Kellogg Finance: Coding Task

William Co*
Department of Economics, University of British Columbia

January 20, 2025

Abstract

This is a data task submitted for a predoctoral application.

*Thank you for the opportunity to complete this data task for my predoctoral application. I appreciate your consideration, and I look forward to meeting and contributing.

1 Introduction

A copy of writing sample can be found here link and will be submitted as well. (Co n.d.)

A copy of the time sheet will be submitted

Code and output data for task 2 will be submitted as well.

2 Task 1: Red Sox Ticket Prices

How do the prices consumers pay for tickets change as the game date approaches (i.e., as the number of days between transaction date and game date declines)? How does this dynamic pattern change across years?

To address the question, we begin by running a preliminary regression to gain a better understanding of our data. This regression includes all available control variables and accounts for the number of tickets purchased. Considering the potential for bulk discounts associated with purchasing multiple tickets, we aim to capture variations in pricing that may occur in different contexts, where such discounts might or might not be offered.

1. **Model with `number_of_tickets`**:

$$\text{price\_per\_ticket} = \beta_0 + \beta_1 \cdot \text{days\_until\_game} + \beta_2 \cdot \text{controls} + \beta_3 \cdot \text{num\_tickets} + \epsilon$$

2. **Model without `number_of_tickets`**:

$$\text{price\_per\_ticket} = \beta_0 + \beta_1 \cdot \text{days\_until\_game} + \beta_2 \cdot \text{controls} + \epsilon$$

3. **Model with `number_of_tickets` using log price**:

$$\log(\text{price\_per\_ticket}) = \beta_0 + \beta_1 \cdot \text{days\_until\_game} + \beta_2 \cdot \text{controls} + \beta_3 \cdot \text{num\_tickets} + \epsilon$$

4. **Model without `number_of_tickets` using log price**:

$$\log(\text{price\_per\_ticket}) = \beta_0 + \beta_1 \cdot \text{days\_until\_game} + \beta_2 \cdot \text{controls} + \epsilon$$

Where:

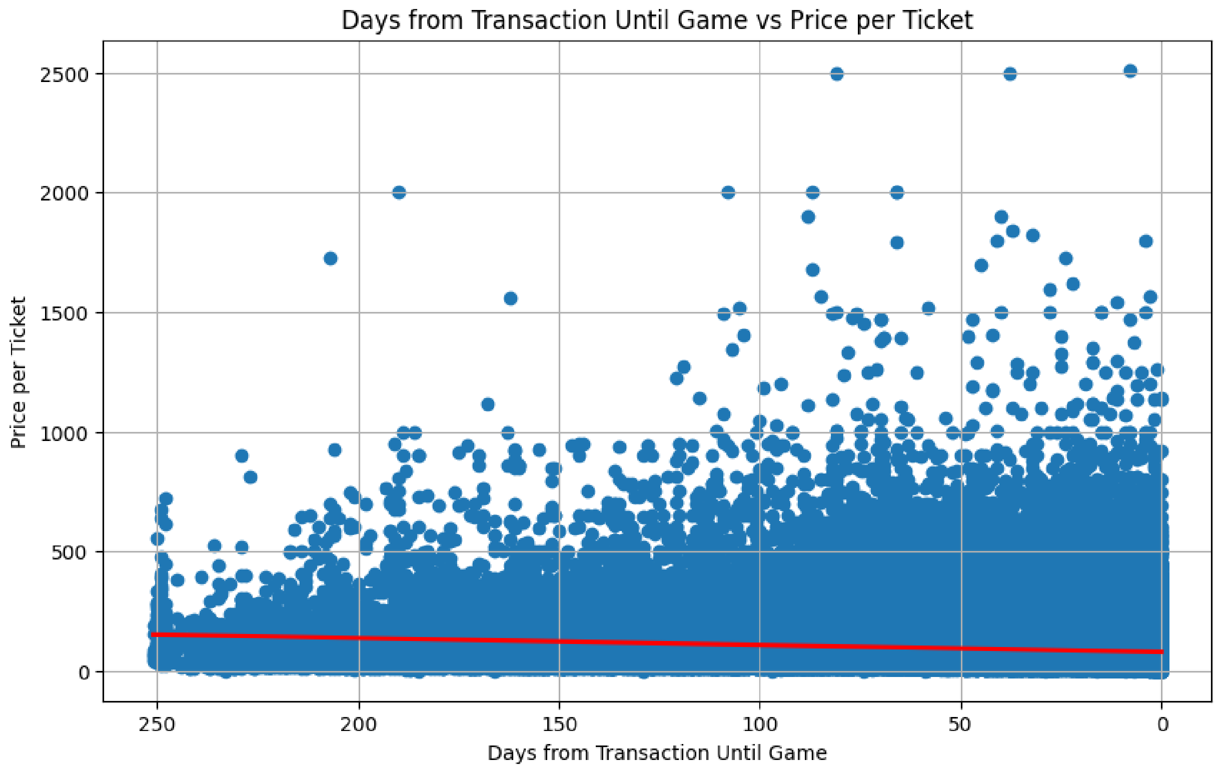
$$\text{controls} = \text{day\_game, weekend\_game, sectiontype, gamemonth, team, year}$$

