Project2

Joseph Kahadze

March 27, 2019

SQL

PROBLEM 1

To calculate the total payroll and the winning percentage I joined the Salaries and Teams tables on the teamID and the yearID and then filtered all the enteties with yearID >= 1990 or less than <= 2014.

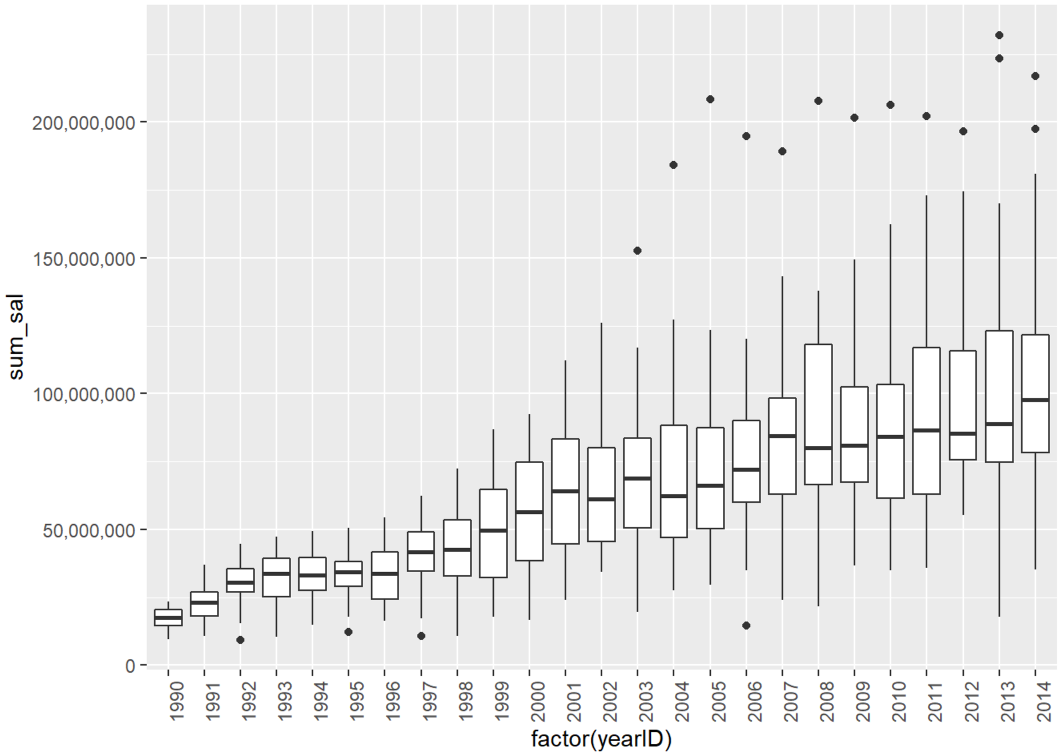
I used an inner join to filter out values with missing data so that only values with corresponding yearID and teamID. Furthermore, I reviewed the table myself to make sure there were no missing enteries.

```
SELECT t.teamID, t.franchID, t.yearID, t.W, t.L, t.G, s.sum_sal, ((t.W * 1.0)/(t.G * 1.0)*100.0) AS WPERC FR
OM Teams as t inner join (SELECT teamID, yearID, sum(salary) as sum_sal FROM Salaries GROUP BY teamID, yearI
D) as s ON t.teamID = s.teamID and t.yearID = s.yearID WHERE t.yearID >= 1990 and t.yearID <=2014
```

