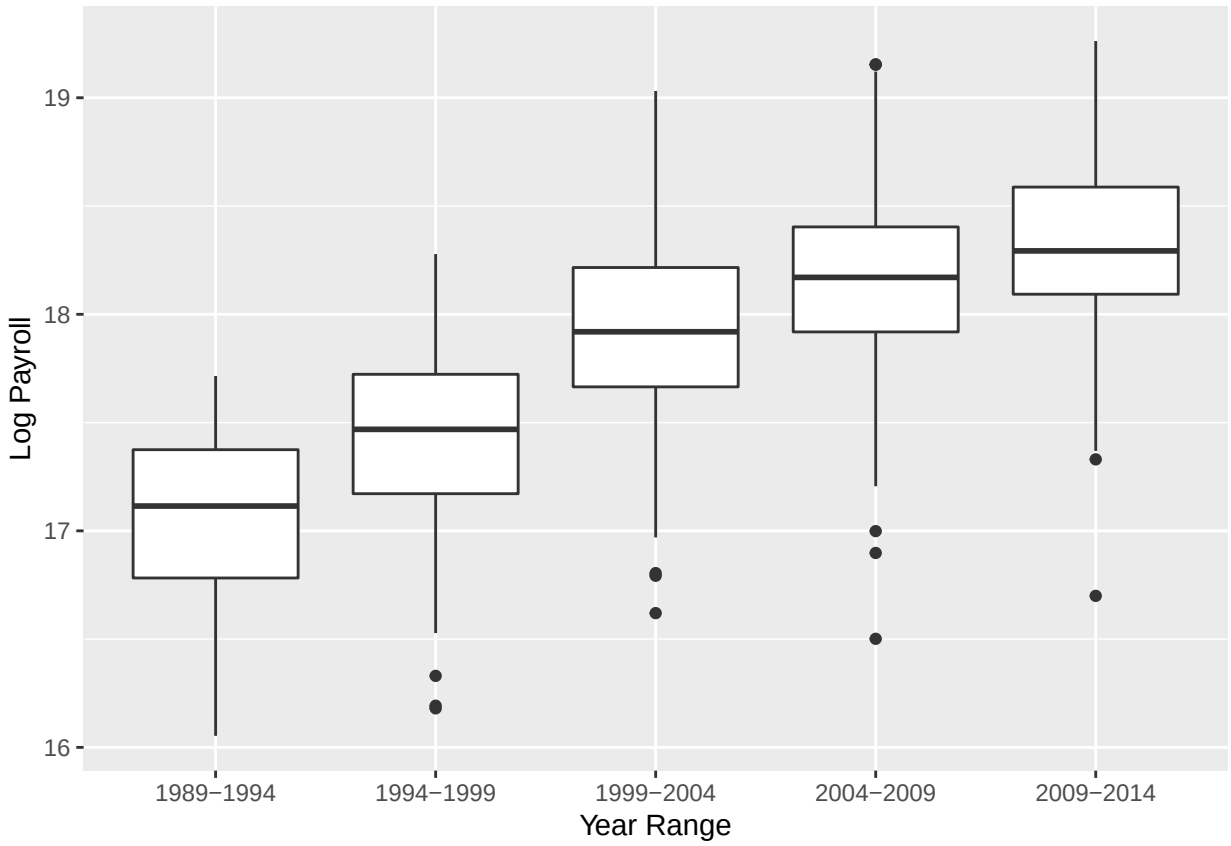
Payroll distribution

Problem 2. Write code to produce a plot (or plots) that shows the distribution of payrolls across teams conditioned on year (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).

```
select Salaries.yearID as year,
  Teams.teamID as team,
  sum(Salaries.salary) as payroll
from Salaries
join Teams
on Salaries.teamID = Teams.teamID
and Salaries.yearID = Teams.yearID
where year between 1990 and 2014
group by year, team
```

```
library(ggplot2)

winpay %>%
  group_by(team) %>%
  mutate(log_payroll = log(payroll)) %>%
  mutate(yearRange = ifelse(1989 < year & year <= 1994, paste("1989-1994"),
                ifelse(1994 < year & year <= 1999, paste("1994-1999"),
                ifelse(1999 < year & year <= 2004, paste("1999-2004"),
                ifelse(2004 < year & year <= 2009, paste("2004-2009"),
                       paste("2009-2014")))))) %>%
