Project2

Joon

April 4, 2018

setting up db

sql

probably join with teamid,lgid,year after summing the salaries table to find payroll

I use sl statements to create multiple tables, payroll which is the sum of payrolls and thus gets the total payroll per team per year and win percentage with the same criteria

```
with payroll as
(select yearID,teamID,lgID,
sum(salary) as payroll
from salaries
group by yearID,teamID,lgID),
Percentage as
(select yearID,teamID,lgID,franchID,G,W,L, (100.00 * W/G) as winp
from teams
group by yearID,teamID,lgID)
select *
from payroll natural join Percentage
```

```
library(tidyverse)
```

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

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

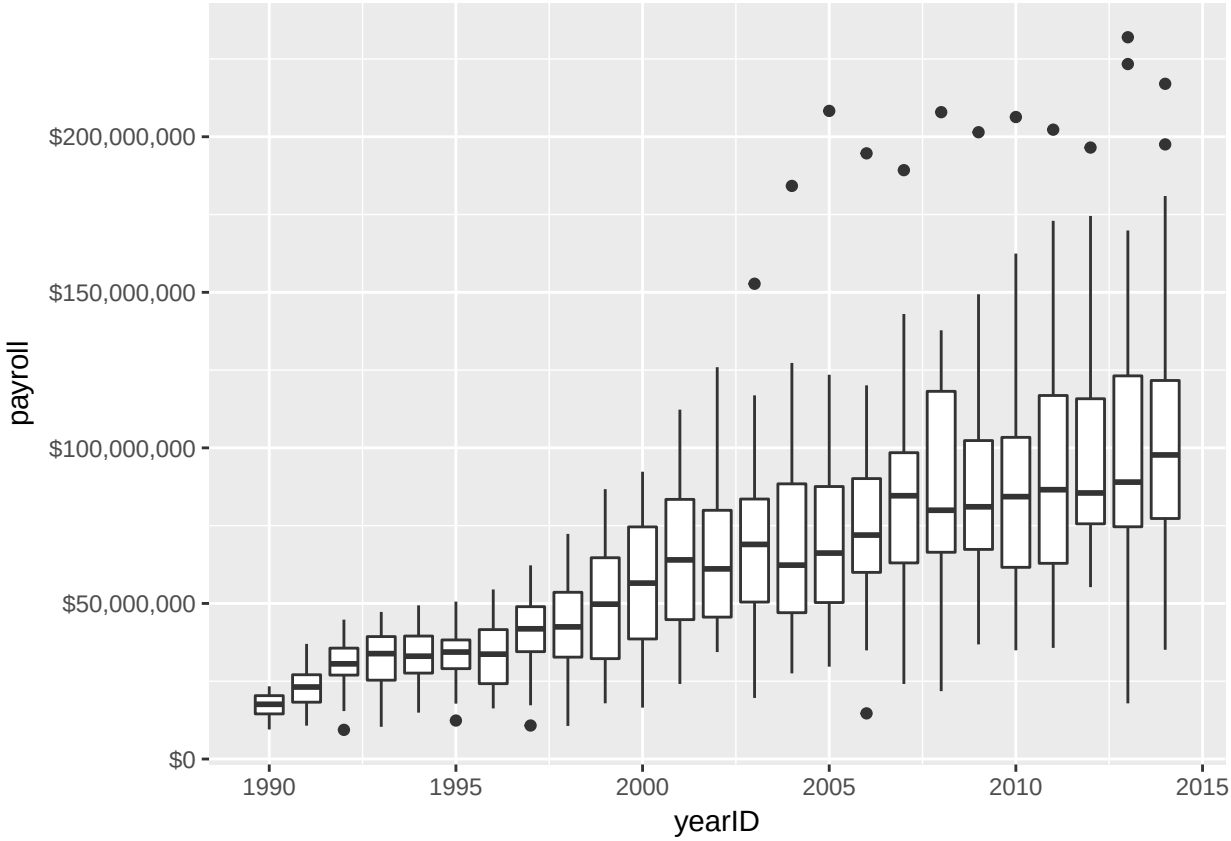
```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
result %>%
  head()
```

```
##   yearID teamID lgID  payroll franchID   G  W  L     winp
## 1   1985    ATL   NL 14807000      ATL 162 66 96 40.74074
## 2   1985    BAL   AL 11560712      BAL 161 83 78 51.55280
## 3   1985    BOS   AL 10897560      BOS 163 81 81 49.69325
## 4   1985    CAL   AL 14427894      ANA 162 90 72 55.55556
## 5   1985    CHA   AL  9846178      CHW 163 85 77 52.14724
## 6   1985    CHN   NL 12702917      CHC 162 77 84 47.53086
```

