IST687 – Applied Data Science \ Prof. John Santerre

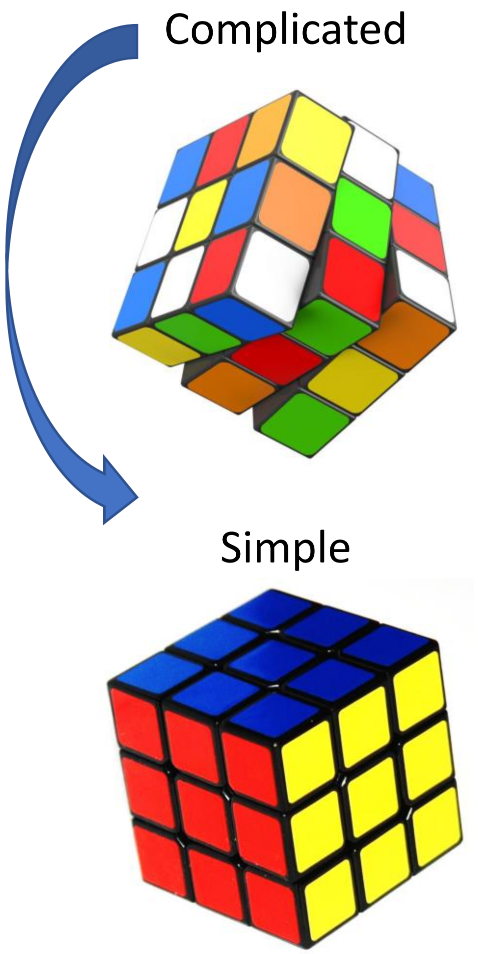
Team: Bradley Coy, Brian Hogan, Jason Maloney & Joel Whitney

Final Project

Getting Swept Up in LA Parking ~ 2018 Ticket Analysis ~

Situation

Team had issues with timestamps in another data set leading to selection of a LA ticket database. Performed the following to make suitable for analysis:



- Original environment:
  - ➢19,000,000 records x 19 descriptor variables x several years
- Team focused on 2018
- Selected ticket violation type: *street cleaning*
  - ➢… 30+ *other* violations
- Learning focus: ticket discriminating factors
  - ➢Street route, car color, time of day, day of week

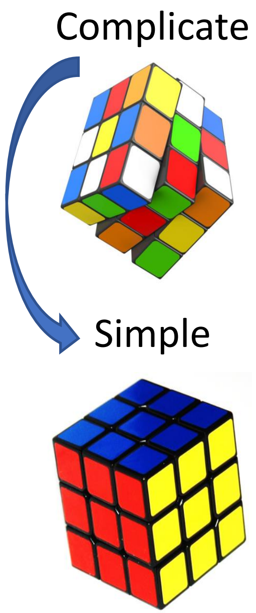
✓Resulting data set: ~600,000 records x 27 variables

https://www.kaggle.com/cityofLA/los-angeles-parking-citations#parking-citations.csv

Data Munging

Cleaning was found to be necessary to support numeric calculations, addressing NAs and data binning to enable data visualizing for anomalies

Munging…
1) Drop X column and check for duplicate ticket.number values
2) Break issue.date into months, weekday, keep issue date 'as is'
3) Convert issue.time into an actual timestamp, and bin into parts of day (morning, early afternoon, evening, etc.)
4) Clean up null, blank and NA values in all columns
   -> insert 0 when necessary
5) Break up plate.expiry.date into month/year
   -> flag for expired plates
6) Drop VIN column because all values are NA
7) Add a flag for import/domestic vehicle makes (checking for data quality issues along the way)
8) Simplify car color levels – 40 original colors
9) Clean up violation.code so formatted same ("80.69BS")
10) Run the lat and lon conversion (below)



Convert measurements from feet to lat & long with library proj4

transform alpha/numeric fields for calculations

ML

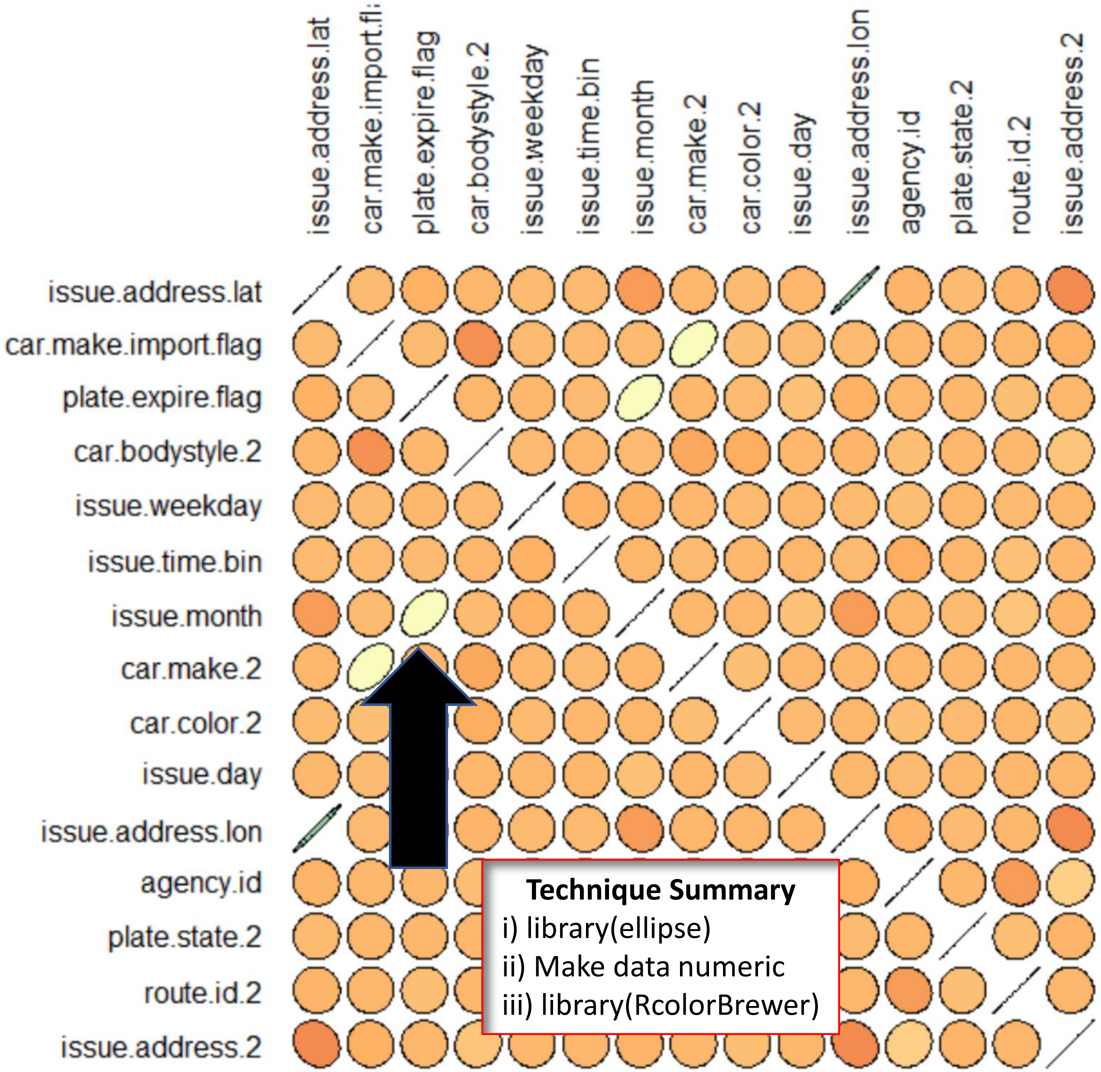
challenges w 53 levels…

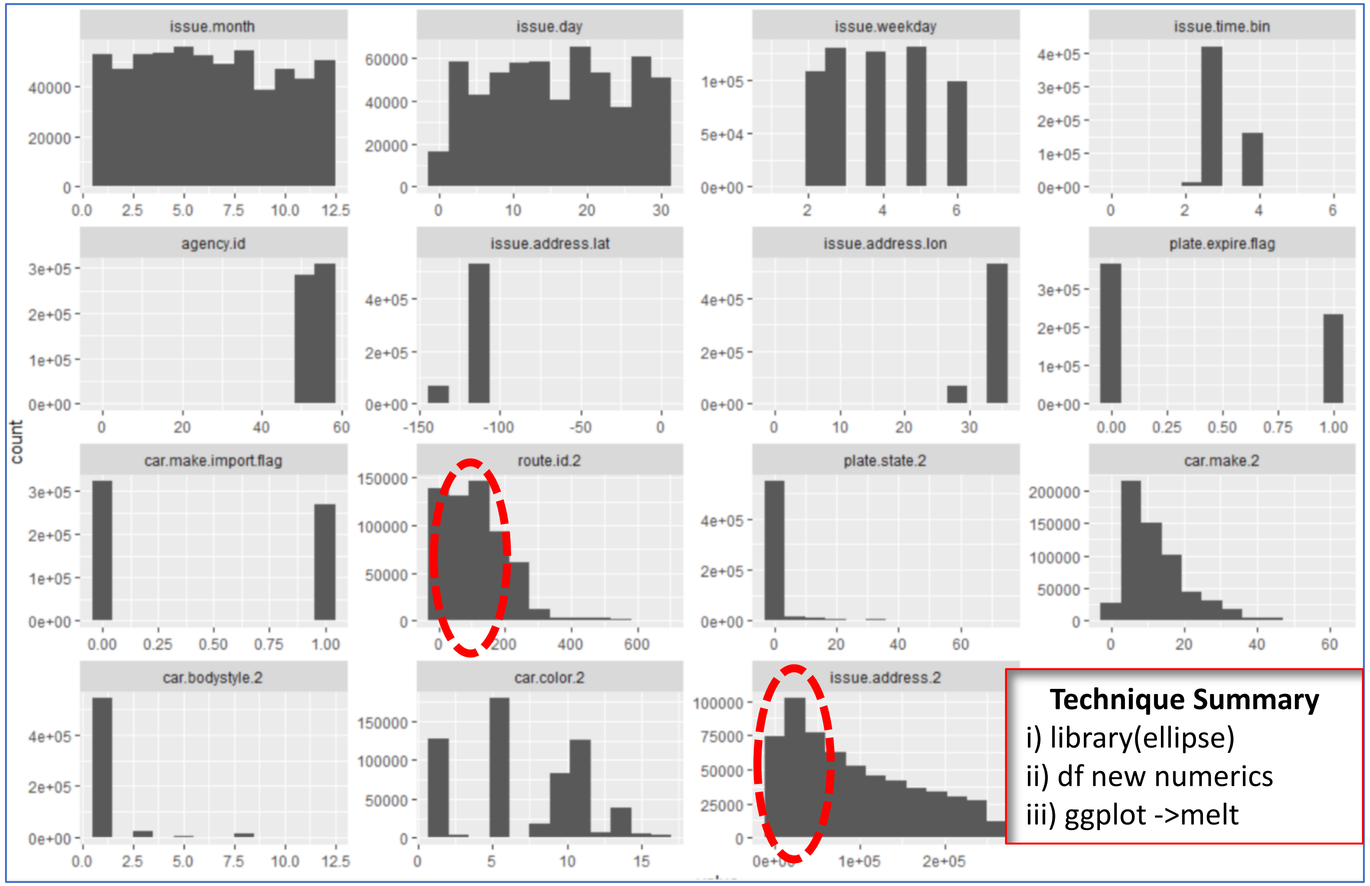
Lat/lon conversion:
```
pj <- "+proj=lcc +lat_1=34.03333333333333 +lat_2=35.46666666666667 +lat_0=33.5 +lon_0=-118 +x_0=2000000
    +y_0=500000.0000000002 +ellps=GRS80 +datum=NAD83 +to_meter=0.3048006096012192 no_defs"
Park <- cbind(park, data.frame(project(data.frame(park$Latitude, park$Longitude), proj = pj, inverse = TRUE)))
```

