Baseball Analytics

Ryan Mavilia

April 1, 2018

```
# make sure you write the path to your sqlite path here
db <- DBI::dbConnect(RSQLite::SQLite(), "lahman2016.sqlite")
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(tidyr)
library(ggplot2)
```

Here we connect to the baseball database and import libraries which will be needed later on.

```
SELECT MAX(100.0 * team.W / team.G) AS winPercentage, SUM(sal.salary) AS payroll, team.W, team.G, team.y
FROM Teams AS team, Salaries AS sal
INNER JOIN teamsFranchises AS teamfran ON
      team.yearID = sal.yearID
      AND team.teamID = sal.teamID
GROUP BY sal.yearID, sal.teamID
```

I run an SQL query which will ask the database to return a table with the calculated payroll & win percentages for the different teams on a per year basis.

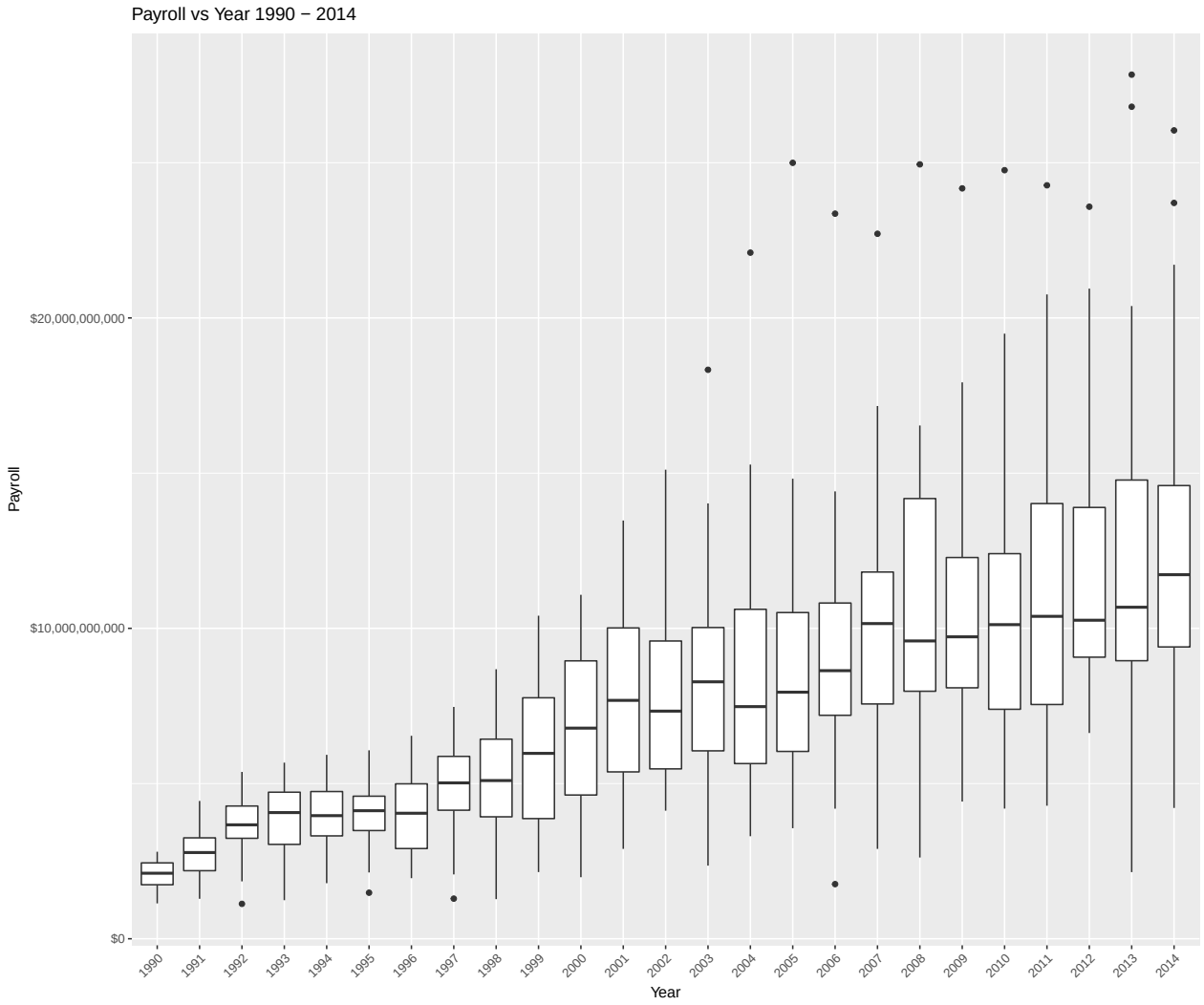
```
payroll_df %>%
    head()
```

```
##   winPercentage    payroll  W   G yearID teamID
## 1      40.74074 1776840000 66 162   1985    ATL
## 2      51.55280 1387285440 83 161   1985    BAL
## 3      49.69325 1307707200 81 163   1985    BOS
## 4      55.55556 1731347280 90 162   1985    CAL
## 5      52.14724 1181541360 85 163   1985    CHA
## 6      47.53086 1524350040 77 162   1985    CHN
```

