# Distribution of Bicycles Optimisation

Using Data Science Analysis to help Optimise Bicycle Share Schemes' Distribution Plans


An IFN 702 Artefact Project by **Joel Schwaber (10241337)**

**Supervisors:**

Dr. Dimitri Perrin (Dimitri.perrin@qut.edu.au)

School of Electrical Engineering and Computer Science,
Queensland University of Technology

## Table of Contents

## *Introduction*

Bicycle share companies provide an important, robust, ecologically-green system of transit for urban areas. They add a valuable solution to the 'last mile' puzzle without relying on an automobile, as well as flexibility to urban multimodal transit systems. A major problem for a bicycle share company such as Citibike is availability parking for its rented bicycles, and availability of bicycles for customers to hire and to take from the station.

Station-based bicycle share schemes rely heavily on bicycles being "redistributed", or "rebalanced", to maintain this 'balance' of bicycles available for rental and spaces for a customer to park their rented bicycle. The aim and objective of this Development Project is primarily to produce data that is both immediately understandable to even a quick glance by Citibike bicycle rebalancing teams located in Jersey City, in a way that provides value to both the business and their rebalancing teams.

The data from the project will be used to design a tool that will assist in ensuring that a bicycle share scheme's bicycle rebalancing team is delivering value. Specifically, it will advise distribution teams in choosing which stations to elect to empty or fill with bicycles by using data that has been made publicly available courtesy of the Jersey City Citibike System. Over time, certain stations trend towards having bicycles taken from them, while others trend toward being filled. This tool will track which stations trend in either direction, so that (for example), they are made aware not to remove bikes from a 75% full station that typically empties out. If it empties out after they removed bikes from it to 50% full, then their efforts in rebalancing team could be considered detrimental instead of helpful to the Bicycle Share Company's attempts to keep bicycles available to the public. By consulting the data that this tool will provide, such a scenario could be avoided.

## *Project Management Approach*

The project is based centrally around Agile Scrum, while borrowing elements of elements of Prince2/DSDM for guidelines and the creation of an 'intended final use,' with intended features. Scrum's Iterative Development Cycle will be utilized to continually deliver a "done," product at the end of every sprint. This will be marked by the product being a stable, usable, item of value. A Prince2 Gantt chart will be used (*fig.1*) will be used to plan out an intended trajectory for the product, along with a rough estimation of a timetable and feature progression, though sprint durations and cancellations will conform only to a Scrum's sprint timebox and burndown chart. These will be determined during the Sprint Planning stage, when scrum team velocity can be estimated more accurately based off of past performance. Scrum's fixed timetable for sprints allows for a better estimation and prevents overrun or over-commitment on features that prove either unfeasible or less valuable than initially believed, which ties in to Scrum being able to cancel a sprint if a feature is decided to be dropped. It also ensures less time is spent in the planning stages and that work can commence more quickly. Given that the semester started late, this was another important reason for choosing Scrum as a methodology.

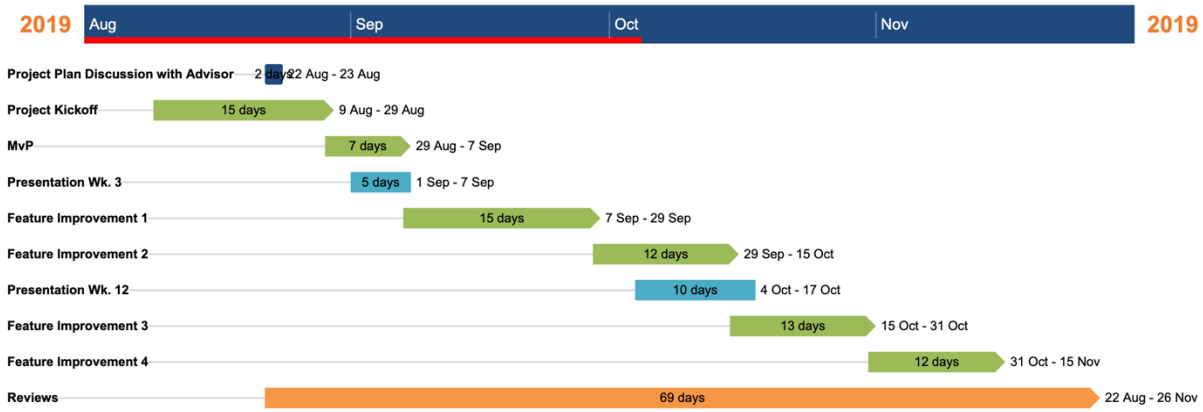


*fig.1* Prince2 Gantt Chart with soft date estimations

In Scrum, Features are flexible, ensuring that if a feature proves difficult to implement or unlikely to satisfy a consumer need, then it can be scrubbed and other elements can be developed in their stead without "breaking" much. Team velocity can be measured more accurately by Scrum after the Scrum Team has been in practice for a period of time and resources have been gathered, too. The added flexibility of Scrum is the driving reason for it being my methodology of choice. Timeline slippage is a possibility, but it will accommodate for that possibility by adjusting the features.

