MLB Payroll analysis

Group 9 Yun An Chen, Hsin Yun Li, Po wei Lin

2025-10-31

Contents

1 Importing and Exploring the Dataset **2**

2 Data Cleaning and Variable Preparation **3**

3 Payroll vs. Winning Percentage (Scatterplot with Regression Line) **4**

4 Winning Percentage by Payroll Quartile (Boxplot) **5**

5 Playoff qualification rate by payroll quartile (proportion barplot) **6**

6 Payroll vs playoff probability (loess or logistic preview) **8**

7 Distribution of MLB Team Payrolls (Histogram & Density) **10**

8 Executive Summary **12**

9 OLS **13**

10 Payroll vs. Playoff Probability — Logistic Regression **16**

11 Logistic regression model + Odds Ratios **16**

12 LOESS smoother: non-parametric preview **18**

13 Binned playoff rates (non-parametric) **19**

14 Predicted curves by year (optional) **20**

1 Importing and Exploring the Dataset

```
#Import the dataset
raw_url <- "https://raw.githubusercontent.com/pwlinbernie/MLB-payroll-analysis/refs/heads/main/
mlb <- read_csv(raw_url)

#check
dim(mlb)
```

```
## [1] 420  12
```

```
head(mlb)
```

```
## # A tibble: 6 x 12
##   Team  `Team Name`    Year `Average Age` Total Payroll Alloca~1 `Active 26-Man`
##   <chr> <chr>         <dbl>         <dbl> <chr>                  <chr>
## 1 OAK   Oakland Athl~  2024          26.5 $62,132,581            $28,956,713
## 2 PIT   Pittsburgh P~  2024          27.7 $84,050,989            $51,220,210
## 3 TB    Tampa Bay Ra~  2024          26.8 $89,707,422            $37,691,876
## 4 DET   Detroit Tige~  2024          26   $96,961,614            $33,226,992
## 5 CLE   Cleveland Gu~  2024          26.3 $105,224,582           $50,885,032
## 6 BAL   Baltimore Or~  2024          28.4 $109,335,494           $65,994,548
## # i abbreviated name: 1: `Total Payroll Allocations`
## # i 6 more variables: Injured <chr>, Retained <chr>, Buried <chr>, Wins <dbl>,
## #   Losses <dbl>, Postseason <chr>
```

```
glimpse(mlb)
```

```
## Rows: 420
## Columns: 12
## $ Team                        <chr> "OAK", "PIT", "TB", "DET", "CLE", "BAL", "~
## $ `Team Name`                 <chr> "Oakland Athletics", "Pittsburgh Pirates",~
## $ Year                        <dbl> 2024, 2024, 2024, 2024, 2024, 2024, 2024, ~
## $ `Average Age`               <dbl> 26.5, 27.7, 26.8, 26.0, 26.3, 28.4, 27.9, ~
## $ `Total Payroll Allocations` <chr> "$62,132,581", "$84,050,989", "$89,707,422~
## $ `Active 26-Man`             <chr> "$28,956,713", "$51,220,210", "$37,691,876~
## $ Injured                     <chr> "$15,581,092", "$14,524,211", "$13,179,262~
## $ Retained                    <chr> "$15,557,073", "$15,341,351", "$34,675,167~
## $ Buried                      <chr> "$1,763,221", "$2,965,217", "$1,706,572", ~
## $ Wins                        <dbl> 69, 76, 80, 86, 92, 91, 93, 77, 86, 71, 62~
## $ Losses                      <dbl> 93, 86, 82, 76, 69, 71, 69, 85, 76, 91, 10~
## $ Postseason                  <chr> "No Playoffs", "No Playoffs", "No Playoffs~
```

The dataset contains information on each MLB team's annual payroll and performance from multiple seasons, including total payroll allocations, wins, losses, and postseason results. Before conducting any analysis, it's essential to understand the data structure and verify that the variables are properly imported.

2 Data Cleaning and Variable Preparation

```
# Convert wins/losses to numeric and compute winning percentage
mlb <- mlb |>
    clean_names() |>
    mutate(
      team = coalesce(team, team_name),
      year = as.integer(year),

      payroll = parse_number(total_payroll_allocations),
      active26 = parse_number(active_26_man),

      # Compute total games and winning percentage for each team-season
      wins = as.integer(wins),
      losses = as.integer(losses),
      games = wins + losses,
      win_pct = if_else(games > 0, wins / games, NA_real_),

      #postseason
      postseason =na_if(as.character(postseason), ""),
      playoffs_bin = case_when(
        is.na(postseason) ~0L,
        str_detect(str_to_lower(postseason), "no playoffs") ~0L,
        TRUE ~ 1L
      ),

      log_payroll = log(payroll)
)

# check
names(mlb)
```

```
##  [1] "team"                     "team_name"
##  [3] "year"                     "average_age"
##  [5] "total_payroll_allocations" "active_26_man"
##  [7] "injured"                  "retained"
##  [9] "buried"                   "wins"
## [11] "losses"                   "postseason"
## [13] "payroll"                  "active26"
## [15] "games"                    "win_pct"
## [17] "playoffs_bin"             "log_payroll"
```

```
summary(select(mlb, payroll, active26, wins, losses, win_pct, playoffs_bin))
```

```
##     payroll            active26            wins            losses
```

```
##  Min.   : 23478635   Min.   : 10252278   Min.   : 19.00   Min.   : 17.00
##  1st Qu.: 85220920   1st Qu.: 56963032   1st Qu.: 69.00   1st Qu.: 69.00
##  Median :116035246   Median : 83823420   Median : 80.00   Median : 79.00
##  Mean   :127645094   Mean   : 93876306   Mean   : 77.34   Mean   : 77.34
##  3rd Qu.:162204268   3rd Qu.:124147463   3rd Qu.: 90.00   3rd Qu.: 89.00
##  Max.   :341673777   Max.   :260231924   Max.   :111.00   Max.   :121.00
##     win_pct        playoffs_bin
##  Min.   :0.2531   Min.   :0.0000
##  1st Qu.:0.4444   1st Qu.:0.0000
##  Median :0.5000   Median :0.0000
##  Mean   :0.5000   Mean   :0.3429
##  3rd Qu.:0.5583   3rd Qu.:1.0000
##  Max.   :0.7167   Max.   :1.0000
```

```
sum(is.na(mlb$payroll)); sum(is.na(mlb$active26))
```

```
## [1] 0
```

```
## [1] 0
```

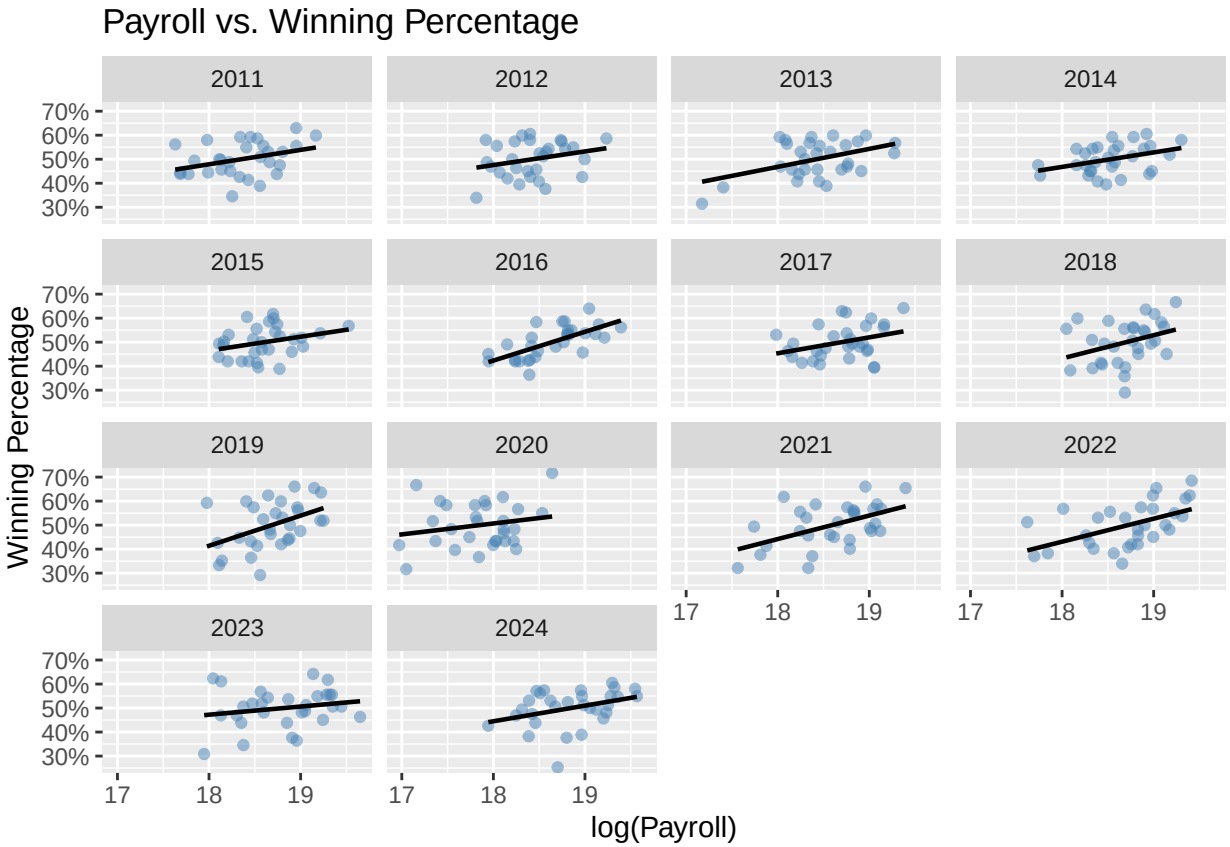
Standardizing column names and converts key variables into numeric formats.

