# FA3_DSC1107_KHAFAJI

## Major League Baseball: Payroll and Wins

Let's analze the payroll and wins of 30 major league baseball teams from 1998 to 2014.

First, let's load the data:

```r
head(ml_pay)
```

```
##    payroll    avgwin       Team.name.2014    p1998    p1999    p2000     p2001
## 1 1.120874 0.4902585 Arizona Diamondbacks 31.61450 70.49600 81.02783  81.20651
## 2 1.381712 0.5527605       Atlanta Braves 61.70800 74.89000 84.53784  91.85169
## 3 1.161212 0.4538250    Baltimore Orioles 71.86092 72.19836 81.44743  72.42633
## 4 1.972359 0.5487172       Boston Red Sox 59.49700 71.72500 77.94033 109.55891
## 5 1.459767 0.4736557         Chicago Cubs 49.81600 42.14276 60.53933  64.01583
## 6 1.315391 0.5111170    Chicago White Sox 35.18000 24.53500 31.13350  62.36300
##       p2002     p2003     p2004     p2005     p2006     p2007     p2008
## 1 102.82000  80.64033  70.20498  63.01583  59.68423  52.06755  66.20271
## 2  93.47037 106.24367  88.50779  85.14858  90.15688  87.29083 102.36568
## 3  60.49349  73.87750  51.21265  74.57054  72.58558  93.55481  67.19625
## 4 108.36606  99.94650 125.20854 121.31194 120.09982 143.02621 133.39004
## 5  75.69083  79.86833  91.10167  87.21093  94.42450  99.67033 118.34583
## 6  57.05283  51.01000  65.21250  75.22800 102.75067 108.67183 121.18933
##       p2009     p2010     p2011     p2012     p2013     p2014 X2014 X2013 X2012
## 1  73.57167  60.71817  53.63983  74.28483  89.10050 112.68867    59    81    81
## 2  96.72617  84.42367  87.00319  83.30994  89.77819 110.89734    73    96    94
## 3  67.10167  81.61250  85.30404  81.42900  90.99333 107.40662    82    85    93
## 4 122.69600 162.74733 161.40748 173.18662 150.65550 162.81741    62    97    69
## 5 135.05000 146.85900 125.48066  88.19703 104.30468  89.00786    64    66    61
## 6  96.06850 108.27320 129.28554  96.91950 119.07328  91.15925    63    63    85
##   X2011 X2010 X2009 X2008 X2007 X2006 X2005 X2004 X2003 X2002 X2001 X2000 X1999
## 1    94    65    70    82    90    76    77    51    84    98    92    85   100
## 2    89    91    86    72    84    79    90    96   101   101    88    95   103
## 3    69    66    64    68    69    70    74    78    71    67    63    74    78
## 4    90    89    95    95    96    86    95    98    95    93    82    85    94
## 5    71    75    83    97    85    66    79    89    88    67    88    65    67
## 6    79    88    79    89    72    90    99    83    86    81    83    95    75
##   X1998 X2014.pct X2013.pct X2012.pct X2011.pct X2010.pct X2009.pct X2008.pct
## 1    65 0.4154930 0.4969325 0.5000000 0.5802469 0.4012346 0.4294479 0.5030675
## 2   106 0.5140845 0.5889571 0.5802469 0.5493827 0.5617284 0.5276074 0.4417178
## 3    79 0.5774648 0.5214724 0.5740741 0.4259259 0.4074074 0.3926380 0.4171779
## 4    92 0.4366197 0.5950920 0.4259259 0.5555556 0.5493827 0.5828221 0.5828221
## 5    90 0.4507042 0.4049080 0.3765432 0.4382716 0.4629630 0.5092025 0.5950920
## 6    80 0.4436620 0.3865031 0.5246914 0.4876543 0.5432099 0.4846626 0.5460123
##   X2007.pct X2006.pct X2005.pct X2004.pct X2003.pct X2002.pct X2001.pct
## 1 0.5521472 0.4691358 0.4753086 0.3148148 0.5185185 0.6049383 0.5679012
## 2 0.5153374 0.4876543 0.5555556 0.5925926 0.6234568 0.6234568 0.5432099
## 3 0.4233129 0.4320988 0.4567901 0.4814815 0.4382716 0.4135802 0.3888889
```

```
## 4 0.5889571 0.5308642 0.5864198 0.6049383 0.5864198 0.5740741 0.5061728
## 5 0.5214724 0.4074074 0.4876543 0.5493827 0.5432099 0.4135802 0.5432099
## 6 0.4417178 0.5555556 0.6111111 0.5123457 0.5308642 0.5000000 0.5123457
##   X2000.pct X1999.pct X1998.pct
## 1 0.5246914 0.6134969 0.3987730
## 2 0.5864198 0.6319018 0.6503067
## 3 0.4567901 0.4785276 0.4846626
## 4 0.5246914 0.5766871 0.5644172
## 5 0.4012346 0.4110429 0.5521472
## 6 0.5864198 0.4601227 0.4907975
```

the payroll column corresponds to the total team payroll (in billion USD) over the years, while the avgwin column is the aggregated win percentage from 1998 to 2014. the Team.name.2014 column corresponds to the team name.

p1998, p1999,. . . , p2014 corresponds to the payroll for each year (in million USD). X1998, X1999, . . . , X2014 corresponds to the number of wins for each year. X1998.pct, X1999.pct, . . . , X2014.pct corresponds to the win percentage for each year.

## Data Cleaning

Let's make 4 tables: Aggregate table - one table for the team name, the total payroll, and the average win rate over the years Payroll table - one table for the payroll for each year, with the respective team name Win Count table - one table for the number of wins for each given year, with the respective team name Win Rate table - one table for the win rate for each given year, with the respective team name.

we can then join the payroll table, win count table, and win rate table, to make a comprehensive "per year" table

```r
aggregate_table_mlb <- ml_pay %>% select(Team.name.2014, payroll, avgwin) %>%
  rename(MLB_Team = Team.name.2014, total_pay = payroll, avg_winrate = avgwin) %>% # rename columns
  mutate(total_pay = total_pay *(10e2)) #to transform into millions USD


aggregate_table_mlb
```

```
##                 MLB_Team total_pay avg_winrate
## 1   Arizona Diamondbacks 1120.8736   0.4902585
## 2         Atlanta Braves 1381.7118   0.5527605
## 3      Baltimore Orioles 1161.2117   0.4538250
## 4         Boston Red Sox 1972.3587   0.5487172
## 5           Chicago Cubs 1459.7668   0.4736557
## 6      Chicago White Sox 1315.3909   0.5111170
## 7        Cincinnati Reds 1024.7816   0.4861602
## 8      Cleveland Indians  999.1810   0.4959225
## 9       Colorado Rockies 1026.1536   0.4633760
## 10        Detroit Tigers 1429.7408   0.4822029
## 11        Houston Astros 1060.1501   0.4687202
## 12    Kansas City Royals  817.7417   0.4342288
## 13    Los Angeles Angels 1562.6224   0.5463819
## 14   Los Angeles Dodgers 1740.2719   0.5308482
## 15         Miami Marlins  667.8019   0.4813631
## 16     Milwaukee Brewers  979.0940   0.4746570
## 17       Minnesota Twins  969.8272   0.5019047
## 18         New York Mets 1588.4288   0.4911388
## 19      New York Yankees 2703.2482   0.5830719
```

```
## 20     Oakland Athletics  840.9340   0.5445067
## 21 Philadelphia Phillies 1630.1209   0.5247021
## 22    Pittsburgh Pirates  733.9057   0.4371254
## 23      San Diego Padres  840.6668   0.4754884
## 24  San Francisco Giants 1416.8770   0.5304369
## 25      Seattle Mariners 1311.1203   0.4925819
## 26   St. Louis Cardinals 1368.1117   0.5595414
## 27        Tampa Bay Rays  710.7894   0.4685176
## 28         Texas Rangers 1269.3201   0.4956494
## 29     Toronto Blue Jays 1129.0219   0.4930823
## 30  Washington Nationals  921.9641   0.4660195
```

First, we retrieved all aggregated data, and renamed the columns. We then converted the total payroll amount from billion USD to million USD, to match the rest of the payroll data. Next, we created the dollars/win column.

