Analysis of the San Fransico TOLE Program

Michael Arango, Yeshwant Chillakuru, Jackson Crum

March 6, 2017

Background Information on Dataset

General Information and Data Collection Process

The Transit-Only Lane Enforcement (TOLE), run by the San Fransico Municipal Transportation Agency (SFMTA) from March 2008 to February 2015, is a pilot program to reduce obstruction of transit-only lanes through increased enforcement. Currently, private and commerical vehicles that drive and park in transit-only lanes can cause delays in bus service, increases in traffic congestion, and inconveniences to customers.

Under the TOLE program, all public buses are equipped with front-facing video cameras to capture the license plates of vehicles parked or stopped in a dedicated transit-lane. The vehicle's registered owner will then recieve the citation in the mail as a means to deter future transit-only lane violations. These citations and their accompanying details form the dataset we are examining.[1]

San Fransico currently operates 26 miles of transit-only lanes with 160,000 daily passsengers and plans future expansion of 22 miles, so the TOLE program has strong potential for long-term impact by increasing the efficiency of the municipal transit system and potentially increasing municipal revenue temporarily as drivers adjust to the enforcement scheme.[2]

Description of Dataset

The unit of observation was an individual citation for a transit-only lane violation reported by municipal drivers under the TOLE program, and the raw dataset contains the following variables regarding the transit-only lane violations:

- `Object.ID` – Unique key that corresponds to each observation
- `Ticket.Number` – Unique number that corresponds to each ticket cataloged with the Superior Court of San Francisco
- `Citation.Issue.Date` – The date (MM/DD/YY) that the traffic only lane violation citation was issued
- `Citation.Issue.Month` – The month (MM) that the traffic only lane violation citation was issued; This column was completely blank when downloaded
- `Citation.Issue.Time` – The time of day (HH:MM) PST that the traffic only lane violation citation was issued
- `Location` – The street address where the traffic only lane violation citation was issued
- `Violation.Code` – A code that corresponds to the type of traffic only lane violation that was committed to warrant issuing a citation
- `Violations` – Description of the traffic only lane violation issued in the citation
- `Fine.Amount` – The dollar amount of to be paid for the traffic only lane violation. Mean: 112.33 USD.
- `Citation.Status` – Status of the citation (Open, Closed, Unapplied)
- `Amount.Paid` – The dollar amount paid for the traffic only lane violation. Mean: 89.59 USD.
- `Amount.Due` – Outstanding balance for the traffic only lane violation. Mean: 19.02 USD.
- `Suspend.Code` – Code that represents reason for license suspension (Suspended until the driver goes to driving school, pays fine, too many points on license, etc.)

[1] Data retrieved from Data.gov (https://catalog.data.gov/dataset/muni-driver-reported-transit-only-lane-violations)

[2] San Francisco Municipal Transportation Agency, "Transit Only Lane Enforcement," SFMTA, accessed on February 24, 2017, https://www.sfmta.com/sites/default/files/projects/2015/TOLE%20one%20pager_v3.pdf

- `Suspend.Process.Date` – Date the license suspension goes into effect
- `Suspend.Until.Date` – Date the license suspension is over
- `Disposition.Code` – The city of San Francisco does not release disposition codes
- `Last.Edited.Date` – Date that represents the last date the corresponding observation was edited; This column was completely blank when downloaded
- `Geom` – The latitude and longitude (latitude, longitude) of the traffic only lane violation

Limitations of Dataset and Data Cleaning

The limitations of the data include the timing of the implementation of the video camera citation system. While SFMTA claims 100% of their buses are equipped with cameras, they fail to indicate if the 100% implemention was at the start or end of the program. We believe there was a trial period at the beginning, so we decided to drop variables in 2008 when the program first started because there were only 154 citations issued that year, but typically several thousand citations issued in the following years. In addition our dataset lacked demographic controls for the regions where citations were issued. This means demographic factors could cause unobserved biases in our data. Similarly to the domestic controls, the dataset lacks observations of citations issued before the program was put in place, which prevents a before and after analysis. While out of the scope of this initial analysis, future analysis can map to the `Geom` variable in the dataset to census tracts to derive demographic information and could be used to compare to other citations.