payroll and active26 are parsed as numerical values. win_pct is calculated from total wins and losses. playoffs_bin is created as a binary indicator (1 = made playoffs, 0 = did not). log_payroll is the natural logarithm of payroll, reducing skewness in later analysis.

#————————————————————————————————————# # First Visualization # #————————————————————————————————————#

3 Payroll vs. Winning Percentage (Scatterplot with Regression Line)

```
ggplot(mlb, aes(x = log_payroll, y = win_pct)) +
  geom_point(alpha = 0.5, color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "black", linewidth = 0.8) +
  scale_y_continuous(labels = percent) +
  labs(title = "Payroll vs. Winning Percentage",
       x = "log(Payroll)", y = "Winning Percentage") +
  facet_wrap(~ year)
```

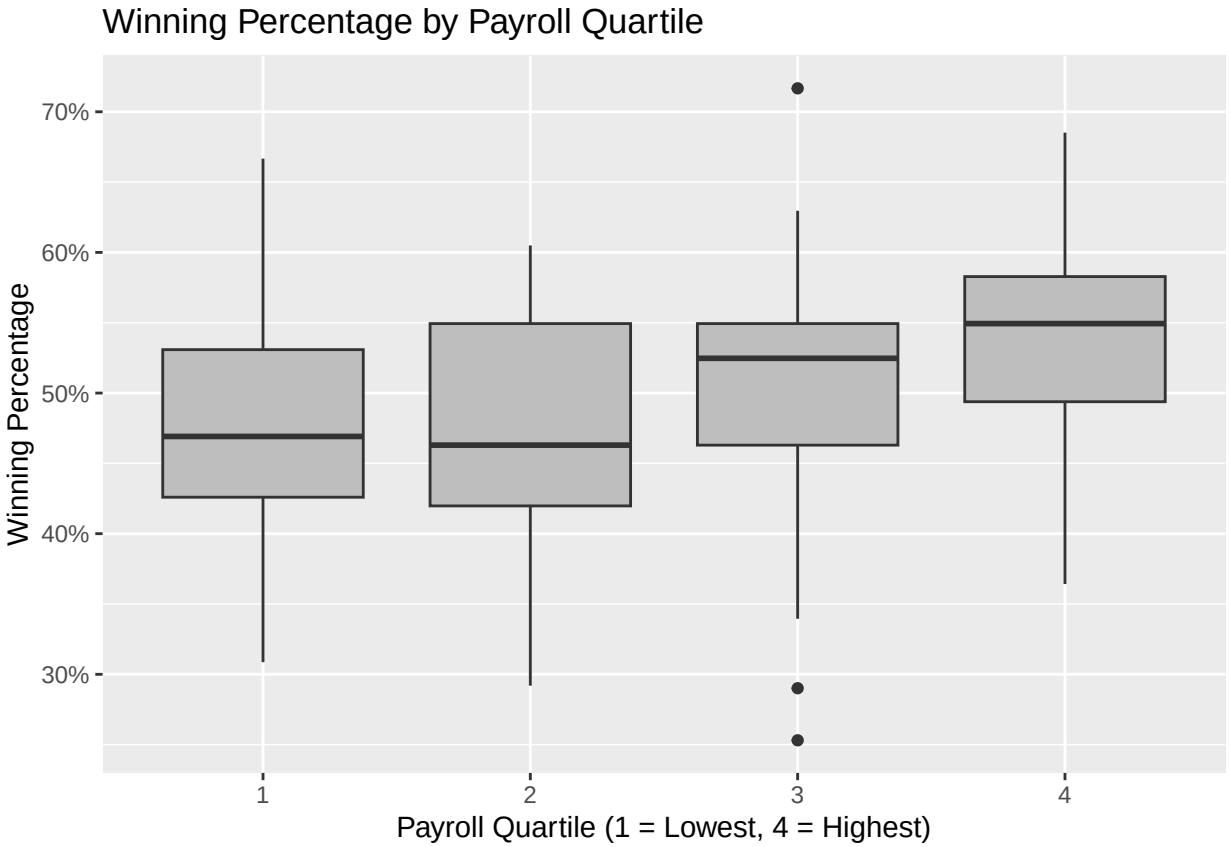


For the scatterplot, we can see that: 1. Across almost all panels, the fitted OLS line has a positive slope, indicating that higher payroll is associated with higher winning percentage in the same season. 2. Teams with very high payrolls cluster toward the upper-right; however, there is substantial vertical spread at many payroll levels, implying that payroll alone does not fully determine performance. 3. A few low-payroll but above-.500 teams (upper-left area) and high-payroll under-.500 teams (lower-right area) illustrate that efficiency, roster construction, injuries, and coaching likely moderate the payroll–performance link.

There is a consistent positive relationship between payroll and winning percentage within seasons, but with meaningful dispersion—payroll helps, yet it does not guarantee success.

4 Winning Percentage by Payroll Quartile (Boxplot)

```
mlb <- mlb |>
  mutate(payroll_q = ntile(payroll, 4))

ggplot(mlb, aes(x = factor(payroll_q), y = win_pct)) +
  geom_boxplot(fill = "gray") +
  scale_y_continuous(labels = percent) +
  labs(title = "Winning Percentage by Payroll Quartile",
       x = "Payroll Quartile (1 = Lowest, 4 = Highest)",
       y = "Winning Percentage")
```



