Jimmy Aspras
Data Analytics Challenge
December 2021

Introduction

The Chicago Tickets dataset published by ProPublic was initially part of a dataset collected to analyze how steep fines for parking tickets and the ways the laws surrounding them are enforced target black citizens in Chicago. This analysis will attempt to see if this could also be applicable to red light tickets.

Red light cameras themselves are inherently impartial, however, an analysis could reveal problem intersections or zip codes that tend to see more violations or owe more than others. It is important to review this information and make adjustments as necessary, which could include a different fine structure based on income or re-timing traffic lights, among other options.

This analysis begins with basic visualizations to determine whether and where patterns in the data might exist and then uses supervised and unsupervised models to help augment these findings. It is important to note that this analysis is limited, as there is a lack of exact location for red light camera intersections, as well as demographic information for offending drivers.

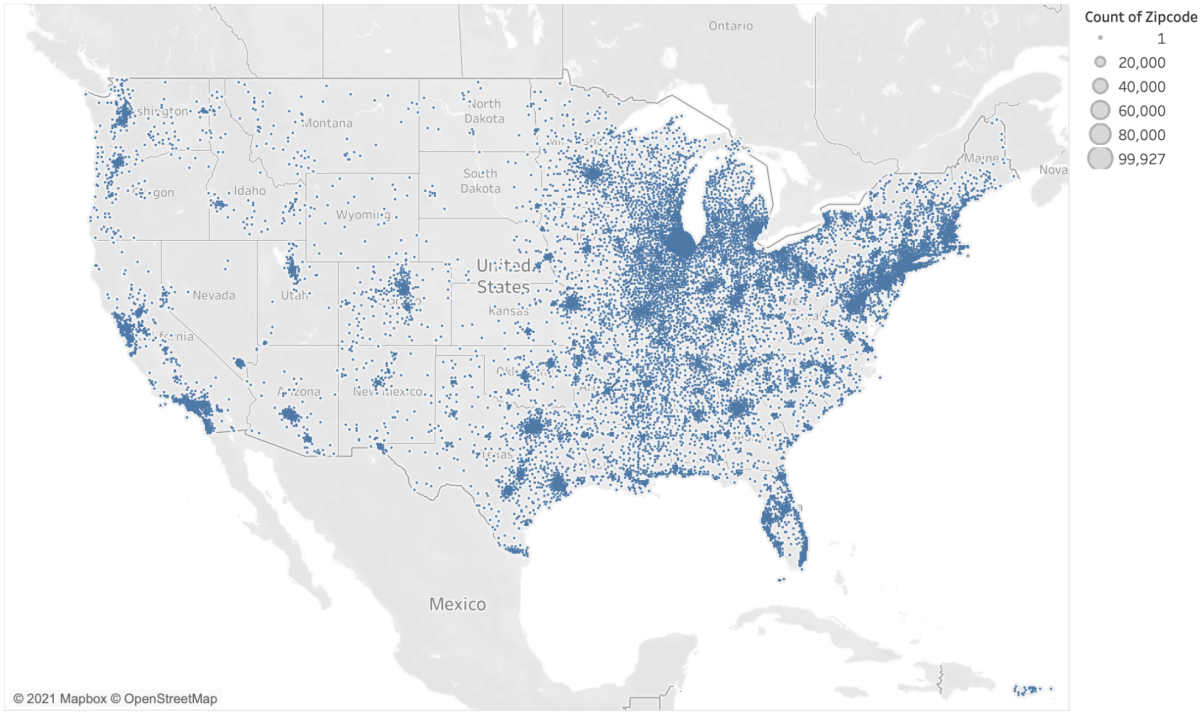
Visualizations

All visualizations were created with Tableau using the full Chicago tickets dataset.

Driver Zip Code Map

Plotting all of the offending drivers' zip codes on a map provides a general idea as to which drivers have the most offenses. Interestingly, many drivers from all 50 states and Puerto Rico have red light offenses in Chicago.

Driver Zipcode

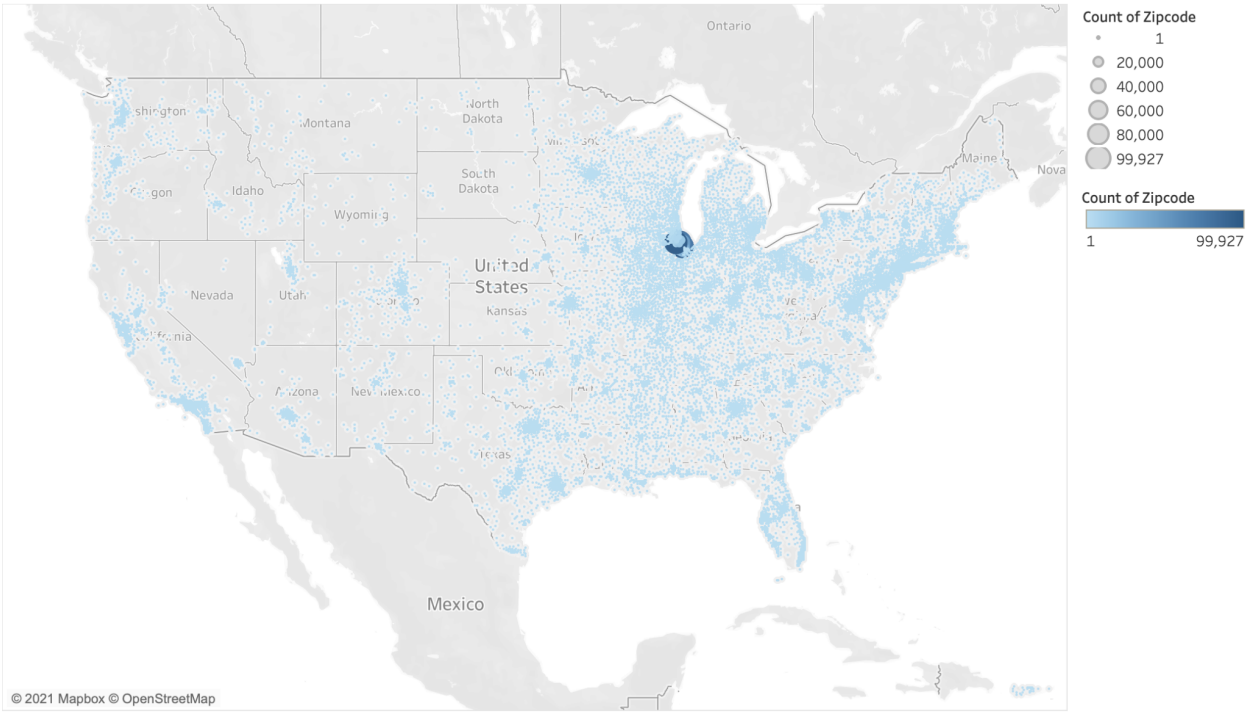


Map based on Longitude (generated) and Latitude (generated). Size shows count of Zipcode. Details are shown for Zipcode. The view is filtered on Latitude (generated) and Longitude (generated). The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

Color Coded Zip Code Map

Color coding the map by count reveals that the vast majority of tickets were issued to drivers with Chicago-area zip codes.