Payroll distribution

```
payroll_df %>%
  group_by(yearID) %>%
    filter(1990 <= yearID & yearID <= 2014) %>%
    ggplot(mapping=aes(y=payroll, x=factor(yearID))) +
    geom_boxplot() +
```

```
    xlab("Year") +
    ylab("Payroll") +
    ggtitle("Payroll vs Year 1990 - 2014") +
    scale_y_continuous(labels = scales::dollar) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

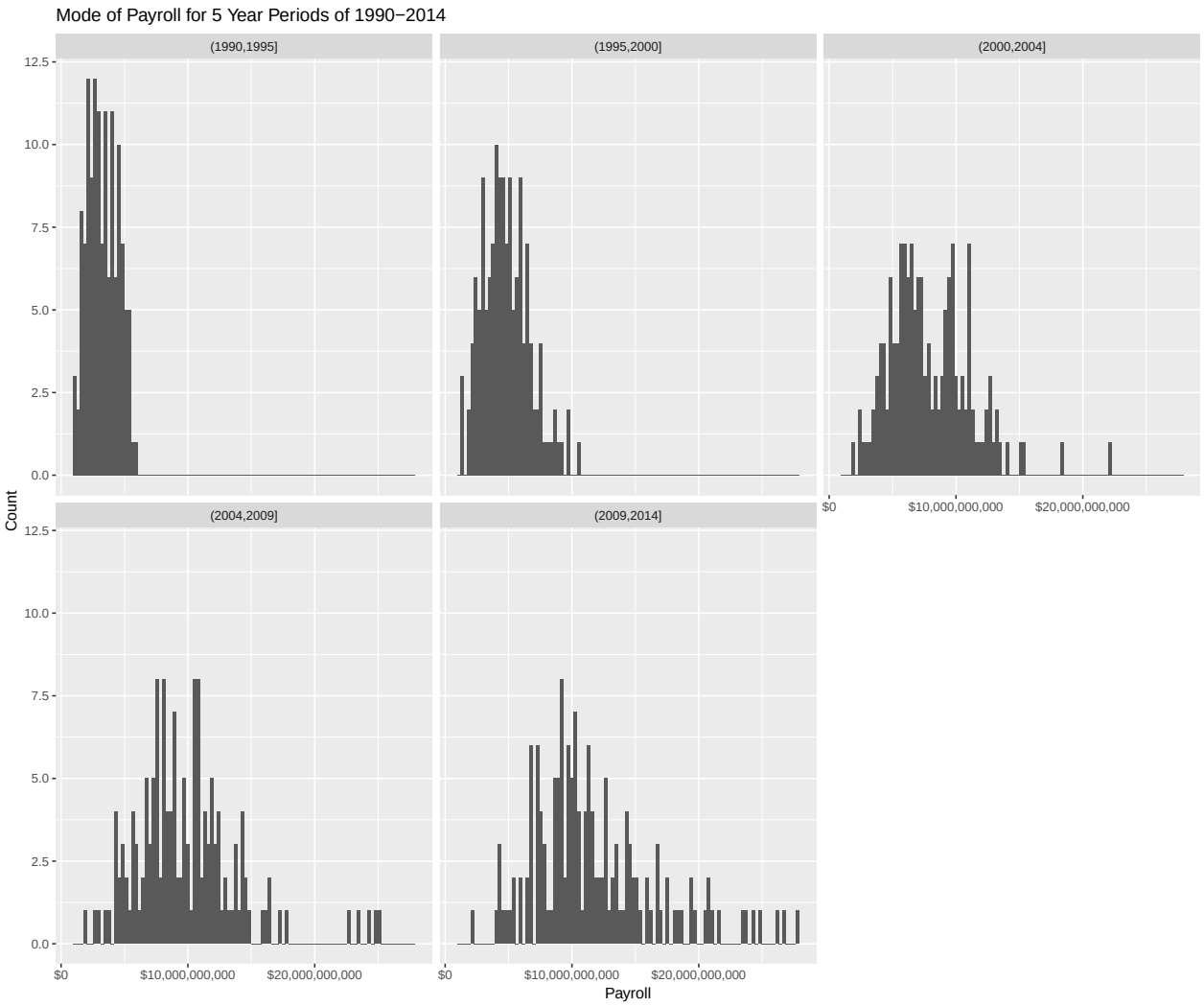


In order to graph the payroll in regards to the year we must first filter out years we don't want which is anything outside of 1990-2014. After that we create a boxplot with payroll as our Y-axis and yearID as our X-axis (using the factor because R reads yearID as a continuous variable). We also label the axes, title, add dolar signs to the payroll labels, and rotate the years to make them easier to read.

As time goes on payrolls for the league have increased in both mean and spread. We can see how large the boxes get towards the the end of the 2000's and that there is an upward trend. I would say that payroll increased greatly from 1990 to 2007 and then evened out which reflects the projects reflection on how things evened out after 2005 when other teams "caught up".

