Apoorv Gupta and Jonathan Zinman: Coding Task

William Co*
Department of Economics, University of British Columbia

January 15, 2025

Abstract

This is a data task submitted for a predoctoral application, in accordance with the guidelines outlined in the following document https://github.com/WilliamClintC/Coding-Task-GuZi/blob/main/GZ_RA_StataTaskDescription.pdf .

*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

I will be using quarto to construct this document per instruction (Co 2025*a*)

The code for task 2 can be found here. link (Co 2025*b*)

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:

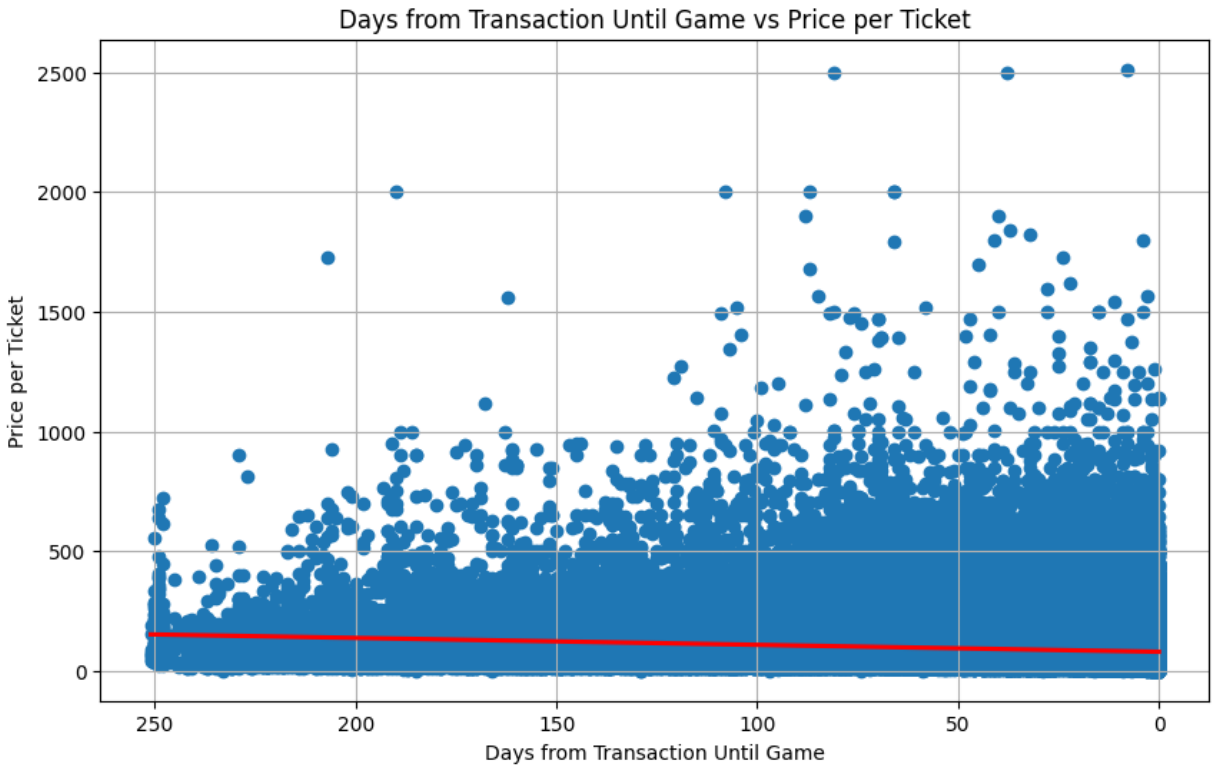
$$\text{controls} = \text{day\_game, weekend\_game, sectiontype, gamemonth, team, year}$$

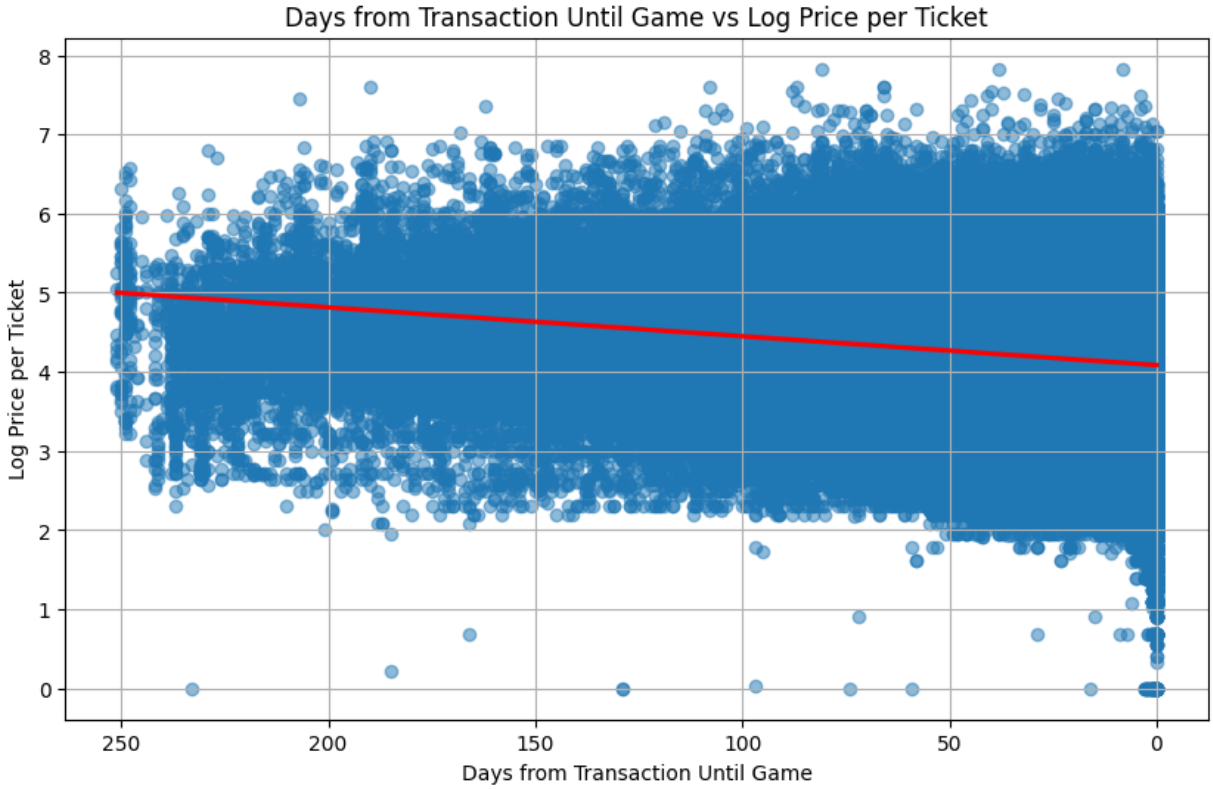
- β_0 is the intercept.
- β_1, β_3 are the coefficients for the respective predictor variables.
- β_2 is a vector of coefficients for control variables.
- ϵ is the error term.

	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.

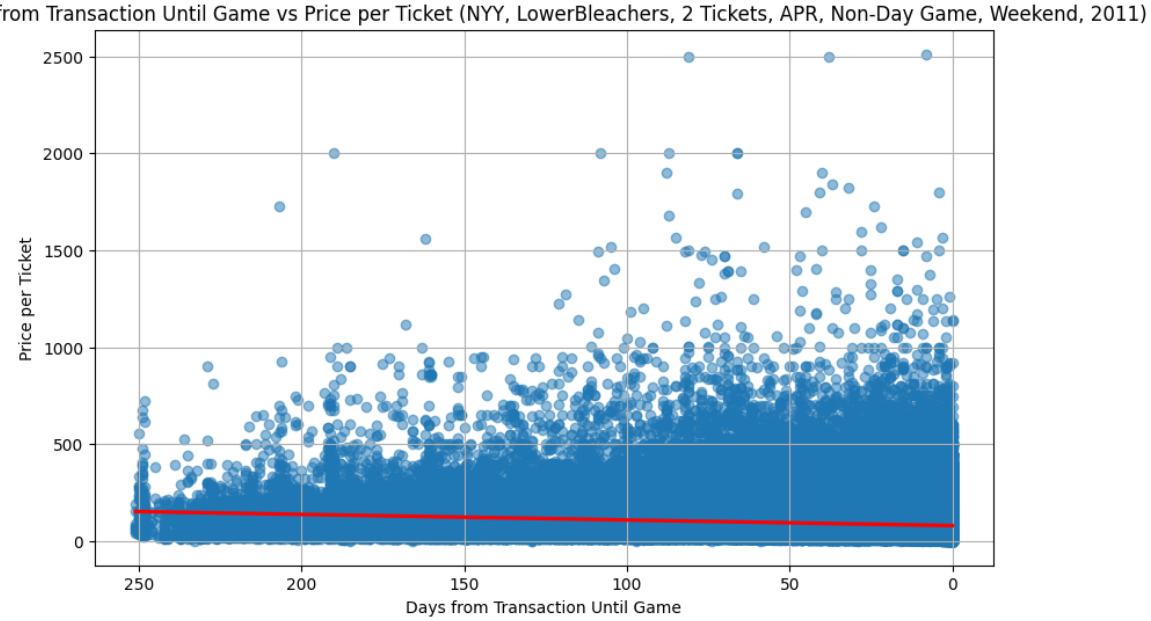
To explore this unexpected result further, we conduct an investigation using scatter plots.





