FinalProj_BStriano

Brendan M. Striano

Before we dive into the analytical question of interest, let's bring in the packages that we will use later on in this code.

Introduction:

Question of Interest

In the course of this analysis, we will be exploring idea and motivation of traffic violations and parking tickets. Penalties of this type are one mechanism through which municipalities attempt to manage the behavior of its citizens. The concept behind these fines is that when citizens violate laws, a fine is levied in order to help discourage violating the rules.

In the idealized version of this system of municipalities encouraging legal behavior, traffic violations and parking tickets are given out in every scenario where an individual has violated the law, irrespective of other circumstantial considerations. This idealized conceptualization is ostensibly the most fair scenario.

However, guiding citizen behavior is not the only potential motivation for levying tickets and violations. These fines are also an important income stream that municipalities utilize in order to maintain a balanced budget. This dual functionality of tickets and violations is where some citizens express concerns.

The main form of this concern is the idea that there may be certain scenarios wherein the function of balancing a budget takes precedent over the functionality of guiding citizen behavior. Citizens who express this concern often remark that when it comes time for budgets to be balanced (e.g. -at the end of a month), citizens are more likely to be cited with parking tickets and traffic violations.

This concern can be assessed with data, but unfortunately that data is unrealistically difficult to obtain. To do this, one would need to have data on every time that a citizen is committing an illegal behavior (e.g. -parking illegally), know whether he/she was penalized for that behavior, and whether this act occurred at a time point where a concerned citizen would expect increased rates of fines being levied. In the absence of the ability to generate this ideal data, we will make use of the best available data to perform an analysis to approximate whether there is a difference in the amount of parking tickets given out based on the time of the month.

This analysis will take advantage of publicly available data from the city of Philadelphia, Pennsylvania. The data is made available through opendataphilly.org (1,2), which is a catalog of open access data of many types from the Philadelphia area and is referenced at the end of this document. One of the major benefits of this data source is that it is quite large and therefore, for ease of conducting this preliminary analysis, we will utilize data just from a single year.

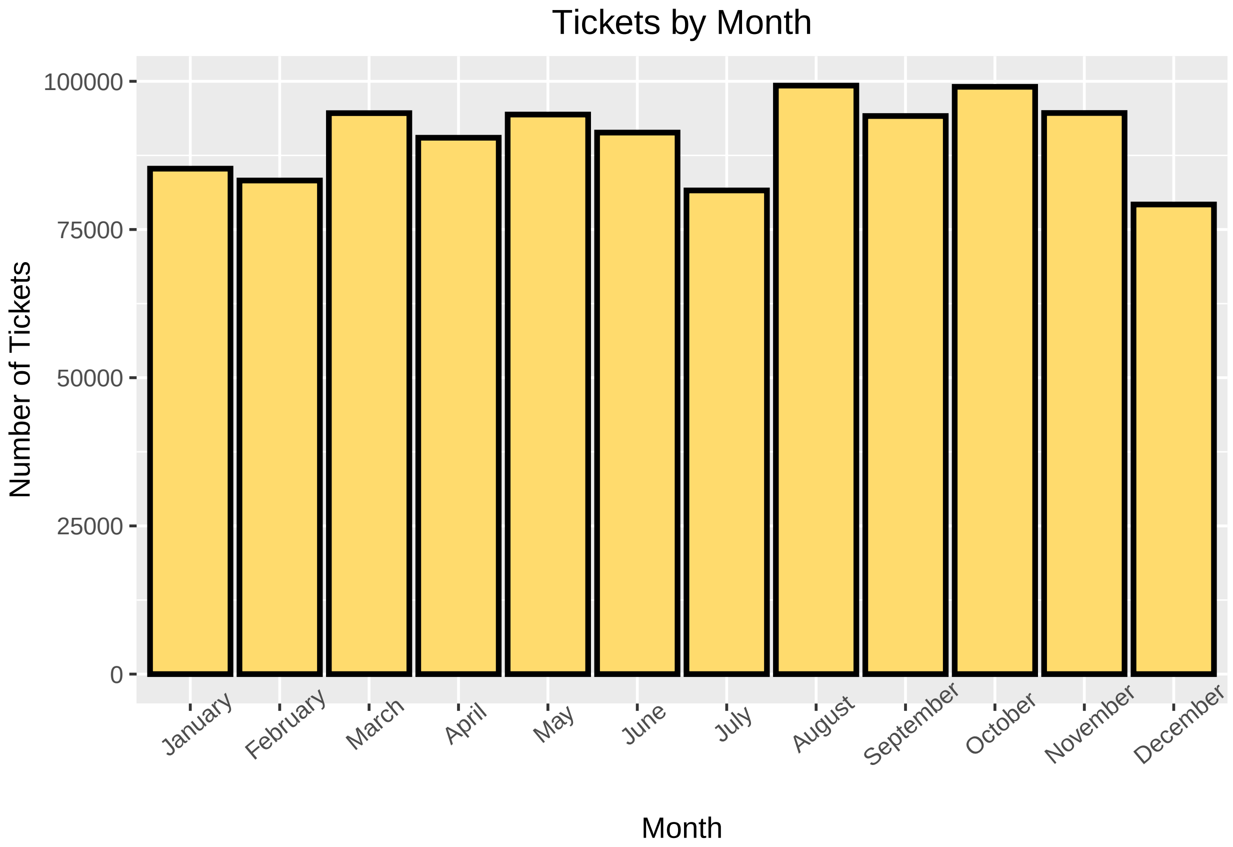
To analyze this data critically, we will begin with simple tests of association to assess the relationship between the number and total fines of tickets and the time of the month. Later, we will perform a regression in order to see if there is a relationship between tickets or fines and time of the month when controlling for additional variables, such as month of the year.