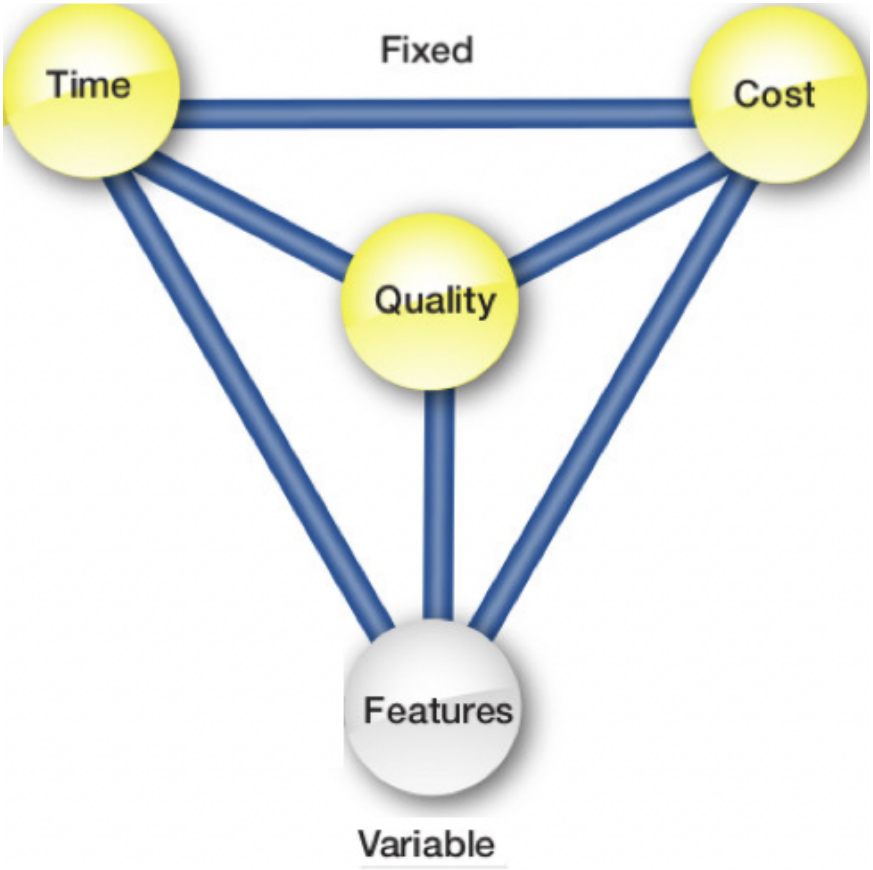


Fig.2 Scrum Variable-vs.-Fixed Chart

In the chart above, the 'Cost' is 'tuition,' which has been paid already. *'Cost'* is then where it ought be- toward the upper end of the chart, "Fixed." *'Time'* is fixed in the

sense that the sprint length is non-adjustable once commenced, and can be no longer than a month. There is a final due date for the delivery of the project as well, though extensions may be granted by QUT to delay delivery. '*Features,*' is highly variable, as continual feedback is between those working in the bicycle share industry will help determine the selection and estimate the value of having those features. This aspect of running a Scrum methodology became highly relevant when features were adjusted after Feature 2 was developed, and the iterative development cycle meant a total change in the way the data was collated and displayed in Python. There were intended to be four features across sprints, however it is at the discretion of the Product Owner to take the project to market and assess the value of the product.

