Univariate analyses

Check for daylight saving shift! Ok!

Bivariate analyses

Study the relation between **overall and specific place check-ins for:**

- check-in count
- time
- location
- accuracy

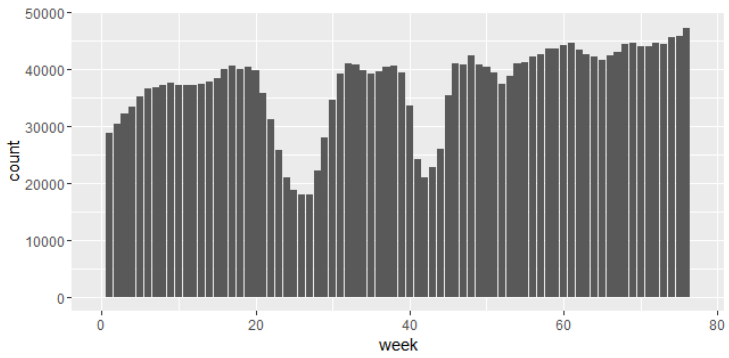
Time is analyzed based on the hour of the day, the day of the week and the week, yearly effects are researched as well.

Overall analysis

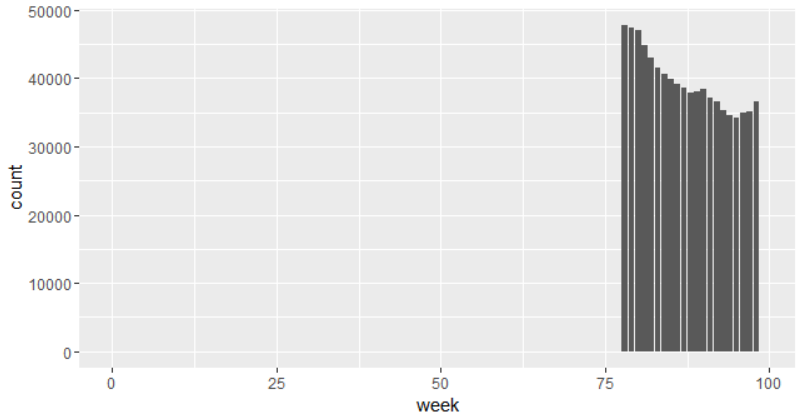
Analysis I: Check-in count versus time (univariate)

No obvious relation with hour and day of week. But related to the week!

Train



Test

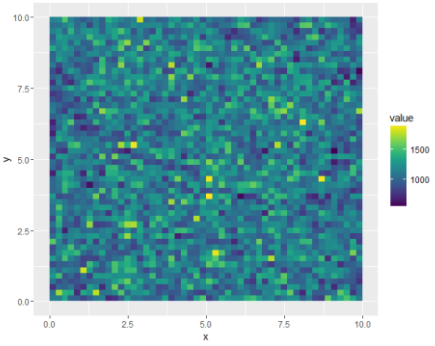


Maybe some places are more popular during popular times?

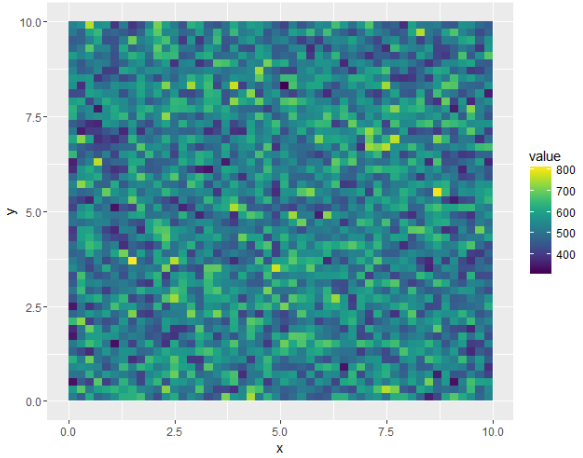
Add predictor that captures total week density (extrapolate last week) just in case!

Analysis II: Check-in count versus location (univariate)

Total density in region and size of place



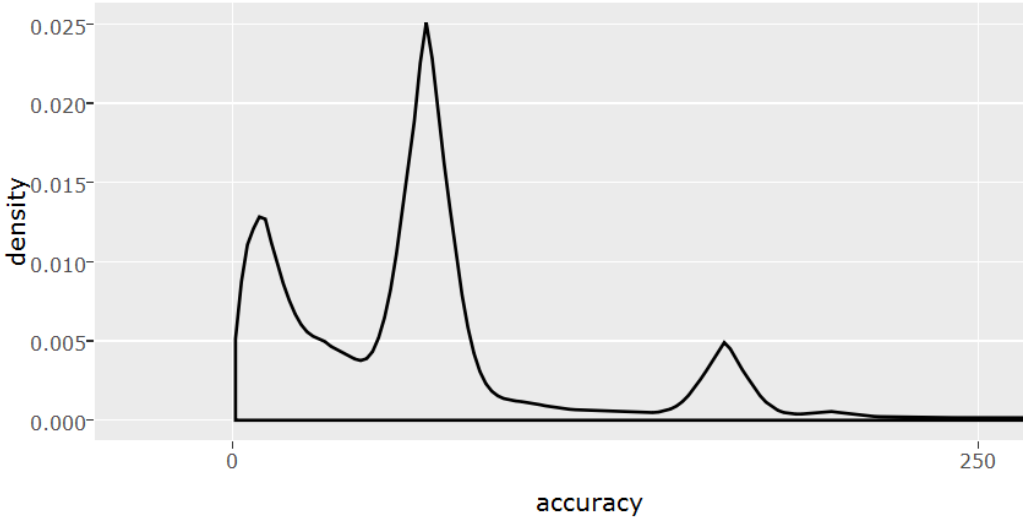
Shows the observation count => no obvious pattern



Shows the mean place count, no obvious pattern

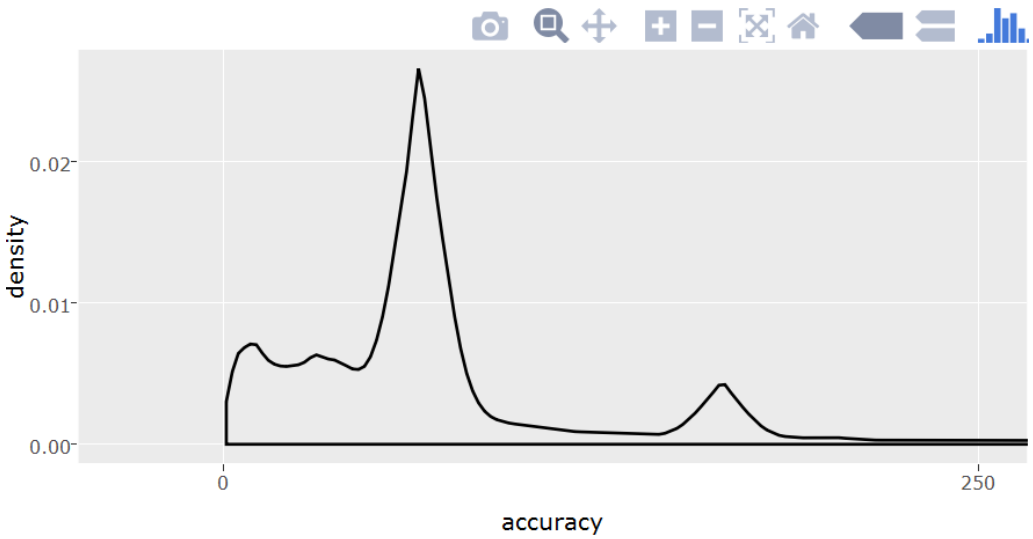
Analysis III: Check-in count versus accuracy (univariate)