```
payroll_df <- filter(payroll_df, 1990 <= yearID & yearID <= 2014)
plot1 <- payroll_df %>%
  mutate(yearRange = cut(payroll_df$yearID, breaks = 5)) %>%
  group_by(yearRange) %>%
  ggplot(mapping = aes(x=payroll)) +
```

```
  geom_histogram(bins = 100) +
  scale_x_continuous(labels = scales::dollar) +
  xlab(label = "Payroll") +
  ylab(label="Count")+
  ggtitle("Mode of Payroll for 5 Year Periods of 1990-2014")
plot1 + facet_wrap(~yearRange)
```



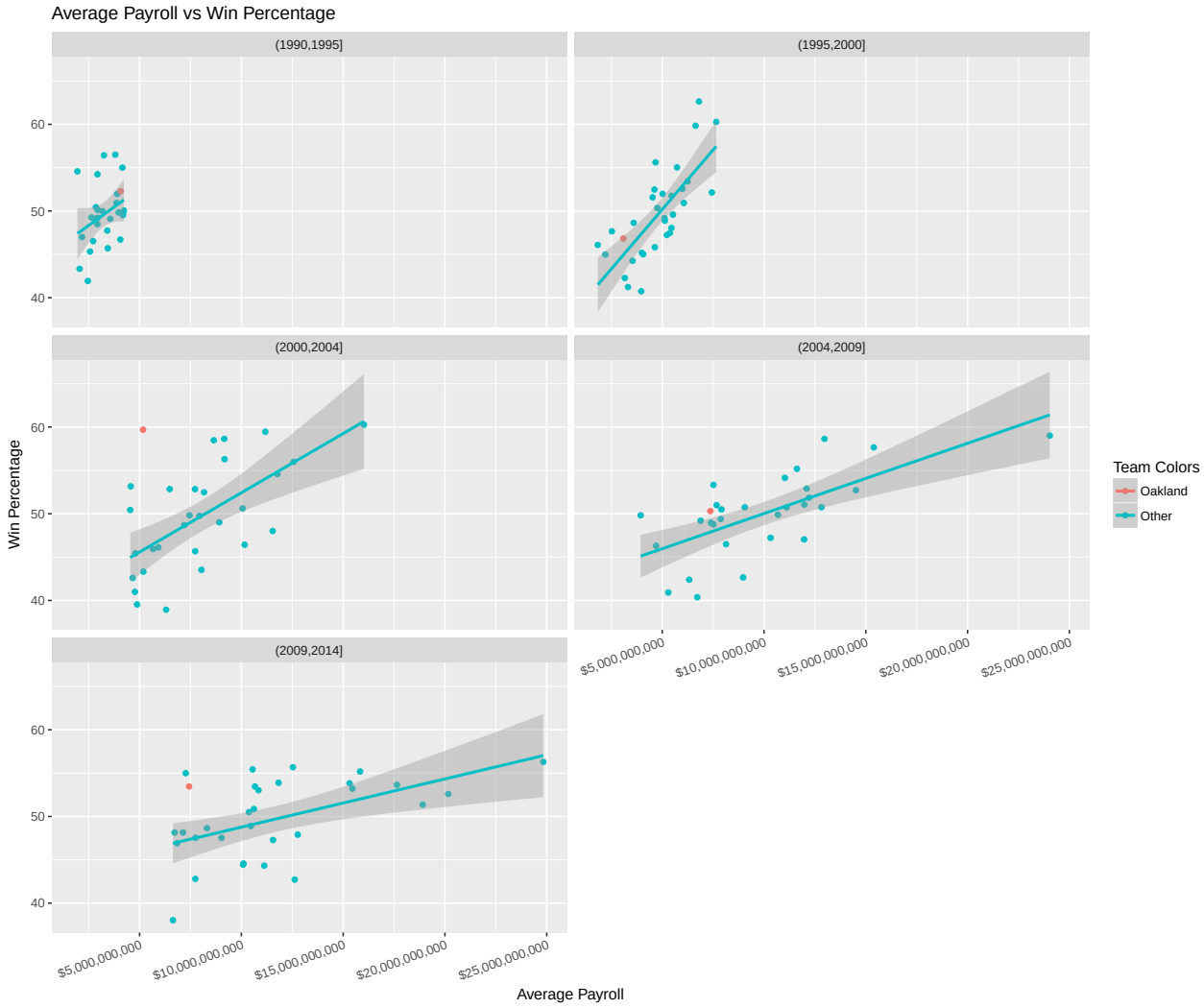
As we can see from these graphs the spread is increasing and the amount being payed annually by teams has a large increase during the 1990's. I used the mean of the payrolls in each range to create the bars. We can see from here that the spread increases dramatically from 1990-2004 and pay increases greatly as well.

Correlation between payroll and winning percentage

```
payroll_df$yearRange <- cut(payroll_df$yearID, breaks=5)

plot1 <- payroll_df %>%
  group_by(teamID, yearRange) %>%
  summarize(m = mean(payroll), n=mean(winPercentage))%>%
    ggplot(mapping=aes(x=m, y=n, col = ifelse(teamID=="OAK", "Oakland", "Other"))) +
```