Driver Zipcode

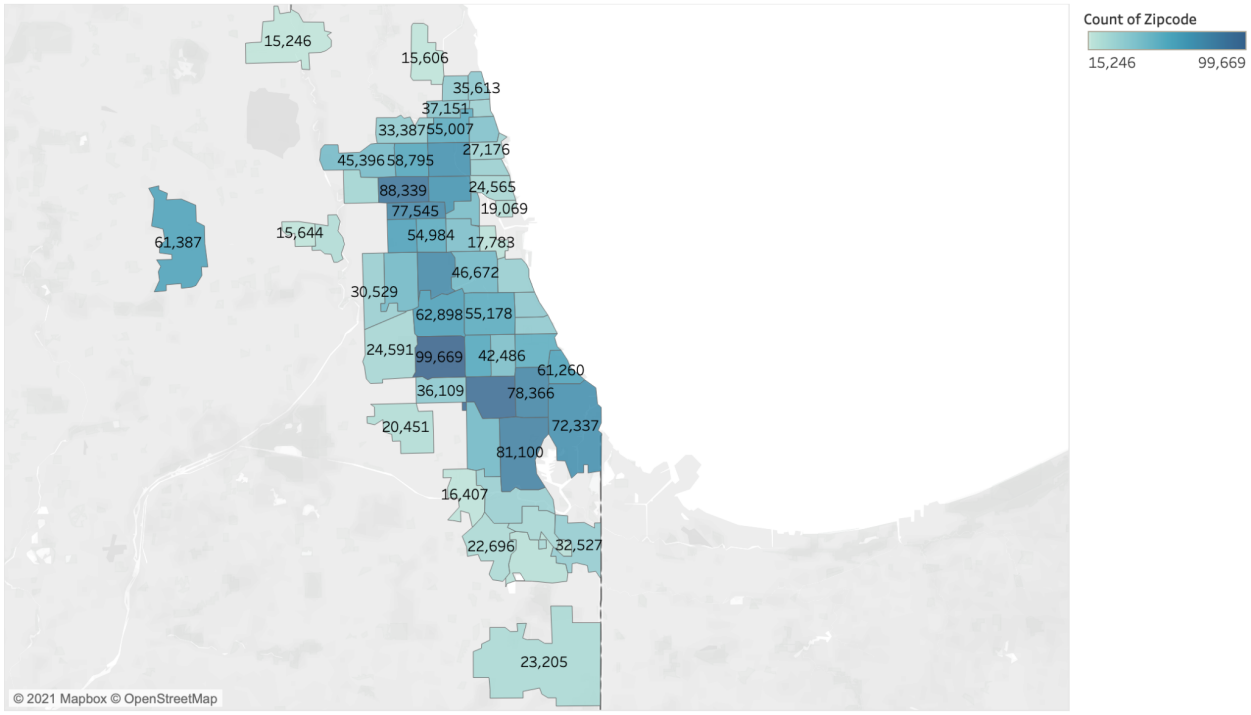


Map based on Longitude (generated) and Latitude (generated). Color shows count of Zipcode. Size shows count of Zipcode. Details are shown for Zipcode. The view is filtered on Latitude (generated) and Longitude (generated). The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

Chicago Offending Zip Codes

To get a better idea of offending zip codes, data was filtered so that only zip codes with >15,000 offenses were displayed. The resulting map reveals only zip codes in the Chicago area.