  ggplot(aes(x=yearRange, y=log_payroll)) +
  geom_boxplot() +
  labs(x="Year Range", y="Log Payroll")
```



Question 1. 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.

There is a clear increasing central tendency, although it seems the spread of the data has also slightly increased over time as well. It can also be seen that the upper end of the payroll scale seemed to flatten out after 2004. Most of the outliers reside on the lower end of the payroll scale, whereas at the upper end there is only one.

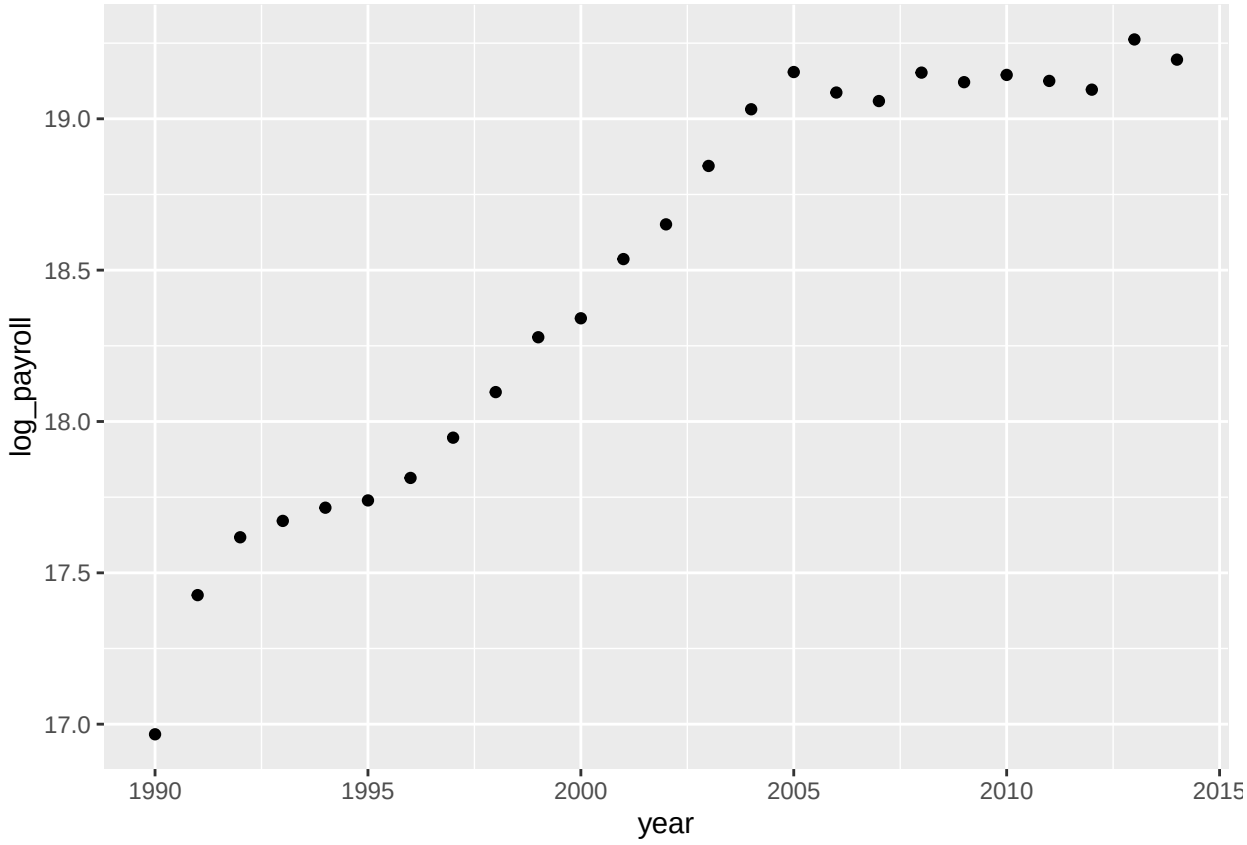
Problem 3. Write code to produce a plot (or plots) that specifically shows at least one of the statements you made in Question 1. For example, if you make a statement that there is a trend for payrolls to decrease over time, make a plot of a statistic for central tendency (e.g., mean payroll) vs. time to show that specifically.