```
      geom_point() +
      geom_smooth(method=lm) +
     labs(color='Team Colors') +
  theme(axis.text.x = element_text(angle = 20, hjust = 1)) +
    scale_x_continuous(labels = scales::dollar) +
  ylab(label="Win Percentage") +
  xlab(label="Average Payroll") +
    ggtitle("Average Payroll vs Win Percentage")
plot1+facet_wrap(~yearRange, nrow = 3, ncol = 2)
```



I used geom_smooth and scatter plots to show the regression lines of pay vs winningness over the 5 periods between 1990-2014.

The Yankees are the most consistent in terms of spending more and landing near the top of the win percentage axis. As for Oakland their performance varied. Oakland across the time periods: 90-95' high pay low win 95-00' low pay low win 00-04' low pay very high win 05-09' low pay low win 09-14' low pay high win

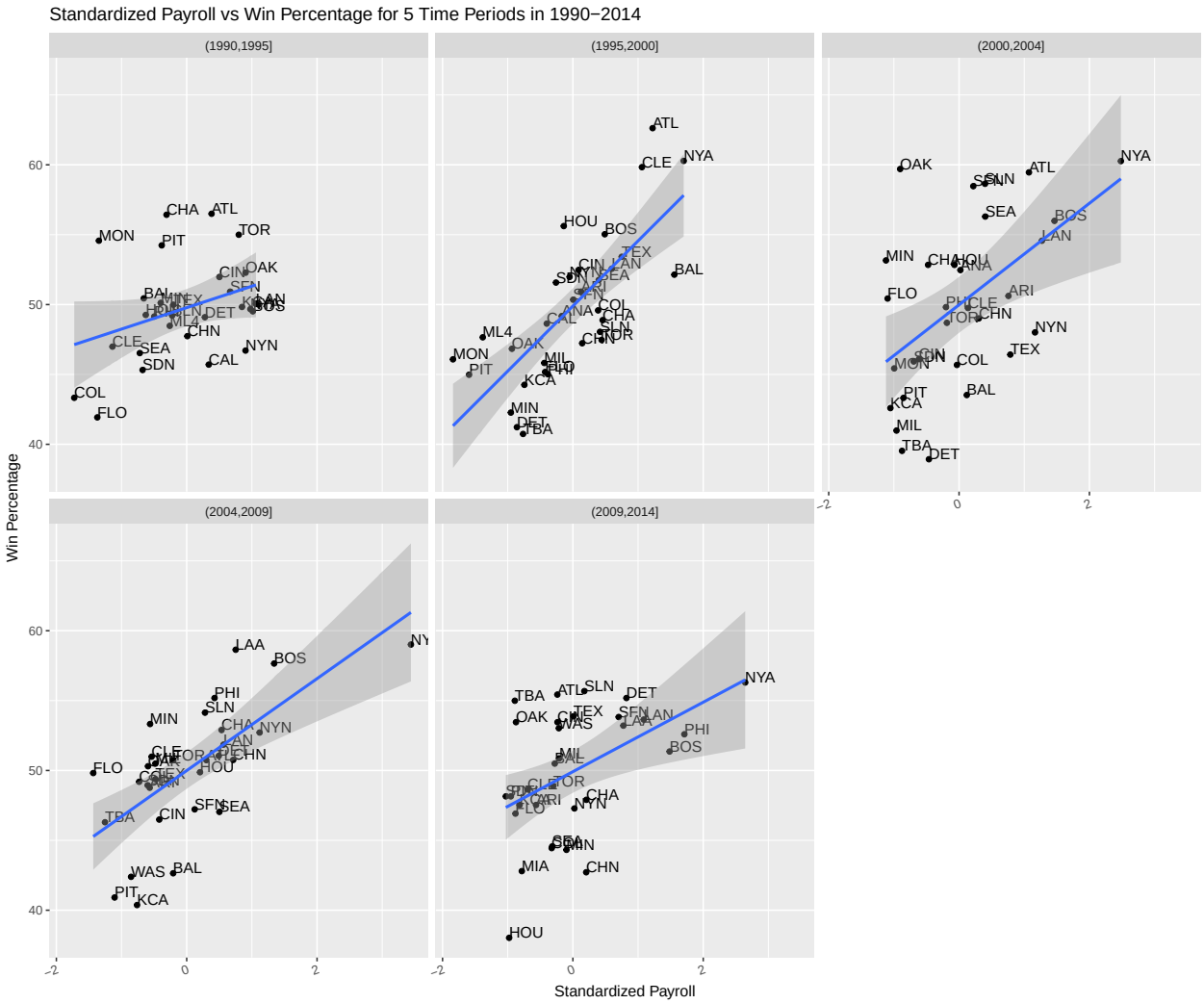
Data transformations

Standardization across years

```
payroll_df <- payroll_df %>%
  group_by(yearID) %>%
  mutate(standardized_payroll = (payroll-mean(payroll))/sd(payroll))
payroll_df[,c(5:8)]
```