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.

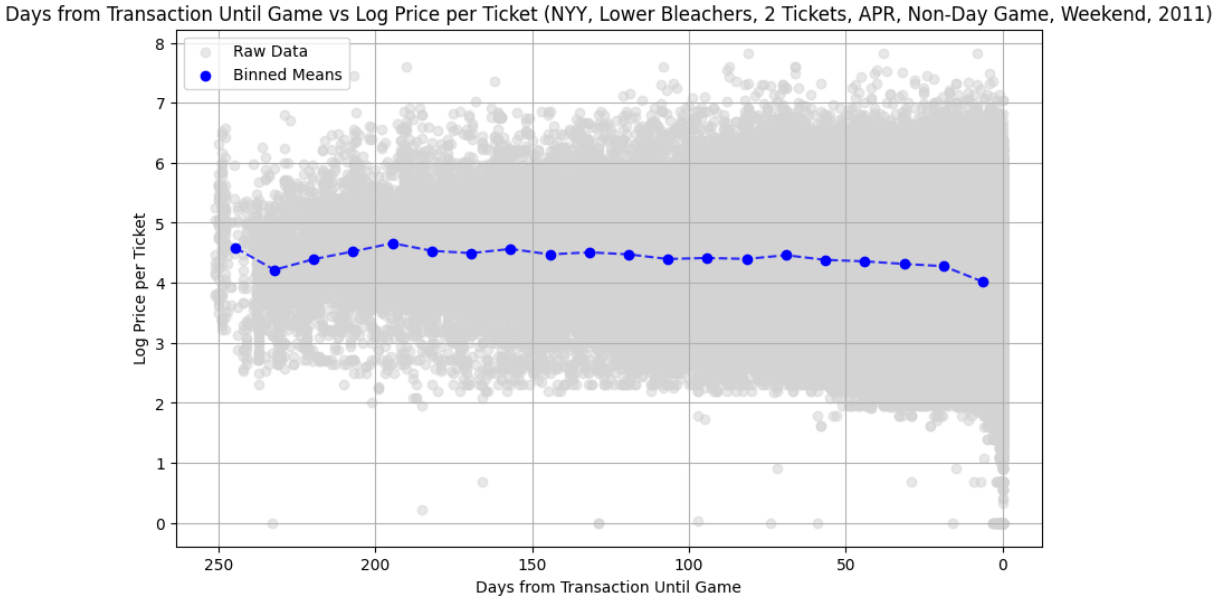


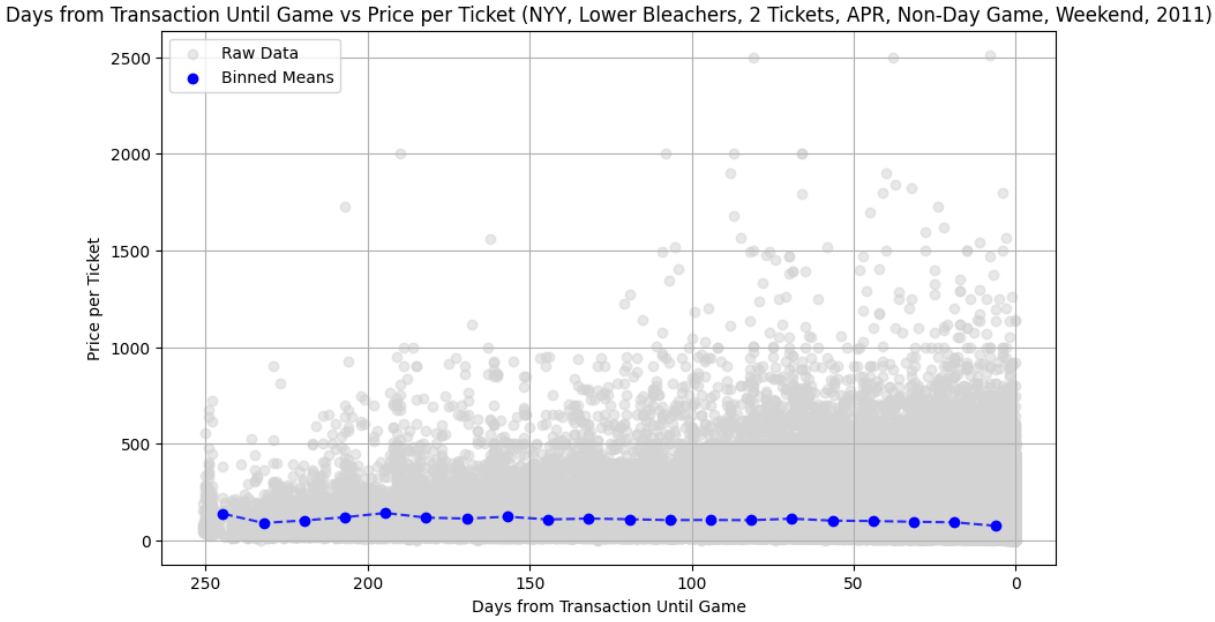
The results show similarity to the initial plots we observed from earlier. I suspect the results from the data are due to some noise introduced by price outliers. I clean the data of price outliers manually and rerun the same regressions, mentioned prior.

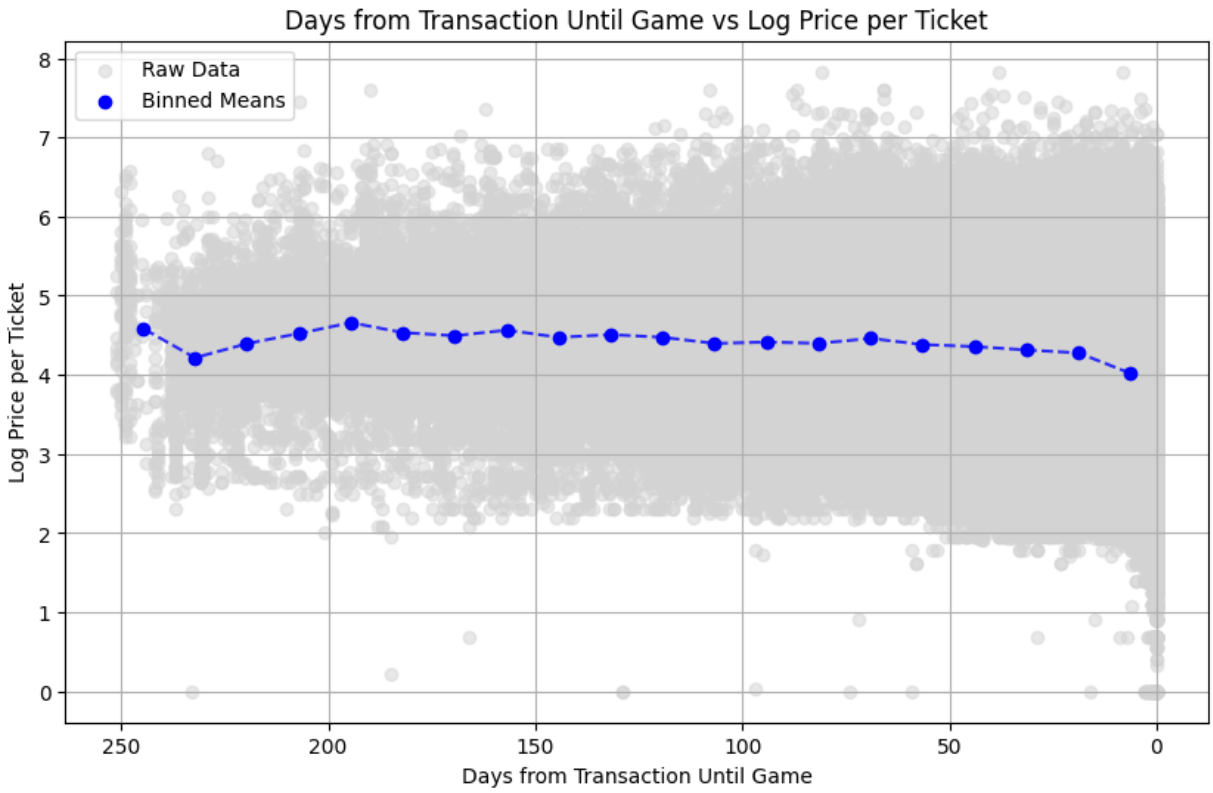
	Dependent variable:			
	price_per_ticket		logprice	
	(1)	(2)	(3)	(4)
days_from_transaction_until_game	0.228***	0.233***	0.003***	0.003***
	(0.002)	(0.002)	(0.00002)	(0.00002)
number_of_tickets	1.885***		0.022***	
	(0.043)		(0.0004)	
Constant	416.412***	420.354***	5.833***	5.880***
	(1.260)	(1.259)	(0.011)	(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