Analysis: Initial Correlations & Variable Inspection

- Team was hopeful of seeing stronger correlations with "route.id"
- Plate expiry and month *towed* positively correlated

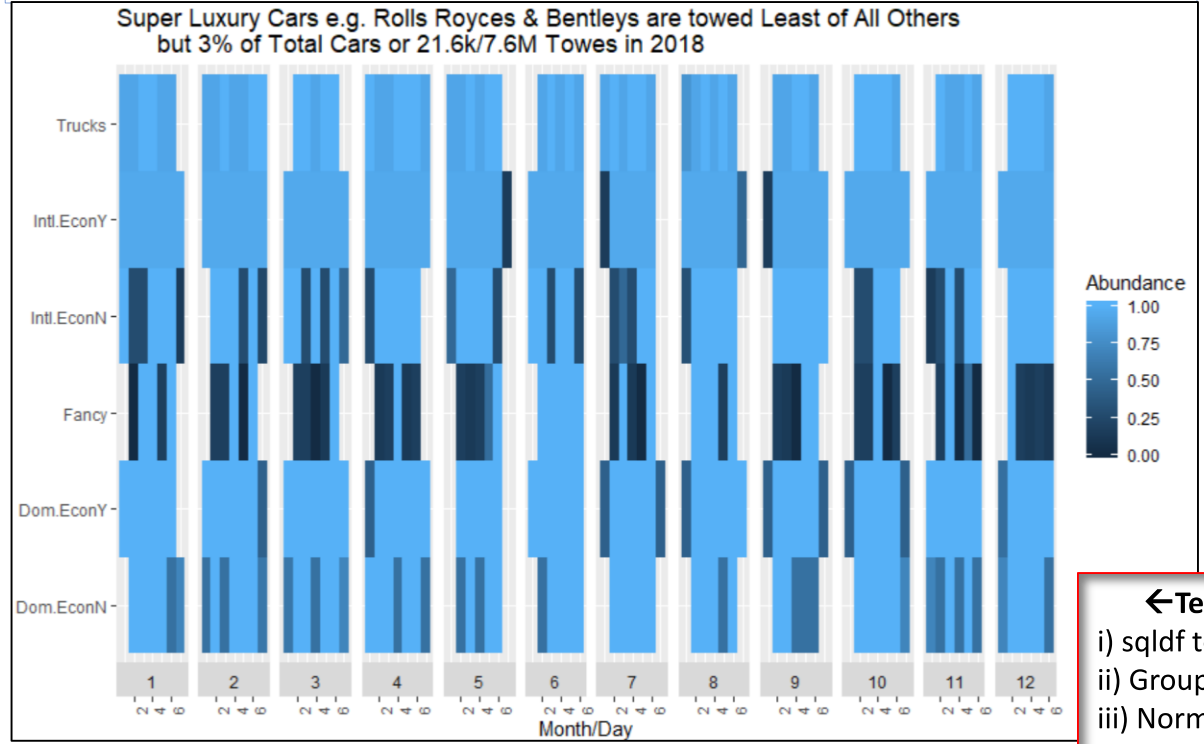
- route.id & issue.address have left hand grouping lead team speculate there are key locations, perhaps higher traffic or visitor areas, where vehicles are being towed…
- Team initially speculated whether more tows performed on out-of-state visitors in these groupings



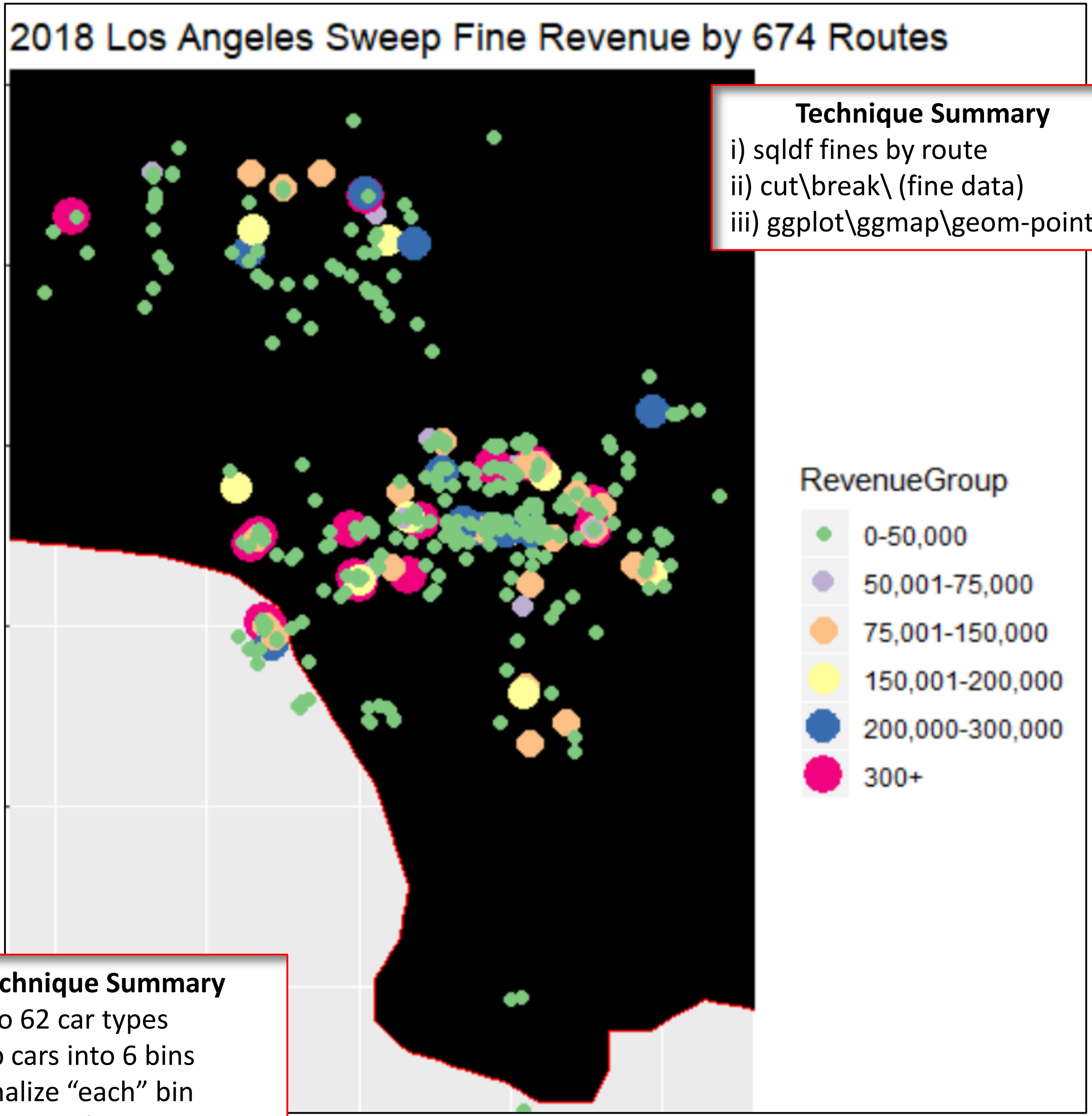


Analysis: Descriptive Statistics

- Street sweeping does not bend the rules on fancy cars; everything gets towed!
- Street parking is an enormous revenue generator for the city with some routes more substantial than others



←Technique Summary
i) sqldf to 62 car types
ii) Group cars into 6 bins
iii) Normalize "each" bin
iv) library(tidyr)\abundance



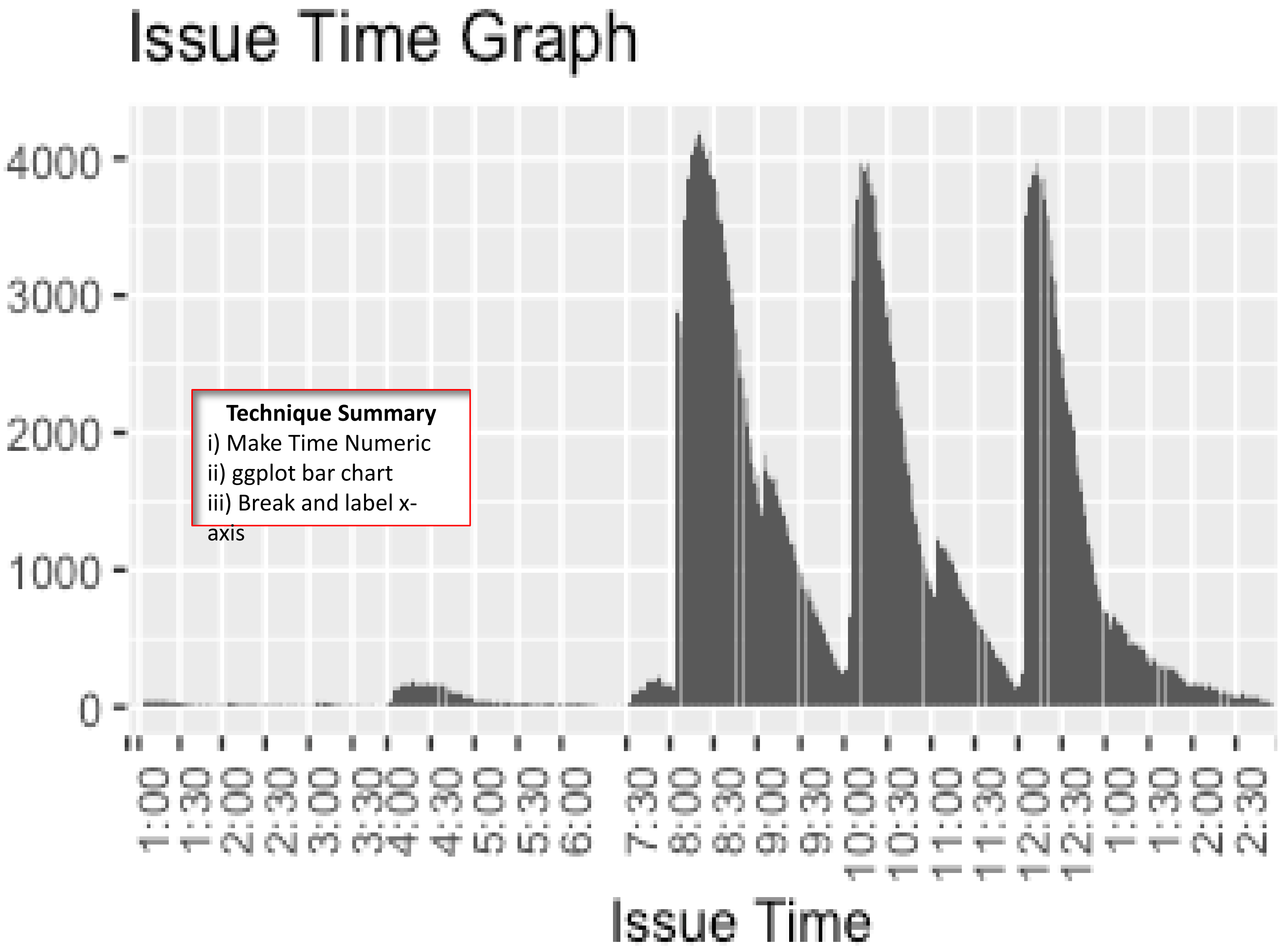
Technique Summary
i) sqldf fines by route
ii) cut\break\ (fine data)
iii) ggplot\ggmap\geom-point

References: https://jcoliver.github.io/learn-r/006-heatmaps.html

Analysis & Visualization

- Illustrating typical street cleaning behavior… represented like a planned spike when cleaning kicks off at 8, 10, &12

- Subsequent hours, 9, 11,12 likely associated with ticket & towing on secondary streets in cleaning zones