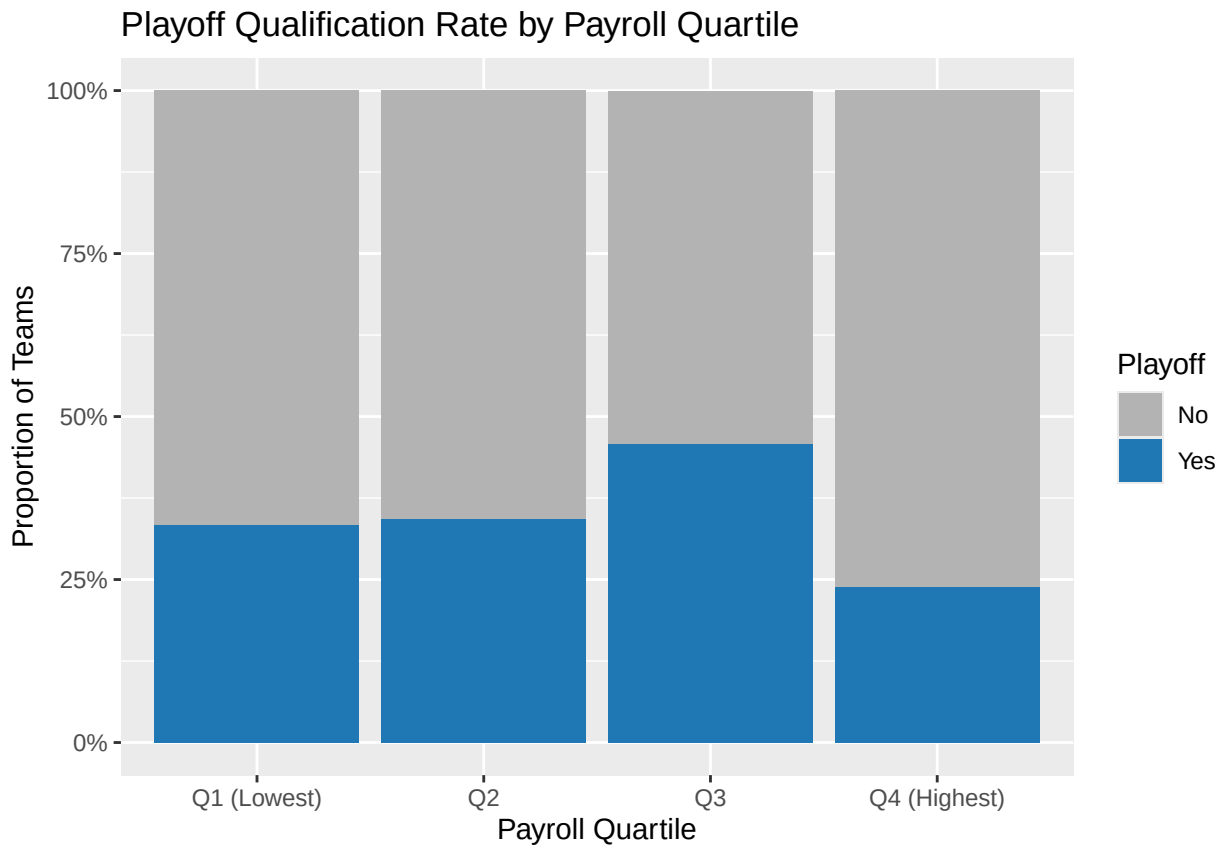
For the boxplot, we can see that: 1. The median winning percentage rises from Q1 (lowest payroll) to Q4 (highest payroll), consistent with the hypothesis that higher payrolls are associated with better performance on average. 2. Both the interquartile range (IQR) and the upper whiskers shift upward across quartiles, meaning not only the typical team but also the upper performers tend to improve with payroll. 3. Despite the upward shift, the boxes/whiskers overlap across quartiles, showing that some lower-payroll teams outperform higher-payroll teams and vice versa. Payroll is a strong correlate, not a guarantee.

The quartile comparison provides clear descriptive evidence that higher payroll groups tend to win more, yet variance and overlap indicate that other factors (roster age, injuries, development, front-office quality) remain important.

5 Playoff qualification rate by payroll quartile (proportion barplot)

```
mlb$Payroll_Group <- ntile(mlb$total_payroll_allocations, 4)
mlb$Playoff <- ifelse(mlb$postseason == "No Playoffs", "No", "Yes")

mlb$Payroll_Group <- factor(mlb$Payroll_Group,
                            labels = c("Q1 (Lowest)", "Q2", "Q3", "Q4 (Highest)"))

ggplot(mlb, aes(x = factor(Payroll_Group), fill = Playoff)) +
```

```
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Playoff Qualification Rate by Payroll Quartile",
       x = "Payroll Quartile",
       y = "Proportion of Teams") +
  scale_fill_manual(values = c("No" = "gray70", "Yes" = "#1f78b4"))
```

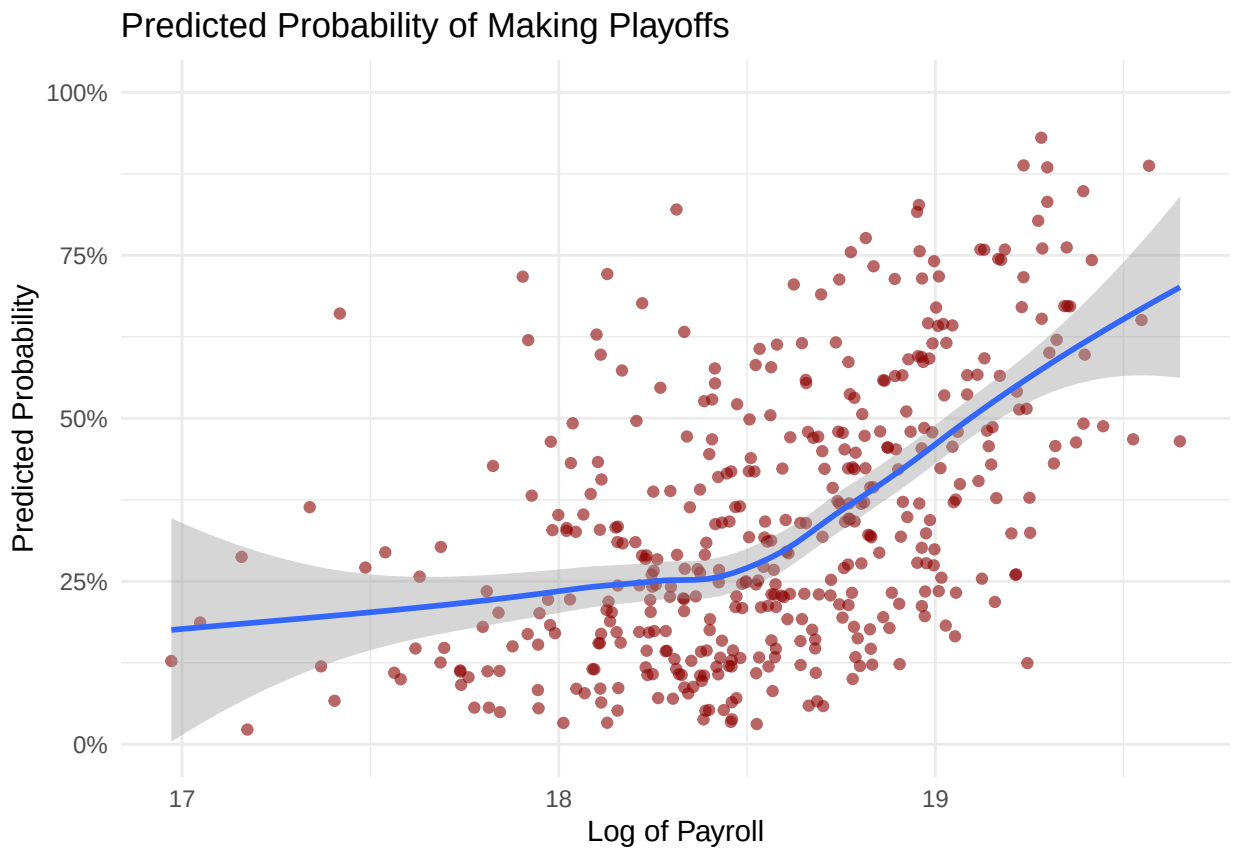


For the proportion barplot, we can see that: 1.Each bar represents a payroll quartile group (Q1–Q4), ranked from lowest to highest payroll. 2.The blue section shows the proportion of teams that qualified for the playoffs, while the gray section represents those that did not. 3.Higher-payroll teams (Q3–Q4) generally exhibit slightly higher qualification rates than lower-payroll teams (Q1–Q2). 4.However, the differences are not very large—some low-payroll teams still managed to reach the postseason. 5.This pattern indicates that payroll alone does not fully determine success; other factors such as player development, management strategy, and team efficiency also influence playoff outcomes.