filtered the result to get only the years between 1989 and 2015 noninclusive

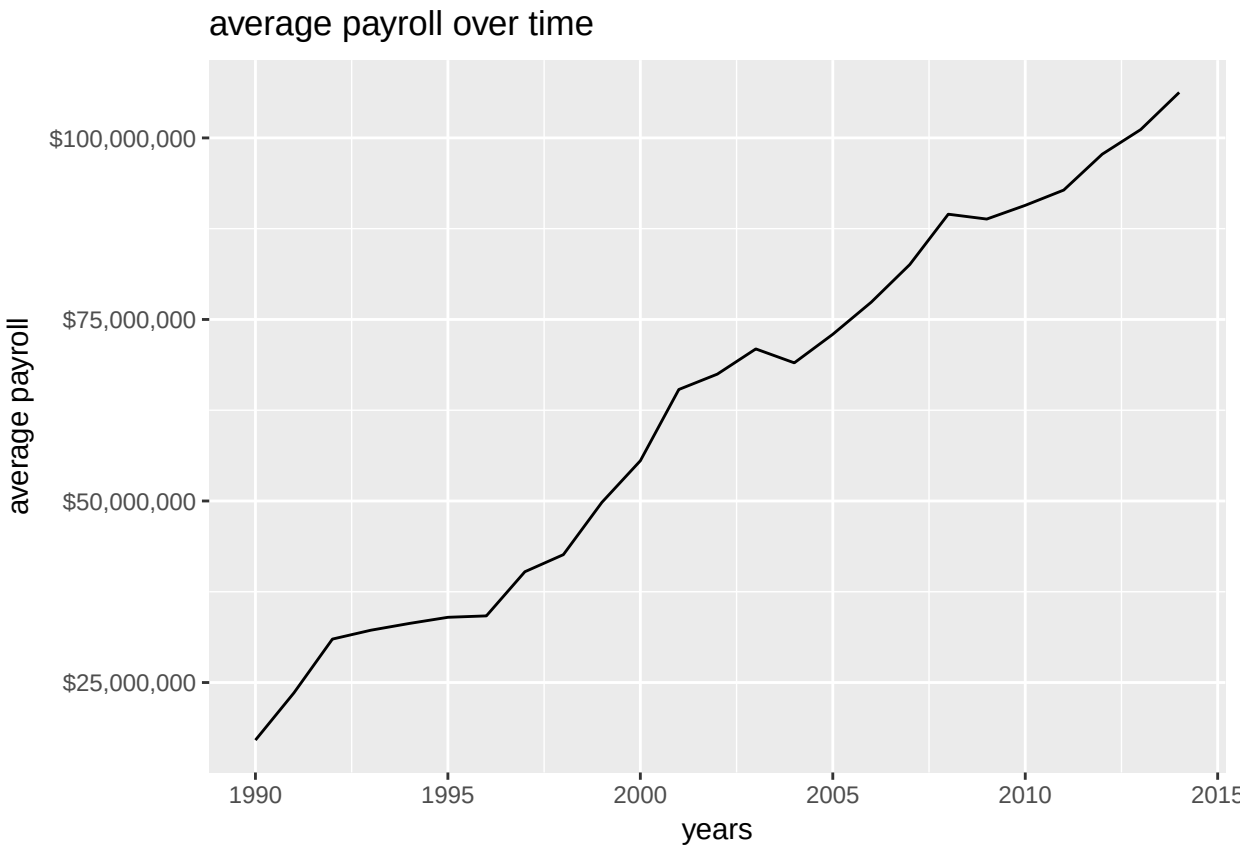
```
library(tidyverse)
filtered_result <- filter(result,yearID < 2015 & yearID > 1989)
    ggplot(filtered_result,mapping=aes(group=yearID,x=yearID,y=payroll)) +
      geom_boxplot() + scale_y_continuous(labels = scales::dollar)
```