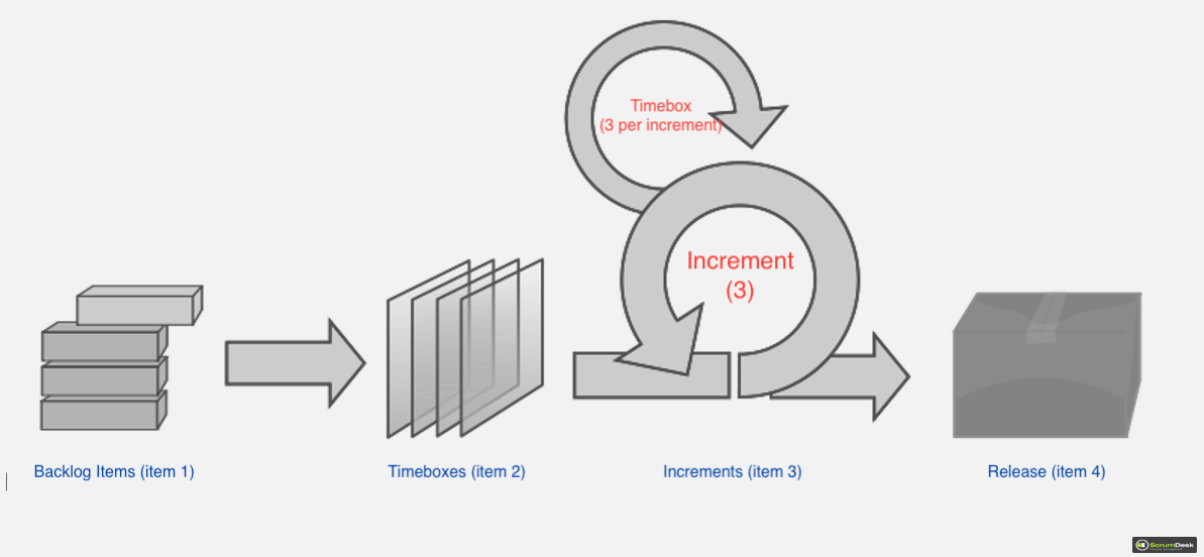


*Fig3. Sprint Timeline Chart*

Each Backlog Item (item 1) is a step that must be performed (e.g., "cleaning the data.") These then are split into the Timeboxes (item 2). At least three Timeboxes will form an increment. Each Increment forms a "Feature," with the feature list as described above, but collated below for easier and quicker understanding, to then form the "Release Item." (item 4).

## *Project Methodology*

This project relies heavily on Python 3 and RStudio for programming, mathematics, and visualisation. Interviews were gathered from individuals employed in Bicycle Share Companies in both management and operations, to better understand challenges that rebalancing teams face, and methods employed for dealing with those challenges.

Data cleaning was performed, with "rebalancing team" efforts wiped from the data prior to it being released to the public. Certain data points regarding demography of the rider and 'customer type' were excluded (e.g., gender of customer, length of ride, and customer subscription type) by data cleansing and removing columns in RStudio. Those elements fell outside the scope of the project, and had no potential to be useful in the project.

Each station has a unique number attached to it as a string integer. Each instance of a bicycle being removed and ridden from a station with that integer will be given a value count

of "-1" for that station's "count" of bicycle balance. A bicycle being docked at that station will count as a "+1" for that station's "count." Then a third function will read the tally and produce a mathematical addition (1 + -1 = 0, so it would be a "self-balancing" station. A station that has +1 +1 -1 = 1, and is a station that fills an average of 1 bicycle over the course of a month.) The conditions for the "MVP" sprint to be qualified as "done," are met at that point.

```
In [130]: f = open("test.txt","w+")
          for key in stores.keys():
              if key in stores2.keys():
                  f.write("{} : in {} out {} : dif {}\n".format(key,stores[key],stores2[key],stores2[key]+stores[key]))
          f.close()
```

*Fig 4. Raw Python addition of columns using keys to gain a raw 'count' of columns from stripped-down data files. This was deemed too inflexible moving forward. Python's Pandas and Numpy provided greater flexibility.*

The next step was changed at this point during the sprint planning and a meeting with a rebalancing technician. The GPS coordinates would be related to the station name to provide a map that held meaning from navigator to driver, and the value would be graphically visible on the map for "balance." This would require the use of ggmap, in either RStudio or Python.

After this point, stations would be assessed "in real time," and historical data values with the online API, made available from Citibike's StreamData: https://streamdata.io/developers/api-gallery/new-york-citibike-api/ updated every 15 seconds. The data would then again math out the figure - a station that typically fills and is full, would be the "top priority" for removals, and a station which typically empties and has no bikes in it would be the "top priority" for deliveries of the bicycles that were picked up from the full station, delivering a "full use" product.

## Task Breakdown and Weekly Plan

| | |
|---|---|
| MVP | Combine datasets, create a counting function. |
| Feature 1 | Create a mathematical function to determine outlier stations and create ways to order the data into a user-readable format. |
| Feature 2 | Add a graphical function with a map overlay using GPS Coordinates |
| Feature 3 | Clean graphical function, add live data updates. Use the expected trend data to prioritise which stations need servicing most urgently. |
| Feature 4 | Add a predictive model to the present number of bicycles that will determine future "problem spots," using the API. Further graphical improvements. |

*Table 1. Task Breakdown*

### MVP

A Minimal Viable Product (MVP) would be to establish a program that can determine a station's *usual* flow- e.g., how many net bicycles usually arrive or depart from a station over the course of a day.

To accomplish the MVP, files were merged to account for the length of time of the data. The naming conventions of the files changed over time, so the file names need to be changed to match a universal case and more consistent calling pattern for any programmed attempt. The re-named CSV files were merged using RStudio's "rbind" command, and after

checking for consistency, all unnecessary columns (such as the rider's gender and type of customer) were dropped, as that data fell outside the scope of the project. RStudio's "write" command was used to output a Comma Separated Value ("csv") file of varying lengths of time, ranging from a month to a few months, to an 'All time' data, which was from the program's inception to its latest file update, a period spanning over a year. This would then be used to view data over lengthy periods to account for seasonal patterns (e.g., school holidays in mid-winter or summer break). A basic python script was run that established a count for every instance of a docking or rental, and then the rentals were subtracted from bicycles docked as a proof of concept and to finish the MVP. This sprint met its objectives and was "done."