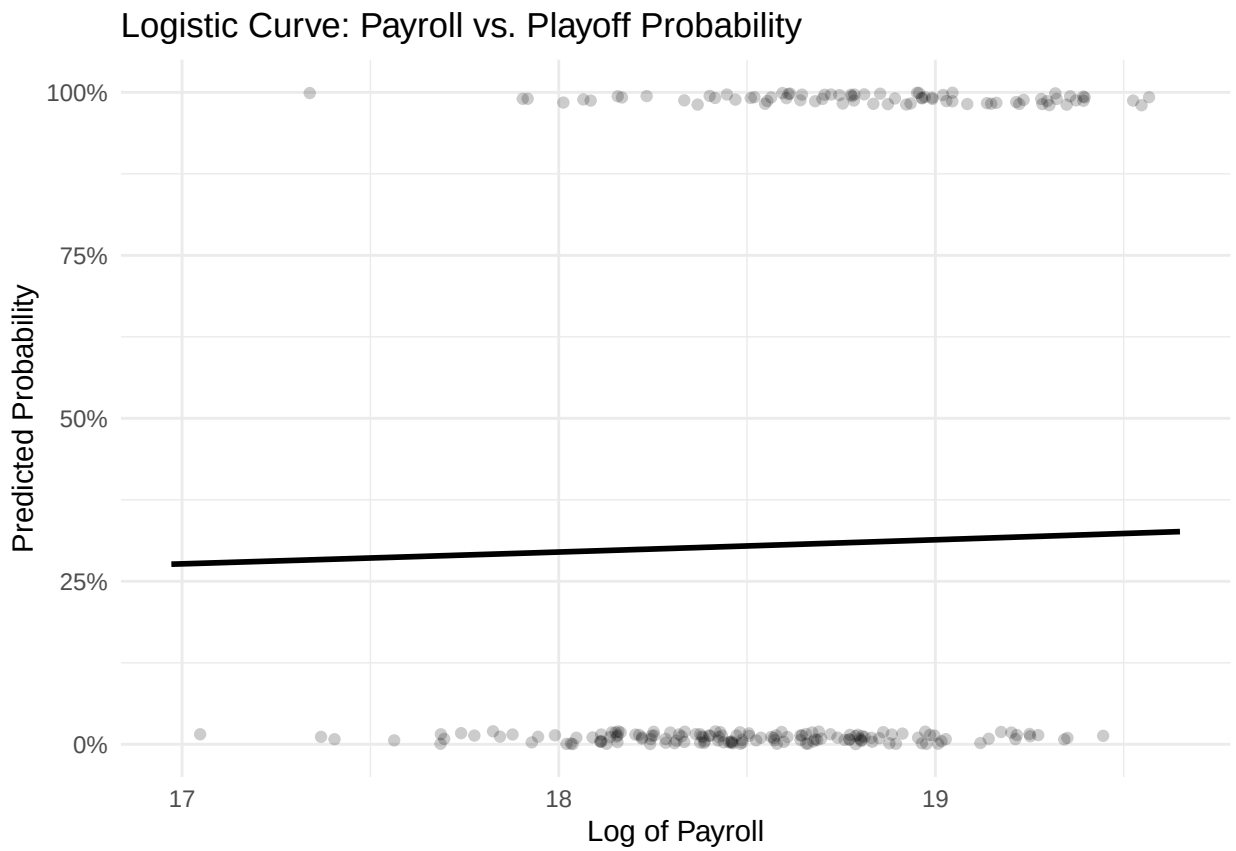
In summary,while there is a positive association between team payroll and playoff qualification, the relationship is not deterministic. Payroll provides a competitive advantage, but effective management and team strategy remain key to achieving postseason success.

6 Payroll vs playoff probability (loess or logistic preview)

```
mlb <- mlb %>%
  mutate(
    payroll_num  = parse_number(total_payroll_allocations),
    average_age  = as.numeric(average_age),
    year         = as.integer(year),
    Playoff_bin  = ifelse(postseason %in% c("No Play", "No Playoffs", "No"), 0, 1)
  )


model_logit <- glm(Playoff_bin ~ log(payroll_num) + average_age + year,
                   data = mlb, family = "binomial")


mlb$predicted_prob <- predict(model_logit, type = "response")


ggplot(mlb, aes(x = log(payroll_num), y = predicted_prob)) +
  geom_point(color = "darkred", alpha = 0.6) +
  geom_smooth(method = "loess") +
  scale_y_continuous(labels = percent, limits = c(0,1)) +
  labs(title = "Predicted Probability of Making Playoffs",
       x = "Log of Payroll", y = "Predicted Probability") +
  theme_minimal()
```




```
newdat <- with(mlb, data.frame(
  payroll_num = seq(min(payroll_num, na.rm = TRUE),
                    max(payroll_num, na.rm = TRUE), length.out = 200),
  average_age = mean(average_age, na.rm = TRUE),
  year        = max(year, na.rm = TRUE)
))
newdat$pred <- predict(model_logit, newdata = newdat, type = "response")

ggplot() +
  geom_point(data = mlb,
             aes(x = log(payroll_num), y = Playoff_bin),
             alpha = 0.2, position = position_jitter(height = 0.02)) +
  geom_line(data = newdat,
            aes(x = log(payroll_num), y = pred), linewidth = 1) +
  scale_y_continuous(labels = percent, limits = c(0,1)) +
  labs(title = "Logistic Curve: Payroll vs. Playoff Probability",
       x = "Log of Payroll", y = "Predicted Probability") +
  theme_minimal()
```



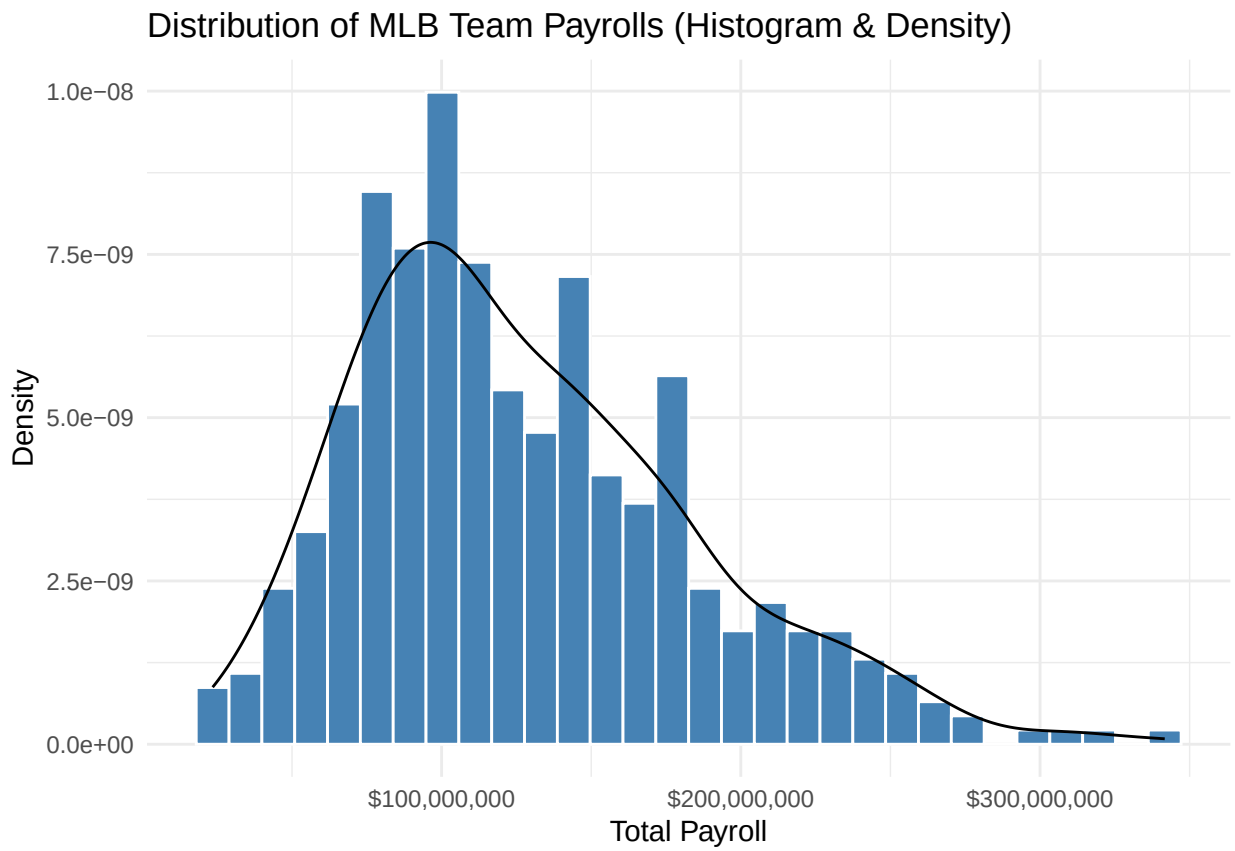
For the playoff probability plots, we can see that: 1.The nonlinear loess smoothing plot (left) illustrates the actual data trend between team payroll and the probability of making the playoffs. 2.The blue curve shows a clear upward trend: as payroll increases, the predicted playoff probability rises noticeably, especially when the logarithm of payroll exceeds approximately 18.5. 3.The widening gray confidence bands at both ends indicate higher uncertainty due to fewer teams at extreme payroll levels. 4.The logistic regression curve (right) depicts the model-based linear-in-logit relationship between payroll and playoff probability. 5.The nearly flat black line suggests that, after controlling for year and team age, the marginal effect of payroll on playoff qualification is relatively small. 6.Comparing both plots reveals that the true relationship is nonlinear—the logistic model underestimates the sharp increase in playoff probability among top-spending teams.