Q1: the plot above shows that the average payroll per year seems to increase over time Also the std deviation tends to increase as well over the years so we have more variance

the plot below shows the mean over the years and how it increases over time I basically just grouped by year and found the average payroll for that year then graphed that average per year

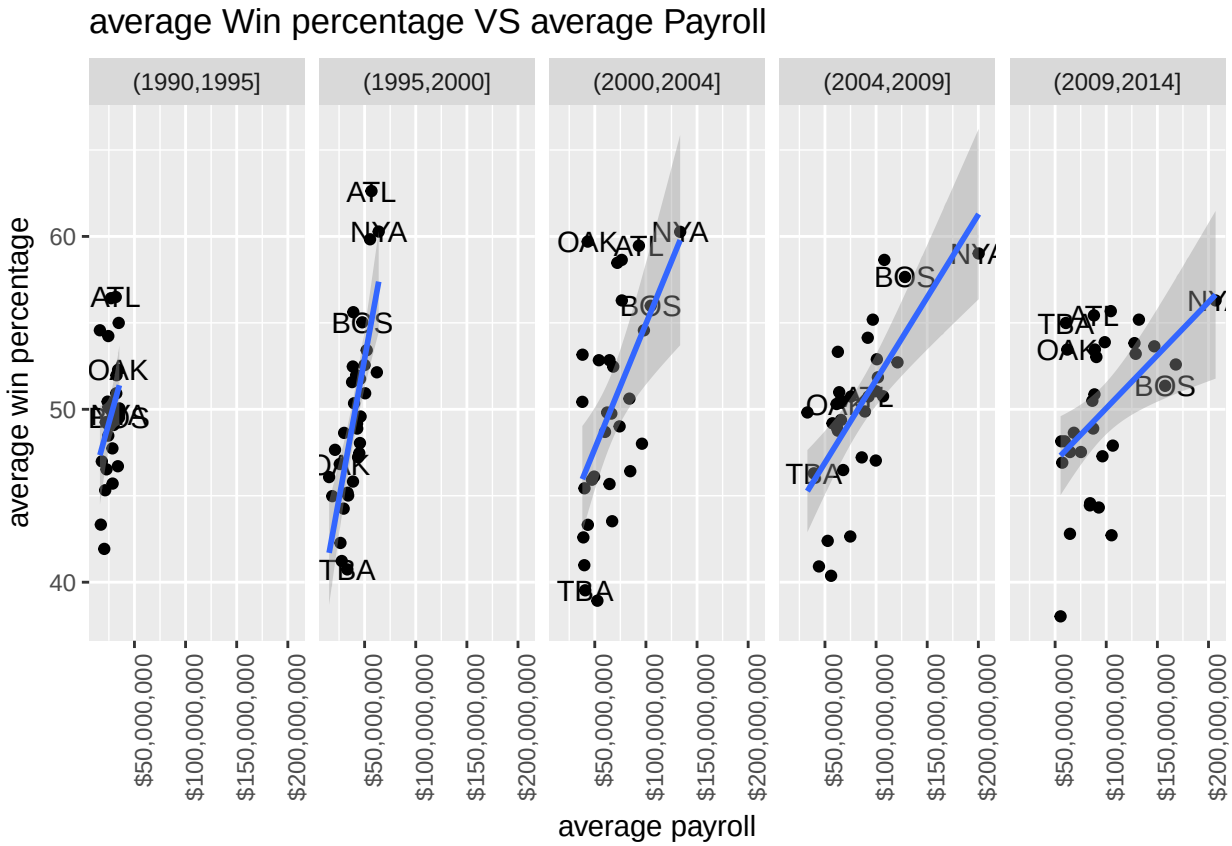
```
filtered_result %>%
  group_by(yearID) %>%
  summarize(avg_payroll = mean(payroll)) %>%
  ggplot(aes(x=yearID,y=avg_payroll)) + geom_line() +
  ggtitle("average payroll over time") +
  xlab("years") +
  ylab("average payroll") +
  scale_y_continuous(labels = scales::dollar)
```



