Matt Bewley
Public Transportation Planning and Management
Tweeting Bus Cancellations
December 14, 2020

Summary

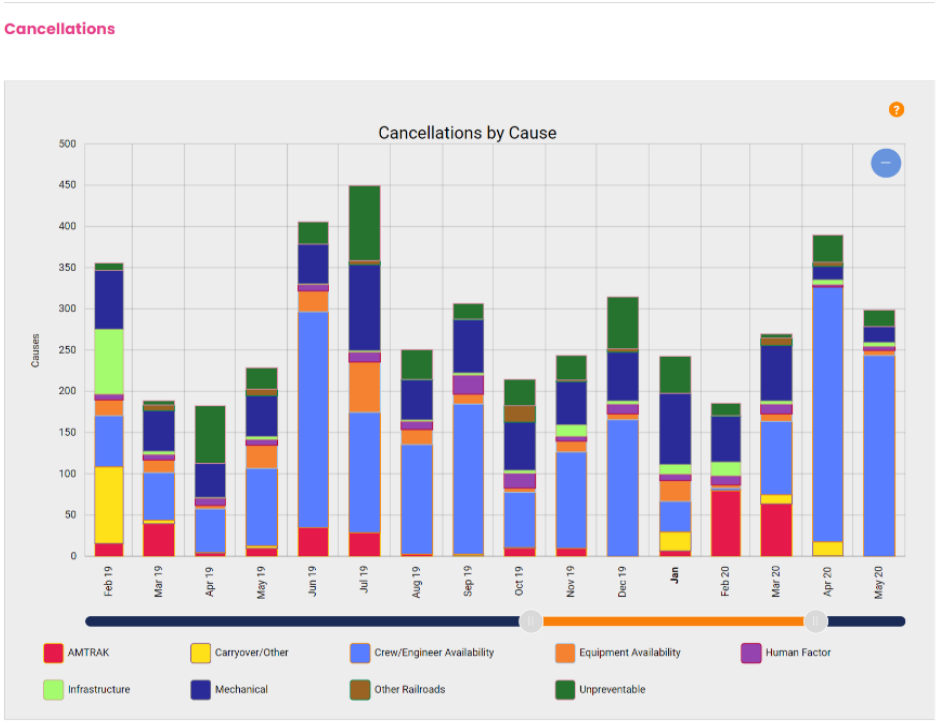
There is currently no public count of New Jersey Transit bus trip cancellations, but the agency announces bus trip cancellations on two Twitter accounts (@NJTRANSIT_NBUS and @NJTRANSIT_SBUS). This paper describes a process for collecting and analyzing bus trip cancellation announcements from those accounts. Among routes with cancellations, most have fewer than 100 over the 7-year period covered by the data. But a smaller number of routes have cancellations numbering in the hundreds, suggesting those routes might be good targets for efforts to reduce the occurrence of cancellations. In addition, cancellations show significant variation by cancellation reason, time of day, and route. The dataset also suggests variation in how NJ Transit has reported bus trip cancellations over the last several years—in particular, operator availability has been cited as a cancellation reason only since 2019. This and other observations suggest that NJ Transit should clearly define a set of cancellation categories, which would be necessary before a reliable cancellation dashboard could be released. Overall, the paper shows that NJ Transit's cancellation announcements can be used to answer pressing questions about bus service in New Jersey, but only tentatively.

Introduction

In 2019, New Jersey Governor Phil Murphy issued Executive Order No. 80, one of several orders aimed at improving New Jersey Transit's transparency and service provision.[1] The order required NJ Transit to publish information about specific performance measures, including on-time performance, mean distance between failures, and causes for delays and cancellations.[2] In response to the order, NJ Transit launched an online dashboard summarizing its performance on each of the required measures.[3] The rail dashboard, for example, allows users to view all of the measures, for the system as a whole and for individual rail lines.[4]

The bus dashboard, however, is more limited.[5] Users can find the same measures for on-time performance and mean distance between failures, but these cannot be viewed by line or division. In addition, there appears to be no measure for bus delays and cancellations by reason. Executive Order No. 80 explicitly required that rail cancellations be counted, and it even specified categories for classifying cancelled rail trips. The order had no such requirement for NJ Transit's bus service, so there remains a lack of any official count of cancelled bus trips in New Jersey. This paper tries to fill that gap by analyzing NJ Transit's tweets about bus trip cancellations.