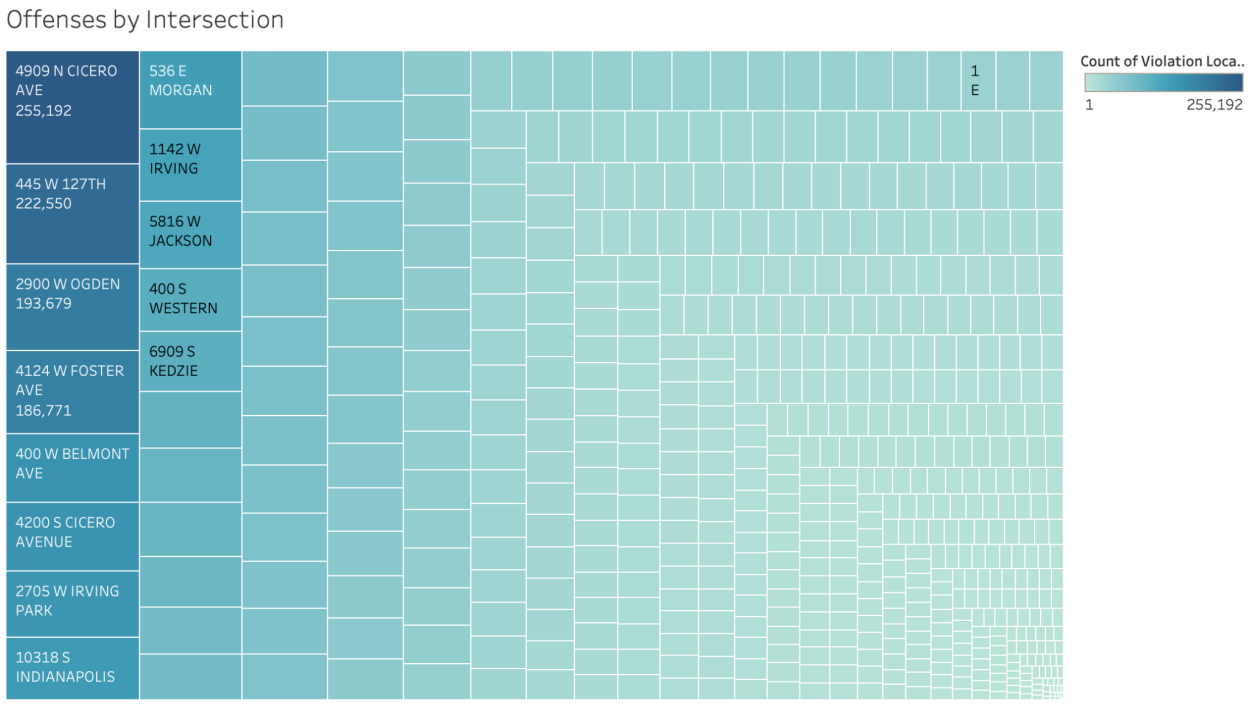
Chicago-area Offenses by Zipcode



Map based on Longitude (generated) and Latitude (generated). Color shows count of Zipcode. The marks are labeled by count of Zipcode. Details are shown for Zipcode. The data is filtered on License Plate State, which keeps IL. The view is filtered on count of Zipcode, which includes values greater than or equal to 15,000.

Intersection Treemap

Creating a treemap of the issuing intersections is also helpful to see if there are some intersections that seem to have more violations than others. There do seem to be around 8 high-offending intersections with a much greater violation rate than the rest. These intersections appear to have issued nearly 10% of all tickets.

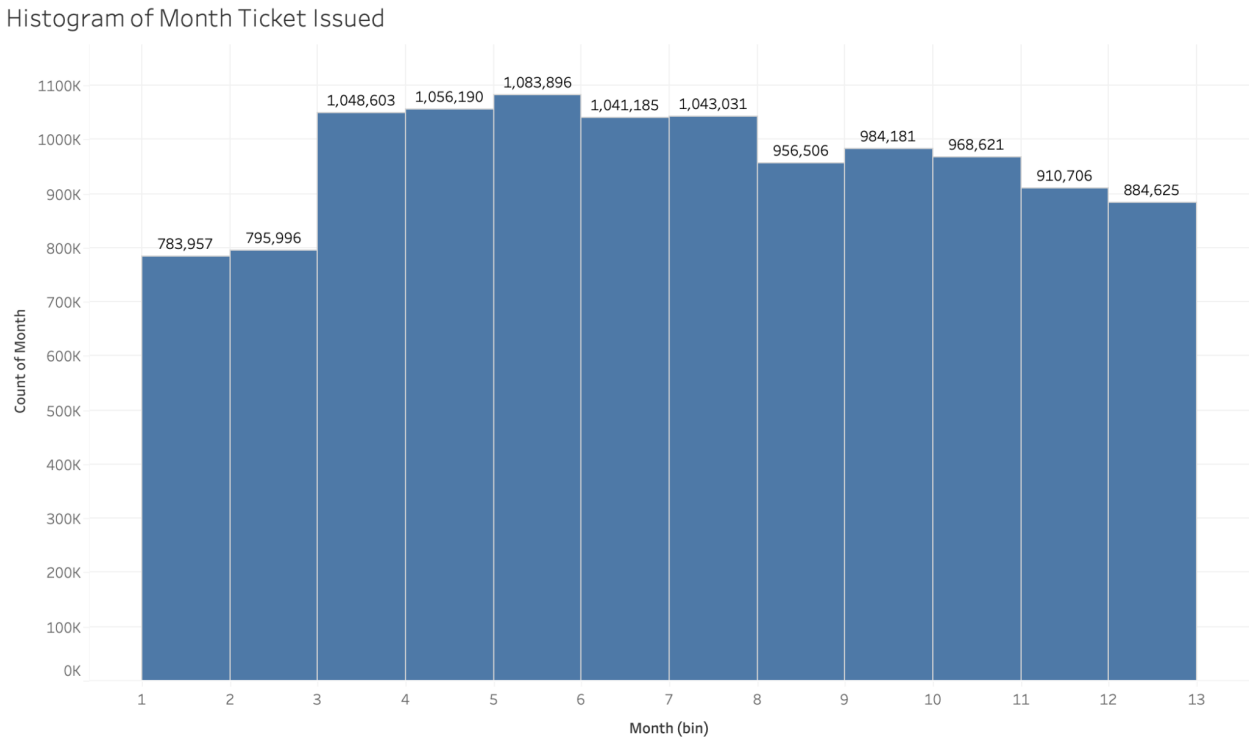


Violation Location and count of Violation Location. Color shows count of Violation Location. Size shows count of Violation Location. The marks are labeled by Violation Location and count of Violation Location.

Month Histogram

It has been established that there are certain driver zip codes that have more overall violations as well as intersections that tend to see more violations. But are there months of the year or times of day that tend to see more violations as well?

