Nishant Arora

CMSC 320

Project 2: Wrangling and Exploratory Data Analysis

Goal:

To apply data wrangling and exploratory data analysis skills to baseball data. In particular, to know how well did Moneyball work for the Oakland A's. Was it worthy of a movie?

Background:

We'll be looking at data about teams in Major League Baseball. A couple of important points to remember:

- Major League Baseball is a professional baseball league, where teams pay players to play baseball.
- The goal of each team is to win as many games out of a 162 game season as possible.
- Teams win games by scoring more runs than their adversary.
- In principle, better players are costlier, so teams that want good players need to spend more money.
- Teams that spend the most, frequently win the most.

So, the question is, how can a team that can't spend so much win? The basic idea that Oakland (and other teams) used is to redefine what makes a player good. I.e., figure out what player characteristics translated into wins. Once they realized that teams were not really pricing players using these characteristics, they could exploit this to pay for undervalued players, players that were good according to their metrics, but were not recognized as such by other teams, and therefore not as expensive.

The Data:

We will be using a useful database on baseball teams, players and seasons curated by Sean Lahman available at http://www.seanlahman.com/baseball-archive/statistics/ (http://www.seanlahman.com/baseball-archive/statistics/). The database has been made available as a sqlite database at https://github.com/jknecht/baseball-archive-sqlite (https://github.com/jknecht/baseball-archive-sqlite).

The Question:

We want to understand how efficient teams have been historically at spending money and getting wins in return. In the case of Moneyball, one would expect that Oakland was not much more efficient than other teams in their spending before 2000, were much more efficient (they made a movie about it after all) between 2000 and 2005, and by then other teams may have caught up.

How is this reflected in the data we have?

Wrangling:

***Problem 1:** Using SQL compute a relation containing 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).*

Include a sentence or two indicating how you dealt with any missing data in these two relations. Specifically, indicate if there is missing data in either table, and how the type of join you used determines how you dealt with this missing data.

```
# Import Database
db <- src_sqlite("C:\\CS\\data_science\\p2\\lahman2016.sqlite")

# SQL Query
query <-
  "with total_payroll as
      (select teamID, sum(salary) as payroll, yearID
       from Salaries
       group by teamID, yearID)
    select Teams.teamID, Teams.yearID, Teams.lgID, payroll, franchID,
           W, G, ((W * 1.0 / G) * 100) as win_percentage
    from total_payroll, Teams
    where total_payroll.teamID=Teams.teamID and
          total_payroll.yearID=Teams.yearID"

# Apply the Query
result <- db %>% tbl(sql(query))

# Convert to a Table for R
payroll_tab <- collect(result)

# View the Result
head(payroll_tab)
```

```
## # A tibble: 6 x 8
##   teamID yearID lgID    payroll franchID     W     G win_percentage
##   <chr>   <int> <chr>     <dbl> <chr>    <int> <int>          <dbl>
## 1 ATL      1985 NL    14807000. ATL         66   162           40.7
## 2 BAL      1985 AL    11560712. BAL         83   161           51.6
## 3 BOS      1985 AL    10897560. BOS         81   163           49.7
## 4 CAL      1985 AL    14427894. ANA         90   162           55.6
## 5 CHA      1985 AL     9846178. CHW         85   163           52.1
## 6 CHN      1985 NL    12702917. CHC         77   162           47.5
```

SQL WITH Clause Documentaton:

```
WITH <alias_name> AS (sql_subquery_statement)
SELECT column_list FROM <alias_name>[,table_name]
[WHERE <join_condition>]
```

Using the WITH…AS clause we can automatically join necessary data after pefroming subqueries. We select teamID, sum of the salary, and yearID first. In the second select statement, we perform an automatic join on total_payroll with Teams and select necersarry attributes such as teamID, yearID, lgID, etc. to compute a relation containing the total payroll and winning percentage.

Note on Missing Data:

- Team data is available since 1871 yet salary data is only available since 1985, therefore, team data between 1871 and 1984 is missing when a join is performed on Teams and total_payroll.
- The (inner) join is only performed on shared teamIDs and yearIDs between the tables.
- Teams with missing payroll information are ignored.