The development of the MVP was originally done via "raw python," in a series of simple mathematical steps, and the data combined in rBind to produce a raw "count." However, this was deemed unsatisfactory in the sprint review.

## Feature 1

The Feature 1 upgrade from the MVP was to use packages such as numpy and pandas rather than simple python. Pandas read the chosen csv files, and parsed dates rather than inferring the dates as a matter of efficiency. Pandas was used again to determine the start and end point locations of a bicycle rental. The in and out counts were created, using station IDs. Every time a station ID was mentioned, +1 was added to that station's count (as identified by 'station_id' column, which is treated as a string/name). The bikes taken away was then subtracted from the bikes delivered, creating a 'balance' figure.

Holidays, weekends, and other days proved difficult to completely comb out of the data at the data cleaning stage, and the value of doing so in the analysis of the results from the Feature 1 upgrade revealed that there was limited value in doing so. Analysing separate days or separate months did not alter which Citibike stations were the most egregious outliers in balance, nor did it alter the direction in which those outliers lay, and so the feature was dropped beyond a measurable outcome. This is in-keeping with Scrum's feature flexibility, while useful to know, there was little point in presenting data that functionally had the same outcome.

## Feature 2

The Feature 2 upgrade implemented a map overlay using GPS coordinates that the JC Citibike system provided as part of its publicly available data for its station locations and ggmap in Python. It also improved on the Feature 1 with graphical improvements and changing the colour palette to better visualise stations which approached a more colour-neutral value. The first "deep dive" into data takes place in Feature 2, providing insights into future feature development based on feedback on which stations are outliers.

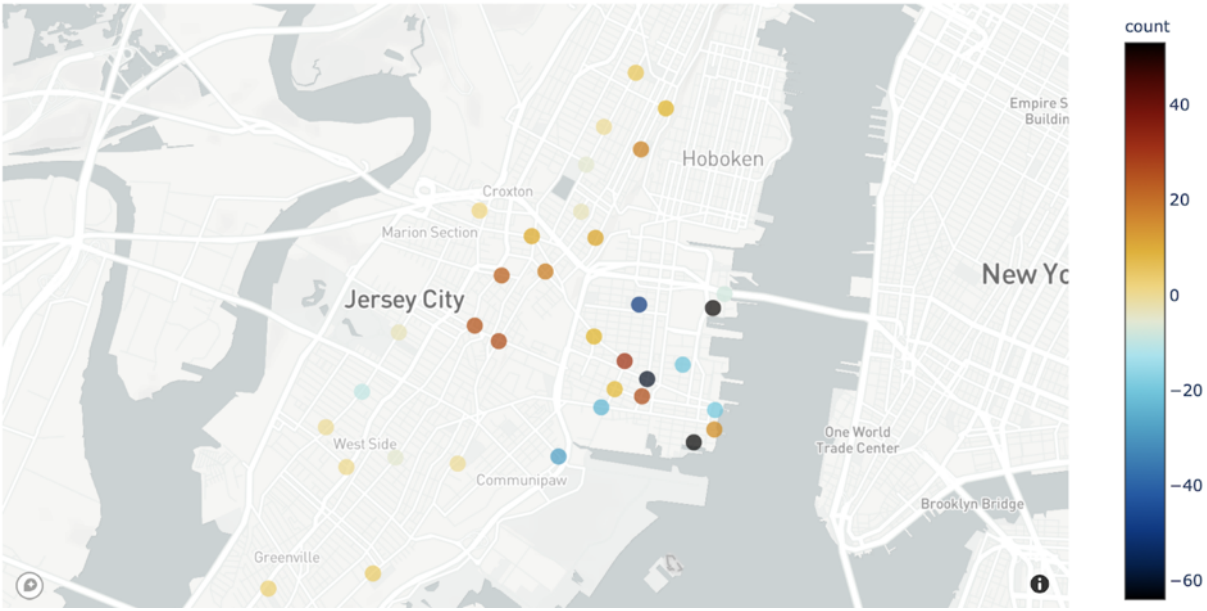


*Fig. 5. Feature 2. A single month's data, using default Px (Plotly Express) colour scheme IceFire*

## Feature 3

Feature 3 was intended to deliver a more refined map appearance after issues cropped up in Feature 2 and a sufficient graphical improvement in Plotly Express was unable to be located. Creating a custom colour palette was possible but required a different plugin than Plotly Express. This was attempted, but the feature was dropped after the difficulty score was revealed to massively be mis-ranked. The sprint was cancelled and the item was removed from the backlog.

At this point, I was also informed that the stations had changed location, and that the end of publicly available data reflected the point in time in which they had relocated stations with the lowest total patronage at the southern end. Therefore, there would be no use in implementing current data or using the Citibike API, and Feature 3/4 were scrapped, with simply a change in colour scheme and a change in shifting to data analysis to measure total patronage, elevation, and to attempt to find trends in the data.