Figure 1: Rail cancellations by cause, from NJ Transit's rail dashboard



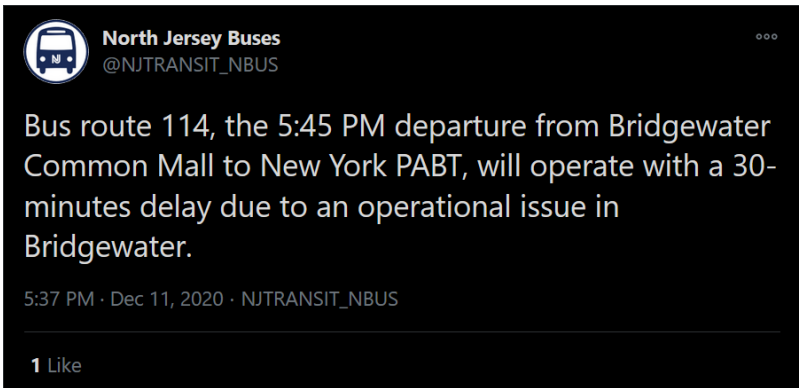
[1] New Jersey Transit. November 26, 2019. "NJ Transit Launches New Online Performance Dashboard." Accessed from: https://www.njtransit.com/press-releases/nj-transit-launches-new-online-performance-dashboard.
[2] Ibid.
[3] New Jersey Transit. 2019. "Progress by the Numbers." Accessed from: https://www.njtransit.com/improve/.
[4] New Jersey Transit. 2019. "NJ Transit Performance Dashboard – Rail." Accessed from: https://www.njtransit.com/improve/on-time-performance/rail.
[5] New Jersey Transit. 2019. "NJ Transit Performance Dashboard – Bus." Accessed from: https://www.njtransit.com/improve/on-time-performance/bus.

Methodology

NJ Transit operates the Twitter accounts @NJTRANSIT_NBUS and @NJTRANSIT_SBUS, which post service announcements. The announcements include cancellations, but they also include delays, detours, and other service changes. An example of a delay announcement is provided in Figure 2, below.

To allow analysis of NJ Transit's cancellation announcements, all of the tweets from both accounts were downloaded and processed. Twitter's API allows downloading large batches of tweets at once, but it does not allow downloading entire archives of users' tweets.[6] Instead, users are limited to batches no larger than 3200 tweets and no older than 7 days. To get around these limits, tweets can be scraped for their unique IDs, which can then be submitted to the API to get all of the tweets' metadata in return.[7]

Figure 2: Delay announcement tweet (retrieved from https://twitter.com/NJTRANSIT_NBUS/status/1337541968021352449)



The tweet in Figure 2, for example, has ID 1337541968021352449. When that ID is submitted to the Twitter API, the API will return metadata that includes the text of the tweet, the time and date of the tweet's creation, the number of likes it received, and the number of retweets it received, among other information. All of the analysis described below is based on tweet metadata returned by the API, for all of the tweets ever posted by the two NJ Transit accounts.

Data Collection and Processing

Altogether, the two accounts posted 98,129 tweets between their creation in 2013 and data collection on December 12, 2020. The North Jersey account (@NJTRANSIT_NBUS) posted 73,995 tweets in that time, while the South Jersey account (@NJTRANSIT_SBUS) posted 24,134. With the metadata from these tweets, each announcement can be sorted into two categories: a cancellation announcement, or some other announcement.