Exploratory Data Analysis

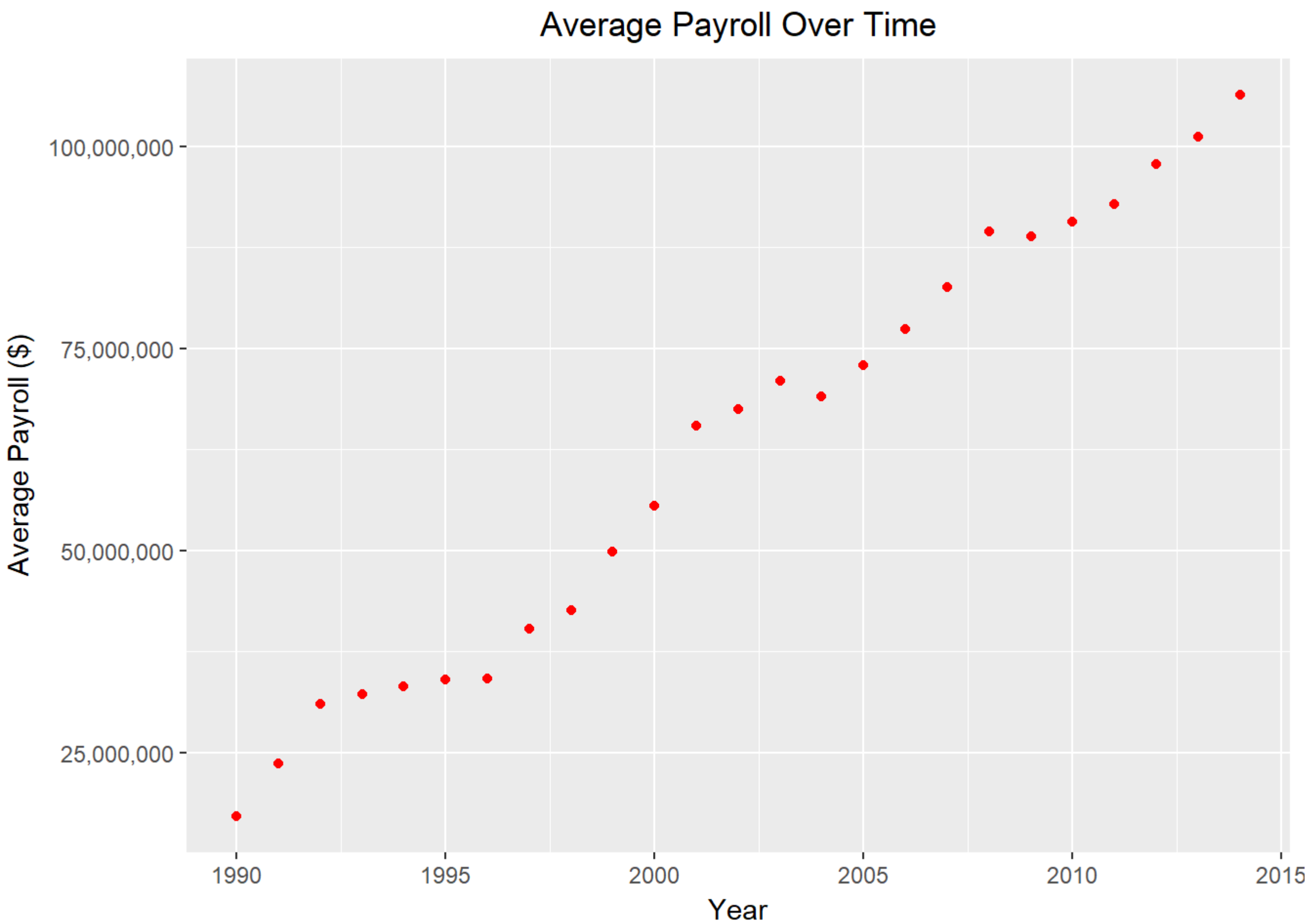
Payroll distribution

Problem 2: *Write code to produce a plot(s) that illustrate the distribution of payrolls across teams conditioned on time (from 1990-2014).*

Plot 1: Average Payroll Over Time

This plot shows the average overall payroll across all teams between 1990 and 2014.

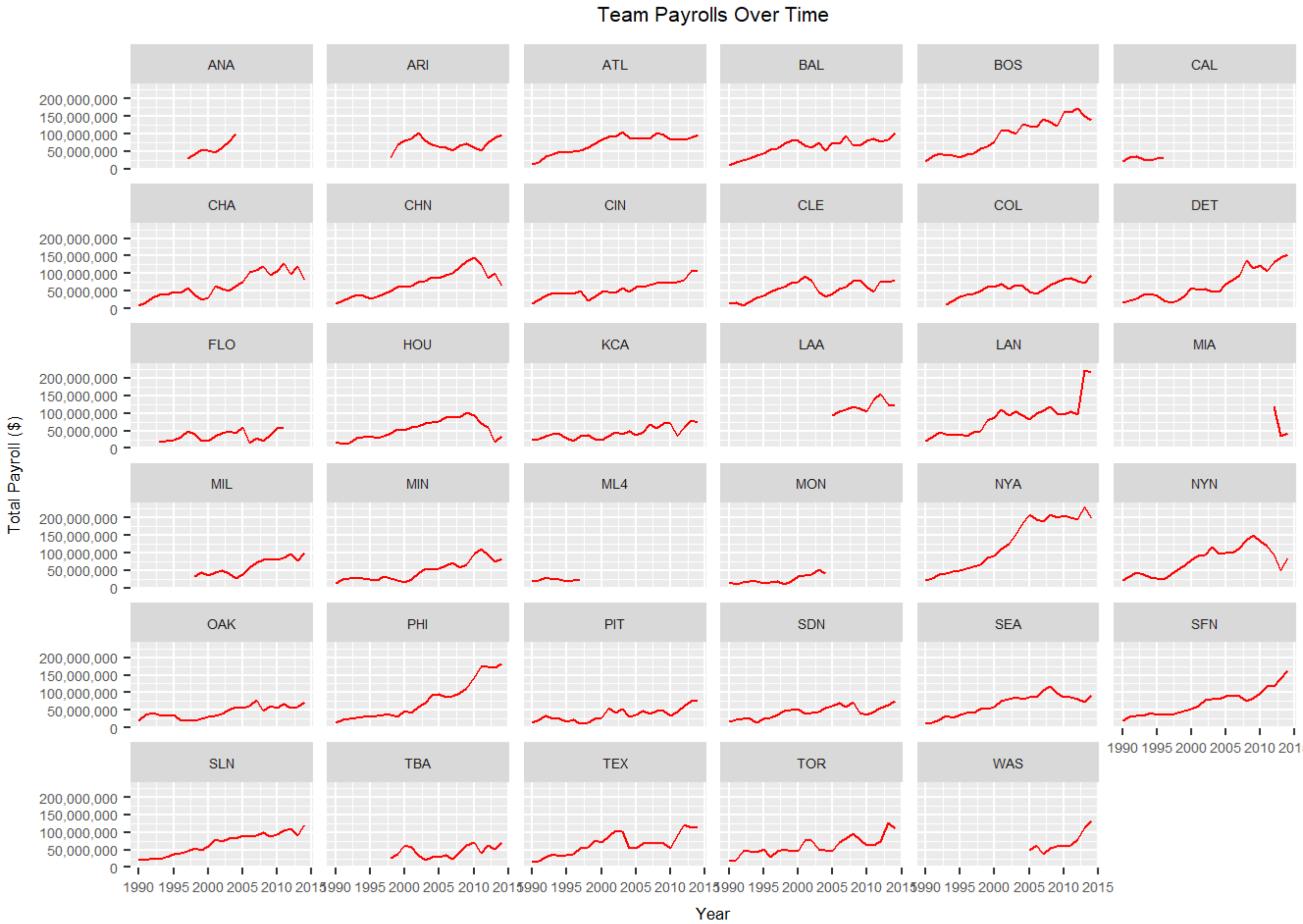
```
payroll_tab %>%
  filter(yearID >= 1990, yearID <= 2014) %>%   # specify years
  group_by(yearID) %>%
  summarize(avg_payroll=mean(payroll)) %>%   # calculate mean payroll
  ggplot(aes(x=yearID, y=avg_payroll)) +  # plot
    geom_point(color="red") +
    scale_y_continuous(labels=comma) +
    labs(title="Average Payroll Over Time", x="Year", y="Average Payroll ($)") +
    theme(plot.title=element_text(hjust=0.5))
```



Plot 2: Team Payrolls Over Time

This plot shows the total payroll for each individual team between 1990 and 2014.

```
payroll_tab %>%
  filter(yearID >= 1990, yearID <= 2014) %>%   # specify years
  ggplot(aes(x=yearID, y=payroll)) +   # plot
    facet_wrap(~teamID) +   # separate plot per team
    geom_line(color="red") +
    scale_y_continuous(labels=comma) +
    labs(title="Team Payrolls Over Time", x="Year", y="Total Payroll ($)") +
    theme(text=element_text(size=6.5), plot.title=element_text(hjust=0.5))
```