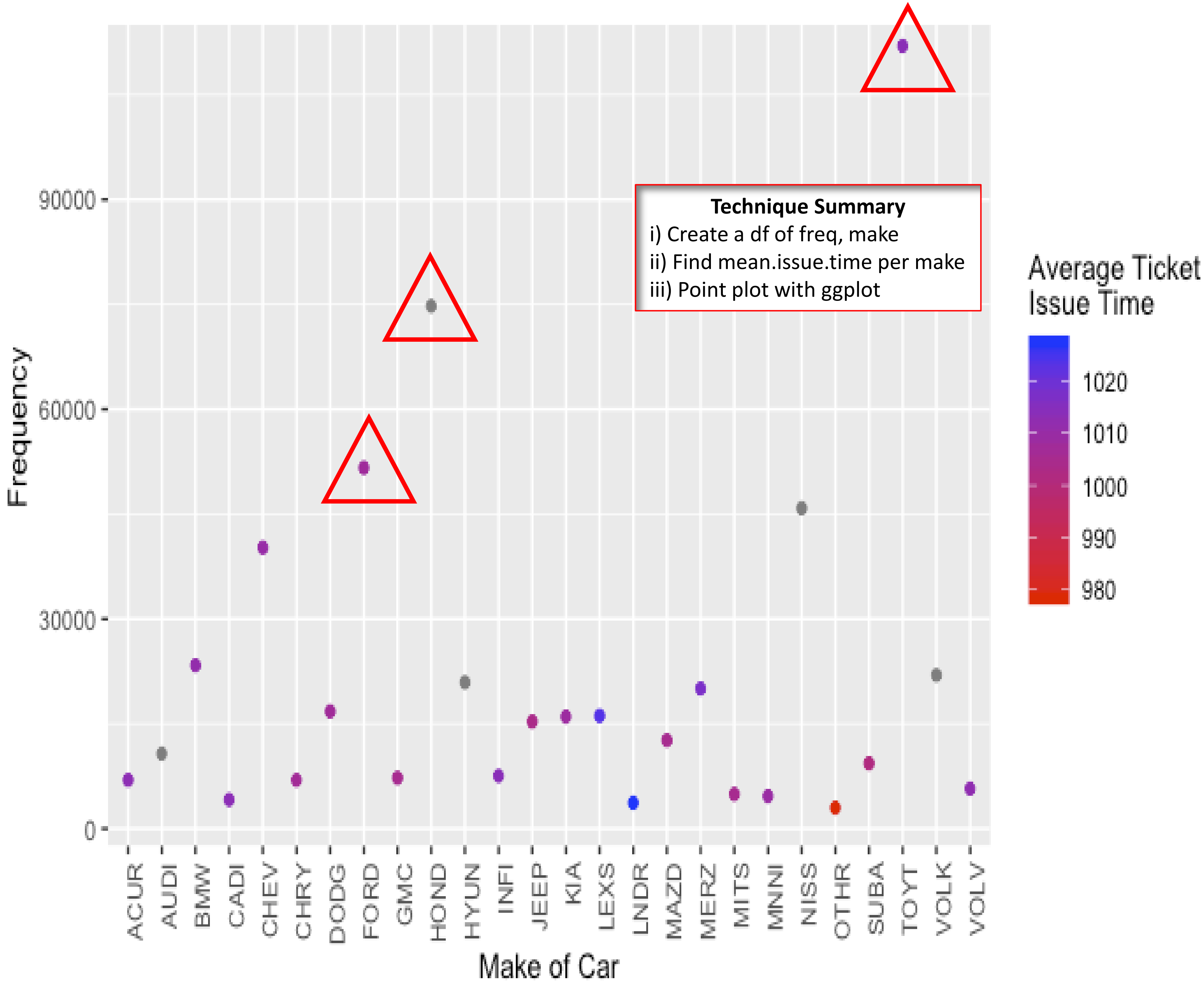
Technique Summary
i) Make Time Numeric
ii) ggplot bar chart
iii) Break and label x-axis

Analysis & Visualization

- Team was hypothesizing predicting car makes more likely to reflect consumer behavior resulting in towing tickets

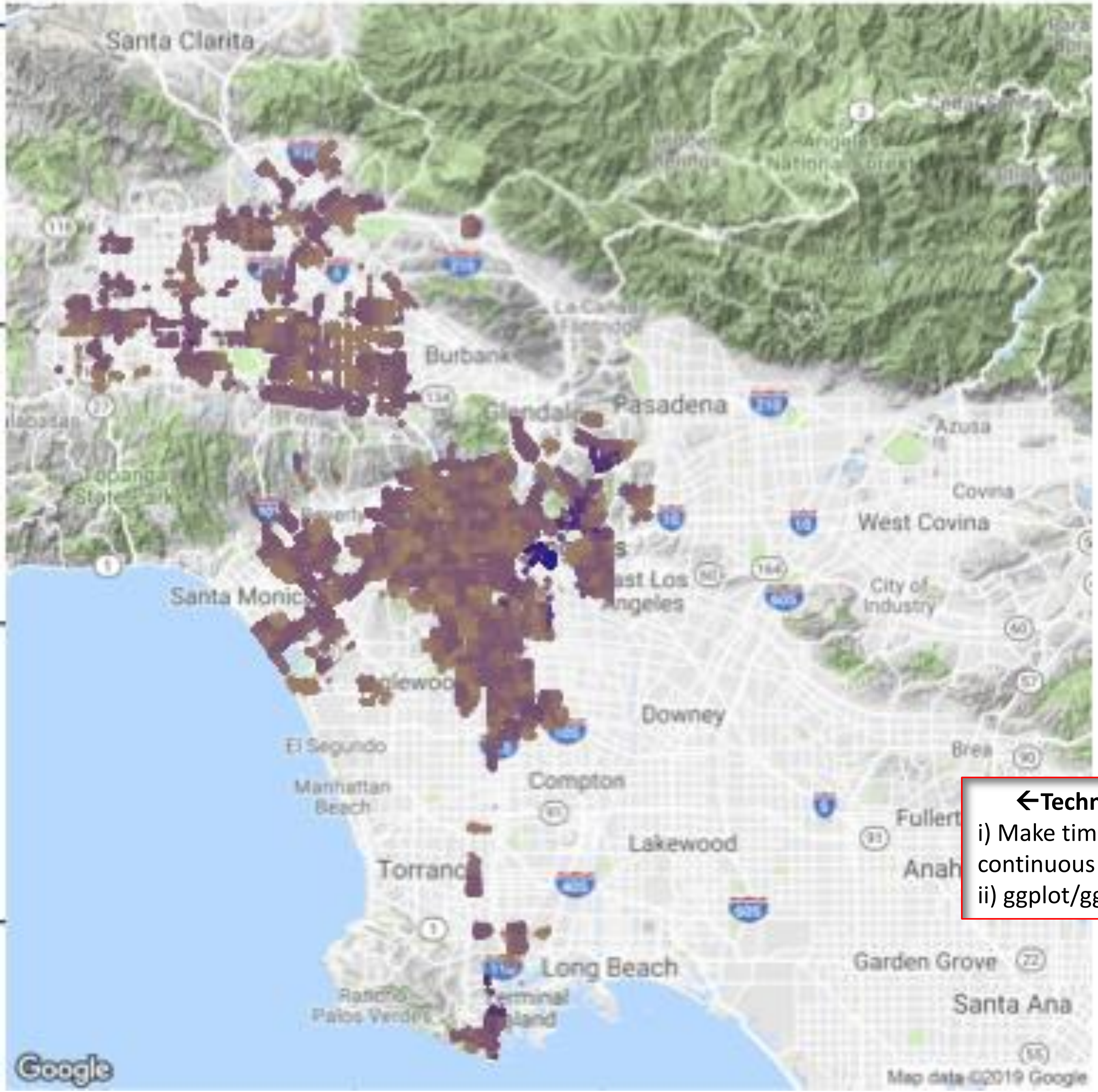
- More Toyotas, Hondas, and Fords get tickets than other makes

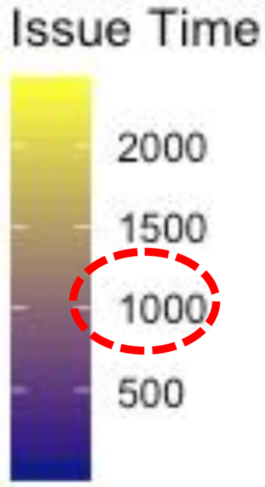
- Team speculates model affordability associated with street parking than any other behavior



Analysis & Visualization

- Dot plot shows ticket grouped by location and spread by issue time .
- Data time stamps buckets spread over day
- Most tickets issued around 10 am in the downtown, and prominent beach area
- Clearly street cleaning signs less effective

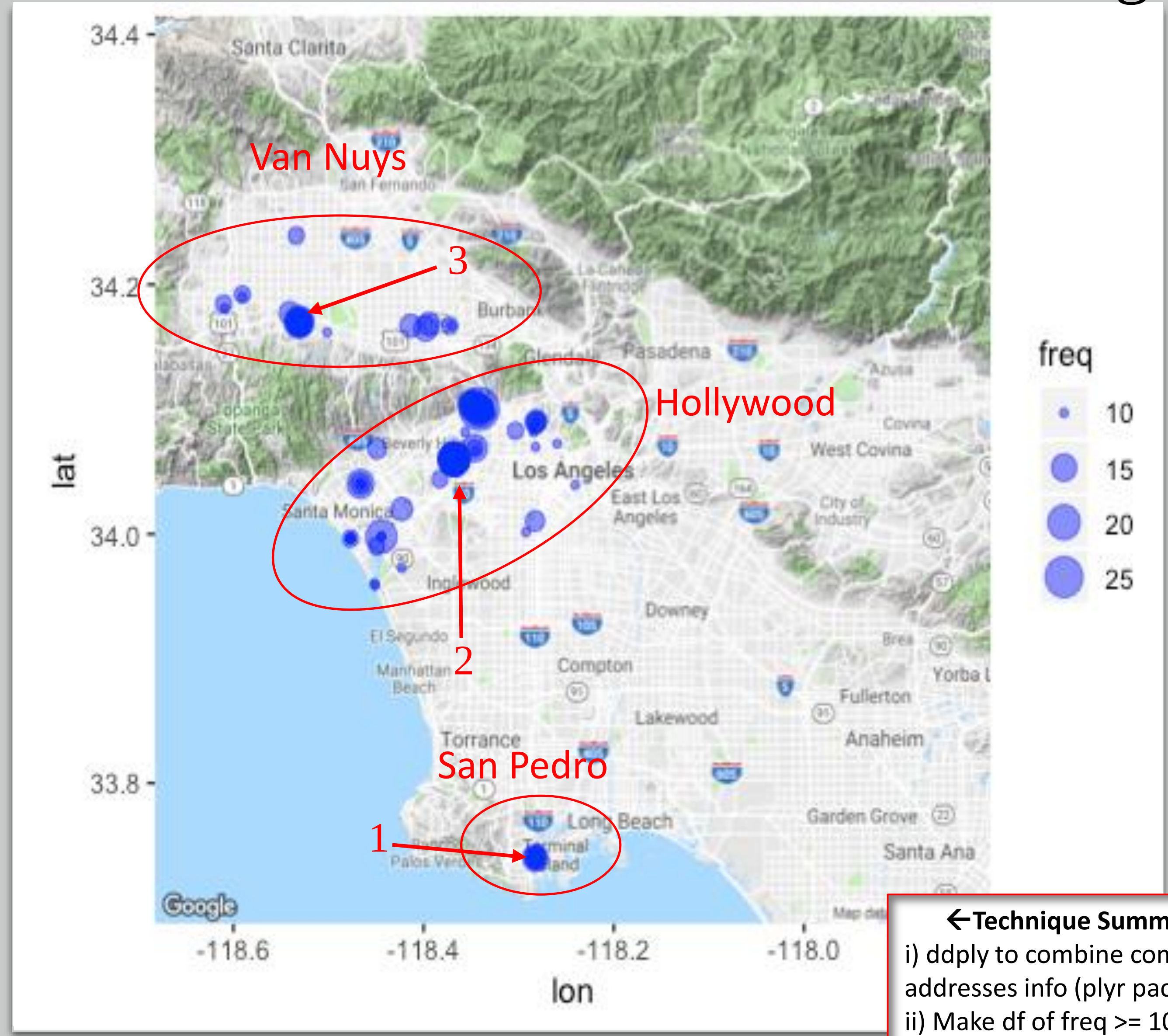




←Technique Summary
i) Make time numeric and continuous
ii) ggplot/ggmap

	AM		PM	
Time	4--9	9--12	12--3	3--12
Bin(s)	0,1,2	3	4	5,6
#Tick	14,526	418,344	161,636	40

