Analysing the Citi Bike NYC data

Shaun McGirr

February 8, 2016

1 Context of the data

Citi Bike, New York City's bike-sharing system, was announced in September 2011 and launched in May 2013. Initially, thousands of bikes were stationed in Manhattan and Brooklyn, and in July 2015 a major expansion began: from 6,000 to 12,000 bikes and extended station coverage.[1]

As with all bike-sharing programmes, both the physical and electronic systems experienced problems. Physical maintenance of thousands of bikes is a complex challenge when they are moving all the time.[2] And the existing software stack has some key deficiencies, for example a station shows as having bikes available even when all are 'locked' for repairs.[3]

From the system's evolution so far, there are two main classes of question:

1. Is the system, as currently laid out, fit for purpose?
2. What can be done to improve it?

2 Potentially interesting questions

Examining the background of Citi Bike NYC, and what others have already done with the data, I developed several questions. Much of the existing analysis online confirms what is already known from other information, for example that bikes tend to be ridden in to Manhattan in mornings but back out in the evening.[4]

This leaves good scope for other questions, and my initial list was:

[1] See `https://www.citibikenyc.com/about`

[2] See background information at `http://www.crainsnewyork.com/article/20150426/TRANSPORTATION/150429891`

[3] See user complaints at `http://cbsloc.al/1nRp2Bh`

[4] `http://bit.ly/1JKvIv6`

- What is the relationship between bike-share usage and taxi usage? What about with subway usage? At what distance of trip is there maximum substitution across modes? (Might be observable from new station openings, and the answer could inform placement of stations in 'hot-spot' areas of excessive demand for shorter taxi trips, or chronic subway crowding.)
- Has increased bike share usage led to increased acccidents (car-bike and bike-bike) or decreased average traffic speeds?
- Does proximity to a bike share station increase the price sellers of property ask for? Is there an actual increase in sale prices?
- Can we synthesize a 'synthetic unit record file' of users from the combination of rider age and birth year, predominant journeys from/to a city area, and demographic data for these areas?
- How well is the system managed? Where are the shortages of available bikes and how has network expansion affected this problem?

Most of these are too ambitious for the limited time I had available. I decided to focus on what seems to be the greatest cost for the operator besides maintaining heavy-duty, hard-to-ride bikes: the need to relocate bikes from 'surplus' to 'deficit' stations. The former tend to (net) gain bikes over a given period, while the latter (net) lose.

Making visible where these are is the first step towards a business action, that could be as simple as scheduling more bike transfers, or as complex as offering discounts on particular routes to help rebalance the system using pedal-power.

3 Obtaining the data and making them ready

As is often the case, this took considerable time. I decided early on to program a reusable pipeline, rather than just 'grab data' individually from the links at `https://www.citibikenyc.com/system-data`.

My pipeline follows these steps (see *code_grooming* folder for scripts):

1. Scrape the web-page for the underlying s3 links, filter these
2. Download and unzip the CSVs to *data_raw/systemdata*
3. Put files in a list and read them in without blowing memory
4. Parse the columns so they are most useful for analysis (date was most painful)

Then I subset the full dataset to a more reliable set (see *code_analysis*), making these decisions:

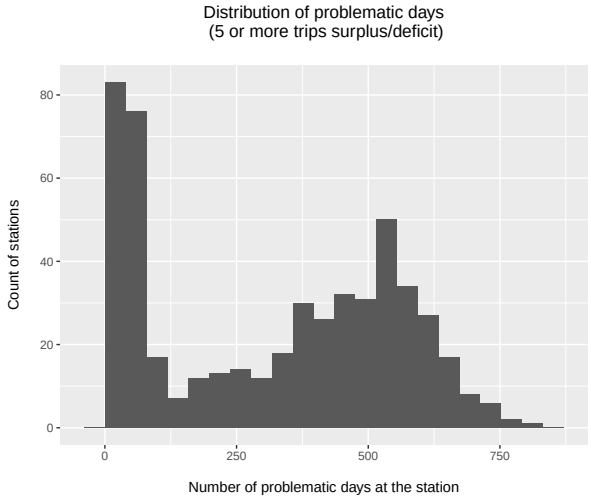
- Discard trips over two hours (following recent business rule change)
- Discard Jan 2016 data (incomplete)
- Retain trips starting/ending same station (similar distribution)
- Add several different date formats for different questions

4 Visualising Citibike's main business problem

As discussed above, the key cost to the business besides bike maintenance is moving these quite heavy bikes around New York City when inventory becomes unbalanced across stations.