Some observations were missing key data. We decided to drop all observations with no `Geom` data, so that future geospatial and demographic analysis could be done (21 observations). We dropped all observations where `Citation.Status` was not specificied or "unapplied", as we determined "unapplied" was out of the scope of this project (84 observations). We dropped `Violations` where the listed violation was determined to be "No Violation" (1 observation) or occured less than or equal to 70 times (77 observations). We also dropped citations in the year 2015 to avoid biases in monthly and yearly analysis because only 2 months were included for 2015.

Overall, we dropped 374 of the original 17,178 observations in the dataset. This only accounts for 2.18 percent of the dataset. Thus, we are confident that we did not introduce any inherent biases to the data by dropping the aforementioned observations.

Literature Review

Most of the the analyses found in the literature compared TOLE data to traffic flow rates. According to SFMTA, certain corridors experienced reductions in delays after TOLE implemention. Westbound Geary Street corridor saw a 3% daily and 7% afternoon reduction in delays. Westbound Stutter Street saw a 15% daily and 20% afternoon reduction in delays. However, this analysis lacks significant meaning because it cherry picks certain areas and also lacks a control comparison for other areas that may not have been in the TOLE program. Additionally, their analysis has some clear flaws as they state that of the 1700 different vehicles that recieved a citation from March to August of 2012, only 2% recieved a citation during the "same three-month period in 2013." Again this analysis lacks a control comparison and suffers a clear inconsistancy as March to August is a six-month period—not three month as SFMTA stated.[3]

According to Michael Rhodes, San Francisco's transit-only lane violations are generally unenforced. While SFMTA had not released citation data prior to Rhodes' article in 2009, he cites anecdotal evidence from interviews with municipal drivers and San Francisco Police Department. Based on the municipality's weak analysis and the anecdotal evidence provided by Rhodes of the need for an effective transit-only lane violation enforcement scheme, a further analysis of municipality data must be conducted. While we cannot analyze a before and after comparison of the TOLE program on transit-only lane violations or traffic flow improvement,

[3]San Francisco Municipal Transportation Agency, "Transit Only Lane Enforcement," SFMTA, accessed on February 24, 2017, https://www.sfmta.com/sites/default/files/projects/2015/TOLE%20one%20pager_v3.pdf

we can examine the distribution of violations by violation type, time of day, day of the week, and revenue potential.[4]

Broad Question

The TOLE program was implemented in 2008 as a pilot set to expire in 2012, unless extended upon an evaluation in 2011. In 2011, the City of San Francisco voted to implement the TOLE program permanently. Therefore, understanding the effects and dimensions of the program are vital to constantly improving it. We believed an exploratory analysis that at revenue generation would be an ideal means of understanding this program. Revenue optimization both increases funding for the state and creates economic incentive for drivers' to avoid citations. Under properly placed incentives and fine structures, revenue optimization can lead to ideal traffic outcomes.[5]

Broad Question: *Under what paramters does the TOLE program optimize TOLE program revenues?*