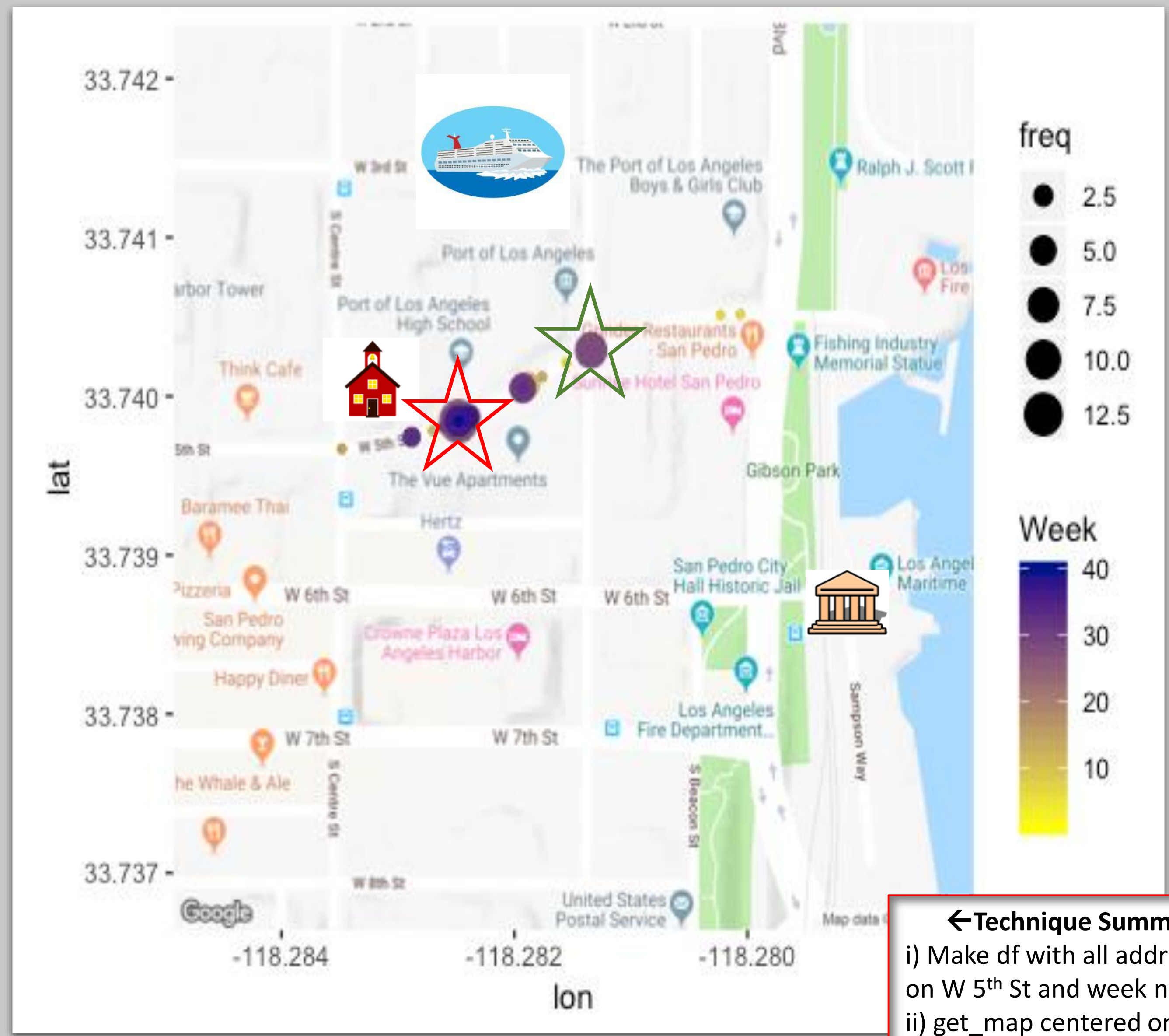
Areas with 10 or more tickets issued in a given week.



←Technique Summary
i) ddply to combine common addresses info (plyr package)
ii) Make df of freq >= 10, lat, lon, issue.address
iii) ggplot/geom_point

- Darker circles indicate locations with several weeks of 10 or more tickets issued.
- Locations with high ticket numbers include:
- 1. 255 West 5th Street in San Pedro (13)
- 2. 7000 Hawthorn Avenue in Hollywood (27)
- 3. 5525 Etiwanda Avenue in Van Nuys (18)

San Pedro: W 5th Street



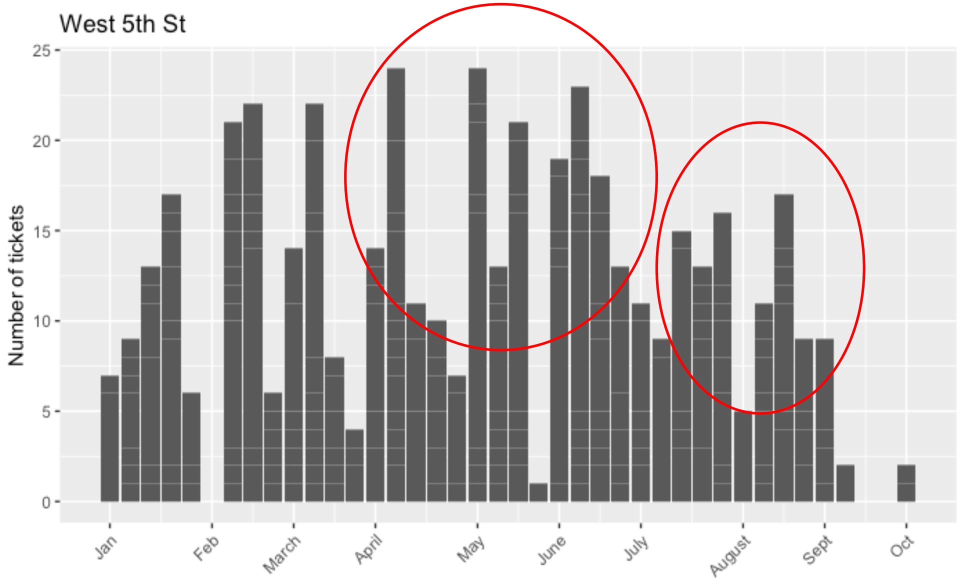
←Technique Summary

i) Make df with all addresses on W 5th St and week number

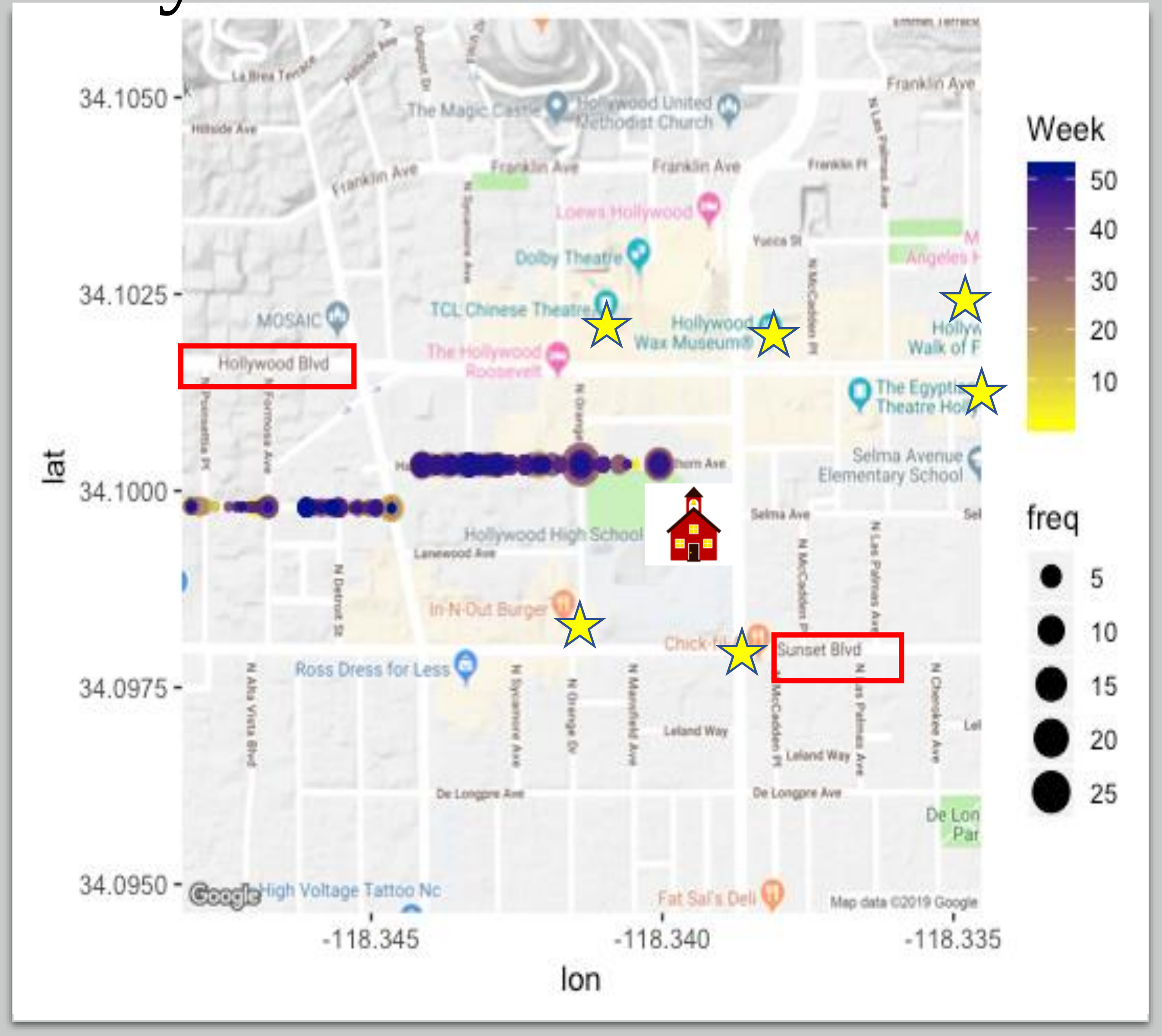
ii) get_map centered on most frequent address

iii) ggplot/geom_point

- On West 5th Street is an apartment complex (under the red star), Port of Los Angeles High School (above the red star) and other businesses.
- The largest circles are in front of Port of Los Angeles High School and the Port of Los Angeles.
- Most of these circles are purple to dark blue, indicating cars are getting more tickets in the middle of and toward the end of the year.