To get a sense of the scale of this problem, I calculated how many trips each station is in 'surplus' or 'deficit' at the end of each day[5]. I then set a threshold of 5 trips surplus/deficit as 'problematic' and calculated the days each station exceeded this threshold.

Here is the distribution of 'problematic days': most stations experience this on relatively few days, while a core group require closer examination. There is a seasonal pattern in the underlying data, but it is weakening, as the Dec/Jan/Feb drop-off in trips declines.



The analysis so far merely confirms what anybody would expect, that most stations experience unbalanced inventory at some point. Furthermore, a significant group are in this status for 40-60% of days (the minor peak is centered around 500 'problematic days', for 900 days data Jul 2013-Dec 2015). This gives me confidence, at least, that this is the core business problem for Citi Bike NYC.

Analytics should always clarify the next best action of business decision-makers. The phrase has fallen out of fashion, but this implies analytics must generate 'decision support tools' to be useful.

In this case, such a tool would ideally have the following properties:

- Make the business problem and its impact clear

[5] I discarded the 1% of trips that cross midnight. I calculated at trip level because it simplifies the problem compared to the bike or trip-bike level, and only 2% of trips start and end at the same station.

- Narrow the focus (drastically, if necessary) to the worst cases
- Imply a remedial action, the effects of which are measureable

To meet these criteria, my tool will focus on the stations where the imbalance (number of days in surplus minus number of days in deficit) is highest historically, and where it is growing fastest. This narrows the focus and makes clear an emerging problem, rather than one likely already apparent from operational reporting. It will use a map to help decision-makers plan a remediation strategy.

Question: Over the life of the system, which stations have been the most imbalanced?

The map on the following page sets the scene for the decision-maker. It shows the 100 most imbalanced stations, colouring the worse business outcome (too many days with increased risk of no bikes) red, and the other poor outcome (too many days with increased risk of no parks) blue. More purple stations have less of an imbalance problem.

As shown by others' analysis, there is a strong tendency for stations in mid-town Manhattan to have many departures than arrivals across a day: Broadway & W 55 St, for example, has 560 days where at least 5 more trips were started than ended, and only 57 days the other way around. This and other stations coloured red are therefore at high risk of holding insufficient bikes to meet demand.