Train



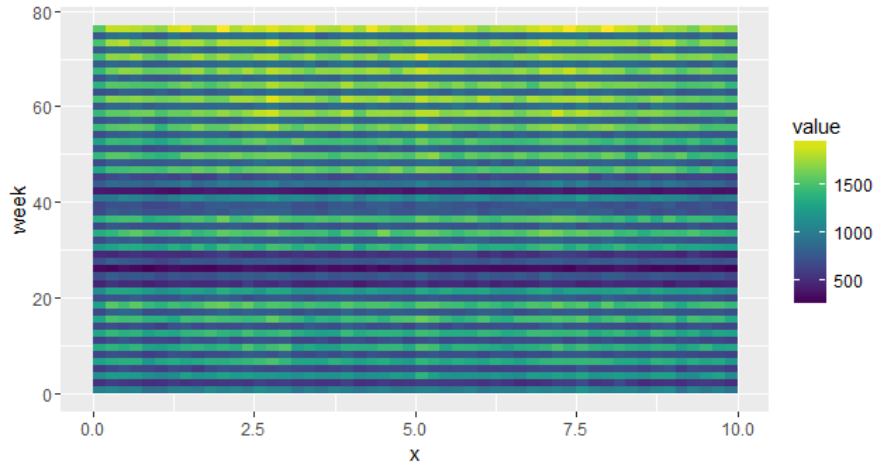
Interesting! Can these three peaks be explained??

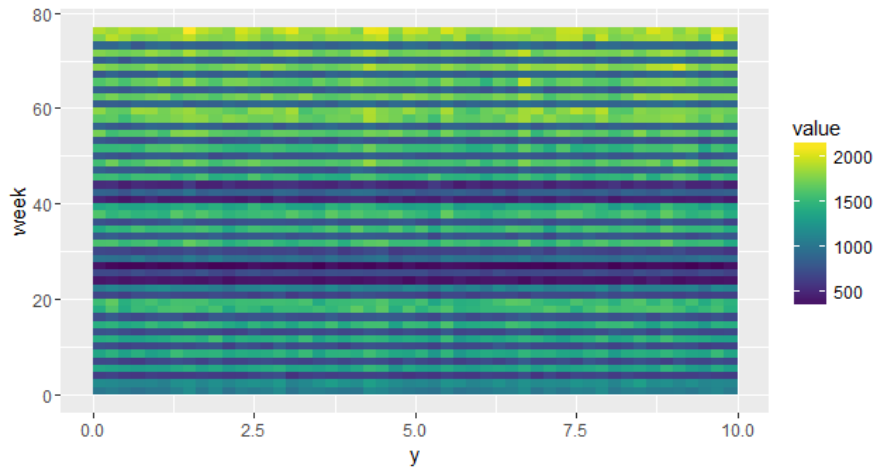
Test – No initial peak



Analysis IV: Week versus location

Weekly frequency visible, lower check-in rate at x and y edges but CRAZY time pattern – plot artefact or real? It was verified to be a plot artefact.





Analysis V: Day of the week versus location

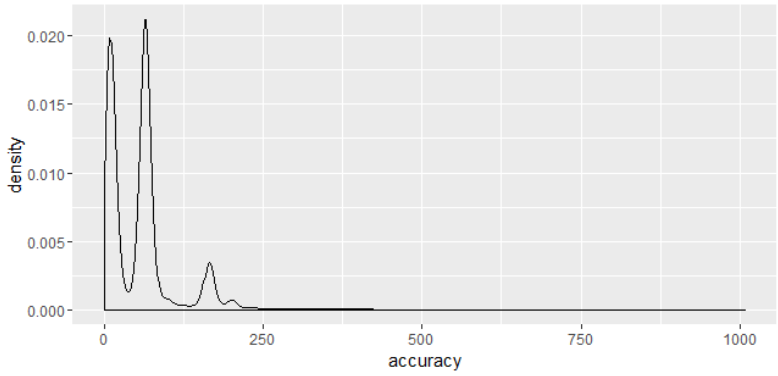
No obvious patterns

Analysis VI: Hour of the day versus location

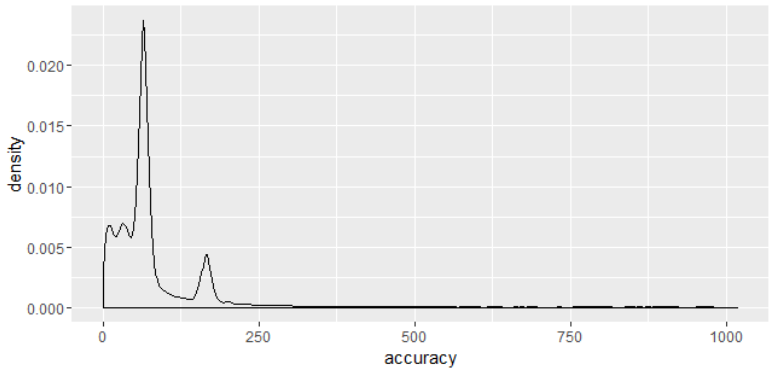
No obvious patterns

Analysis VII: Week versus accuracy – sliding density

Week 1-4:

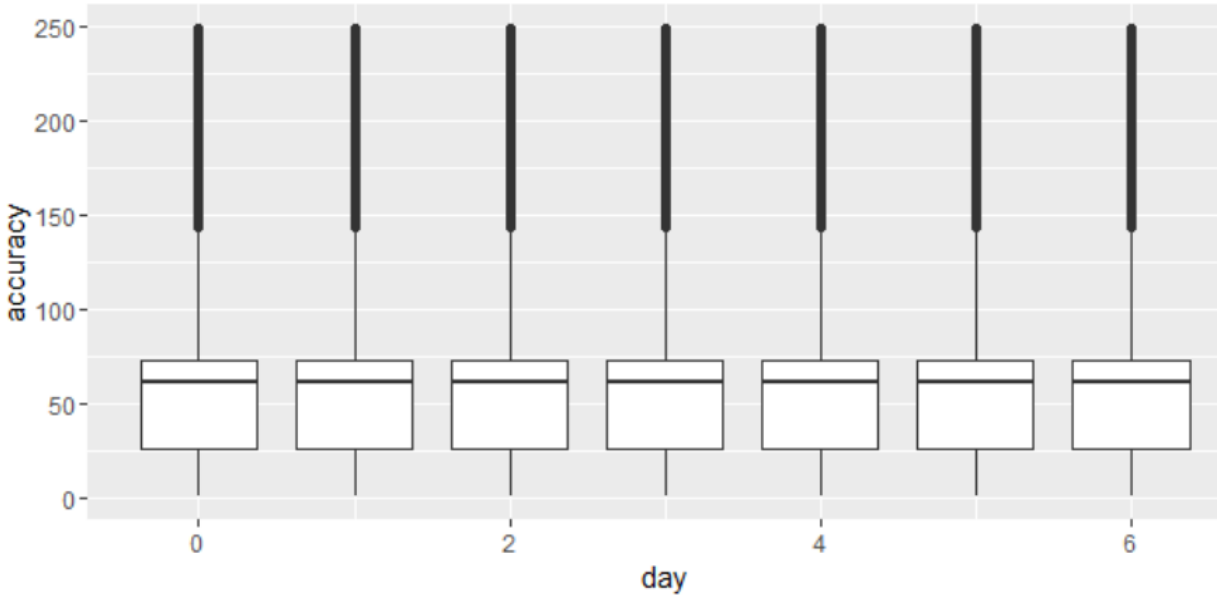


Week 72-76:

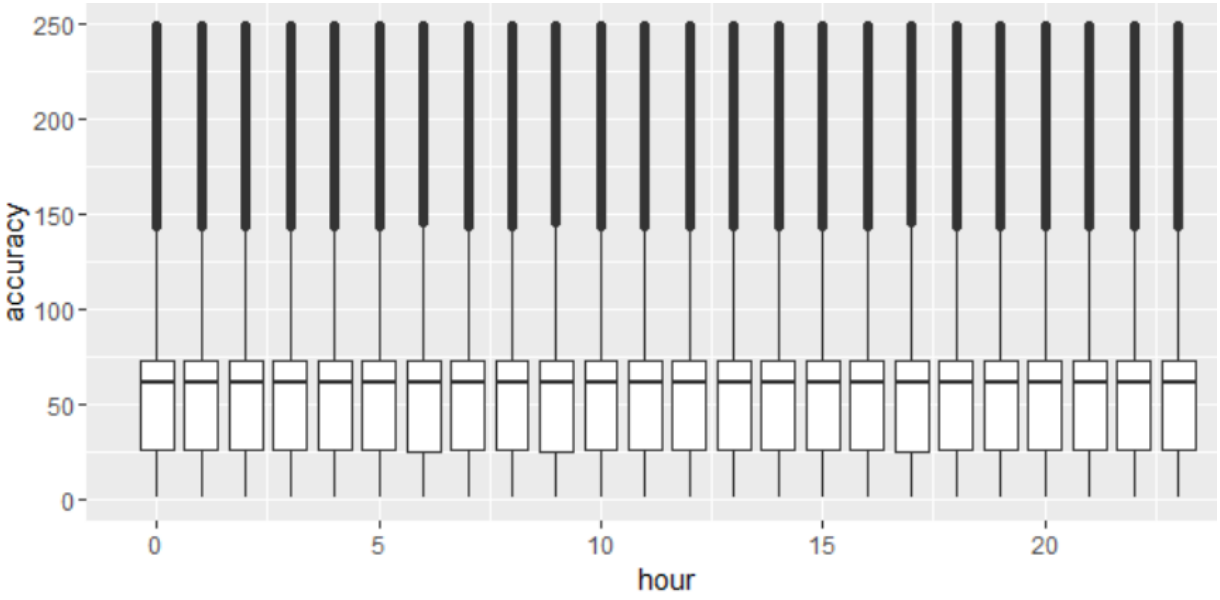


Calculate quantile feature of accuracy in week!

Analysis VIII: Day of the week versus accuracy

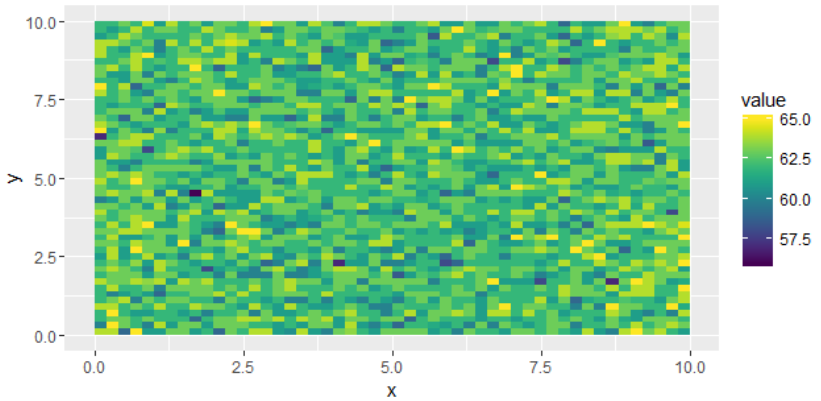


Analysis IX: Hour of the day versus accuracy



Analysis X: Location versus accuracy

Median accuracy



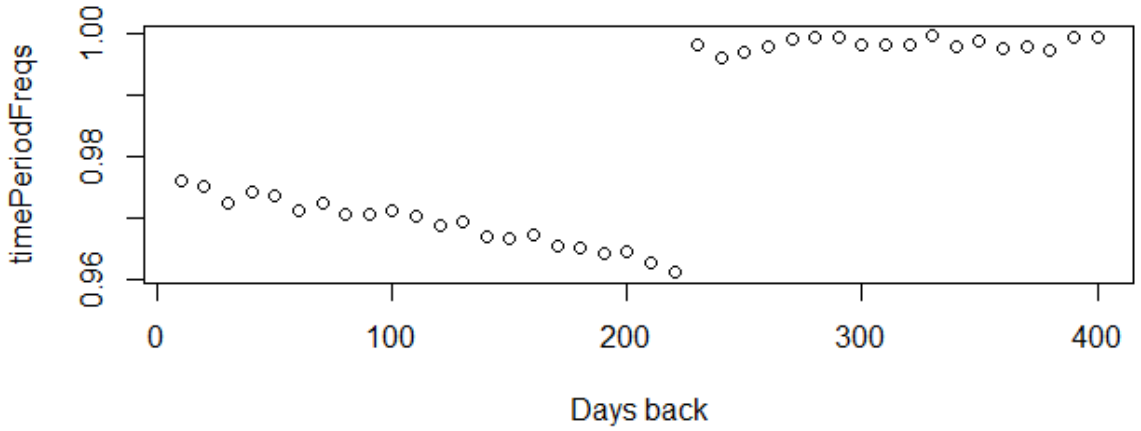
Specific place analysis

Analysis I: Check-in count versus time

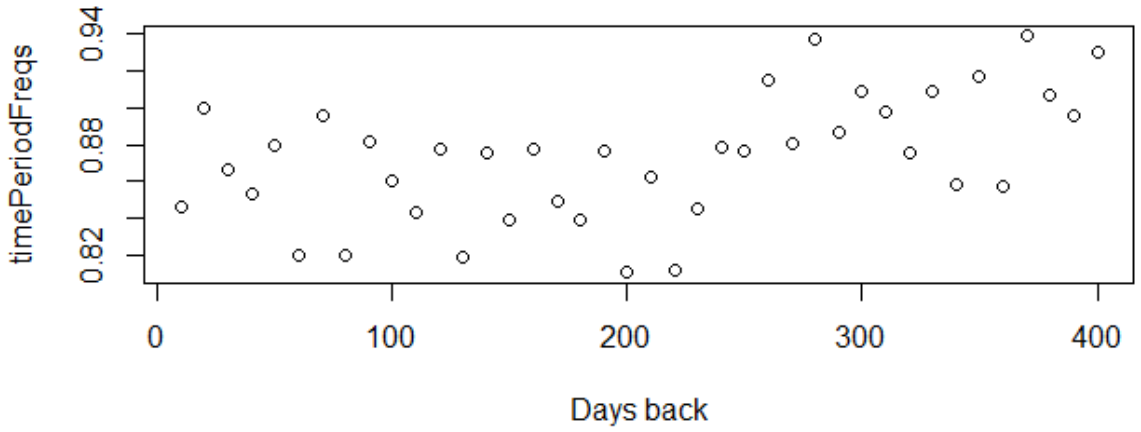
Year

No elevated density when looking a calendar year back in time

BUT this one is interesting: looking at a +- 7 days window, days back, what fraction of the places did observe a checkin during that time frame



Window of +- 1 day gives a more distorted picture:



It seems like the historical activity (>225 days) is a better indicator than recent check-in events!!

Other interpretation: hardly any new places after 225 days?

Add time period density rescaled indicators of historical activity with focus on days back>225!

