project_2

Team Members : Payyavula Jaya Chandar and Oscar Lomibao Jr

```
# problem 1

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(RSQLite)
```

```
## Loading required package: DBI
```

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
lahman_con <- src_sqlite("C:\\Users\\jay\\Desktop\\lahman2014.sqlite")

query <-
 "SELECT Teams.yearID, Teams.teamID, Teams.franchID, W, G,
 ((CAST(W AS Double) / CAST(G AS Double)) * 100) as wins_percentage, sum(salary) as total_payroll
FROM Teams
LEFT JOIN Salaries
ON Salaries.teamID = Teams.teamID AND Salaries.yearID = Teams.yearID
GROUP BY Teams.yearID, Teams.teamID, Salaries.yearID, Salaries.teamID"

query_result <- lahman_con %>% tbl(sql(query))
result <- collect(query_result)

result1 <- result

head(result1)
```

```
## # A tibble: 6 × 7
##   yearID teamID franchID     W     G wins_percentage total_payroll
##    <int>  <chr>    <chr> <int> <int>           <dbl>         <chr>
## 1   1871    BS1      BNA    20    31        64.51613          <NA>
## 2   1871    CH1      CNA    19    28        67.85714          <NA>
## 3   1871    CL1      CFC    10    29        34.48276          <NA>
## 4   1871    FW1      KEK     7    19        36.84211          <NA>
## 5   1871    NY2      NNA    16    33        48.48485          <NA>
## 6   1871    PH1      PNA    21    28        75.00000          <NA>
```

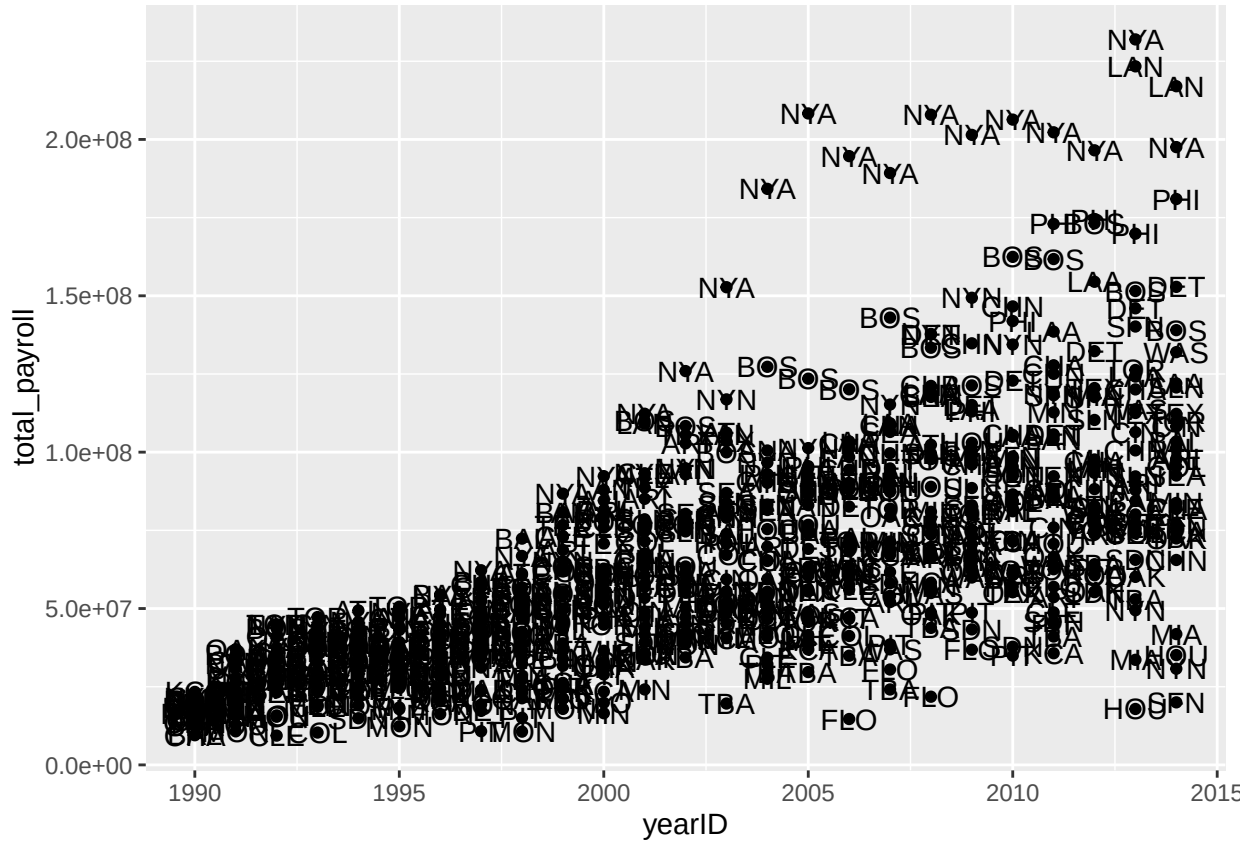
```
# problem 2