Now, let's get the payroll table:

```r
payroll_mlb <- ml_pay %>% select(Team.name.2014, num_range("p",1998:2014)) %>%
  rename(MLB_Team = Team.name.2014) %>%
  pivot_longer(starts_with("p"), names_to = "year", values_to = "payroll") %>%
  mutate(year = str_remove_all(year, c("p"))) %>%
  mutate_at(c("year"), as.integer)

head(payroll_mlb)
```

```
## # A tibble: 6 x 3
##   MLB_Team              year payroll
##   <fct>                <int>   <dbl>
## 1 Arizona Diamondbacks  1998    31.6
## 2 Arizona Diamondbacks  1999    70.5
## 3 Arizona Diamondbacks  2000    81.0
## 4 Arizona Diamondbacks  2001    81.2
## 5 Arizona Diamondbacks  2002   103.
## 6 Arizona Diamondbacks  2003    80.6
```

We simply retrieved the columns that contained yearly payroll data and pivoted it. We then cleaned the values containing the year so that it could serve as our year column.

Let's then create the win count table:

```r
wincount_mlb <- ml_pay %>% select(Team.name.2014, num_range("X",1998:2014)) %>%
  rename(MLB_Team = Team.name.2014) %>%
  pivot_longer(num_range("X",1998:2014), names_to = "year", values_to = "win_Count") %>%
  mutate(year = str_remove_all(year, c("X"))) %>%
  mutate_at(c("year"), as.integer)

head(wincount_mlb)
```

```
## # A tibble: 6 x 3
##   MLB_Team              year win_Count
##   <fct>                <int>     <int>
## 1 Arizona Diamondbacks  1998        65
## 2 Arizona Diamondbacks  1999       100
## 3 Arizona Diamondbacks  2000        85
## 4 Arizona Diamondbacks  2001        92
## 5 Arizona Diamondbacks  2002        98
```

```
## 6 Arizona Diamondbacks  2003       84
```

Similar to how we cleaned the yearly payroll table

Lastly, lets get the table for the winrate

```
mlb_winrate <- ml_pay %>% select(Team.name.2014, ends_with(".pct")) %>%
  rename(MLB_Team = Team.name.2014) %>%
  pivot_longer(ends_with(".pct"), names_to = "year", values_to = "win_Rate") %>%
  mutate(year = str_remove_all(year, "X|\\.pct" )) %>%
  mutate_at(c("year"), as.integer)

head(mlb_winrate)
```

```
## # A tibble: 6 x 3
##   MLB_Team              year win_Rate
##   <fct>                <int>    <dbl>
## 1 Arizona Diamondbacks  2014    0.415
## 2 Arizona Diamondbacks  2013    0.497
## 3 Arizona Diamondbacks  2012    0.5
## 4 Arizona Diamondbacks  2011    0.580
## 5 Arizona Diamondbacks  2010    0.401
## 6 Arizona Diamondbacks  2009    0.429
```

What we did was similar to the last 2 tables.

We can now join the three tables that we made:

```
mlb_pay_wincount_winrate <- left_join(payroll_mlb, wincount_mlb, join_by("MLB_Team", "year")) %>%
  left_join(., mlb_winrate, join_by("MLB_Team", "year")) %>%
  mutate(total_games = as.integer(win_Count/win_Rate)) %>% #create total games table
  mutate(dollars_per_win = payroll/win_Count) # create dollars/win column

head(mlb_pay_wincount_winrate)
```

```
## # A tibble: 6 x 7
##   MLB_Team           year payroll win_Count win_Rate total_games dollars_per_win
##   <fct>             <int>   <dbl>     <int>    <dbl>       <int>           <dbl>
## 1 Arizona Diamondb~  1998    31.6        65    0.399         163           0.486
## 2 Arizona Diamondb~  1999    70.5       100    0.613         163           0.705
## 3 Arizona Diamondb~  2000    81.0        85    0.525         162           0.953
## 4 Arizona Diamondb~  2001    81.2        92    0.568         162           0.883
## 5 Arizona Diamondb~  2002   103.         98    0.605         162           1.05
## 6 Arizona Diamondb~  2003    80.6        84    0.519         162           0.960
```

We also went ahead and created total_games column, getting the total games played for each team and each year, as well as creating the dollars/win column.

Now, let's add a total games column in the aggregate table using the yearly table.

```
total_win_df <- mlb_pay_wincount_winrate %>%
  group_by(MLB_Team) %>%
  summarise(total_Win = sum(win_Count))

aggregate_table_mlb <- left_join(aggregate_table_mlb, total_win_df, join_by("MLB_Team")) %>%
  mutate(dollars_per_win = total_pay/total_Win) # create dollars/win,

head(aggregate_table_mlb)
```

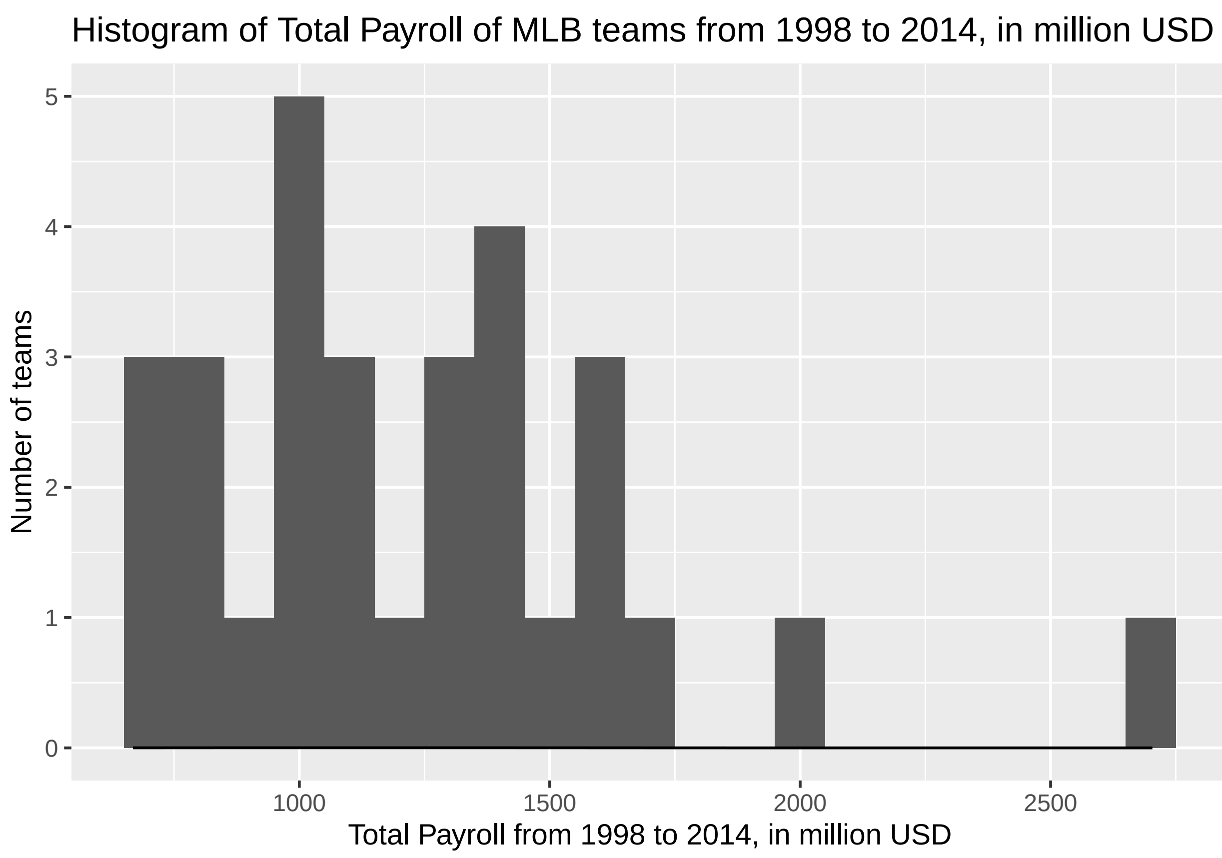
```
##              MLB_Team total_pay avg_winrate total_Win dollars_per_win
## 1 Arizona Diamondbacks  1120.874   0.4902585      1350       0.8302768
## 2       Atlanta Braves  1381.712   0.5527605      1544       0.8948911
## 3    Baltimore Orioles  1161.212   0.4538250      1250       0.9289694
## 4       Boston Red Sox  1972.359   0.5487172      1513       1.3036079
## 5         Chicago Cubs  1459.767   0.4736557      1301       1.1220345
## 6    Chicago White Sox  1315.391   0.5111170      1390       0.9463244
```