PROBLEM 2

I used a boxplot to graph the distribution of payrolls across teams because it visually shows the mean and variance of the payroll per year.

```
sum_df <- payroll_df

sum_df %>%
  ggplot(mapping=aes(x=factor(yearID), y=sum_sal)) +
    geom_boxplot() + scale_y_continuous(labels = scales::comma) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



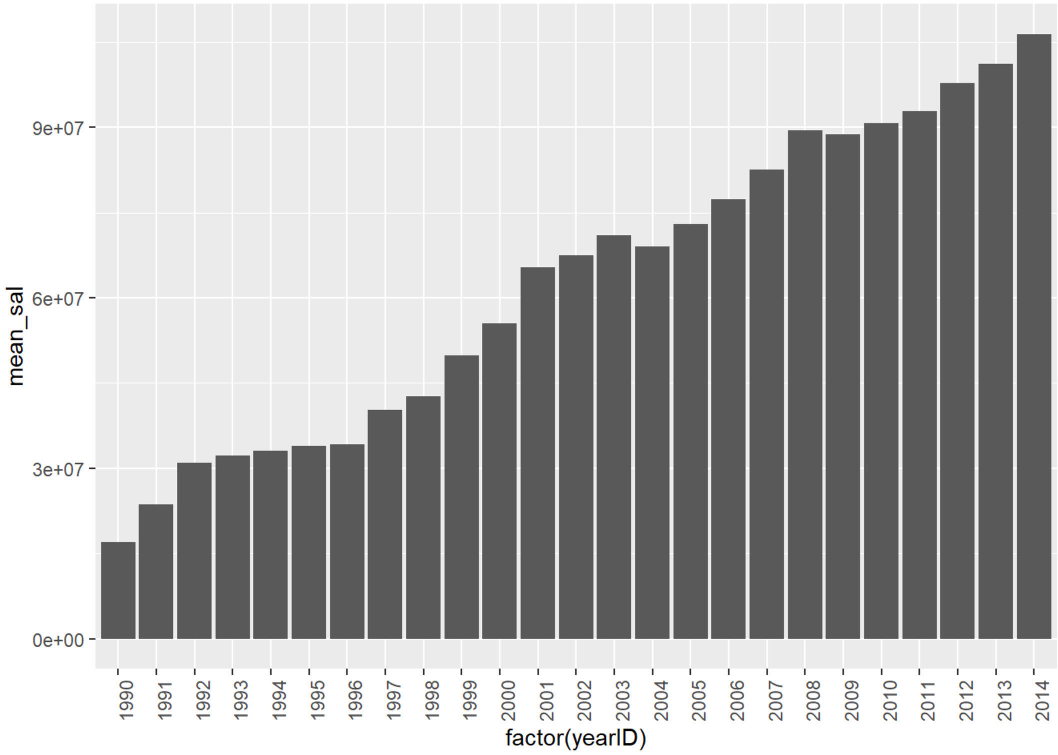
Question 1

The mean payroll in the MLB increased from 1990 to 2014 from about 25,000,000 to about 100,000,000. Furthermore, the variation in payroll between teams also increased from 1990 to 2014.

PROBLEM 3

I calculated the mean payroll for each year and then plotted it for every year to show that the payroll has a tendency to increase over time.

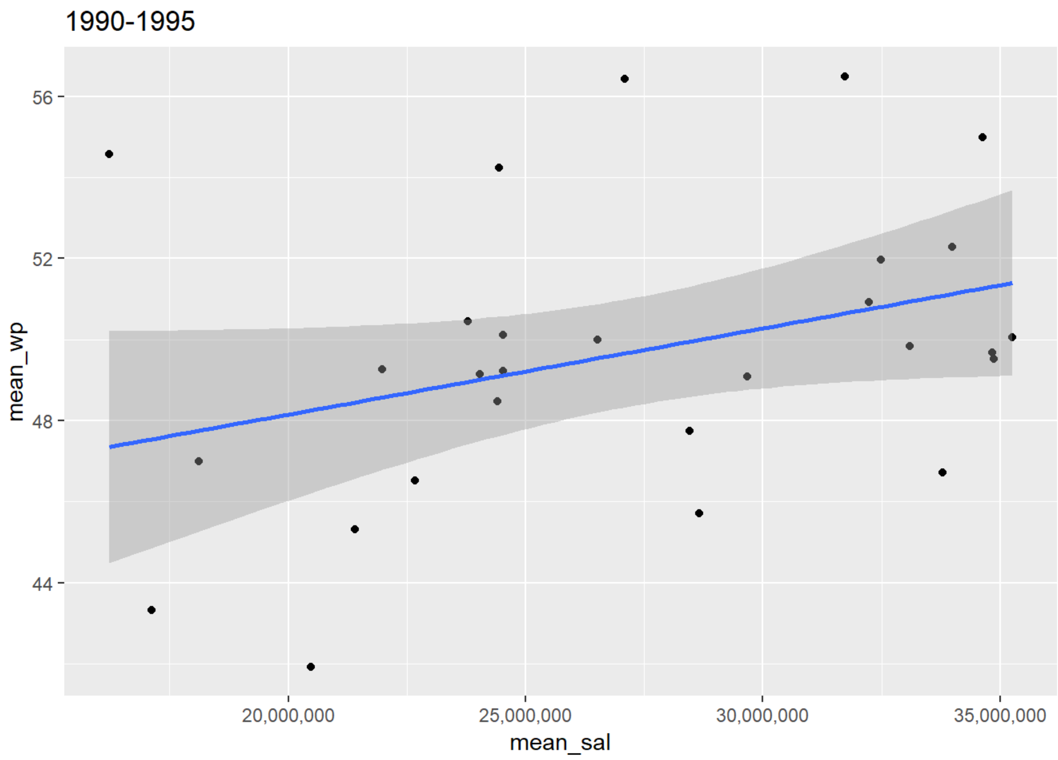
```
mean_df <- payroll_df %>%
  group_by(yearID) %>%
  dplyr::summarise(mean_sal = mean(sum_sal))


mean_df %>%
  ggplot(mapping=aes(x=factor(yearID), y=mean_sal)) +
    geom_bar(stat="identity")  +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



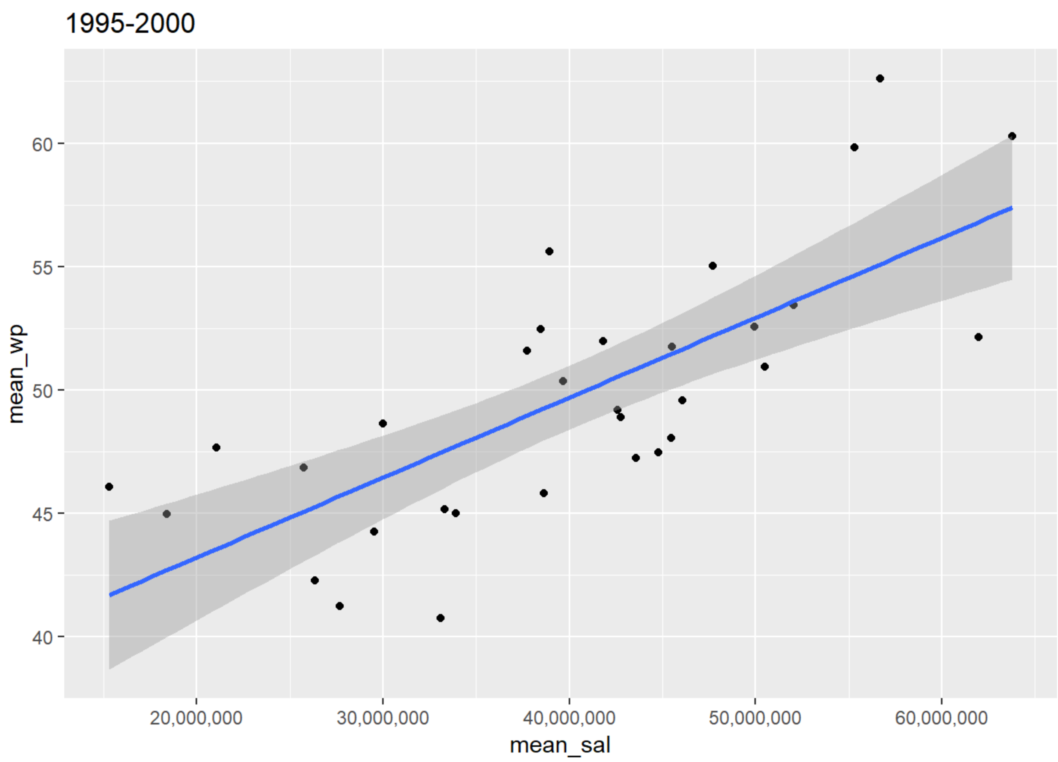
PROBLEM 4

I split the payroll_df into five categories based on yearID in 5 year ranges. Then I calculated the mean winning percentage and mean payroll for each of the 5 tables and plotted them.