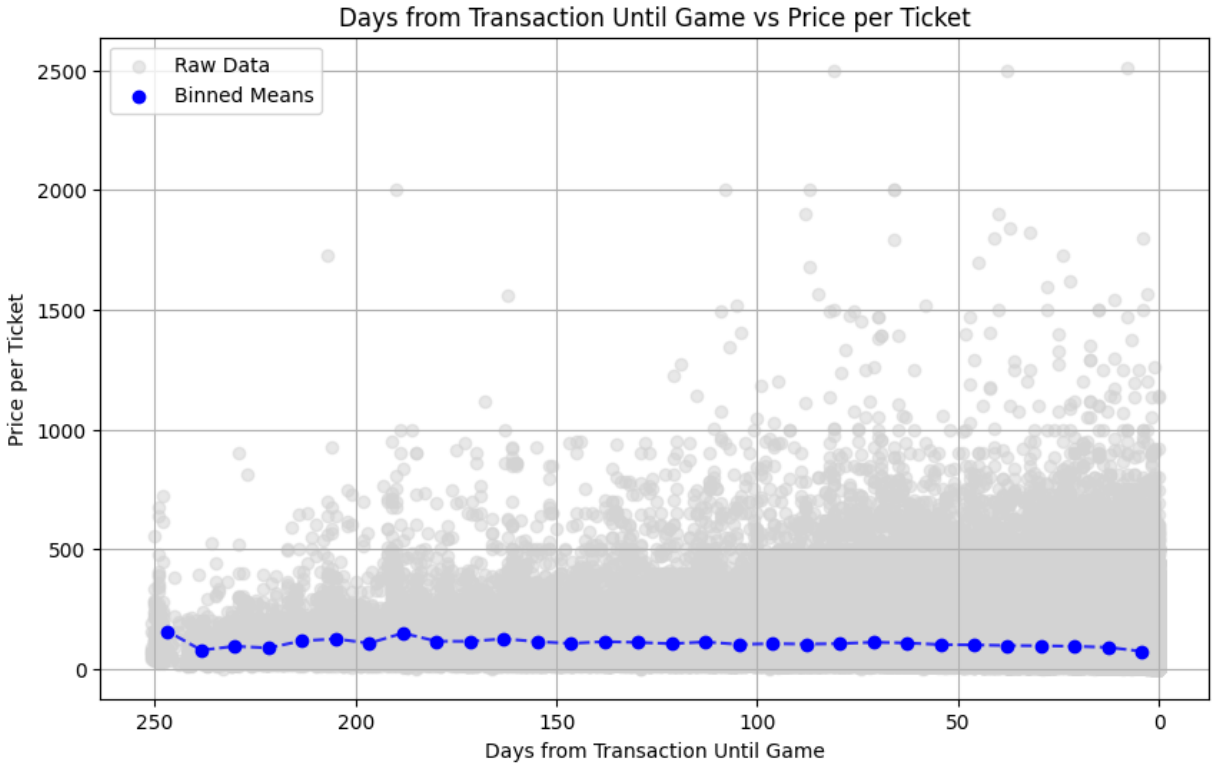
There does not seem to be any significant effect on the results. To further investigate, we

use bin scatter plots and show our results again.



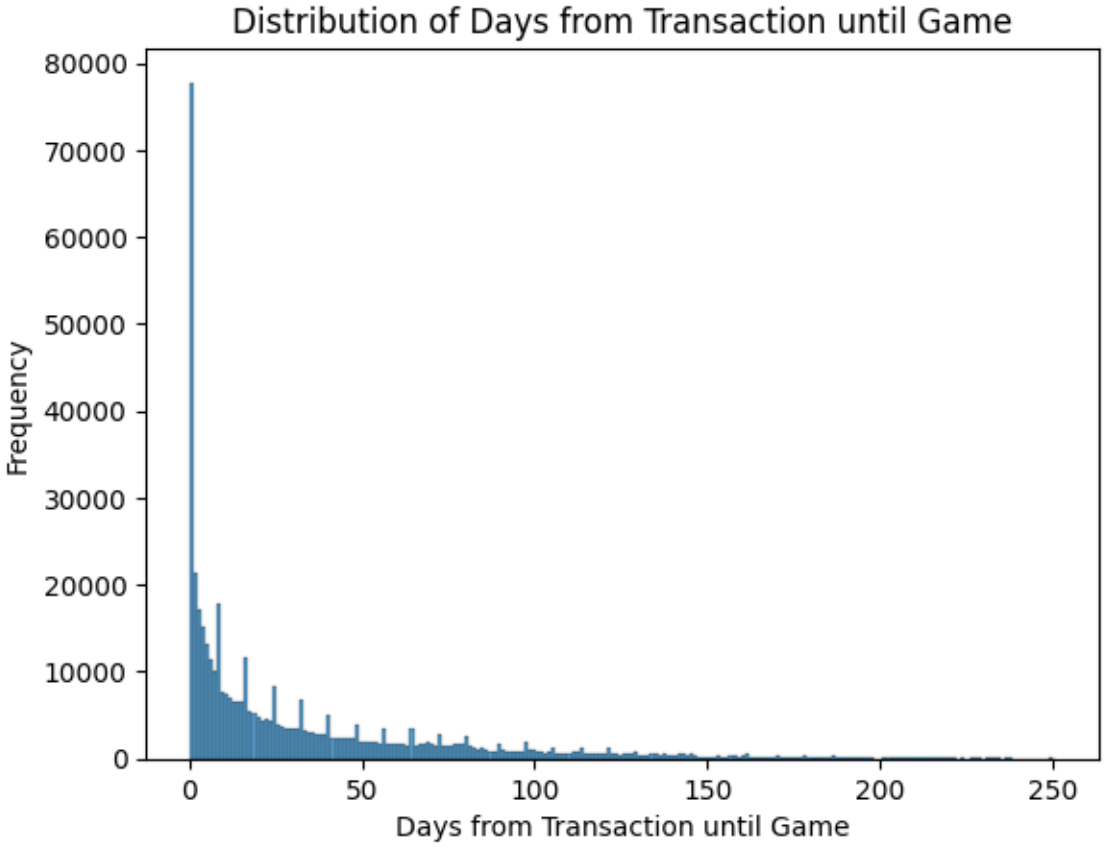






From this we see our regression results makes sense now. There are huge outliers of people who pay more when game day is near (relative to when game day is far). But on average

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.



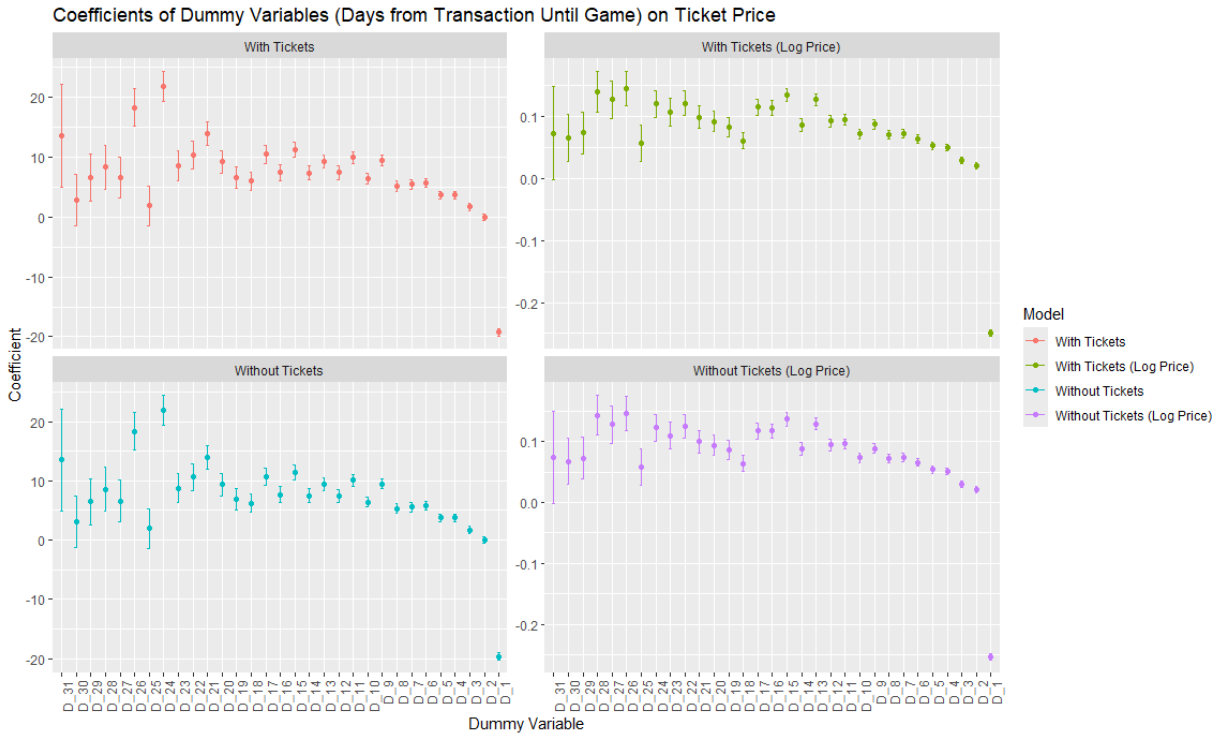
Using this observation we create our model.