## Data Exploration

### Payroll across years

First, let's get the histogram of the total payroll amount across the years.

```r
aggregate_table_mlb %>% ggplot(aes(x=total_pay)) +
  geom_histogram(binwidth = 100) +
  geom_density(alpha=.2, fill="blue")+
  xlab("Total Payroll from 1998 to 2014, in million USD")+
  ylab("Number of teams") +
  ggtitle("Histogram of Total Payroll of MLB teams from 1998 to 2014, in million USD")
```




Let's figure out the top 5 biggest and top 5 smallest spenders across all years:

```r
aggregate_table_mlb %>% arrange(desc(total_pay)) %>%
  slice(sort(c(seq_len(5), n() - seq_len(5) +1))) %>%
  select(c("MLB_Team", "total_pay"))
```

```
##                 MLB_Team total_pay
```

```
## 1       New York Yankees 2703.2482
## 2         Boston Red Sox 1972.3587
## 3    Los Angeles Dodgers 1740.2719
## 4  Philadelphia Phillies 1630.1209
## 5          New York Mets 1588.4288
## 6       San Diego Padres  840.6668
## 7     Kansas City Royals  817.7417
## 8     Pittsburgh Pirates  733.9057
## 9         Tampa Bay Rays  710.7894
## 10         Miami Marlins  667.8019
```

```
teams_lowest_payroll <- aggregate_table_mlb %>% arrange(desc(total_pay)) %>%
  slice(sort(c(n() - seq_len(5) +1))) %>%
  select(c("MLB_Team", "total_pay"))

teams_highest_payroll <- aggregate_table_mlb %>% arrange(desc(total_pay)) %>%
  slice(sort(c(seq_len(5)))) %>%
  select(c("MLB_Team", "total_pay"))
```

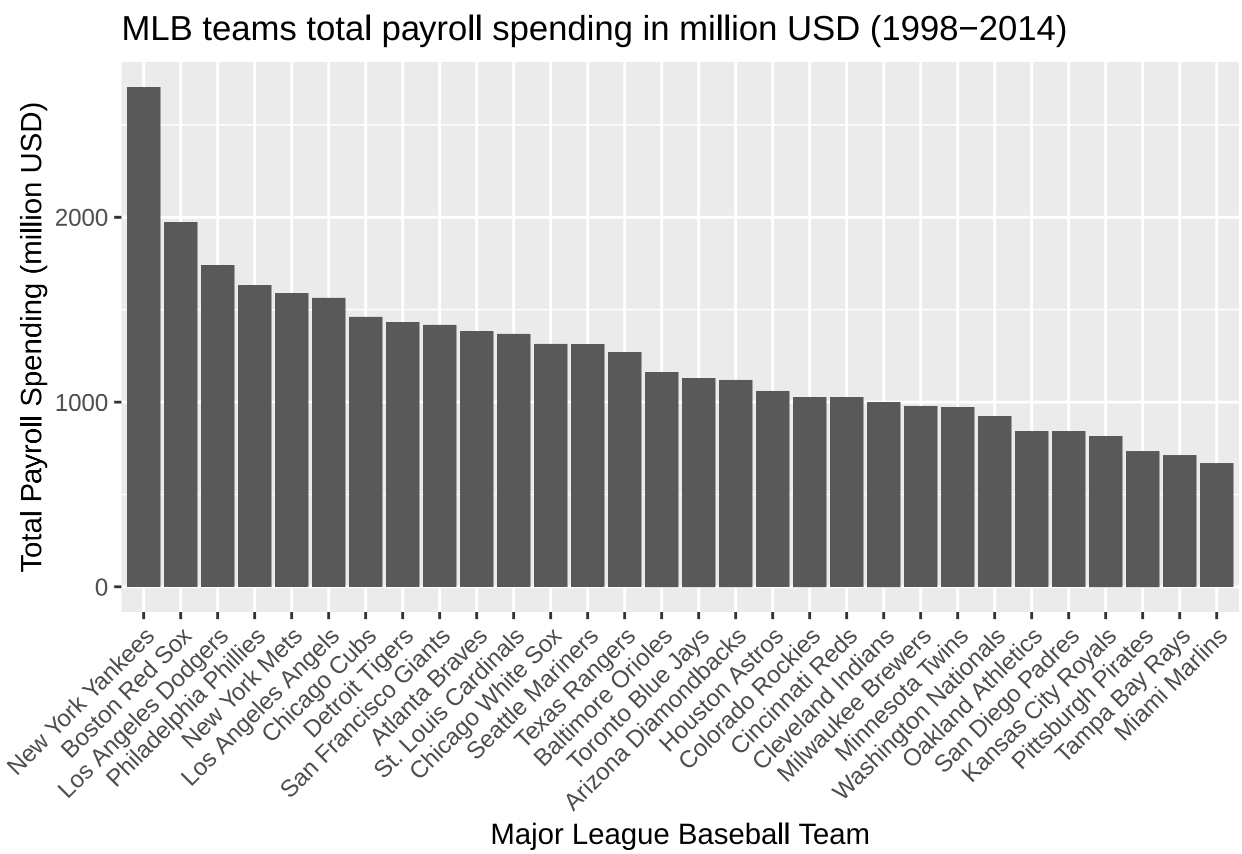
We can see that the New York Yankees are the highest spenders in terms of payroll, paying a total of 2.7 billion USD across the years.

After the Yankees, the figure drops to below 2 billion USD, with a 700 million USD gap between the Yankees and the 2nd highest spender, the Boston Red Sox. But even the Sox have some quarter billion USD gap compared to the next highest spending team, the Los Angeles Dodgers. The Philadelphia Phillies, and the New York Mets, the 4th and 5th highest spending, respectively, have spent close to that the Dodgers have spent, with less than a 100 million dollar difference.

The lowest spenders have few difference in their spending, indicating that there is a lower bound that an MLB team is willing to spend for the payroll of their players. Among them, the Miami Marlins are lowest with 667.8 million USD total spending across the years. It is followed by the Tampa Bay Rays, Pittsburgh Pirates, Kansas City Royals, and the San Diego Padres.