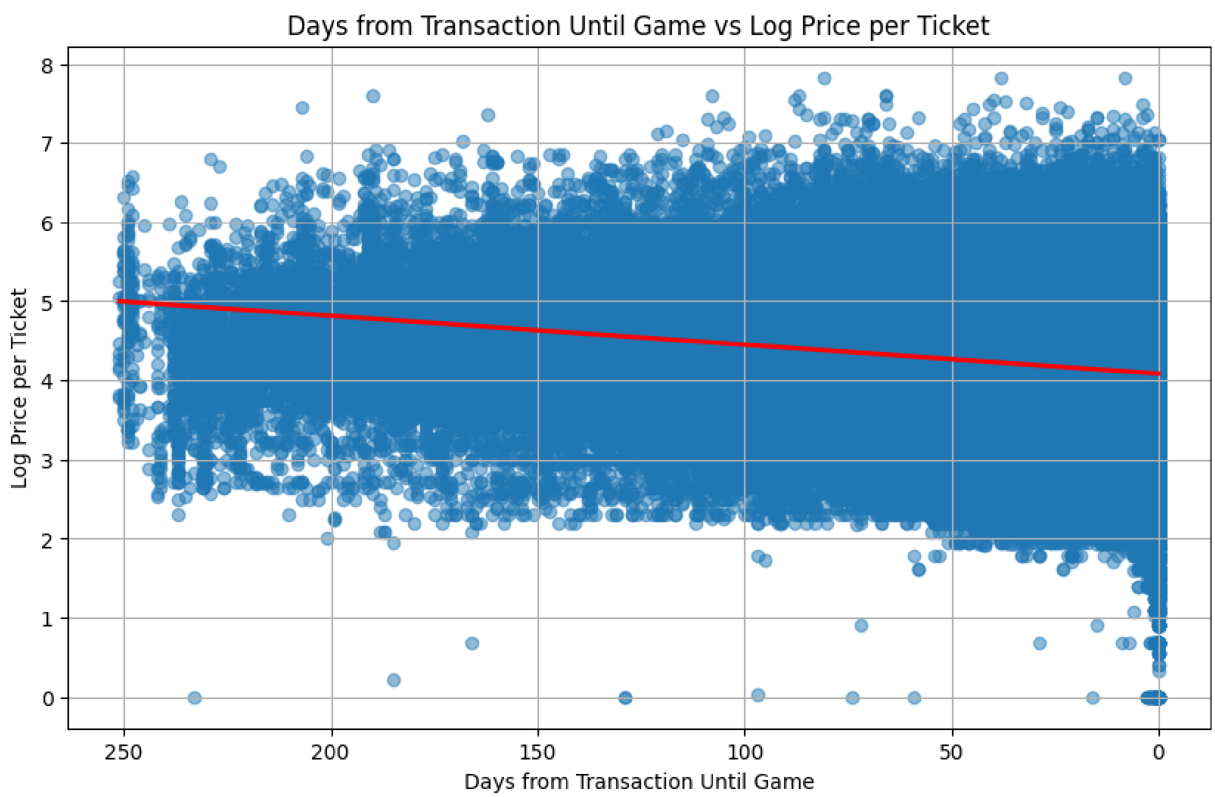
- β_0 is the intercept.
- β_1, β_3 are the coefficients for the respective predictor variables.
- β_2 is a vector of coefficients for control variables.
- ϵ is the error term.

	Dependent variable:			
	price_per_ticket		logprice	
	(1)	(2)	(3)	(4)
days_from_transaction_until_game	0.227^{***}	0.232^{***}	0.003^{***}	0.003^{***}
	(0.002)	(0.002)	(0.00002)	(0.00002)
number_of_tickets	1.865^{***}		0.022^{***}	
	(0.046)		(0.0004)	
Constant	416.553^{***}	420.454^{***}	5.833^{***}	5.880^{***}
	(1.335)	(1.334)	(0.011)	(0.011)
Day Game Controls	Yes	Yes	Yes	Yes
Weekend Game Controls	Yes	Yes	Yes	Yes
Section Controls	Yes	Yes	Yes	Yes
Game Month Controls	Yes	Yes	Yes	Yes
Team Controls	Yes	Yes	Yes	Yes
Year Controls	Yes	Yes	Yes	Yes
Number of Tickets Controls	Yes	No	Yes	No
Observations	452,936	452,936	452,936	452,936
R^2	0.657	0.656	0.725	0.723
Adjusted R^2	0.657	0.656	0.725	0.723
Residual Std. Error	48.028 (df = 452879)	48.116 (df = 452880)	0.404 (df = 452879)	0.405 (df = 452880)
F Statistic	$15{,}511.650^{***}$ (df = 56; 452879)	$15{,}705.800^{***}$ (df = 55; 452880)	$21{,}331.100^{***}$ (df = 56; 452879)	$21{,}500.570^{***}$ (df = 55; 452880)
Note:				*p<0.1; **p<0.05; ***p<0.01

