

# Task 1: Red Sox Ticket Price Analysis

Yuexin (Joy) Wang

2025-11-21

## Contents

1. Executive Summary . . . . .	1
2. Data Loading and Preparation . . . . .	1
3. Price Dynamics Over Time (Full Sample) . . . . .	3
4. Addressing Composition Bias: Focusing on a Single Seat Type . . . . .	4
5. Formalizing the Trend: A Simple Regression . . . . .	6
6. Discussion and Conclusion . . . . .	7

## 1. Executive Summary

This report uses secondary-market transaction data for Boston Red Sox home games from 2009–2012 to study how ticket prices evolve as the game date approaches.

Two main patterns emerge. First, across all four seasons, average ticket prices tend to decline as the game date gets closer, especially within the last 10–15 days before first pitch. Buyers who wait until the final week typically pay less than those purchasing 2–4 weeks in advance.

Second, the magnitude of this “last-minute discount” increases over time. In 2009, tickets bought in the last seven days are only modestly cheaper than those purchased 8–30 days before the game. By 2012, however, the average last-week discount exceeds \$30 relative to earlier purchases. This suggests that sellers become more aggressive in cutting prices close to game day as the resale market matures.

Because overall averages can be driven by changes in the mix of seats sold at different times, I also examine price dynamics within a single, common seat type. Within this seat type, prices are more stable over time, and the decline near game day is smaller. This indicates that part of the apparent discounting reflects composition effects—higher-priced seats tend to sell earlier—rather than pure markdowns for identical seats.

## 2. Data Loading and Preparation

I begin by loading all four yearly files from the Red\_Sox directory and stacking them into a single master dataset. For each file I recover the season year from the filename and create a factor version for plotting.

```
# List all yearly files such as "red_sox_2009.csv"

file_list <- list.files(
  path    = data_path,
  pattern = "^red_sox_\\d{4}\\..csv$",
  full.names = TRUE
)

# Read and combine, attach year information

raw_df <- purrr::map_dfr(file_list, function(file_path) {
```

```

year_from_file <- stringr::str_extract(basename(file_path), "\\d{4}") |> as.numeric()

readr::read_csv(file_path, show_col_types = FALSE) |>
mutate(
  year      = year_from_file,
  year_factor = factor(year)
)
})

# Basic data-quality checks: missing values and duplicates

na_summary <- raw_df %>%
summarise(across(everything(), ~ sum(is.na(.))))

na_summary

## # A tibble: 1 x 13
##   transaction_date sectiontype price_per_ticket number_of_tickets gamemonth
##             <int>         <int>           <int>           <int>      <int>
## 1             0             0             0             0            0
## # i 8 more variables: team <int>, day_game <int>, weekend_game <int>,
## #   gamedate <int>, logprice <int>, days_from_transaction_until_game <int>,
## #   year <int>, year_factor <int>

total_n    <- nrow(raw_df)
distinct_n <- raw_df %>% distinct() %>% nrow()

cat("Total rows:", total_n, "\nDistinct rows:", distinct_n, "\n")

## Total rows: 453171
## Distinct rows: 414145

# Keep clearly valid transactions

clean_df <- raw_df %>%
filter(
  price_per_ticket > 0,
  days_from_transaction_until_game >= 0
)

# Basic sample size by year

clean_df %>%
count(year) %>%
arrange(year)

## # A tibble: 4 x 2
##   year      n
##   <dbl> <int>
## 1  2009 105673
## 2  2010 118895
## 3  2011 152525
## 4  2012  76078