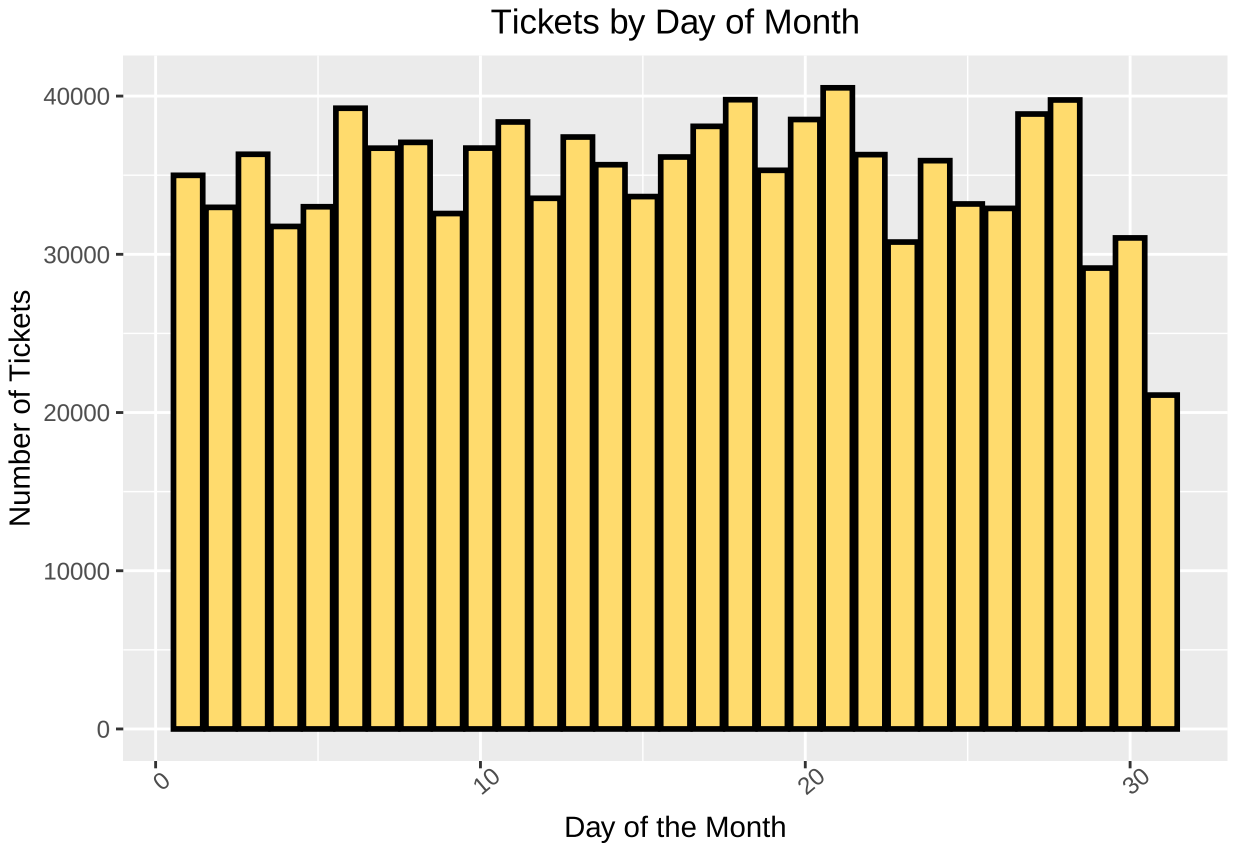
To make this analysis feasible, we will constrain the data to tickets dispensed in the first and last week of each month. This will allow for maximum contrast between the hypothetical financial "motivation" to dispense a ticket. Because we will be looking at the number of tickets and the total dollar amount of tickets given out, we will utilize t-tests and linear regression. The data set is sufficiently large to utilize parametric statistics (3).

Exploratory Data Analysis Initial exploration of the data demonstrates that we have 7 columns the names of which are fairly self-explanatory. The format of the data is such that each row is a ticket and the columns are the information for that individual ticket. The columns of the unmodified data are:

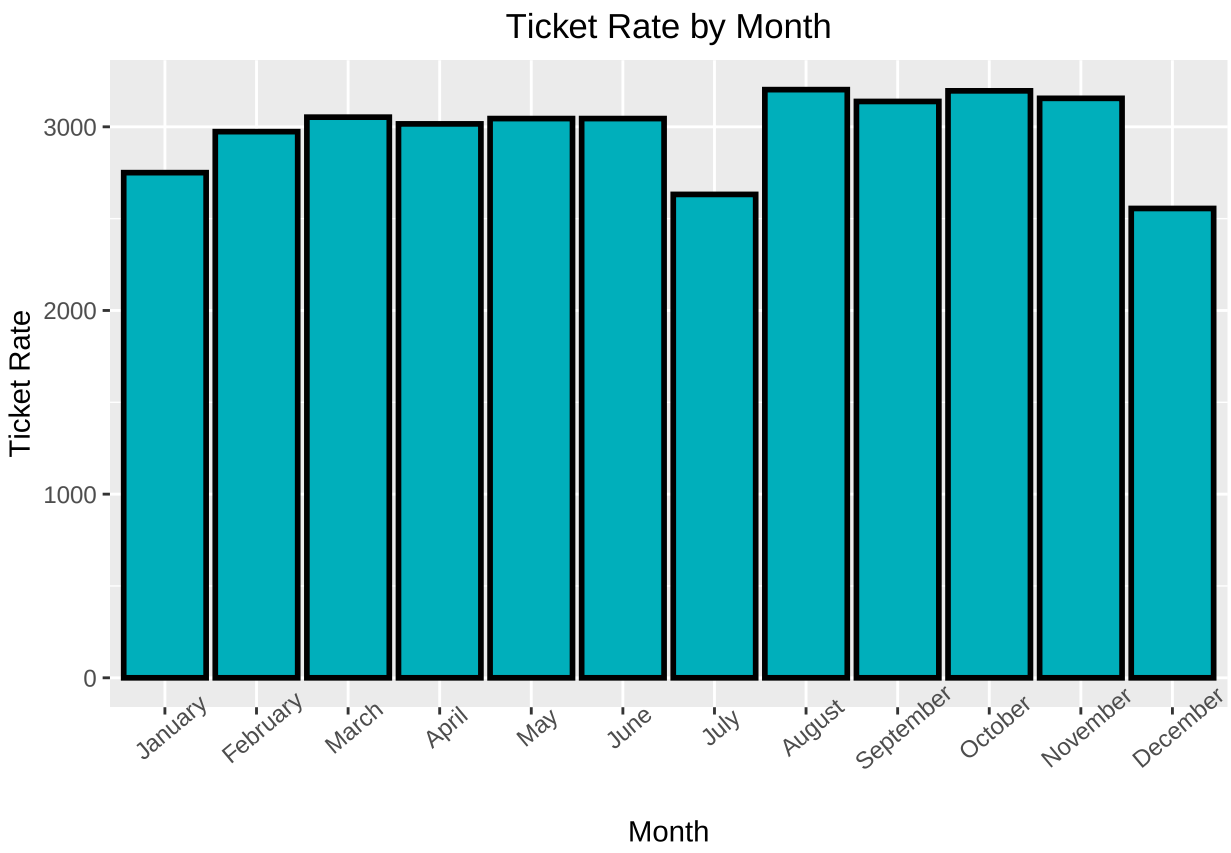
1) The type of violation for which the ticket was given
2) The date and time at which ticket was given
3) The dollar amount of the fine for the ticket
4) The agency that gave out the ticket
5) The geographic latitude of the violation
6) the geographic longitude of the violation
7) The zip code in which the infraction occurred

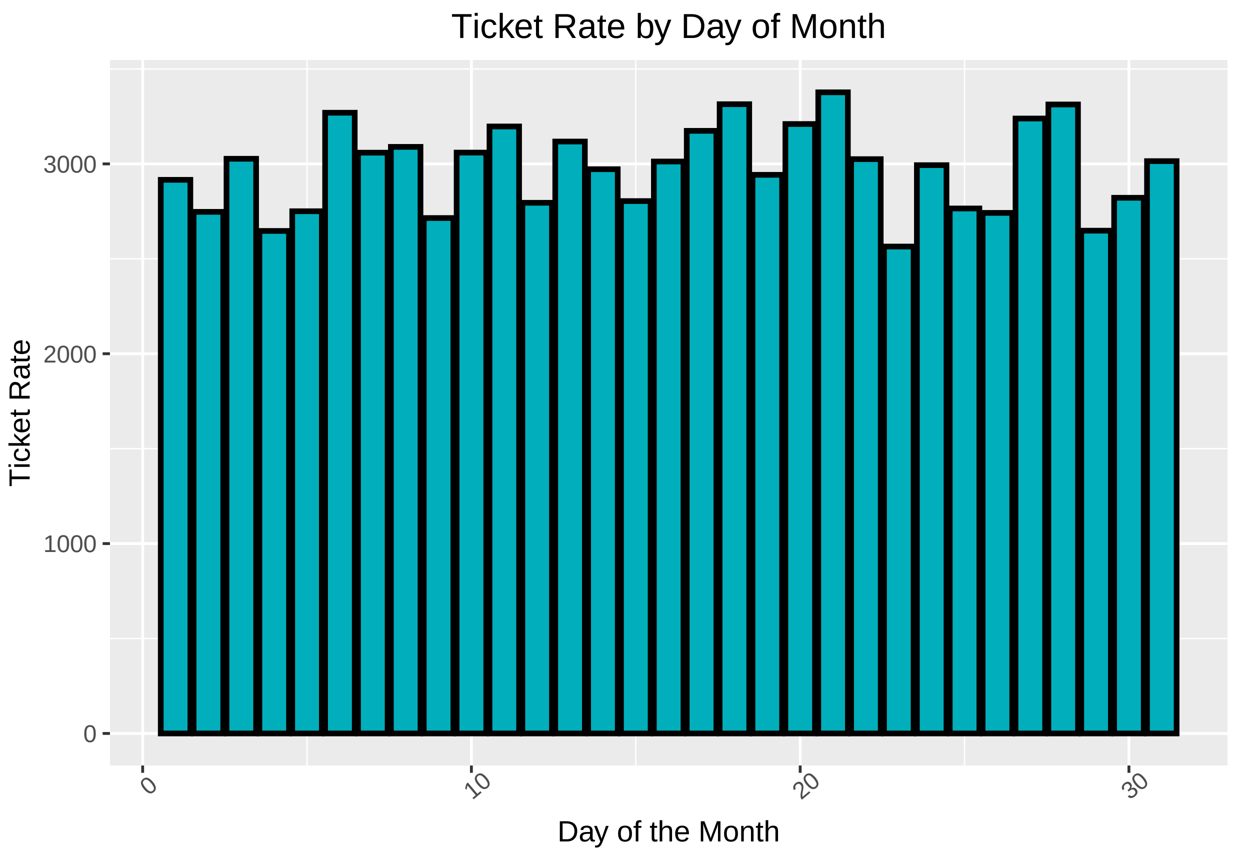
After inspecting the data for missing values/blanks, NAs, outlier data points, reassigning data types, managing dates, and transforming data where necessary, we have an appropriately cleaned data set that will allow us to begin to visualize the data.



