Homework #7

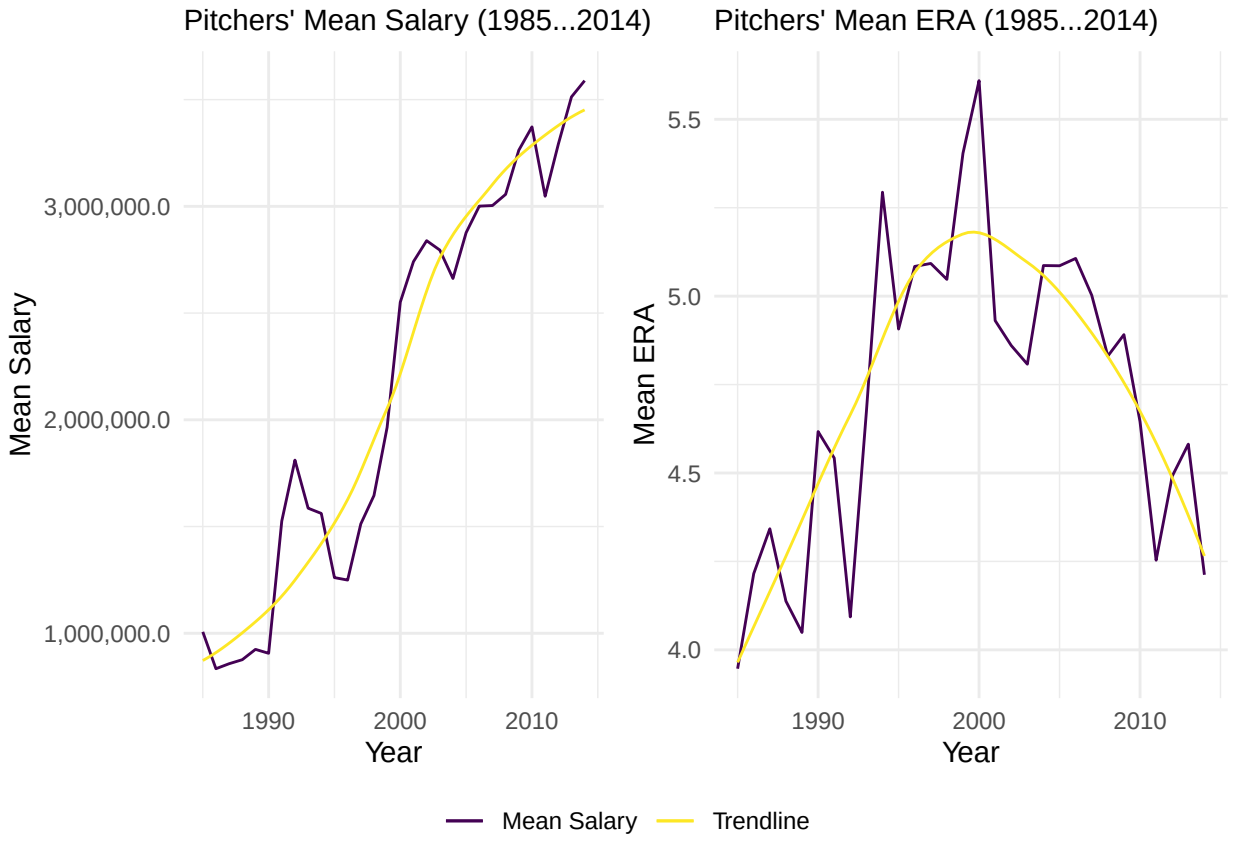
Sang Doan

10/16/2020

Pitcher Salaries and ERAs Over Time

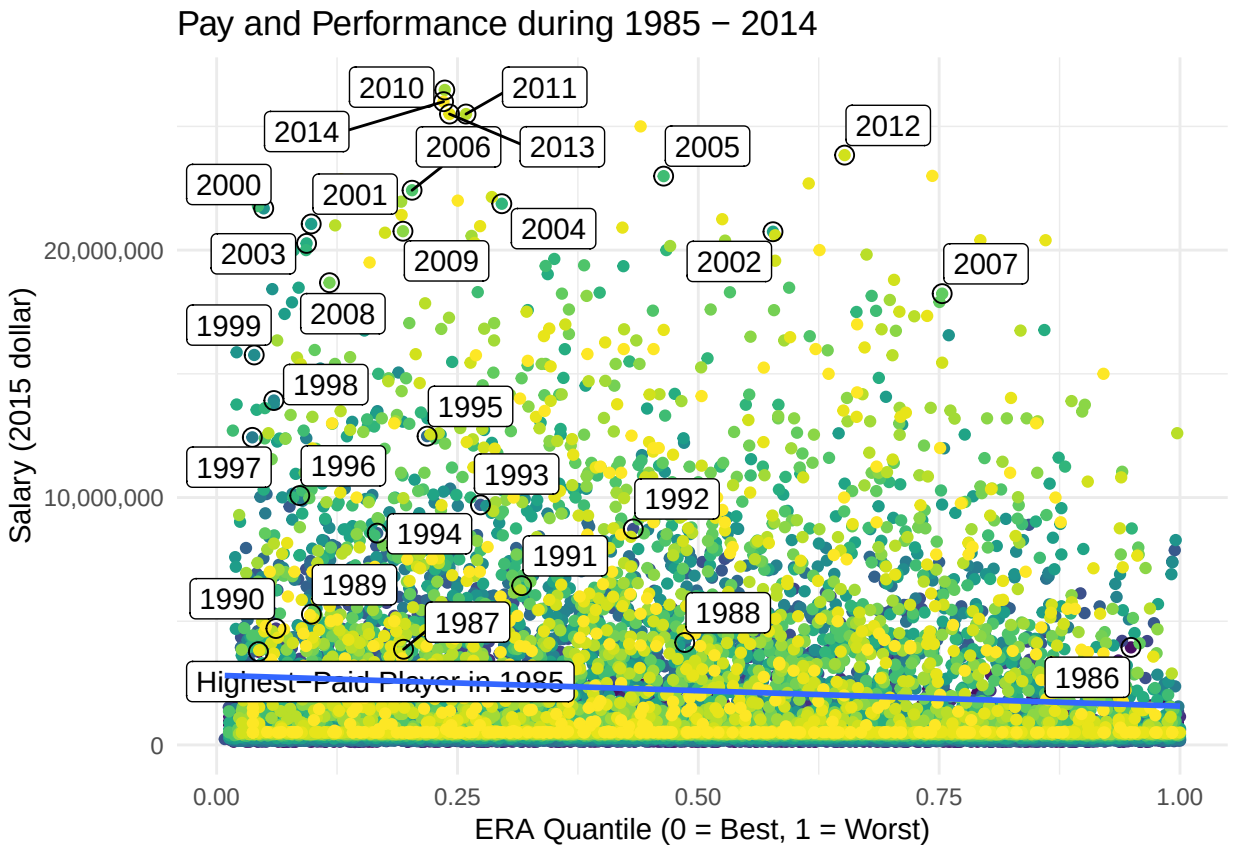
Although mean salary for pitchers quite consistently grew from 1985 to 2014, their mean ERA increased by 25% during 1985 and 2000, before decreasing back to near the 1985 level. An average pitcher is paid 3.5 times higher in 2014 than in 1985, despite their slightly worse ERA.

```
dat <- get_processed_data()
salary <- salera_over_time(dat, 'Salary')
eras <- salera_over_time(dat, 'ERA')

ggarrange(salary, eras, ncol = 2, nrow = 1,
          common.legend = TRUE, legend = 'bottom')
```