```

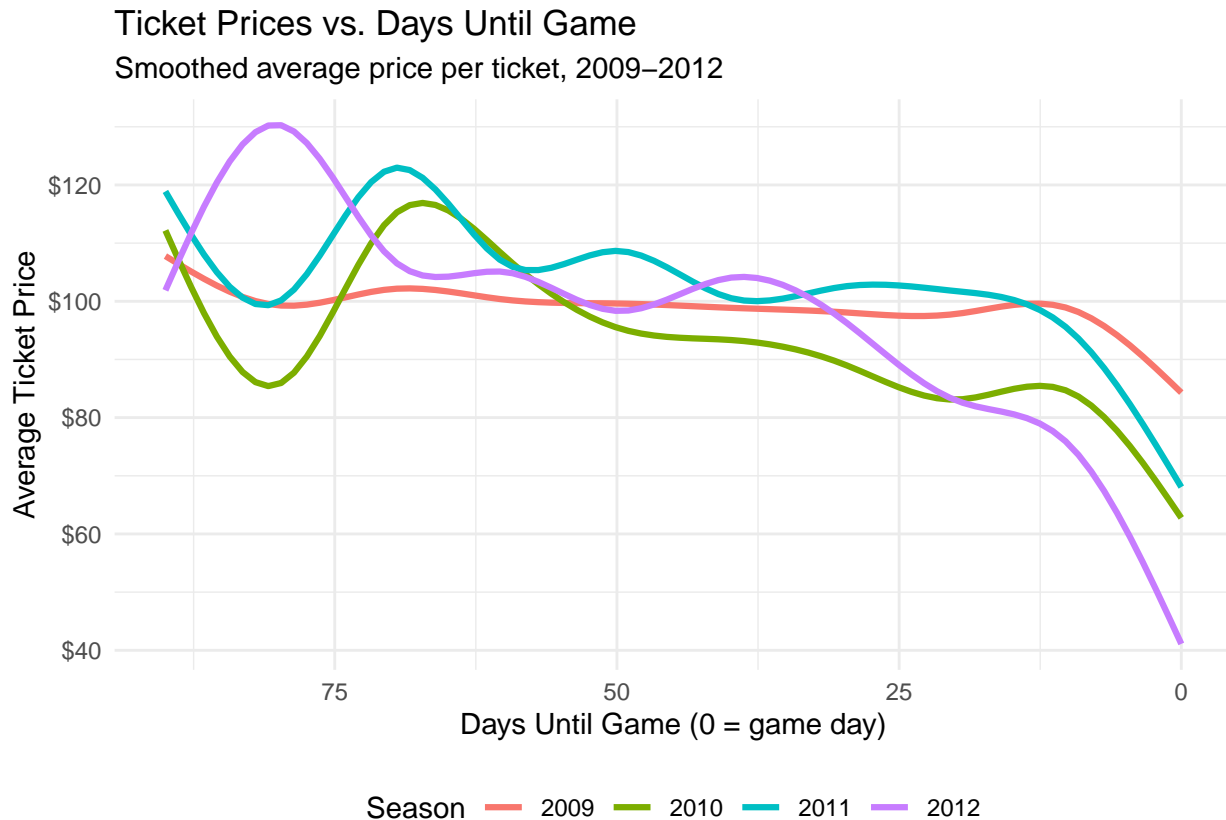
The raw files already contain a per-ticket price (`price_per_ticket`) and the number of days between the transaction date and the game date (`days_from_transaction_until_game`), so no additional construction is

needed for these key variables. I drop observations with non-positive prices and negative days-until-game, which likely reflect data-entry errors or special transactions.

### 3. Price Dynamics Over Time (Full Sample)

To study how prices move as the game date approaches, I plot average prices against the number of days until the game, allowing for a flexible smooth relationship by year.

```
ggplot(clean_df,
aes(x = days_from_transaction_until_game,
y = price_per_ticket,
color = year_factor)) +
geom_smooth(se = FALSE, size = 1.1) +
scale_x_reverse(limits = c(90, 0)) + # show last 90 days, with game day on the right
scale_y_continuous(labels = dollar_format()) +
labs(
title = "Ticket Prices vs. Days Until Game",
subtitle = "Smoothed average price per ticket, 2009–2012",
x = "Days Until Game (0 = game day)",
y = "Average Ticket Price",
color = "Season"
) +
theme_minimal() +
theme(legend.position = "bottom")
```



Interpretation.

Across seasons, the smoothed profiles generally slope downward as the game date approaches, indicating lower average prices closer to game day. Visually, the 2011–2012 curves appear somewhat flatter than 2009, especially at longer horizons. However, when I focus on the last 30 days and explicitly compare the final

week to days 8–30 (Table 1), the average last-week discount is actually larger in later seasons. In other words, prices do not become more stable over time; instead, last-minute markdowns become more pronounced, even though the overall shape of the curves looks relatively smooth.