```
## # A tibble: 728 x 4
## # Groups:   yearID [25]
##    yearID teamID yearRange   standardized_payroll
##     <int> <chr>  <fct>                      <dbl>
##  1   1990 ATL    (1990,1995]               -0.667
##  2   1990 BAL    (1990,1995]               -1.96
##  3   1990 BOS    (1990,1995]                0.924
##  4   1990 CAL    (1990,1995]                1.23
##  5   1990 CHA    (1990,1995]               -2.01
##  6   1990 CHN    (1990,1995]               -0.914
##  7   1990 CIN    (1990,1995]               -0.716
##  8   1990 CLE    (1990,1995]               -0.685
##  9   1990 DET    (1990,1995]                0.138
## 10   1990 HOU    (1990,1995]                0.333
## # ... with 718 more rows
```

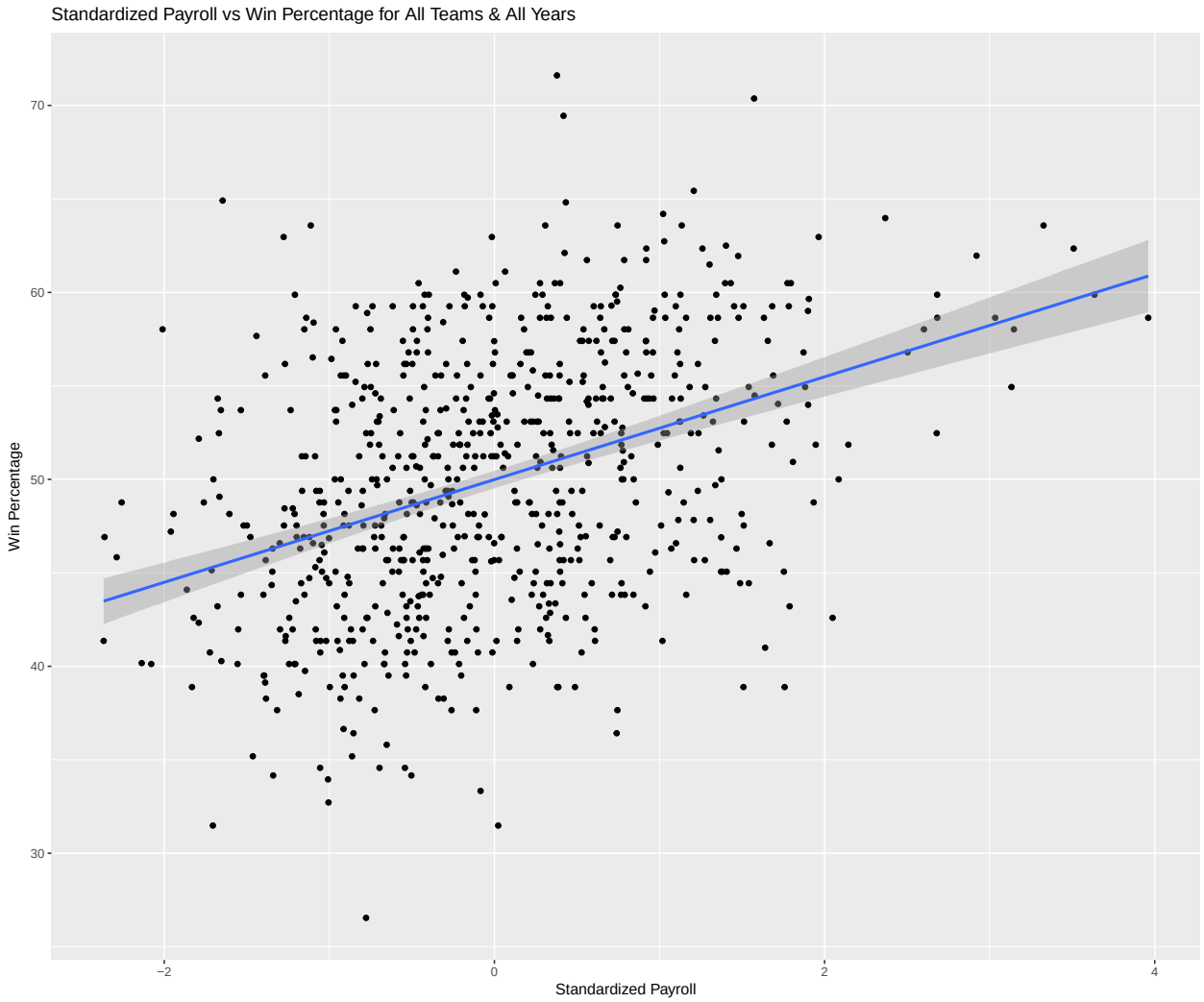
```
plot5 <- payroll_df %>%
  group_by(teamID, yearRange) %>%
  summarise(n = mean(standardized_payroll), m = mean(winPercentage)) %>%
    ggplot(aes(x=n, y=m)) +
      geom_point() +
      geom_text(aes(label=teamID),hjust=0, vjust=0) +
      geom_smooth(method=lm) +
  theme(axis.text.x = element_text(angle = 20, hjust = 1)) +
  xlab(label="Standardized Payroll") +
  ylab(label="Win Percentage") +
    ggtitle("Standardized Payroll vs Win Percentage for 5 Time Periods in 1990-2014")
plot5+facet_wrap(~yearRange)
```



Here I have recomoputed the same plots from problem 4 with the new standardized payroll data.

This shows us which teams were really spending the most vs least and how that affected their wins. For example in 2000-2004 Oakland is spending way less than other teams but still winning a lot. We can see this because Oakland is far to the left denoting a they are to the left of the continuous payroll plot so they are spending the least. It works the same as a line graph where the farther left the lower your number (payroll) is.