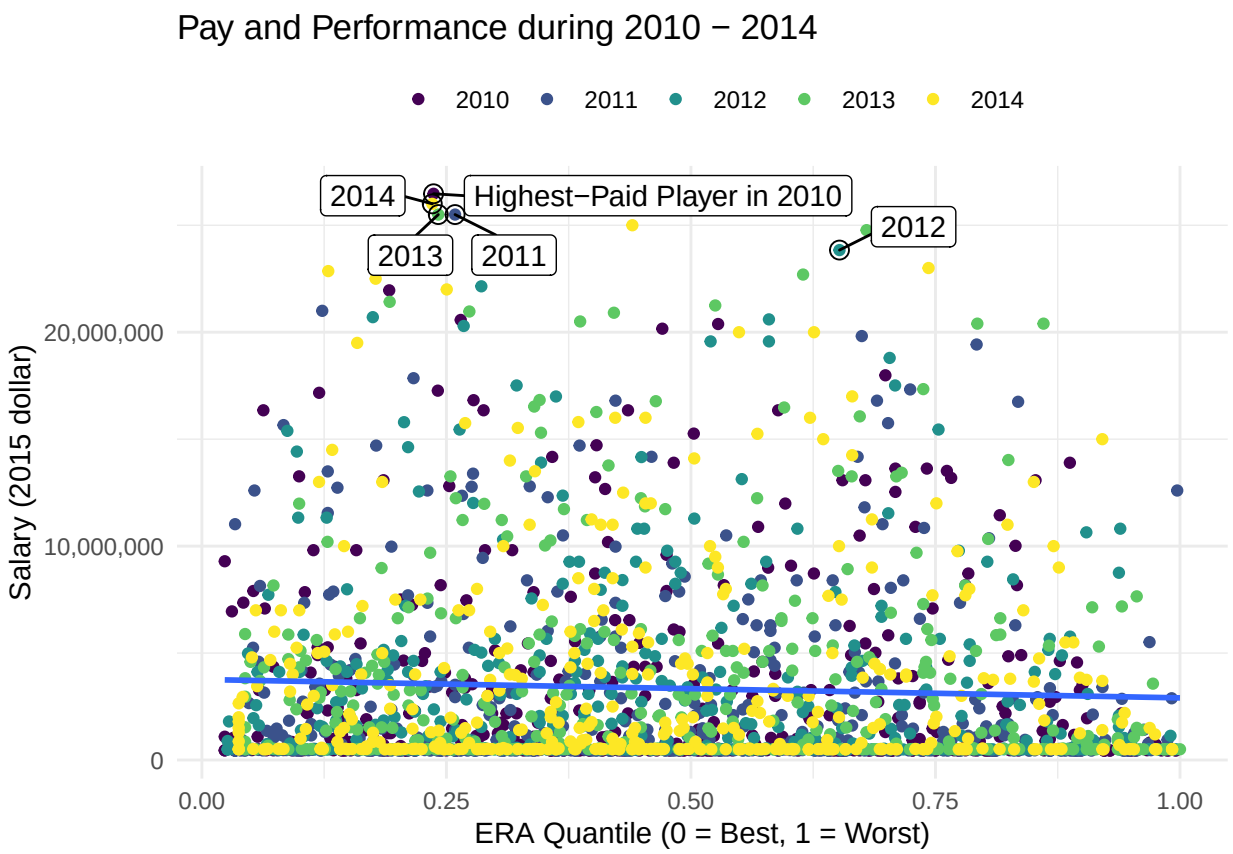
Relationship between Salary and ERA

```
pp <- sal_era(dat, 1985:2014)
ggarrange(pp, legend = 'none')
```



We can see from the blue trendline that during the considered time 1985—2014, a player's ERA quantile is a bad predictor of their salary. Highest paid players of a particular season (highlighted datapoints) did not necessarily have a really low ERA. In fact, they were in very different ERA quantiles of their respective season. In recent years, the correlation between salary and ERA has become almost non-existent. As the graph below shows, from 2010-2014, none of the highest-paid players were among the top 20% regarding ERAs; one of them fell far below the median (2012).