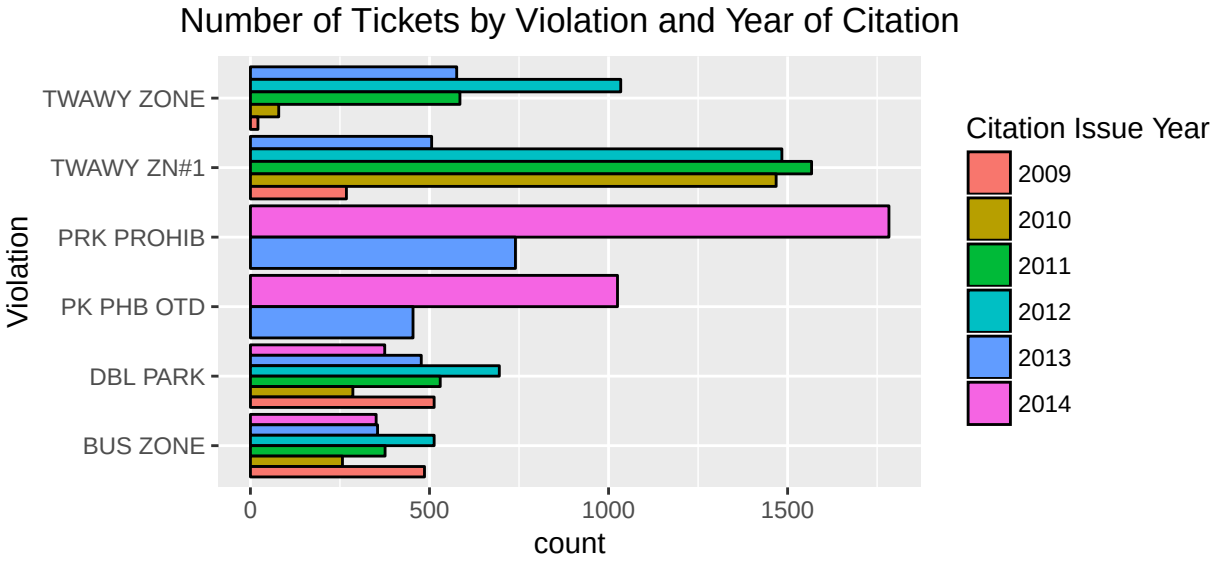
We started with this broad question before choosing a fully established SMART (Specific, Measurable, Answerable, Relevant, Time bound) question to allow EDA to inform the development of the SMART question.

EDA & SMART Question

We first examined basic summary statistics involving our dataset. Our cleaned dataset had 16804 observations (i.e. citations) and 16 variables. There were 6 different types of violations:

- 'BUS ZONE' – Parking in Bus Zone – 2338 citations
- 'DBL Park' – Double Parking – 2876 citations
- 'PK PHB OTD' – Tow-Away Zone - Outside Downtown Core – 1479 citations
- 'PRK PROHIB' – Tow-Away Zone - Downtown Core – 2523 citations
- 'TWAWY ZN#1' – Towaway Zone 1 and Towaway Zone – 5293 citations
- 'TWAWY ZONE' – Towaway Zone – 2295 citations

We then examined if violations vary by year, and saw that `PRK PROHIB` and `PK PHB OTD` occur only after 2013, while `TWAWY ZN#1` and `TWAWY ZONE` violations were issued only until 2013.