# gets years 1990-2014
result1 <- filter(result1, yearID > 1989)

# converts total payroll into numeric
result1$total_payroll <- as.numeric(result1$total_payroll)

# Problem 2 (plotting)
result$total_payroll <- as.numeric(result$total_payroll)

result %>%
 filter(yearID > 1989) %>%
 ggplot(aes(x = yearID, y = total_payroll, label = teamID)) +
 geom_text() +
 geom_point()
```



```r
head(result1)
```

```
## # A tibble: 6 × 7
##   yearID teamID franchID     W     G wins_percentage total_payroll
##    <int>  <chr>    <chr> <int> <int>           <dbl>         <dbl>
## 1   1990    ATL      ATL    65   162        40.12346      14555501
## 2   1990    BAL      BAL    76   161        47.20497       9680084
## 3   1990    BOS      BOS    88   162        54.32099      20558333
## 4   1990    CAL      ANA    80   162        49.38272      21720000
## 5   1990    CHA      CHW    94   162        58.02469       9491500
## 6   1990    CHN      CHC    77   162        47.53086      13624000
```

Problem 2 Question 1

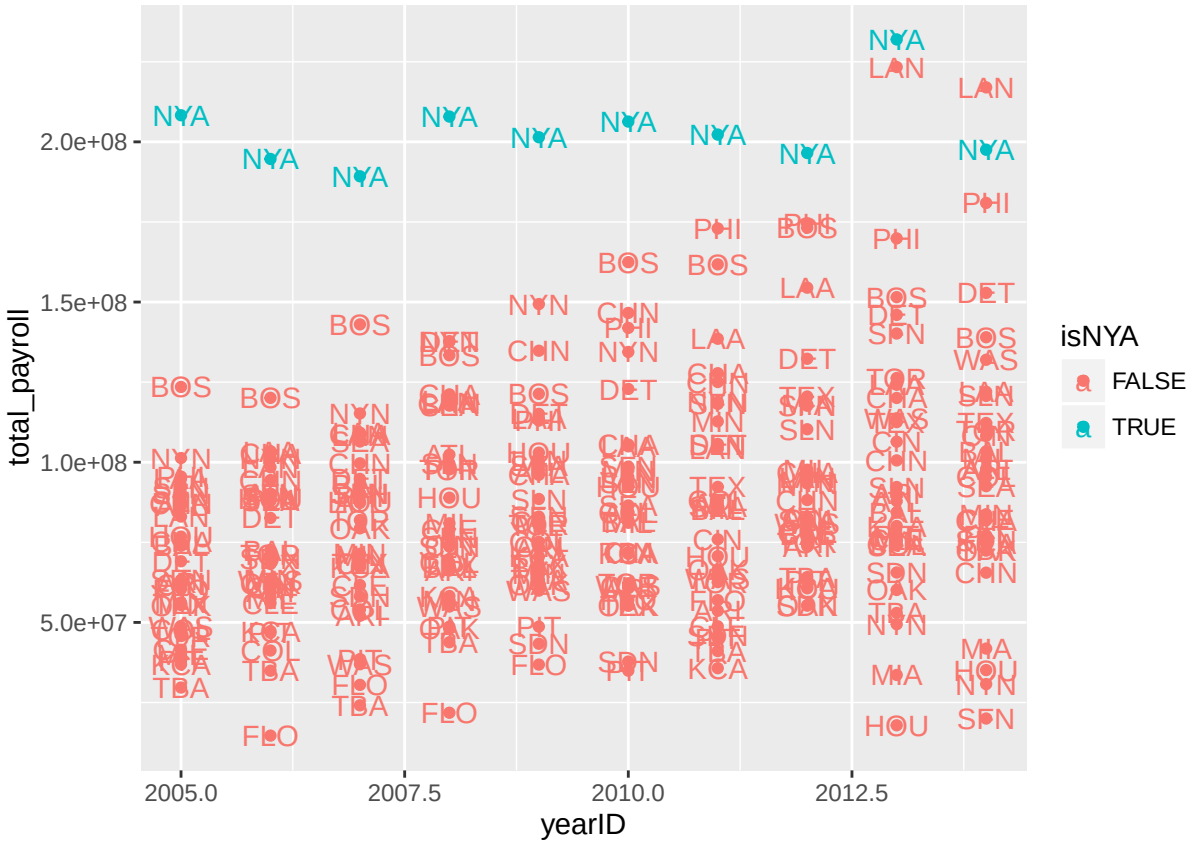
NYA has been spending the most money to buy top players since 2005 and it stands out in this category. Between 2005-2014 NYA was nine times the highest spender out of 15 times among all other teams.