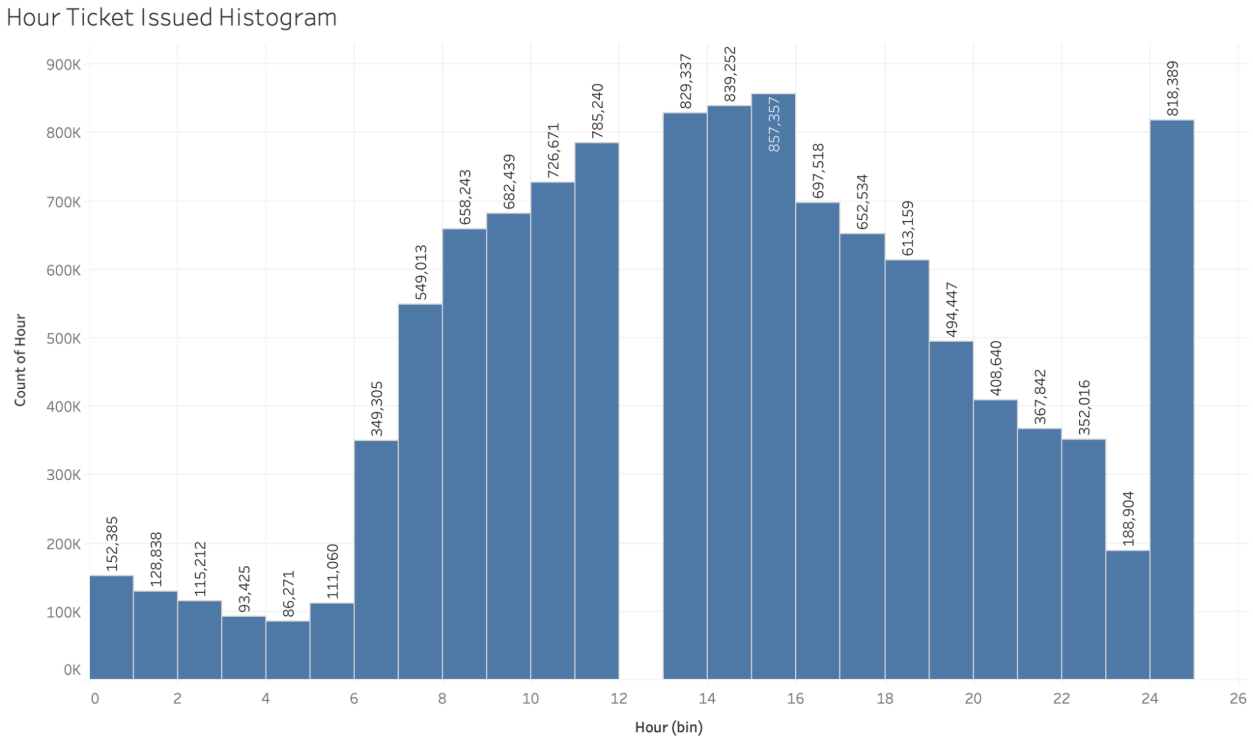
There seems to be a slight downward trend in month of violation that peaks in spring and continues downward into January and February before jumping up again in March.



The trend of count of Month for Month (bin). The marks are labeled by count of Month.

Hour Histogram

Immediately, it appears as if there could be an error in this data. Further analysis would be required to determine if there is an issue. However, there is a clear trend that more tickets are issued during waking hours, following a bell curve from the AM to PM commutes. If the data is to be believed, there is a spike of violations that occurs between 12-1 am.



The trend of count of Hour for Hour (bin). The marks are labeled by count of Hour.

Models

The preceding visualizations indicate some trends that should be investigated further. All analyses in R were conducted using the sample data set. This section seeks to determine if a model can be constructed to mathematically predict:

1) The amount a given zip code is likely to owe
2) The amount of violations during a given month
3) The amount of violations during a given hour
4) Whether there are any additional patterns that exist in the data

Zip Code as a Predictor of Amount Due

As the sample dataset is about 10% the size of the full dataset, the data was filtered to include those zip codes that have greater than 1,500 violations (10% of the 15,000 used in the Tableau analyses). This returned 60 zip codes. Dummy variables were created for each zip code, and a linear regression was run using current_amount_due as the dependent variable and the 60 zip codes used as the independent variables.

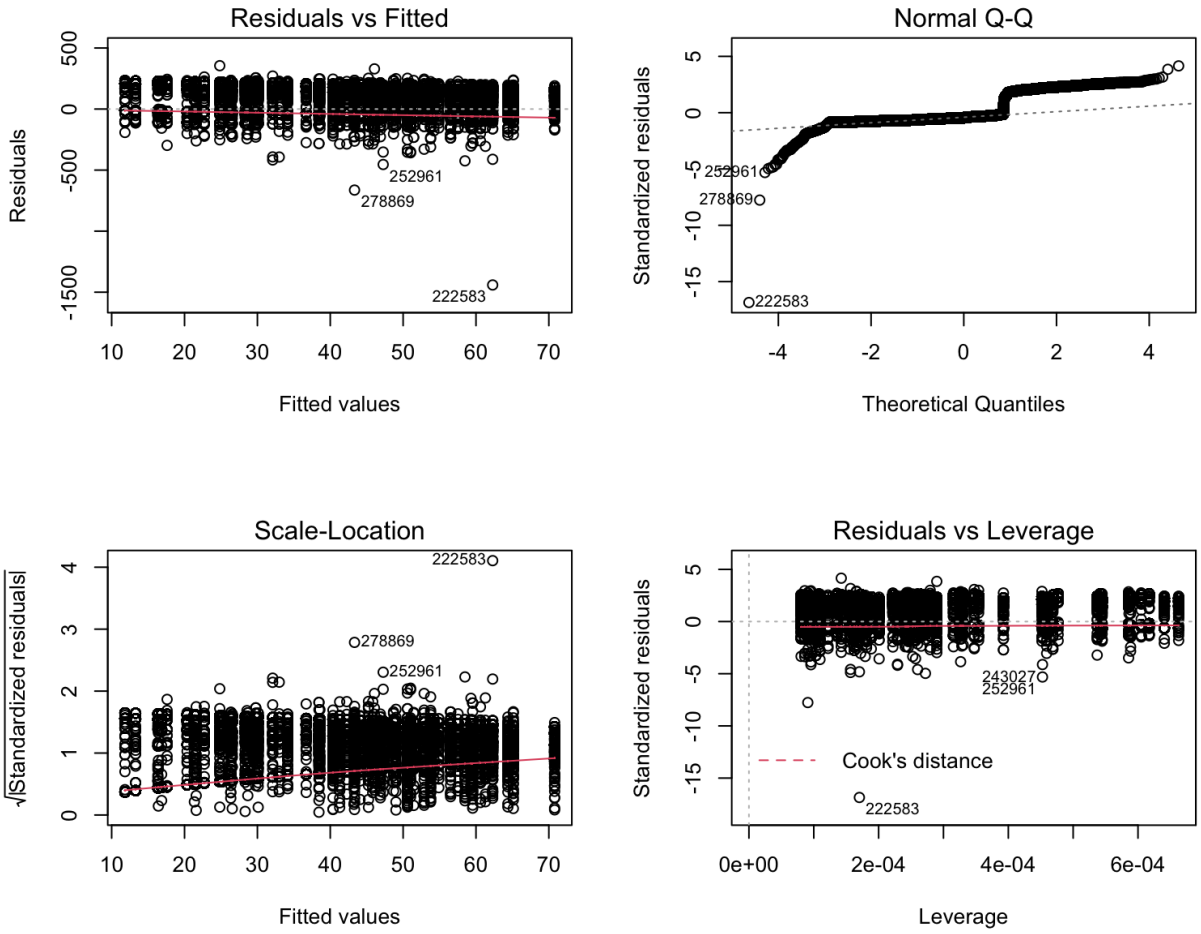
Most of the analyzed zip codes are statistically significant predictors of current amount due. It can be concluded that there are several zip codes that could be classified as “hardest hit” by the red light cameras. There are 9 zip codes (all statistically significant) where the model predicts residents would owe in excess of $30, and one zip code, 60426, where residents would be predicted to owe $42 at any given time using this dataset. Further data should be collected, and further analysis should be conducted to determine whether this is a result of population density, ability to pay, or other demographic information not available in this analysis.