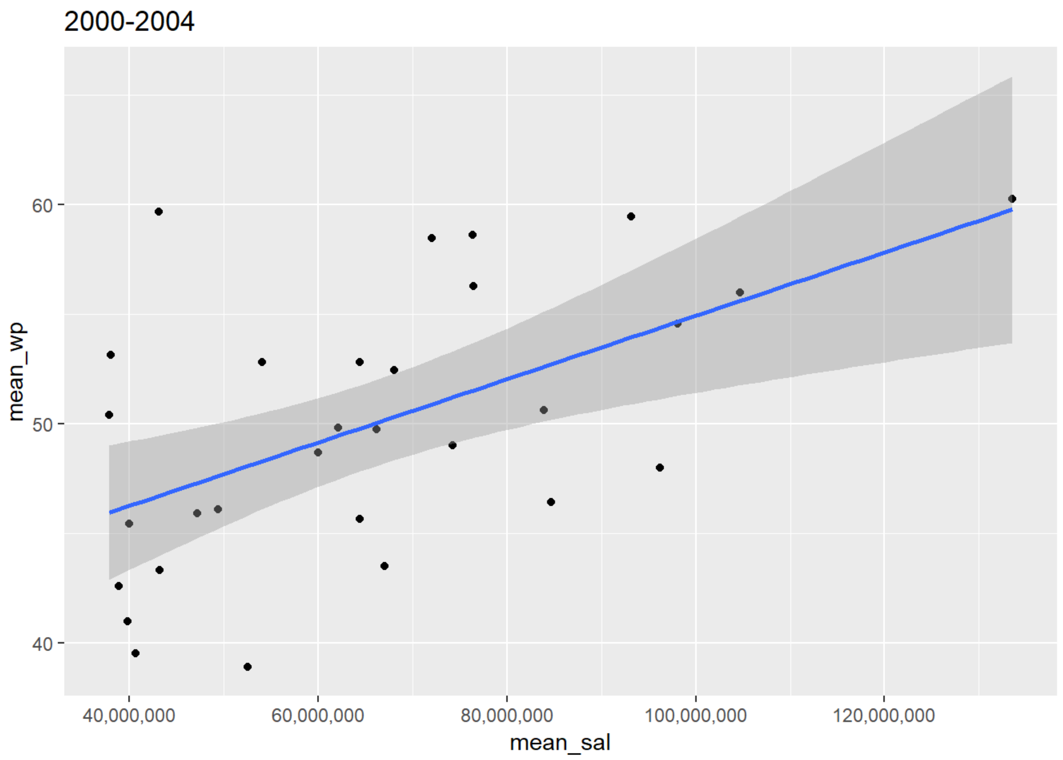
```
cut_df <- payroll_df

cut_df$group <- cut_df$yearID %>%
  cut(breaks=5)

X <- split(cut_df, cut_df$group)

per1 <- X[[1]]
per1 <- per1 %>% group_by(teamID) %>% dplyr::summarise(mean_sal = mean(sum_sal), mean_wp = mean(WPERC))
per2 <- X[[2]]
per2 <- per2 %>% group_by(teamID) %>% dplyr::summarise(mean_sal = mean(sum_sal), mean_wp = mean(WPERC))
per3 <- X[[3]]
per3 <- per3 %>% group_by(teamID) %>% dplyr::summarise(mean_sal = mean(sum_sal), mean_wp = mean(WPERC))
per4 <- X[[4]]
per4 <- per4 %>% group_by(teamID) %>% dplyr::summarise(mean_sal = mean(sum_sal), mean_wp = mean(WPERC))
per5 <- X[[5]]
per5 <- per5 %>% group_by(teamID) %>% dplyr::summarise(mean_sal = mean(sum_sal), mean_wp = mean(WPERC))


per1 %>%
  ggplot(mapping=aes(x=mean_sal, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("1990-1995") + geom_smooth(me
thod=lm)
```



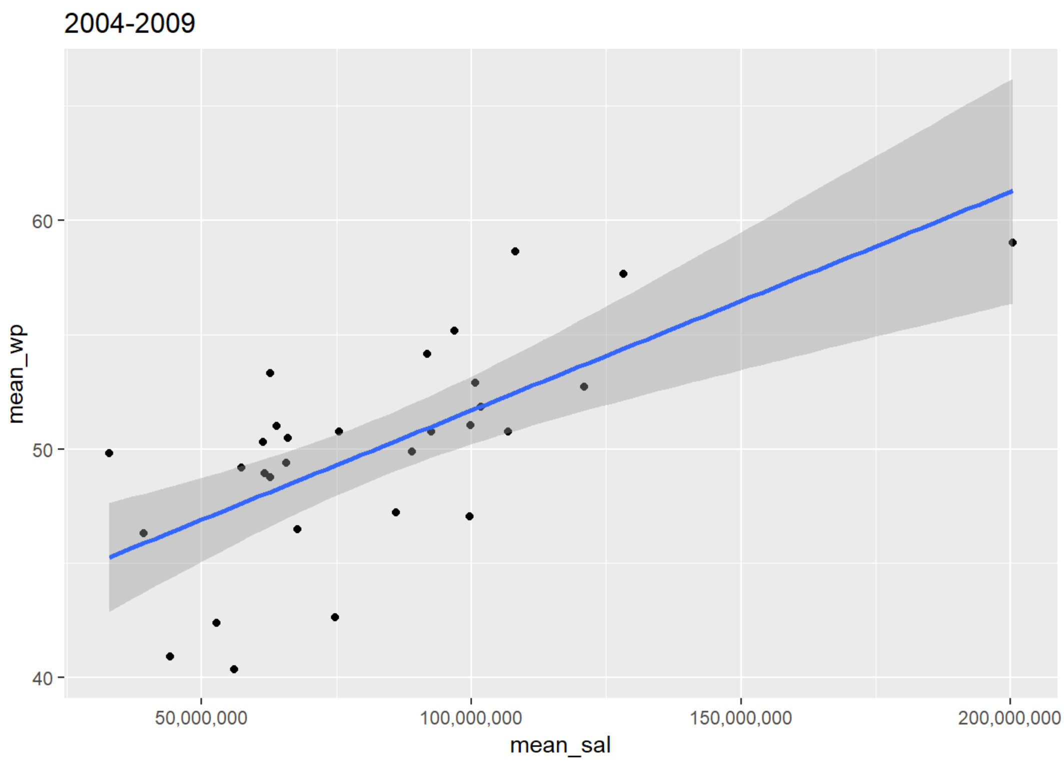

```
per2 %>%
  ggplot(mapping=aes(x=mean_sal, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("1995-2000") + geom_smooth(me
thod=lm)
```