Question 1: *What statements can you make about the distribution of payrolls across time based on these plots? Remember you can make statements in terms of central tendency, spread, etc.*

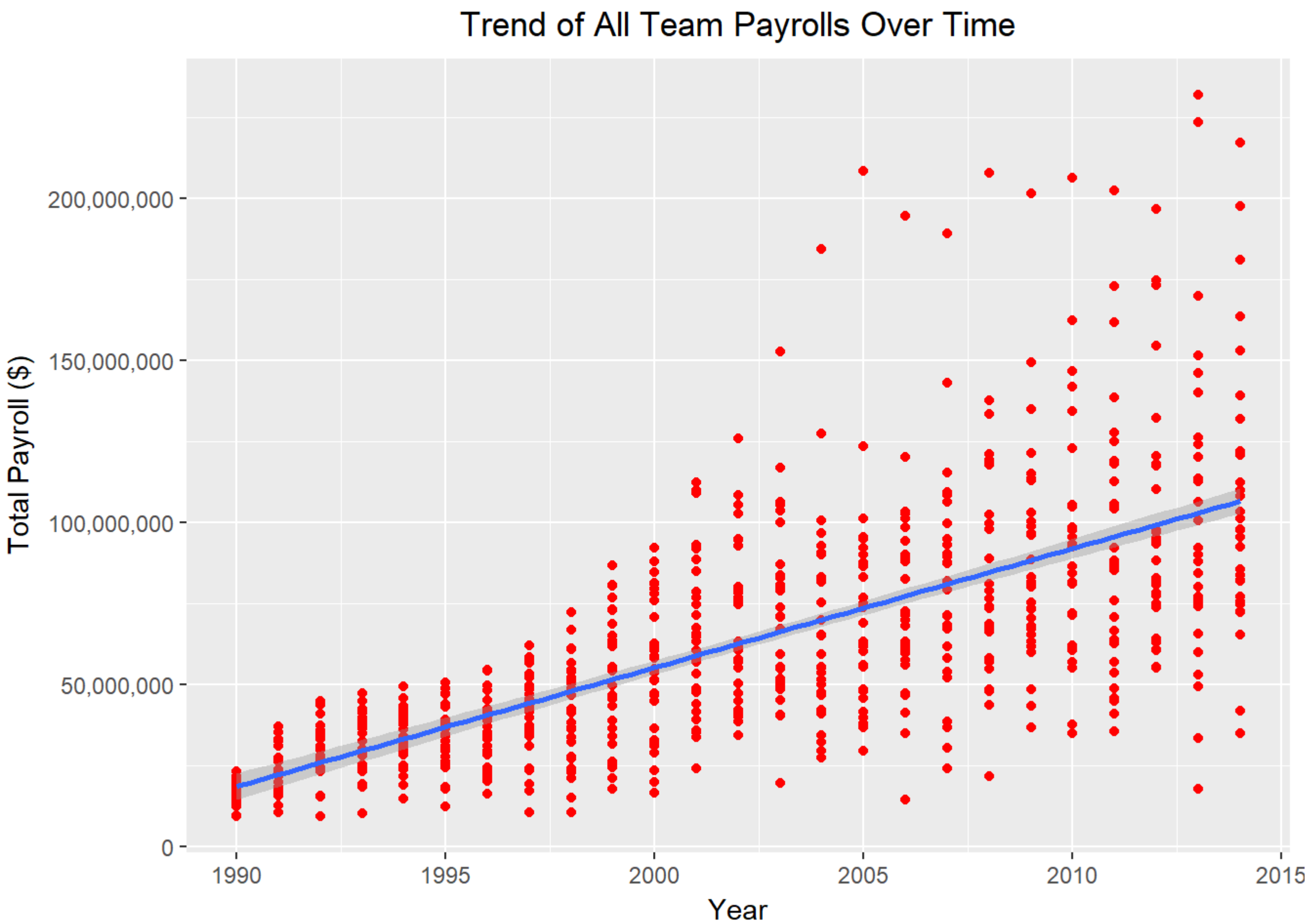
- Plot 1: Mean payroll has increased over time. (which makes sense economically)
- Plot 2: It seems as if average payrolls of teams are increasing over time. (issue: spread and skew are difficult to see)

Problem 3: *Write code to produce a plot(s) that specifically show at least one of the statements you made in Question 1.*

Plot 3: Trend of All Team Payrolls Over Time

This plot takes the data from Plot 2 and converts it to a scatterplot with a trend line for team payrolls between 1990 and 2014.

```
# aggregate plot
payroll_tab %>%
  filter(yearID >= 1990, yearID <= 2014) %>%   # specify years
    ggplot(aes(x=yearID, y=payroll)) +   # plot
      geom_point(color="red") +
      geom_smooth(method="lm") +
      scale_y_continuous(labels=comma) +
      labs(title="Trend of All Team Payrolls Over Time", x="Year", y="Total Payroll ($)") +
      theme(plot.title=element_text(hjust=0.5))
```