Looking at the model plots, there are 5 observations that are highly influential; the scale-location plot indicates that the assumption of equal variance is not violated; the Normal Q-Q plot shows residuals that are fairly normally distributed; finally, the plot of residuals vs fitted shows that the residuals follow a linear pattern. Therefore, this appears to be a useful model for predicting the current amount due by zip code.

Model Output

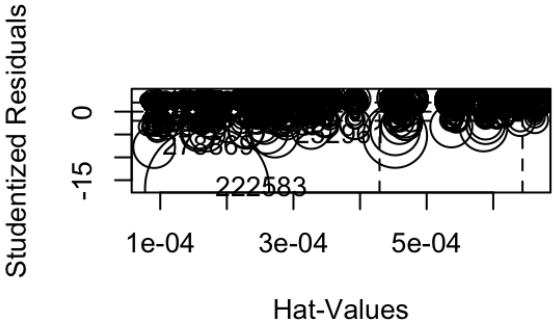
	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	.data_60426	42.2	2.53	16.7	2.28e-62
2	.data_60621	36.6	2.47	14.8	1.39e-49
3	.data_60644	35.8	2.37	15.1	1.77e-51
4	.data_60636	33.7	2.35	14.3	1.67e-46
5	.data_60624	32.7	2.40	13.6	2.84e-42
6	.data_60153	32.0	2.77	11.6	5.15e-31
7	.data_60623	31.0	2.25	13.8	3.51e-43
8	.data_60411	30.9	2.77	11.2	7.19e-29
9	.data_60637	30.7	2.49	12.4	3.92e-35
10	.data_60827	29.9	2.51	11.9	8.31e-33

Model Plots


Residuals vs Fitted
Residuals
Fitted values
252961
278869
222583
Normal Q-Q
Standardized residuals
Theoretical Quantiles
252961
278869
222583
Scale-Location
√|Standardized residuals|
Fitted values
222583
278869
252961
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
2430271
252961
222583

Influence Plot

	StudRes	Hat	CookD
207	-0.642408	6.626905e-04	4.561107e-06
311	-0.642408	6.626905e-04	4.561107e-06
222583	-16.865783	1.703287e-04	8.068325e-04
252961	-5.313725	4.526935e-04	2.131108e-04
278869	-7.763844	9.080178e-05	9.121032e-05



Tickets Given per Month

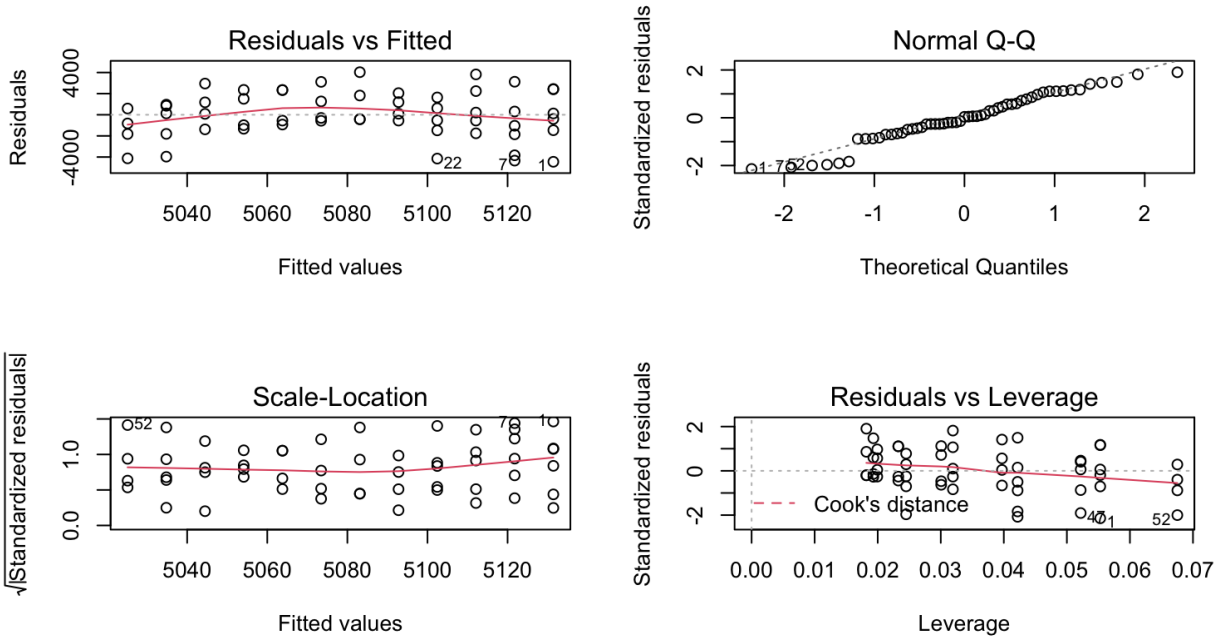
The Tableau analysis showed a slight but inconclusive trend in tickets by month. To see if there is any meaningful trend, data was grouped by month, year, and ticket count, and a simple linear regression was run to determine whether month can be used as a predictor of tickets issued.

The results show that month is not a statistically significant predictor of tickets issued. Analysis of the model plots indicates that the residuals have a small but somewhat nonlinear relationship; the residuals are normally distributed; the residuals are equally spread; and there are three points of high influence, observations 1, 7, 52, and 53. However, because the p-value for month is greater than 0.05, the model can be concluded to be of no use in predicting tickets issued by month.