I cut the years into 5 year chunks using the cut method and summarized the average win percentage and average payroll based on team then graphed them over each 5 year period

Q2: as time goes on the payroll seems to increase while the average win percentage seems to have the same distribution. Atlanta seems to have the best efficiency in terms of cost per win for the first 15 years while the Oakland A's seem to have maximum efficiency during the 2000-2004 period and the worst in 1995-2000.

```
library(tidyverse)
filtered_result %>%
  mutate(interval=cut(yearID,breaks=5)) %>%
  group_by(teamID,interval) %>%
  summarize(avg_winp = mean(winp),avg_payroll=mean(payroll)) %>%
  ggplot(aes(x=avg_payroll,y=avg_winp,label=teamID)) + facet_grid(.~interval) +
  ggtitle("average Win percentage VS average Payroll") +
  geom_text(aes(label=ifelse(teamID=="OAK" | teamID=="ATL"
                      | teamID=="NYA" | teamID=="BOS" | teamID=="TBA", as.character(teamID),''))) +
  geom_point() +
  xlab("average payroll") +
  ylab("average win percentage") +
  scale_x_continuous(labels = scales::dollar) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  geom_smooth(method =lm)
```



This code just finds the average payroll and standard deviation of payroll per year and then merges it to the original dataframe we were using that was filtered from 1990-2015. It then calculates the standardized payroll