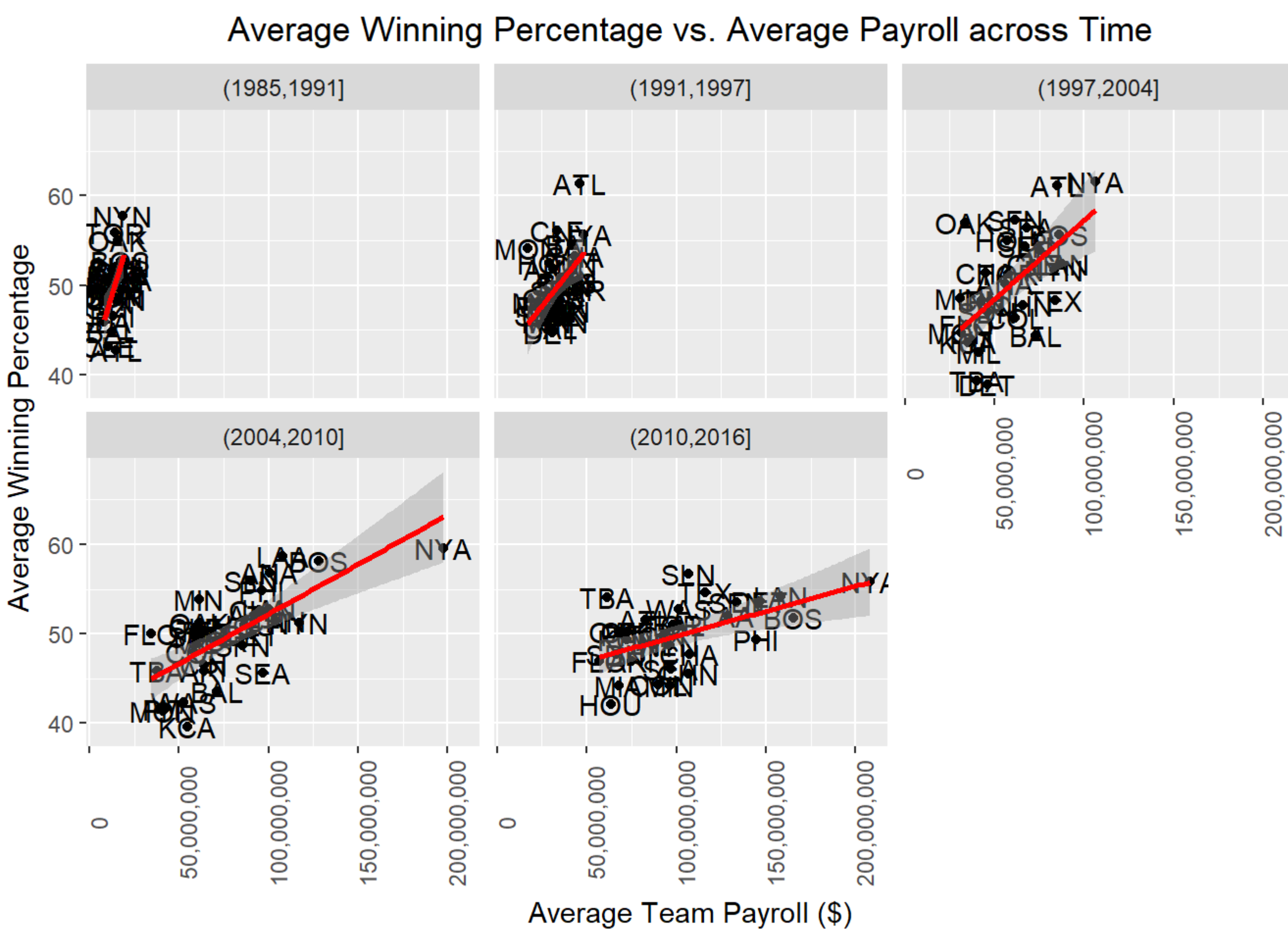


By combining the data from Plot 2 into one plot, we can definitively see that average payrolls of teams are increasing over time. Spread and skew are also visible now. The spread of team payrolls is increasing across most, if not all teams over time. There is also some skew among payrolls since the value of these teams change over time, affecting the salaries the teams are able to pay their athletes.

Correlation Between Payroll and Winning Percentage

Problem 4: *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.*

```
# use cut to create 5 time periods
payroll_tab$time_period <- cut(payroll_tab$yearID, breaks=5)

# data frame of all teams with average payroll and average win percentage
mean_stats <- payroll_tab %>%
  group_by(time_period, teamID) %>%
  summarize(avg_pay_over_time=mean(payroll), avg_win_percent_over_time=mean(win_percentage, na.r
m=TRUE))

# plot the teams average payroll and win percentage across time periods
mean_stats %>%
  ggplot(aes(x=avg_pay_over_time, y=avg_win_percent_over_time, label=teamID)) +   # plot
    geom_point() +
    geom_text() +
    facet_wrap(~time_period) +
    labs(x="Average Team Payroll ($)",
         y="Average Winning Percentage",
         title="Average Winning Percentage vs. Average Payroll across Time") +
    geom_smooth(method='lm', color="red") + scale_x_continuous(labels=comma) +
    theme(axis.text.x=element_text(angle=90), plot.title=element_text(hjust=0.5))
```



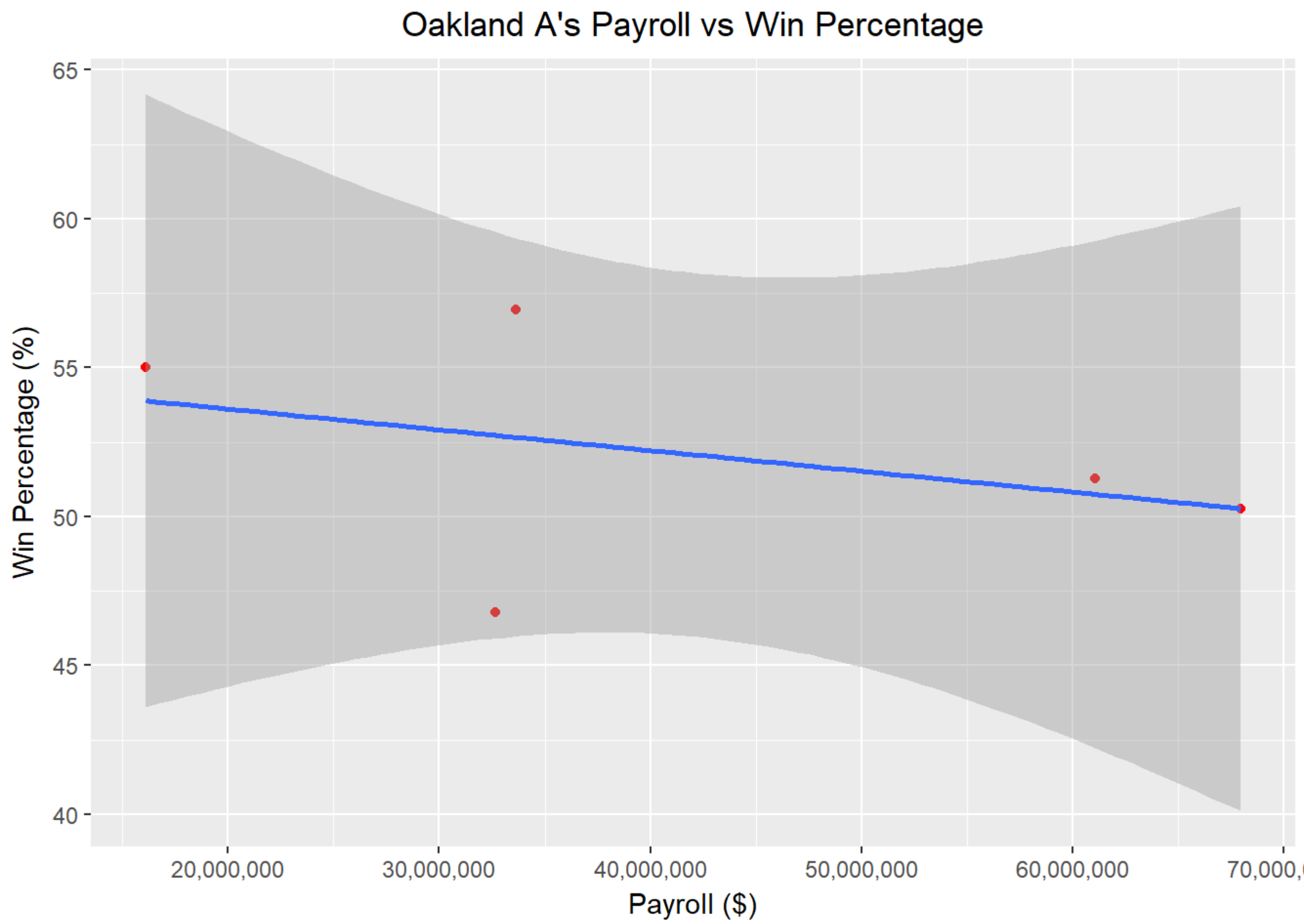
To make this plot, a new attribute (time_period) was created to group teams, years, and win percentages into 5 year ranges. We then created a new data frama (mean_stats) with average payroll and average winning percentage for teams in each year range. This plot is composed of all of the previous data in one plot.