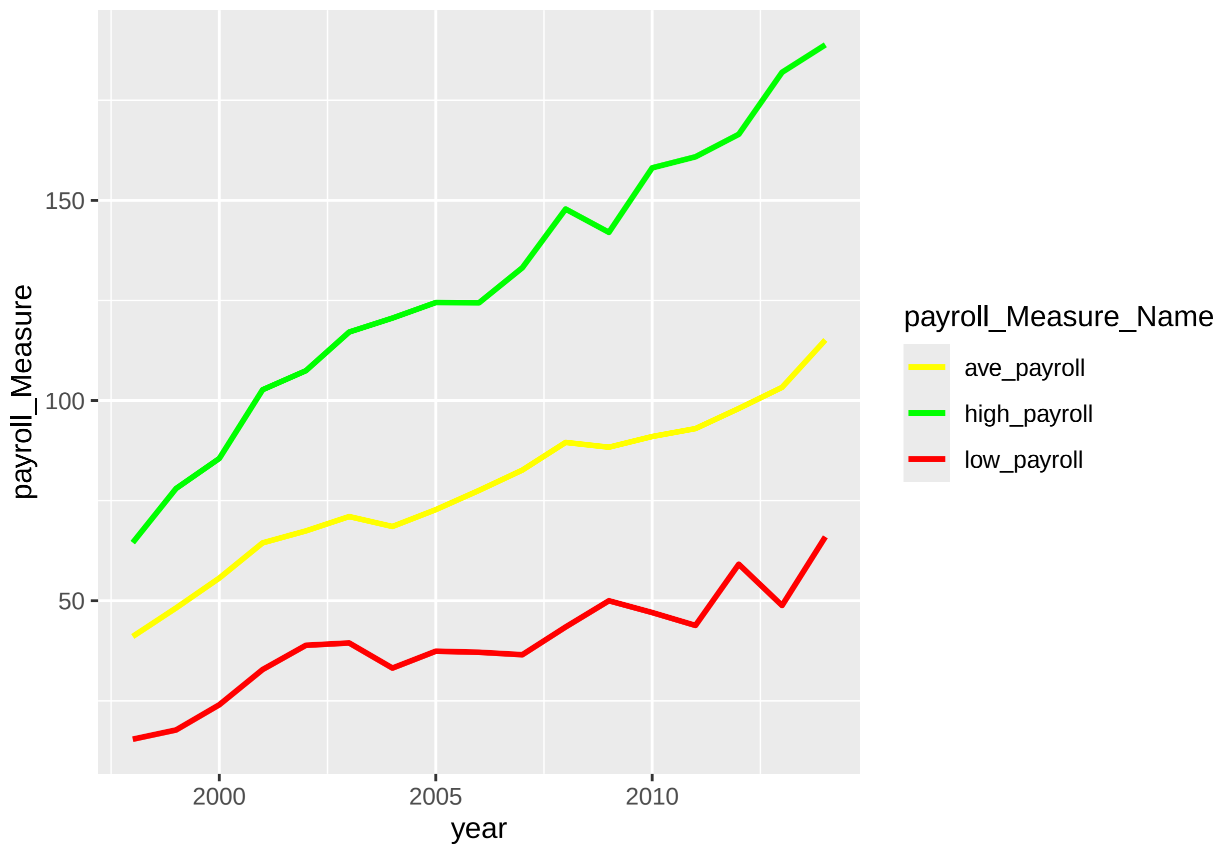
We can visualize this better using a bar graph:

```
aggregate_table_mlb %>% arrange(desc(total_pay)) %>%
  ggplot( aes(x=reorder(MLB_Team, -total_pay)  , y=total_pay) ) +
  geom_bar(stat = "identity")+
  theme(axis.text.x=element_text(angle=45,hjust=1, vjust = 1))+
  xlab("Major League Baseball Team") +
  ylab("Total Payroll Spending (million USD)") +
  ggtitle("MLB teams total payroll spending in million USD (1998-2014)")
```




Now, let's graph the year vs payroll for the league across the years. Note that this payroll data is in million USD. This also uses the top 5 spender teams and bottom 5 spender teams for the high payroll and low payroll values, respectively.

```
mlb_pay_wincount_winrate %>% group_by(year) %>%
  summarise(
    ave_payroll = mean(payroll),
    high_payroll = mean(sort(payroll, decreasing = TRUE)[1:5]),
    low_payroll = mean(sort(payroll, decreasing = FALSE)[1:5])
    ) %>%
  pivot_longer(
      c("ave_payroll", "high_payroll", "low_payroll"),
      names_to = "payroll_Measure_Name",
      values_to = "payroll_Measure"
      ) %>%
  ggplot(aes(x=year, y=payroll_Measure, color=payroll_Measure_Name))+
  geom_line(linewidth=1)+
  scale_color_manual(values =
                       c("ave_payroll" = "yellow",
                         "high_payroll"="green",
                         "low_payroll"= "red"))
```

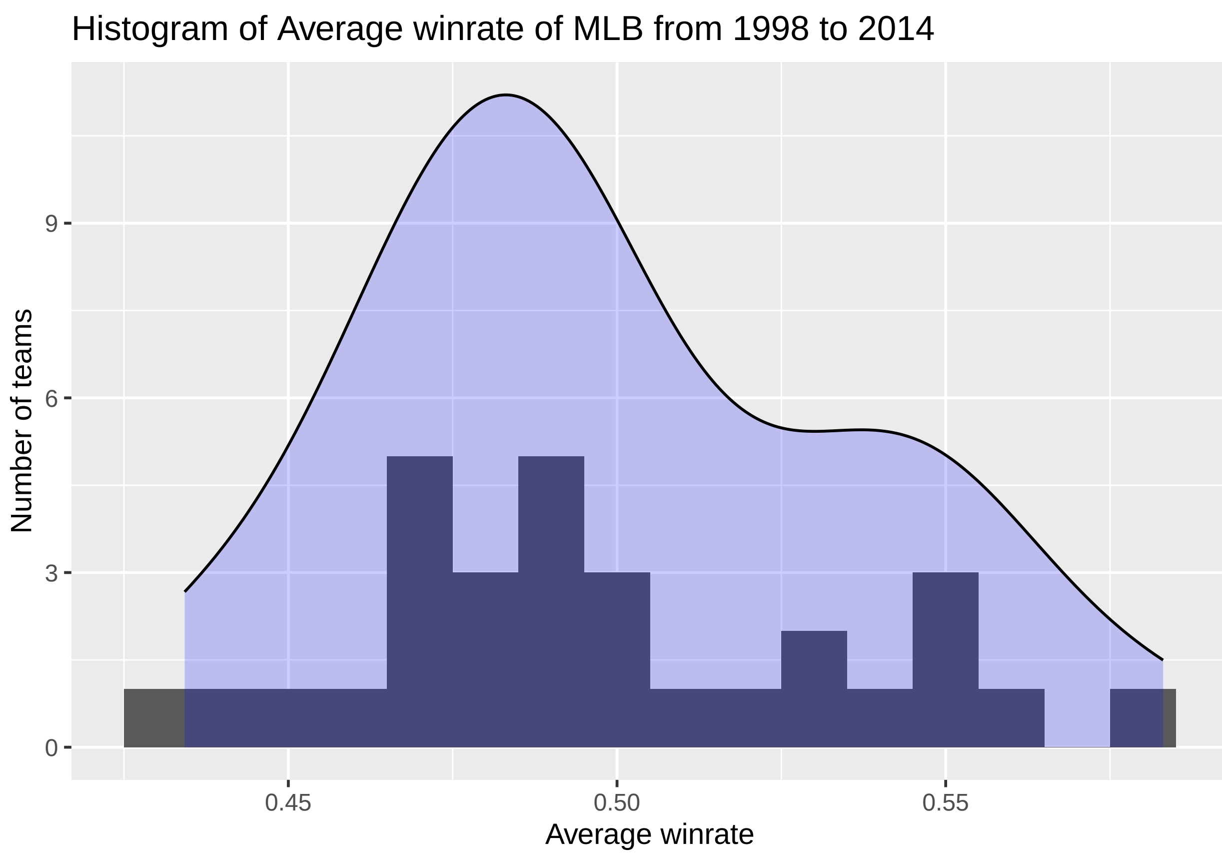



Early on in our data, we can see that there is a uniform gap between the minimum, average, and maximum payroll spending. As the years went by, the average spending on payroll remained close to the minimum payroll spending for each year. In contrast, the maximum payroll spending shot up, creating a huge gap.