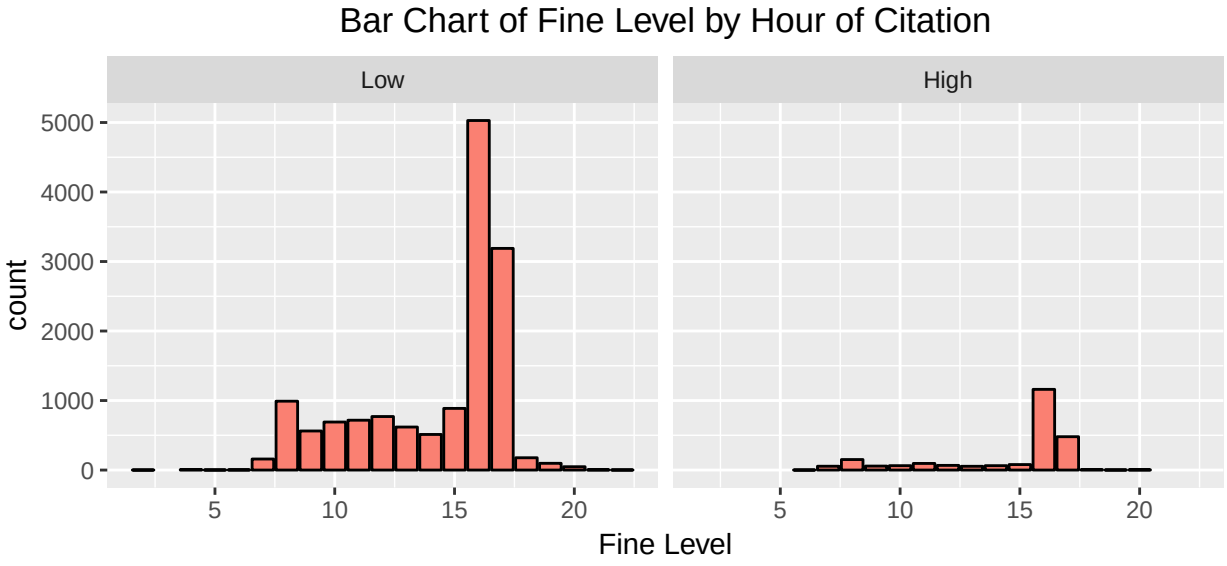


[4]Michael Rhodes, "Violations in SF's Transit-Only Lanes Rampant and Rarely Enforced," StreetsBlogSF, August 11, 2009, http://sf.streetsblog.org/2009/08/11/violations-in-sfs-transit-only-lanes-rampant-and-rarely-enforced/

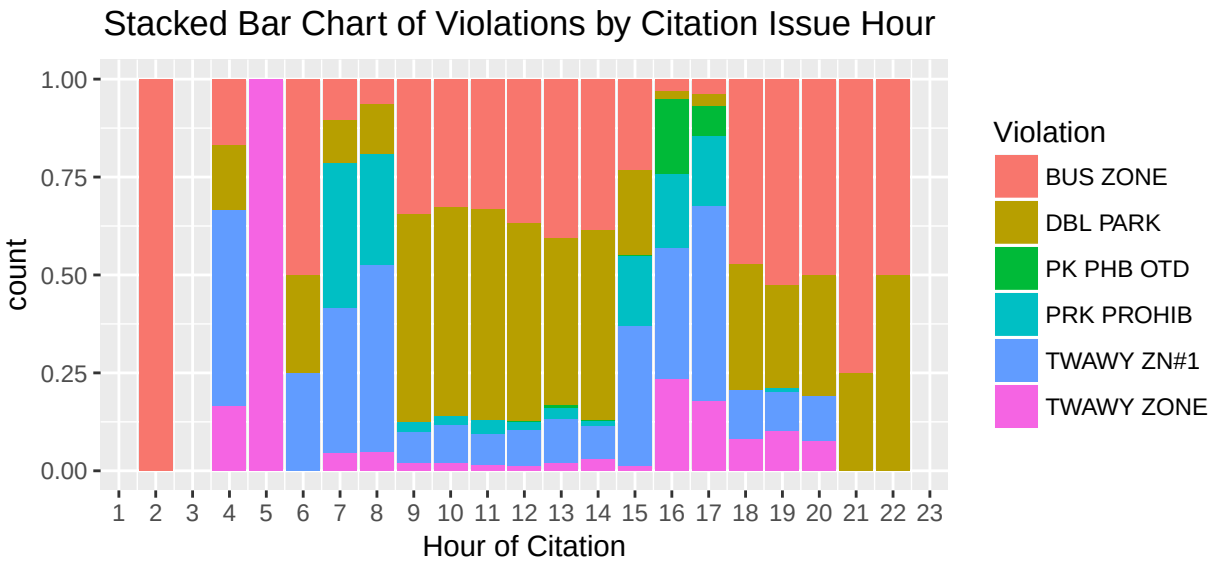
[5]Official California Legislative Information, "AB 1041 Assembly Bill—Bill Analysis," Assembly Third Reading, April 27, 2011, http://www.leginfo.ca.gov/pub/11-12/bill/asm/ab_1001-1050/ab_1041_cfa_20110429_113836_asm_floor.html

We decided to examine how fine amount varies by type of violation and discovered a bimodal distribution with a "high" and "low" level fine for each type of violation.