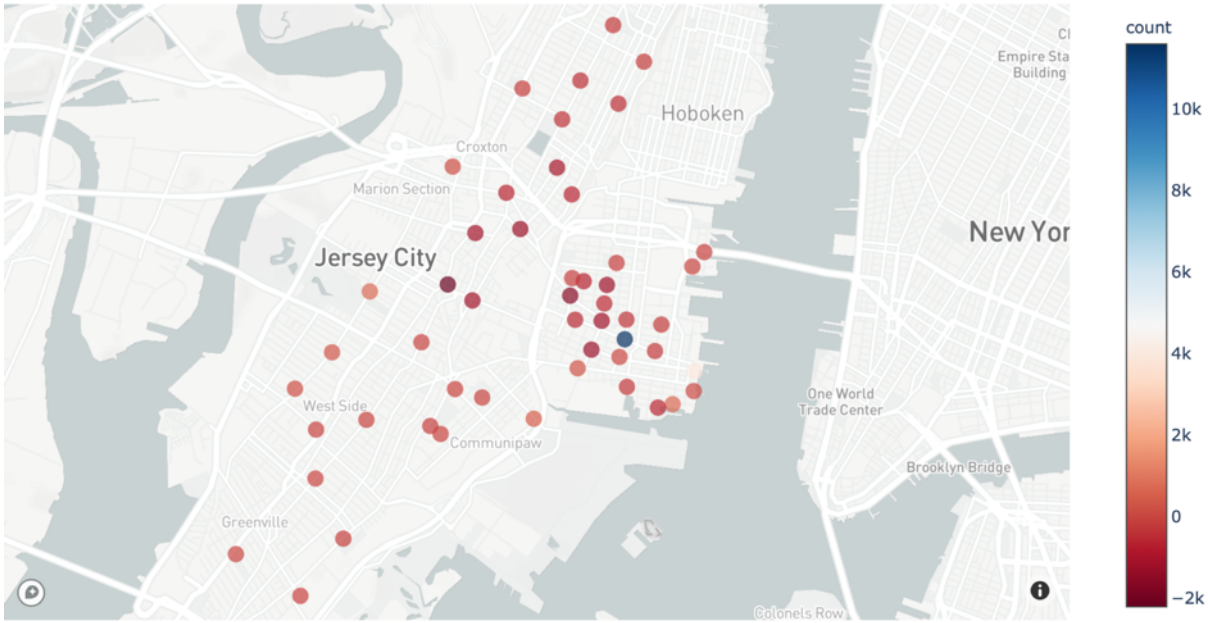


*Fig. 6 RdBu colour feature applied to total data.*

### Feature 4

The Feature 3 upgrade was intended to be an "if" statement for any station that scored less than a user-specified number of spaces available or spaces remaining 'projected within the next 24 hours,' to *simulate* the station's expected loss or gain of bicycles, after reading the Citibike API to measure its present number of bicycles. If it was going to empty entirely or to become completely full, then it would make itself a "priority" target for the distribution team. This was intended to be done by taking the data for imbalance and dividing by number of days measured to determine "bikes per day," and rounding to the next integer value (e.g., -0.6 would round to -1 bike expected, and +2.5 would round to +3 bikes expected to arrive within the next 24 hours). The 'elevator pitch' for the feature was to create a weighting system. The weighting system would measure the current fill for all stations, and then prioritise based on how soon the station might fill or empty out if it kept with its usual trend of tending to fill or empty out. However, this was scrapped for the same reason as the initial measure of Feature 3, with the stations having relocated rendering the live data feed useless.

## *Outcomes/Results/Findings*

### Main Findings

Whilst the outcome of the project has been a less than fully-featured product, it is still marketable and does contain more features than the 'MVP' (Minimum Viable Product). The product provides meaningful historical data and enables evaluation of the bicycle share system's usage.

Determining patterns within the data proved difficult, though some outliers did emerged, they were not without caveats. The most heavily patronised stations were those near to transit hubs that led directly to New York City along a Port Authority Trans Hudson (PATH) station. However, in terms of 'balance,' there was variability. The largest "delivery" spot, by a considerable margin was Grove Street at +11,584 over the time span of the measured data. Exchange Place received +3,974. Both of these stations are adjacent to PATH

stations. However, the aforementioned caveat is that stations such as Newport station, (also adjacent to a PATH station) received high total ridership figures but maintained an exactly perfect balance of 0. That is to say that it received precisely as many bikes as were ridden away from it. Sip Ave, adjacent to a PATH station at Journal Square, had a score of -1,112 bicycles, and was situated atop an incline. The station adjacent to it, McGinley Square, was the largest departure station, with -2,215 bicycles ridden from it.