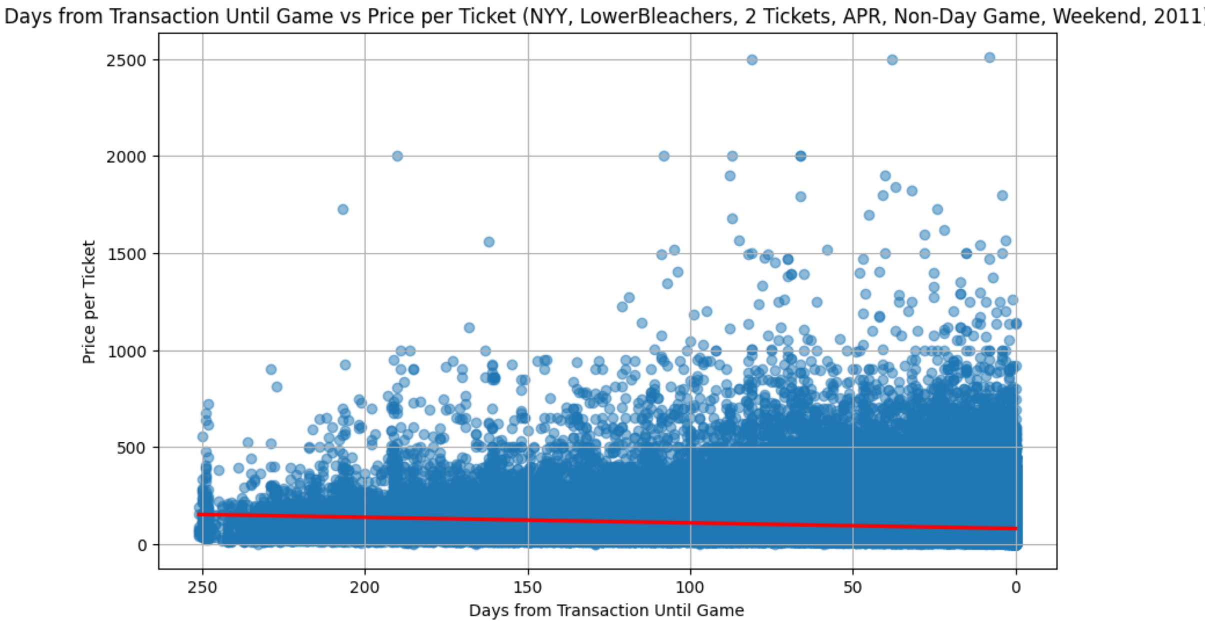
The results indicate that the earlier tickets are purchased, the more expensive they tend to be, as evidenced by the positive and significant coefficient on the variable `days_from_transaction_until_game`. This finding is counter intuitive. Typically, purchasing tickets earlier is expected to be less expensive because it reduces the risk for the ticket vendor, smooths out their cash flow, and provides an early cash injection. Additionally, the time value of money suggests that receiving payments earlier should incentivise lower prices.

To explore this unexpected result further, we conduct an investigation using scatter plots.





The relationship may be biased by noise. So this would warrant further investigation. To investigate further, I look into the team with the most observations (NYY). I also control for other attributes with the most observation within NYY (LowerBleachers, 2 Tickets, April, Evening and Weekend games in 2011). The following are my