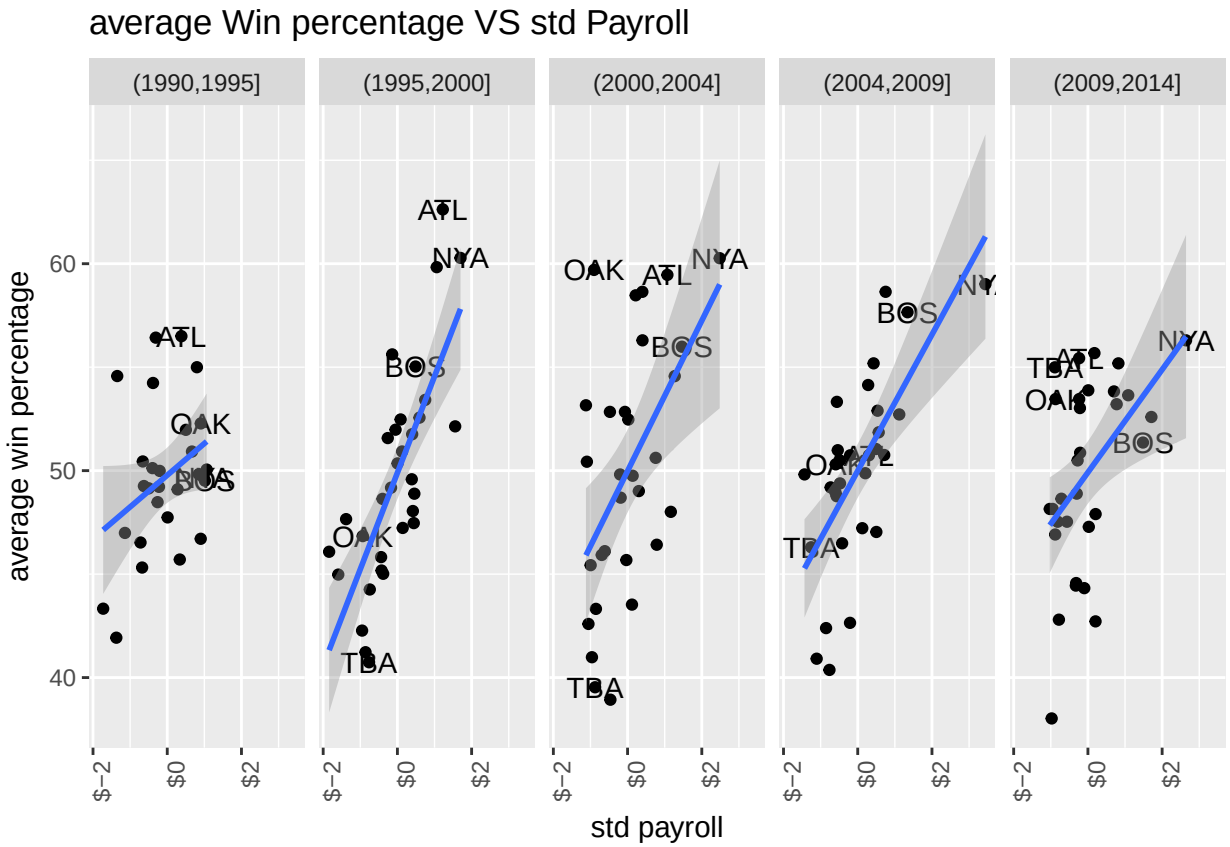
```
library(tidyverse)
library(dplyr)
avg_std <- filtered_result %>%
  mutate(interval=cut(yearID,breaks=5)) %>%
  group_by(yearID,interval) %>%
  summarize(avg_payroll = mean(payroll),std_dev = sd(payroll)) %>% as.data.frame()


standard <- merge(filtered_result,avg_std,by="yearID", all = TRUE) %>% as.data.frame() %>%
        mutate(std_payroll =(as.integer(payroll)- as.integer(avg_payroll))/std_dev ) %>% as.data.frame()
```

This codes does the same thing as q4 however the only difference is we calculate the average of the standard payroll that we calculated in the earlier chunk

Q3: The plots seem to mainly have the same same distribution and not have changed too much. which makes sense as the the payroll has gone through a transformation onthing else has really changed