Question 2: *What statements can you make about the distribution of payrolls across time based on these plots? Remember you can make statements in terms of central tendency, spread, etc.*

- It seems as if the spread of average payroll increases over time since the teams are paying their players more as time goes on.
- Througout year ranges the trend line gets less steep over time period, showing that spending more money on players is more likely to result in a team winning more games.
- The New York Yankees have the overall highest payroll which translates into them having the highest win percentage.
- Team payroll definitely increases over time.
- The Oakland A’s had a high win percentage while spending much less money than other teams

Analyzing the Oakland A’s:

```
# plot this data for just the A's
mean_stats %>%
  filter(teamID == "OAK") %>%
  ggplot(aes(x=avg_pay_over_time, y=avg_win_percent_over_time)) +
    geom_point(color="red") +
    geom_smooth(method=lm) +
    scale_x_continuous(labels=comma) +
    labs(title="Oakland A's Payroll vs Win Percentage", x="Payroll ($)", y="Win Percentage (%)")
 +
    theme(plot.title=element_text(hjust=0.5))
```



The Oakland A’s spending efficiency peaked in the 1997-2002 time period, giving them a high win percentage for a low cost. Following this time period however, in 2003-2007 the A’s had their worst spending effiency leading to a negative trend between spending money and winning games. The Oakland A’s started off like all of the other teams from 1985 to 1997, but in 1997 to 2002, the they were doing significantly better than other teams who were spending the same amount of money as them. This has leveled off over time.

Data Transformations:

Standardization Across Years:

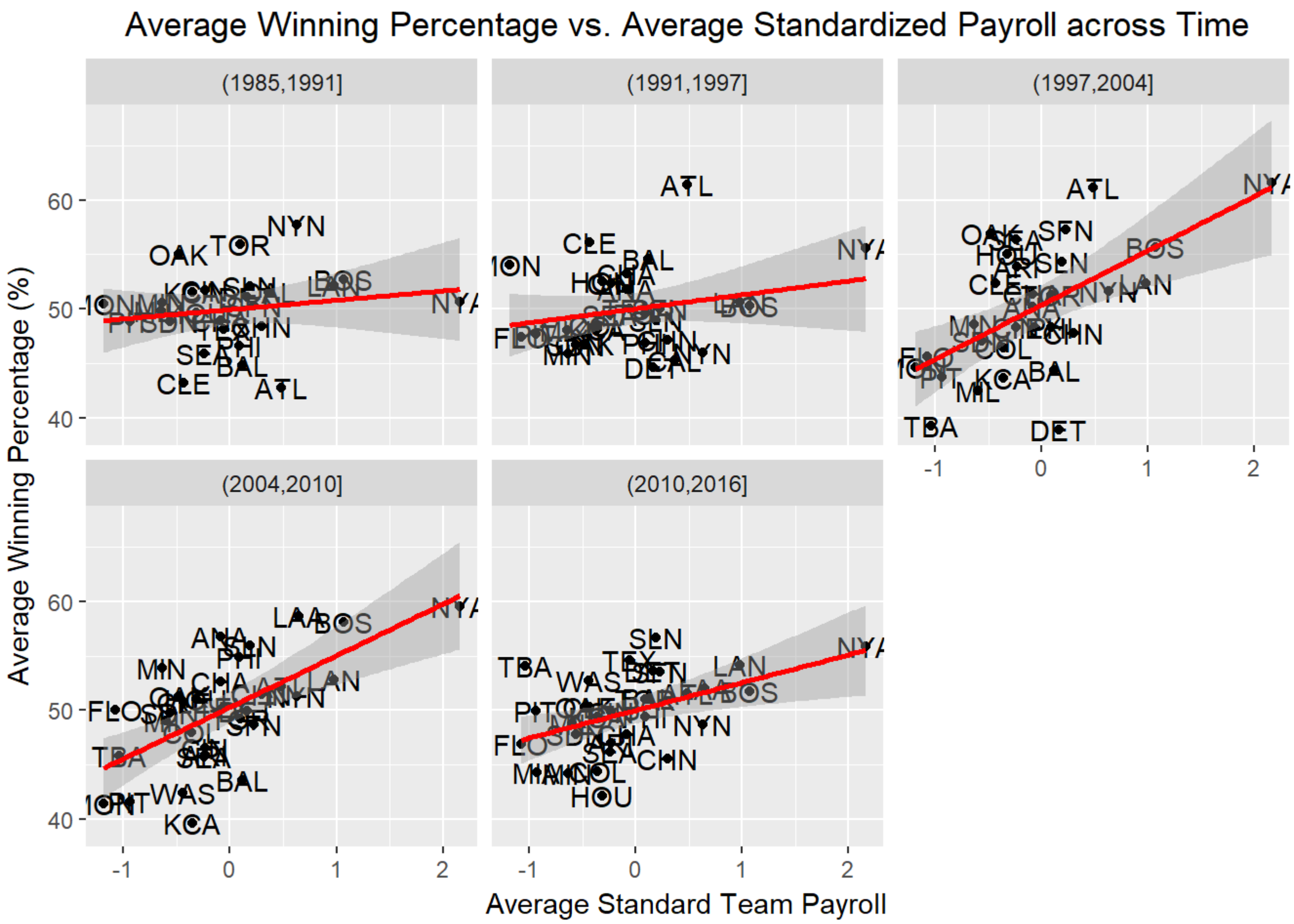
Problem 5: *Write dplyr code to create a new variable in your dataset that standardizes payroll conditioned on year.*

```
# year, average, and standard deviation for payrolls by team
team_payrolls <- payroll_tab %>%
  group_by(yearID, teamID) %>%
  summarize(team_payroll=sum(payroll)) %>%
  inner_join(mean_stats)

avg_payroll_tab <- team_payrolls %>%
  group_by(yearID) %>%
  summarize(mean_payroll=mean(team_payroll), sd_payroll=sd(team_payroll))

# standardized data table
std_tab <- inner_join(team_payrolls, avg_payroll_tab) %>%
  mutate(std_payroll=((team_payroll - mean_payroll) / sd_payroll))

# view it
sample_n(std_tab, 10)
```

```
## # A tibble: 320 x 9
## # Groups:   yearID [32]
##    yearID teamID team_payroll time_period avg_pay_over_time
##     <int> <chr>         <dbl> <fct>                   <dbl>
##  1   1985 PIT        9227500. (1985,1991]         12235667.
##  2   1985 BAL       11560712. (1985,1991]         12495510.
##  3   1985 CHA        9846178. (2010,2016]        107417996.
##  4   1985 SDN       11036583. (2004,2010]         60615615.
##  5   1985 BOS       10897560. (1991,1997]         39499670.
##  6   1985 CHN       12702917. (1991,1997]         35040975.
##  7   1985 PHI       10124966. (2004,2010]         96171106.
##  8   1985 KCA        9321179. (2010,2016]         72471043.
##  9   1985 ML4       11284107. (1991,1997]         23725861.
## 10   1985 TEX        7676500. (2004,2010]         63889646.
## # ... with 310 more rows, and 4 more variables:
## #   avg_win_percent_over_time <dbl>, mean_payroll <dbl>, sd_payroll <dbl>,
## #   std_payroll <dbl>
```