However, upon further examination, we saw there was no obvious difference in `Fine.Amount` or the `Fine.Level` when controlling for other factors like time of day.

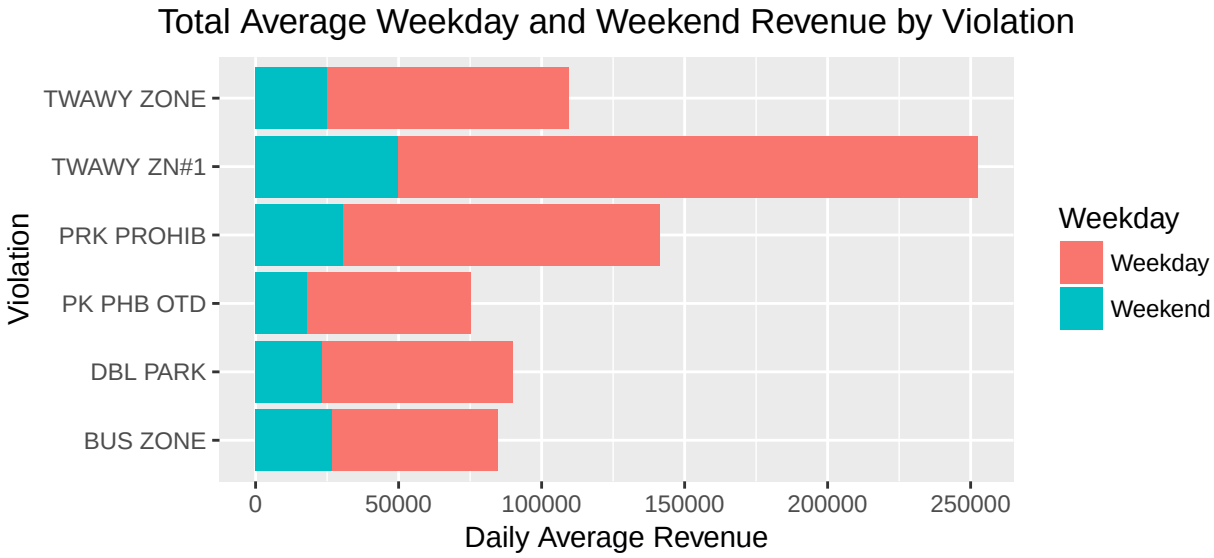


We then began to shift our analysis towards looking at the interaction between time and revenue generated by fines. We looked at the distribution of different citations over time in a stacked bar chart:



This graph shows that Double Parking (`DBL PARK`) and Bus Zone (`BUS ZONE`) Violations compromise most non-rush hour violations, while Towaway Zone (`TWAWY ZN#1` and `TWAWY ZONE`) and Downtown Towaway Zone (`PRK PROHIB`) Violations make up most of rush hour violations.

We further investigated which violations produce the most revenue. On a daily basis, Towaway Zone 1 produces the most revenue. even when broken down by weekdays and weekends.