The statement I am going to prove is that the maximum payrolls stopped increasing significantly after 2004. I did this by graphing strictly the max payrolls between the years 1990 and 2014. The graph clearly shows a steep decline in the slope of the graph after 2005, showing that the payrolls stopped increasing as sharply at that point.

```
select year, max(payroll) as max_payroll
from
  (select Salaries.yearID as year,
    Teams.teamID as team,
    sum(Salaries.salary) as payroll
  from Salaries
  join Teams
```

```
  on Salaries.teamID = Teams.teamID
  and Salaries.yearID = Teams.yearID
  where year between 1990 and 2014
  group by year, team)
group by year
```

```
max_df %>%
  mutate(log_payroll = log(max_payroll)) %>%
  ggplot(aes(x=year, y=log_payroll)) +
  geom_point()
```

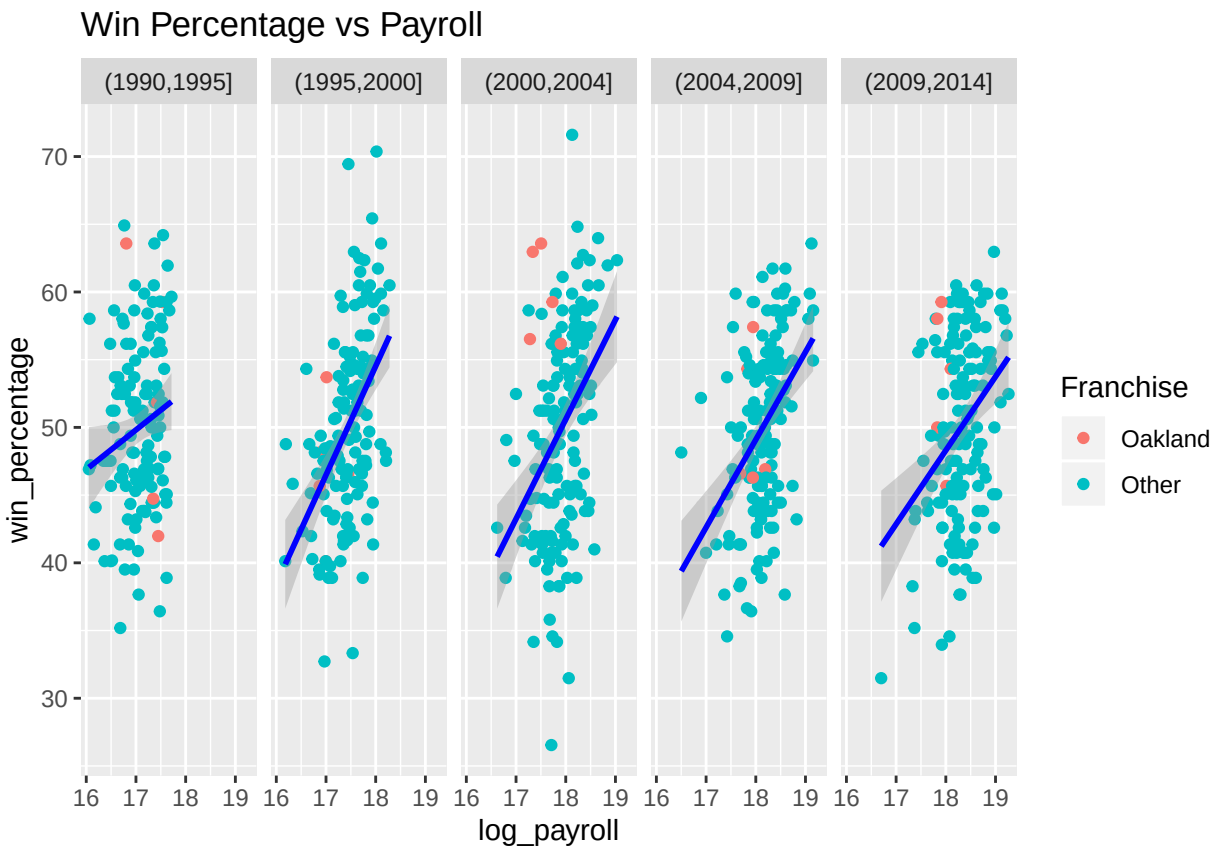


Correlation between payroll and winning percentage

Problem 4. Write code to discretize year into five time periods (e.g., using the cut function with parameter breaks=5 (in R, bins=5 in python) 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.

```
winpay$year_range <- cut(winpay$year, breaks=5)
winpay$Franchise[winpay$franchName != 'Oakland Athletics'] <- 'Other'
winpay$Franchise[winpay$franchName == 'Oakland Athletics'] <- 'Oakland'
```

```
winpay %>%
  mutate(log_payroll = log(payroll)) %>%
  ggplot(aes(x=log_payroll, y=win_percentage, color=Franchise)) +
  labs(title='Win Percentage vs Payroll') +
  facet_grid(~year_range) +
  geom_point() +
  geom_smooth(method=lm, color='blue')
```



Question 2. 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 some points in the scatterplot can help interpretation).

I can say that as a general trend, win percentages increase as a team's payroll increases, but there's so much variance in these plots that it is hard to call this a strong causation relationship. It seems that the time period when these two variables were most strongly correlated was the 1995 to 2000 time period. The Oakland A's spending efficiency seems to be solid. Acroess the five time periods, they generally end to be either above the mean, or very close to it, except for the 1990-1995 time range.