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

```
# plot the average payroll and win percentage across time period for teams
std_tab %>%
  group_by(time_period, teamID) %>%
  summarize(avg_pay=mean(std_payroll),
            avg_win_percentage=mean(avg_win_percent_over_time,
            na.rm=TRUE)) %>%
    ggplot(aes(x=avg_pay, y=avg_win_percentage, label=teamID)) +
      geom_point() +
      geom_text() +
      facet_wrap(~time_period) +
      labs(x="Average Standard Team Payroll",
           y="Average Winning Percentage (%)",
           title="Average Winning Percentage vs. Average Standardized Payroll across Time") +
      geom_smooth(method='lm', color="red") +
      theme(plot.title=element_text(hjust=0.5))
```



In this plot we again grouped teams by their time period and teamID and plotted the data based on the average standard payrolls and the average win percentage.

Question 3: *Discuss how the plots from Problem 4 and Problem 6 reflect the transformation you did on the payroll variable.*

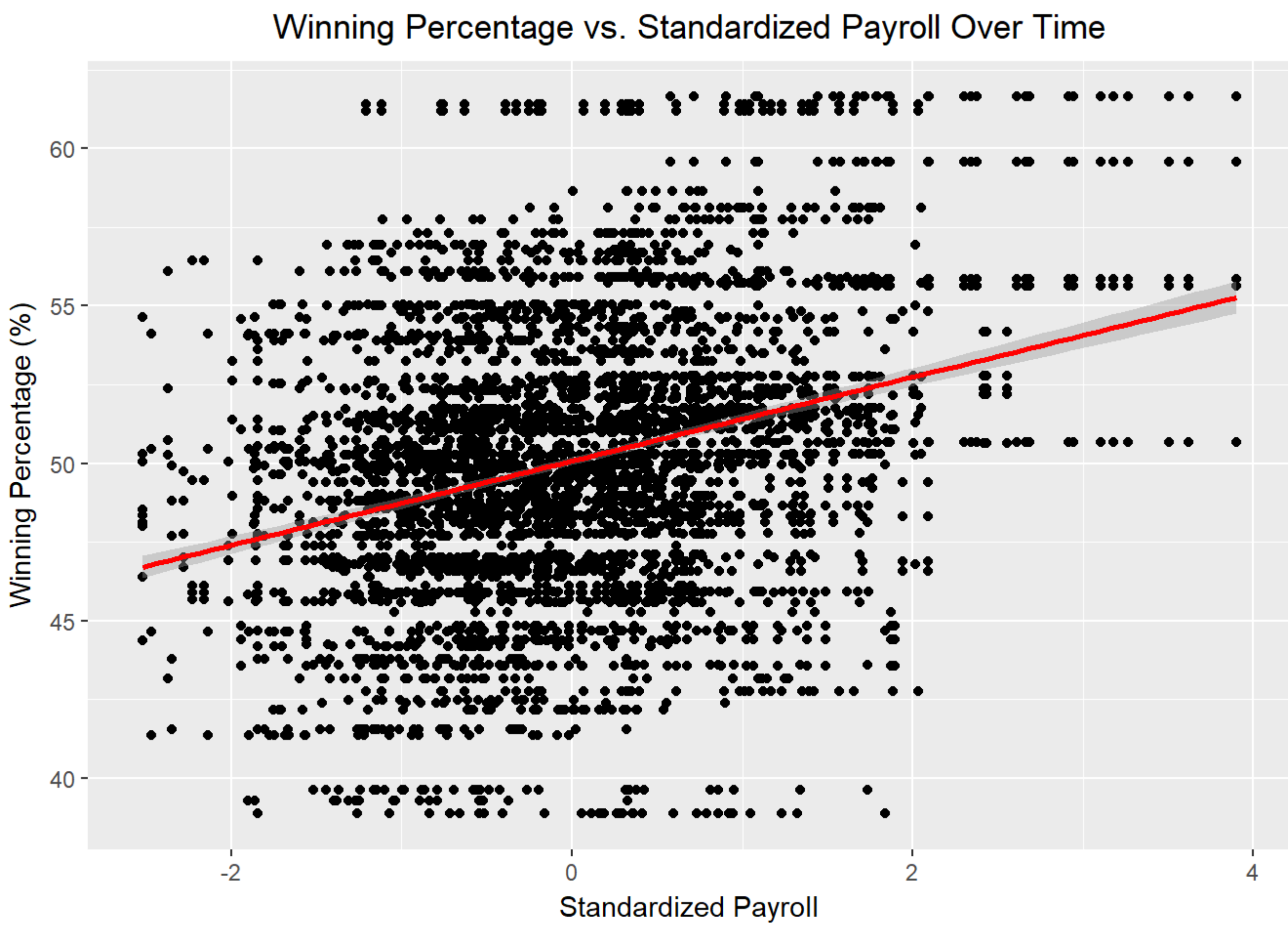
These new plots represent the transformation on payroll. Each data point is relative to the others on a standardized scale. The normalization of the data based on mean payroll centers the data at 0 and makes the standard deviation 1. This normalization allows us to easily see if a team in a specific time period has an above average payroll resulting in a certain winning percentage.