In summary,payroll has a positive association with playoff qualification, but the effect is nonlinear. High-payroll teams enjoy a substantially greater likelihood of reaching the postseason, while payroll advantages become significant only beyond the top spending levels.

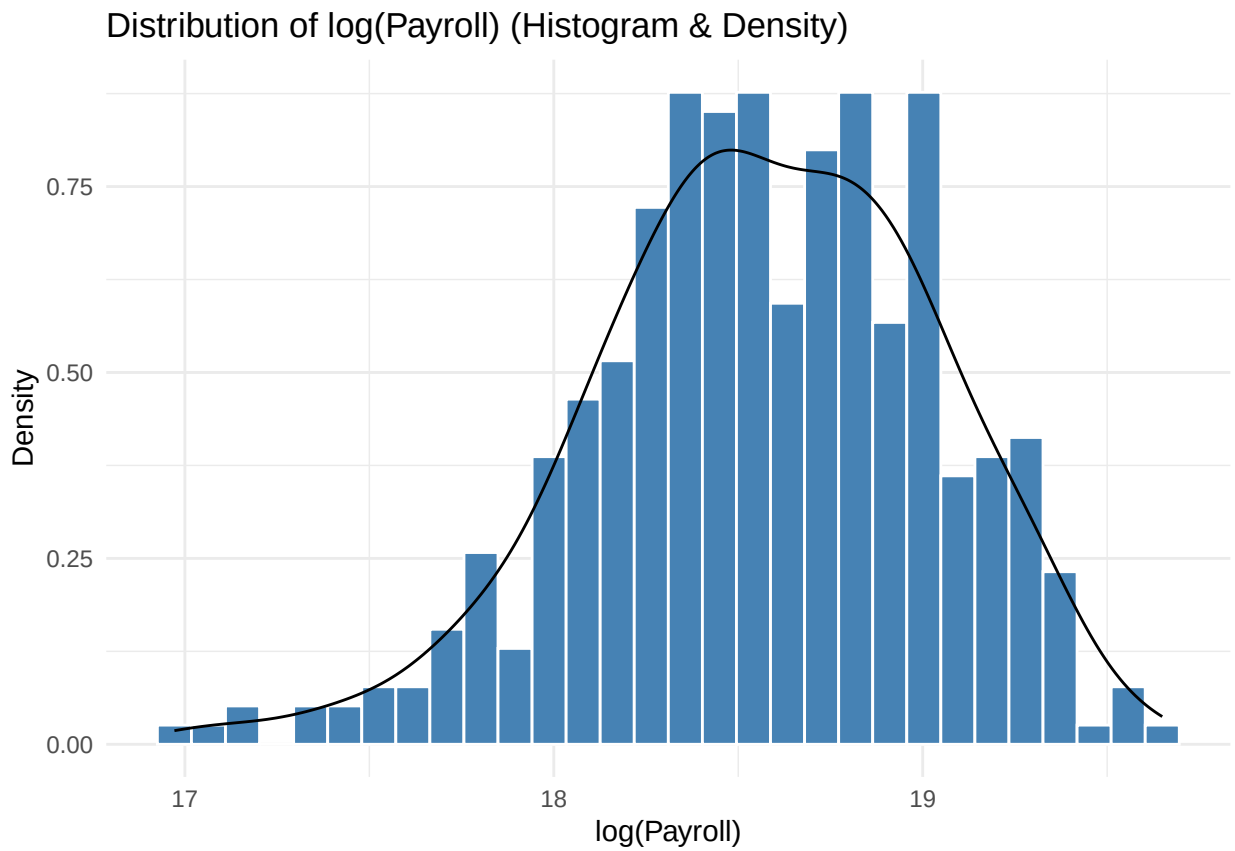
7 Distribution of MLB Team Payrolls (Histogram & Density)

```
# Raw payroll: histogram on density scale + density curve
ggplot(mlb, aes(x = payroll)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30,
                 fill = "steelblue", color = "white") +
  geom_density(adjust = 1.1) +
```

```
  scale_x_continuous(labels = scales::label_dollar()) +
  labs(x = "Total Payroll", y = "Density",
       title = "Distribution of MLB Team Payrolls (Histogram & Density)") +
  theme_minimal()
```




```
# Log payroll: same idea
ggplot(mlb, aes(x = log_payroll)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30,
                 fill = "steelblue", color = "white") +
  geom_density(adjust = 1.1) +
  labs(x = "log(Payroll)", y = "Density",
       title = "Distribution of log(Payroll) (Histogram & Density)") +
  theme_minimal()
```



For the payroll distribution plots, we can see that: 1.Raw payroll (left) — The histogram with a density overlay is heavily right-skewed: most teams cluster at low–mid payrolls, with a long right tail of a few top spenders. In such skewed data, the mean typically exceeds the median, and outliers can dominate raw-scale comparisons. 2.Using density scale means bar heights reflect probability density (areas integrate to 1), so the histogram and the smooth curve are directly comparable in shape, not in counts. 3.Log payroll (right) — After taking logs, the long tail is compressed, producing a distribution that is more concentrated and closer to bell-shaped. This makes patterns easier to see and supports more stable, interpretable linear/GLM modeling. 4.The contrast between the two panels motivates using log(payroll) in subsequent plots and regressions to reduce skewness and improve comparability across teams (and seasons). MLB payrolls are strongly right-skewed; transforming to log(payroll) yields cleaner visuals and more reliable inference.

8 Executive Summary

Objective. Assess how MLB team payroll relates to on-field performance (winning percentage and playoff qualification) across seasons, and clarify whether the relationship is linear or exhibits thresholds.

Data & Preparation. Team–season panel containing total payroll allocations, active-roster payroll, wins/losses, and postseason results. Variables were standardized and parsed to numeric; key derived fields include win_pct = wins/(wins+losses), a binary playoffs_bin (1 = made playoffs, 0 = did not), and log_payroll = log(payroll) to mitigate right-skew.

Key Findings.

1.Payroll vs. Winning % (scatter + OLS, by season). The fitted OLS line is consistently upward within seasons: higher payrolls are associated with higher winning percentages. Vertical dispersion is large at most payroll levels: payroll helps, but it does not fully determine outcomes. Counterexamples (low-payroll $>$ .500; high-payroll $<$ .500) imply moderating factors such as roster construction, injuries, coaching, and schedule strength.

2.Winning % by Payroll Quartile (boxplots). Medians and upper whiskers rise from Q1 to Q4, indicating better average performance for higher-payroll groups. Overlap across quartiles remains meaningful, reinforcing that payroll is a strong correlate—not a guarantee.

3.Playoff Qualification Rates by Payroll Quartile (proportion bars).

Higher quartiles exhibit higher playoff rates, though differences are not overwhelming; low-payroll playoff teams persist. Interpreted together with (1)–(2), this supports a positive association while underscoring significant variance.

4.Playoff Probability vs. Payroll (LOESS & logistic views).

The LOESS curve reveals a positive but nonlinear pattern: playoff odds rise with payroll and steepen near the top-spending range. A simple logistic regression (controlling for year and average age) increases more gently across most of the range and can understate the sharp uptick among elite payrolls. Net: the payoff to spending is concentrated at the very top; mid-range marginal effects are modest after controls.