Data transformations

Standardization across years

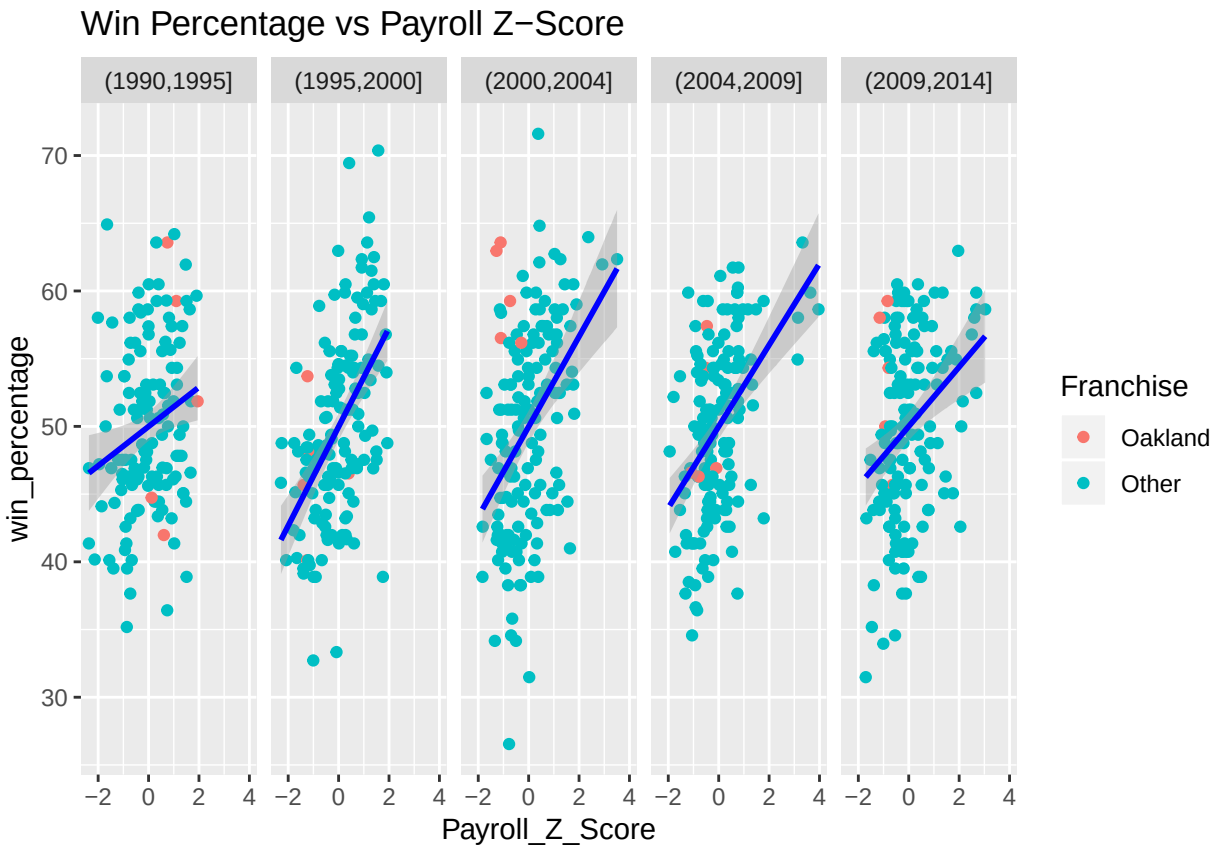
Problem 5. Write code to create a new variable in your dataset that standardizes payroll conditioned on year.

```
payroll_mew = array()
payroll_sigma = array()
for(i in 1990:2014) {
  payroll_mew[i - 1990 + 1] <- mean(filter(winpay, year==i)$payroll)
  payroll_sigma[i - 1990 + 1] <- sd(filter(winpay, year==i)$payroll)
}


winpay <- winpay %>%
  mutate(Payroll_Z_Score =
(payroll - payroll_mew[year - 1990 + 1]) / payroll_sigma[year - 1990 + 1])
```

Problem 6. Repeat the same plots as Problem 4, but use this new standardized payroll variable.

```
winpay %>%
  ggplot(aes(x=Payroll_Z_Score, y=win_percentage, color=Franchise)) +
  labs(title='Win Percentage vs Payroll Z-Score') +
  facet_grid(~year_range) +
  geom_point() +
  geom_smooth(method=lm, color='blue')
```