To quantify this, I compare average prices in the last week before the game to prices 8–30 days out.

```
by_year_period <- clean_df %>%
  filter(days_from_transaction_until_game <= 30) %>%
  mutate(
    period = if_else(days_from_transaction_until_game <= 7,
      "last_7_days", "days_8_30")
  ) %>%
  group_by(year, period) %>%
  summarise(
    mean_price = mean(price_per_ticket, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  tidyr::pivot_wider(
    names_from = period,
    values_from = mean_price
  ) %>%
  mutate(
    diff_last7_vs_8_30 = last_7_days - days_8_30
  )

by_year_period %>%
  kable(
    col.names = c("Season", "Price (8-30 Days Out)", "Price (Last 7 Days)", "Difference ($)"),
    digits = 2,
    caption = "Comparison of Average Ticket Prices (Early vs. Late Purchase)",
    align = "c"
  )
```

Table 1: Comparison of Average Ticket Prices (Early vs. Late Purchase)

Season	Price (8-30 Days Out)	Price (Last 7 Days)	Difference (\$)
2009	97.98	89.54	-8.45
2010	83.67	70.23	-13.44
2011	98.74	75.23	-23.51
2012	82.26	49.76	-32.50

For example, in 2009 the average price in the last seven days was only \$8.45 lower than 8–30 days before the game, while in 2012 the difference increased to \$32.50. This supports the graphical impression that late-game discounting became progressively stronger over the sample period, with sellers offering steeper discounts in later seasons compared to 2009.

#### 4. Addressing Composition Bias: Focusing on a Single Seat Type

A potential concern with the preceding analysis is composition bias. High-end seats (e.g., club boxes) may tend to sell earlier, while cheaper seats sell later. Even if each seat type has constant prices over time, the average price could fall simply because the mix of seats changes.

To address this, I restrict attention to a single, relatively common seat type. The table below shows the most frequently traded sections:

```
section_counts <- clean_df %>%
count(sectiontype, sort = TRUE)

head(section_counts, 10)
```

```
## # A tibble: 10 x 2
##   sectiontype      n
##   <chr>         <int>
## 1 LowerBleachers 103256
## 2 IFGS           79634
## 3 RFGS           63931
## 4 RFFieldBox     46208
## 5 UpperBleachers 42625
## 6 LogeBox        36210
## 7 FieldBox       28916
## 8 SRO            12348
## 9 RFR00FBOX      7254
## 10 FamilyGS      6678
```

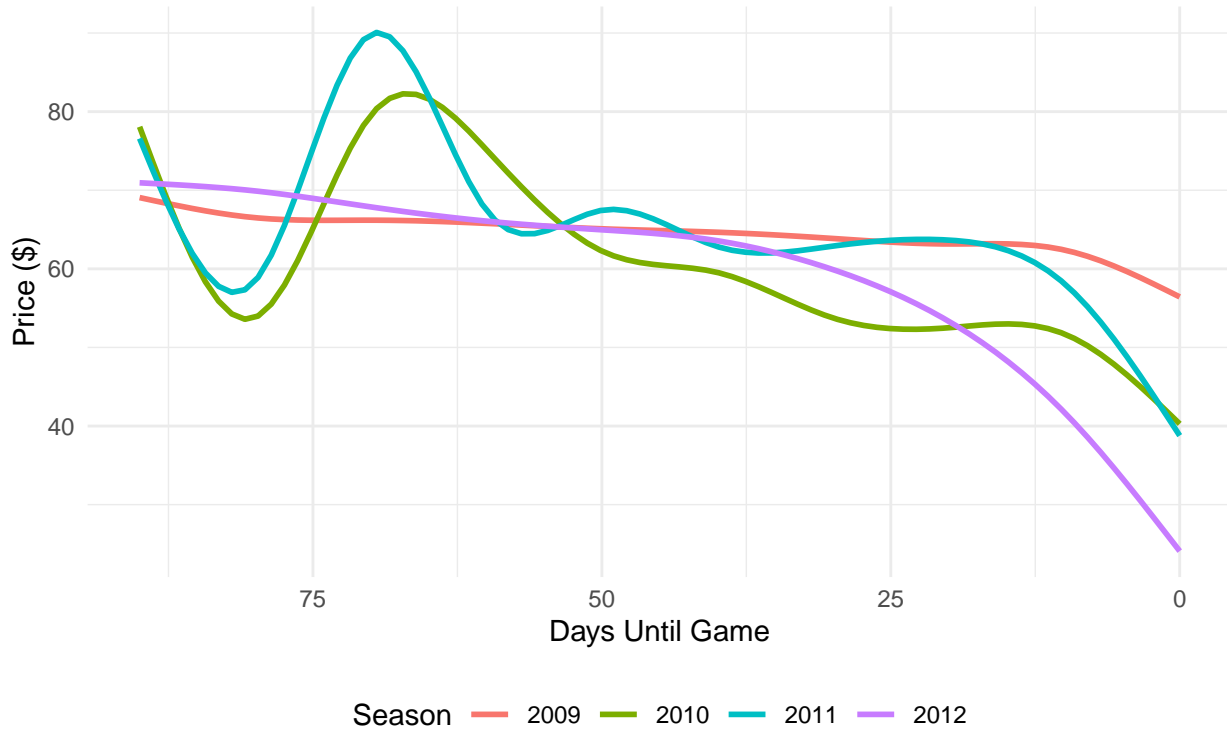
One common category across seasons is RFGS (Right Field Grandstand). I focus on this section to hold seat quality roughly fixed and re-examine the price-time profile.

```
target_section <- "RFGS"

clean_df %>%
filter(sectiontype == target_section) %>%
ggplot(aes(x = days_from_transaction_until_game,
y = price_per_ticket,
color = year_factor)) +
geom_smooth(se = FALSE, span = 0.8) +
scale_x_reverse(limits = c(90, 0)) +
labs(
title = "Price Dynamics Within a Single Seat Type (RFGS)",
subtitle = "Holding seat type fixed reduces composition concerns",
x = "Days Until Game",
y = "Price ($)",
color = "Season"
) +
theme_minimal() +
theme(legend.position = "bottom")
```