Model Output

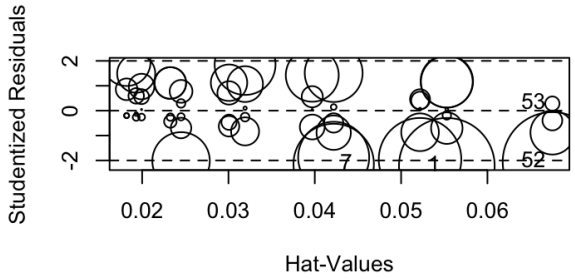
term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1 (Intercept)	5141.	570.	9.02	2.75e-12
2 month	-9.66	80.5	-0.120	9.05e- 1

Model Plot



Influence Plot

	StudRes	Hat	CookD
1	-2.2213204	0.05529418	0.134424177
7	-2.1454268	0.04218798	0.094916877
52	-2.0602907	0.06752146	0.144818311
53	0.2844553	0.06752146	0.002981249



Probability of Ticket Given Hour

The histogram of tickets by hour created in Tableau seemed to reveal a significant relationship between hour of day and number of tickets issued. To further investigate this, data was grouped by hour and ticket count, and a simple linear regression was run.

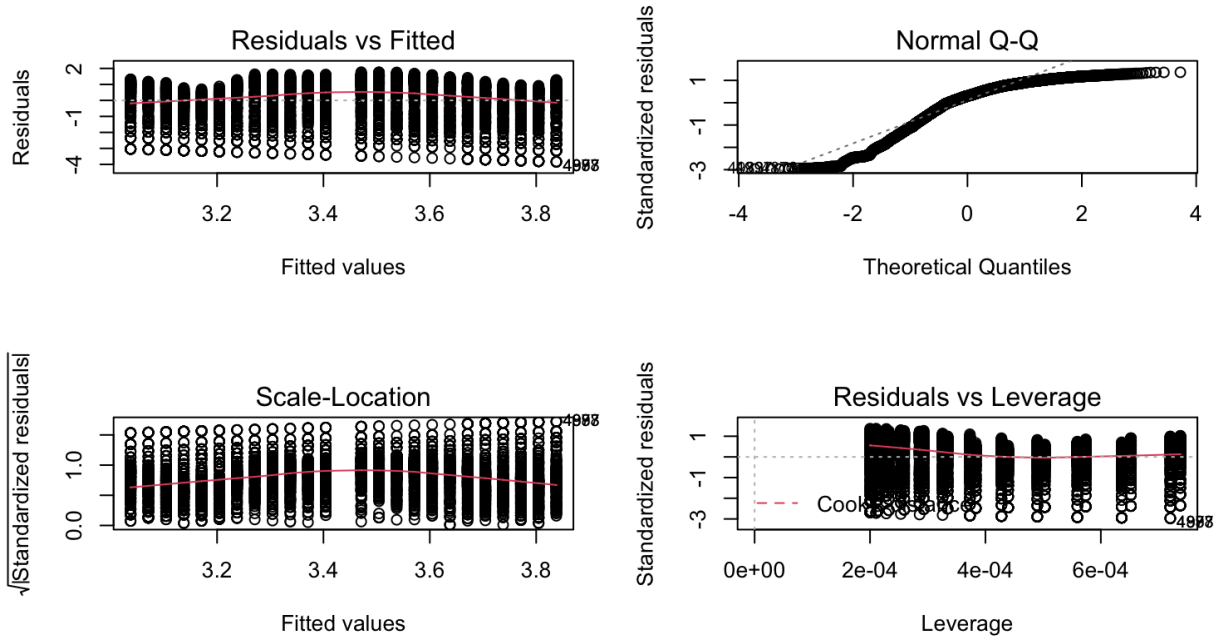
There is obviously not a linear relationship between violations and tickets issued, so the data must be transformed to attempt to draw any meaningful conclusions. Models were constructed by taking the log(1), log(10), sqrt(), and reciprocal of the dependent variable, ticket_count. The models constructed with log(1), log(10), and sqrt() transformations were fairly identical in terms of visual observation of the model plots. The reciprocal model caused a violation of assumptions of the linear model. The charts that appear below are from the log(1) model.

With a p-value of 0, hour is a statistically significant predictor of ticket count. Even with the transformed variable, there is a muted but noticeable pattern in the residuals; the Normal Q-Q plot also indicates that the residuals are not normally distributed; the Scale-Location plot does show residuals spread evenly among the range of predictors, so the assumption of equal variance seems to be satisfied; finally, there are 4 highly influential values that would alter the model if they were removed.

Model Output

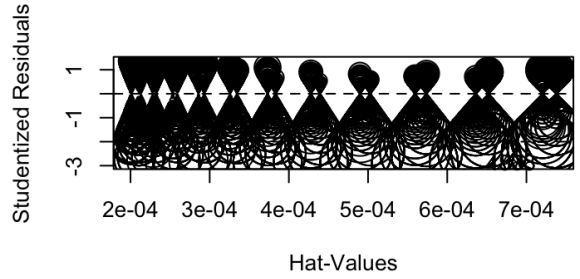
term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1 (Intercept)	3.04	0.0351	86.4	0
2 hour	0.0334	0.00249	13.4	2.36e-40

Model Plots



Influence Plot

	StudRes	Hat	CookD
3	0.9690146	0.0007378050	0.0003466554
4	0.7776619	0.0007378050	0.0002232789
4878	-2.9702603	0.0007203642	0.0031750801
4983	-2.9702603	0.0007203642	0.0031750801