#### Win percentage across years

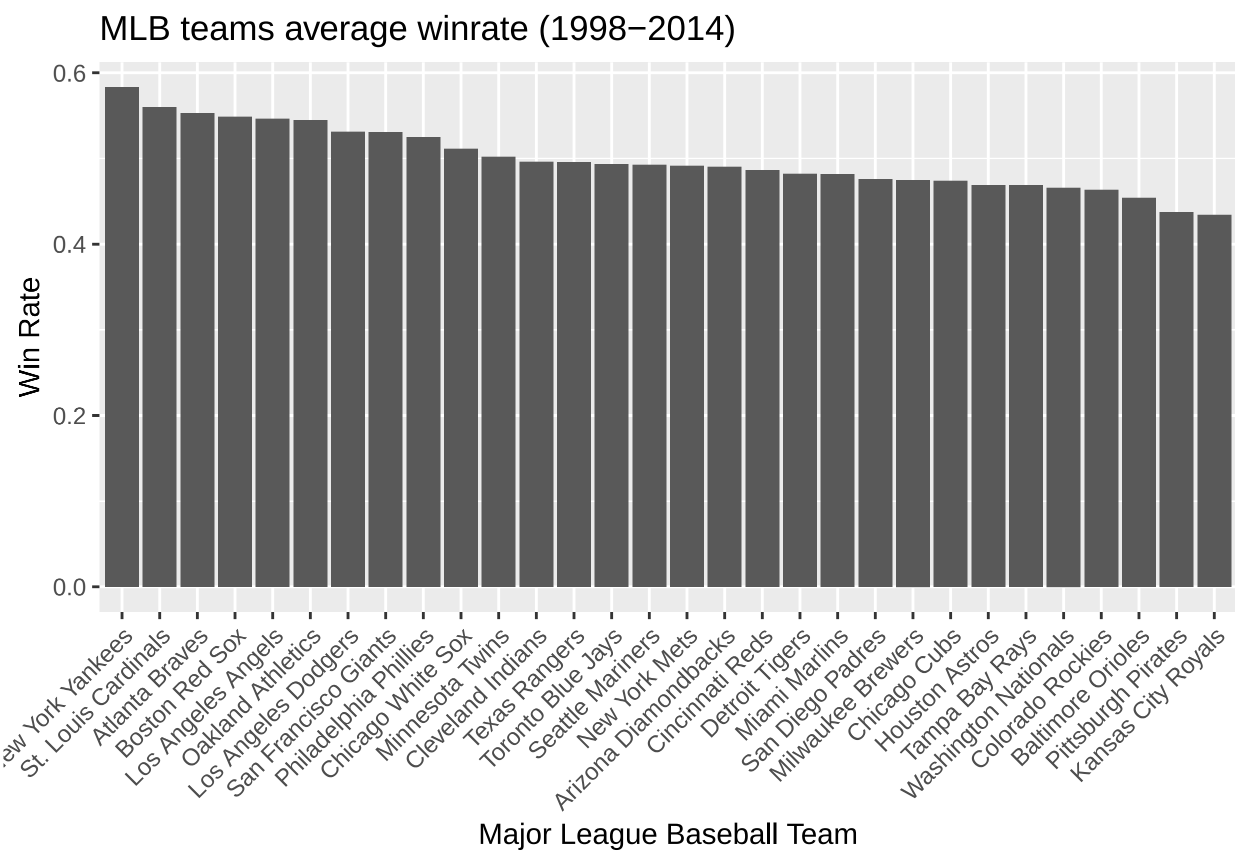
First, we want to look at the histogram of the average win rates from 1998 to 2014.

```
aggregate_table_mlb %>%
  ggplot(aes(x=avg_winrate)) +
  geom_histogram(binwidth = 0.01) +
  geom_density(alpha=.2, fill="blue")+
  xlab("Average winrate")+
  ylab("Number of teams")+
  ggtitle("Histogram of Average winrate of MLB from 1998 to 2014")
```




Next, we want to look at the teams with the highest and lowest win rates

```
aggregate_table_mlb %>% arrange(desc(avg_winrate)) %>%
  ggplot( aes(x=reorder(MLB_Team, -avg_winrate)  , y=avg_winrate) ) +
  geom_bar(stat = "identity")+
  theme(axis.text.x=element_text(angle=45,hjust=1, vjust = 1))+
  xlab("Major League Baseball Team") +
  ylab("Win Rate") +
  ggtitle("MLB teams average winrate (1998-2014)")
```




```r
aggregate_table_mlb %>% arrange(desc(avg_winrate)) %>%
  slice(sort(c(seq_len(5), n() - seq_len(5) +1))) %>%
  select(c("MLB_Team", "avg_winrate"))
```

```
##                 MLB_Team avg_winrate
## 1       New York Yankees   0.5830719
## 2    St. Louis Cardinals   0.5595414
## 3         Atlanta Braves   0.5527605
## 4         Boston Red Sox   0.5487172
## 5     Los Angeles Angels   0.5463819
## 6   Washington Nationals   0.4660195
## 7       Colorado Rockies   0.4633760
## 8      Baltimore Orioles   0.4538250
## 9     Pittsburgh Pirates   0.4371254
## 10    Kansas City Royals   0.4342288
```

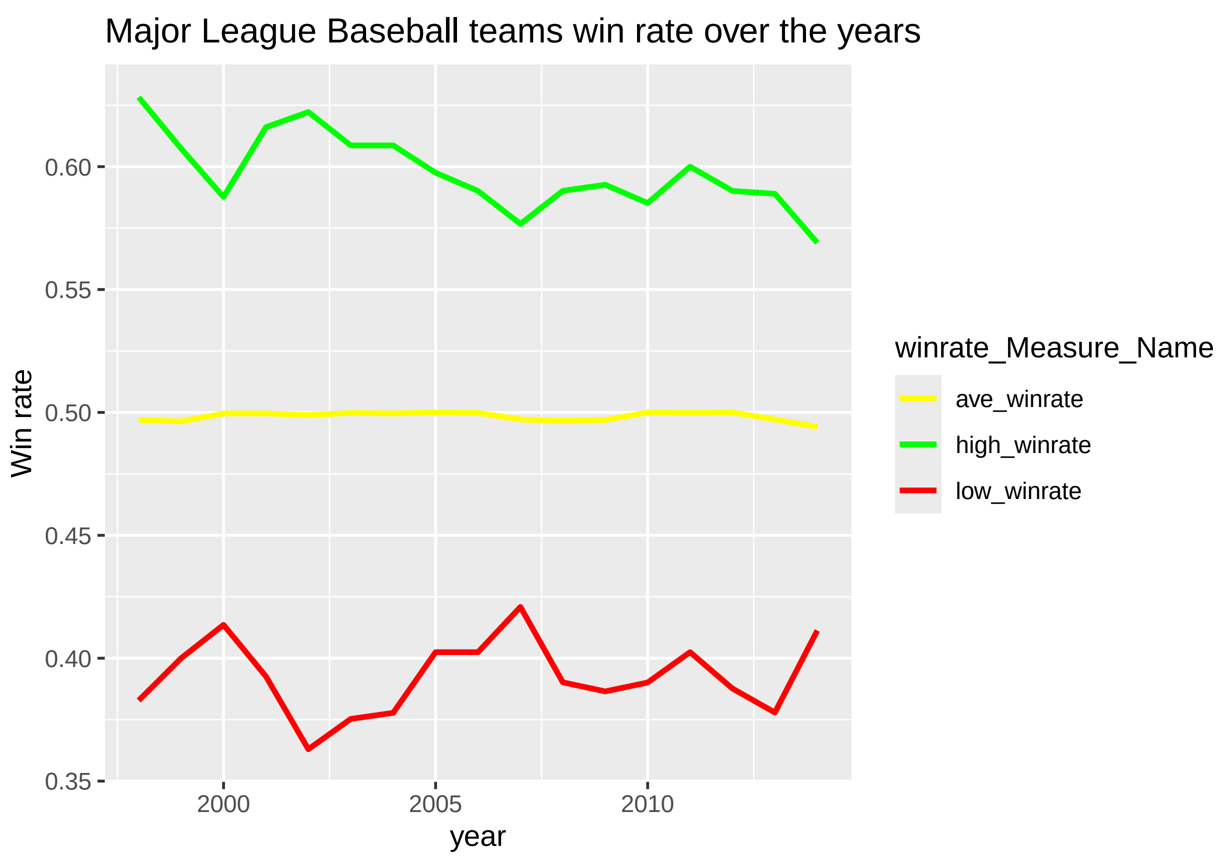
```r
highest_winrate <- aggregate_table_mlb %>% arrange(desc(avg_winrate)) %>%
  slice(sort(c(seq_len(5)))) %>%
  select(c("MLB_Team", "avg_winrate"))

lowest_winrate <- aggregate_table_mlb %>% arrange(desc(avg_winrate)) %>%
  slice(sort(c(n() - seq_len(5) +1))) %>%
  select(c("MLB_Team", "avg_winrate"))
```

The most successful team has been the New York Yankees, with a 58.31% win rate. They are followed by the St. Louis Cardinals, the Atlanta Braves, the Boston Red Sox, and the Los Angeles Angels.