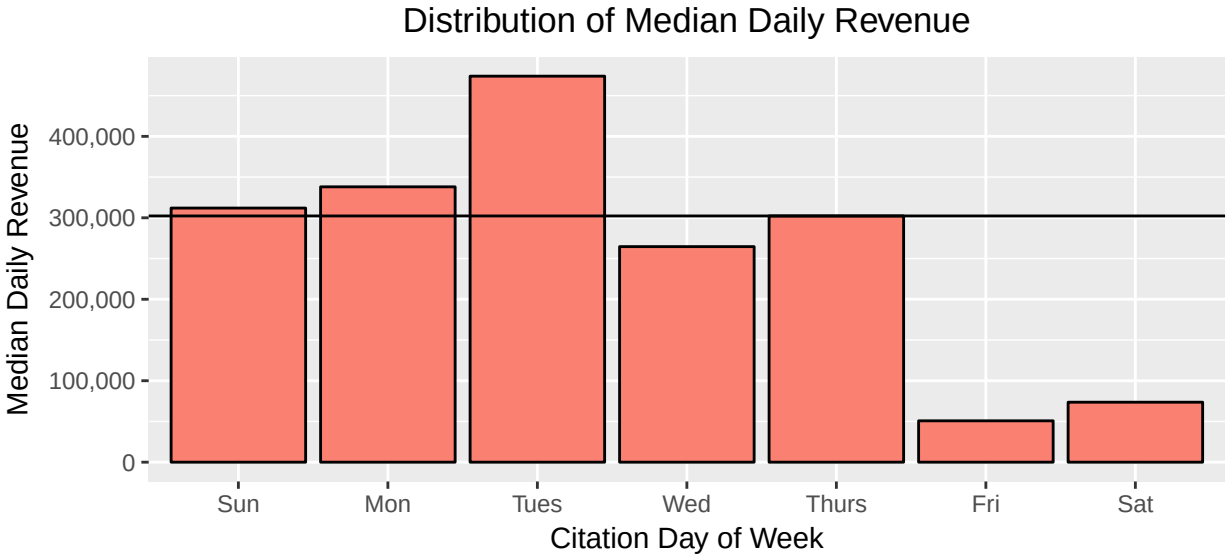
We saw that there was potential for certain days of the week to generate more revenue than others so we decided to focus on how revenue can vary based on each day of the week. Understanding the deviation of median revenue for weekdays from the median daily revenue is an ideal starting point for further analysis. If deviation is present, then future work could examine variation based on violation type, time of day, rush hour vs non-rush hour, etc.. Additionally, using median is justified to account for the biasing effect of the uneven bimodal distribution of fine amounts (the high and low fines) and other outliers. Ideally, if the dataset could be connected to traffic flow/congestion rate data, then this analysis could be examined to better understand the impact of this program on actual traffic flows.

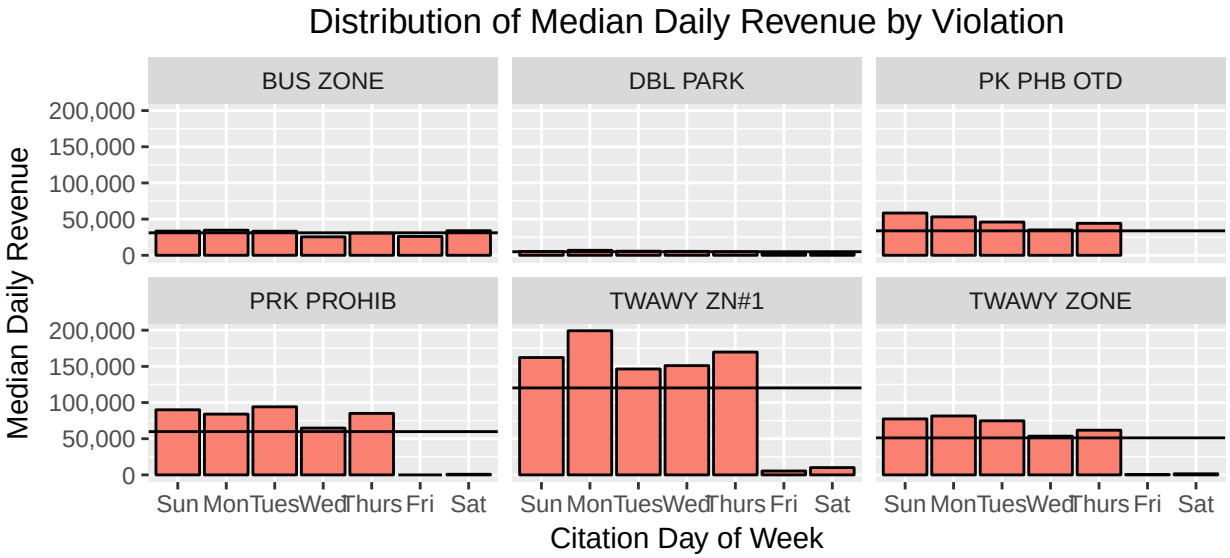
SMART Questions

1. ***In the TOLE program, which days of the week earn significantly more or less revenue than the overall median daily revenue?***
2. ***In the San Francisco TOLE program, is the type of violation issued in the citation independent from the day of the week?***