5.Distribution of Payrolls (histogram & density). Raw payrolls are heavily right-skewed with a long tail of top spenders; means exceed medians and outliers dominate raw-scale comparisons.

Using log(payroll) compresses the tail, yielding a more concentrated, near-bell shape and more stable, interpretable modeling.

This motivates expressing scatterplots/regressions in log-payroll space.

Conclusions. The relationship between payroll and performance is reliably positive but meaningfully dispersed and nonlinear. Spending helps, yet the largest gains accrue only at the very top of the spending distribution. At similar payrolls, outcomes vary widely—execution quality (scouting, development, health, in-game strategy, coaching) remains pivotal.

#————————————————————————————# # First Analysis # #————————————————————————————#

Earlier in the visualization section, we included a plot called “Payroll vs playoff probability (loess or logistic preview)”. That plot was mainly part of our initial exploration, just to get a basic sense of how payroll might relate to the chance of making the playoffs. It wasn’t meant to be a final or formal model — more like a first look at the data. In the following analysis, we revisit the same question using a cleaner dataset structure and more appropriate modeling choices, so the next few plots represent our formal version of the analysis.

9 OLS

```
mlb_reg <- mlb %>%
  filter(!is.na(wins), !is.na(payroll), !is.na(average_age))
```

```r
ols_model <- lm(wins ~ log_payroll + average_age + factor(year), data = mlb_reg)

summary(ols_model)
```

```
##
## Call:
## lm(formula = wins ~ log_payroll + average_age + factor(year),
##     data = mlb_reg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.1166  -6.6251  -0.4516   6.1612  29.5308
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -143.48156   22.88424  -6.270 9.29e-10 ***
## log_payroll         1.42783    1.56197   0.914    0.361
## average_age         7.10212    0.65375  10.864  < 2e-16 ***
## factor(year)2012    0.72949    2.59856   0.281    0.779
## factor(year)2013    0.01174    2.59525   0.005    0.996
## factor(year)2014    0.41789    2.61211   0.160    0.873
## factor(year)2015    1.01034    2.63751   0.383    0.702
## factor(year)2016    0.72967    2.64524   0.276    0.783
## factor(year)2017    0.91337    2.66678   0.343    0.732
## factor(year)2018    1.65440    2.67935   0.617    0.537
## factor(year)2019    2.60268    2.69274   0.967    0.334
## factor(year)2020  -51.59481    2.73185 -18.886  < 2e-16 ***
## factor(year)2021   -2.47542    2.61021  -0.948    0.344
## factor(year)2022   -1.58216    2.64461  -0.598    0.550
## factor(year)2023   -1.31443    2.68216  -0.490    0.624
## factor(year)2024   -2.36383    2.68329  -0.881    0.379
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.04 on 404 degrees of freedom
## Multiple R-squared:  0.6965, Adjusted R-squared:  0.6852
## F-statistic:  61.8 on 15 and 404 DF,  p-value: < 2.2e-16
```

```r
tidy_ols <- tidy(ols_model)
tidy_ols
```

```
## # A tibble: 16 x 5
##    term              estimate std.error statistic  p.value
##    <chr>                <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)      -143.        22.9    -6.27    9.29e-10
##  2 log_payroll         1.43       1.56    0.914   3.61e- 1
```

```
##  3 average_age          7.10       0.654  10.9     2.72e-24
##  4 factor(year)2012     0.729      2.60    0.281   7.79e- 1
##  5 factor(year)2013     0.0117     2.60    0.00452 9.96e- 1
##  6 factor(year)2014     0.418      2.61    0.160   8.73e- 1
##  7 factor(year)2015     1.01       2.64    0.383   7.02e- 1
##  8 factor(year)2016     0.730      2.65    0.276   7.83e- 1
##  9 factor(year)2017     0.913      2.67    0.343   7.32e- 1
## 10 factor(year)2018     1.65       2.68    0.617   5.37e- 1
## 11 factor(year)2019     2.60       2.69    0.967   3.34e- 1
## 12 factor(year)2020   -51.6        2.73  -18.9     1.76e-57
## 13 factor(year)2021    -2.48       2.61   -0.948   3.44e- 1
## 14 factor(year)2022    -1.58       2.64   -0.598   5.50e- 1
## 15 factor(year)2023    -1.31       2.68   -0.490   6.24e- 1
## 16 factor(year)2024    -2.36       2.68   -0.881   3.79e- 1
```

OLS Regression Results The OLS regression examines how team payroll influences regular-season wins while controlling for roster age and year fixed effects. The model achieves a relatively strong fit, with an R-squared of 0.696, indicating that approximately 70% of the variation in team wins can be explained by payroll, roster age, and season effects.

1. Effect of Payroll The coefficient on log_payroll is positive (= 1.43) but not statistically significant (p = 0.361). Interpretation: After controlling for roster age and season effects, differences in payroll do not significantly predict differences in win totals. A positive but nonsignificant effect suggests that payroll may still matter directionally, but once we account for league-wide factors and age, the effect is too weak or too variable to be distinguished from zero. This also aligns with modern MLB trends where player development, analytics, and injuries can overshadow raw payroll size.

2. Effect of Average Roster Age The coefficient on average_age is positive and highly significant (= 7.10, $p < 0.001$). Interpretation: A one-year increase in average roster age is associated with about 7 additional wins, holding payroll and year constant. This is a large and meaningful effect, suggesting that: Older rosters tend to be more experienced or stable Younger teams may be rebuilding or rely heavily on developing players Age may proxy for accumulated MLB service time, star player stability, or fewer rookie innings This is the strongest predictor in the model.

3. Year Fixed Effects Most year coefficients are statistically insignificant, meaning most seasons do not differ drastically from the baseline year after accounting for payroll and age. The major exception is: 2020 has a very large and significant negative coefficient (= -51.59, $p < 0.001$) Interpretation: MLB's shortened 2020 season (60 games due to COVID-19) produced systematically lower win totals, which the model correctly captures. Other years show small, insignificant effects—this indicates that typical season-to-season changes do not meaningfully alter win totals after controlling for payroll and age.

4. Model Fit $R^2 = 0.6965$ Adjusted $R^2 = 0.6852$ This indicates the model explains a large portion of win variation, driven mainly by roster age and year (especially 2020) rather than payroll.

Overall Interpretation The OLS results suggest: Payroll does not significantly predict team wins once other factors are controlled. Average roster age is the strongest and most significant predictor of

wins, indicating that experience and player maturity are crucial for team performance. Year effects matter mainly for unusual seasons, such as the COVID-shortened 2020 season. Team success is influenced more by experience, development cycles, and season contexts, rather than raw financial spending. These results challenge the assumption that higher payroll always leads to better regular-season performance.

10 Payroll vs. Playoff Probability — Logistic Regression

```
mlb_logit <- mlb |>
  mutate(
    average_age = as.numeric(average_age)
  ) |>
  filter(!is.na(playoffs_bin),
         !is.na(log_payroll),
         !is.na(average_age),
         !is.na(year))
nrow(mlb_logit)
```

```
## [1] 420
```

This step prepares the dataset for logistic regression by converting roster age to a numeric variable and removing rows with missing values in key fields. This ensures that the model is estimated using complete and consistent information across payroll, age, year, and playoff outcomes.