The plots in Problem 4 were difficult to interpret since all of the teams had different mean payrolls and standard deviations, making it hard to compare data across time periods. The standardized transformation makes this easier since they are all on the normal scale.

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.*

```
# plot
std_tab %>%
  ggplot(aes(x=std_payroll, y=avg_win_percent_over_time)) +
    geom_point() +
    geom_smooth(method=lm, color="red") +
    labs(x="Standardized Payroll", y="Winning Percentage (%)", title="Winning Percentage vs. Sta
ndardized Payroll Over Time") +
    theme(plot.title=element_text(hjust=0.5)) +
    scale_x_continuous(labels=comma)
```



This data, specifically the regression line, shows that if a team spends the average payroll on players, they will likely win approximately 50% of their games on average.

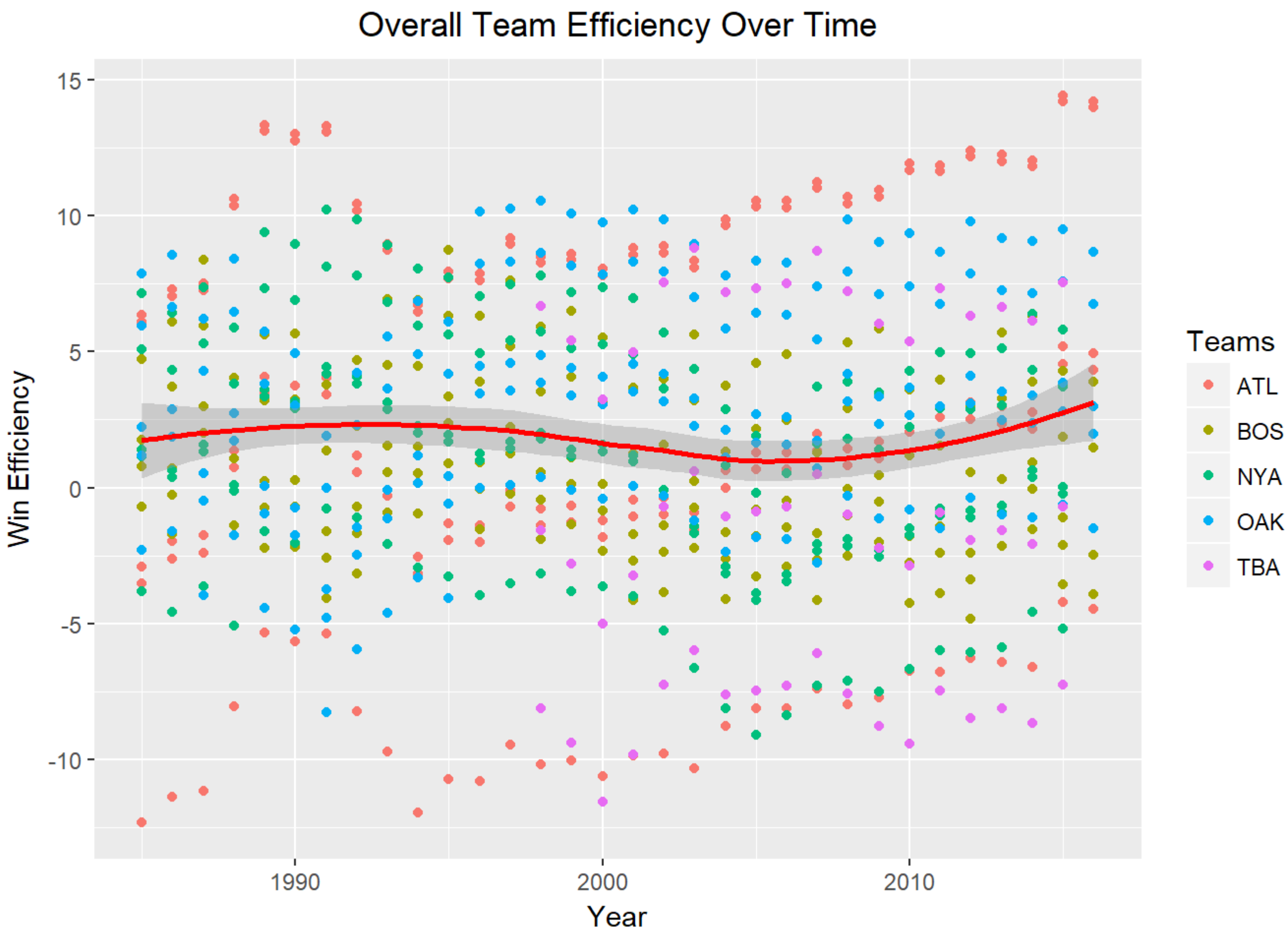
Spending Efficiency:

Problem 8: *Write dplyr code to calculate spending efficiency for each team. Make a line plot with year on the x-axis and efficiency on the y-axis.*