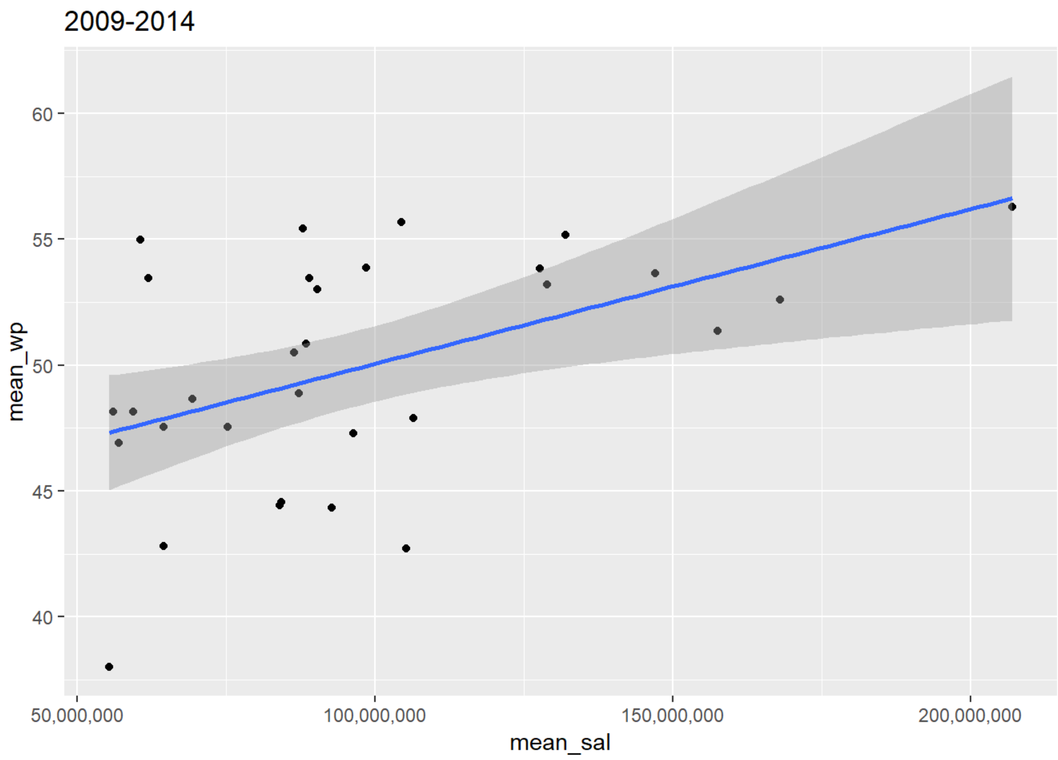

```
per3 %>%
  ggplot(mapping=aes(x=mean_sal, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("2000-2004") + geom_smooth(me
thod=lm)
```




```
per4 %>%
  ggplot(mapping=aes(x=mean_sal, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("2004-2009") + geom_smooth(me
thod=lm)
```




```
per5 %>%
  ggplot(mapping=aes(x=mean_sal, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("2009-2014") + geom_smooth(me
thod=lm)
```



QUESTION 2

There is a positive correlation between the mean payroll and mean winning percentage for every 5 year period. The most significant positive correlation in the periods 1995 to 2000 and 2004 to 2009.

PROBLEM 5

I calculated the standardized formula using the given formula st_payroll = (pay_roll - mean_payroll)/st_dev

```
values <- payroll_df %>%
  group_by(yearID) %>%
  dplyr::summarise(mean_sal = mean(sum_sal), sd_sal = sd(sum_sal)) %>%
  inner_join(payroll_df, by="yearID") %>%
  mutate (z = (((sum_sal * 1.0) - (mean_sal*1.0)) / (sd_sal*1.0) ))

values
```

```
## # A tibble: 728 x 11
##    yearID mean_sal sd_sal teamID franchID     W     L     G sum_sal WPERC
##     <int>    <dbl>  <dbl> <chr>  <chr>    <int> <int> <int>   <dbl> <dbl>
##  1   1990   1.71e7 3.77e6 ATL    ATL         65    97   162  1.46e7  40.1
##  2   1990   1.71e7 3.77e6 BAL    BAL         76    85   161  9.68e6  47.2
##  3   1990   1.71e7 3.77e6 BOS    BOS         88    74   162  2.06e7  54.3
##  4   1990   1.71e7 3.77e6 CAL    ANA         80    82   162  2.17e7  49.4
##  5   1990   1.71e7 3.77e6 CHA    CHW         94    68   162  9.49e6  58.0
##  6   1990   1.71e7 3.77e6 CHN    CHC         77    85   162  1.36e7  47.5
##  7   1990   1.71e7 3.77e6 CIN    CIN         91    71   162  1.44e7  56.2
##  8   1990   1.71e7 3.77e6 CLE    CLE         77    85   162  1.45e7  47.5
##  9   1990   1.71e7 3.77e6 DET    DET         79    83   162  1.76e7  48.8
## 10   1990   1.71e7 3.77e6 HOU    HOU         75    87   162  1.83e7  46.3
## # ... with 718 more rows, and 1 more variable: z <dbl>
```