11 Logistic regression model + Odds Ratios

```
fit_logit <- glm(playoffs_bin ~log_payroll + average_age + factor(year),
                 data = mlb_logit, family = binomial)

coefs_or <- tidy(fit_logit, conf.int = TRUE, conf.level = 0.95, exponentiate = TRUE)
kable(coefs_or, digits = 3, caption = "Logistic regression (odds ratios)")
```

Table 1: Logistic regression (odds ratios)

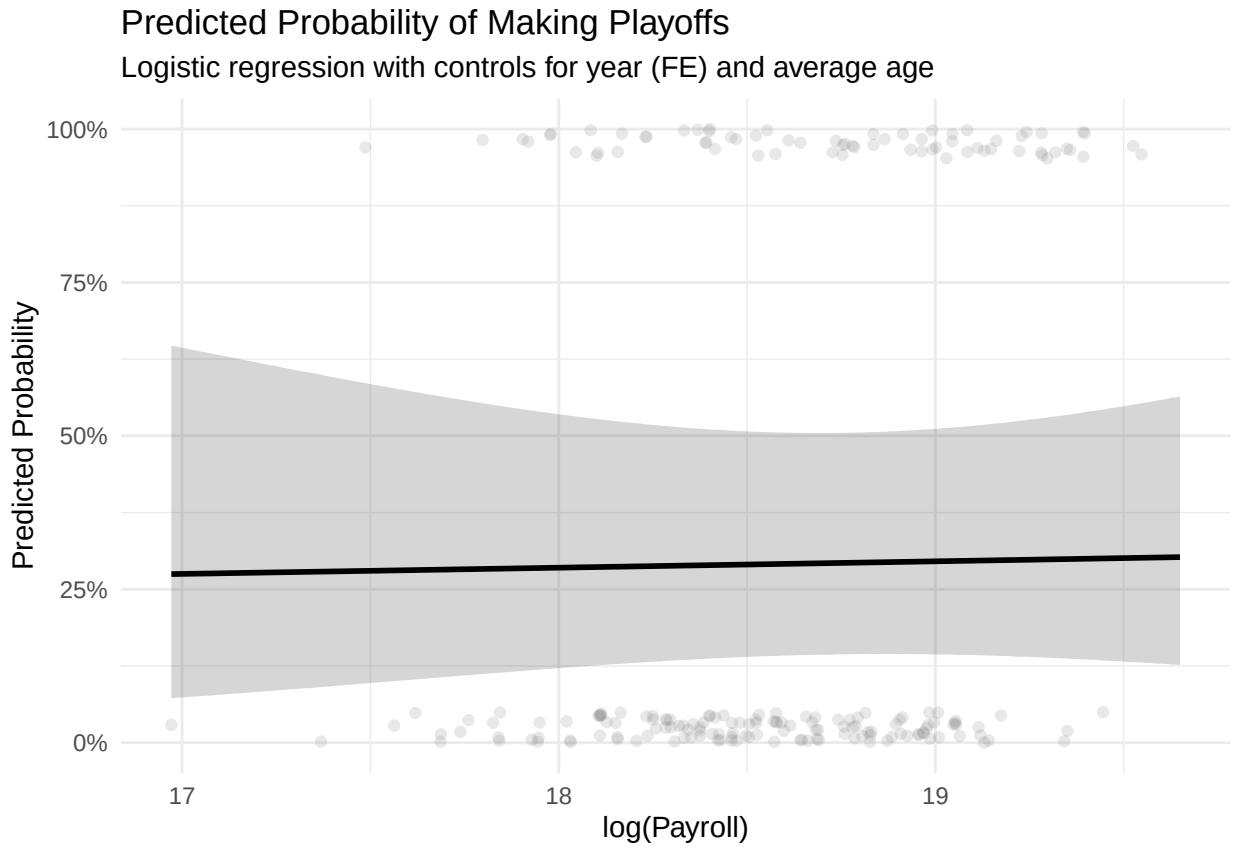
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.000	5.932	-5.851	0.000	0.000	0.000
log_payroll	1.052	0.365	0.138	0.890	0.514	2.159
average_age	3.211	0.173	6.744	0.000	2.312	4.562
factor(year)2012	1.678	0.626	0.826	0.409	0.492	5.850
factor(year)2013	1.411	0.620	0.555	0.579	0.419	4.860
factor(year)2014	1.651	0.622	0.806	0.421	0.489	5.715

term	estimate	std.error	statistic	p.value	conf.low	conf.high
factor(year)2015	1.876	0.627	1.004	0.315	0.552	6.555
factor(year)2016	1.774	0.631	0.909	0.363	0.517	6.240
factor(year)2017	1.735	0.629	0.875	0.382	0.508	6.092
factor(year)2018	1.940	0.650	1.020	0.308	0.547	7.098
factor(year)2019	2.303	0.643	1.299	0.194	0.658	8.307
factor(year)2020	1.723	0.641	0.848	0.396	0.495	6.217
factor(year)2021	0.990	0.629	-0.016	0.987	0.288	3.454
factor(year)2022	1.590	0.633	0.733	0.464	0.463	5.616
factor(year)2023	1.165	0.648	0.235	0.814	0.326	4.216
factor(year)2024	1.433	0.658	0.546	0.585	0.396	5.313

We fit a logistic regression model to estimate how payroll affects the likelihood of making the playoffs, controlling for team age and year fixed effects. The table shows odds ratios, which indicate how much the odds of reaching the postseason change for each unit increase in log payroll. # Logistic prediction curve + 95% confidence interval

```
#    vs        year=   age=
logp_seq <- seq(min(mlb_logit$log_payroll, na.rm = TRUE),
                max(mlb_logit$log_payroll, na.rm = TRUE), length.out = 200)

newdat <- tibble(
  log_payroll = logp_seq,
  average_age = mean(mlb_logit$average_age, na.rm = TRUE),
  year        = max(mlb_logit$year, na.rm = TRUE)
)

#  SE
pred_link <- predict(fit_logit, newdata = newdat, type = "link", se.fit = TRUE)
newdat <- newdat |>
  mutate(
    fit  = plogis(pred_link$fit),
    lwr  = plogis(pred_link$fit - 1.96 * pred_link$se.fit),
    upr  = plogis(pred_link$fit + 1.96 * pred_link$se.fit)
  )

ggplot() +

  geom_jitter(data = mlb_logit,
              aes(x = log_payroll, y = playoffs_bin),
              height = 0.05, alpha = 0.15, color = "grey40") +
  #      + 95% CI
  geom_ribbon(data = newdat, aes(x = log_payroll, ymin = lwr, ymax = upr),
              alpha = 0.2) +
  geom_line(data = newdat, aes(x = log_payroll, y = fit), linewidth = 1) +
  scale_y_continuous(labels = scales::percent, limits = c(0, 1)) +
```

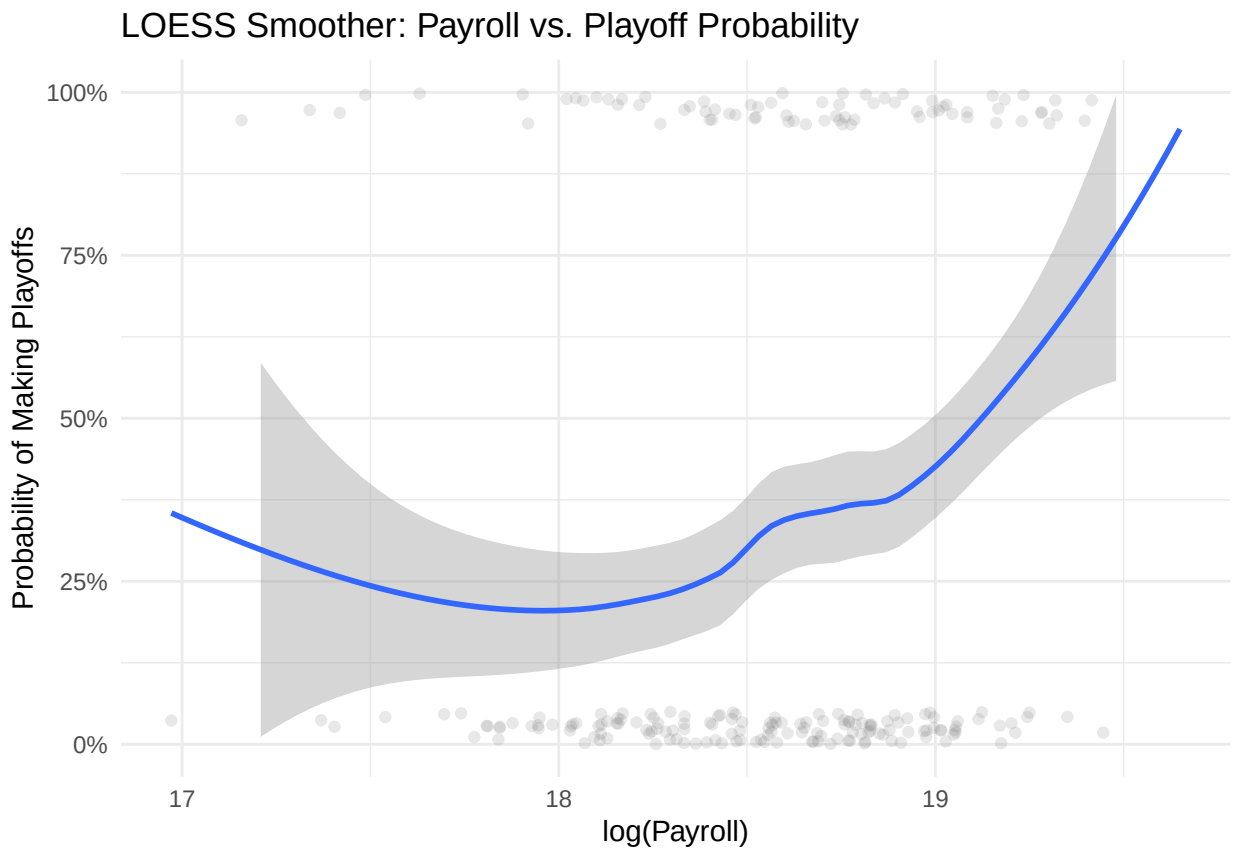
```
  labs(title = "Predicted Probability of Making Playoffs",
       subtitle = "Logistic regression with controls for year (FE) and average age",
       x = "log(Payroll)", y = "Predicted Probability") +
  theme_minimal()
```