## Price Dynamics Within a Single Seat Type (RFGS)

Holding seat type fixed reduces composition concerns



Interpretation.

Within RFGS, the time pattern becomes flatter, especially in later years. Prices still tend to be slightly lower in the last few days before the game, but the steep drop seen in the full-sample averages is muted. This suggests that part of the overall decline is due to changes in the mix of seats sold over time, not just true discounting for identical seats.

## 5. Formalizing the Trend: A Simple Regression

```
m1 <- feols(log(price_per_ticket) ~ days_from_transaction_until_game | year_factor,
            data = clean_df)

m2 <- feols(log(price_per_ticket) ~ days_from_transaction_until_game | year_factor
            + sectiontype, data = clean_df)

modelsummary(
  list("Baseline" = m1, "With Section FE" = m2),
  stars = TRUE,
  gof_map = c("nobs", "r.squared"),
  coef_map = c("days_from_transaction_until_game" = "Days Until Game"),
  title = "Regression of Log Ticket Prices on Timing"
)
```

Interpretation.

Table 2 presents a simple regression analysis. Even after controlling for unobserved heterogeneity across seat sections (Column 2), the coefficient on Days Until Game remains positive and statistically significant. The point estimate of about 0.004 implies that tickets bought 10 days earlier are roughly 4 percent more expensive, holding season and seat type fixed.

Table 2: Regression of Log Ticket Prices on Timing

	Baseline	With Section FE
Days Until Game	0.004*** (0.000)	0.004*** (0.000)
Num.Obs.	453 171	453 171
R2	0.060	0.506

+ p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

## 6. Discussion and Conclusion

This short analysis yields several takeaways:

1. **Moderate last-minute discounting in early seasons.** In 2009, both the smoothed curves and the early-versus-late summary statistics show that buyers who wait until the last week before the game typically pay less than those buying 2–4 weeks in advance.
2. **Stronger last-minute discounting in later seasons.** The early-versus-late comparison in Table 1 shows that the average price gap between tickets bought 8–30 days ahead and in the last week widens over time, reaching more than \$30 in 2012. Rather than disappearing, last-minute markdowns become more aggressive in later seasons.

When I restrict the analysis to a single, common seat type (RFGS), the downward slope of prices over time becomes noticeably weaker. This indicates that part of the apparent “markdown” over time comes from high-priced seats being sold earlier and cheaper seats remaining closer to game day.

### Limitations.

The analysis is descriptive; I do not attempt to estimate a structural demand model or control for all confounders.

Several important factors are omitted, such as opponent quality, day of the week, weather, or whether the game has playoff implications. These could affect both pricing and purchase timing.

Finally, the results are specific to Red Sox home games in 2009–2012 and may not generalize to other teams, stadiums, or time periods.

Given the time constraint of the assignment, I focus on clear, reproducible descriptive patterns. With more time, one could extend the analysis by adding richer controls, estimating regressions for log prices on days-to-game and game characteristics, or exploring heterogeneity across weekday vs. weekend games and different opponents.