The most unsuccessful was Kansas City Royals with a 43.42% winrate, followed by the Pittsburgh Pirates, the Baltimore Orioles, the Colorado Rockies, and the Washington Nationals, all with a win rate below 47%

Now, let's graph the average win rate for each year, as well as the average win rate of the top 5 best performing teams that year, and the average win rate of the bottom 5 worst performing teams.

```
mlb_pay_wincount_winrate %>% group_by(year) %>%
  summarise(
    ave_winrate = mean(win_Rate),
    high_winrate = mean(sort(win_Rate, decreasing = TRUE)[1:5]),
    low_winrate = mean(sort(win_Rate, decreasing = FALSE)[1:5])
    ) %>%
  pivot_longer(
      c("ave_winrate", "high_winrate", "low_winrate"),
      names_to = "winrate_Measure_Name",
      values_to = "winrate_Measure"
      ) %>%
  ggplot(aes(x=year, y=winrate_Measure, color=winrate_Measure_Name))+
  geom_line(linewidth=1)+
  scale_color_manual(values =
                       c("ave_winrate" = "yellow",
                         "high_winrate"="green",
                         "low_winrate"= "red")) +
  ylab("Win rate")+
  ggtitle("Major League Baseball teams win rate over the years")
```




Of course, since there are no ties in Major League Baseball, the average win rate stays at around 50%.

From 1998 to 2014, while erratic, the average win rate of the top 5 best performing teams have actually trended closer to the 50% win rate line, albeit slightly.

However, the average win rate of the 5 teams with the lowest win rates for each season have been erratic, with varying levels of success.

It can also be noted that, when the average win rate of the worst performing teams increase, the average win rate of the best performing teams for that year decreases. The opposite is also observed.

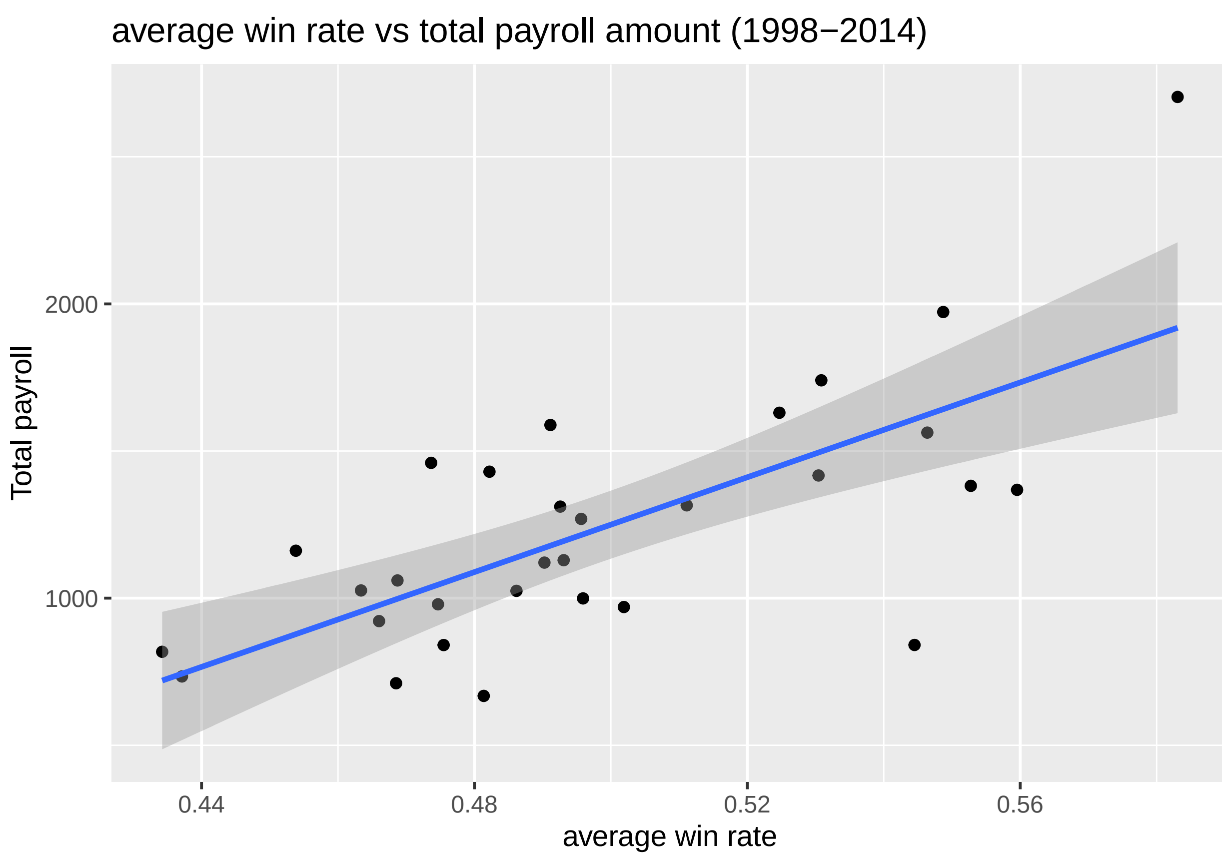
### Win percentage versus payroll

We have already explored the payroll and win rate variables. Now, let's see if they have any interaction.

Since we're dealing with continuous variables, let's use a scatter plot to visualize them.

let's start with the aggregate/overall data

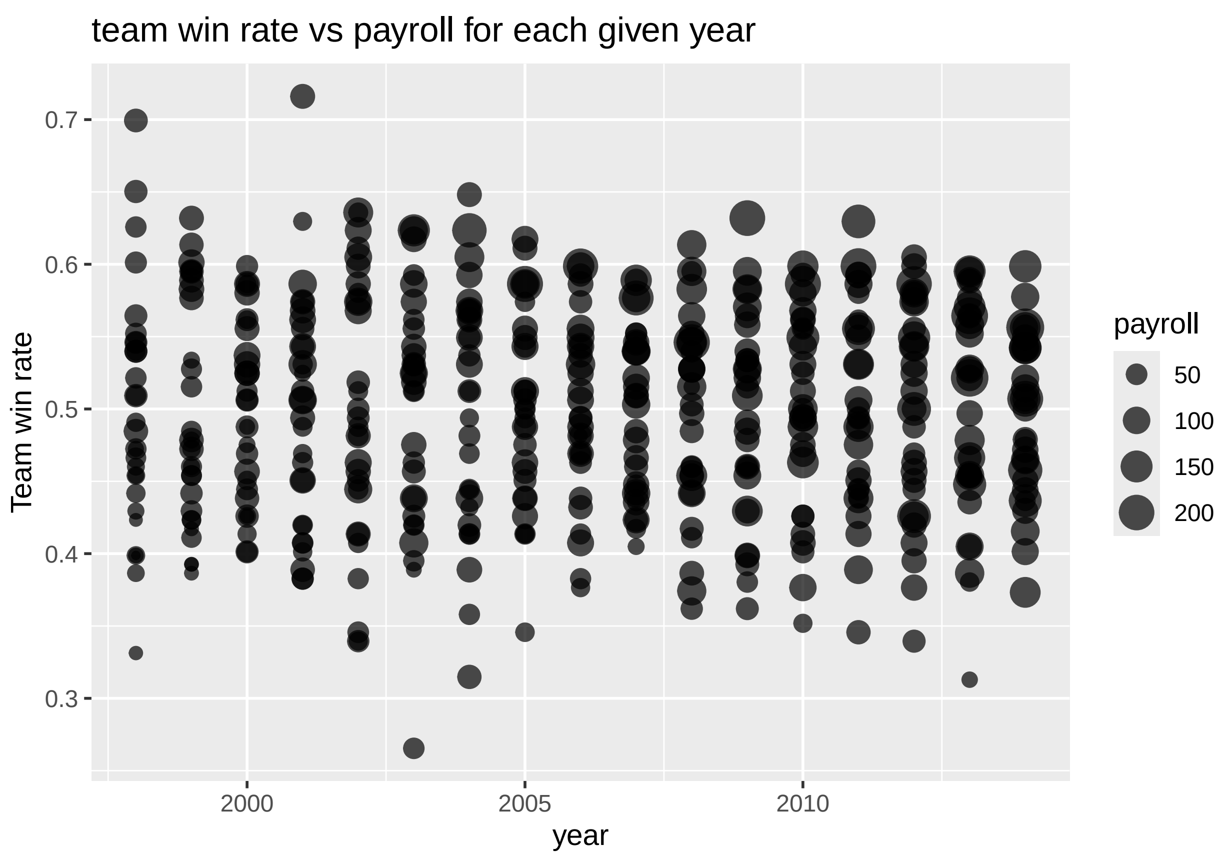
```
aggregate_table_mlb %>%
  ggplot(aes(x=avg_winrate, y=total_pay)) +
  geom_point() +
  stat_smooth(
    method = "lm",
    formula = y ~ x,
    geom = "smooth"
    )+
  ylab("Total payroll")+
  xlab("average win rate")+
  ggtitle("average win rate vs total payroll amount (1998-2014) ")
```