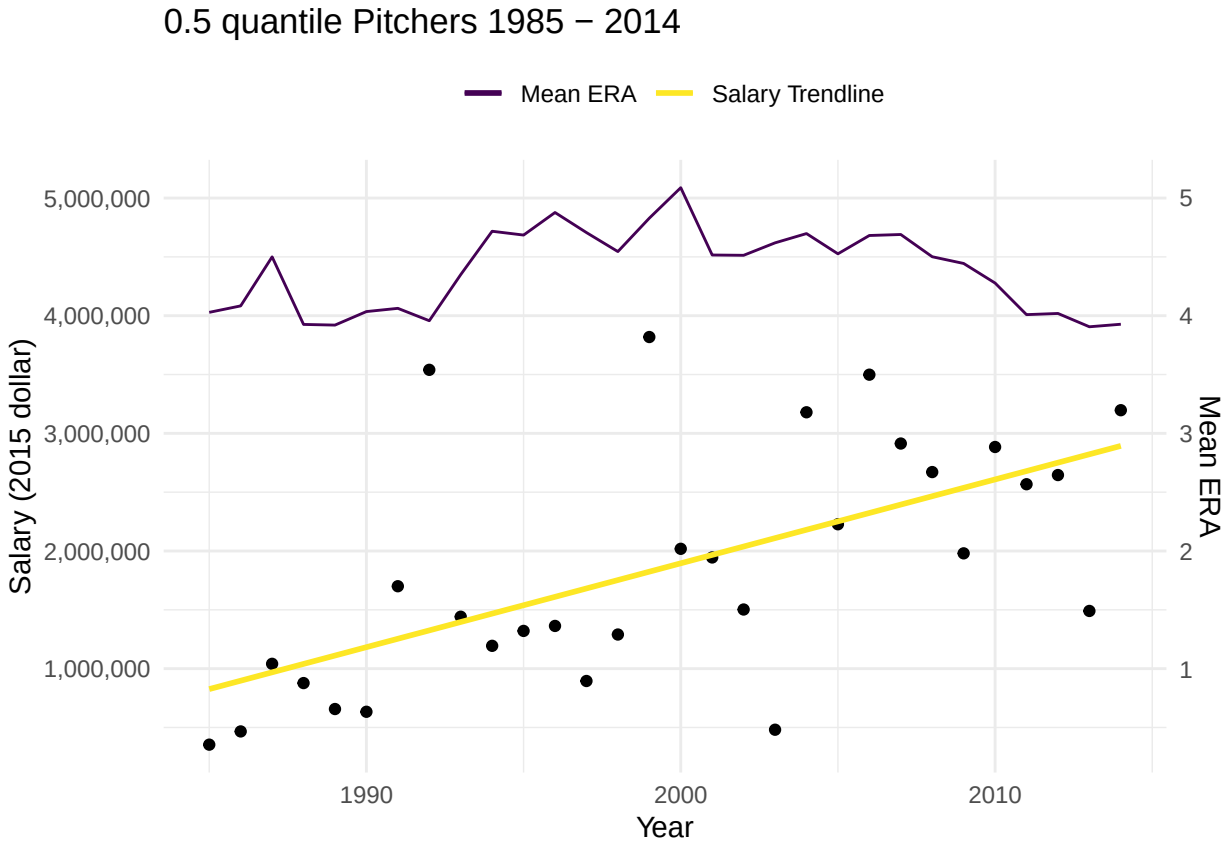
```
pp <- sal_era(dat, 2010:2014)
ggarrange(pp, legend = 'top')
```


Pay and Performance during 2010 − 2014
2010
2011
2012
2013
2014
2014
Highest−Paid Player in 2010
2012
2013
2011
20,000,000
10,000,000
0
Salary (2015 dollar)
0.00
0.25
0.50
0.75
1.00
ERA Quantile (0 = Best, 1 = Worst)