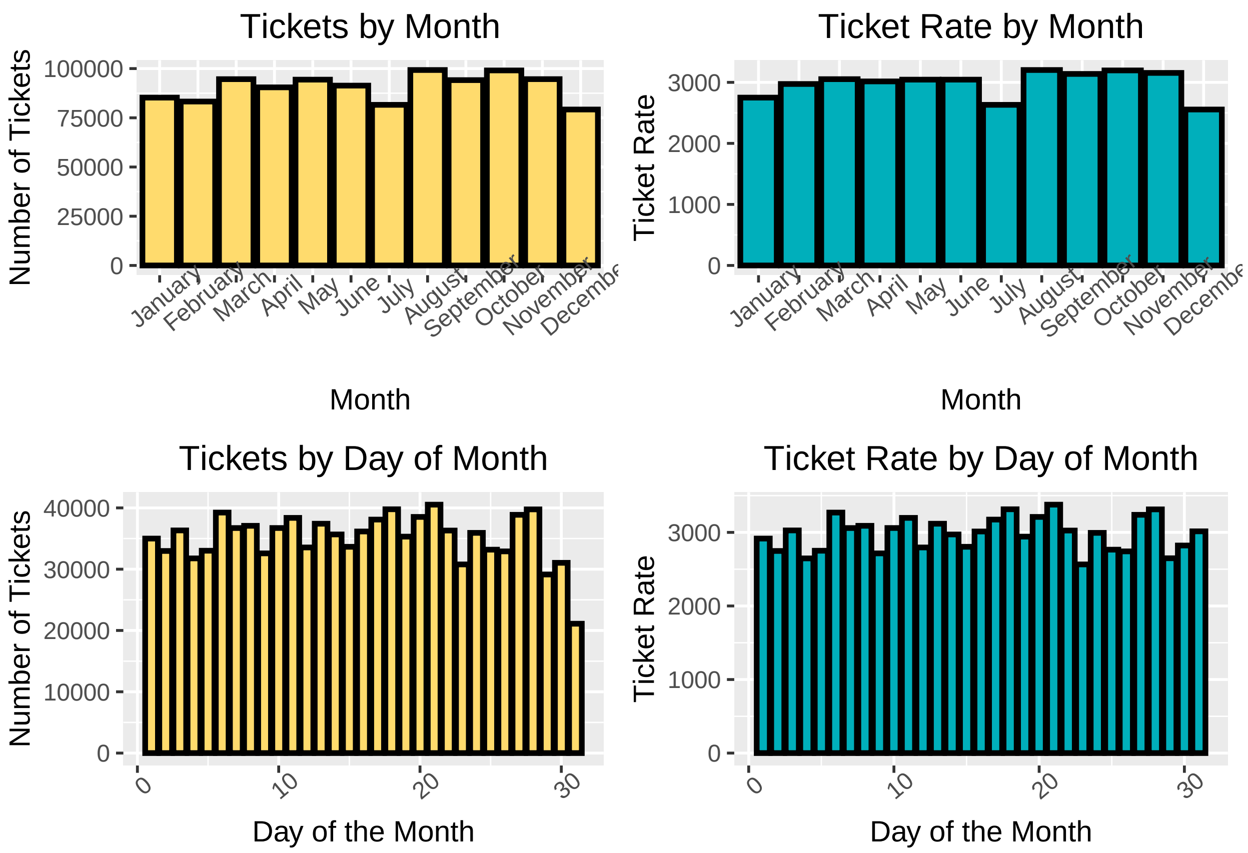


This initial visualization suggests that there are a fairly consistent number of tickets given out each month and at each day of the month. These are only preliminary representations, so these figures do not address the fact that there are different numbers of days in each month, or different representations of the different days of the month.