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

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., $241 - 250$).
- $\beta_2 \cdot \text{controls}$ 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.

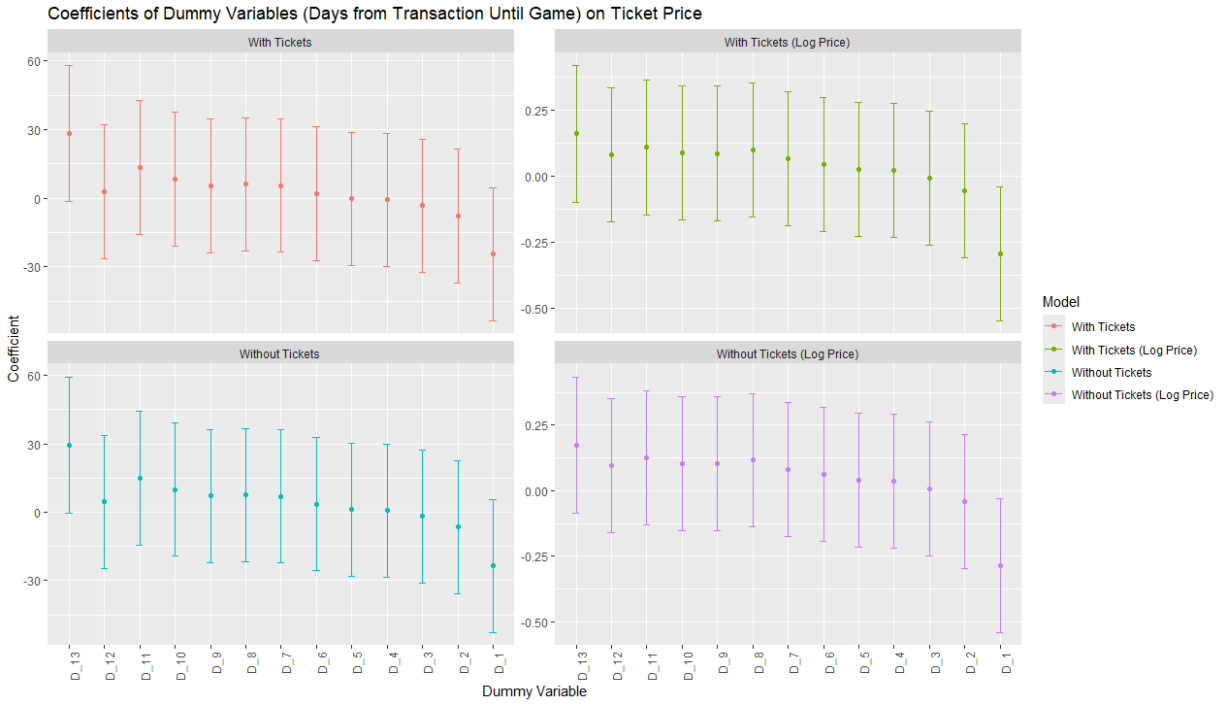
$$\text{price\_per\_ticket} = \beta_0 + \sum_{j=1}^{n} \beta_{1j} \cdot D_j \cdot \text{days\_until\_game} + \beta_2 \cdot \text{controls} + \beta_3 \cdot \text{num\_tickets} + \epsilon$$

2.2 Where:

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

- β_0 is the intercept.

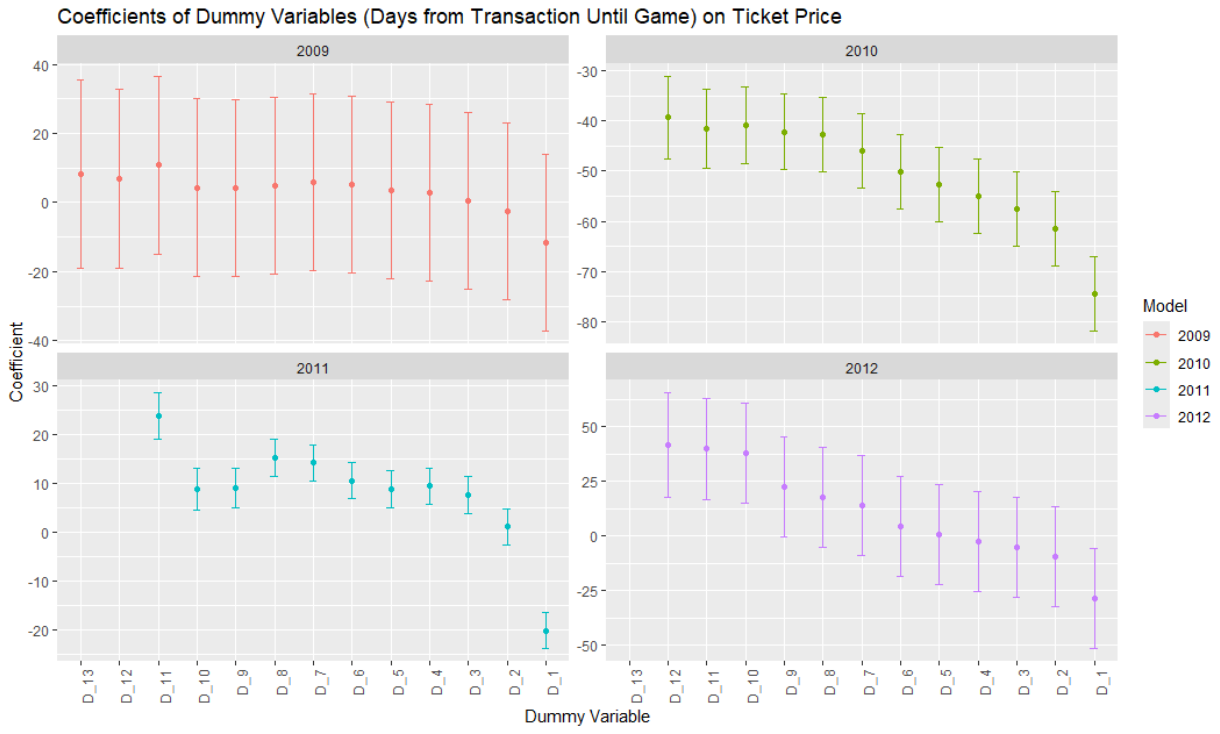
- β_{1j} is the coefficient for each range j of weeks and months from the transaction until the game:
  - $D_1 = 1$ if the time until the game is in the range of 0-1 week
  - $D_2 = 1$ if the time until the game is in the range of 1-2 weeks
  - $D_3 = 1$ if the time until the game is in the range of 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: Lottery Study

3.1 Part I: Do any of the observations seem suspicious to you?

3.1.1 1.Descriptive Data Analysis

The following are all my observations of the data. In summary, **`Expend_Total`** variable has a outlier with implausible values, which has been removed. Duplicates are remoed as well. More details are as follows.

- **Income and Age Variables:** The distributions of income and age appear normal based on the plots. These variables seem acceptable.
- **Race**: The values for `black`, `hispanic`, and `white` align with the expected categories.
- **Gender**: The data contains `male` and `female`, but the manual specifies values should be:
  - `1` for male
  - `2` for female
  - `3` for "other."

    This inconsistency needs to be addressed.
- **Marital Status**: The data includes values such as `married` and `0.0`, whereas the manual specifies:
  - `1` for married
  - `0` for not married.
- **Urban Residence**: The column contains `Metro Area` and `Non-Metro Area`, but it should be:

  - `1` for urban (residing in a metropolitan area)
  - `0` otherwise.

- **Employment and Religion Variables:** These variables seem acceptable.
- **Years of Education:** This variable appears normal, and a plot has been attached for reference.
- **Ideology:** The data contains values such as `Extremely Conservative`, `Moderate`, `Slightly Conservative`, etc., but the manual specifies:
  - `1` for extremely liberal
  - `2` for liberal
  - `3` for slightly liberal
  - `4` for moderate
  - `5` for slightly conservative
  - `6` for conservative
  - `7` for extremely conservative
- **State:** The data lists 51 states, which is incorrect as there are only 50 states in the U.S. This discrepancy needs investigation.
- **Expend_Total:** This variable is categorized as categorical in the data but should be numeric. Additionally, there are significant outliers, including an entry with a value of `100,000.0`. Some entries are not even numeric, requiring cleaning and conversion.
- **Income_Delta:** This variable appears normal, and a plot has been attached for verification.