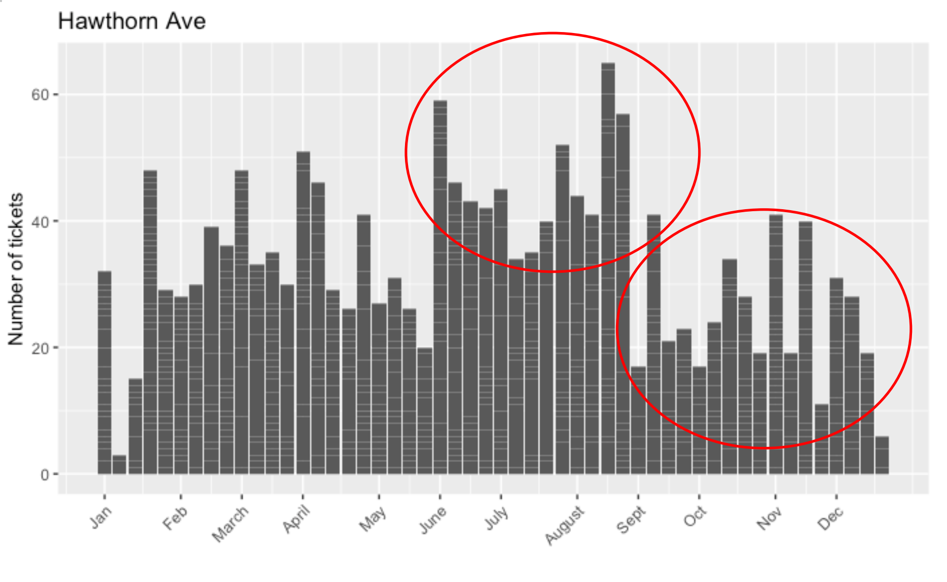
Hollywood: Hawthorn Ave.



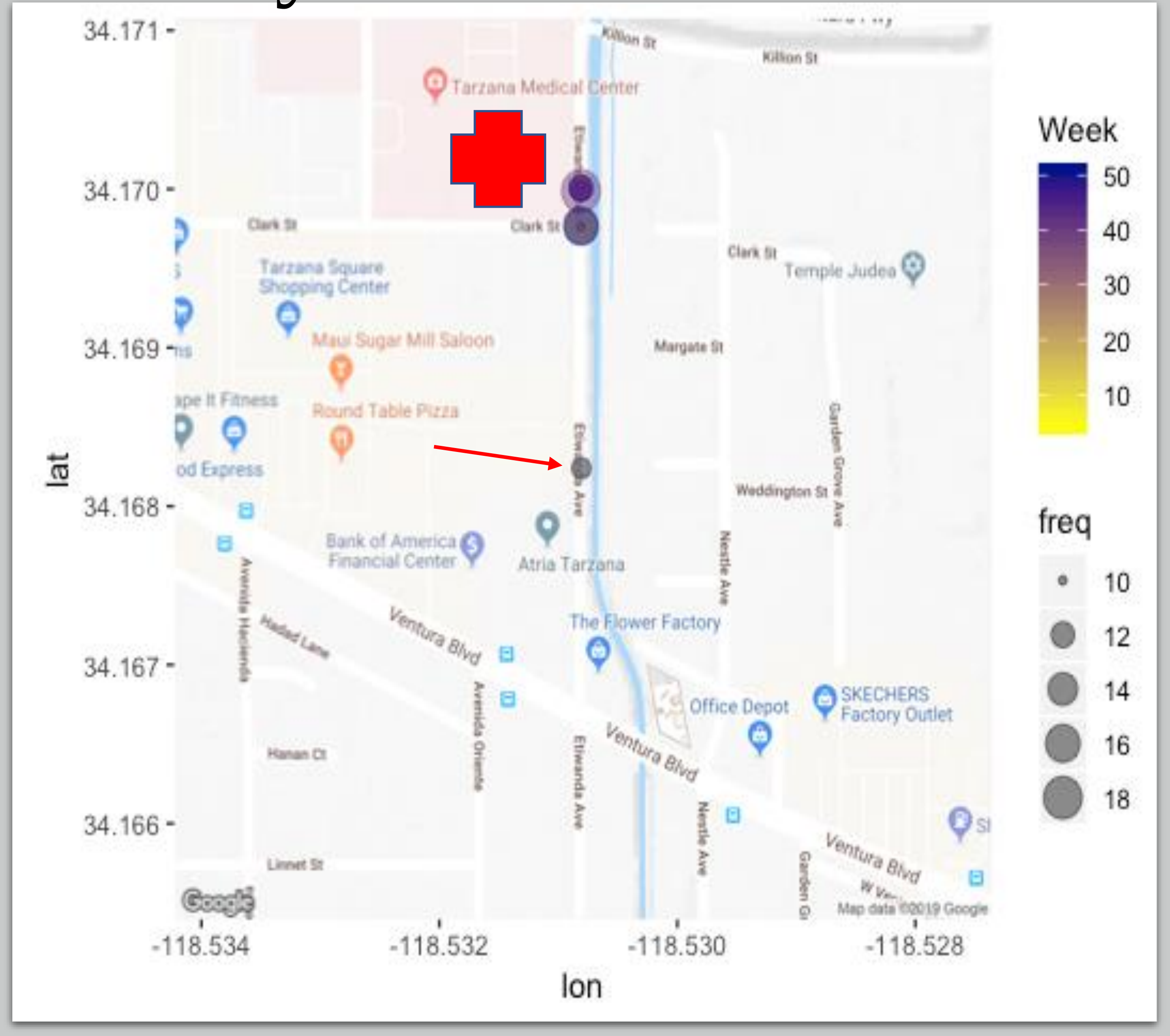
- Hawthorn Avenue is between Hollywood Blvd and Sunset Blvd near several popular tourist attractions and restaurants. This makes it an attractive street to park on for the day and walk around Hollywood.

- The largest circles are in front of Hollywood High School.

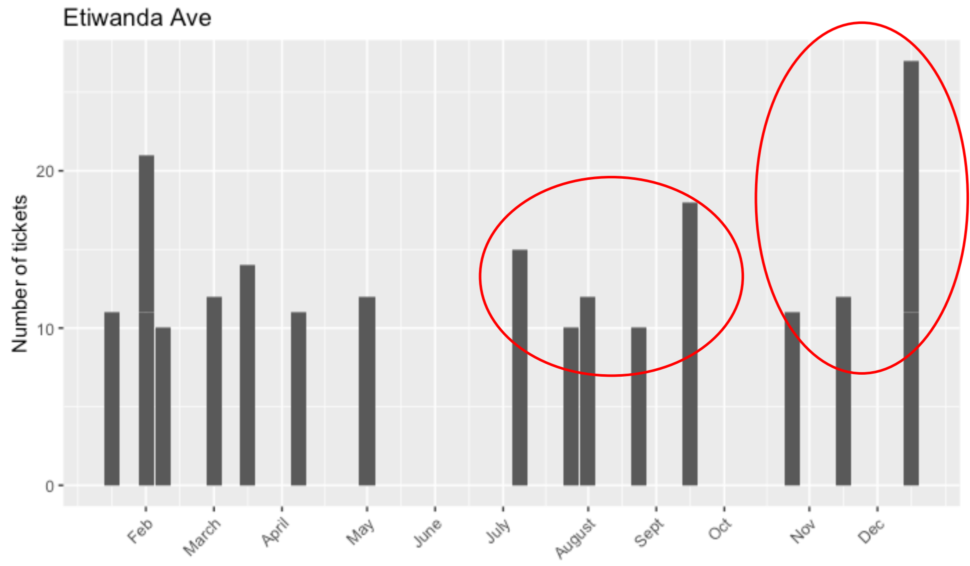
- Again, most of these circles are purple to dark blue, indicating cars are getting more tickets in the middle of and toward the end of the year.



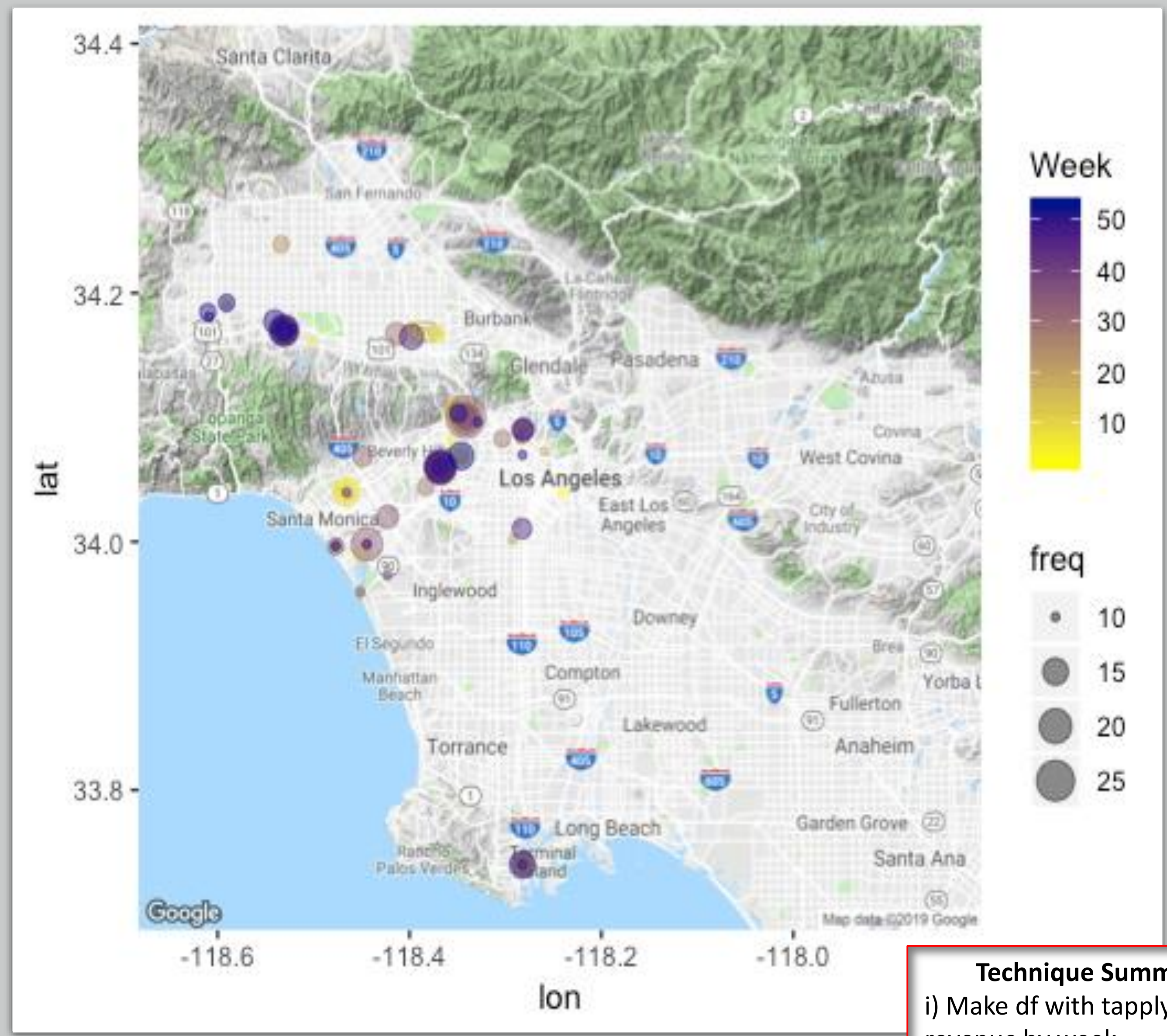
Van Nuys: Etiwanda Ave.