```r
# problem 3

# Produces a plot that shows what we stated in problem 2 question 1.
# This plot shows that NYA has been spending the most money to buy top players since 2005.

# creates a new column (in result1) to tell whether a team is NYA or not
result1$isNYA <- ifelse(stringr::str_detect(result1$teamID, "NYA") , TRUE, FALSE)

# plots all team's payrolls throughout 2005-2014, emphasizing the NYA (as a different color);
# so that it is easier for viewers to see that NYA has been spending the most money to buy top players.
result1 %>%
 filter(yearID >= 2005) %>%
 ggplot(aes(x = yearID, y = total_payroll, label = teamID, color = isNYA)) +
 geom_text() +
 geom_point()
```




```
# problem 4

# breaks yearIDs into 5 periods (using the cur function)
result1$period <- cut(result1$yearID,breaks = 5, c('1990-1994','1995-1999','2000-2004','2005-2009','201

# gets the mean of winning percentages of each period
means <- result1 %>% group_by(period,teamID) %>% summarise(mean_wins = mean(wins_percentage))

# gets the mean of total payyrolls of each period
means_payroll <- result1 %>% group_by(period,teamID) %>% summarise(mean_payrolls = mean(total_payroll))

# joins the means into one data frame w/ the 5 periods
final_mean <- full_join(means_payroll, means)
```

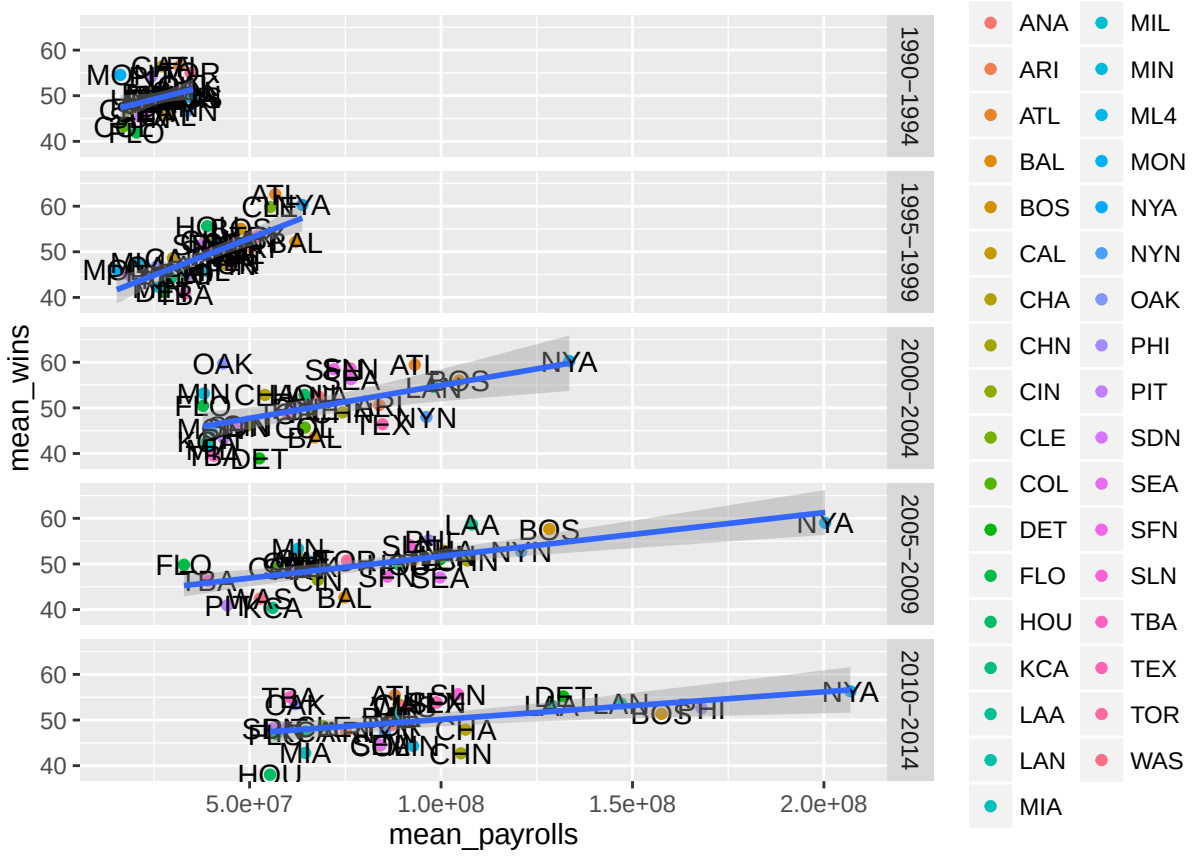
```
## Joining, by = c("period", "teamID")
```

```
head(final_mean)
```

```
## Source: local data frame [6 x 4]
## Groups: period [1]
##
##      period teamID mean_payrolls mean_wins
##      <fctr>  <chr>         <dbl>     <dbl>
## 1 1990-1994    ATL      31721853  56.49773
## 2 1990-1994    BAL      23785204  50.44408
```

```
## 3 1990-1994    BOS      34863217  49.51476
## 4 1990-1994    CAL      28654777  45.70478
## 5 1990-1994    CHA      27090400  56.42631
## 6 1990-1994    CHN      28460670  47.74012
```