Although it isn't clear, the line of best fit shows that the average win rate increases with the total payment

Now, lets use our yearly data too see if time makes a difference.

```
mlb_pay_wincount_winrate %>%
  ggplot(aes(x=year, y=win_Rate, size = payroll)) +
  geom_point(alpha=0.7) +
  scale_size()+
  ylab("Team win rate")+
  ggtitle("team win rate vs payroll for each given year")
```



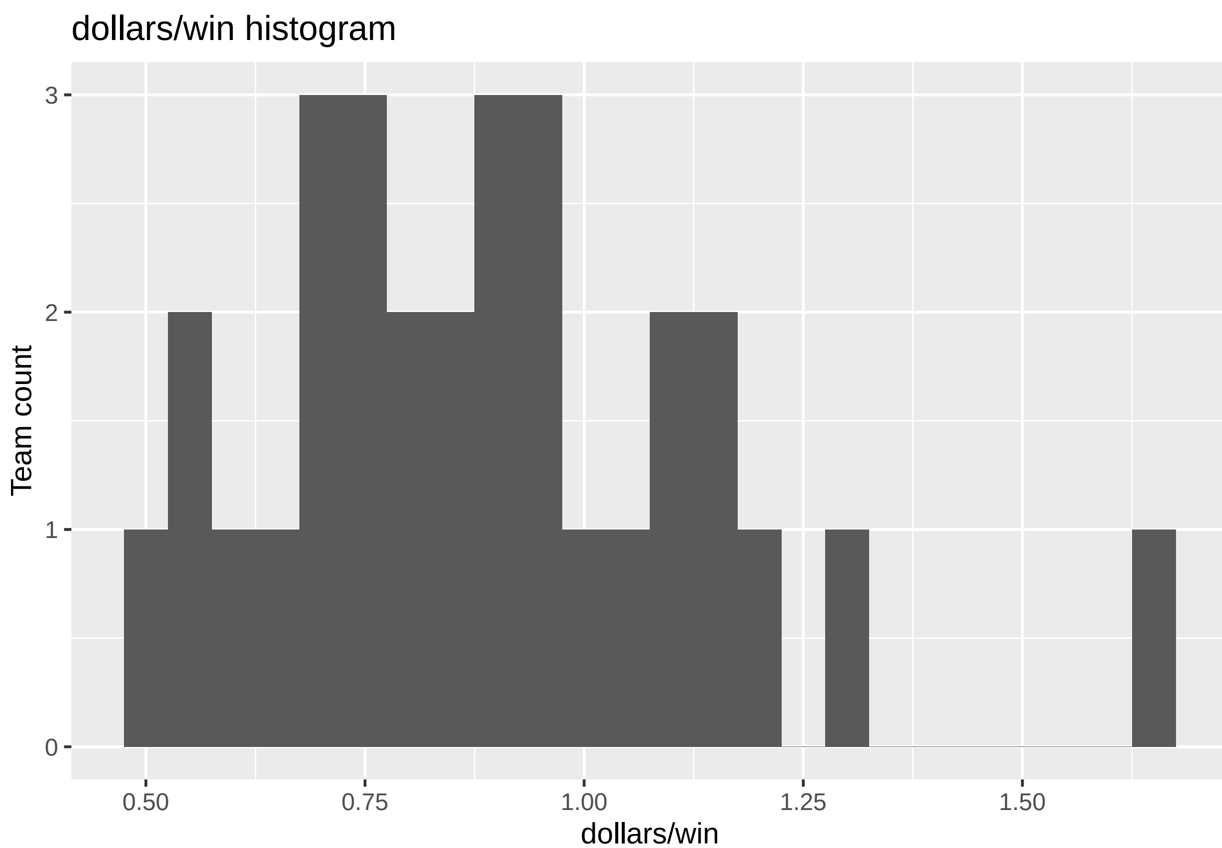

As we can see from the bubble chart, as the year goes by, the discrepancy of the win rates between the teams actually went down. This coincides with the increase of the average payroll per year. We can also see that the teams with the highest payroll is often among the top teams in terms of win rates, but that teams with a smaller payroll budget for that year can sometimes perform better.

### Team efficiency

In team efficiency, we are using dollars per win, which we would be prudent to remember is actually million dollars/win.

First, let's create a histogram for the aggregate efficiency

```
aggregate_table_mlb %>% ggplot(aes(x=dollars_per_win)) +
  geom_histogram(binwidth = 0.05)+
  xlab("dollars/win")+
  ylab("Team count")+
  ggtitle("dollars/win histogram")
```




We can see that most teams spend an aggregate amount of 750 thousand to 1.25 million USD for each win from 1998 to 2014. In that regard, we have one outlier, spending around 1.6 million per win, which is very inefficient.