This plot visualizes the predicted probability of making the playoffs across different payroll levels, based on a logistic regression model that includes log-transformed payroll as the main predictor and controls for team average age and year fixed effects. The black line shows the estimated probability, while the shaded band gives the 95% confidence interval. The gray jittered points represent actual team-season outcomes (0 = missed playoffs; 1 = made playoffs), helping us see how the model-based curve aligns with the observed data.

12 LOESS smoother: non-parametric preview

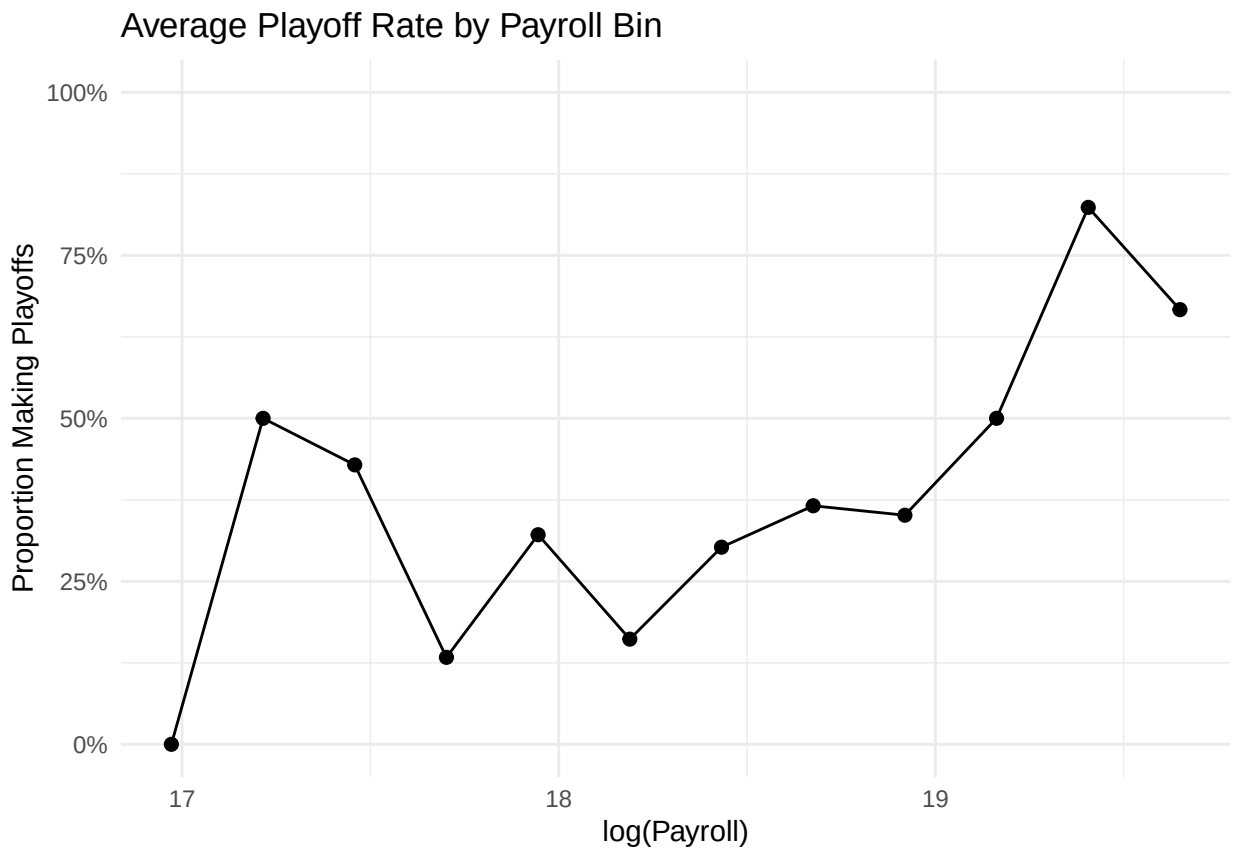
```
ggplot(mlb_logit, aes(x = log_payroll, y = playoffs_bin)) +
  geom_jitter(height = 0.05, alpha = 0.15, color = "grey40") +
  geom_smooth(method = "loess") +
  scale_y_continuous(labels = percent, limits = c(0, 1)) +
  labs(title = "LOESS Smoother: Payroll vs. Playoff Probability",
       x = "log(Payroll)", y = "Probability of Making Playoffs") +
  theme_minimal()
```



This LOESS curve provides a flexible, non-parametric view of the payroll–playoff relationship. It does not rely on any statistical model assumptions and instead captures the trend directly from the data. The smoother suggests a slight upward pattern, supporting the idea that higher payroll may help, although the overall relationship appears fairly weak or noisy.

13 Binned playoff rates (non-parametric)

```
#
ggplot(mlb_logit, aes(x = log_payroll, y = playoffs_bin)) +
  stat_summary_bin(fun = mean, bins = 12, geom = "point", size = 2) +
  stat_summary_bin(fun = mean, bins = 12, geom = "line") +
  scale_y_continuous(labels = scales::percent, limits = c(0,1)) +
  labs(title = "Average Playoff Rate by Payroll Bin",
       x = "log(Payroll)", y = "Proportion Making Playoffs") +
  theme_minimal()
```



This figure shows the empirical proportion of teams that made the playoffs within evenly spaced bins of log payroll. Each point represents the average playoff rate among teams falling into that payroll range, and the line connects these bin-level averages to highlight the trend without imposing a functional form.

14 Predicted curves by year (optional)

```
#
yr_keep <- sort(unique(mlb_logit$year))
grid <- expand.grid(
  log_payroll = logp_seq,
  average_age = mean(mlb_logit$average_age, na.rm = TRUE),
  year        = yr_keep
)
pred_by_year <- predict(fit_logit, newdata = grid, type = "response")
grid$pred <- pred_by_year

ggplot(grid, aes(x = log_payroll, y = pred, color = factor(year))) +
  geom_line() +
  scale_y_continuous(labels = scales::percent, limits = c(0,1)) +
  labs(title = "Payroll → Playoff Probability by Year",
```

```
      x = "log(Payroll)", y = "Predicted Probability", color = "Year") +
  theme_minimal()
```

This plot shows the logistic regression predictions separately for each season. While the slopes remain modest in all years, the vertical shifts between lines indicate that some seasons were generally easier or harder for teams to reach the postseason. This helps contextualize payroll effects within year-to-year league differences.

While logistic regression with linear log-payroll shows only a weak association, the non-parametric binned plot reveals a clear non-linear pattern: playoff probability rises sharply only among the highest-spending teams. Together, these results highlight that payroll does matter, but its impact is concentrated among top spenders rather than increasing steadily across the entire distribution.