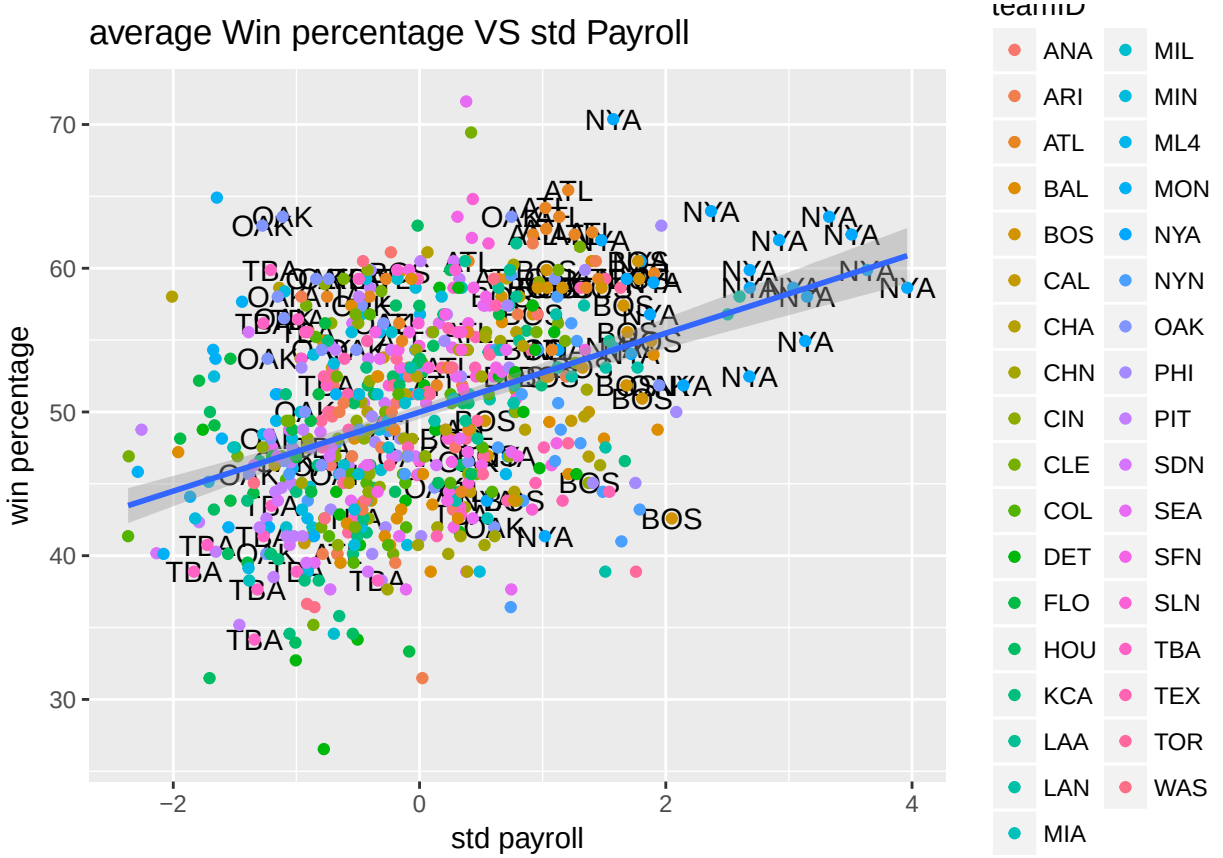
```
library(tidyverse)
standard %>%
  mutate(interval=cut(yearID,breaks=5)) %>%
  group_by(teamID,interval) %>%
  summarize(avg_winp = mean(winp), avg_std_payroll = mean(std_payroll)) %>%
  ggplot(aes(x=avg_std_payroll,y=avg_winp)) + facet_grid(.~interval) +
  ggtitle("average Win percentage VS std Payroll") +
  geom_text(aes(label=ifelse(teamID=="OAK" | teamID=="ATL" | teamID=="NYA"
```

```
                          | teamID=="BOS" | teamID=="TBA", as.character(teamID),'''))) +
  geom_point() +
  xlab("std payroll") +
  ylab("average win percentage") +
  scale_x_continuous(labels = scales::dollar) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  geom_smooth(method =lm)
```



This chunk just plots the standard payroll over the win percentage however this plot seems really clutterd it would be cooler maybe that indicates that there is no correlation between how big the payroll is and win percentage