```
# scatter plot showing mean winning percentage (y-axis) vs. mean payroll (x-axis)
# for each of the five time periods; this also includes a regression line.
final_mean %>%
filter(period %in% c("1990-1994","1995-1999","2000-2004","2005-2009","2010-2014")) %>%
ggplot(aes(x = mean_payrolls, y = mean_wins)) +
facet_grid(period ~ .) +
geom_point(aes(color = teamID)) +
geom_text(aes(label = teamID)) +
geom_smooth(method=lm)
```



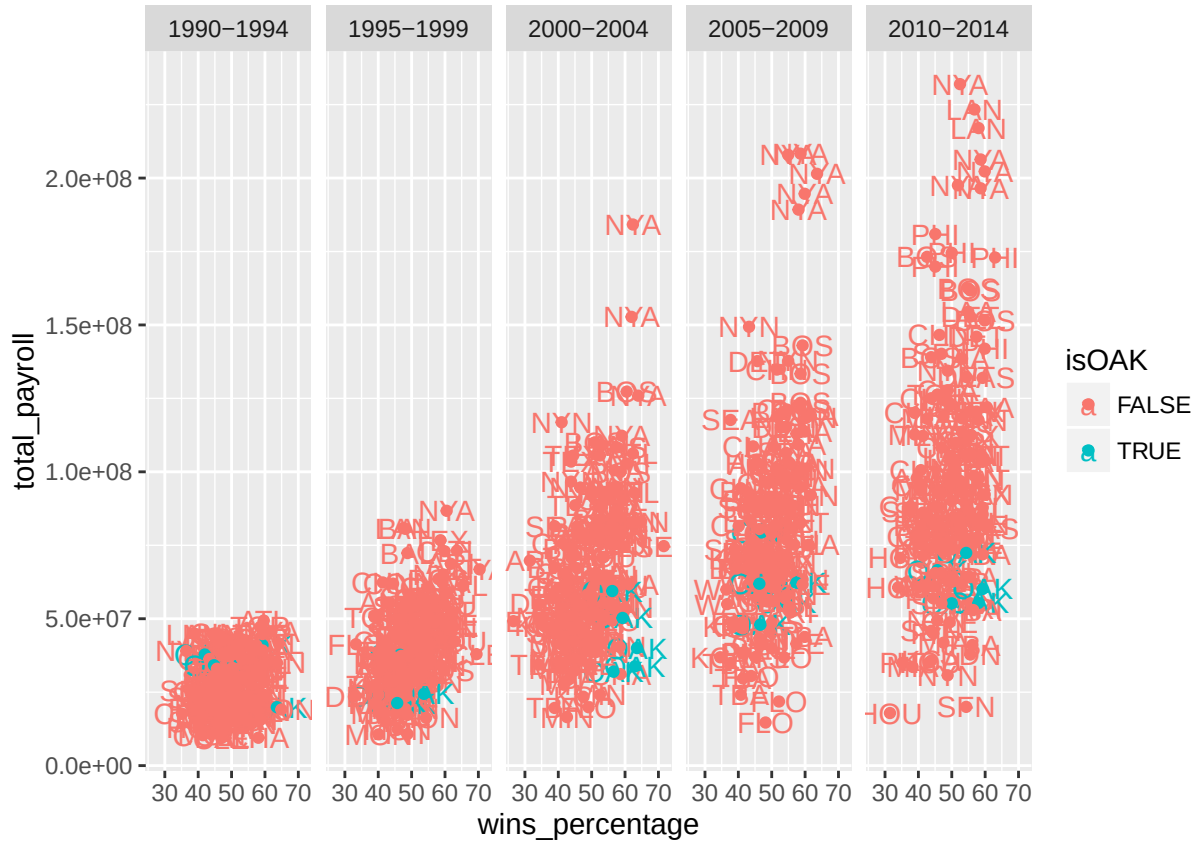
Question 2a The team that stands out as being particularly good at paying for wins across these time periods is NYA.Overall the payroll of all teams have increased since 1990.Below is the code on how Oakland's pay spending efficency.

```
# Question 2


# creates a new column (in result1) to tell whether a team is OAK or not
result1$isOAK <- ifelse(stringr::str_detect(result1$teamID, "OAK") , TRUE, FALSE)

# plots all team's payrolls throughout the 5 time periods, emphasizing OAK (as a different color);
```

```
# so that it is easier for viewers to see
result1 %>%
 filter(period %in% c("1990-1994","1995-1999","2000-2004","2005-2009","2010-2014")) %>%
 ggplot(aes(x = wins_percentage, y = total_payroll, label = teamID, color = isOAK)) +
 facet_grid(. ~ period) +
 geom_text() +
 geom_point()
```




```
head(result1)
```

```
## # A tibble: 6 × 10
##   yearID teamID franchID     W     G wins_percentage total_payroll isNYA
##    <int>  <chr>    <chr> <int> <int>           <dbl>         <dbl> <lgl>
## 1   1990    ATL      ATL    65   162        40.12346      14555501 FALSE
## 2   1990    BAL      BAL    76   161        47.20497       9680084 FALSE
## 3   1990    BOS      BOS    88   162        54.32099      20558333 FALSE
## 4   1990    CAL      ANA    80   162        49.38272      21720000 FALSE
## 5   1990    CHA      CHW    94   162        58.02469       9491500 FALSE
## 6   1990    CHN      CHC    77   162        47.53086      13624000 FALSE
## # ... with 2 more variables: period <fctr>, isOAK <lgl>
```

Oakland A's spending efficiency across these time periods is pretty good. We have observed that they don't spend as much as most teams, but at the same time their win percentage becomes pretty high througout these periods.