Maybe weekly density for each place id? => Matrix rather than vector

Week

It's very hard to identify patterns or even cluster of patterns => non parametric approach:

Time density corrected counts and relative time difference between new observations and KNN counts

Day of week

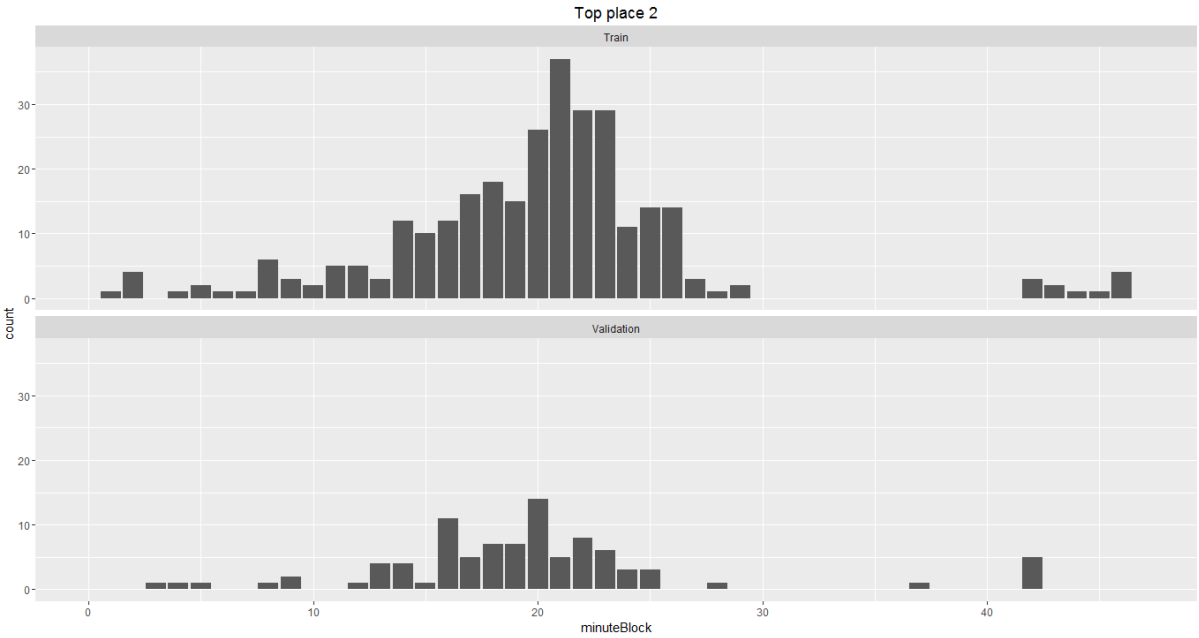
Obvious pattern. Also include a feature that smooths the counts and a feature that uses the relaxed densities

Hour of day

Obvious pattern. Also include a feature that smooths the counts and a feature that uses the relaxed densities

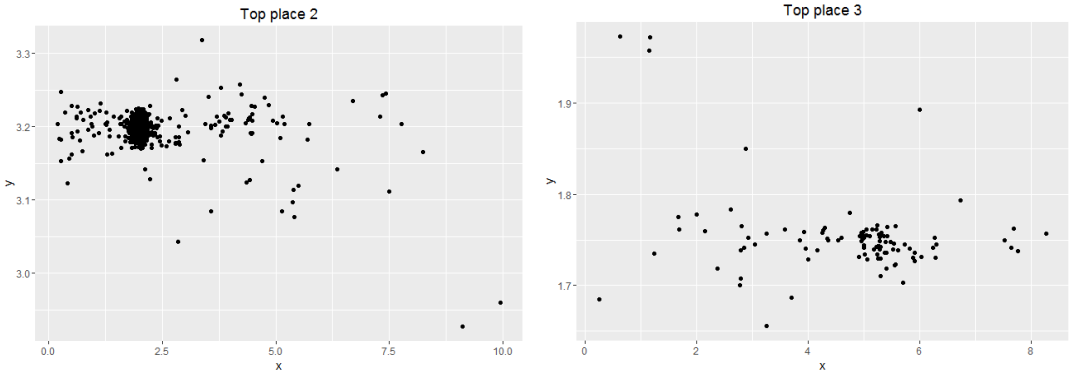
X-minute period of day

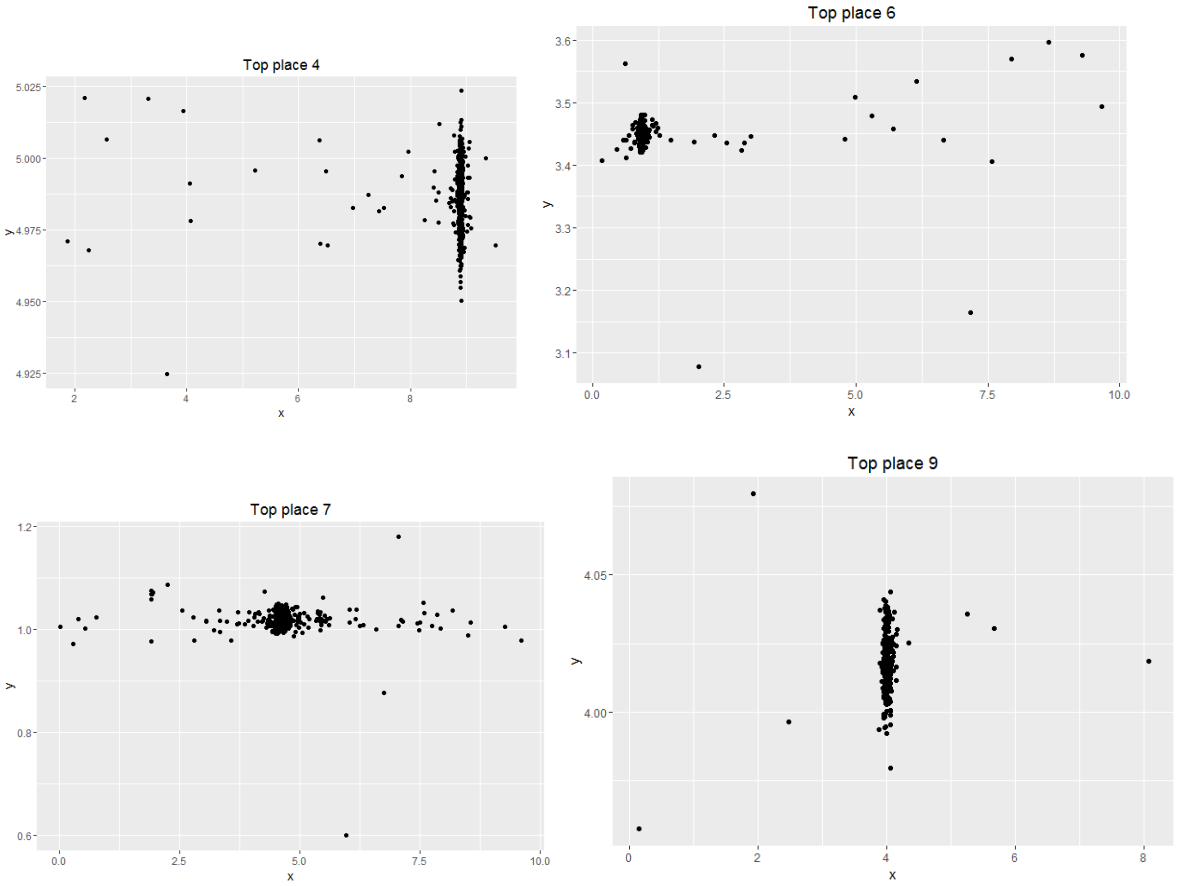
Going beyond hour does not seem to make sense since the densities between the train and validation period do not align. Half hour plot:



Analysis II: Check-in count versus location

Same observation here as with the week trend analysis: it is very hard to derive general patterns. There are often strong outliers. Non parametric approach seems most reasonable here as well (multiple KNN counts). The robust Z score using med and mad are also likely relevant.