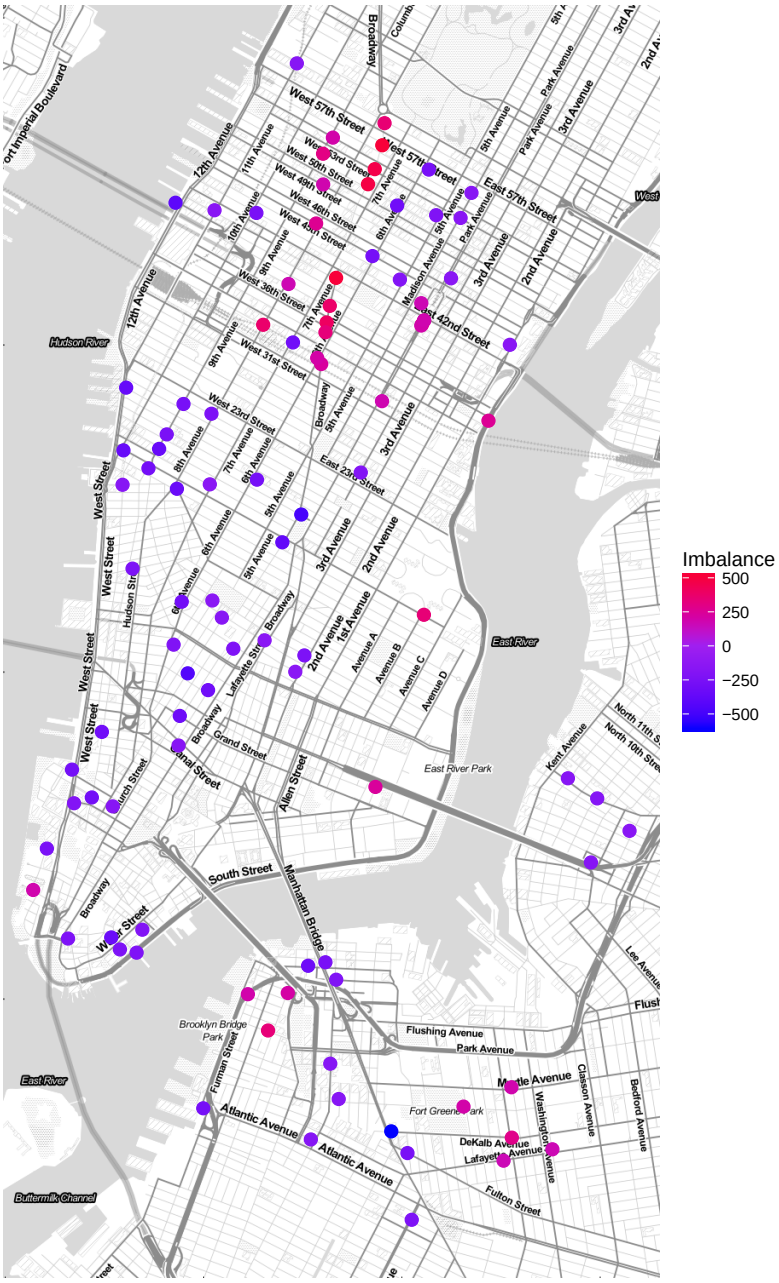
Based on this tool the decision-maker can immediately begin to, quite literally, join the dots. In Brooklyn there is one dark-blue station, meaning on most days, many more trips end there than start, creating a risk of too-few parks. Nearby are several stations with the opposite problem.

Bikes are likely already moved between these Brooklyn stations outside peak times, but given the bike-friendly distances involved I would recommend considering differential pricing: incentivise users taking short, within-Brooklyn trips in the 'helpful' direction to use a bike instead of other modes. This has the potential to rebalance inventory at lower cost and with greater flexibility across the day.