- **Expend_Delta:** Although there are some outliers, the variable seems normal otherwise.
- **Income_Effects_Delta_Pct:** This variable is generally normal, apart from a few outliers.
- **Risk-Seeking:** The data contains entries like `-3`, `-4`, `-1 - Very unwilling`, and `-7 - Very willing`. However, the manual specifies the scale should range from `-7` to `-1`, where `-1` is "very unwilling" and `-7` is "very willing." The inconsistent format needs correction.
- **Risk Aversion:** This variable contains text entries, but it should include categorical numerical values:
  - `1` for "Substantial financial risks expecting to earn substantial returns"
  - `2` for "Above-average financial risks expecting to earn above-average returns"
  - `3` for "Average financial risks expecting to earn average returns"
  - `4` for "No financial risks."
- **Seems_Fun:** Entries include `-3 - Strongly Disagree`, `0 - Neutral`, and `3 - Strongly Agree`, but the values should be on a scale from `-3` (strongly disagree) to `3` (strongly agree). The inconsistency requires recording.
  - Similar issues are present in `enjoy_thinking` and `self_control`.
- **Financial Literacy, Numeracy, and Related Columns:** Variables such as `financial_literacy`, `financial_numeracy`, `non_belief_lln`, `ev_miscalculation`, `overconfidence`, and `lottery_payout` appear normal.
- **Happiness:** This column contains NaN values, which should be addressed during

data cleaning.

3.1.2 2. tabulation of income and gender

gender	Male	Female
income		
5.0	18	28
7.5	17	44
12.5	42	62
17.5	39	69
22.5	75	84
27.5	72	94
32.5	76	83
37.5	61	73
45.0	128	124
55.0	142	127
67.5	186	159
80.0	81	64
92.5	144	138
112.5	137	129
137.5	97	61
162.5	57	38
187.5	21	23
250.0	53	41

3.1.3 3. Report detailed summary statistics

```
age_summary = df['age'].describe()
age_summary
```

```
count    2887.000000
mean       48.879806
std        16.901751
min        18.000000
25%        34.000000
50%        49.000000
75%        63.000000
max       145.000000
Name: age, dtype: float64
```

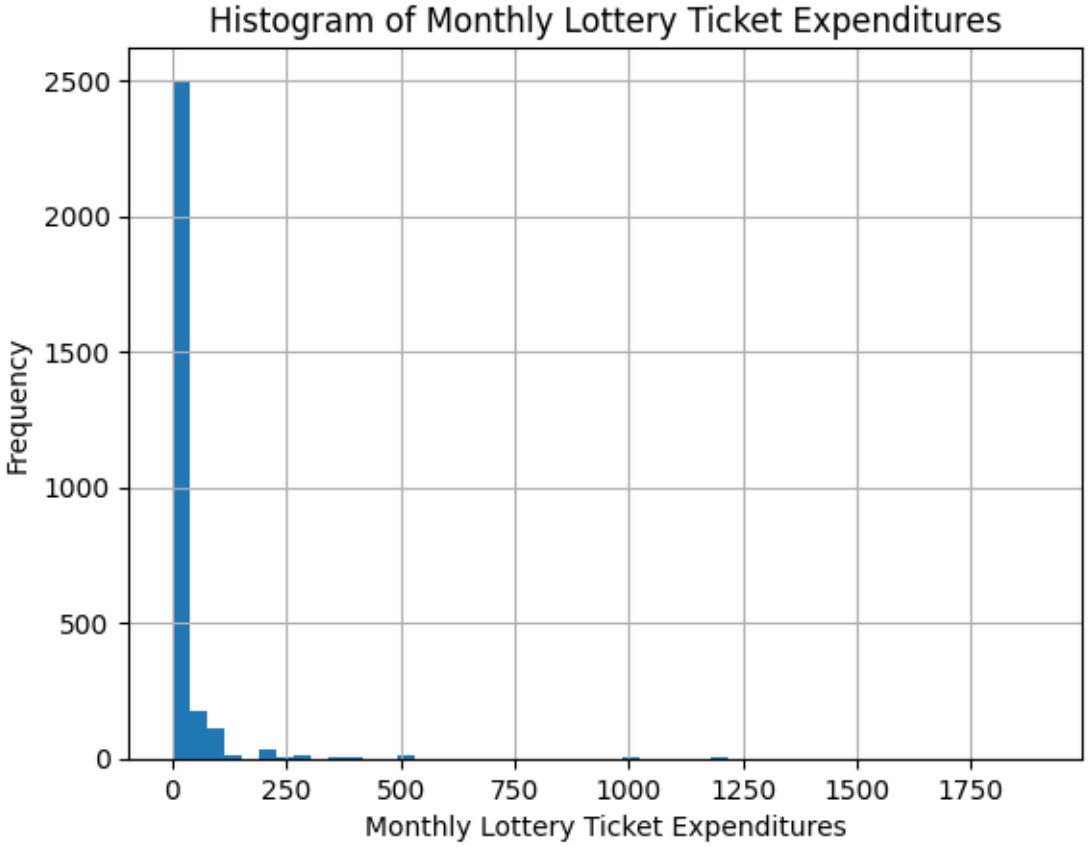
```
mean_age = age_summary['mean']
print(f"The mean age in the sample is {mean_age:.2f}")
```

```
The mean age in the sample is 48.88
```

3.1.4 4. Report summary statistics

```
count    2887.000000
mean       24.101489
std        88.288585
min         0.000000
25%         0.000000
50%         2.000000
75%        15.000000
max      1900.000000
Name: expend_total, dtype: float64
```

```
count    2887.000000
mean        0.761211
std         3.537560
min         0.000000
25%         0.000000
50%         0.029630
75%         0.285714
max        69.090909
Name: expenditures_share_income, dtype: float64
```



3.1.5 5. Indicator Variable

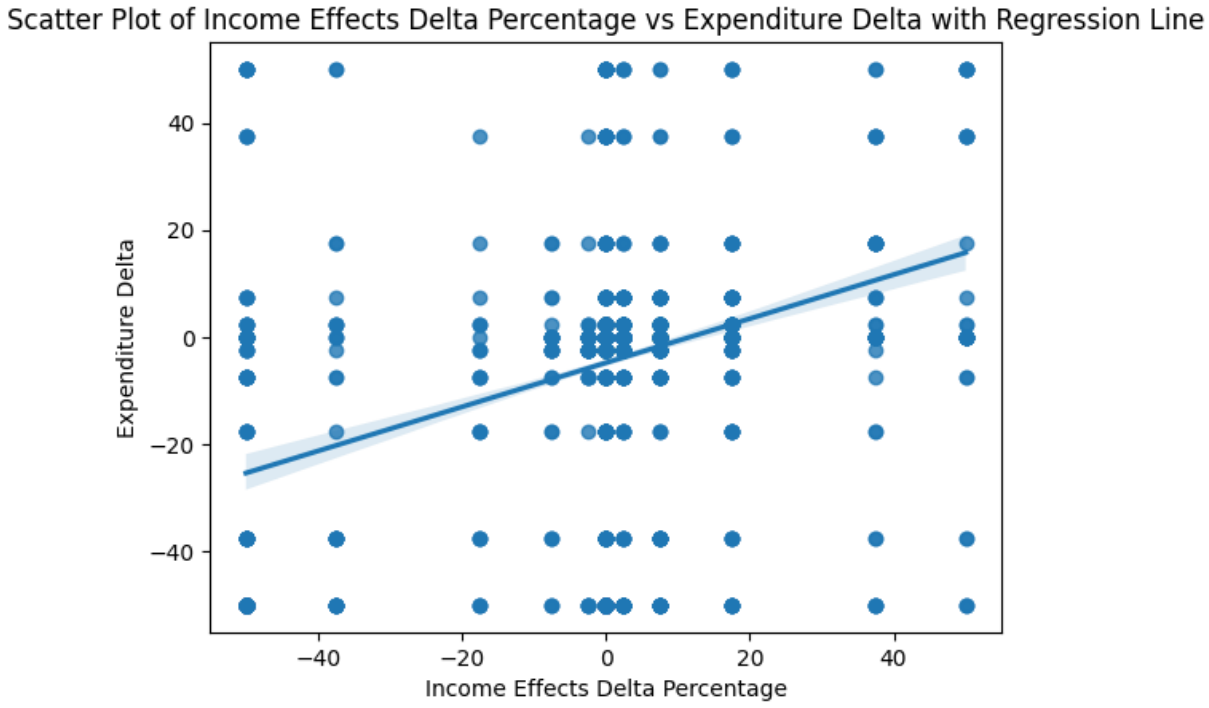
```python
education_thresholds = [5, 10, 15, 20]

for y in education_thresholds:
    df[f'education_{y}'] = df['education'].apply(lambda x: 1 if x >= y else 0)

df[['education', 'education_5', 'education_10', 'education_15', 'education_20']]
```

3.1.6 6. How did changes in income from 2018 to 2019 vary

We see a positive relationship between income delta and expenditure delta.