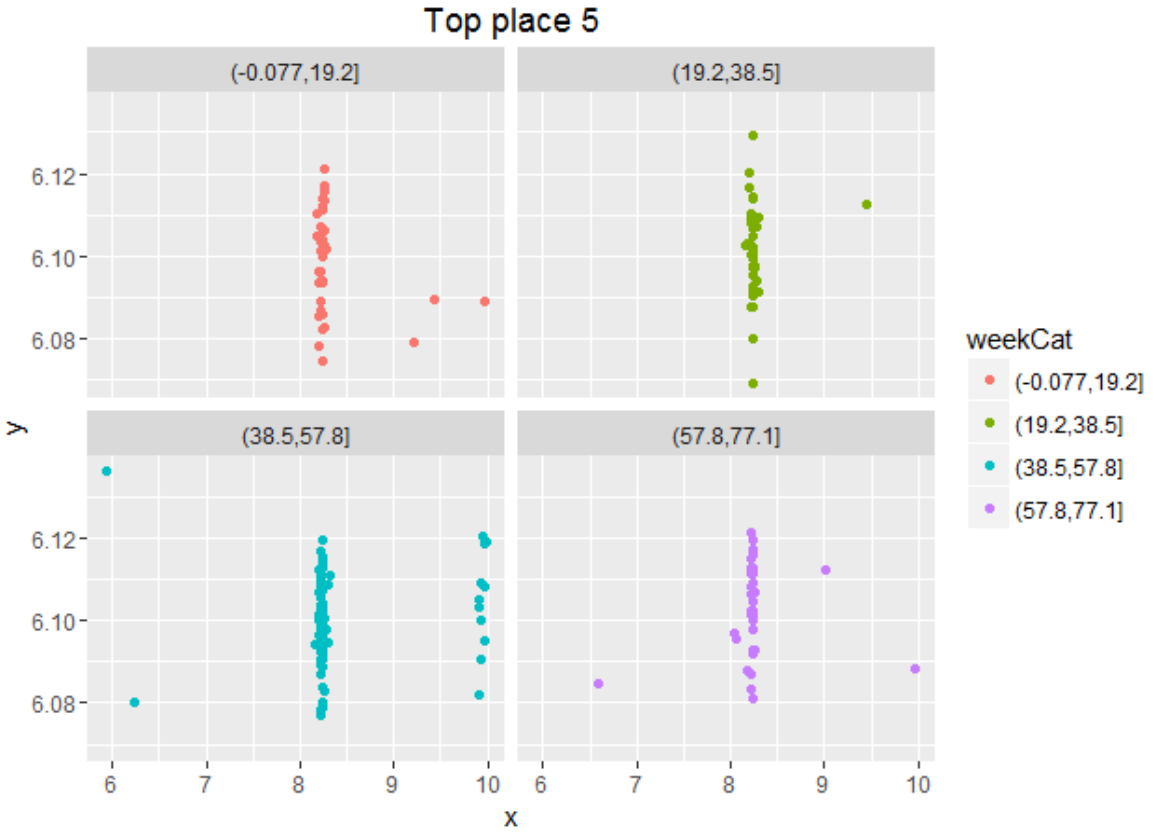


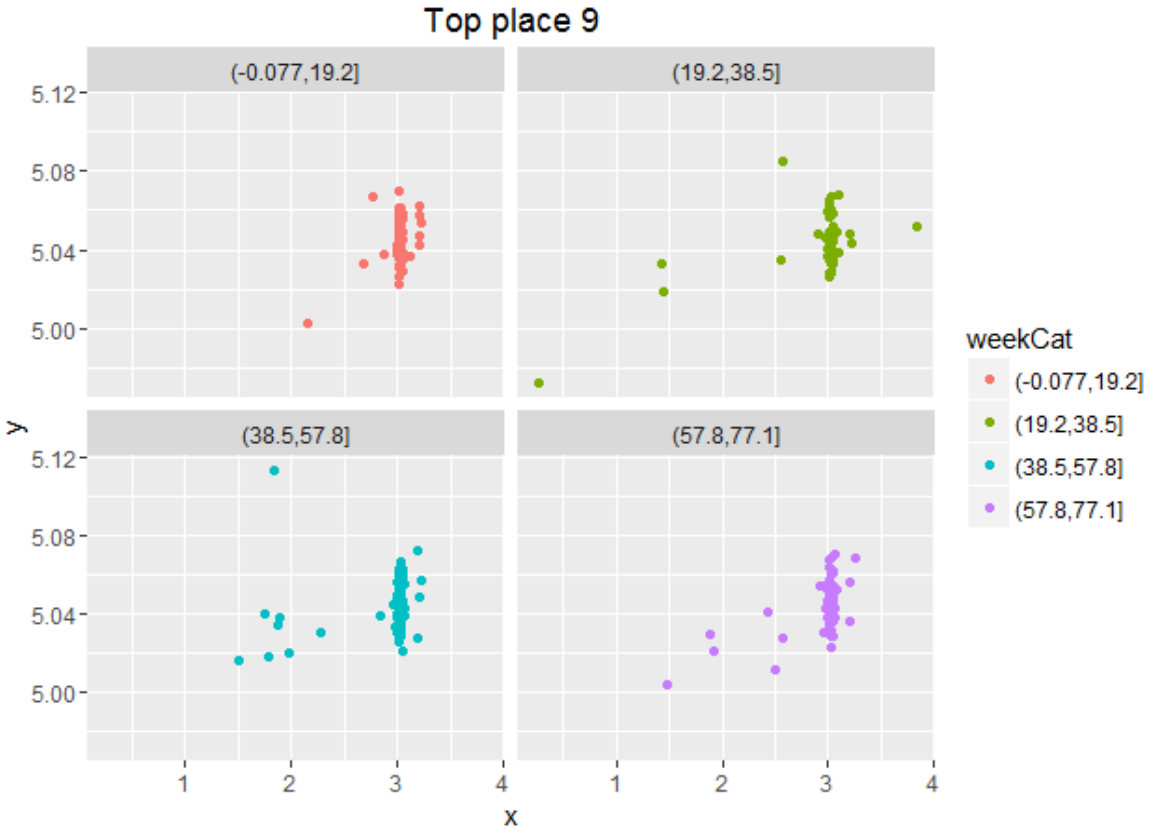
Analysis III: Check-in count versus accuracy

All different densities, as suggested before: store quantiles for places?

Analysis IV: Week versus location

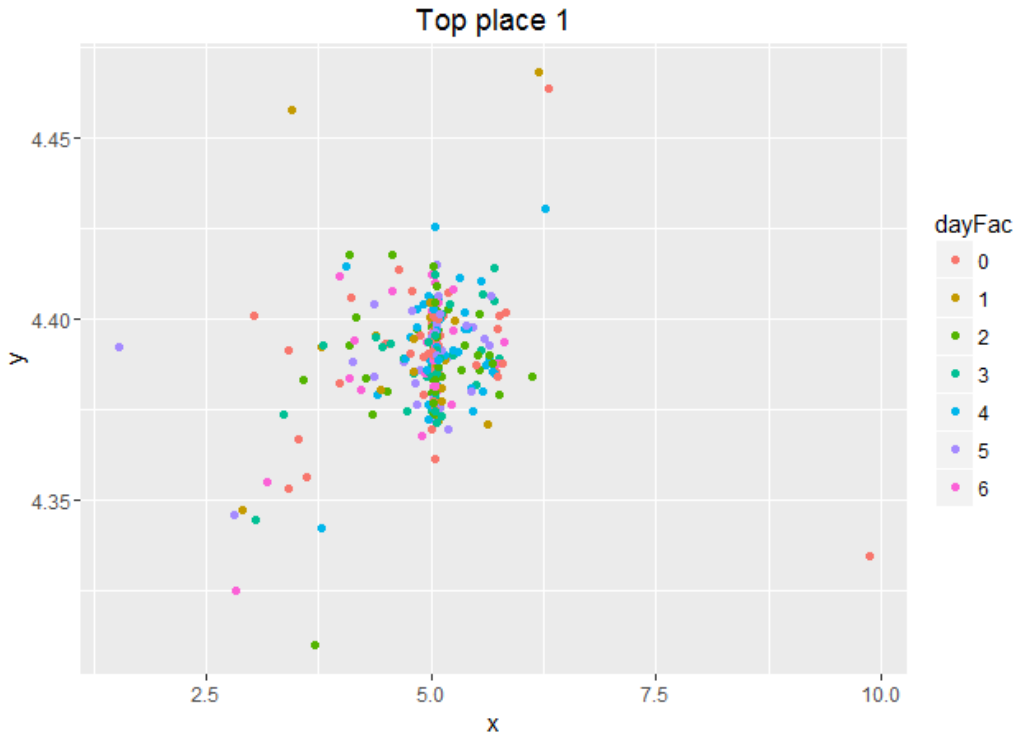
No super obvious patterns but it seems like there is a slight relation between location and time