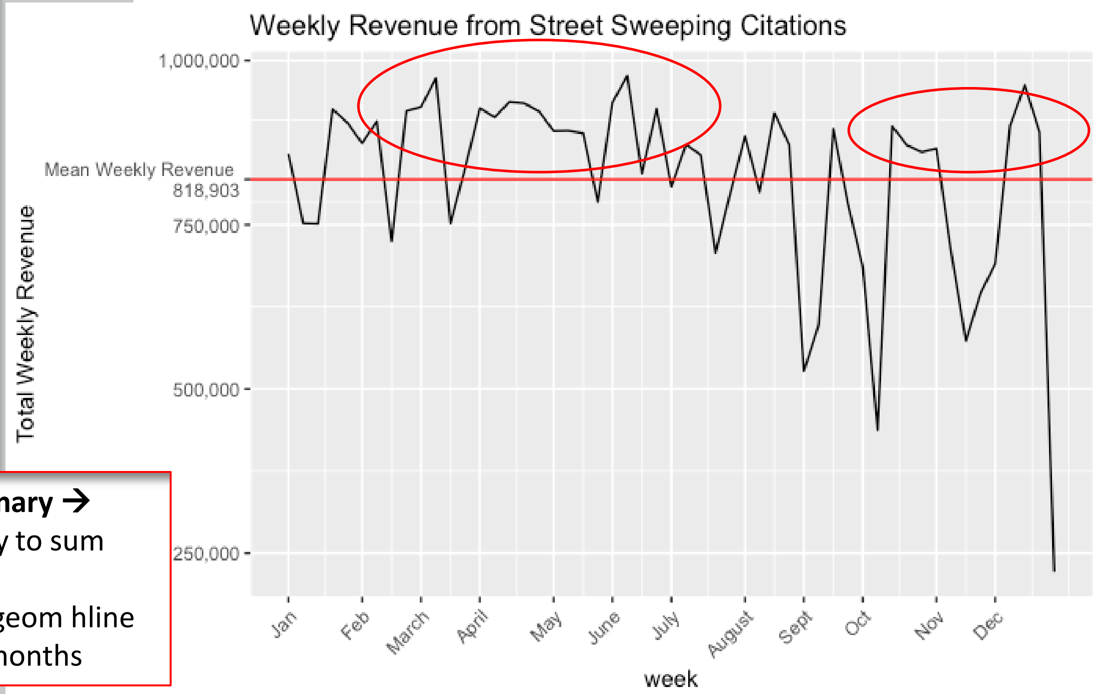
- The largest circles are near Tarzana Medical Center.
- The circles are generally darker which indicates weeks of high frequency are toward the end of the year.
- Does this trend continue throughout L.A. County?



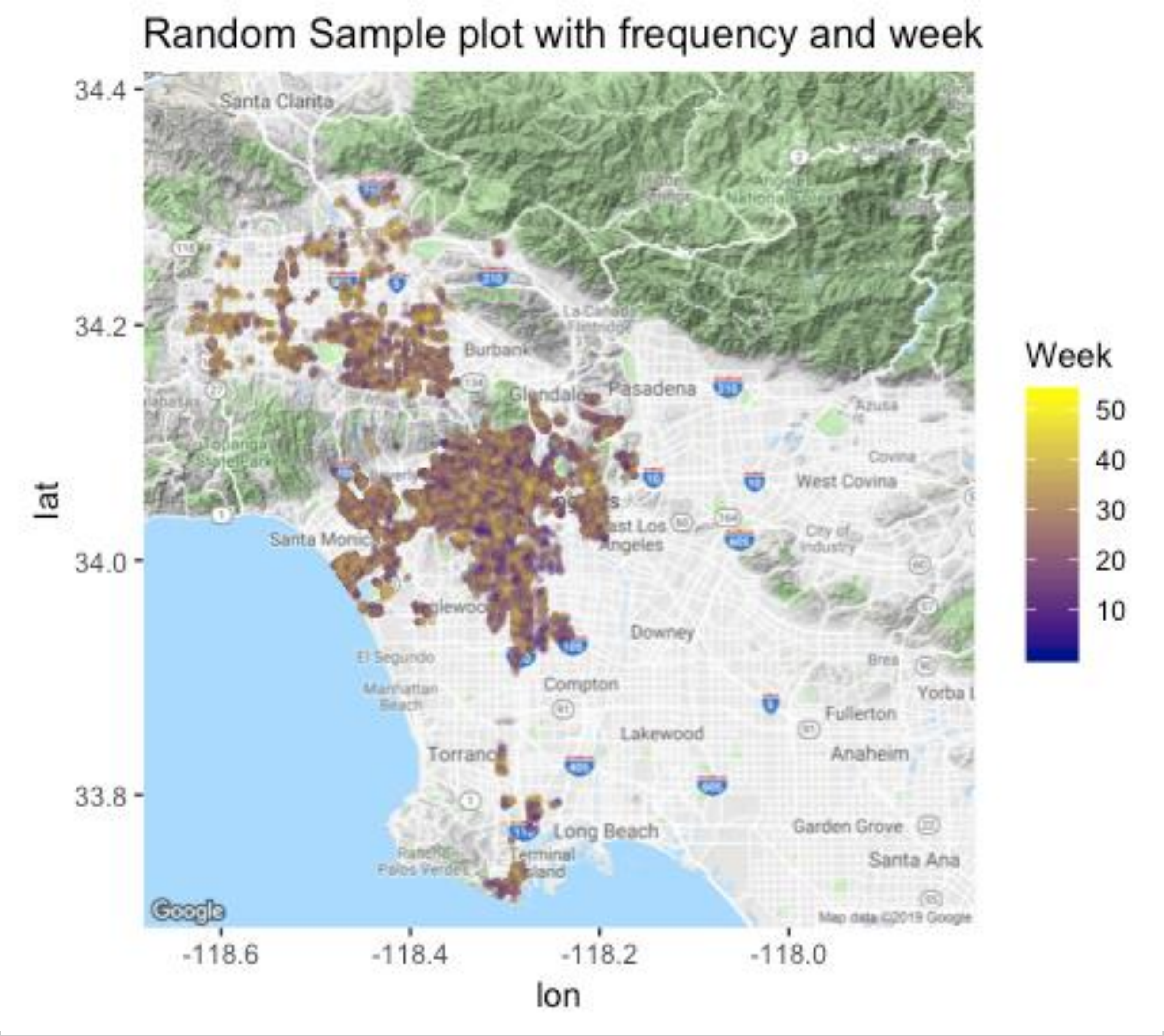
Areas with 10 or more tickets issued in one week



- ➢ This trend of the majority of tickets issued during the middle and at the end of the year is common throughout LA County as indicated by more of the purple to dark blue circles.
- ➢ This is further evidenced in the trends the weekly revenue earned.

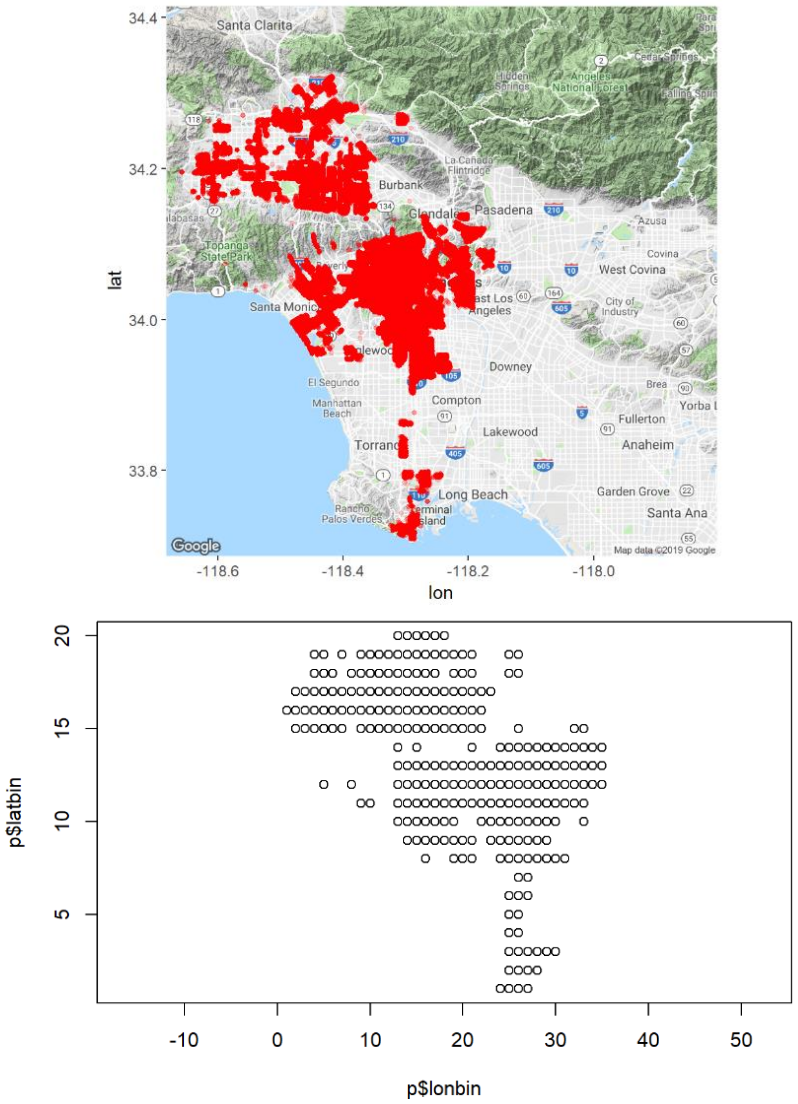


Technique Summary →
i) Make df with tapply to sum revenue by week
ii) ggplot/geom line/geom hline
iii) Break x-axis into months



- Random sample of 10% of the tickets (59,454) plotted by week.
- The trend continues. There are some dark blue and blue-purple dots mixed in with a majority of lighter yellow and lighter purple dots.
- When visiting L.A., especially April – August and November – December, be sure to double check the parking regulations.

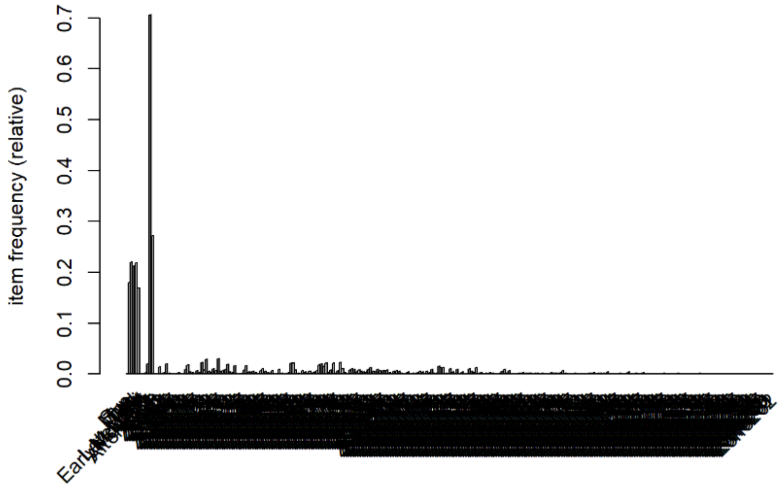
Association Rules - Geography



- Goal:
  - Develop an association rules model to examine relationships between day of week, time of day, and location
- Difficulty:
  - Location is coded as latitude and longitude, continuous numerical data
  - Association rules require logical data
- Solution:
  - Binning
    - Index of zones