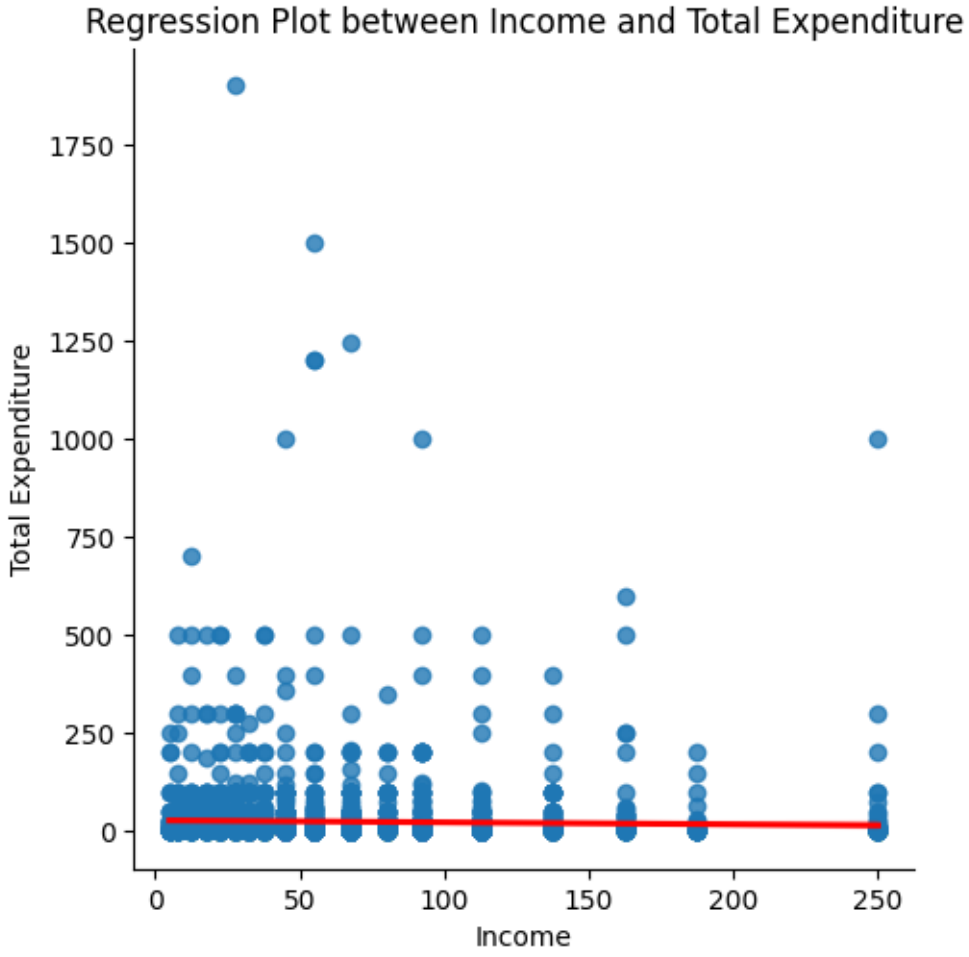


We see income effect delta and expenditure delta has a steeper positive slope than income delta against expenditure delta. Overall, both plots have similar variation.

3.2 Part II Unpacking the determinants and correlates of lottery expenditures

3.2.1 1.relationship between the monthly lottery expenditure variable and income.

Below are my preliminary plots.



Previous knowledge on the topic suggests that higher-income individuals tend to spend less on lotteries. However, based on our initial observations, this does not appear to be the case. When plotting income on the x-axis and lottery expenditure on the y-axis, we expect a negative slope, reflecting the anticipated inverse relationship. Instead, our observations reveal a nearly flat line, although there is a slight downward trend.

Though, when examining lottery expenditure as a share of income, we find that lower-income individuals allocate a larger proportion of their income to lottery tickets compared to higher-income individuals. Some observations show extreme cases where up to 70% of income is spent on lottery tickets, aligning with preconceived notions about lottery spending behavior among low-income groups.

To investigate this further, we will conduct a multiple regression analysis using cross-sectional data.

- **Dependent Variable**: Lottery expenditure

- **Independent Variable**: Income

- **Control Variables**: Age, race (Black, Hispanic, White), gender, marital status, urban residence, employment status, religion, education level, ideology, and state of residence.

To account for age-related life-cycle effects, we will create dummy variables representing different age groups:

1. **Not Working Age**
2. **Early Working Age**
3. **Late Working Age**
4. **Retired**

These dummy variables will capture the differential effects across the specified age categories. This approach provides a more nuanced understanding of the relationships between income, demographic factors, and lottery expenditure.

Additionally, I will use ideology as a continuous sliding scale ranging from left (liberal) to right (conservative) as one of the variables in the analysis.

The model is:

$$Y_i = \beta_0 + \beta_1 \text{Income}_i + \sum_{j=1}^{p} \gamma_j C_{ij} + \sum_{k=1}^{q} \delta_k A_{ik} + \epsilon_i$$

Where:

- Y_i: Lottery expenditure (dependent variable)

- Income_i: Individual income (main independent variable)

- $\sum_{j=1}^{p} \gamma_j C_{ij}$: Summation over p control variables (C_{ij}) with coefficients γ_j. Control variables include:

  - Race ($\text{Black}_i, \text{Hispanic}_i, \text{White}_i$) - dummy variable
  - Gender - dummy variable
  - Marital status - dummy variable
  - Urban residence - dummy variable
  - Employment status - dummy variable
  - Religion - dummy variable
  - Education
  - Ideology
  - State - dummy variables
  - $\sum_{k=1}^{q} \delta_k A_{ik}$: Summation over q age life-cycle dummies (A_{ik}) with coefficients δ_k. These variables are:
    - **NotWorkingAge**: $A_{i1} = 1$ if age < 15, else $A_{i1} = 0$
    - **EarlyWorkingAge**: $A_{i2} = 1$ if $15 \leq \text{age}_i < 30$, else $A_{i2} = 0$
    - **LateWorkingAge**: $A_{i3} = 1$ if $30 \leq \text{age}_i < 60$, else $A_{i3} = 0$

– **Retired**: $A_{i4} = 1$ if $\text{age}_i \geq 60$, else $A_{i4} = 0$

- ϵ_i: Error term

	expend_total		expenditures_share_income	
	(1)	(2)	(3)	(4)
income	0.044	-0.051	-0.008***	-0.011***
	(0.037)	(0.032)	(0.001)	(0.001)
black1	26.405***		1.274***	
	(8.451)		(0.332)	
hispanic1	11.889		0.086	
	(7.800)		(0.307)	
white1	-5.537		-0.255	
	(6.714)		(0.264)	
genderMale	10.808***		0.211	
	(3.433)		(0.135)	
maritalMarried	-8.483**		-0.185	
	(3.790)		(0.149)	
urbanNon-Metro Area	11.839**		0.175	
	(4.990)		(0.196)	
employment1	-4.596		-0.573***	
	(3.971)		(0.156)	
religion1	0.940		0.029	
	(3.726)		(0.146)	
education	-1.901***		-0.026	
	(0.729)		(0.029)	
ideology	-0.437		-0.046	
	(1.132)		(0.044)	
not_working_age				
early_working_age	-4.200		-0.288	
	(6.610)		(0.260)	
late_working_age	17.221*		0.289	
	(9.226)		(0.363)	
retired	11.291		-0.260	
	(10.122)		(0.398)	
Constant	23.882	27.988***	1.756	1.546***
	(47.221)	(2.861)	(1.857)	(0.113)
State controls	Yes	No	Yes	No
Observations	2,772	2,772	2,772	2,772
R^2	0.043	0.001	0.071	0.026
Adjusted R^2	0.021	0.001	0.050	0.026

Our findings indicate that **the relationship between income and lottery ticket expenditure is weak.** After controlling for variables such as race, gender, marital status,

urban location, employment, religion, education, ideology, and age, we identified that more significant determinants of lottery expenditure include education, working age, race, gender, marital status, urban area, and age life cycle. When we use expenditure as a share of income as our dependent variable, the results align with our previous observations from the scatter plot. Specifically, **individuals with higher incomes tend to spend a smaller proportion of their income on lottery tickets**. This effect remains consistent even with all the controls mentioned earlier. Additionally, the results suggest that employment and race are more significant variables when it comes to ticket expenditure as a share of income.

3.2.2 2. Monthly lottery expenditure variable and bias proxy variables

I will explore the relationship between the monthly lottery expenditure variable and the preference and bias proxy variables. To achieve this, I will utilize the following variables: risk_seeking, risk_aversion, seems_fun, enjoy_thinking, overconfidence, and happiness. I plan to employ Principal Component Analysis (PCA) to combine these variables into a single composite variable that captures an individual's overall likelihood of being interested in participating in a lottery ticket.

Next, I will incorporate a self_control variable to represent an individual's impulse control, distinguishing it from their interest in participating in a lottery.

Lastly, I will again use PCA to create a composite variable from the following state variables: financial_literacy, financial_numeracy, gamblers_fallacy, non_belief_lln, and ev_miscalculation. This composite variable will capture an individual's understanding of risk and reward when participating in a lottery ticket.