Specifically, the text of cancellation announcements almost always involves certain standard phrasings. When a tweet is posted before a cancelled trip was scheduled to depart, the tweet notes that the trip "will not operate." When a tweet is posted after a cancelled trip was scheduled to depart, the tweets notes that the trip "did not operate." No other types of announcements use these phrases, so their presence in a tweet makes them reliable indicators that the tweet is announcing a cancelled bus trip. In addition, a thorough review of the collected tweets suggested

[6] Twitter API Documentation. Accessed from: https://developer.twitter.com/en/docs.
[7] Python modules used for this process included *snscrape* (a module that allows downloading social media posts), *Selenium* (a module that allows a user to automate the process of accessing webpages), and *Tweepy* (a module that provides simplified tools for submitting requests to and processing information from the Twitter API).

that cancellations have never been announced without use of the phrase "not operate." In other words, it is unlikely that selecting for tweets that contain those words will miss many, or any, cancellation announcements. The typical cancellation announcement also includes details like the route number of the cancelled trip, the trip's scheduled departure time, the trip's origin and destination, and the reason for the cancellation. After reviewing a large subset of the collected tweets to identify common patterns, the following regular expressions (regex) expressions were used to extract the relevant details:

Variable	**Source**	**Extraction**
Date and time of Tweet	Tweet metadata	Provided in metadata
Cancelled (y/n)	Tweet text	Regex expression: 'not operate'
Route number	Tweet text	Regex expression: '[0-9]{1,3}[A-Z]{0,1}'
Scheduled departure time of cancelled trip	Tweet text	Regex expression: '[0-9]{1,2}:[0-5][0-9].{0,2}[AaPp]\.{0,1}[Mm]\.{0,1}'
Trip cancellation reason	Tweet text	Split on regex expression '[Dd]ue [Tt]o|[Dd]ue|do to', followed by selection of last element in resulting array
Trip origin/destination	Tweet text	Split on 'from' and 'to' (variable not used for analysis)

For variables extracted from the tweet text, some additional cleaning was required after extracting the relevant segment of text. Otherwise, the regex expressions listed above produced few cases in which a tweet was identified as a cancellation announcement while the other variables could not be clearly identified. The main exceptions were the cancellation reason and the trip origin and destination. Cancellation reason will be discussed in more detail below.

Origin and destination, however, were not included in the final, cleaned dataset. Unlike the other variables, the trip origin and destination were often phrased idiosyncratically. Sometimes parentheses might be added around a phrase like, "from Englewood Cliffs to NY PABT." Sometimes, the destination in a phrase like that one might be rendered, "New York PABT," sometimes, "New York Port Authority Bus Terminal," and other times, "Port Authority Bus Terminal." Without a preexisting list of trip termini as a basis for creating alternative regex patterns, creating expressions that would capture all or even most cases was prohibitively time consuming.

After the data were cleaned and used to derive additional variables, the final analytical dataset consisted of 16,983 tweets and the following variables:

Variable	**Description**
Date and time of Tweet	Used to derive hour, day of the week, month, and year
Cancelled (y/n)	A Boolean indicating a tweet was a cancellation. Used to calculate counts of cancellations by time period, route, etc.
Hour	Hour of the day during which a tweet was posted
Day of the week	Day of the week on which a tweet was posted
Month	Month in which a tweet was posted
Year	Year during which a tweet was posted
Route number	Route number for the route to which the cancelled trip belonged
Scheduled departure time of cancelled trip	The time (in hh:mm am/pm) at which a cancelled trip was scheduled to depart
Trip cancellation reason	The given reason, if any, for the cancellation
N =	16,983

Analysis

Cancellations by year

Although the lack of origin and destination limits the usefulness of the dataset, it can still be used to answer many questions. For example, a notable pattern shows up even in a simple table of the number of cancellation tweets posted by year:

Year	**Cancelled trips**
2013	397
2014	1,552
2015	225
2016	1,239
2017	2,527
2018	2,521
2019	2,823
2020	5,699
Total	16,983

The number of cancellation tweets is very inconsistent across years. If not due to uncaught errors in collecting or processing the tweets, this probably either suggests that not all cancelled bus trips are announced on the Twitter feeds, or that the number of cancellations really has changed wildly over the last seven years. This means that the apparent patterns discussed below need to be treated with caution—they may reflect a tendency to announce some cancellations and not others on these two Twitter feeds, rather than any particular pattern in which some trips are actually cancelled more than others.