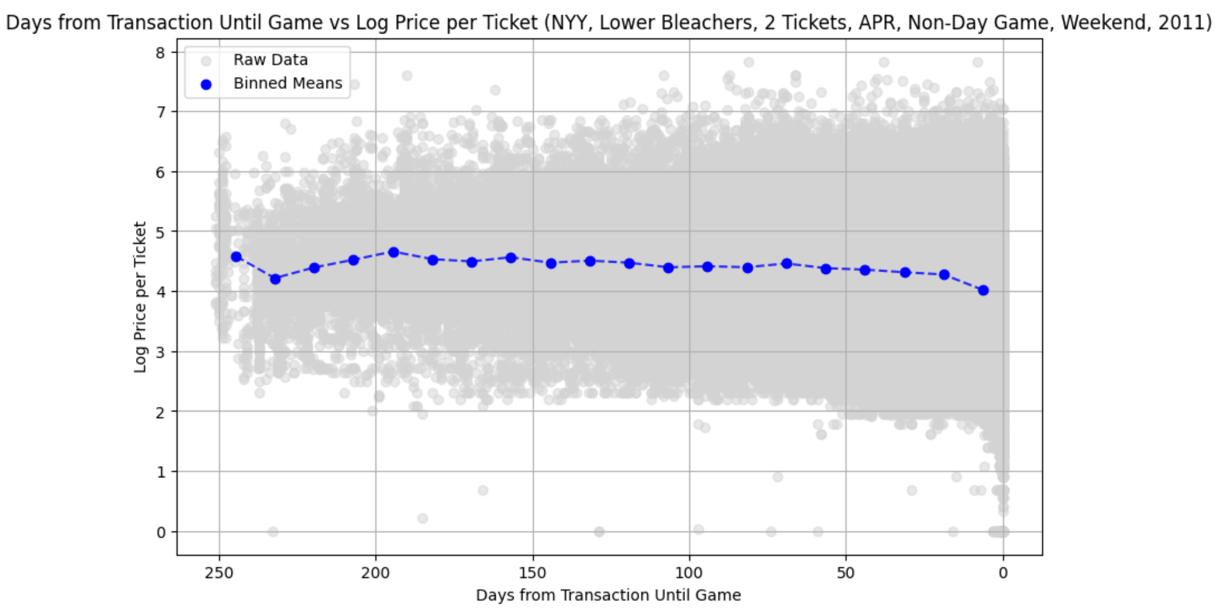


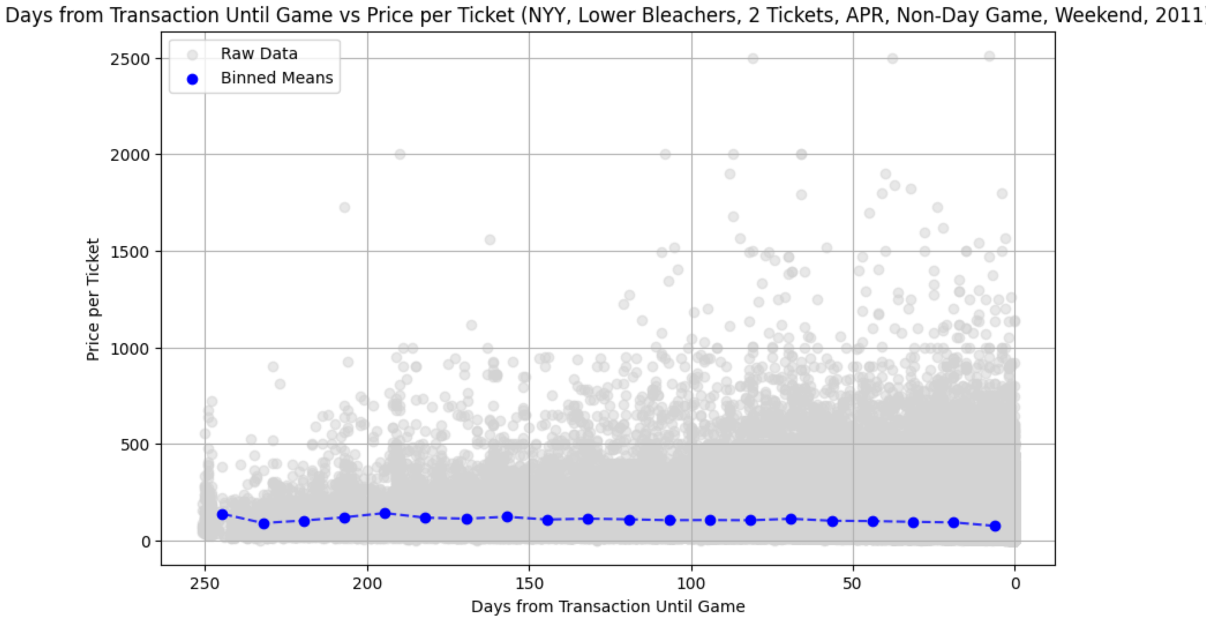
results.

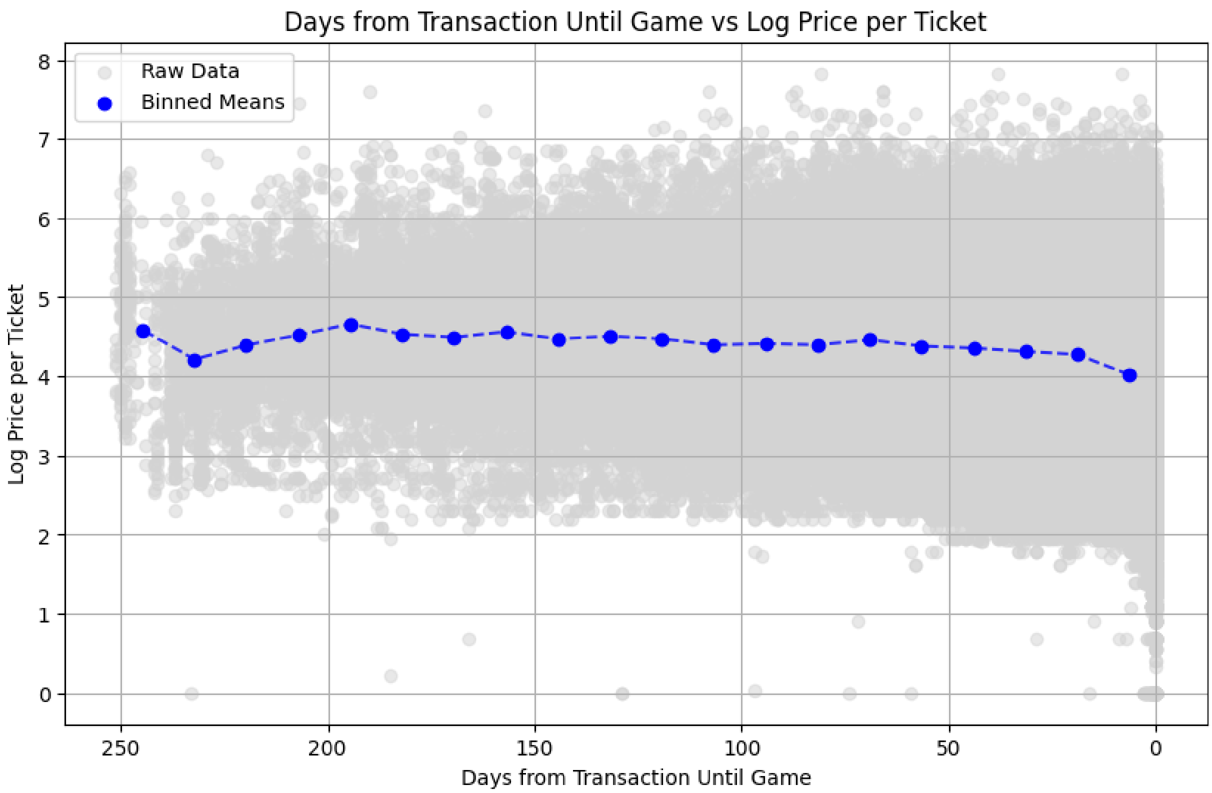
The results show similarity to the initial plots we observed from earlier. I suspect the results from the data are due to some noise introduced by price outliers. I clean the data of price outliers manually and rerun the same regressions, mentioned prior.