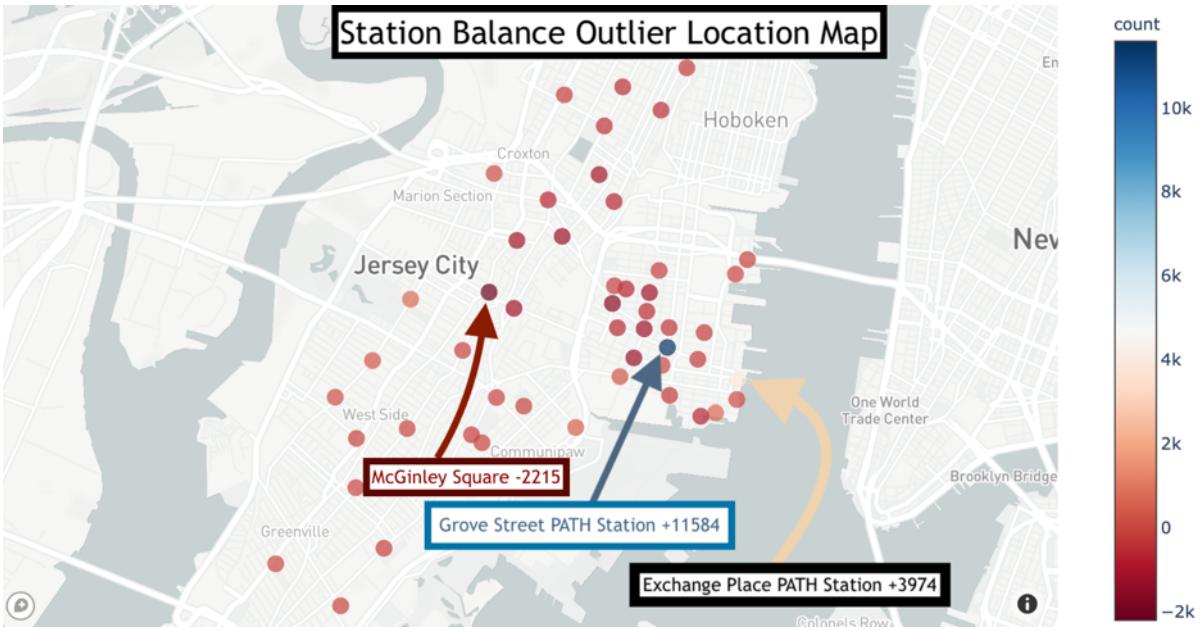


*Fig. 7. All balance outliers*

Given the limited size of the JC Citibike program, each had extenuating circumstances which will now be explored.

## Elevation as a Potential Factor for Station Imbalance

Citibikes weigh about 40 pounds (18 kilograms), whereas a normal bike (sans rack, fenders, etc.,) weighs a little over half that amount (Abad-Santos, 2013). The stations located in the suburb "The Heights," each scored a negative total value in bicycle balance. A conclusion might be reached that riders seemed to simply be more willing to ride downhill than uphill, which then reflects why McGinley Square, adjacent to Journal Square tended to have more outbound bicycles rather than receive.

Due to the small sample size in Citibike stations which can be analyzed, a similar docking-system based bicycle share company was researched with interviews for supportive data. This is anecdotally reinforced by information provided from Brisbane's CityCycle program, in which Station 127 (atop Annerly Road in the Dutton Park suburb), is a station which is considered to empty out the greatest amount of bicycles relative to the number of bikes it receives. This was according to interviews with 7-year veterans of the CityCycle Rebalancing teams (Brady, 2019; , Vary, 2018). Its relative elevation to stations 27 and 97, which are among the stations most frequently emptied out, is pictured in the Fig 5. below.


*Fig. 8 Brisbane's CityCycle Station 127 atop Dutton Park reportedly empties out, whereas Stations 27 and 97, at a lower relative elevation, routinely fill out. (Brady, 2019) ©OpenStreetMap Red marks higher elevation, blue marks lower elevation.*

This pattern was also noted to be present atop CityCycle's *Fortitude Valley Station 53*, which is perched on the 'upper' abutment of the Story Bridge at Bowen Terrace, and stations which are adjacent to it are at a relatively lower elevation. However, quantifiable trip data was not made publicly available, so this could not be confirmed mathematically and evaluated.

While CityCycle is part of a separate system, it does nevertheless lend some credence to the notion that the elevation of one station relative to surrounding stations may play a factor in its propensity to either fill or empty out. Further exploration would be needed across other bicycle share systems, or additional data to confirm this theory

Analysis can be performed in a few different ways, depending on the intention of use. The stated use of the project is to deliver value to the Citibike organization in optimizing its rebalancing teams. In this sense, a raw numerical 'positive' (more bicycles docked than checked out from station) and 'negative' (more bicycles checked out than bicycles docked) outliers can be identified. Those stations are located at Grove Street, Exchange Place for 'positive,' and 'McGinley Square' as the furthest 'negative' outlier. The advice provided to a new Bicycle Re-Distribution Crew would be then to primarily remove bicycles from Grove Street and Exchange Place stations and to expect those stations to fill over time, and to ensure that McGinley Square/Sip Avenue retains a healthy stock of bicycles, and to expect McGinley Square/Sip Avenue stations to run out of bicycles if it is not re-stocked periodically.