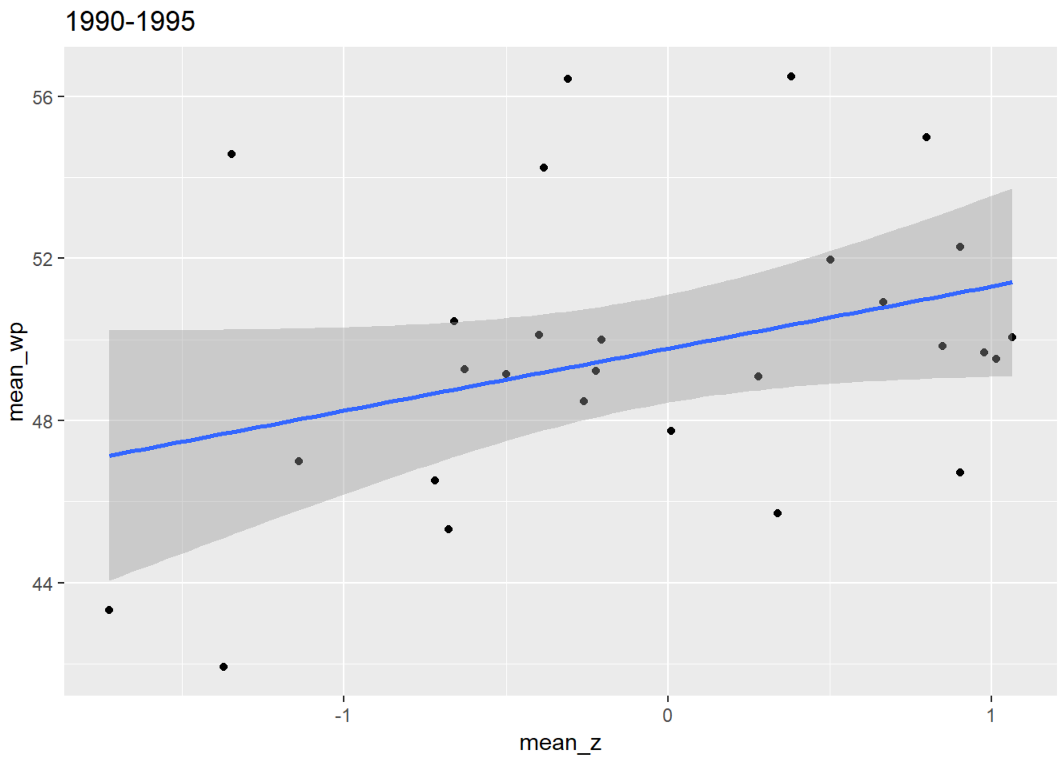
PROBLEM 6

I split the payroll_df into five categories based on yearID in 5 year ranges. Then I created a standardized variable z to standardize payroll. Finally I graphed all 5 split tables with mean standardized value on the x-axis and the mean winning percentage on the y-axis.

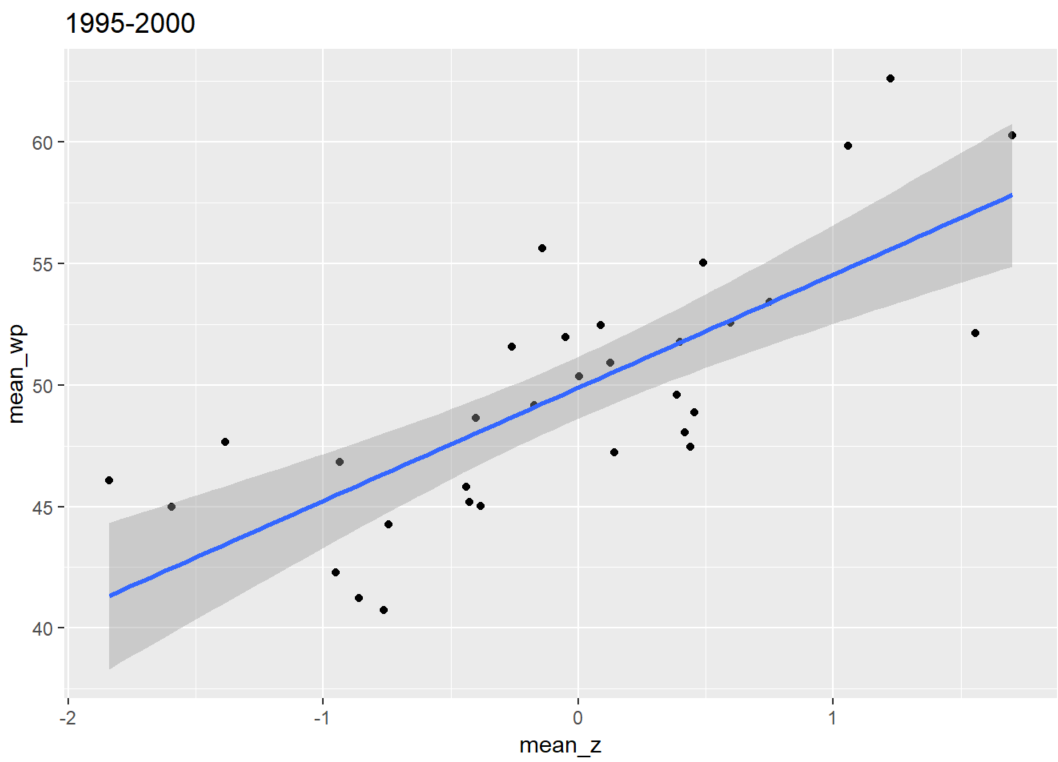
```r
values <- payroll_df %>%
  group_by(yearID) %>%
  dplyr::summarise(mean_sal = mean(sum_sal), sd_sal = sd(sum_sal)) %>%
  inner_join(payroll_df, by="yearID") %>%
  mutate (z = (((sum_sal * 1.0) - (mean_sal*1.0)) / (sd_sal*1.0) ))

values
```

```
## # A tibble: 728 x 11
##    yearID mean_sal sd_sal teamID franchID     W     L     G sum_sal WPERC
##     <int>    <dbl>  <dbl> <chr>  <chr>    <int> <int> <int>   <dbl> <dbl>
##  1   1990   1.71e7 3.77e6 ATL    ATL         65    97   162  1.46e7  40.1
##  2   1990   1.71e7 3.77e6 BAL    BAL         76    85   161  9.68e6  47.2
##  3   1990   1.71e7 3.77e6 BOS    BOS         88    74   162  2.06e7  54.3
##  4   1990   1.71e7 3.77e6 CAL    ANA         80    82   162  2.17e7  49.4
##  5   1990   1.71e7 3.77e6 CHA    CHW         94    68   162  9.49e6  58.0
##  6   1990   1.71e7 3.77e6 CHN    CHC         77    85   162  1.36e7  47.5
##  7   1990   1.71e7 3.77e6 CIN    CIN         91    71   162  1.44e7  56.2
##  8   1990   1.71e7 3.77e6 CLE    CLE         77    85   162  1.45e7  47.5
##  9   1990   1.71e7 3.77e6 DET    DET         79    83   162  1.76e7  48.8
## 10   1990   1.71e7 3.77e6 HOU    HOU         75    87   162  1.83e7  46.3
## # ... with 718 more rows, and 1 more variable: z <dbl>
```