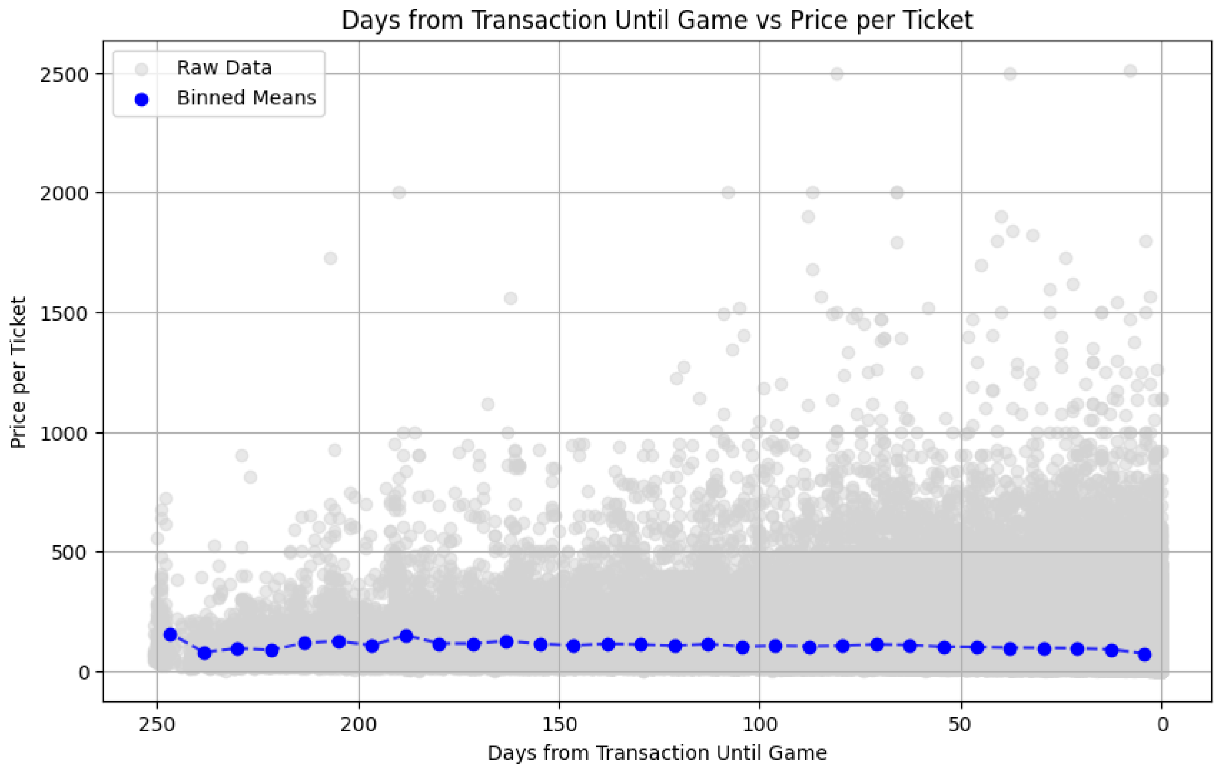
	Dependent variable:			
	price_per_ticket		logprice	
	(1)	(2)	(3)	(4)
days_from_transaction_until_game	0.228^{***}	0.233^{***}	0.003^{***}	0.003^{***}
	(0.002)	(0.002)	(0.00002)	(0.00002)
number_of_tickets	1.885^{***}		0.022^{***}	
	(0.043)		(0.0004)	
Constant	416.412^{***}	420.354^{***}	5.833^{***}	5.880^{***}
	(1.260)	(1.259)	(0.011)	(0.011)
Day Game Controls	Yes	Yes	Yes	Yes
Weekend Game Controls	Yes	Yes	Yes	Yes
Section Controls	Yes	Yes	Yes	Yes
Game Month Controls	Yes	Yes	Yes	Yes
Team Controls	Yes	Yes	Yes	Yes
Year Controls	Yes	Yes	Yes	Yes
Number of Tickets Controls	Yes	No	Yes	No
Observations	452,935	452,935	452,935	452,935
R^2	0.683	0.682	0.725	0.723
Adjusted R^2	0.683	0.682	0.725	0.723
Residual Std. Error	45.315 (df = 452878)	45.411 (df = 452879)	0.404 (df = 452878)	0.405 (df = 452879)
F Statistic	$17{,}434.380^{***}$ (df = 56; 452878)	$17{,}642.400^{***}$ (df = 55; 452879)	$21{,}341.330^{***}$ (df = 56; 452878)	$21{,}510.620^{***}$ (df = 55; 452879)
Note:				*p<0.1; **p<0.05; ***p<0.01