Unsupervised Analysis

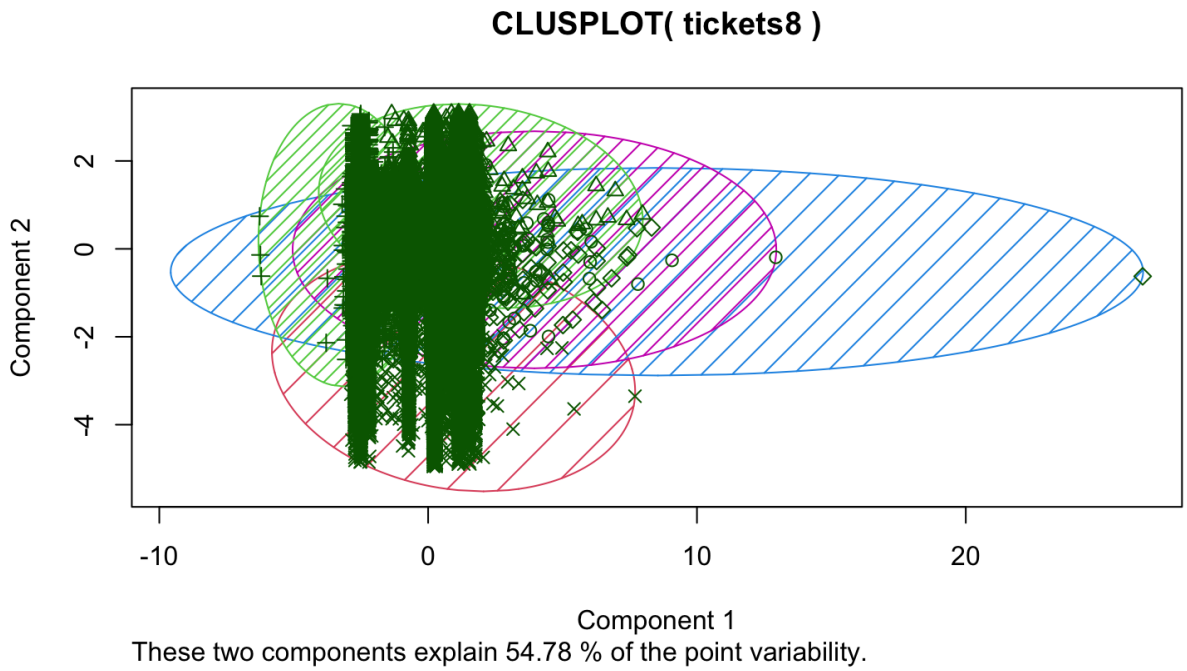
Finally, in order to determine whether there may be any hidden relationships in the data, a k-means analysis was run. The analysis was conducted using the variables zipcode, current_amount_due, total_payments, month, and hour and built 5 clusters. The plots below indicate that there are no clear relationships using these variables.

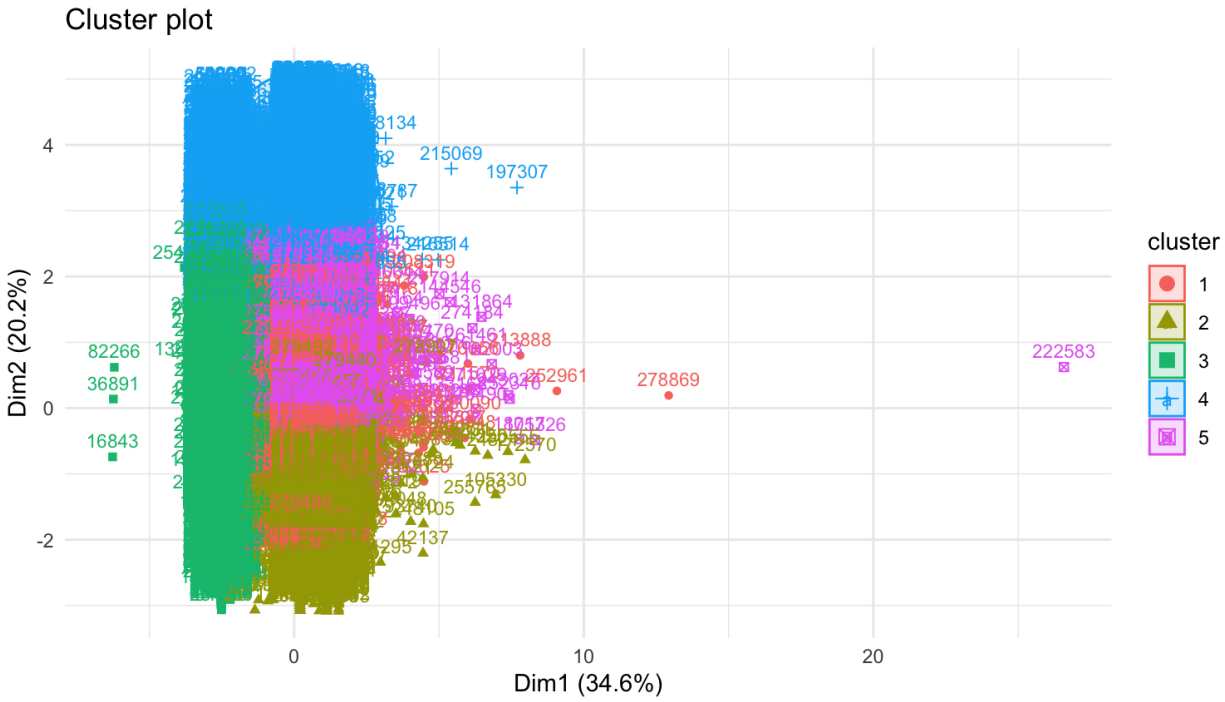
Model Output

Cluster means:

	zipcode	current_amount_due	total_payments	month	hour
1	0.1590405	-0.4848513	0.325775986	1.007119861	0.41834357
2	0.1299550	2.0261463	-1.483205324	-0.037758165	-0.01722353
3	0.1748674	-0.4838891	0.352292664	-0.190681612	-1.17899513
4	0.1519904	-0.4855979	0.395968066	-0.816434222	0.59614388
5	-4.0624807	-0.1022125	0.002571837	-0.001722987	0.01252070

Model Plots




Cluster plot
Dim2 (20.2%)
Dim1 (34.6%)
4
2
0
-2
0
10
20
cluster
1
2
3
4
5
82266
36891
16843
222583
278869
252961
215069
197307
279440
279480
42137

Conclusion

Through visualizations and statistical analysis, it has been established that there are several clear patterns in the data:

- There are several intersections that are the source of a proportionally high number of tickets
- There are zip codes that have been issued substantially more tickets than their neighbors
- There are zip codes that are significantly more likely to pay or be able to pay their ticket than others
- Time of day could be a meaningful predictor of tickets, but month is not
- An unsupervised analysis was not able to find meaningful relationships among a handful of numeric variables

This analysis raises some additional questions that were initially raised when ProPublica published their original article. Further research and analysis should be conducted to link zip codes to intersections and demographic data to driver zip codes to determine whether there are any groups that are being disproportionately harmed by red light traffic cameras.

Furthermore, the fact that there is a relationship between zip code and, possibly, ability to pay, the city should investigate alternative penalty structures for offending drivers (income based, percentage of car value, etc.), or possibly re-timing the lights to allow for a greater volume of drivers to traverse the intersections during peak hours.