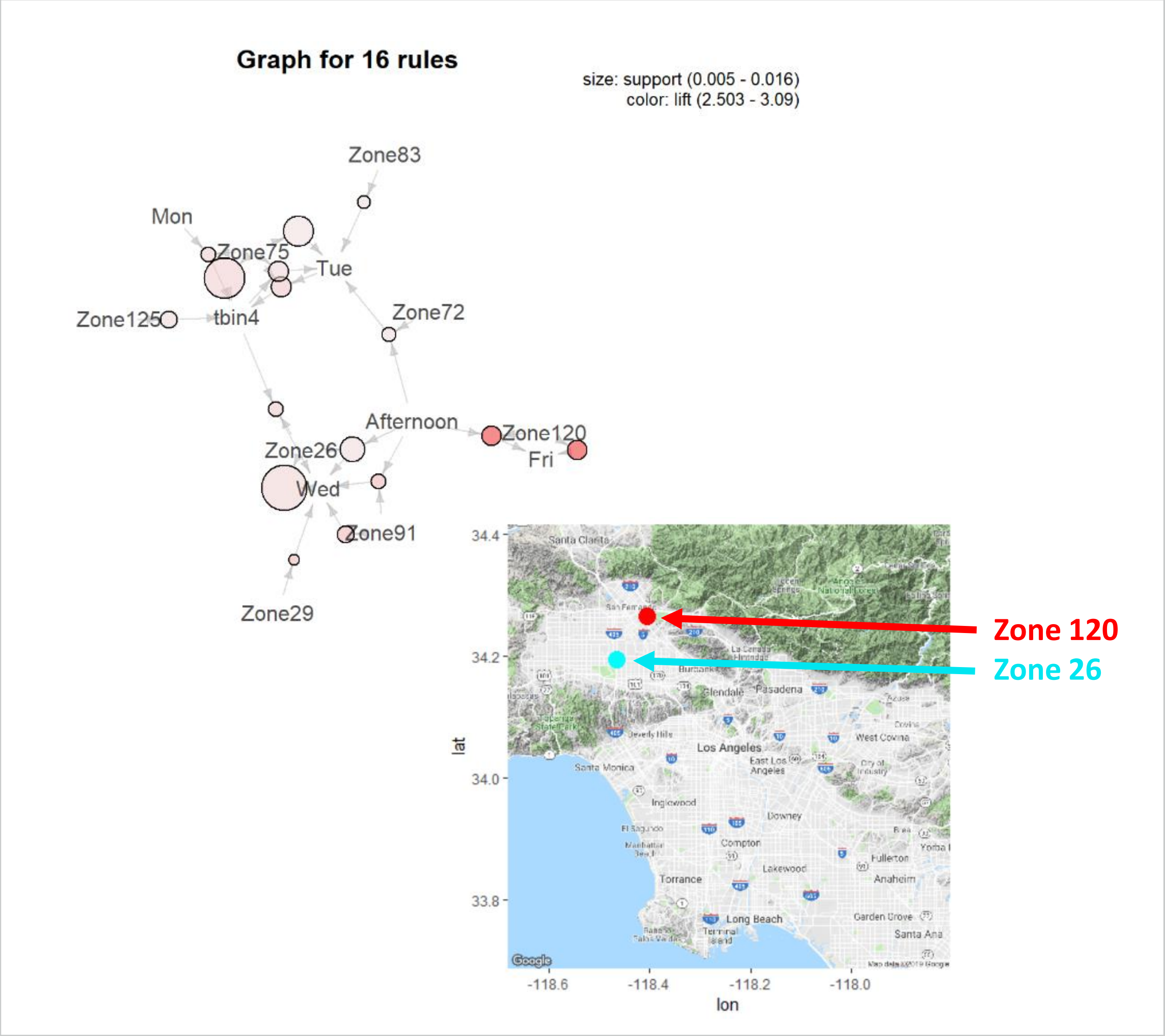
Association Rules - Logic

	day1	day2	day3	day4	day5	day6	day7	tbin1
1	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
4	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
7	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
8	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
9	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
10	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
11	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
12	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
13	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
14	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
15	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE



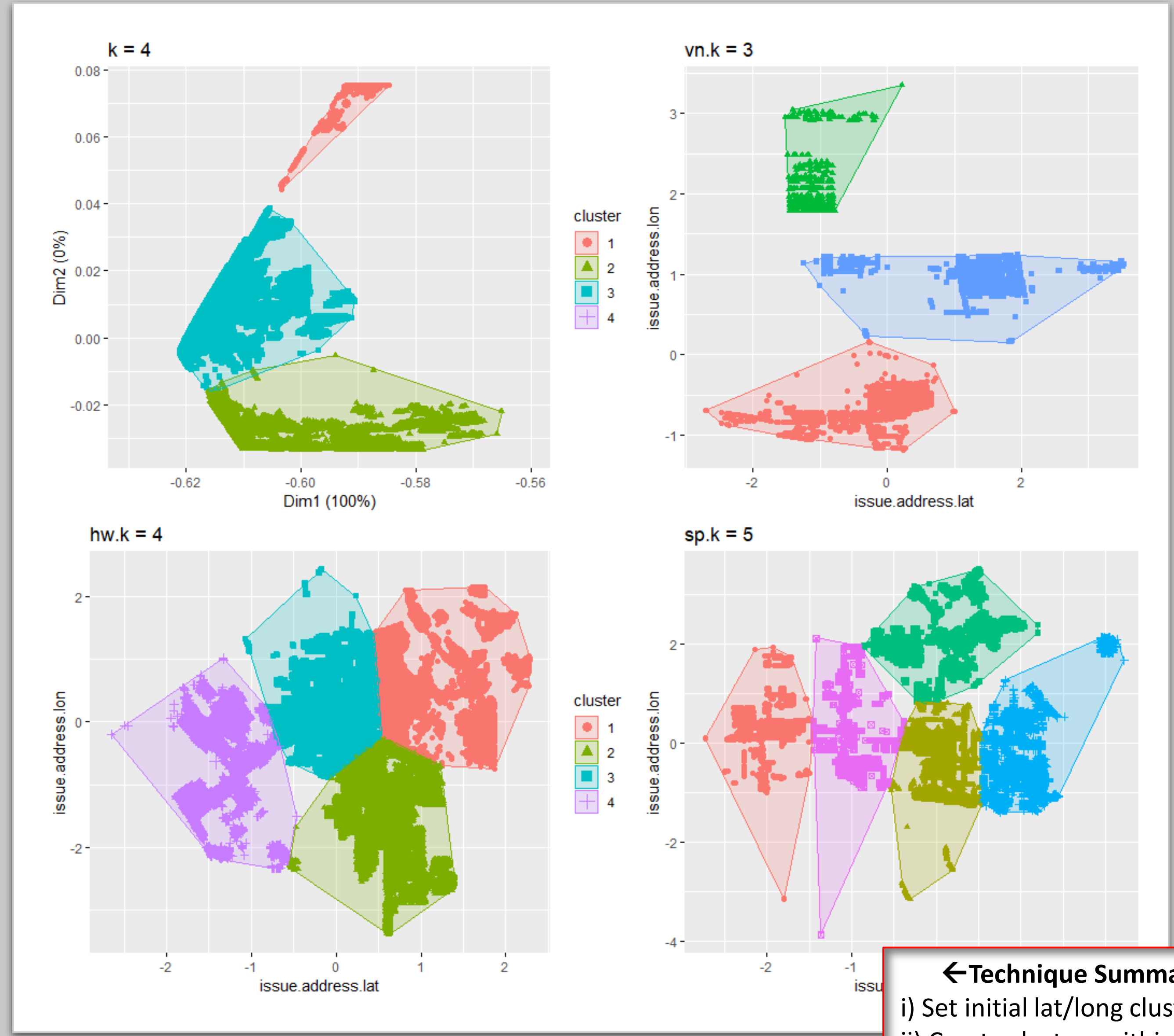
- Goal:
  - Develop an association rules model to examine relationships between day of week, time of day, and location
- Difficulty:
  - Location is coded as latitude and longitude, continuous numerical data
  - Association rules require logical data
- Solution:
  - Binning
    - Index of zones
  - Build a logical transaction grid
    - Each day of week, each time of day, each zone
    - For 244 possible variables
      - Of which each entry has only three TRUEs
    - Note: item frequency chart intentionally left impenetrable

Association Rules - The Rules



- The graph demonstrates that the relationship between Zone120 and tickets issued on Fridays is very strong – one would be advised to park elsewhere on that day
- The higher support (but lower lift) of the relationship between Zone26 and Wednesdays demonstrates that while the argument for causation is not as strong, there are a great many tickets issues in that zone on that day

Grouping Location: kMeans Clustering



←Technique Summary

i) Set initial lat/long clusters

ii) Create clusters within each initial cluster

iii) factoextra/cluster

➢ kMeans clustering results

The 'k = 4' plot shows that the citations issued in L.A. center around three distinct neighborhoods; Van Nuys (c1), Hollywood (c3), and San Pedro (c2). A fourth cluster is identified, which highlights data with missing lat/long values.

Following the initial city center clustering ('k = 4'), sub-clusters are built in each neighborhood. For example, Van Nuys ('vn.k = 3') has 3 clear centers where citations are most densely issued. The same sentiment is repeated for Hollywood and San Pedro.

The goal of using kMeans, in this situation, is to statistically identify districts within each neighborhood which can be used within a Random Forest prediction model. The prediction model will attempt to provide the number of citations issued per day. Having a concise location grouping may turn out to be predictive.

Data Aggregation: Model Prep

	X	ticket.number	issue.date	issue.year	issue.month	issue.day	issue.weekday	issue.time	issue.time.bin	agency.id	meter.id	route.id	issue.address	issue.address.lat
1	1	4323144652	2018-01-12T00:00:00	2018	1	12	6	01:05:00	1	56	0	00600	470 3RD ST E	-118.2500
2	2	4323144663	2018-01-12T00:00:00	2018	1	12	6	01:08:00	1	56	0	00600	400 SAN PEDRO ST S	-118.2430
3	3	4323144674	2018-01-12T00:00:00	2018	1	12	6	01:10:00	1	56	0	00600	400 SAN PEDRO ST S	-118.2430
4	4	4323144685	2018-01-12T00:00:00	2018	1	12	6	01:12:00	1	56	0	00600	400 SAN PEDRO ST S	-118.2430
5	5	4323144696	2018-01-12T00:00:00	2018	1	12	6	01:13:00	1	56	0	00600	400 SAN PEDRO ST S	-118.2430
6	6	4323144700	2018-01-12T00:00:00	2018	1	12	6	01:14:00	1	56	0	00600	400 SAN PEDRO ST S	-118.2430
7	7	4323144711	2018-01-12T00:00:00	2018	1	12	6	01:18:00	1	56	0	00600	515 6TH ST E	-118.2543
8	8	4323144722	2018-01-12T00:00:00	2018	1	12	6	01:21:00	1	56	0	00600	801 6TH ST E	-118.2574
9	9	4323144744	2018-01-12T00:00:00	2018	1	12	6	01:24:00	1	56	0	00600	600 KOHLER ST	-118.2405
10	10	4323144755	2018-01-12T00:00:00	2018	1	12	6	01:25:00	1	56	0	00600	600 KOHLER ST	-118.2405
11	11	4323144766	2018-01-12T00:00:00	2018	1	12	6	01:26:00	1	56	0	00600	600 KOHLER ST	-118.2405
12	12	4323144770	2018-01-12T00:00:00	2018	1	12	6	01:27:00	1	56	0	00600	600 KOHLER ST	-118.2405
13	13	4323144781	2018-01-12T00:00:00	2018	1	12	6	01:28:00	1	56	0	00600	600 KOHLER ST	-118.2405
15	15	4323144814	2018-01-12T00:00:00	2018	1	12	6	01:33:00	1	56	0	00600	600 KOHLER ST	-118.2405
16	16	4323144825	2018-01-12T00:00:00	2018	1	12	6	01:34:00	1	56	0	00600	600 KOHLER ST	-118.2405

	fineCnt	time3p	time4p	wkDay2p	wkDay3p	wkDay4p	wkDay5p	wkDay6p	holidayp	cityCntVn1p	cityCntVn2p	cityCntVn3p
1	3040	0.7092105	0.2746711	0	1	0	0	0	0	0.039802632	0.0006578947	0.0013157895
2	2547	0.7051433	0.2536317	0	0	1	0	0	0	0.025520220	0.0047114252	0.0019630938
3	2464	0.6911526	0.2966721	0	0	0	1	0	0	0.013798701	0.0008116883	0.0012175325
4	1744	0.7522936	0.2247706	0	0	0	0	1	0	0.005733945	0.0017201835	0.0063073394
7	2086	0.7104506	0.2818792	1	0	0	0	0	0	0.028283797	0.0000000000	0.0009587728
8	47	0.6808511	0.0000000	0	1	0	0	0	0	0.000000000	0.0212765957	0.0212765957
9	2199	0.6807640	0.2792178	0	0	1	0	0	0	0.025466121	0.0000000000	0.0090950432
10	2363	0.6969953	0.2814219	0	0	0	1	0	0	0.007617435	0.0004231909	0.0050782903
11	1854	0.6990291	0.2729234	0	0	0	0	1	0	0.009169364	0.0000000000	0.0021574973
14	2636	0.6904401	0.3031108	0	1	0	0	0	0	0.020485584	0.0000000000	0.0121396055
15	2175	0.6565517	0.3066667	0	0	1	0	0	0	0.023448276	0.0068965517	0.0013793103
16	2200	0.7000000	0.2736364	0	0	0	1	0	0	0.014090909	0.0018181818	0.0040909091
17	1623	0.7473814	0.2230437	0	0	0	0	1	0	0.009242144	0.0018484288	0.0043130006
20	2139	0.7040673	0.2875175	1	0	0	0	0	0	0.003740065	0.0000000000	0.0014025245
21	2241	0.6742526	0.3154842	0	1	0	0	0	0	0.018295404	0.0022311468	0.0071396698

- Data preparation for Random Forest

The original Los Angeles street sweeping citations data provided information for each individual citation.