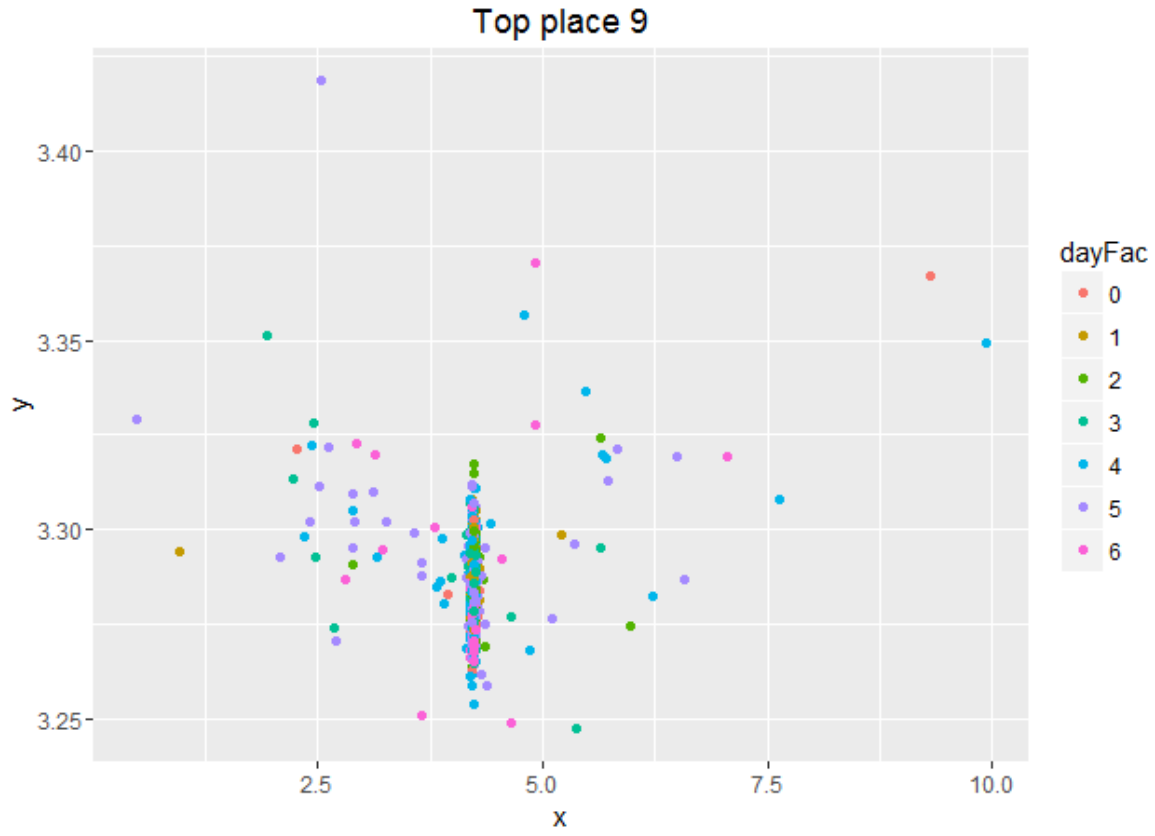

Top place 5
(-0.077,19.2]
(19.2,38.5]
(38.5,57.8]
(57.8,77.1]
6.12
6.10
6.08
y
6
7
8
9
10
x
weekCat
(-0.077,19.2]
(19.2,38.5]
(38.5,57.8]
(57.8,77.1]


Top place 9
(-0.077,19.2]
(19.2,38.5]
(38.5,57.8]
(57.8,77.1]
5.12
5.08
5.04
5.00
y
1
2
3
4
x
weekCat
(-0.077,19.2]
(19.2,38.5]
(38.5,57.8]
(57.8,77.1]