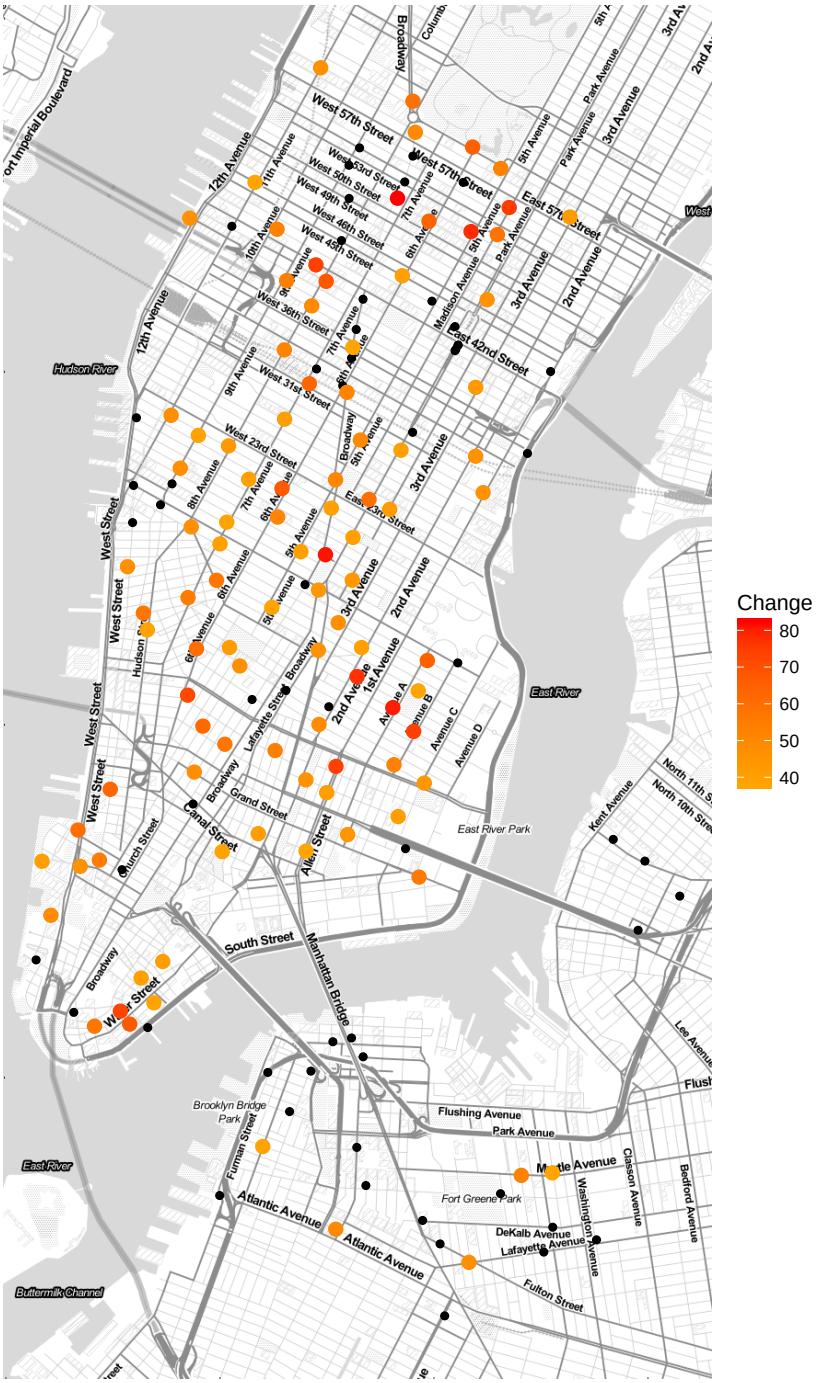
Next steps for this tool could include:

- Replicating analysis for the stations where this imbalance is growing fastest (see next graph)
- Separating weekday from weekend/holiday analysis
- Changing the grain from per-day to per-hour or per-period to drill inside daily patterns (this would need to be accompanied by further simplification)
- Animating the balance between trip starts/ends across the day

100 most imbalanced stations Jul 2013−Dec 2015
(days with surplus starts − days with surplues ends)



100 stations with greatest month−on−month change in imbalance, 2015
(imbalance = days with surplus starts − days with surplues ends)



I repeated a similar analysis but calculated which stations experienced the greatest month-on-month change in the number of 'imbalanced days' (at least five more trip starts than ends) across 2015. These present a different challenge to business decision-makers: on the one-hand they indicate areas of strong growth in usage, but also areas of emerging inventory problems.

The picture is quite different. While some of the most concerning stations from the previous graph appear again, some of the fastest-growing in terms of imbalanced days[6] do not. They highlight areas, such as Alpabet City (Avenues A/B/C/D) where pre-emptive operational changes might be wise.

5 Final thoughts

This is a very interesting dataset, and once wrangled, lends itself to further exploration.

Given more time, I would want to model the demand for bikes, which could be used to give business decision-makers a 'heads up' when their current allocation of inventory to a given station is unlikely to be sufficient over the next 3-24 hours. This would allow them to proactively move bikes, or choose to alter pricing strategy and have others move bikes for them.

Further value could be added by joining these data to other open data, such as the similar dataset on taxi rides in NYC, to gain an understanding of subsititution effects between modes.

[6] Those stations at which, across 2015, the number of days ending with more trip starts than ends (or vice-versa), grew by 70+.