```r
#problem 5

# gets the average payroll of all teams for each year
avg_payroll <- result1 %>% group_by(yearID) %>% summarise(mean_per_year = mean(total_payroll))

# gets the standard deviation payroll of all teams for each year
sd_payroll <- result1 %>% group_by(yearID) %>% summarise(sd_per_year = sd(total_payroll))

# joins the average payroll and standard deviation payroll into 1 dataframe (joins by year)
avg_and_sd <- full_join(avg_payroll, sd_payroll)
```

```
## Joining, by = "yearID"
```

```r
# adds the new column (avg_and_sd) to result1
result1 <- full_join(result1, avg_and_sd)
```

```
## Joining, by = "yearID"
```

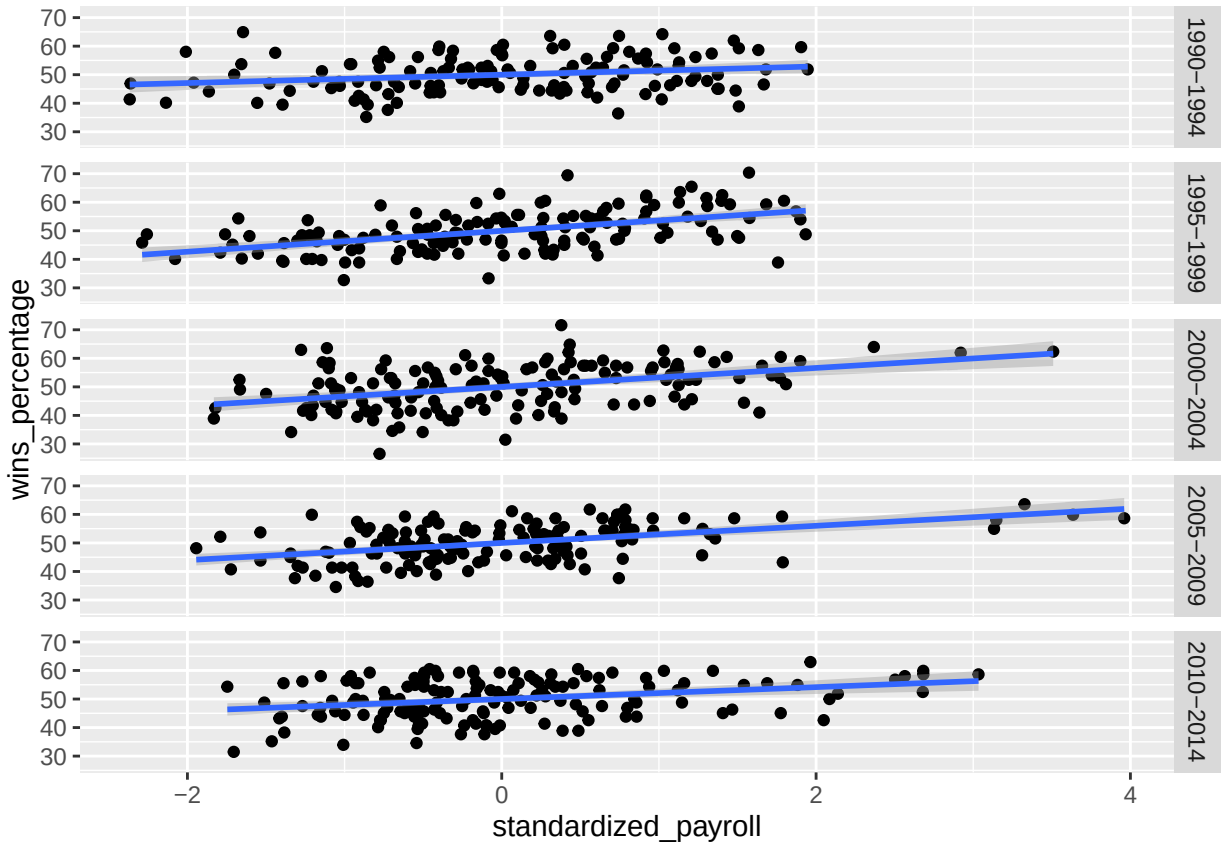
```r
# gets the standardized payroll for each payroll
# and adds a column of these standarized payrolls into result1
result1$standardized_payroll <- (result1$total_payroll - result1$mean_per_year) / result1$sd_per_year