Analysis V: Day of the week versus location

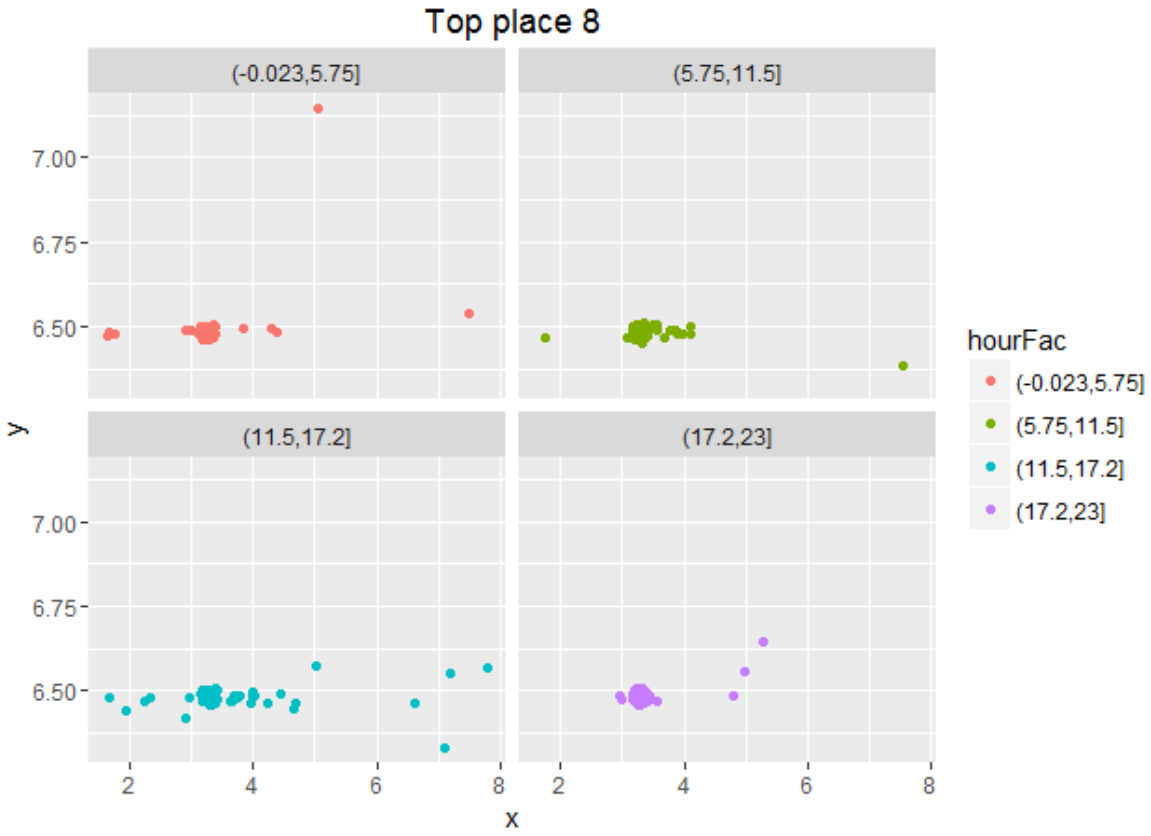
No obvious patterns

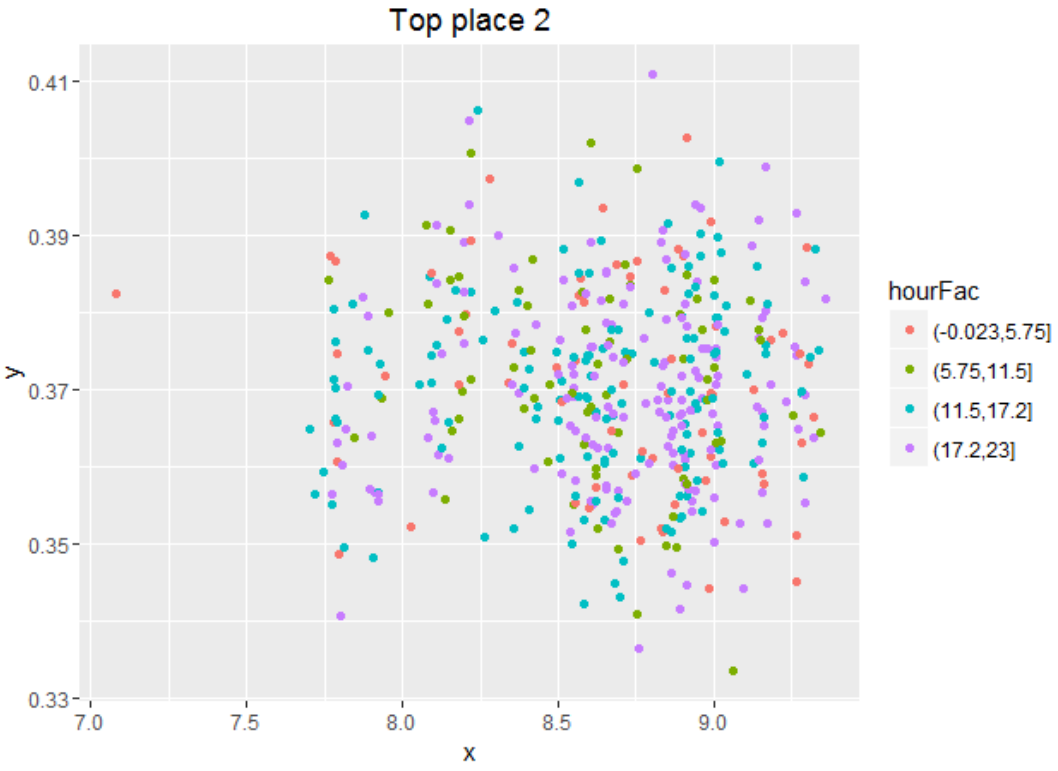




Analysis VI: Hour of the day versus location

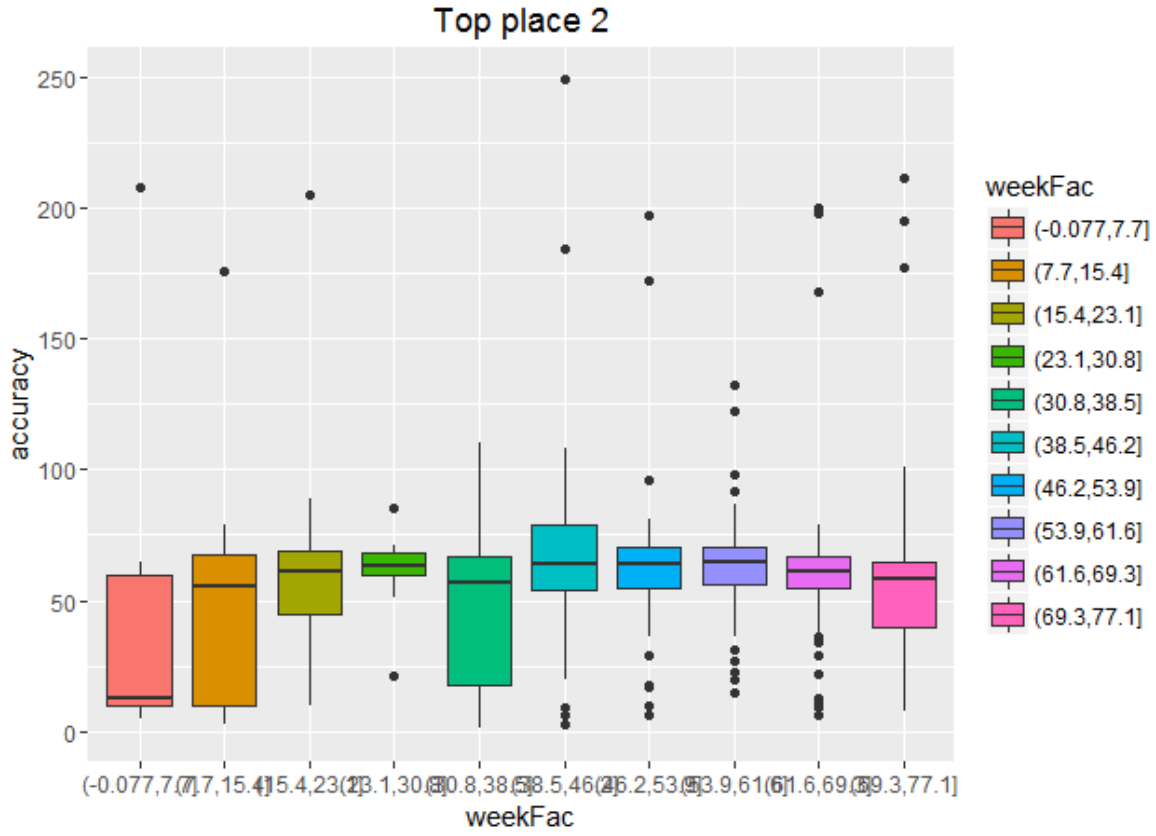
I tried hard to tell myself there is a pattern but in the end it's not obvious ☹. It looked like there was a pattern since some hour ranges are more populated.





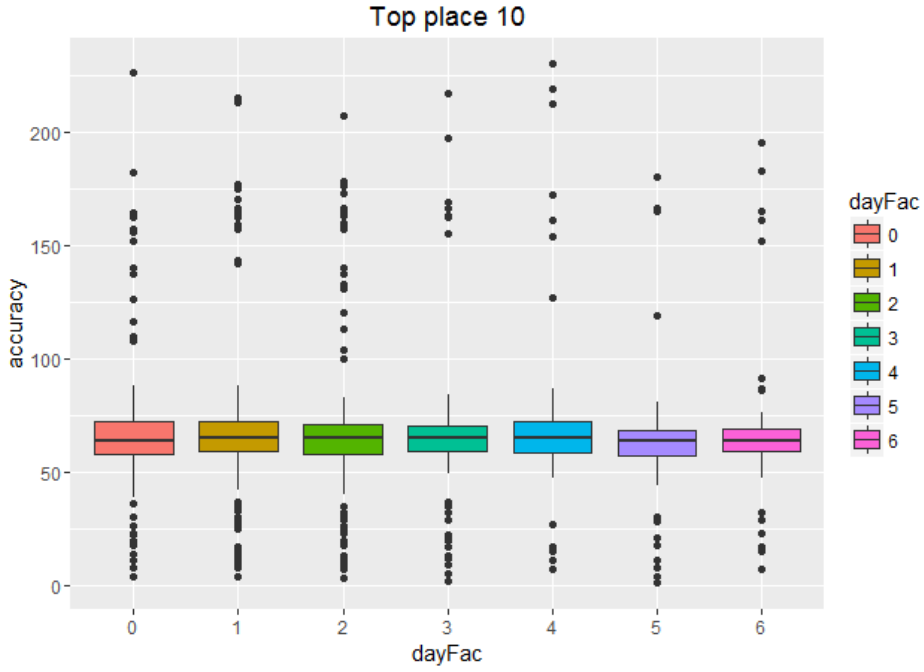
Analysis VII: Week versus accuracy – sliding density

Looking at the quantiles would be appropriate here (TODO)