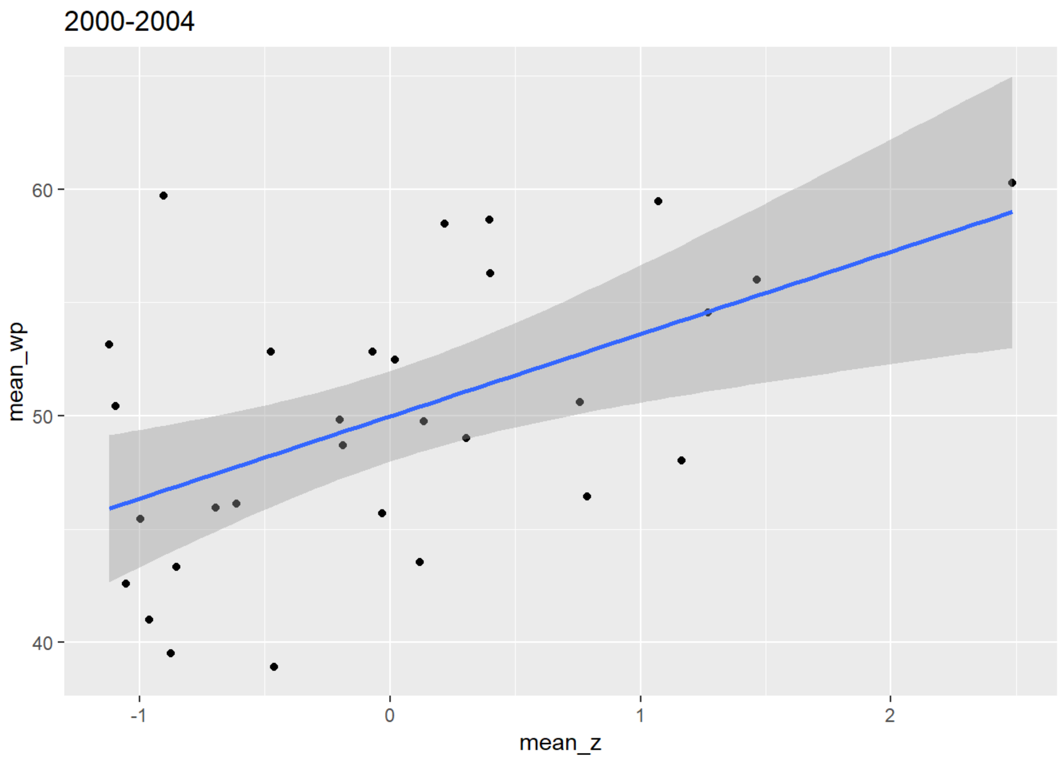
```r
scut_df <- values

scut_df$group <- scut_df$yearID %>%
  cut(breaks=5)

X <- split(scut_df, scut_df$group)

per1 <- X[[1]]
per1 <- per1 %>% group_by(teamID) %>% dplyr::summarise(mean_z = mean(z), mean_wp = mean(WPERC))
per2 <- X[[2]]
per2 <- per2 %>% group_by(teamID) %>% dplyr::summarise(mean_z = mean(z), mean_wp = mean(WPERC))
per3 <- X[[3]]
per3 <- per3 %>% group_by(teamID) %>% dplyr::summarise(mean_z = mean(z), mean_wp = mean(WPERC))
per4 <- X[[4]]
per4 <- per4 %>% group_by(teamID) %>% dplyr::summarise(mean_z = mean(z), mean_wp = mean(WPERC))
per5 <- X[[5]]
per5 <- per5 %>% group_by(teamID) %>% dplyr::summarise(mean_z = mean(z), mean_wp = mean(WPERC))


per1 %>%
  ggplot(mapping=aes(x=mean_z, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("1990-1995") + geom_smooth(me
thod=lm)
```