Ticket Rate by Month
Ticket Rate
3000
2000
1000
0
January
February
March
April
May
June
July
August
September
October
November
December
Month

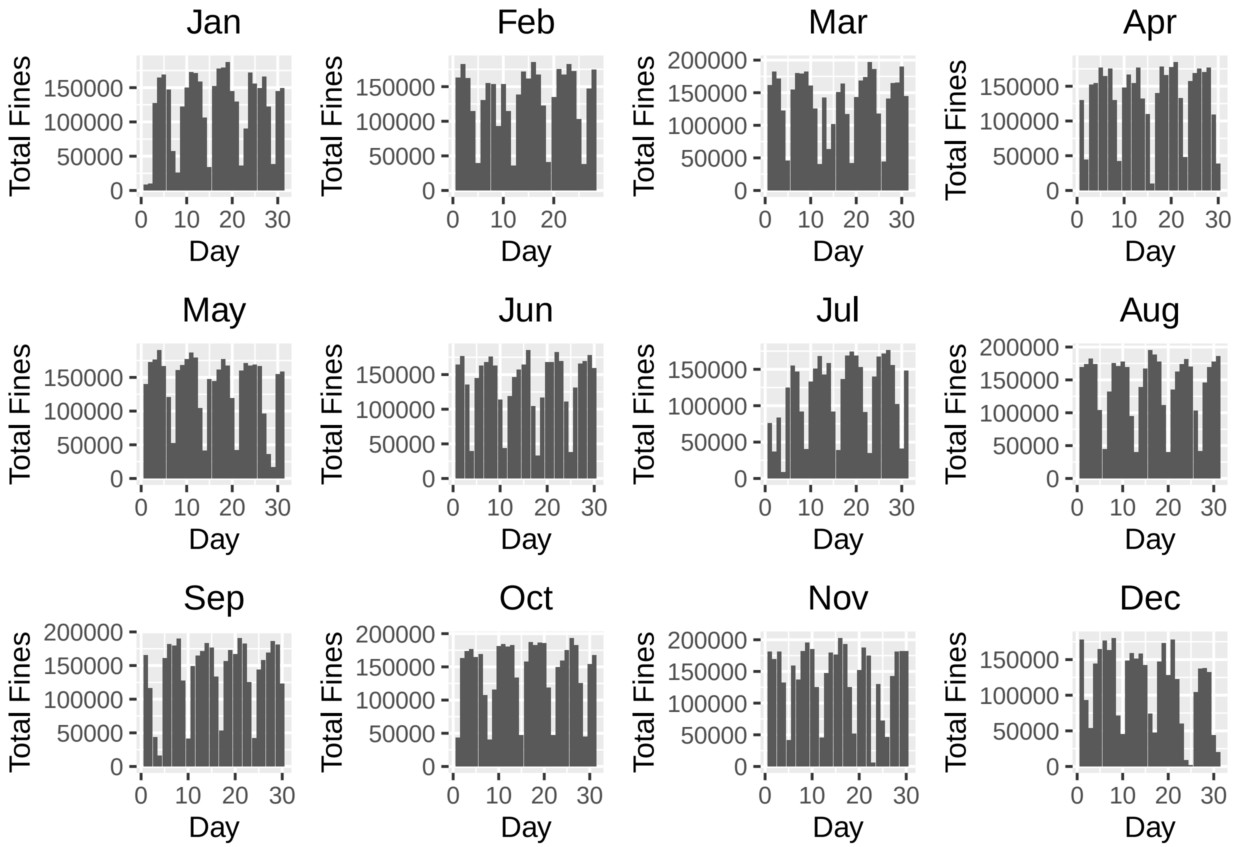


In order to account for the aforementioned differences, here, the data is presented in rates so they columns are therefore all scaled appropriately. With this adjustment, it still seems that there is no difference in how many tickets are being given out across the course of the year or the course of the month.



For visual comparison purposes, this paneled figure places the unadjusted figures alongside the adjusted figures and again we see that there does not appear to be a difference, visually.

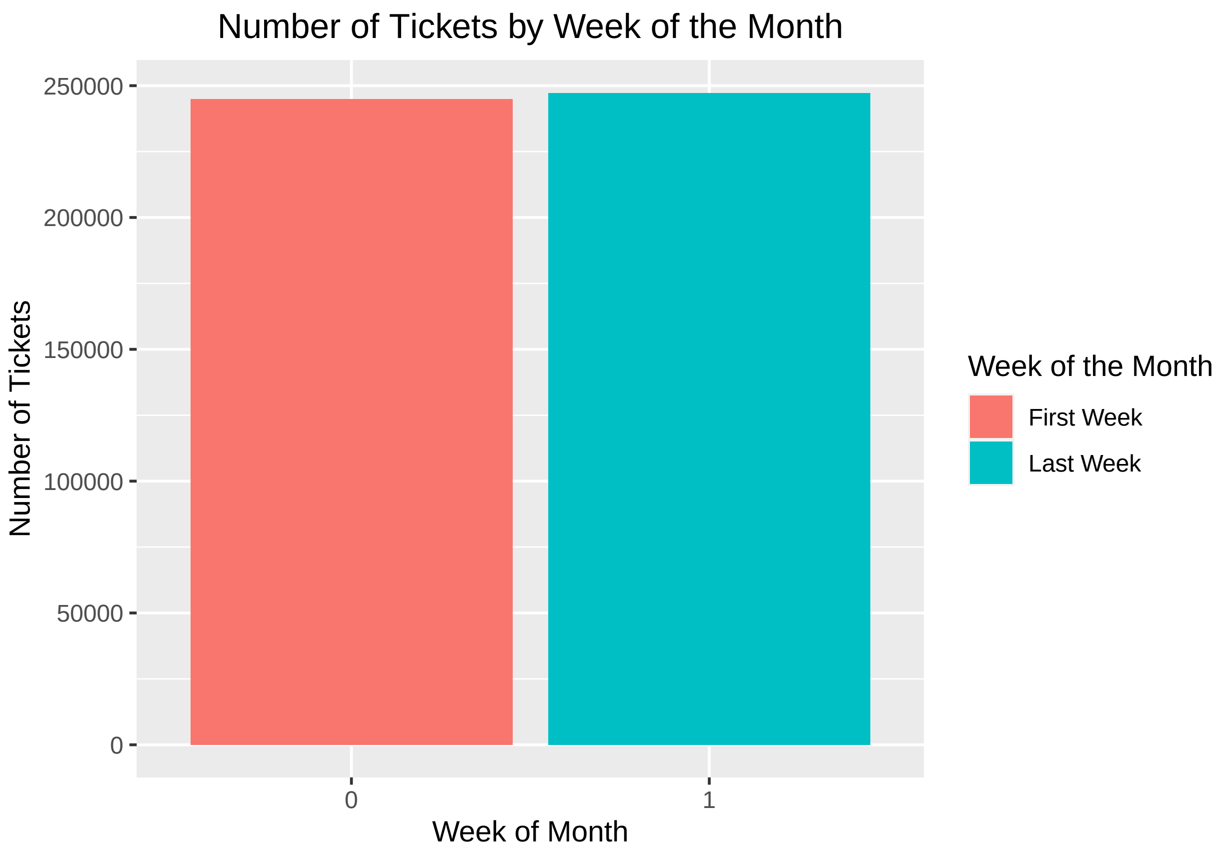
The prior data presented the number of tickets, but it made no exploration of the amount of dollars levied by the tickets. It is possible that in order to help balance the budget, municipalities are not increasing the number or rate of tickets being given out, but rather increasingly giving more expensive tickets. This would be one way to generate increased revenue around times to balance budgets.