Through this analysis, I aim to identify specific biases that influence lottery participation

and determine which aspects of bias play a more significant role in this behavior.

Regression Results

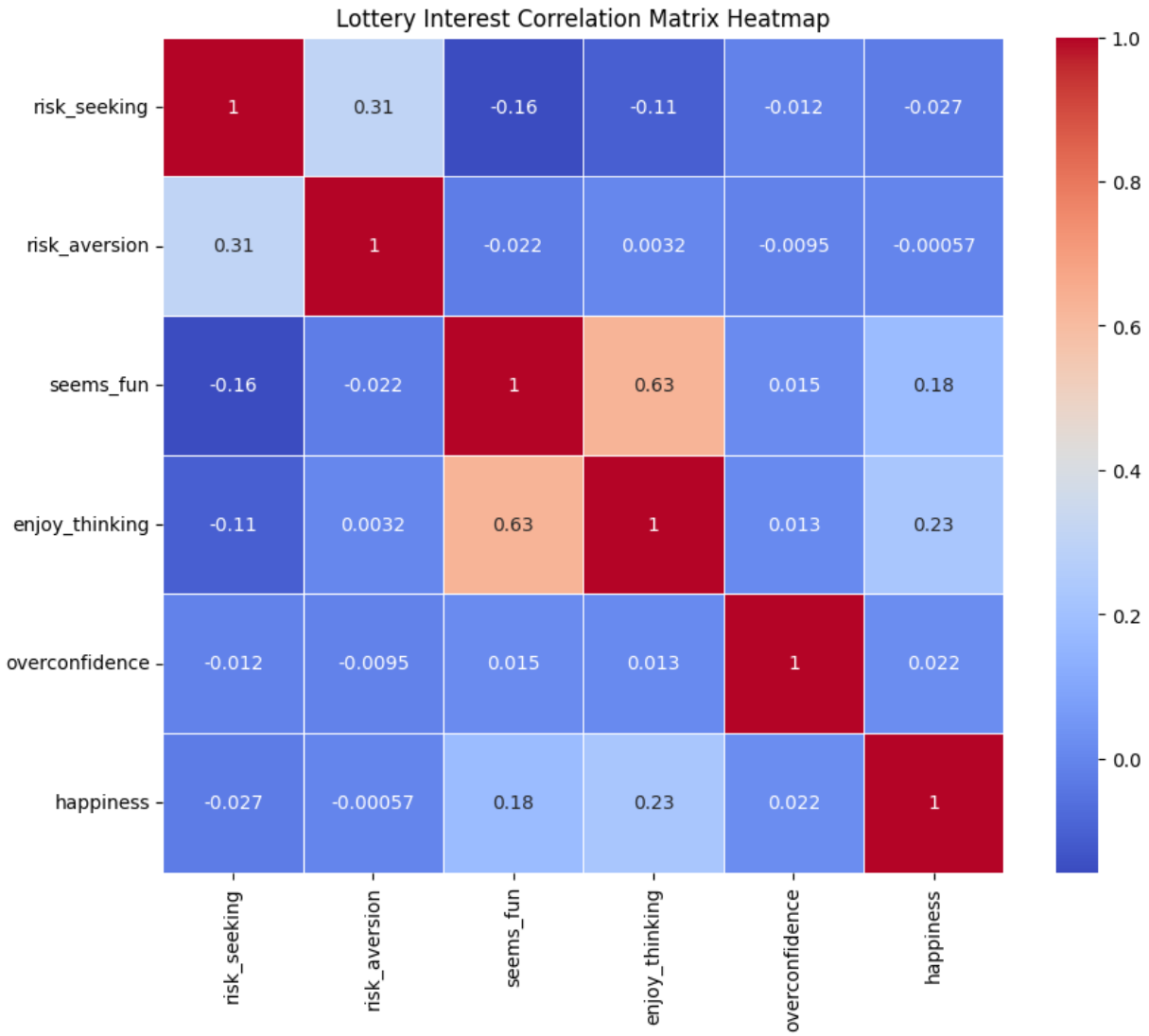
	Dependent variable:			
	expend_total	expenditures_share_income	expend_total	expenditures_share_income
	(1)	(2)	(3)	(4)
self_control	14.222***	0.433***	13.399***	0.408***
	(1.586)	(0.064)	(1.571)	(0.063)
PCA_Lottery_Understanding	0.547	-0.057		
	(1.117)	(0.045)		
PCA_Lottery_Interest	1.619	0.017		
	(1.276)	(0.052)		
financial_literacy			-24.019***	-1.213***
			(8.486)	(0.338)
financial_numeracy			-10.487	-0.852***
			(6.477)	(0.258)
gamblers_fallacy			3.065	-0.027
			(4.612)	(0.184)
non_belief_lln			7.747	0.193
			(9.885)	(0.393)
ev_miscalculation			5.618	-0.134
			(5.171)	(0.206)
risk_seeking			-1.582	-0.066
			(1.301)	(0.052)
risk_aversion			-1.456	-0.125
			(2.236)	(0.089)
seems_fun			6.448***	0.193***
			(1.176)	(0.047)
enjoy_thinking			1.397	0.012
			(1.124)	(0.045)
overconfidence			5.411*	-0.010
			(3.240)	(0.129)
happiness			0.025	-0.003
			(0.355)	(0.014)
Constant	43.604	2.340	26.871	2.906
	(46.319)	(1.876)	(47.675)	(1.898)
Income controls	Yes	Yes	Yes	Yes
Race controls	Yes	Yes	Yes	Yes
Gender controls	Yes	Yes	Yes	Yes
Marital status controls	Yes	Yes	Yes	Yes
Urban controls	Yes	Yes	Yes	Yes
Employment controls	Yes	Yes	Yes	Yes
Religion controls	Yes	Yes	Yes	Yes
Education controls	Yes	Yes	Yes	Yes
Ideology controls	Yes	Yes	Yes	Yes
State controls	Yes	Yes	Yes	Yes
Age controls	Yes	Yes	Yes	Yes
Observations	2,660	2,660	2,772	2,772
R^2	0.071	0.087	0.110	0.114
Adjusted R^2	0.049	0.065	0.085	0.089
Residual Std. Error	86.297 (df = 2595)	3.496 (df = 2595)	85.317 (df = 2695)	3.396 (df = 2695)
F Statistic	3.120*** (df = 64; 2595)	3.869*** (df = 64; 2595)	4.373*** (df = 76; 2695)	4.573*** (df = 76; 2695)

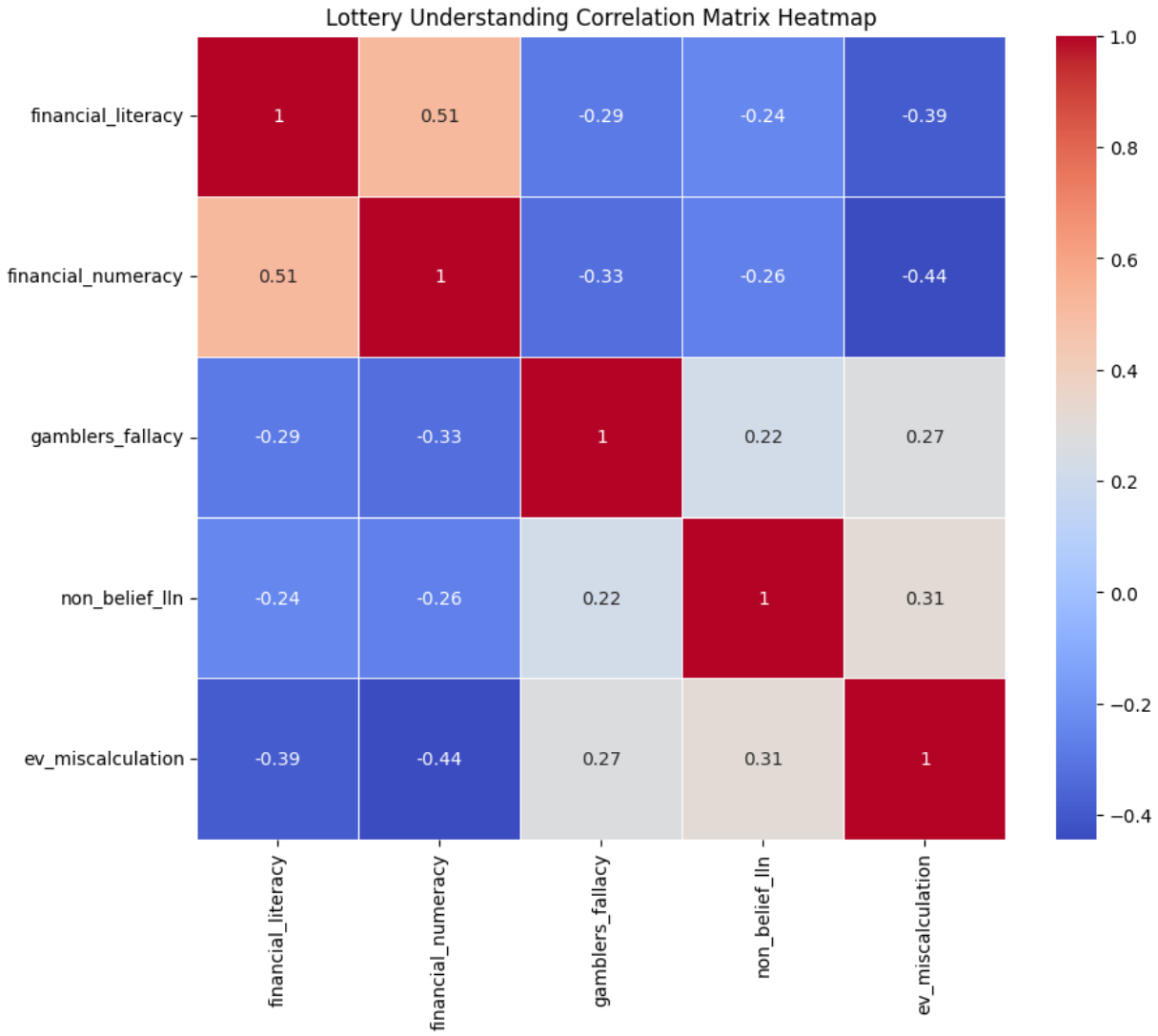
Examining our PCA variables, we find that the measures for lottery understanding and