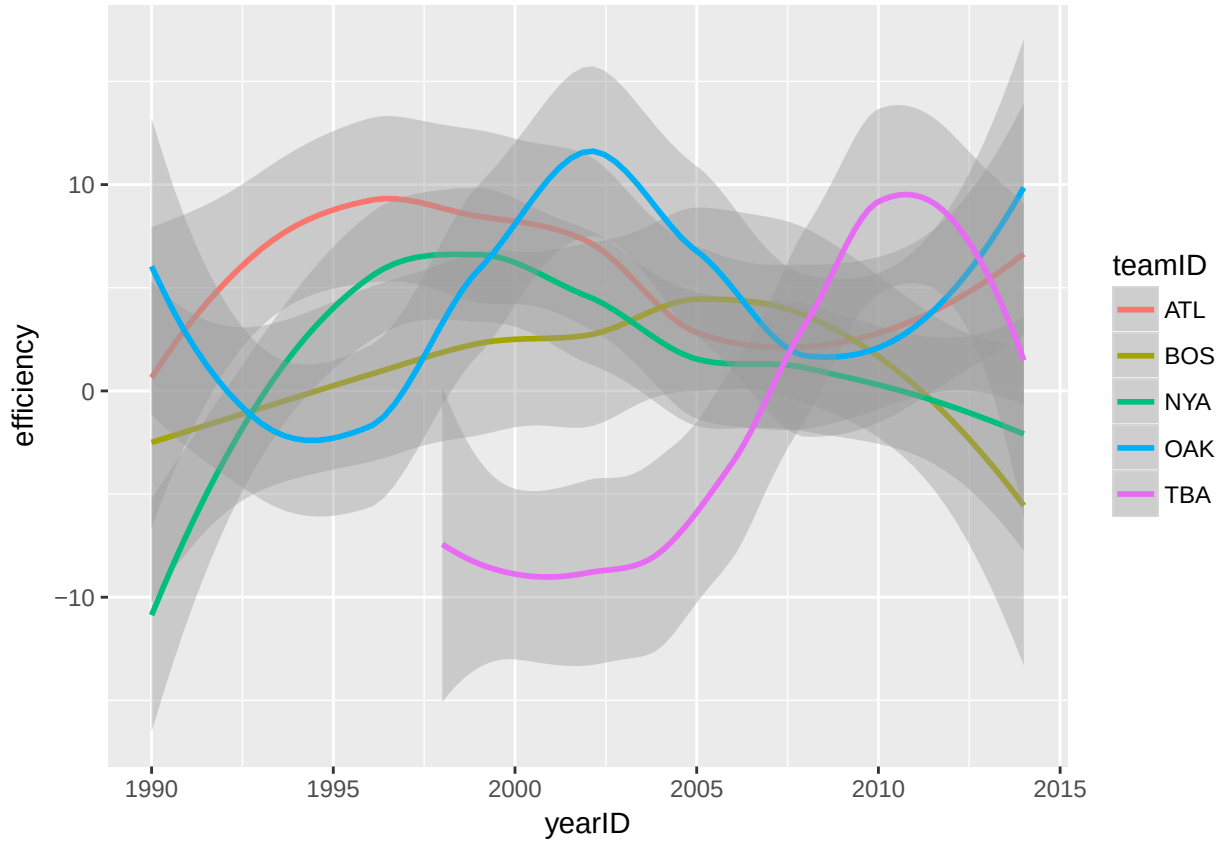
```
standard %>% ggplot(aes(x=std_payroll,y=winp))+
ggtitle("average Win percentage VS std Payroll") +
  geom_text(aes(label=ifelse(teamID=="OAK" | teamID=="ATL" | teamID=="NYA"
                          | teamID=="BOS" | teamID=="TBA", as.character(teamID),''))) +
  geom_point(aes(color=teamID)) +
  xlab("std payroll") +
  ylab("win percentage") +
  geom_smooth(method=lm)
```



Q4: from this plot we can see the teams efficiency over time but not necessarily how the cost affects win percentage. It does show trends of win percentages over time. The Oakland A's were making cash money Moolah during the money ball period. Their efficiency was at an all time high, however from their efficiency seems to die down .

```
standard %>%
  mutate(expected_win_pct= 50 + (2.5 * std_payroll)) %>%
  mutate(efficiency = winp - expected_win_pct) %>%
  filter(teamID %in% c("OAK","BOS","NYA","ATL","TBA")) %>%
  ggplot(aes(x=yearID,y=efficiency)) +geom_smooth(aes(color=teamID))
```

```
## `geom_smooth()` using method = 'loess'
```


efficiency
10
0
−10
1990
1995
2000
2005
2010
2015
yearID
teamID
ATL
BOS
NYA
OAK
TBA