Salaries for Different ERA Quantiles

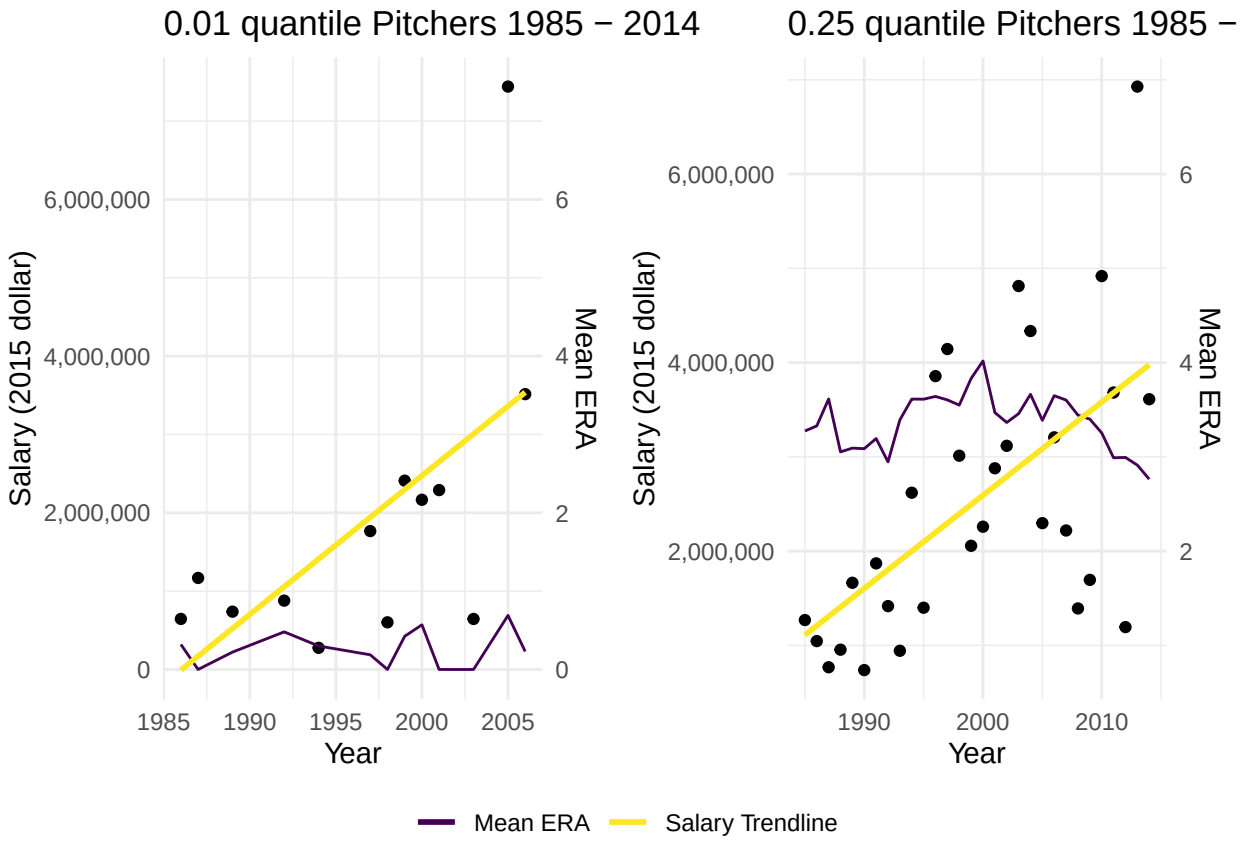
The Median Pitcher

```
pp <- qtile_pitcher(dat, 1985:2014, .5)
ggarrange(pp, legend = 'top')
```



.01 Quantile (The very top players) and .25 Quantile (who surprisingly earned more than the 0.01 quantile group)