### Proximity to Transit Stations as an indicator of use and balance

The most extreme outliers in the Jersey City Citibike system are located at Grove Street, Exchange Place, and Newport PATH stations. The Port Authority Trans-Hudson railway line, or, "PATH" line, operates between New York City and New Jersey, connecting the two states across the Hudson River via subway.

Grove Street and Exchange Place are each near PATH stations and at low elevations. However, again disrupting any possible trend, the Newport PATH Citibike Station is situated at a similar elevation but was neutral in its bikes received vs. taken away values. The difference then between Newport PATH and Exchange Place/Grove Street is that the latter two are located toward the centre of the bicycle share system, rather than rather remote position, along its edge due with Hoboken, NJ (which has no Citibike stations). Proximity to

the centre of the system then seems to be a contributing factor as well, indicating heavily that the working hours of the employees ending before 5 o'clock rush hour plays a heavy part in redistributing the bicycles to a numerical balance by the end of shift- riders check bikes out from the residential areas, with a particularly pronounced effect especially if those stations are located up top a hill, and then through the working day, the distribution crews pick up the bicycles.

### Additional Perspective: Total Usage per Station

Balance vs. Patronage as a figure as was a simple matter to establish. A change from making "bikes removed from station-" subtracted from bikes delivered to it being an additive function established the total usage of the station. Not coincidentally, the stations with the lowest patronage and highest imbalance-to-usage ratio were the ones to be relocated in 2018. The outlier station, Grove Street, was an extreme outlier once again, with over 81,000 bicycles either arriving or departing, Exchange Place at over 50,000, and the station near Newport PATH 33,000 and received *exactly* as many bicycles as it has received over the lifespan of the Citibike JC program.

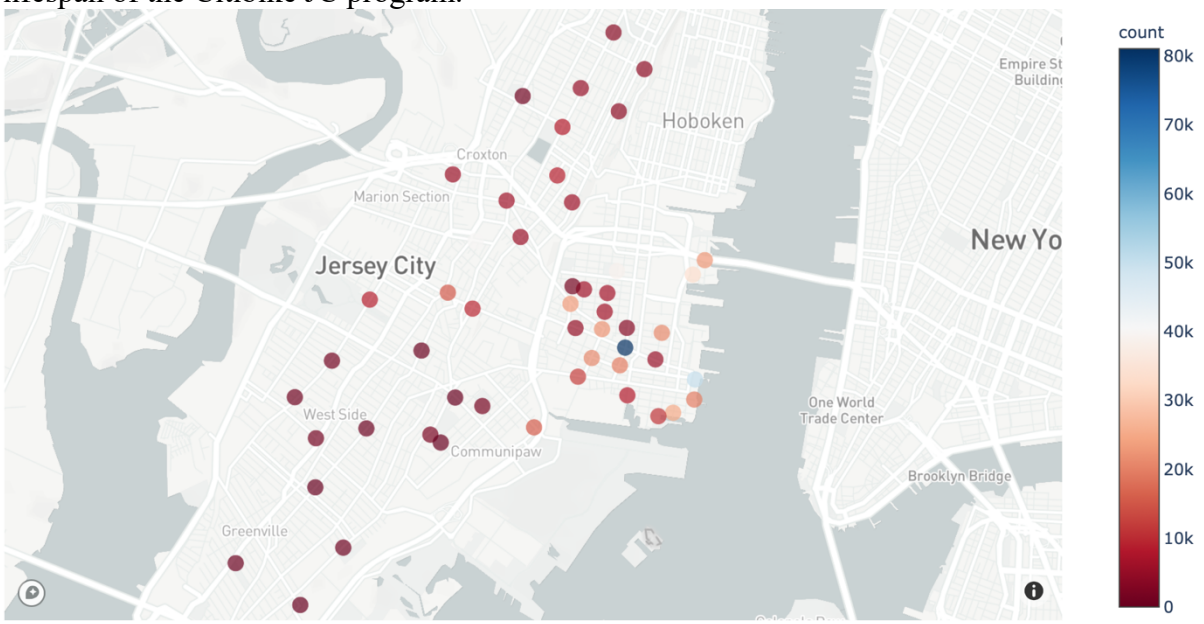


*Fig. 9 Raw Patronage, Visualized*

From a sheer 'patronage' to 'effort-required-to-rebalance' ratio, then, the Newport PATH station is a clear leader, with it requiring practically no rebalancing through its life-span from the launch of the company through all the data made available. What defines Newport Station from a Station Layout Perspective is that it possesses a high number of docking terminals, proximity to a PATH station, is on the 'edge' compared to other stations. Its relative low elevation seems to be partially offset by being of the same elevation in terms of proximity to stations that are most adjacent to it.