There does not seem to be any significant effect on the results. To further investigate, we use bin scatter plots and show our results again.

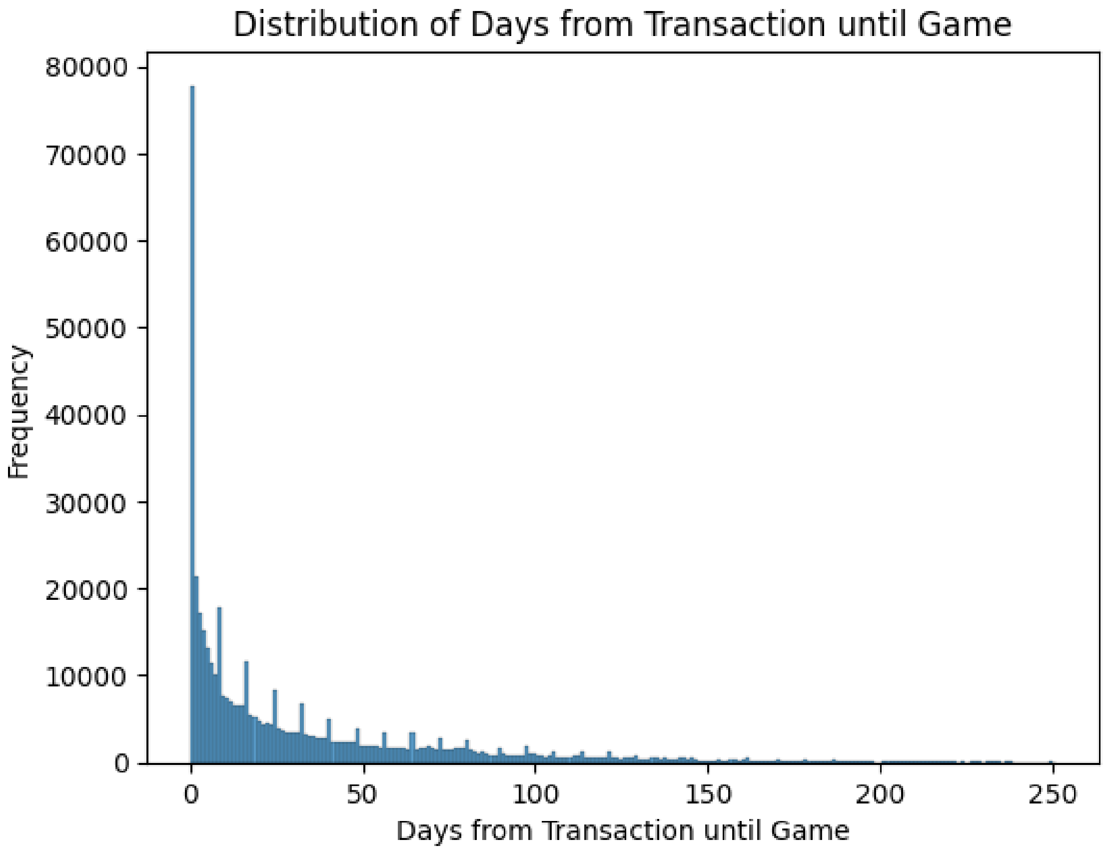



Days from Transaction Until Game vs Price per Ticket (NYY, Lower Bleachers, 2 Tickets, APR, Non-Day Game, Weekend, 2011)
Raw Data
Binned Means
Price per Ticket
Days from Transaction Until Game


Days from Transaction Until Game vs Log Price per Ticket
Raw Data
Binned Means
Log Price per Ticket
Days from Transaction Until Game



From this we see our regression results makes sense now. There are huge outliers of people who pay more when game day is near (relative to when game day is far). But on average people pay less when buying tickets near game day. So to answer this question How do the prices consumers pay for tickets change as the game date approaches (i.e., as the number of days between transaction date and game date declines)? The initial answer would be the prices **decrease** as game date approaches. To further investigate the dynamic pattern, we would run a quasi "event study" model to investigate. We show the distribution of transaction and see there is bunching happening approximately every 8 days. There are many reasons why this can be happening ranging from discounts/promotions, timing of the games, etc. While we don't know why exactly this is happening we can exploit these observations for our event study model.