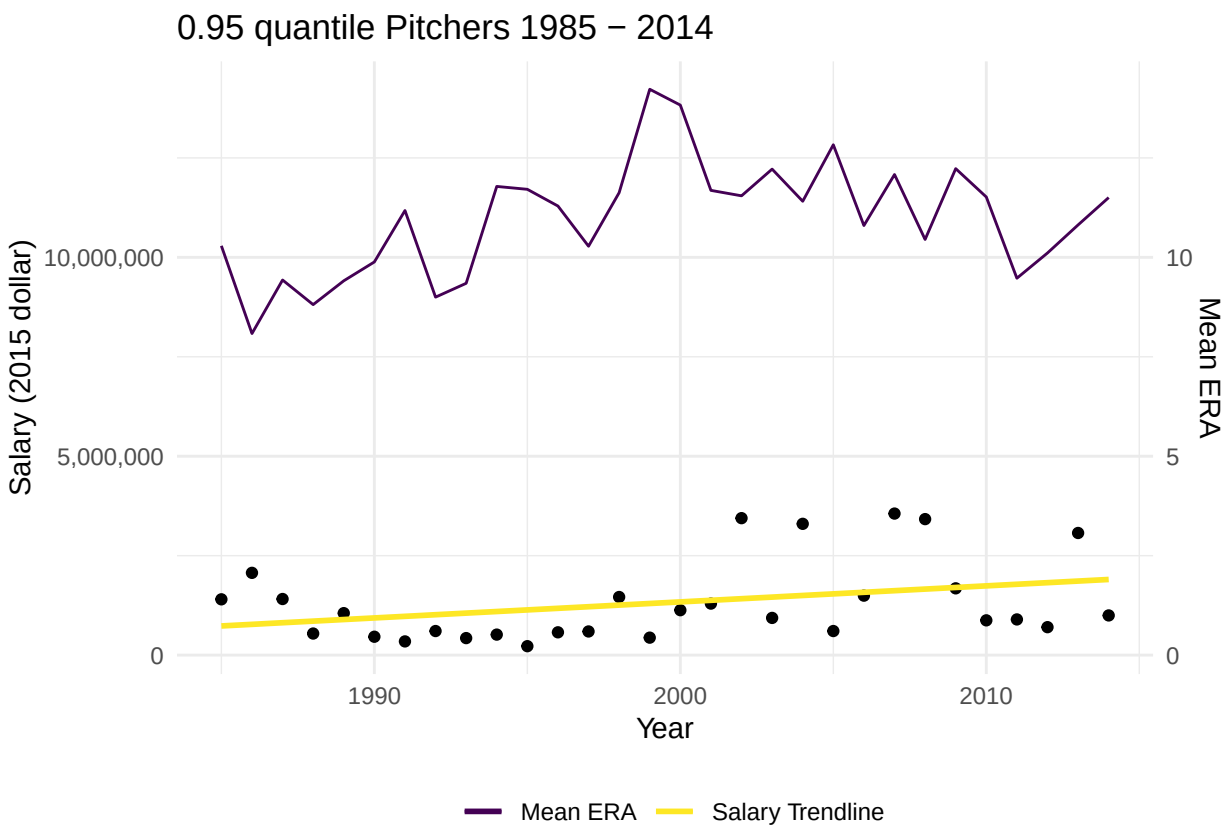
```
firstqtile <- qtile_pitcher(dat, 1985:2014, .01)
tfqtile <- qtile_pitcher(dat, 1985:2014, .25)

ggarrange(firstqtile, tfqtile, ncol = 2, nrow = 1,
          common.legend = TRUE, legend = 'bottom')
```



.95 Quantile (The very bottom players)

Their ERAs went down by around 10%, but their salaries still increased, though not as dramatically as those of top players.

```
pp <- qtile_pitcher(dat, 1985:2014, .95)
ggarrange(pp, legend = 'bottom')
```



Code

data_munge.R

```
salaries <- read.csv('../raw_data/Salaries.csv', header = T, stringsAsFactors = F) %>%
  dplyr::select(yearID, playerID, teamID, salary) %>% filter_dat()
inflation <- read.csv('../raw_data/Inflation.csv', header = T, stringsAsFactors = F)
salaries <- adjust_salaries(salaries, inflation)

pitch <- read.csv('../raw_data/Pitching.csv', header = T, stringsAsFactors = F) %>%
  dplyr::select(yearID, playerID, teamID, ERA) %>% filter_dat()
pitch <- adjust_ERAs(pitch)

processed <- dplyr::left_join(pitch, salaries,
                              by = c('year' = 'year', 'player' = 'player', 'team' = 'team'))
write.csv(drop_na(processed), '../processed_data/processed.csv', row.names = F)
```

data.R

```
get_processed_data <- function() {
  d <- read.csv('../processed_data/processed.csv', header = T, stringsAsFactors = F)
  return(d)
}
```

