Proj2

Danny Brewer

March 27, 2019

```
library(rvest)
```

```
## Loading required package: xml2
```

```
library(tidyverse)
```

```
## -- Attaching packages ----------------
```

```
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts -------------------------
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## x purrr::pluck()          masks rvest::pluck()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(dplyr)
library(tidyr)
library(readr)
```

```
db <- DBI::dbConnect(RSQLite::SQLite(), "lahman2016.sqlite")
```

SQL Query to join Teams and Salaries, while filtering out older years

```
select t.*, payroll, (t.W * 100.0 / t.G) as win_percentage from (select teamID, yearID, sum(salary) as payroll
from

Salaries

where yearID >= 1990 and yearID <= 2014

group by TeamID, yearID) as m join Teams t where m.teamID = t.teamID and m.yearID = t.yearid
```

Payroll Dataframe

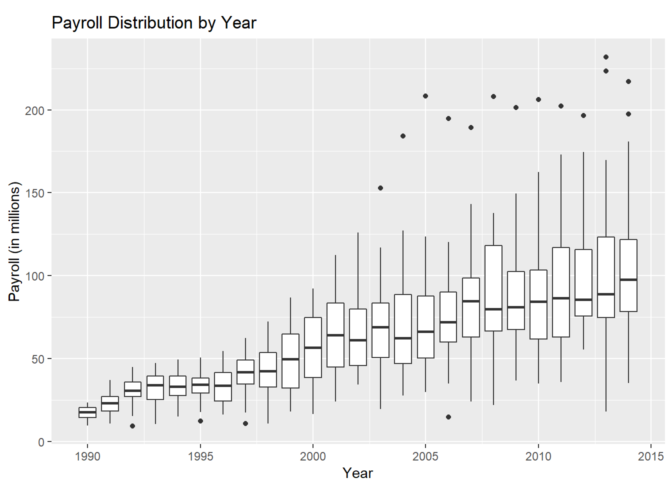
```
payroll_df %>%
  head()
```

```
##   yearID lgID teamID franchID divID Rank   G Ghome  W  L DivWin WCWin
## 1   1990   NL    ATL      ATL     W    6 162    81 65 97      N  <NA>
## 2   1990   AL    BAL      BAL     E    5 161    80 76 85      N  <NA>
## 3   1990   AL    BOS      BOS     E    1 162    81 88 74      Y  <NA>
## 4   1990   AL    CAL      ANA     W    4 162    81 80 82      N  <NA>
## 5   1990   AL    CHA      CHW     W    2 162    80 94 68      N  <NA>
## 6   1990   NL    CHN      CHC     E    4 162    81 77 85      N  <NA>
##   LgWin WSWin   R   AB    H  2B 3B  HR  BB   SO  SB CS HBP SF  RA  ER  ERA
## 1     N     N 682 5504 1376 263 26 162 473 1010  92 55  NA NA 821 727 4.58
## 2     N     N 669 5410 1328 234 22 132 660  962  94 52  NA NA 698 644 4.04
## 3     N     N 699 5516 1502 298 31 106 598  795  53 52  NA NA 664 596 3.72
## 4     N     N 690 5570 1448 237 27 147 566 1000  69 43  NA NA 706 612 3.79
## 5     N     N 682 5402 1393 251 44 106 478  903 140 90  NA NA 633 581 3.61
## 6     N     N 690 5600 1474 240 36 136 406  869 151 50  NA NA 774 695 4.34
##   CG SHO SV IPouts   HA HRA BBA SOA   E  DP    FP              name
## 1 17   8 30   4287 1527 128 579 938 158 133 0.974    Atlanta Braves
## 2 10   5 43   4305 1445 161 537 776  91 151 0.985 Baltimore Orioles
## 3 15  13 44   4326 1439  92 519 997 123 154 0.980    Boston Red Sox
## 4 21  13 42   4362 1482 106 544 944 140 186 0.977 California Angels
## 5 17  10 68   4347 1313 106 548 914 124 169 0.980 Chicago White Sox
## 6 13   7 42   4326 1510 121 572 877 122 136 0.980      Chicago Cubs
##                          park attendance BPF PPF teamIDBR teamIDlahman45
## 1 Atlanta-Fulton County Stadium     980129 105 106      ATL            ATL
## 2              Memorial Stadium    2415189  97  98      BAL            BAL
## 3                Fenway Park II    2528986 105 105      BOS            BOS
## 4               Anaheim Stadium    2555688  97  97      CAL            CAL
## 5                 Comiskey Park    2002357  98  98      CHW            CHA
## 6                Wrigley Field    2243791 108 108      CHC            CHN
##   teamIDretro  payroll win_percentage
## 1         ATL 14555501       40.12346
## 2         BAL  9680084       47.20497
## 3         BOS 20558333       54.32099
## 4         CAL 21720000       49.38272
## 5         CHA  9491500       58.02469
## 6         CHN 13624000       47.53086
```