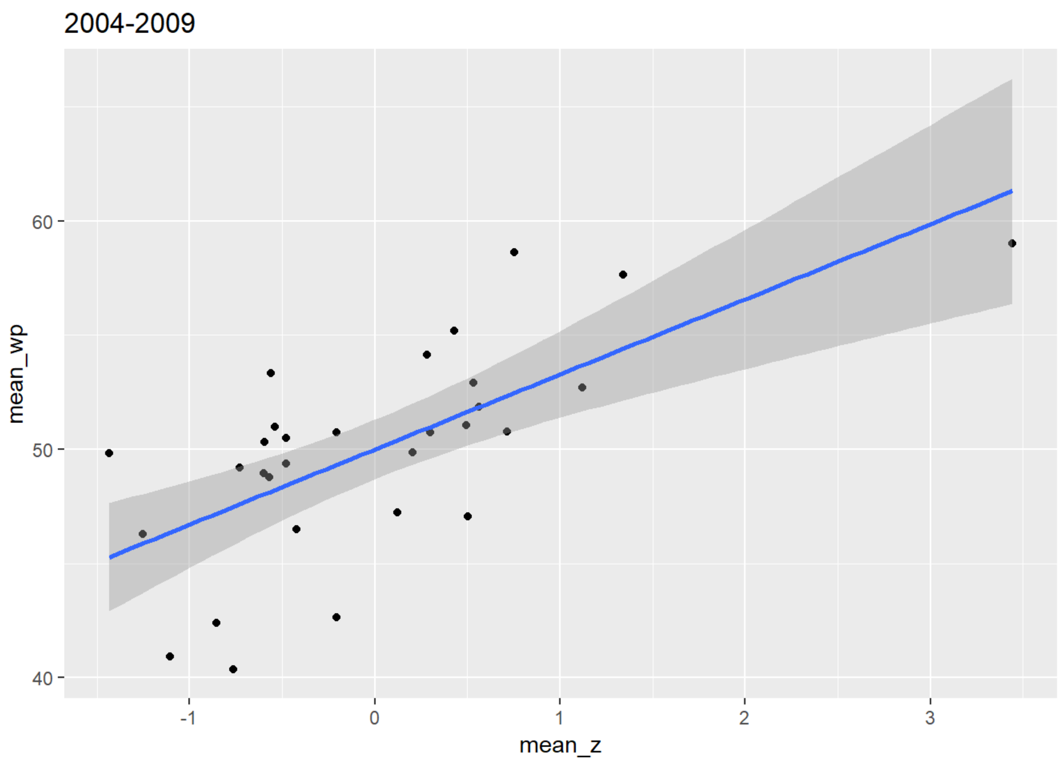

```
per2 %>%
  ggplot(mapping=aes(x=mean_z, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("1995-2000") + geom_smooth(me
thod=lm)
```



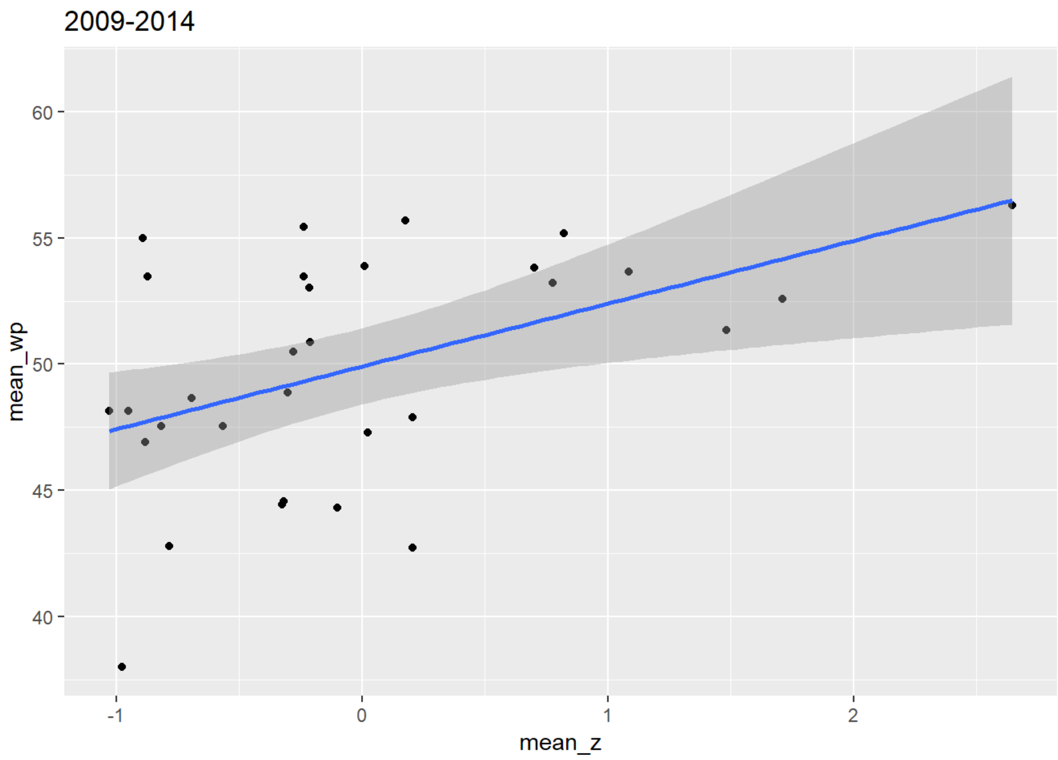

```
per3 %>%
  ggplot(mapping=aes(x=mean_z, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("2000-2004") + geom_smooth(me
thod=lm)
```




```
per4 %>%
  ggplot(mapping=aes(x=mean_z, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("2004-2009") + geom_smooth(me
thod=lm)
```




```
per5 %>%
  ggplot(mapping=aes(x=mean_z, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("2009-2014") + geom_smooth(me
thod=lm)
```