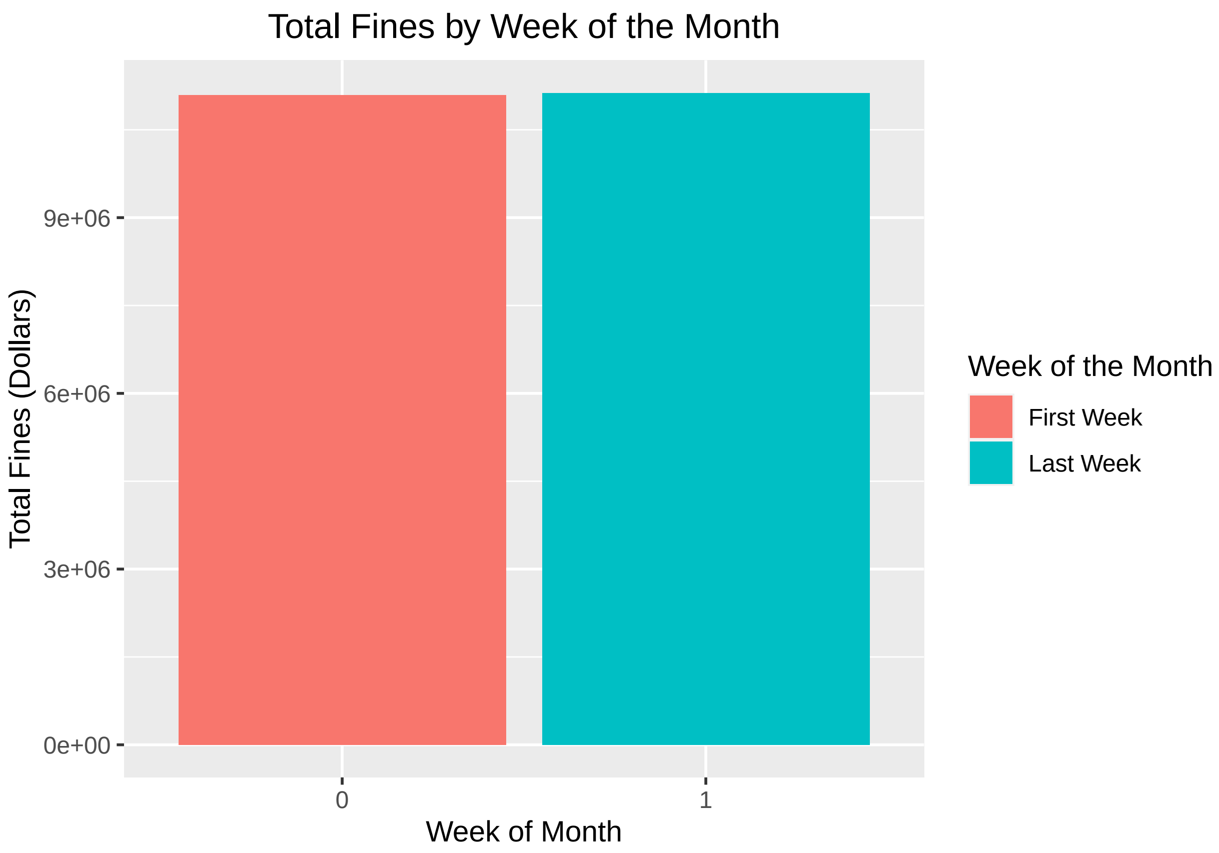


With this plot, we see that there does not seem to be a consistent increasing trend throughout the course of the month as one would suspect if municipalities were attempting to balance a budget with more expensive fines at the end of the month. We do see that there appear to be some trends along the course of an individual week, but we will not be exploring them here.

RESULTS

The above visualizations were helpful to get a general lay of the data, but they do not specifically address the main analytic question: is there a different in the number of tickets or the dollars of fines dispensed in the first v. last month of the year. Those data can be seen below.