head(result1)
```

```
## # A tibble: 6 × 13
##   yearID teamID franchID     W     G wins_percentage total_payroll isNYA
##    <int>  <chr>    <chr> <int> <int>           <dbl>         <dbl> <lgl>
## 1   1990    ATL      ATL    65   162        40.12346      14555501 FALSE
## 2   1990    BAL      BAL    76   161        47.20497       9680084 FALSE
## 3   1990    BOS      BOS    88   162        54.32099      20558333 FALSE
## 4   1990    CAL      ANA    80   162        49.38272      21720000 FALSE
## 5   1990    CHA      CHW    94   162        58.02469       9491500 FALSE
## 6   1990    CHN      CHC    77   162        47.53086      13624000 FALSE
## # ... with 5 more variables: period <fctr>, isOAK <lgl>,
## #   mean_per_year <dbl>, sd_per_year <dbl>, standardized_payroll <dbl>
```

```r
#problem 6

# scatter plot showing mean winning percentage (y-axis) vs. standardized payroll (x-axis)
# for each of the five time periods; this also includes a regression line.
result1 %>%
filter(period %in% c("1990-1994","1995-1999","2000-2004","2005-2009","2010-2014")) %>%
ggplot(aes(x = standardized_payroll, y = wins_percentage)) +
 facet_grid(period ~ .) +
geom_point() +
geom_smooth(method=lm)
```



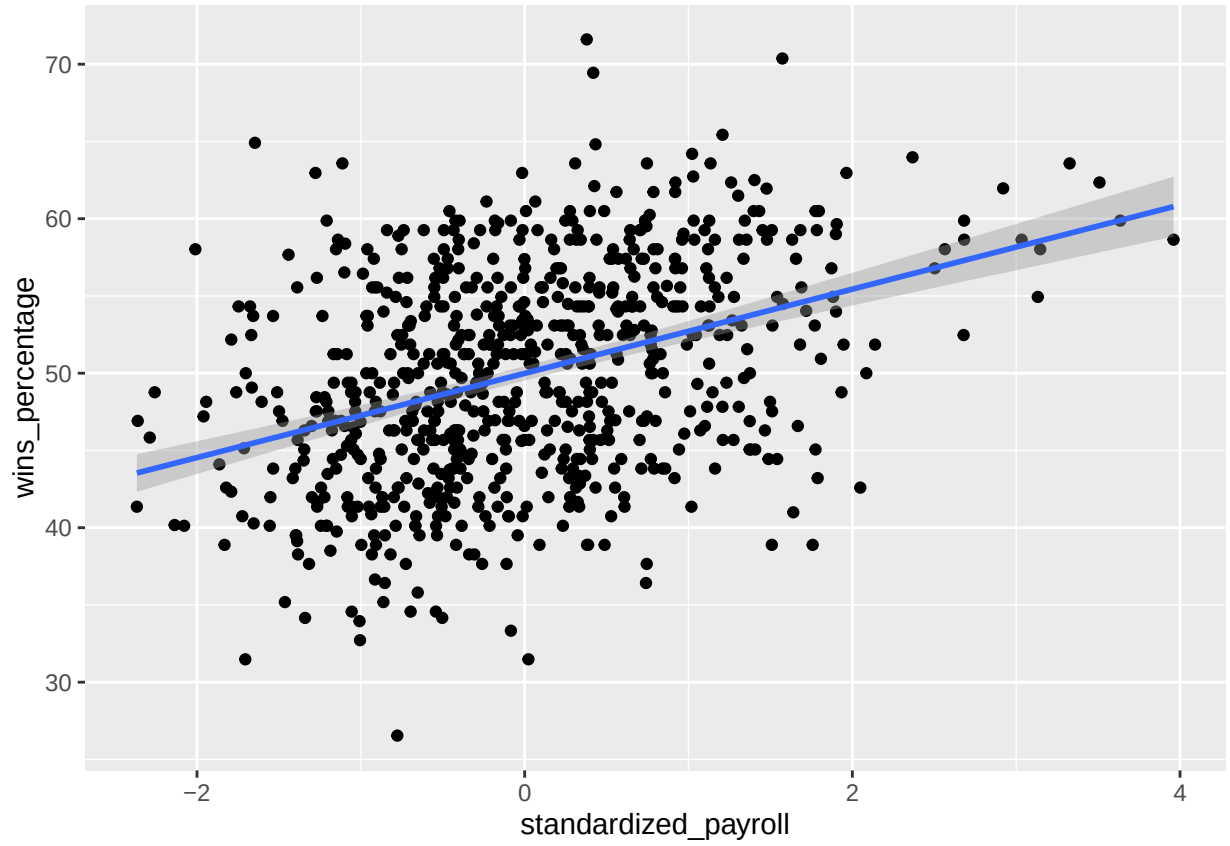
Question 3)

Discuss how the plots from Problem 4 and Problem 6 reflect the transformation you did on the payroll variable ?

There trends in the graphs are pretty much identical. But after the transformation all the graphs in 6 are much more centered due to the heavy change in the intervals ie in plot 4 the x axis were in millions but in plot 6 they are just between [-2,4]