analysis.R

```
#Pay for Pitcher at qth tile over Time
qtile_pitcher <- function(dat, duration, qtile) {
  mp <- dat %>% filter(year %in% duration, between(adjERA, qtile - 0.01, qtile + 0.01)) %>%
```

```
    group_by(year) %>% summarize (adjSal = mean(adjSal), ERA = mean(ERA))

  out <- ggplot(data = mp, aes(x = year, y = adjSal)) +
    geom_point() +
    geom_line(aes(x = year, y = ERA * 1000000, color = 'Mean ERA')) +
    geom_smooth(method = lm, se = F, aes(color = 'Salary Trendline')) +
    labs(
      title = paste(qtile, 'quantile Pitchers', duration[1], '-', dplyr::last(duration)),
      x = 'Year',
      y = 'Salary (2015 dollar)',
      color = ''
    ) + theme_minimal() +
    scale_y_continuous(
      labels = scales::comma,
      sec.axis = sec_axis(trans = ~./1000000, name = 'Mean ERA')) +
    scale_color_viridis(discrete = T)

  return(out)
}

#Salary and ERA
sal_era <- function(dat, duration) {
  d <- filter(dat, year %in% duration)
  highest_paid <- d %>% group_by(year) %>% filter(row_number(desc(adjSal)) == 1)
  highest_paid$year[1] <- paste('Highest-Paid Player in', highest_paid$year[1])

  out <- ggplot(data = d, aes(x = adjERA, y = adjSal)) +
    geom_point(aes(color = factor(year))) +
    labs(
      title = paste('Pay and Performance during', duration[1], '-', dplyr::last(duration)),
      x = 'ERA Quantile (0 = Best, 1 = Worst)',
      y = 'Salary (2015 dollar)',
      color = ''
    ) + theme_minimal() +
    scale_y_continuous(labels = scales::comma) +
    scale_color_viridis(discrete = T) +
    geom_point(data = highest_paid, size = 3, shape = 1) +
    ggrepel::geom_label_repel(data = highest_paid, aes(label = year)) +
    geom_smooth(method = lm, se = F)

  return(out)
}

#Pitcher Salaries over Time
salera_over_time <- function(dat, salera) {
  d <- mean_salera_by_year(dat, 1985:2014, salera)

  out <- ggplot(data = d, aes(x = yr, y = meanSE)) +
    geom_line(aes(color = paste('Mean', salera))) + theme_minimal() +
    labs(
      subtitle = paste("Pitchers' Mean", salera, '(1985-2014)'),
      x = 'Year',
      y = paste('Mean', salera),
```

```
      color = ''
    ) +
  geom_smooth(se = F, size = 0.5, aes(color = 'Trendline')) +
    scale_color_viridis(discrete = T) +
    scale_y_continuous(labels = scales::comma_format(accuracy = 0.5))

  return(out)
}

mean_salera_by_year <- function(dat, duration, salera) {
  SE <- paste('filter(dat, year == i)$', ifelse(salera == 'Salary', 'adjSal', 'ERA'))
  d <- data.frame(yr = duration, meanSE = rep(NA, length(duration)), stringsAsFactors = F)
  for(i in duration)
    d$meanSE[d$yr == i] <- mean(eval(parse(text = SE)))
  return(d)
}

#Data Munging Functions
filter_dat <- function(dat) {
  return(dplyr::filter(dat, yearID %in% 1985:2014))
}

adjust_salaries <- function(salaries, inflation) {
  for(i in 1985:2014) {
    thesePlayers <- which(salaries$yearID %in% i)
    salaries$salary[thesePlayers] <- salaries$salary[thesePlayers] *
      inflation$inflation2015[inflation$year == i]
  }
  names(salaries) <- c('year', 'player', 'team', 'adjSal')
  return(salaries)
}

adjust_ERAs <- function(pitch) {
  pitch$adjERA <- NA
  for(i in 1985:2014) {
    thisYear <- pitch$ERA[pitch$yearID == i]
    pitch$adjERA[pitch$yearID == i] <- thisYear %>%
      dplyr::cume_dist() #Cumulative distribution transformation
  }
  names(pitch) <- c('year', 'player', 'team', 'ERA', 'adjERA')
  return(pitch)
}
```

config.R

```
source('data.R')
source('analysis.R')

library(tidyverse)
library(magrittr)
library(viridis)
library(ggrepel)
library(ggpubr)
```