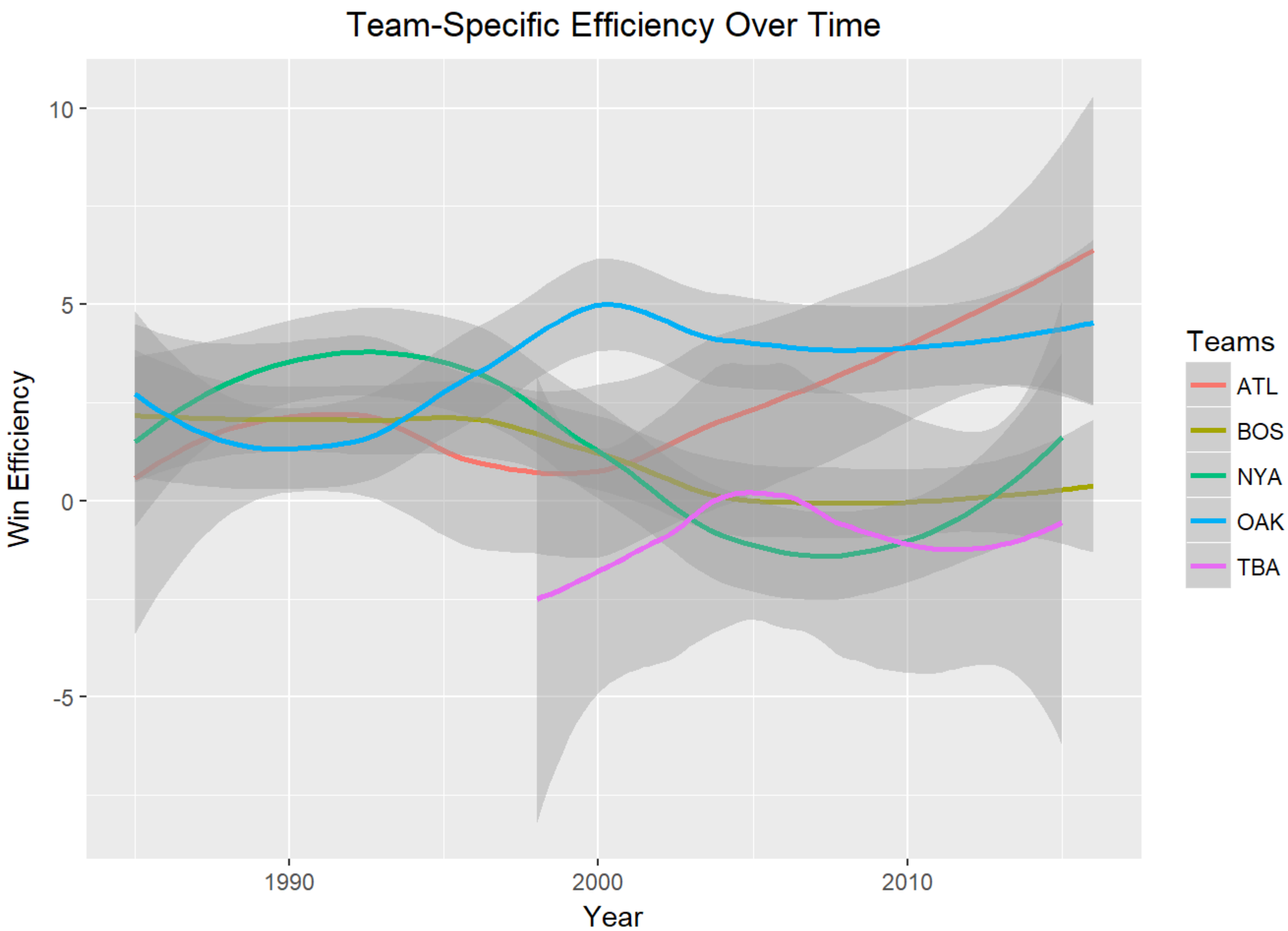
```
# calculate expected winning percentage
std_tab <- std_tab %>%
  mutate(exp_win_percentage=(50 + 2.5  * std_payroll))

# view it
sample_n(std_tab %>% select(teamID, yearID, avg_win_percent_over_time, exp_win_percentage), 10)
```

```
## # A tibble: 320 x 4
## # Groups:   yearID [32]
##    teamID yearID avg_win_percent_over_time exp_win_percentage
##    <chr>   <int>                     <dbl>              <dbl>
##  1 NYA      1985                      55.9               54.5
##  2 MIN      1985                      44.2               45.6
##  3 TEX      1985                      48.1               47.6
##  4 BOS      1985                      58.1               51.0
##  5 NYA      1985                      50.7               54.5
##  6 CHA      1985                      52.6               49.9
##  7 PIT      1985                      49.9               49.3
##  8 MON      1985                      50.4               49.5
##  9 ATL      1985                      42.8               55.1
## 10 CHN      1985                      51.5               52.9
## # ... with 310 more rows
```

```
# calculate efficiency
std_tab <- std_tab %>%
  mutate(efficiency=avg_win_percent_over_time - exp_win_percentage)

# overall efficiency over time
std_tab %>%
  filter(teamID %in% c("OAK", "BOS", "NYA", "ATL", "TBA")) %>%
  ggplot(aes(x=yearID, y=efficiency)) +
    geom_point(aes(color=teamID)) +
    geom_smooth(color="red") +
    labs(x="Year", y="Win Efficiency", title="Overall Team Efficiency Over Time", color="Teams")
 +
    theme(plot.title=element_text(hjust=0.5))
```




```
# team specific efficiency over time
std_tab %>%
  filter(teamID %in% c("OAK", "BOS", "NYA", "ATL", "TBA")) %>%
    ggplot(aes(x=yearID, y=efficiency, color=teamID)) +
      geom_smooth() +
      labs(x="Year", y="Win Efficiency", title="Team-Specific Efficiency Over Time", color="Team
s") +
      theme(plot.title=element_text(hjust=0.5))
```



The expected winning percentage of a team is defined as 50 + 2.5 * the standard payroll. The efficiency of a team is defined as their winning percentage - expected winning percentage. The regression line shows that on average, a team's winning percentage is correlated to the amount spent on their players (payroll). If a team wins more than the expected value, they are above average, if they win less, they are below average.

In the first plot we can observe the average efficiency of 5 teams: Oakland, New York Yankees, Boston, Atlanta, and Tampa Bay.

In the second plot, we can observe how the efficiency of each team has changed over time.

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?*

From these plots we can learn why the Oakland A's were so successful at recruiting afforable players yet still having an above average win rate. Winning efficiency of teams seemed to peak near 2000 and then plateaued after 2005. In questions 2 and 3 we saw that higher payrolls directly correlate to more wins. Oakland is an outlier in this trend. From 2000 to 2005, Oakland the most efficient team, meaning they were able win at an above average rate despite having a below average payroll.

These plots show how spending efficiency changes across teams over time. We can see that among these teams, spending efficiency has increased over the years. There are still some unique trends however. It's interesting that even after the Oakland A's "Moneyball" period, their efficiency trend does not rise at all or fall by much. This could be due to the Oakland A's "Post Moneyball Success" since their incredible efficiency between 1995 and 2000 gave them more money so they did not have to be as efficient as this time period.