Number of Tickets by Week of the Month
Number of Tickets
250000
200000
150000
100000
50000
0
0
1
Week of Month
Week of the Month
First Week
Last Week



Here, we see that there is almost exact parity in the number of tickets given out in the first week and last week of the month. Similarly, the revenue generated in the first week appears nearly identical to the revenue generated in the last week of the month. This suggests that there is no difference between the two different time point. To go beyond visual inspection, we will quantify the relationship with statistical testing first in the form of a t-test.

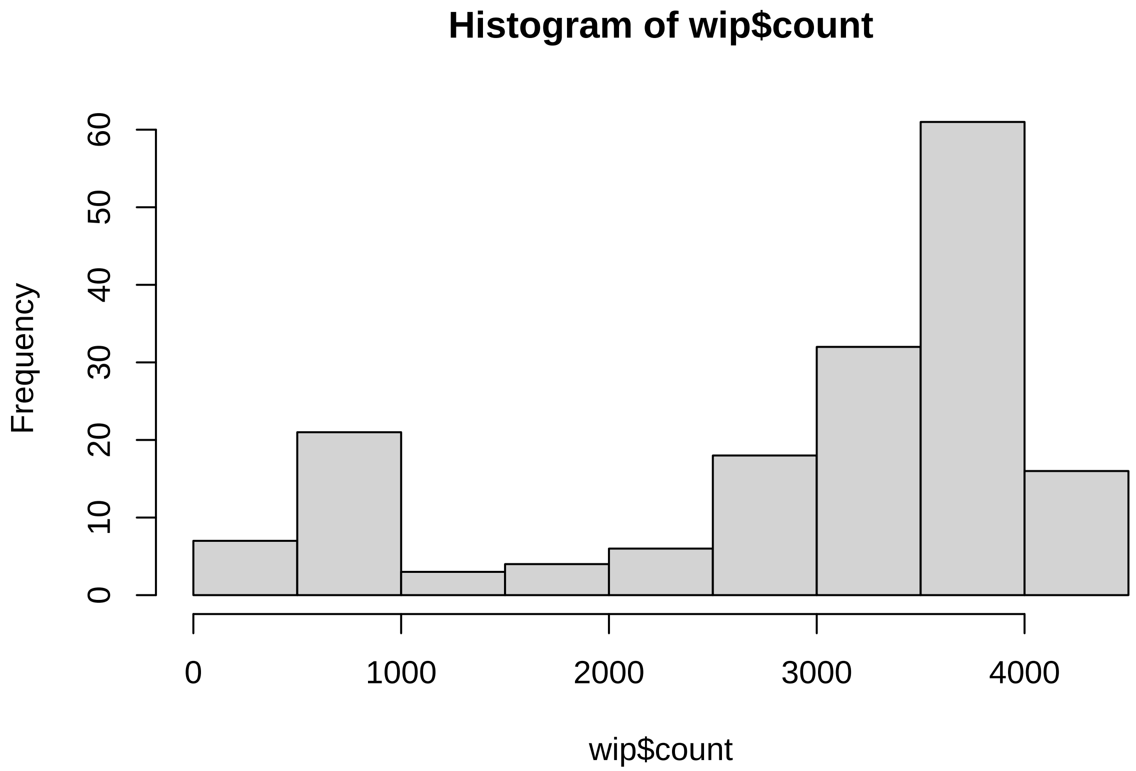
Observation of the output from the ttest on the count data demonstrates the mean number of tickets in the first week of the month was 2916.5 compared to 2943.8 in the last week of the month. This indicated that on average, there were about 43 more tickets given in the last week of the month, compared to the first week of the month. However, this difference is far from significantly different with a p value of 0.88 with a range of possible values (i.e. -95% CI) from -385 tickets to 331 tickets. Therefore, in this unadjusted analysis, there is no difference in the number of tickets given out based on the month of the year.

With regard to the financial data, the daily mean dollar amount given out in the last week of the month was 132510.90 dollars compared to 132102.30 dollars in the first week of the month. This difference of408.60 dollars was not significantly different with a p value of 0.95 and a 95% confidence interval from -16298.92 to 15481.90. Because this difference crosses a difference of 0 dollars, we conclude that there is no consistent difference in the dollar amount of tickets given out on the basis of the week of the month.