```
payroll_df %>%
  ggplot(mapping = aes(x=standardized_payroll, y = winPercentage)) +
  geom_point() +
  geom_smooth(method = lm)+
  xlab(label="Standardized Payroll") +
  ylab(label="Win Percentage") +
    ggtitle("Standardized Payroll vs Win Percentage for All Teams & All Years")
```



Here I created a plot using all of the points and we can see that as spending increases the win percentage increases as well.

```
payroll_df <- payroll_df %>%
  mutate(expected_win_pct = (50 + 2.5 * standardized_payroll)) %>%
  mutate(efficiency = winPercentage - expected_win_pct)

payroll_df
```

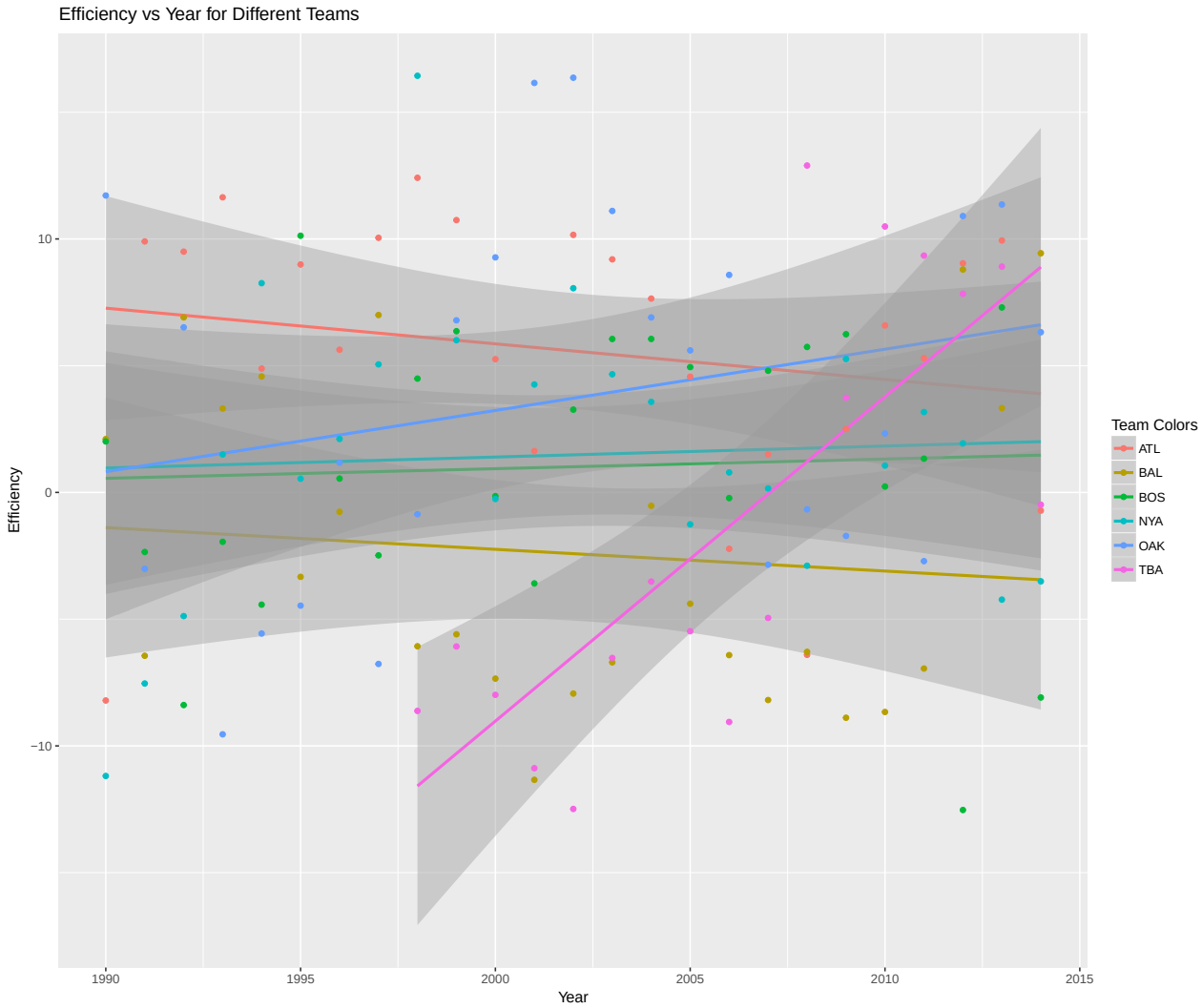
```
## # A tibble: 728 x 10
## # Groups:   yearID [25]
##    winPercentage    payroll     W     G yearID teamID yearRange
##            <dbl>      <dbl> <int> <int>  <int> <chr>  <fct>
##  1          40.1 1746660120    65   162   1990 ATL    (1990,1995]
##  2          47.2 1161610080    76   161   1990 BAL    (1990,1995]
##  3          54.3 2466999960    88   162   1990 BOS    (1990,1995]
##  4          49.4 2606400000    80   162   1990 CAL    (1990,1995]
##  5          58.0 1138980000    94   162   1990 CHA    (1990,1995]
##  6          47.5 1634880000    77   162   1990 CHN    (1990,1995]
##  7          56.2 1724400000    91   162   1990 CIN    (1990,1995]
##  8          47.5 1738440000    77   162   1990 CLE    (1990,1995]
##  9          48.8 2111188560    79   162   1990 DET    (1990,1995]
```

```
## 10          46.3 2199600000    75   162   1990 HOU    (1990,1995]
## # ... with 718 more rows, and 3 more variables:
## #   standardized_payroll <dbl>, expected_win_pct <dbl>, efficiency <dbl>
```

I've created the expected win percentage and efficiency calculations based on the formulas given using the dplyr mutate function.

Spending efficiency

```
payroll_df %>%
  filter(teamID %in% c("OAK", "BOS", "BAL", "NYA", "ATL", "TBA")) %>%
  ggplot(mapping = aes(x=yearID, y=efficiency, color = teamID)) +
  geom_smooth(method=lm) +
  geom_point() +
  labs(color='Team Colors') +
  xlab(label="Year") +
  ylab(label="Efficiency") +
  ggtitle("Efficiency vs Year for Different Teams")
```



Here I've created a plot with the efficiency mapped against the year for the 5 teams mentioned in the

documentation as well as the Baltimore Orioles since they are my favorite team. I used geom_smooth as specified and also provided coloring.

The Oakland Athletics did extremely well during the MoneyBall year. This shows just how well their strategy worked for them before others caught up such as Tampa Bay in 2010. I think this would do really well as a 3D graph with another factor added in such as team size, batting average, etc. so that we can see how efficiency interleaves with other factors.