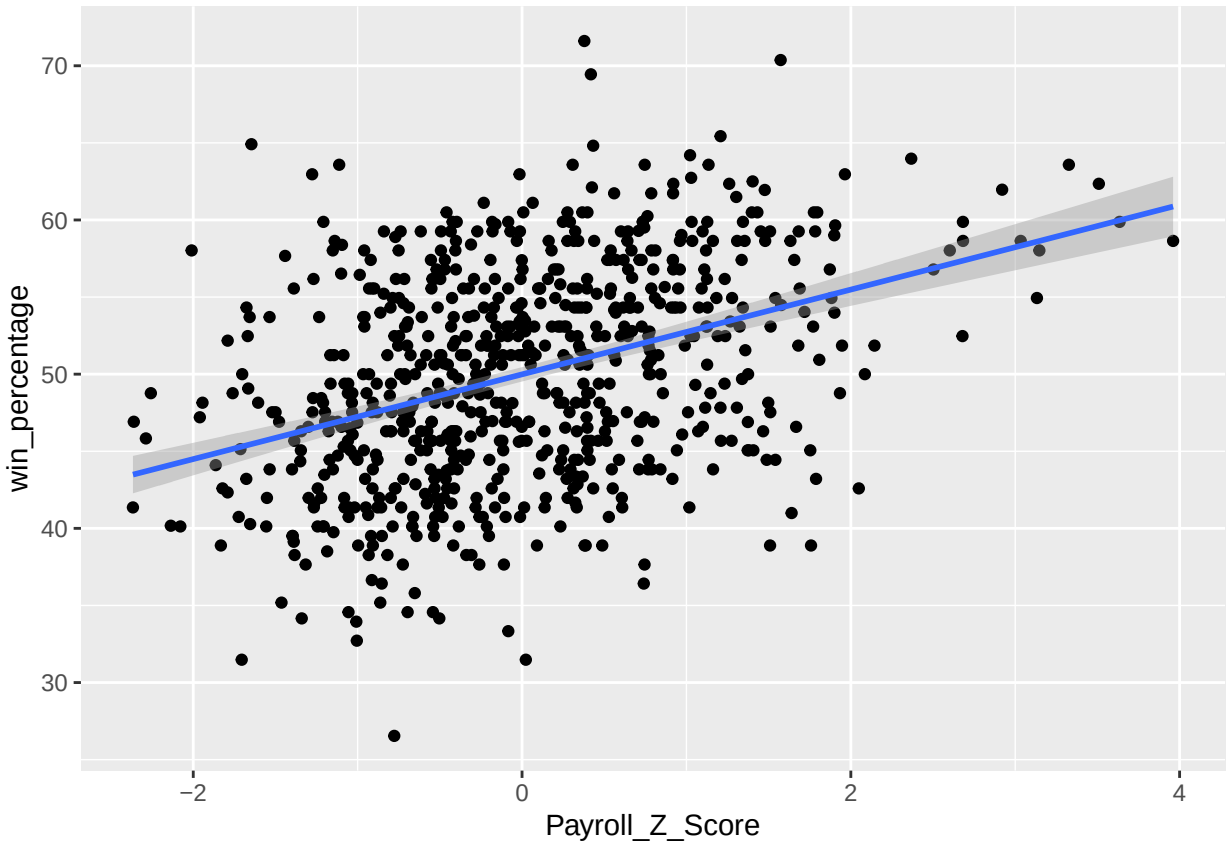
Question 3. Discuss how the plots from Problem 4 and Problem 6 reflect the transformation you did on the payroll variable. Consider data range, center and spread along with observed correlation in your discussion. Some of these change after the transformation, others don't.

The most obvious difference is the slope of the trendline has noticably increased, albeit not dramatically. The variance of the data has increased slightly, contributing to the increase in the trendline, but is still nearly the same. Each graph is also centered around x=0, whereas the graphs in problem 4 each had their own center-point making them harder to compare. Also each graph shows that as x=0, the win percentage is consistently 50%, implying when a team has the average payroll amount, they win 50% of the time.

Expected wins

Problem 7. 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)).

```
winpay %>%
  ggplot(aes(x=Payroll_Z_Score, y=win_percentage)) +
  geom_point() +
  geom_smooth(method=lm)
```