Finally, the intersections with the most tickets should be evaluated to determine if there are extenuating circumstances that cause outsized violations relative to others, such as higher traffic volume, shorter yellow light cycles, lack of left turn lane, etc. Some basic improvements could reduce the number of violations and improve the safety of the intersection.

R Code

Zip code regression

```
library(tidyverse)
library(readr)

options(max.print=50)
tickets=read_csv("/Users/jimmyaspras/Challenge/chicago_camera_tickets_sample.csv")
as_tibble(tickets)

#Amount due by zip code analysis

#Select columns relevant to the analysis

tickets2<-tickets[c("issue_date","violation_location","license_plate_state","zipcode","fine_level1_
amount",
                    "fine_level2_amount","current_amount_due","total_payments",
                    "ticket_queue","year","month","hour")]

#Drop missing data

tickets3<-na.omit(tickets2)

#Since the data primarily pertains to Chicago, filter out offending drivers outside IL

tickets4<-filter(tickets3,license_plate_state=="IL")

#To reduce the size of the data set and to better focus in on the Chicago area, filter all zip codes
with less than 1,500 violations

#find and return zip codes with >1,500 violations

violations<-count(tickets4,zipcode,sort=TRUE,name="max_violations")

#Zip code 60016 was the closest to 15,000 total violations in the Tableau analysis with the full
dataset, which corresponds with just about 1,500 in this sample set.

violations2<-as.data.frame(violations)
violations3<-filter(violations2,max_violations>1500)

#Filter the dataset with the zipcode output in violations3
```

```
tickets5<-semi_join(tickets4, violations3, by = "zipcode")

#Create dummy variables for remaining zip codes

library(fastDummies)
dummies<-dummy_cols(tickets5$zipcode,
                    remove_first_dummy=TRUE,
                    ignore_na=TRUE)

#Merge the dataframes

amountdue_lmdata<-cbind(dummies,tickets5)

#Drop columns irrelevant to the analysis

amountdue_lmdata<-select(amountdue_lmdata,-c(".data","issue_date","violation_location","licen
se_plate_state",
                     "zipcode","fine_level1_amount","fine_level2_amount",
                     "total_payments","ticket_queue","year","month","hour"))

library(tidymodels)
amountdue_lm = linear_reg() %>%
  set_engine("lm") %>%
  fit(current_amount_due ~ ., data = amountdue_lmdata)

library(broom)
cleanamountdue_lm<-tidy(amountdue_lm)
arrange(cleanamountdue_lm,desc(estimate))

par(mfrow=c(2,2))
plot(amountdue_lm$fit)

library(car)
influencePlot(amountdue_lm$fit)
```

Month regression

```
#Tickets by month

monthregdata<-tickets5
monthregdata2<-monthregdata %>%
  group_by(month,year) %>%
  dplyr::summarise(ticket_count=n())
```

```
monthreg = linear_reg() %>%
  set_engine("lm") %>%
  fit(ticket_count ~ month, data = monthregdata2)
tidy(monthreg)

#Model charts

par(mfrow=c(2,2))
plot(monthreg$fit)

#Influence plot

influencePlot(monthreg$fit)
```

Hour regression model (and alternative transformations)

```
#Tickets by hour

hourdata<-tickets5

#Create variable for day

hourdata$issue_date<-format(as.Date(hourdata$issue_date,format="%Y-%m-%d"), format =
"%d")
as.numeric(hourdata$issue_date)

hourregdata<-hourdata %>%
  group_by(hour,issue_date,year) %>%
  dplyr::summarise(ticket_count=n())

as.factor(hourregdata$hour)

#From the Tableau analysis, we know that tickets and hour have a nonlinear relationship and
should therefore be transformed to perform the analysis

hourregdatalog<-hourregdata
hourregdatalog$ticket_count<-log(hourregdatalog$ticket_count,base=exp(1))

hourreg = linear_reg() %>%
  set_engine("lm") %>%
  fit(ticket_count ~ hour, data = hourregdatalog)
tidy(hourreg)
```

```
#Model charts

par(mfrow=c(2,2))
plot(hourreg$fit)

#Influence plot

influencePlot(hourreg$fit)


#Log10

hourregdatalog10<-hourregdata
hourregdatalog10$ticket_count<-log(hourregdatalog10$ticket_count,base=exp(10))

hourreg10 = linear_reg() %>%
  set_engine("lm") %>%
  fit(ticket_count ~ hour, data = hourregdatalog10)
tidy(hourreg10)

#Model charts

par(mfrow=c(2,2))
plot(hourreg10$fit)

#Influence plot

influencePlot(hourreg10$fit)


#SQRT

hourregdatasqrt<-hourregdata
hourregdatasqrt$ticket_count<-log(hourregdatasqrt$ticket_count)

hourregsqrt = linear_reg() %>%
  set_engine("lm") %>%
  fit(ticket_count ~ hour, data = hourregdatasqrt)
tidy(hourregsqrt)

#Model charts

par(mfrow=c(2,2))
```

```
plot(hourregsqrt$fit)

#Influence plot

influencePlot(hourregsqrt$fit)



#Reciprocal

hourregdatarecip<-hourregdata
hourregdatarecip$ticket_count<-(1/(hourregdatarecip$ticket_count))

hourregrecip = linear_reg() %>%
  set_engine("lm") %>%
  fit(ticket_count ~ hour, data = hourregdatarecip)
tidy(hourregrecip)

#Model charts

par(mfrow=c(2,2))
plot(hourregrecip$fit)

#Influence plot

influencePlot(hourregrecip$fit)
```

Unsupervised model

```
#Unsupervised model

tickets6<-tickets5[c("zipcode","current_amount_due","total_payments","month","hour")]

#Convert and scale

tickets6[] = lapply(tickets6, as.numeric)
tickets6 = scale(tickets6)

#Build model

ticketunsupmodel = kmeans(tickets6, 5, nstart = 10)
ticketunsupmodel
```

```
#Cluster plot

library(cluster)
library(fpc)
par(mfrow=c(1,1))
clusplot(tickets6, ticketunsupmodel$cluster, color = TRUE, shade = TRUE)

#Cluster plot 2

library(factoextra)
fviz_cluster(ticketunsupmodel,data = tickets6, labelsize = 9, ellipse.type = "norm",
             choose.vars = c("zipcode","current_amount_due","total_payments","month","hour")) +
theme_minimal()
```