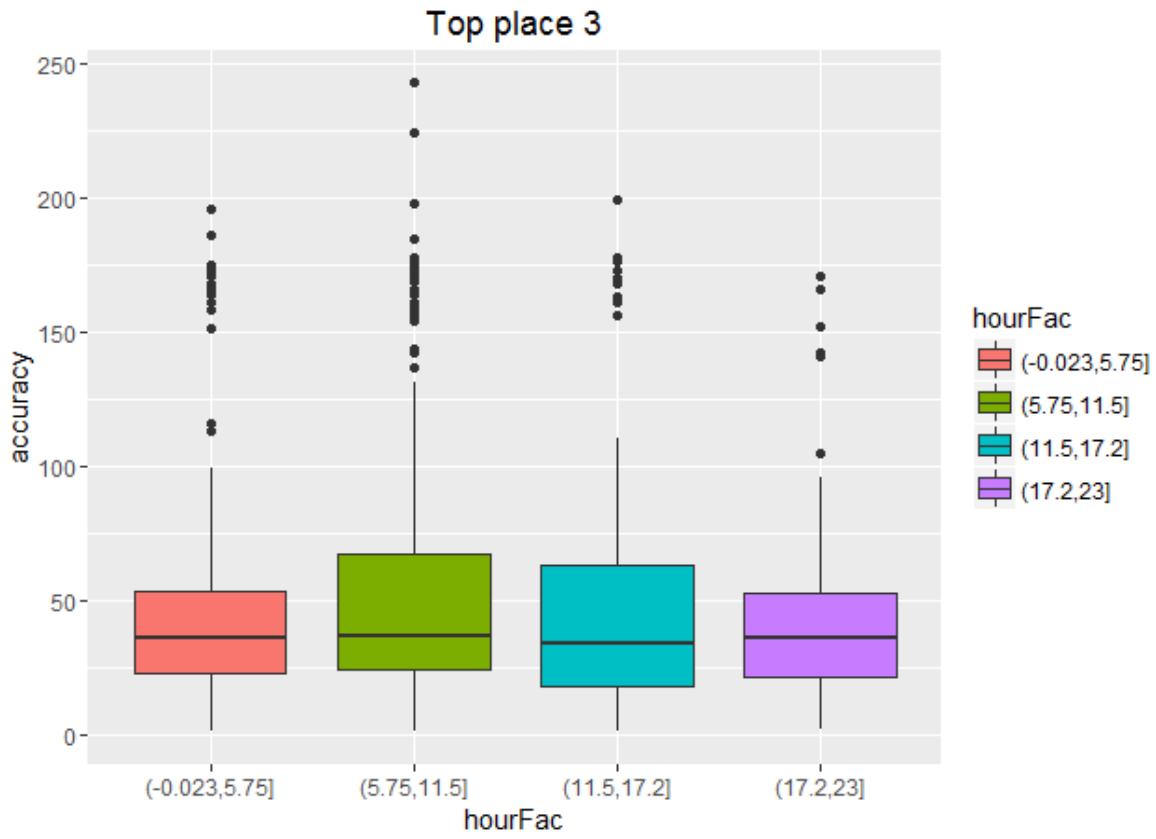
Analysis VIII: Day of the week versus accuracy

Hard to find a clear relation

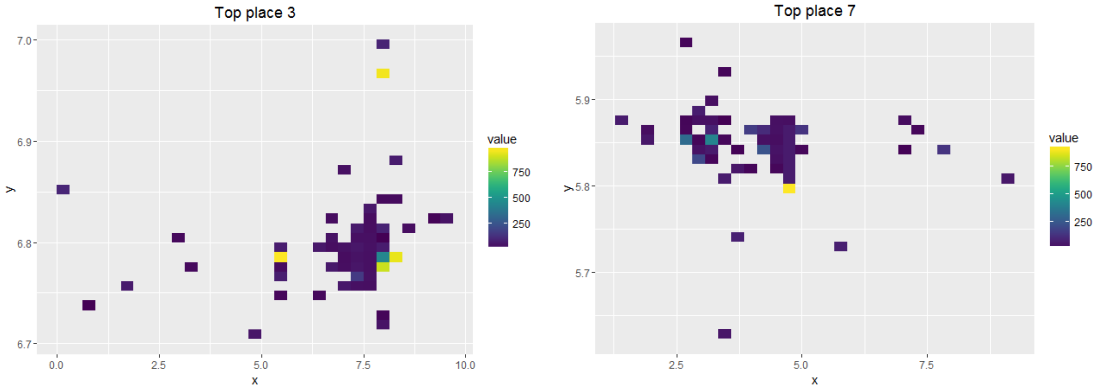


Analysis IX: Hour of the day versus accuracy

No pattern either



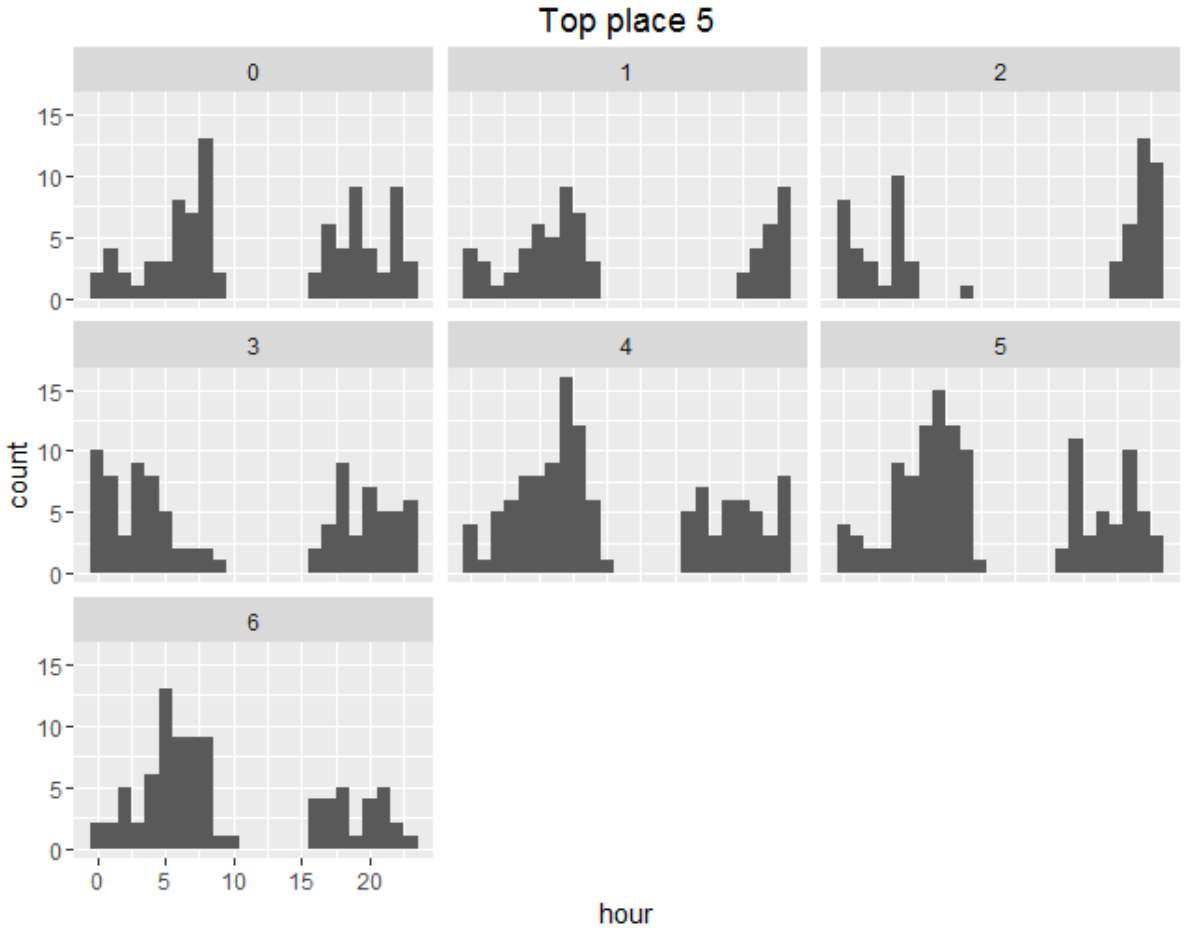
Analysis X: Location versus accuracy



No clear pattern between accuracy and location

Interesting:

Hour seems to interact with day of week – 7*24 blocks! Smooth and relax!


Top place 5
0
1
2
3
4
5
6
count
hour
0
5
10
15
20