Next, let's see the teams with the highest efficiency

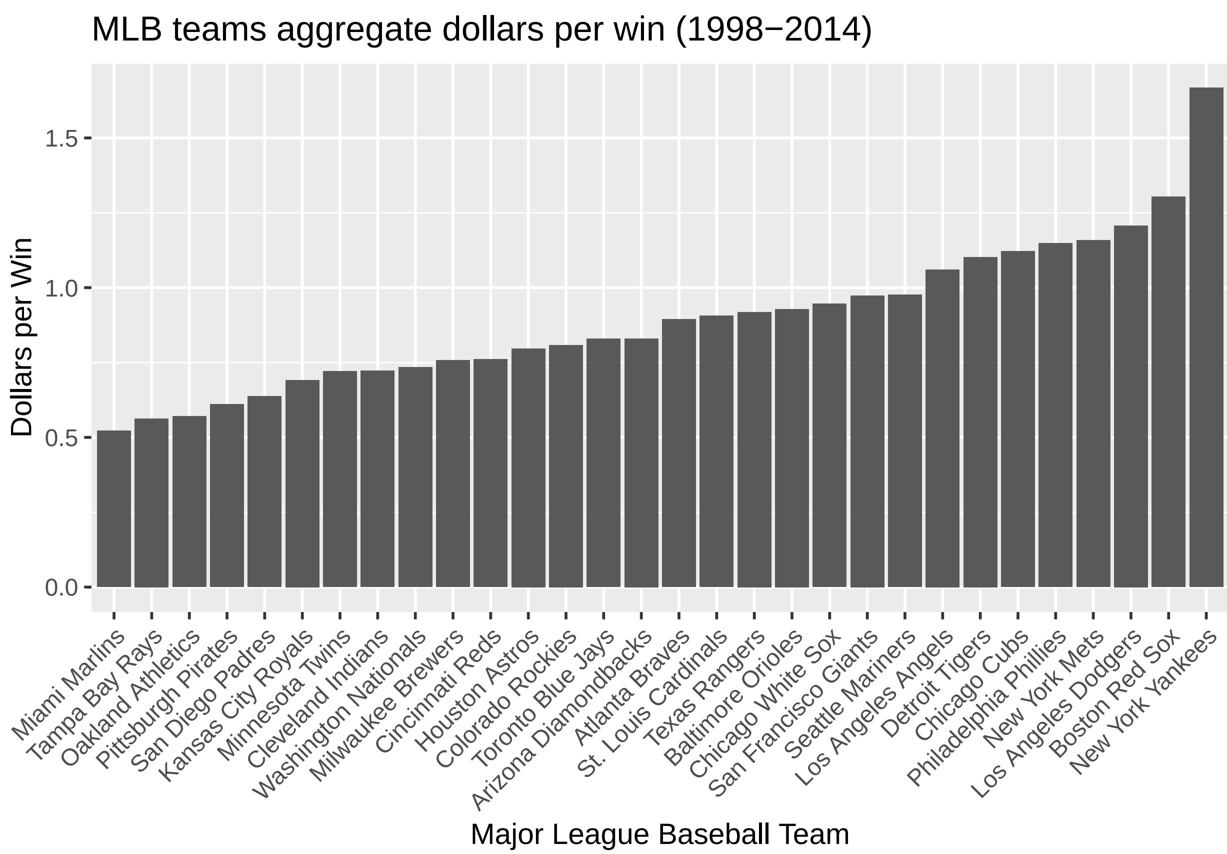
```
aggregate_table_mlb %>% arrange(dollars_per_win) %>%
  slice(sort(c(seq_len(5), n() - seq_len(5) +1))) %>%
  select(c("MLB_Team", "dollars_per_win"))
```

```
##                 MLB_Team dollars_per_win
## 1          Miami Marlins       0.5217202
## 2         Tampa Bay Rays       0.5627786
## 3      Oakland Athletics       0.5701248
## 4     Pittsburgh Pirates       0.6110788
## 5       San Diego Padres       0.6368688
## 6  Philadelphia Phillies       1.1479724
## 7          New York Mets       1.1585914
## 8    Los Angeles Dodgers       1.2076835
## 9         Boston Red Sox       1.3036079
## 10      New York Yankees       1.6676423
```

```
low_dolperwin <- aggregate_table_mlb %>% arrange(dollars_per_win) %>%
  slice(sort(c(seq_len(5)))) %>%
  select(c("MLB_Team", "dollars_per_win"))

high_dolperwin <- aggregate_table_mlb %>% arrange(dollars_per_win) %>%
  slice(sort(c(n() - seq_len(5) +1))) %>%
```

```r
  select(c("MLB_Team", "dollars_per_win"))
```

```r
aggregate_table_mlb %>% arrange(dollars_per_win) %>%
  ggplot( aes(x=reorder(MLB_Team, dollars_per_win)  , y=dollars_per_win) ) +
  geom_bar(stat = "identity")+
  theme(axis.text.x=element_text(angle=45,hjust=1, vjust = 1))+
  xlab("Major League Baseball Team") +
  ylab("Dollars per Win") +
  ggtitle("MLB teams aggregate dollars per win (1998-2014)")
```



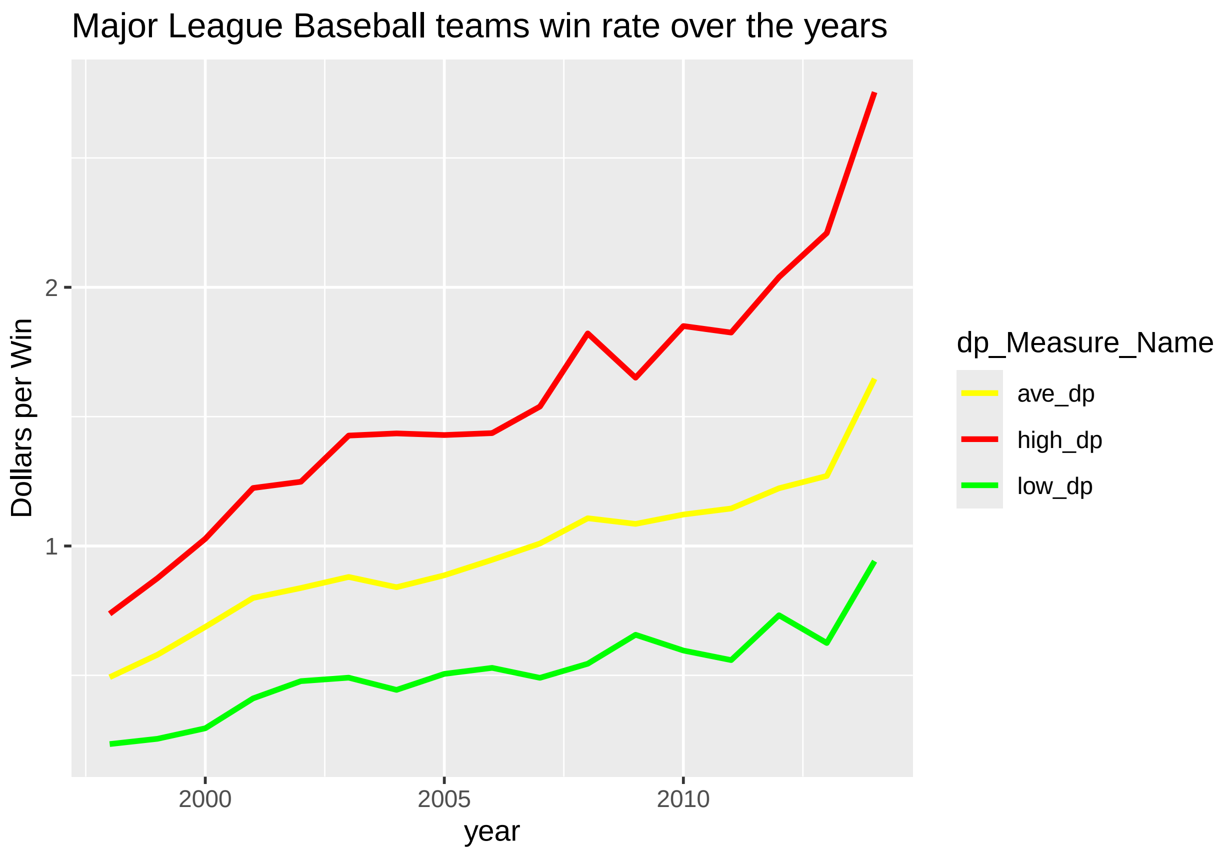

We can see that the Miami Marlins are the most efficient team in the MLB in terms of dollars per win, spending roughly 521.7 thousand USD per win. It is followed by the Tampa Bay Rays, Oakland Athletics, Pittsburgh Pirates, and the San Diego Padres. Notably, these are also the teams have the lowest total payroll spending, and the Pittsburgh Pirates are among those with the lowest average win rate.

In terms of most inefficient, the New York Yankees spends 1.67 million dollars for each win. The Boston Red Sox 1.3 million for each, followed by the Los Angeles Dodgers, New York Mets, and the Philadelphia Phillies.

Next, let's see how the average efficiency changes per year:

```r
mlb_pay_wincount_winrate %>% group_by(year) %>%
  summarise(
    ave_dp = mean(dollars_per_win),
    high_dp = mean(sort(dollars_per_win, decreasing = TRUE)[1:5]),
    low_dp = mean(sort(dollars_per_win, decreasing = FALSE)[1:5])
    ) %>%
  pivot_longer(
      c("ave_dp", "high_dp", "low_dp"),
```

```
    names_to = "dp_Measure_Name",
    values_to = "dp_Measure"
    ) %>%
  ggplot(aes(x=year, y=dp_Measure, color=dp_Measure_Name))+
  geom_line(linewidth=1)+
  scale_color_manual(values =
                      c("ave_dp" = "yellow",
                        "high_dp"="red",
                        "low_dp"= "green")) +
  ylab("Dollars per Win")+
  ggtitle("Major League Baseball teams win rate over the years")
```




The graph shows that the average dollars per win increases within the league for each passing season, a sign that the league is getting more and more competitive after each season. Increase in dollars per win was somewhat uniform, except for the years after 2010, when teams with low efficiency started paying even more for each win than those with high efficiency, although there was a spike all across the board.