Box plot showing the distribution of payrolls between 1990 and 2014

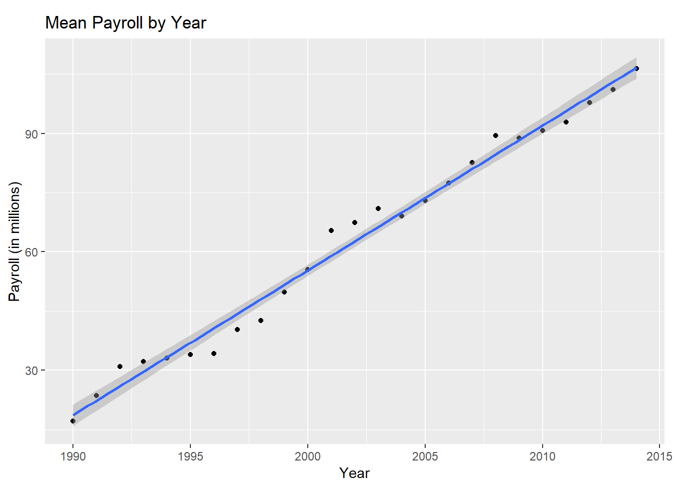
Exploratory statements: As the years progress, the overall spread of the distribution steadily increases. Around 2004, we also see the beginning of statistical outliers that are well above the maximum. The medians per year are steadily increasing, as well as Q1 and Q3 numbers. In general, payrolls are increasing by year.

```
payroll_df %>% ggplot(mapping=aes(y = payroll, x=yearID, group = yearID)) + geom_boxplot() + scale_y_continuous
(label = function(x) format(x/1000000)) + ylab("Payroll (in millions)") + xlab("Year") + ggtitle("Payroll Distr
ibution by Year")
```



A line graph was created representing the trend of the mean payroll of the MLB by year. The graph shows a linear regression line trending upwards as the year progresses.

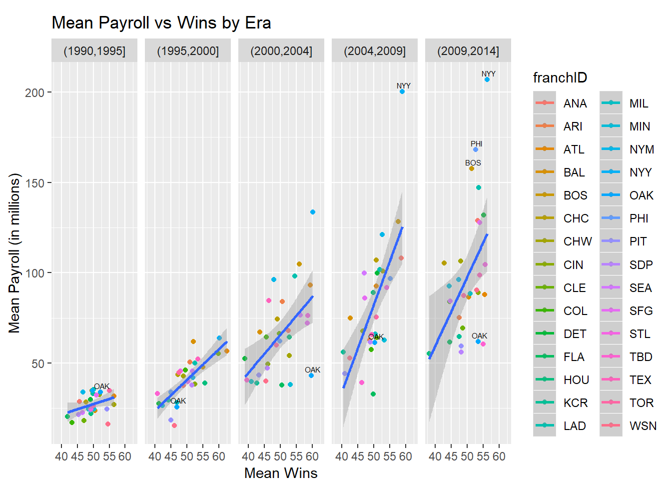
```
payroll_df %>% group_by(yearID) %>% summarize(average_payroll = mean(payroll)) %>% ggplot(mapping=aes(y = avera
ge_payroll, x=yearID)) + geom_point() + geom_smooth(method=lm) + scale_y_continuous(label = function(x) format
(x/1000000)) + ylab("Payroll (in millions)") + xlab("Year") + ggtitle("Mean Payroll by Year")
```



Dataframe is created that groups the years into five eras. A scatter plot is created based on the mean payroll vs mean number of wins for each team in each era

Based on our linear regression line, team payrolls have drastically increased. Our linear regression line has also became steeper, indicating some teams are paying more for less wins. The New York Yankees seem to always have a high amount of wins but also are overpaying. I have labeled the Oakland Athletics specifically due to their ability to consistently stay below the regression line. Between 2000-2004, The Oakland Athletics significantly improved their spending vs wins ratio.

```
group_df <- payroll_df %>% group_by(cut(yearID, breaks= 5), franchID) %>% summarize(mean_win = mean(win_percent
age), mean_payroll = mean(payroll))

group_df <- as.tibble(group_df)

names(group_df)[1] <- "five_group"

group_df %>% ggplot(aes(x = mean_win, y=mean_payroll, color = franchID, group = 1)) + geom_point() + facet_grid
(~five_group) +geom_smooth(method=lm)+ scale_y_continuous(label = function(x) format(x/1000000)) + ylab("Mean P
ayroll (in millions)") + xlab("Mean Wins") + geom_text(aes(label=ifelse(mean_payroll > 150000000 | franchID ==
"OAK", as.character(franchID),'')),hjust=0.4,vjust=-0.6, size = 2, color = "black") + ggtitle("Mean Payroll vs
 Wins by Era")
```