Using this observation we create our model.

$$\text{price\_per\_ticket} = \beta_0 + \sum_{j=1}^{n} \beta_{1j} \cdot D_j \cdot \text{days\_until\_game} + \beta_2 \cdot \text{controls} + \beta_3 \cdot \text{num\_tickets} + \epsilon$$

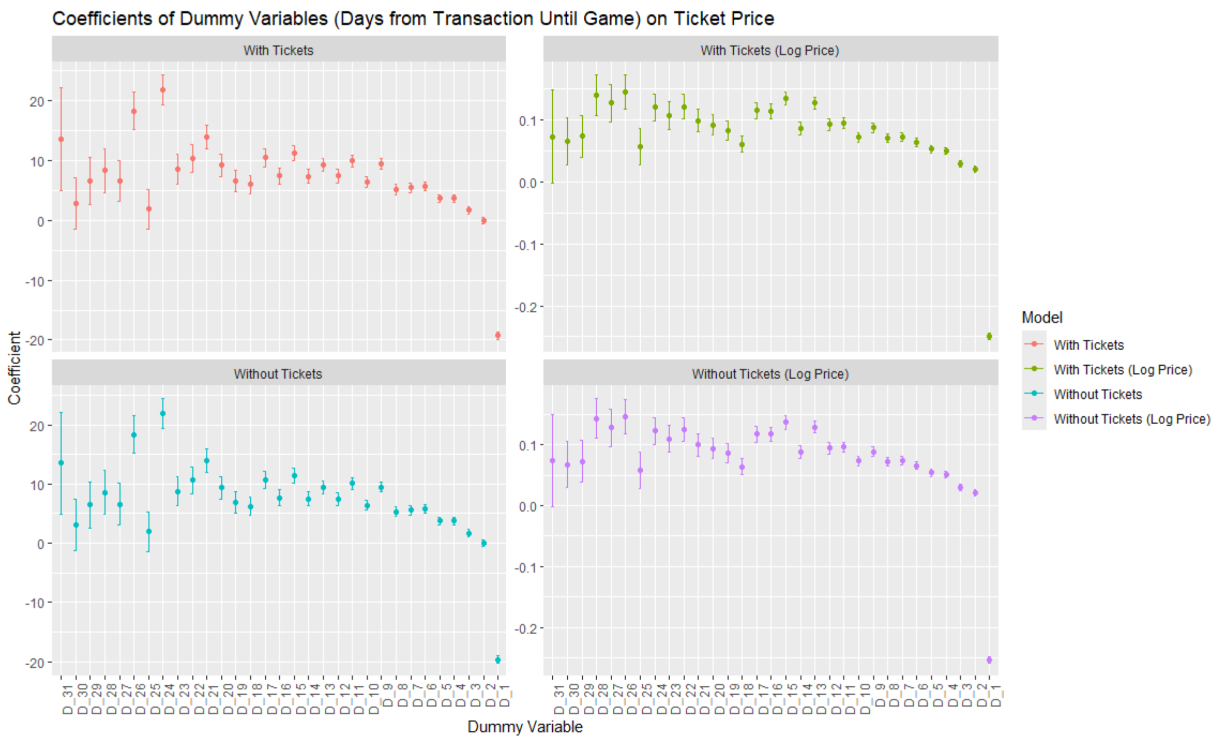
2.1 Where:

$$\text{controls} = \text{day\_game, weekend\_game, sectiontype, gamemonth, team, year}$$

- β_0 is the intercept.
- β_{1j} is the coefficient for each range j of days from transaction until game:
  - $D_1 = 1$ if days are in the range $0 - 8$
  - $D_2 = 1$ if days are in the range $9 - 16$
  - $D_3 = 1$ if days are in the range $17 - 24$

- ...
- $D_n = 1$ for the last specified range (e.g., $241 - 250$).

- $\beta_2 \cdot$ controls represents a vector of control variables included in the model.
- β_3 is the coefficient for the number of tickets.
- ϵ is the error term.



We see from this that the relationship may not be entirely linear. We also see that ticket prices are in fact lower come one week before a game, which supports previous results. One thing we did not take into account for is the bias of human time perception. People typically think of time between weeks, days and months, wherein overly long periods of times are not referred to in weeks but in months. To study this, we run the same model but take into account human biases, instead of cutting the dummy variables in a 8 day basis, we cut up our dummy variables based on human perceptions of months and weeks.

$$\text{price\_per\_ticket} = \beta_0 + \sum_{j=1}^{n} \beta_{1j} \cdot D_j \cdot \text{days\_until\_game} + \beta_2 \cdot \text{controls} + \beta_3 \cdot \text{num\_tickets} + \epsilon$$