Citibike stations that require the most balancing should not then imply that those stations are in any way less valuable because of the additional work that is required to rebalance those stations. A series of stations toward the south were among the lowest outlier values in balance, with none exceeding +/- 101 bikes. However, a glance at the cumulative use of those stations also reveals that they had a far lower total trips taken count, with none of those stations south of Communipaw exceeding 753 bike trips, a low figure given that Grove Street Station had a trip total of 81,034, a figure that is over 107 times greater.

“Bike share is important and we're going to do what we have to do to make the system successful but it really relies on support from the community, you need users.”(Jersey City Mayor Fulop, 2017, when commenting on Jersey City’s decision to relocate those under-utilized southern stations in 2018, after the selected cut-off for data collected). This disadvantages communities which are over-reliant on the automobile (Saelens, 2003), but was necessary due to low overall patronage figures. While relocating the stations to a more central area may invite the need for further rebalancing, it seems that the rises in patronage is also considerable.

## *Reflection*

Delving into why these outliers are present delivers value to the organizational and management arm of the Citibike organization by offering valuable insights. Station layout optimization in the event of future station expansion or reorganization can also trickle down to lower in the organization hierarchy, giving rebalancing teams expectations on likely usage patterns on new stations.

The colour balance on the mapping tool was shifted by the extreme outliers in Feature 2, specifically Grove St., which was by far Citibike’s busiest station. Additionally, interviewing staff members, it was disclosed that ‘neutral’ stations that trended toward filling would still periodically empty out and need restocking, and vice versa, and that having bicycles available took priority over having spaces to dock. As a result, there was little value to the company for changing the colour scale to accurately reflect a perfectly neutral balance. Given that the aims of each sprint was to deliver maximum value, it was then left alone.

Initially, I had hoped that there would be an identifiable trend between Citibike station proximity to PATH Stations and either a need for rebalance or lack of need. However, this was disproven almost right away when the Newport PATH station had an exactly evenly balanced ratio of number of bicycles dropped off to bicycles rented from the station, whereas conversely, the largest outlier station was Grove Street. Data analysis was very inconclusive between the proximity of the Newport PATH Citibike station and why it was balanced. While other fringe stations existed, they did not balance themselves, and they did not have nearly the same patronage figures. In fact their patronage figures were so low that Citibike decided to relocate those stations in 2018, which created a cut-off date for publicly available data.

There were factors that influenced station balance. Having taken IFN509 it would have been possible for me to utilize a notation for how deviant from the mean or median each station was. It should also be noted that there was a differential across all measured Jersey City data of positive-129 bicycles across time. The cause of the discrepancy was not explained by staff when asked, and was told it was ‘various’ causes.

Personally, I think I learned from going into this with an expectation of the value, that sometimes the data might not be straightforward to read. For all the high number of trips taken, ultimately there’s a rather limited sample size in stations to choose and derive patterns from. Had I thought to include other bicycle shares and to then measure across systems, perhaps patterns would have emerged more consistently.

The bicycle share community at large could be served by this formula. If this model is expanded to include other bicycle share systems’ data, it could strengthen the theory of

higher elevation correlating with departures, and assist in planning future bicycle share systems.

Data:
https://www.dropbox.com/s/00xbzg6af0sqq5n/exportUltimateData.csv?dl=0

## *Citations*


Abad-Santos, *No, You're Not Too Fat to Ride New York City's New Bikes,* The Atlantic, May, 2013. https://www.theatlantic.com/national/archive/2013/05/there-are-people-too-fat-ride-new-york-citys-new-bikes/315684/.

Brady, W., et. al., Interview, Citycycle, 2018, 2019.

Fuller, D., Gauvin, L., Kestens, Y. *et al. The potential modal shift and health benefits of implementing a public bicycle share program in Montreal, Canada*. *Int J Behav Nutr Phys Act* **10,** 66, 2013, https://ijbnpa.biomedcentral.com/articles/10.1186/1479-5868-10-66

Higgs, L., *Hoboken and Jersey City have a bike sharing problem. Can companies cross city lines?* NJ.com Traffic, Jan. 2019, https://www.nj.com/traffic/2018/02/survey_asks_can_hoboken_jersey_city_share_each_oth.html

Liu, Z., Jia, X., Cheng, W., *Solving the Last Mile Problem: Ensure the Success of Public Bicycle System in Beijing,* **Procedia - Social and Behavioral Sciences**, Volume 43, Pages 73-78, 2012, https://doi.org/10.1016/j.sbspro.2012.04.079

NYGov, *PATH Ridership Report January 2018* https://www.panynj.gov/path/pdf/2018-PATH-Monthly-Ridership-Report.pdf

Saelens, B. E., Sallis, J. F., Frank, L. D., *Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures*, *Annals of Behavioral Medicine*, Volume 25, Issue 2, April 2003, Pages 80–91, https://doi.org/10.1207/S15324796ABM2502_03

Tedeschi, A., *Rebalancing Citi Bike* Institut für Geoinformatik Feb. 2016, https://run.unl.pt/bitstream/10362/17842/1/TGEO0145.pdf

Vary, J., et. al., Interview, Citycycle, 2018, 2019.