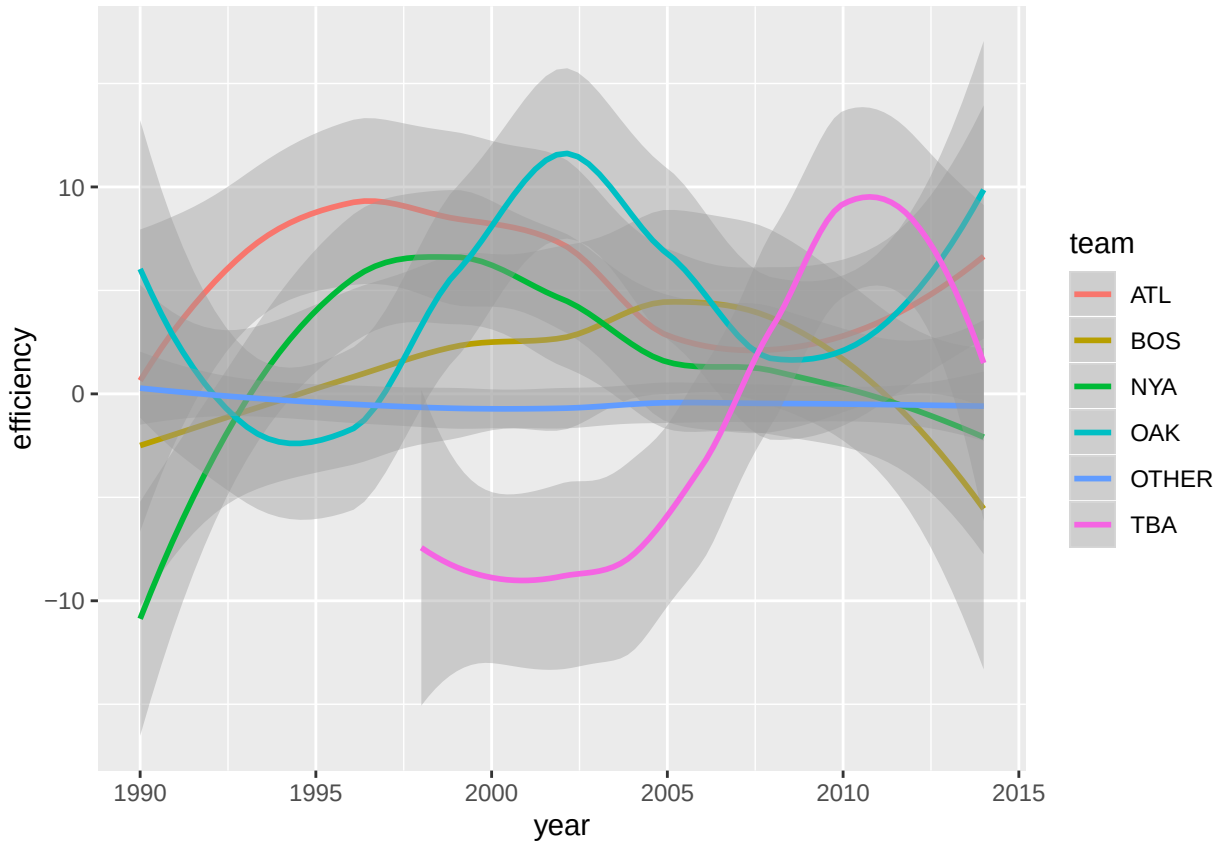


Spending efficiency

Problem 8. Write code to calculate spending efficiency for each team. 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.

```
winpay <- winpay %>%
  mutate(expected_win_pct = 50 + 2.5 * Payroll_Z_Score) %>%
  mutate(efficiency = win_percentage - expected_win_pct) %>%
  mutate(team =
           ifelse(
             team %in% c('OAK', 'BOS', 'NYA', 'ATL', 'TBA'),
             team, 'OTHER'))

winpay %>%
  ggplot(aes(x=year, y=efficiency, color=team)) +
  geom_smooth(method=loess)
```



Question 4. 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?

This plot provides significantly more valuable information than the plots from questions 2 and 3. This plot shows which teams are worth analyzing based on their efficiency throughout time. For example, if I were going to start a baseball team, I would analyze Oakland's strategies for team building in 2002, where they had an efficiency of 16.3% despite consistently being the lower end of the payroll spectrum. Oakland's efficiency during the Moneyball period was excellent. As this chart shows, Oakland started off with about average efficiency, then gradually throughout the 1990s their win percentage went up and up until it peaked in 2002. During the Moneyball period Oakland's win percentage was substantially above average and their win percentage stayed above average at least from roughly 1997 to 2015.