Answering the SMART Questions

Analysis of Median Daily Revenue





From the graph of "Distribution of Median Daily Revenue," we see that Tuesday appears to generate more revenue than the median daily revenue. However, Friday and Saturday appear to generate significantly less than median daily revenue. When broken down by violation type in "Distribution of Median Daily Revenue by Violation", Tuesday's high overall value is not able to be visually attributed to any single violation. However, Friday and Saturday can be attributed to below median revenue daily revenue in generation in all violations

except DBL PARK and BUS ZONE, which both seem consistent across all days. Also, TWAWY ZN#1 has a notably higher than median revenue generation on Monday.

We then decided to calculate the variance of the median daily revenue for citations from the average median daily revenue from citations.

The mean daily median revenue1 is 259301. The variance of the daily median revenue is 22472659383.2857 and the standard deviation is 149909. This high standard deviation is caused by the low median daily revenues of Friday and Saturday. We considered using a t-test to determine which days of the week are significantly different from the mean, but because of a low sample size and the lack of a normal distribution[6], conducting a t-test would not have significant meaning.

Analysis of Violation Type and Day of the Week: χ^2 Test of Independence

We decided to do a χ^2 test in order to determine whether the violation type is independent of the day of the week. The following is the contignency table:

```
## $observed
##
## tickets$Violation  Sun  Mon Tues  Wed Thurs  Fri  Sat
##        BUS ZONE    359  373  358  274   329  280  365
##        DBL PARK    434  572  471  450   404  207  338
##        PK PHB OTD  365  332  287  219   276    0    0
##        PRK PROHIB  542  506  567  390   512    1    5
##        TWAWY ZN#1 1020 1253  915  944  1061   35   65
##        TWAWY ZONE  496  522  479  343   396   23   36
##
## $expected
##
## tickets$Violation       Sun       Mon     Tues      Wed    Thurs       Fri
##        BUS ZONE    447.4535  495.0371 428.1139 364.5299 414.3397  75.96691
##        DBL PARK    550.4175  608.9507 526.6277 448.4123 509.6839  93.44775
##        PK PHB OTD  283.0555  313.1565 270.8214 230.5987 262.1080  48.05606
##        PRK PROHIB  482.8593  534.2082 461.9895 393.3742 447.1253  81.97798
##        TWAWY ZN#1 1012.9902 1120.7149 969.2074 825.2595 938.0239 171.98155
##        TWAWY ZONE  439.2240  485.9325 420.2401 357.8255 406.7192  74.56975
##
## tickets$Violation       Sat
##        BUS ZONE   112.55903
##        DBL PARK   138.46013
##        PK PHB OTD  71.20394
##        PRK PROHIB 121.46554
##        TWAWY ZN#1 254.82248
##        TWAWY ZONE 110.48887
```

The χ^2 test with 30 degrees of freedom resulted in a χ^2 statistic of 2481.004 and a p-value of approximately 0. Thus, we reject the null hypothesis that the violation types are *not* independent of the days of the week.

Conclusion

We found that Tuesday generates significantly more than the median Daily Revenue—even after controlling for type of violation—while revenue for Friday and Saturday falls precipitously. In the future, we hope to

[6](reject the null hypothesis that the sample comes from a population with a normal distribution based on Shapiro Test of Normality where p = 0)

analyze why Friday and Saturday produce significantly lower revenue but Sunday ticket revenue stays fairly consistent with the other days of the week. Furthermore, the χ^2 test resoundingly showed that violation types are *not* independent of the days of the week.

Another question that prompted further interest is the data seem to show a small and even negative correlation between the amount of the fine and the amount paid. This is counterintuitive and even the opposite of our hypothesized relationship between the two variables. This is most likely due to several factors that have not been accounted for in the dataset as we have controlled for all meaningful variables.