lottery interest are insignificant. This suggests that lottery understanding and interest do not significantly influence lottery expenditure. On the other hand, self-control is significant, indicating that the aspect of self-control plays a critical role in driving lottery activity, potentially due to addictive behaviors.

These findings suggest that behavioral biases, particularly those related to self-control, play a substantial role in lottery expenditures. This may point to a form of addiction underlying lottery behavior.

To further study our created PCA variables, I analyze the correlation matrices underlying our PCA variables and compare them to our known controls and fixed effects.


Lottery Interest Correlation Matrix Heatmap
risk_seeking
risk_aversion
seems_fun
enjoy_thinking
overconfidence
happiness
1
0.31
-0.16
-0.11
-0.012
-0.027
0.31
1
-0.022
0.0032
-0.0095
-0.00057
-0.16
-0.022
1
0.63
0.015
0.18
-0.11
0.0032
0.63
1
0.013
0.23
-0.012
-0.0095
0.015
0.013
1
0.022
-0.027
-0.00057
0.18
0.23
0.022
1
risk_seeking
risk_aversion
seems_fun
enjoy_thinking
overconfidence
happiness
1.0
0.8
0.6
0.4
0.2
0.0



We see that majority of the variables are uncorrelated. This could be a reason why our PCA variables are insignificant. With this observation in mind, I decompose my PCA variables into their underlying variables and run the regression with the underlying variables. The result are as follows.

	Dependent variable:			
	expend_total	expenditures_share_income	expend_total	expenditures_share_income
	(1)	(2)	(3)	(4)
self_control	14.222***	0.433***	13.399***	0.408***
	(1.586)	(0.064)	(1.571)	(0.063)
PCA_Lottery_Understanding	0.547	-0.057		
	(1.117)	(0.045)		
PCA_Lottery_Interest	1.619	0.017		
	(1.276)	(0.052)		
financial_literacy			-24.019***	-1.213***
			(8.486)	(0.338)
financial_numeracy			-10.487	-0.852***
			(6.477)	(0.258)
gamblers_fallacy			3.065	-0.027
			(4.612)	(0.184)
non_belief_lln			7.747	0.193
			(9.885)	(0.393)
ev_miscalculation			5.618	-0.134
			(5.171)	(0.206)
risk_seeking			-1.582	-0.066
			(1.301)	(0.052)
risk_aversion			-1.456	-0.125
			(2.236)	(0.089)
seems_fun			6.448***	0.193***
			(1.176)	(0.047)
enjoy_thinking			1.397	0.012
			(1.124)	(0.045)
overconfidence			5.411*	-0.010
			(3.240)	(0.129)
happiness			0.025	-0.003
			(0.355)	(0.014)
income	0.020	-0.008***	0.054	-0.007***
	(0.037)	(0.002)	(0.037)	(0.001)
black1	24.601***	1.280***	19.847**	1.033***
	(8.378)	(0.339)	(8.243)	(0.328)
hispanic1	13.367*	0.115	8.059	-0.070
	(7.714)	(0.313)	(7.582)	(0.302)
white1	-6.289	-0.285	-3.511	-0.170
	(6.678)	(0.271)	(6.535)	(0.260)
genderMale	8.989***	0.149	13.108***	0.294**
	(3.437)	(0.139)	(3.467)	(0.138)
maritalMarried	-7.005*	-0.152	-6.215*	-0.108
	(3.786)	(0.153)	(3.678)	(0.146)
urbanNon-Metro Area	10.359**	0.141	9.724**	0.108
	(5.087)	(0.206)	(4.833)	(0.192)
employment1	-5.270	-0.634***	-6.264	-0.611***
	(3.968)	(0.161)	(3.878)	(0.154)
religion1	0.863	0.019	4.161	0.114
	(3.729)	(0.151)	(3.636)	(0.145)
education	-2.038***	-0.032	-0.844	0.016
	(0.727)	(0.029)	(0.732)	(0.029)
ideology	-0.336	-0.040	-0.136	-0.037
	(1.135)	(0.046)	(1.102)	(0.044)
not_working_age				
early_working_age	-7.928	-0.300	-2.245	-0.255
	(6.605)	(0.268)	(6.447)	(0.257)
late_working_age	12.953	0.246	17.758**	0.280
	(9.140)	(0.370)	(8.929)	(0.355)
retired	7.038	-0.312	17.672*	-0.028
	(10.031)	(0.406)	(9.847)	(0.392)
Constant	43.604	2.340	26.871	2.906
	(46.319)	(1.876)	(47.675)	(1.898)
State controls	Yes	Yes	Yes	Yes

This analysis reveals that some behavioral biases significantly influence lottery expenditures. The coefficients for self-control and financial literacy are comparable, if not greater, than those of previously mentioned significant variables for lottery expenditure, such as race (being Black), gender (being male), and being in the late working age group. Overall, the data suggest that while some behavioral biases do impact lottery expenditure, only very specific biases are significant; most behavioral biases do not have a meaningful effect on lottery expenditures.

3.3 PART III – Short Answer

3.3.1 A. Share of the total prize pool added to sub-pool by Mega Millions

Please see this link, outlining Florida's mega millions lottery for 2020 (Lottery 2020)

Match Per MM Play	Prize Category	Prize Payment	Estimated Percentage of Prize Pool Allocated to Prize Category
Five first set numbers and the one number of the second set	Jackpot Prize	Jackpot Prize	75.3018%*
Five first set numbers and none of the second set	Second Prize	$1,000,000	7.9319%
Four first set numbers and the one number of the second set	Third Prize	$10,000	1.0742%

3.3.2 B Interesting economics paper that you have recently read

An interesting paper I recently came across is *Worth Your Weight: Experimental Evidence on the Benefits of Obesity in Low-Income Countries* by Elisa Macchi (2023), published in the *American Economic Review* (Macchi 2023). I found this particularly fascinating

because it quantifies the economic benefits of obesity in credit markets within developing countries.

Coming from a developing country, I've often heard anecdotally about the association between obesity and wealth or prosperity. However, in more developed countries, obesity is frequently linked to poverty. Living in a developed country, this created a striking contrast in my mind, on the economic structures that lead to this observation.

What stood out to me in this paper was the use of face morphing technology, which I believe represents a relatively new and unique tool in economic research. The study demonstrated that obesity acts as a beneficial signal only in markets where other indicators of wealth or creditworthiness are absent. In such environments, obesity or luxury goods serve as signals to facilitate credit access.

This led me to reflect on the role of wealth signals in both developing and developed countries. Why do these signals work effectively in some settings but not in others? Wouldn't similar mechanisms apply in developed countries as well? If not, why are obesity and luxury goods less prevalent as signals in developed countries compared to developing ones? Is the difference primarily cultural or institutional? I would be very interested in seeing a similar study conducted in developed countries to determine whether these results align across contexts.

References

Co, W. (2025*a*), 'Coding-Task-GuZi/GZ_RA_StataTaskDescription.pdf at main · WilliamClintC/Coding-Task-GuZi', https://github.com/WilliamClintC/Coding-Task-GuZi/blob/main/GZ_RA_StataTaskDescription.pdf.

Co, W. (2025*b*), 'Coding-Task-GuZi/lottery_study/code at main · WilliamClintC/Coding-Task-GuZi', https://github.com/WilliamClintC/Coding-Task-GuZi/tree/main/lottery_study/code.

Lottery, F. (2020), 'MEGA MILLIONS Game Rules', https://files.floridalottery.com/exptkt/megaMillions-GameRules.pdf.

Macchi, E. (2023), 'Worth Your Weight: Experimental Evidence on the Benefits of Obesity in Low-Income Countries', *American Economic Review* **113**(9), 2287–2322.