```
#problem 7

# Make a single scatter plot of winning percentage (y-axis) vs.standardized payroll (x-axis).
result1 %>% filter(yearID >= 1990) %>% ggplot(aes(x=standardized_payroll, y= wins_percentage )) +
  geom_point() +
  geom_smooth(method = lm)
```




```
# gets/calculates expected winning percentage
result1$expctd_win_pct <- 50 + (2.5 * result1$standardized_payroll)

head(result1)
```

```
## # A tibble: 6 × 14
##   yearID teamID franchID     W     G wins_percentage total_payroll isNYA
##    <int>  <chr>    <chr> <int> <int>           <dbl>         <dbl> <lgl>
## 1   1990    ATL      ATL    65   162        40.12346      14555501 FALSE
## 2   1990    BAL      BAL    76   161        47.20497       9680084 FALSE
## 3   1990    BOS      BOS    88   162        54.32099      20558333 FALSE
## 4   1990    CAL      ANA    80   162        49.38272      21720000 FALSE
## 5   1990    CHA      CHW    94   162        58.02469       9491500 FALSE
## 6   1990    CHN      CHC    77   162        47.53086      13624000 FALSE
## # ... with 6 more variables: period <fctr>, isOAK <lgl>,
## #   mean_per_year <dbl>, sd_per_year <dbl>, standardized_payroll <dbl>,
## #   expctd_win_pct <dbl>
```

```
#problem 8
 result1$efficiency <- result1$wins_percentage - result1$expctd_win_pct

head(result1)
```

```
## # A tibble: 6 × 15
##   yearID teamID franchID     W     G wins_percentage total_payroll isNYA
```

```
##    <int>  <chr>    <chr> <int> <int>           <dbl>         <dbl> <lgl>
## 1   1990    ATL      ATL    65   162        40.12346      14555501 FALSE
## 2   1990    BAL      BAL    76   161        47.20497       9680084 FALSE
## 3   1990    BOS      BOS    88   162        54.32099      20558333 FALSE
## 4   1990    CAL      ANA    80   162        49.38272      21720000 FALSE
## 5   1990    CHA      CHW    94   162        58.02469       9491500 FALSE
## 6   1990    CHN      CHC    77   162        47.53086      13624000 FALSE
## # ... with 7 more variables: period <fctr>, isOAK <lgl>,
## #   mean_per_year <dbl>, sd_per_year <dbl>, standardized_payroll <dbl>,
## #   expctd_win_pct <dbl>, efficiency <dbl>
```

```
# scatter plot
result1 %>%
filter(teamID %in% c("OAK","BOS","NYA","ATL","TBA")) %>%
ggplot(aes(x = yearID, y = efficiency)) +
geom_point(aes(color=teamID)) +
geom_smooth(method=lm)
```

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 gives a better perspective about winning efficiency of the teams. As in the calculation of efficinecy we have used the both payroll and winning percentage. Hence this plot is provides a better results than plot 2 and 3.It can be observed from the graph that Oaklands efficiency has increased drastically during the moneyball period.