2.2 Where:

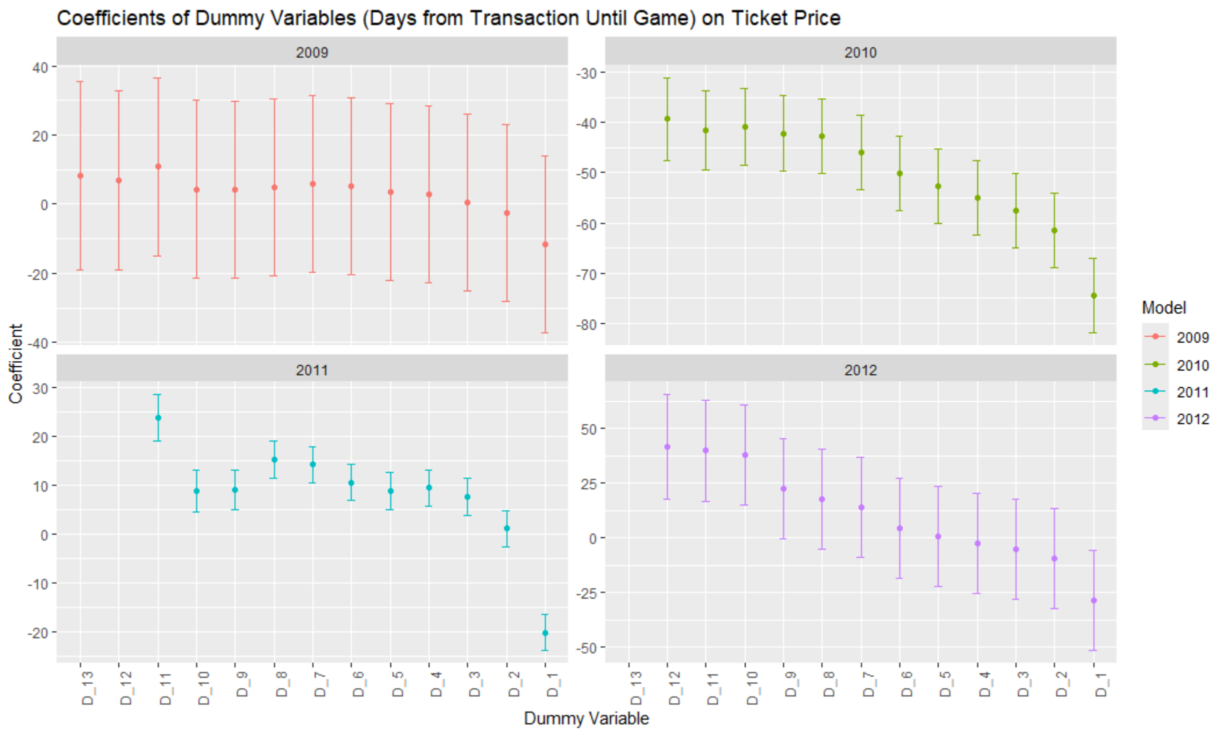
$$\text{controls} = \text{day\_game}, \text{weekend\_game}, \text{sectiontype}, \text{gamemonth}, \text{team}, \text{year}$$

- β_0 is the intercept.
- β_{1j} is the coefficient for each range j of weeks and months from the transaction until the game:
  - $D_1 = 1$ if the time until the game is in the range of 0-1 week
  - $D_2 = 1$ if the time until the game is in the range of 1-2 weeks
  - $D_3 = 1$ if the time until the game is in the range of 2-3 weeks
  - $D_4 = 1$ if the time until the game is in the range of 3-4 weeks
  - $D_5 = 1$ if the time until the game is in the range of 4-5 weeks
  - $D_6 = 1$ if the time until the game is in the range of 1-2 months
  - $D_7 = 1$ if the time until the game is in the range of 2-3 months
  - $D_8 = 1$ if the time until the game is in the range of 3-4 months
  - ...
  - $D_n = 1$ if the time until the game is in the range of 8 to 8.3 months, as the data concludes at 250 days.
- $\beta_2 \cdot$ controls represents a vector of control variables included in the model.
- β_3 is the coefficient for the number of tickets.
- ϵ is the error term.

We see smoother and more observable relationship here. All this to suggest that in fact, on average, the later you buy your tickets the cheaper ticket prices would be. Next we study the year differences. We look at the year coefficients of our main model.

year2010	-7.773***
	(0.208)
year2011	-5.471***
	(0.197)
year2012	-13.459***
	(0.242)

We see that there are significant year fixed effects that could be worth investigating. Next, we run the same analysis restricting our observations within each year.



The observed trends reveal an interesting pattern: purchasing tickets well in advance tends to result in higher ticket prices. However, there are variations in this relationship over the years. For instance, in 2009, the standard error is notably large, indicating substantial variability. During this year, some consumers were able to purchase tickets far in advance (approximately 250 days before the game) at prices comparable to those closer to game day.

In contrast, the dynamics shift in subsequent years, particularly in 2010, 2011, and 2012. The standard error becomes significantly smaller, demonstrating greater consistency in pricing trends. As a result, a clearer pattern emerges: tickets purchased closer to game day are generally cheaper. By 2012, buying tickets just one week before the game consistently results in lower prices, emphasizing the evolving dynamic relationship between purchase timing and ticket cost.

3 Task 2: Data Manipulation

The data task took approximately 40 min.

References

Co, W. (n.d.), 'EvaluativeAssignment_2024/EvaluativeAssignment_2024/Writing Sample.pdf at main · WilliamClintC/EvaluativeAssignment_2024', https://github.com/WilliamClintC/EvaluativeAssignment_2024/blob/main/EvaluativeAssignment_20