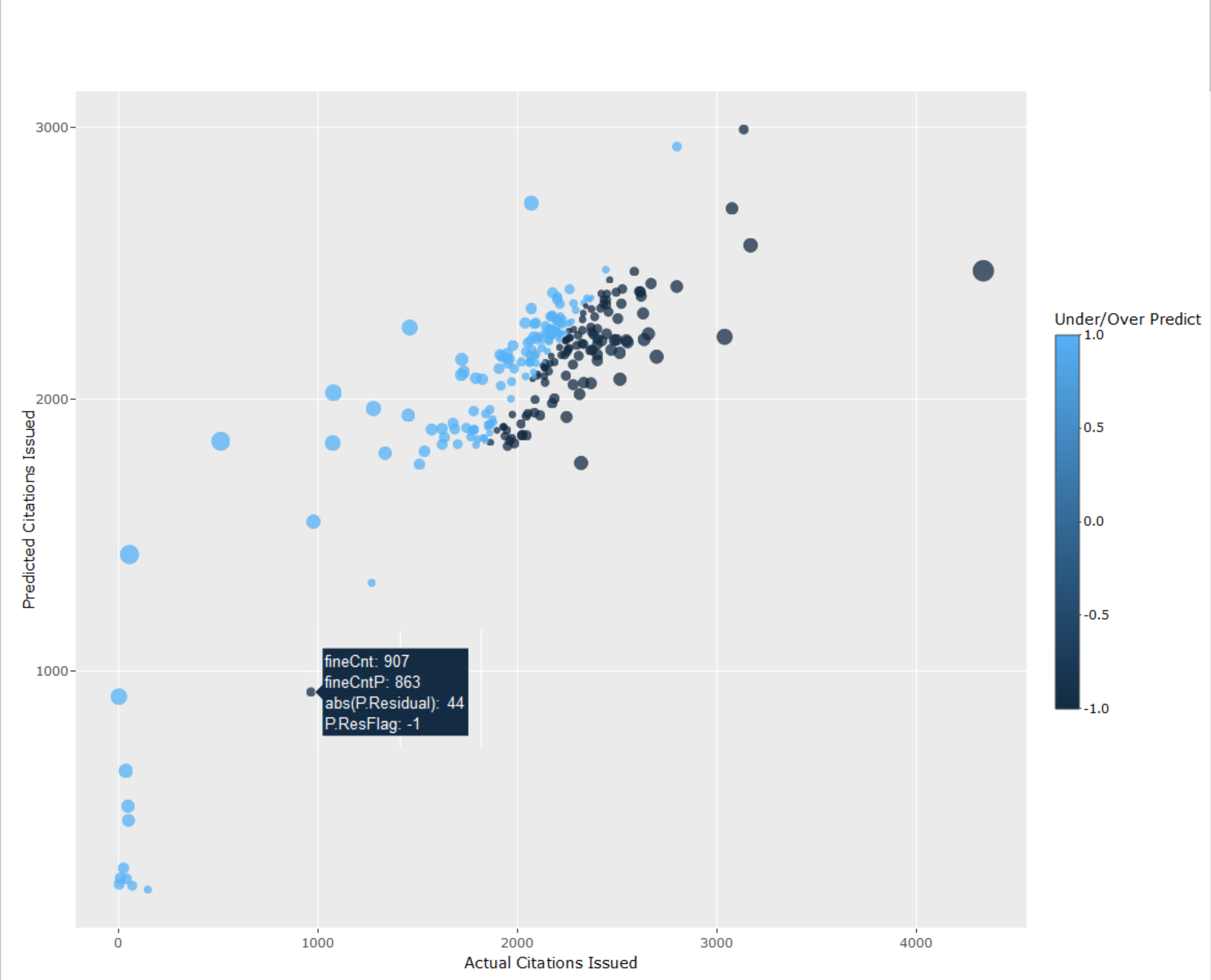
Using the ‘ddply’ function (‘plyr’ package), a summary table showing the number of citations issued by day was created. The original view of the table provided granular citation counts by time bin, weekday, location, LA DOT agency, and cited car descriptions. For example;

```
p.summary2 <- ddply(p,"issue.calday",summarise,
                    fineCnt = NROW(which(X>=0)),
                    fineAmt = NROW(which(X>=0))*73,
                    time3 = NROW(which(issue.time.bin==3)),
                    time4 = NROW(which(issue.time.bin==4)),
                    wkDay2 = NROW(which(issue.weekday==2)),
                    wkDay3 = NROW(which(issue.weekday==3)),
                    wkDay4 = NROW(which(issue.weekday==4)),
                    wkDay5 = NROW(which(issue.weekday==5)),
                    wkDay6 = NROW(which(issue.weekday==6)),
                    holiday = NROW(which(holiday.ind==1)),
```

The initial summary table had to then be transformed to show ratios instead of raw counts if a citation count prediction model is to be built. For example;

```
# Add the 'percent of' ratios to the summary data frame
p.summary2$time3p <- p.summary2$time3/p.summary2$fineCnt
p.summary2$time4p <- p.summary2$time4/p.summary2$fineCnt
```

Prediction Model: randomForest



←Technique Summary

i) Train Random Forest

ii) Create variables in main table for residual and pos/neg

iii) randomForest/plotly

➢ Random Forest Results

Variable Importance (top 10)

1. cityCntSp3p – San Pedro Sub-Cluster 3
2. agency53p – LA City DOT Agency 53
3. cityCntSp4p – San Pedro Sub-Cluster 4
4. agency51p – LA City DOT Agency 51
5. cityCntSp5p – San Pedro Sub-Cluster 5
6. plateCAp – Citation issued to drive with CA plate
7. cityCntSp2p – San Pedro Sub-Cluster 2
8. plateNCAp – Citation issued to driver with an out-of-state plate
9. time3p – Citation issued in the morning
10. agency54p – LA City DOT Agency 51

Random Forest prediction performance

Mean citations issued per weekday – 2,049

Standard Deviation citations issued per weekday – 563.11

<u>Run 1 without observed holiday flags and 500 trees</u>

RMSE: 283.0865

Actual citations / predicted citations: -4.10% [over prediction]

<u>Run 2 with observed holiday flags and 500 trees</u>

RMSE: 283.4215

Actual citations / predicted citations: -4.14% [over prediction]

<u>Run 3 with observed holiday flags and 5000 trees</u>

RMSE: 281.7746

Actual citations / predicted citations: -4.11% [over prediction]

Mean predicted citations issued: 2,072

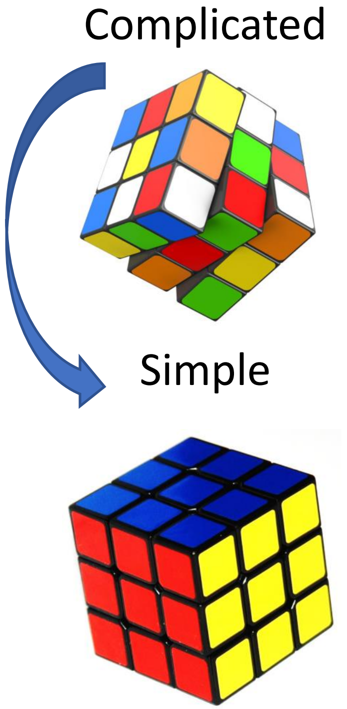
Standard Deviation citations issued: 410.82

Summary

The third model run provides an okay prediction of the number of citations issued per day; with a root mean square error of 282. This means the model that there are more instances where the model has predicted more citations than what actually happened (by 4% on avg.). The plotted residuals show the model does very well when predicting a normal day, during 2018. However, the model has a hard time finding outlier. Therefore a flag for observed holidays was included in the 2nd and 3rd run, at least in part. Since our group's goal is to alert customers of targeted areas/days, it was best to over predict (than under).

Outcomes & Proposals

Parking tickets are a reality of urban dweller living... this study hopes to support angel investor funding for an application to help consumers park more inexpensively and strategically in cities near you. Analysis supports development of a smart phone “notification” app could be developed based on high-quality meter reader data being generated in large metropolitan cities.



Cons \ Heartaches

- Data quality issues \ cleaning for real-time event application

Pros \ Benefits

- Avg tow costs $100
- Price of annual app expense should reduce 1 negative event
- Assist out of state visitors with an adverse event
  - advertise at car rental places at airports

References: https://www.spotangels.com/blog/los-angeles-street-cleaning-holidays-and-rules/

Data Dictionary – Key Fields – 594,546 x 27 Variables

Data Frame – LA Tickets		
Correlation	**data.frame': 594546 obs. of 27 variables:**	**Type**
	$ X : int 1 2 3 4 5 6 7 8 9 10 ...	case count
	$ ticket.number : num 4.32	ticket id
	$ issue.date : Factor w/ 320 levels "2	cas
	$ issue.year : int 2018 2018 20	only 2018
Yes	$ issue.month : int 1 1 1 1 1 1	month
Yes	$ issue.day : int 12 12 12 12	day
Yes	$ issue.weekday : int 6 6 6 6 6	weekday
	$ issue.time : Factor w/ 848 levels "00:00:00"	
Yes	$ issue.time.bin : int 1 1 1 1 1 1	daily time bin
Yes	$ agency.id : int 56 56 56 56	numeric
	$ meter.id : Factor w/ 136 levels "0","48","CP170",.	not used
Yes	$ issue.address.lat : num -118 -118 -118 -118 -118 ...	time
Yes	$ issue.address.lon : num 34.1 34 34 34 34 ...	numeric
	$ violation.id : Factor w/ 1 level "80.69BS": 1 1 1 1 1	numeric
	$ violation.desc : Factor w/ 1 level "NO PARK/STREET CLEAN": 1	all street sweeping
	$ violation.fine.amt : int 73 73 73 73 73 73 73 73 73 73 ...	numeric
Yes	$ plate.expire.date : int 201801 201803 201801	numeric
	$ plate.expire.year : int 2018 2018 2018 2018 2018	
Yes	$ plate.expire.month : int 1 3 1 4 3 3 1 8 10 1 ...	~~all sweep~~
Yes	$ plate.expire.flag : int 0 0 0 0 0 0 0 0 1 0 ...	numeric
Yes	$ car.make.import.flag: int 0 0 1 0 0 0 1 1 1 1 ...	numeric
Yes	$ route.id : Factor w/ 674 levels "0","00001",	numeric
Yes	$ plate.state : Factor w/ 73 levels "AB","AK","AL",..: 7	numeric
Yes	$ car.make : Factor w/ 62 levels "ACUR","ALFA",..: 43	numeric
Yes	$ car.bodystyle : Factor w/ 12 levels "BU","CM","MC",..: 7 7	numeric
Yes	$ car.color : Factor w/ 16 levels "BG",	numeric
Yes	$ issue.address : Factor w/ 262194 levels "! % CULVER BLVD",..:	numeric

https://www.kaggle.com/cityofLA/los-angeles-parking-citations#parking-citations.csv