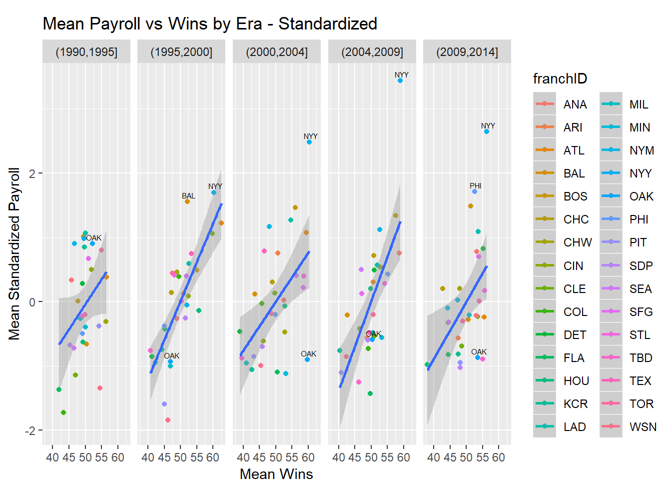


A new dataframe was created similar to the one above, except a standardized payroll value was calculated and graphed instead.

Standardizing the payroll variable was beneficial. It gives a better representation of our model as time varies. The linear regression line is more consistent from era to era, and the data range is more stable. Due to this consistency, it is easier to make predictions.

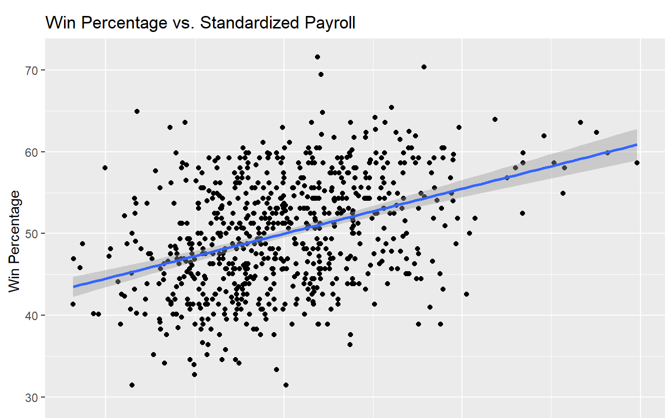
Another thing to note is the ability to see the Yankees as outliers in the 1995 era. This is another characteristic that should give an idea of how the rest of the graph was tweaked. The Oakland A’s datapoint has not changed much, but the idea that they perform efficiently with payroll vs. wins is still there.

```
payroll_df <- payroll_df %>% group_by(yearID) %>% summarize(mean_payroll = mean(payroll), sd_payroll = sd(payro
ll)) %>% inner_join(payroll_df, by="yearID") %>% mutate(standardized_payroll = (payroll-mean_payroll)/sd_payrol
l)

group_df <- payroll_df %>% group_by(cut(yearID, breaks= 5), franchID) %>% summarize(mean_win = mean(win_percent
age), mean_sd_payroll = mean(standardized_payroll))

group_df <- as.tibble(group_df)

names(group_df)[1] <- "five_group"

group_df %>% ggplot(aes(x = mean_win, y=mean_sd_payroll, color = franchID, group = 1)) + geom_point() + facet_g
rid(~five_group) +geom_smooth(method=lm)+ ylab("Mean Standardized Payroll") + xlab("Mean Wins") + geom_text(aes
(label=ifelse(mean_sd_payroll > 1.5 | franchID == "OAK", as.character(franchID),'')),hjust=0.4,vjust=-0.6, size
= 2, color = "black") + ggtitle("Mean Payroll vs Wins by Era - Standardized")
```



Scatter plot with an added regression line to give expected winning percentage as a function of standardized payroll

```
payroll_df %>% ggplot(aes(x = standardized_payroll, y = win_percentage)) + geom_point() + geom_smooth(method=l
m) + ylab("Win Percentage") + xlab("Standardized Payroll") + ggtitle("Win Percentage vs. Standardized Payroll")
```



From the graph above, an expected win percentage was able to be produced, which was used to calculate the efficiency of the team for that year. Popular teams were chosen, including the famous Oakland A’s for their “Moneyball” year. In order to show the extreme variation by year, I stuck with the default geom_smooth method.

During the Moneyball period (around 2002), the A’s outplayed their expected win percentage by more than 10%. In comparison with our linear regression plots by era, this plot directly shows how efficient each team was. Our other plots gave us an idea based on how much each team deviated from the linear regression plot, but could not calculate exactly how efficient a team was during that time period.

```
payroll_df %>% mutate(expected_win_pct = 50 +2.5 * standardized_payroll) %>% mutate(efficiency = win_percentage
- expected_win_pct) %>% filter(teamID == "OAK" | teamID == "BOS" | teamID == "NYA" | teamID == "ATL" | teamID =
= "TBA") %>% ggplot(aes(x=yearID, y = efficiency, color = teamID)) + geom_point() + geom_smooth(se=FALSE)
```

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