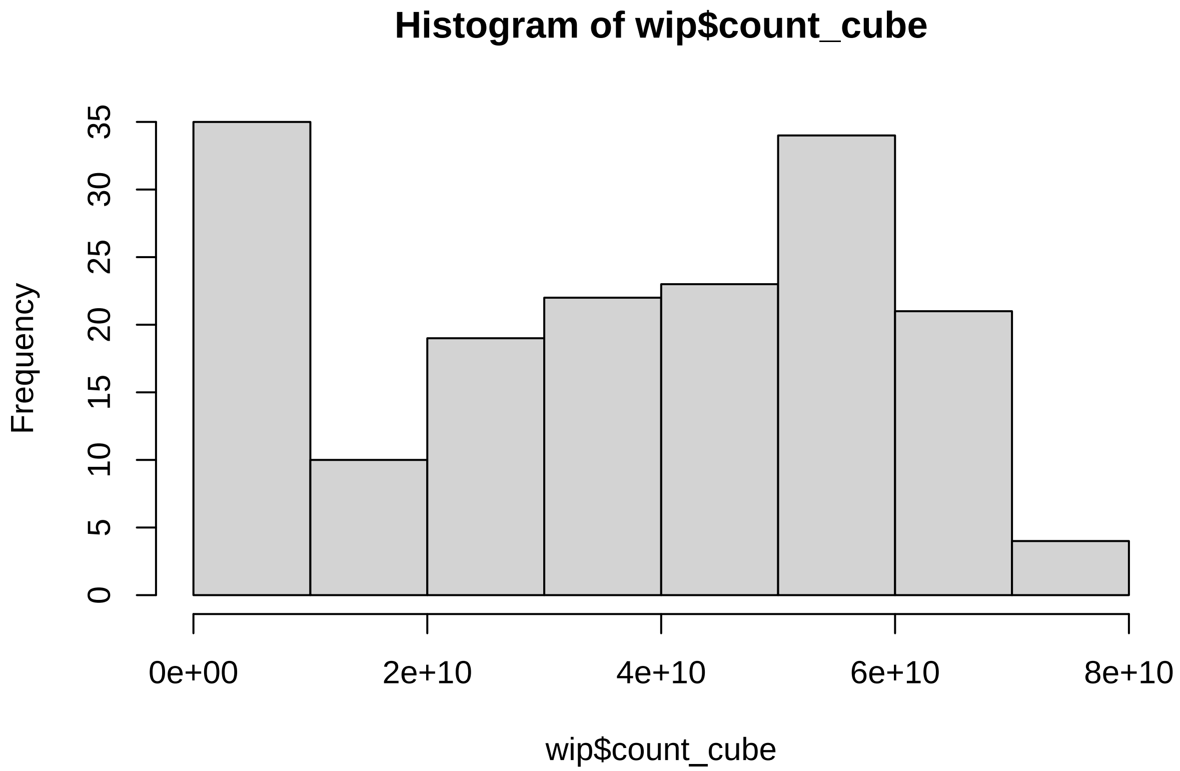
These statistical tests are consistent with our graphical representations which suggested that we would likely not find a difference in either the number of tickets or the total dollar amount of fines levied.

Now that we have assessed our outcomes of interest with bivariate statistics, we will move onto the regression analysis. In this linear regression, we will assess the number of tickets as a function of the week of the month (first or last) while controlling for the month of the year.

```
hist(wip$count)
```



```
hist(wip$count_cube)
```



Running the linear regression for the number of tickets dispensed, we see that when controlling for the month of the year, switching from the first week of the month to the last week of the month only imparts an expected difference of 27.3 tickets. More importantly, the p-value for this association is not significant at p=0.882.

It should be noted that when observing the distribution of the ticket data, it was appreciated that the data were left skewed and therefore, a cubic transformation was perfomed, which did reduce the skewness of the data. However, despite the transformation of the data, the effect of the week of the month was still not signifcant at p=0.879.

Similarly, when evaluating fines based on the week of the month, controlling for the month of the year, we see that there is a difference of 408 dollars. Again, this difference is not statistically significant p=0.96.

As with the ticket count data, a cubic transformation was performed on the fine data to address the leftward skew. Despite improvement in the visual distribution of the data, as assessed via histogram, there was no change in the linear regression with no significant effect of the primary predictor with p=0.987.

Conclusion

In conclusion, during the course of this analysis, we explored the assertion that at different times, municipalities selective dispense tickets with financial motivation. Because it is not possible to directly answer this exact question, we analyzed proxy questions: 1) Do the number of tickets vary depending on the week of the month? and 2) Does the amount of money collected in fined vary depending on the week of the month? We first explored this visually with a collection of crude graphs and then with graphs adjusted for differences in the length of months and the occurrences of different days. This required considerable data manipulation and tidying in order to generate the appropriate data format. Visually, it did not appear that there was any trend wither across different months or within months. We explored further by directly graphing the counts of tickets and total fines levied in the first v. last weeks of months. Again, this appeared to not have a difference.

In order to provided statistical comparisons concordant with these graphs, t-tests were performed on both the ticket and fine data demonstrating no significant difference. These bivariate analyses do not adjust for other factors, such as the time of the year. That is to say that it is possible that there was some effect of the week of the month that was being hidden by the effect of different months of the year. In order to evaluate this possibility, linear regression on both the ticket number and fine data were run with the week of the month used as the primary predictor and controlling for differences in the month of the year. Again, these analyses demonstrated that there was no significant effect of week of the month on the number of tickets of total amount of fines. Even when cubic transformations were performed to address the leftward skew of the data, there was no difference in the result.

As best as possible with the available data, these graphical representations, bivariate statistics, and multivariable statistics demonstrate that there does not appear to be any financial motivation to the way in which parking violations are dispensed in the city of Philadelphia in 2017.

Future analyses could include additional covariates in order to attempt to better control for additional confounders. Further, this analysis could be enhanced by the addition of data from other years to assess beyond just 2017. Finally, this question could be assayed in a different fashion by employing machine learning and attempting to see if with different algorithms, the machine can be trained to identify whether a ticket was given in the first or last week of the month. If these initial analyses are correct, I would expect that the model would not be able to distinguish a first week ticket from a last week ticket with any reliability.

References

1) Data source: https://www.opendataphilly.org/dataset/parking-violations
2) Additional sources: https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-12-03
3) Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health. 2002;23:151-69. doi: 10.1146/annurev.publhealth.23.100901.140546. Epub 2001 Oct 25. PMID: 11910059.