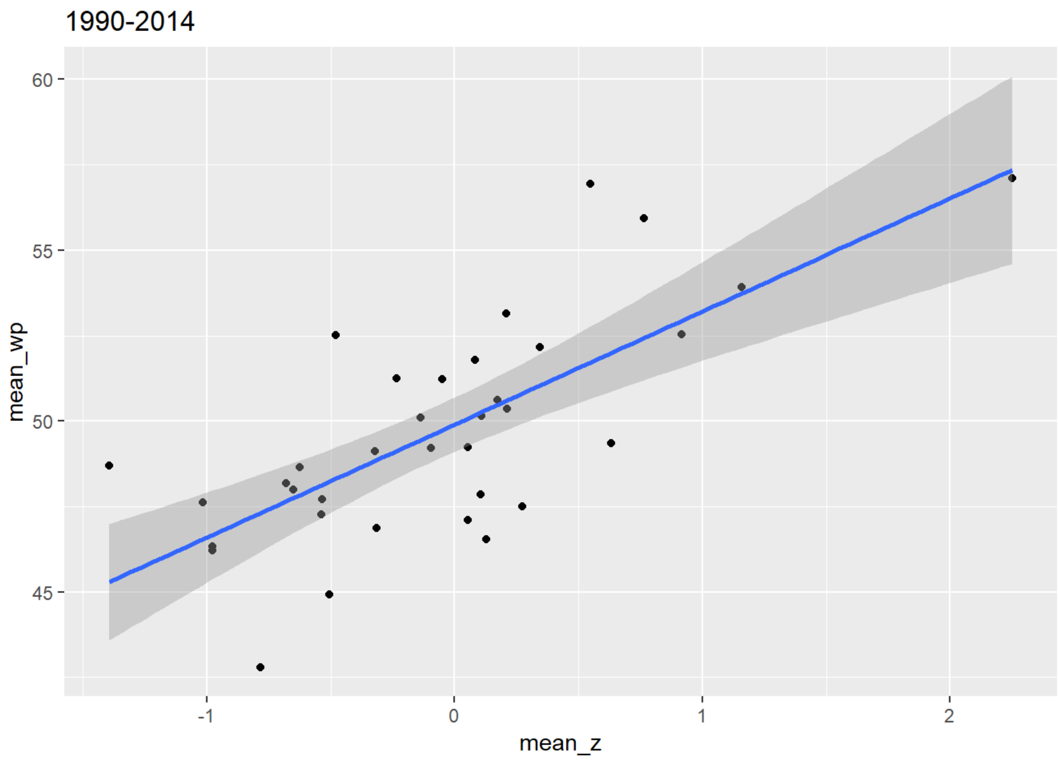
QUESTION 3

The standardized plots follow similar trends but have a mean closer to 0. You can see parallels between specific time periods in the dots spread and patter.

PROBLEM 7

To plot the overall standardized value for the entire time period I dind't split the data frame. I just calculated the stardadized value and graphed it using a scatterplot from 1994 to 2014.

```
values <- payroll_df %>%
  group_by(yearID) %>%
  dplyr::summarise(mean_sal = mean(sum_sal), sd_sal = sd(sum_sal)) %>%
  inner_join(payroll_df, by="yearID") %>%
  mutate (z = (((sum_sal * 1.0) - (mean_sal*1.0)) / (sd_sal*1.0) ))

values <- values %>% group_by(teamID) %>% dplyr::summarise(mean_z = mean(z), mean_wp = mean(WPERC))

values %>%
  ggplot(mapping=aes(x=mean_z, y=mean_wp)) +
    geom_point() + scale_x_continuous(labels = scales::comma) +        ggtitle("1990-2014") + geom_smooth(me
thod=lm)
```

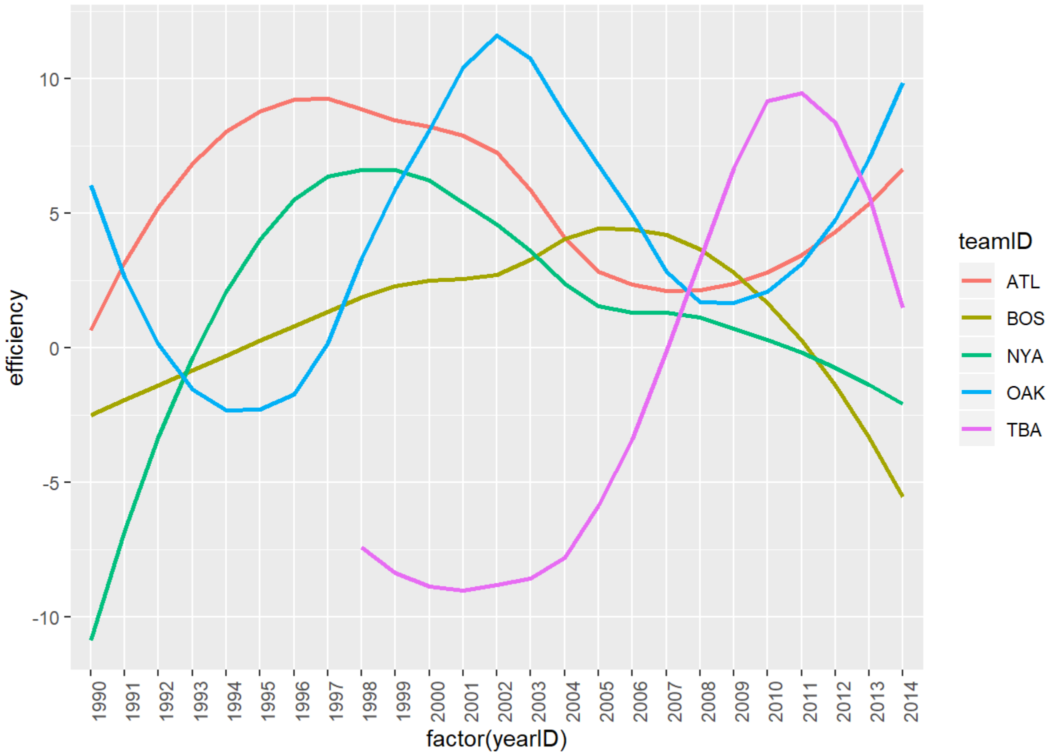


PROBLEM 8

I calculated the effeciency using the expected winning percentage and the actual winning percentage based on the payroll and the plotted the data for the 5 given teams using a line graph. I plotted the efficiency on the y-axis and year on the x-axis.

```
values <- payroll_df %>%
  group_by(yearID) %>%
  dplyr::summarise(mean_sal = mean(sum_sal), sd_sal = sd(sum_sal)) %>%
  inner_join(payroll_df, by="yearID") %>%
  mutate (z = (((sum_sal * 1.0) - (mean_sal*1.0)) / (sd_sal*1.0) ))

values <- values %>%
  mutate(exp_wp = 50.0 + (2.5 * z)) %>%
  mutate(efficiency = WPERC - exp_wp) %>%
  filter(teamID == "OAK" | teamID == "BOS" | teamID == "NYA" | teamID == "ATL" | teamID == "TBA")

values %>%
  ggplot(aes(x=factor(yearID), y=efficiency, group=teamID)) +
  geom_smooth(aes(color=teamID), se=FALSE)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Question 4

This plot shows the effieciency of the teams from 1990 to 2014. Compared to plots 2 and 3 which showed the relationship between winning percentage and payroll, this graph shows the caluclated efficiency for specific teams over a period of time.

The graph shows a clear peak in Oaklands efficiency during the “Moneyball period,” from 2000 to 2005. But, then the efficiency dips in later years.