It may be, for example, that trips that are cancelled for a particular reason have been counted or announced more fully since 2013. To consider whether this may be the case, it is helpful to consider how cancellation reasons were identified and assigned to the tweets.

Cancellation reason

Cancellation reasons were much simpler to identify than origin/destination pairs, but they were also more idiosyncratic than variables like departure time or route number. Importantly, a manual inspection of the tweets' text showed that not all cancellation announcements included a reason. When they did include a reason (universally preceded by the phrase, "due to"), some categories appeared more obvious than others. Operator availability, for example, was an obvious candidate even before examining the tweets, due to the ongoing operator shortage affecting NJ Transit and other agencies.[8] Several other reasons suggested themselves, too, based on common operational issues that are as familiar to regular transit riders as they are to agency staff—these reasons include traffic, police activity, and medical emergencies, among others.

On examining the text of cancellation announcements, all of these categories and more were apparent, in addition to the fact that the tweets usually used standardized phrasing for each of them. To take one example, no tweet about operator availability did not include the word "operator." Most other reasons were similarly simple to classify using the regex expressions in the table below:

Cancellation reason	**Extraction expression**	**Matched tweets**
Operator availability	'[Oo]perator|[Mm]anpower|[Aa]vailability'	6783
Operational issue	'[Oo]perational|[Oo]perating [Dd]ifficulty|[Oo]peration issue'	4283
Mechanical issue	'[Mm]echanical|[Vv]ehicle[\-][Rr]elated'	2163
Police activity	'[Pp]olice'	205
Weather	'[Ff]looding|[Ww]eather|[Rr]ain|[Ii]cy'	43
Accident (i.e. collision)	'[Aa]ccident'	212
Traffic	'[Cc]ongestion|[Tt]raffic'	99
Accessibility issue	'[Aa]ccessibility'	151
Schedule adjustment	'[Ss]chedule'	831
Medical emergency	'[Mm]edical|[Ii]njury'	98
Other or truncated	Inverse of '[Oo]perator|[Mm]anpower|[Aa]vailability|[Oo]perational|[Oo]perating [Dd]ifficulty|[Oo]peration issue|[Mm]echanical|[Vv]ehicle[\-][Rr]elated|[Pp]olice|[Ff]looding|[Ww]eather|[Rr]ain|[Ii]cy|[Ss]chedule|[Aa]ccident|[Cc]ongestion|[Tt]raffic|[Aa]ccessibility|[Mm]edical'	2115

[8] Bliss, Laura. June 28, 2018. "There's a Bus Driver Shortage. And No Wonder." CityLab. Accessed from: https://www.bloomberg.com/news/articles/2018-06-28/there-s-a-bus-driver-shortage-and-no-wonder.

The table makes clear that operator availability is indeed a major issue affecting NJ Transit's service reliability. However, it also suggests that the categories above are far from comprehensive, because the "Other" category is the fourth largest, after operator availability, operational issues, and mechanical issues. Since the "Other" category is defined as all tweets that are not captured by one of the other categories, the size of the category suggests one of three possibilities: the category is catching many announcements for which a reason was not provided at all; the category is catching many announcements whose provided reason does not fall into one of the other designated categories; or, the category is catching tweets for which no reason could be identified for some other reason.

The last possibility is certainly true for many of the announcements, mostly due to the truncation of a large number of tweets. A short URL was appended to the truncated tweets, but the links (originally to the full-length announcement) are now dead. A likely reason for the truncation is that, prior to late 2017, Twitter's default limit of 140 characters per tweet prevented most of the provided reasons from showing up in the text of the tweets themselves.[9] These are counted up in the table below, which shows that link-shortened tweets no longer appear after 2018. (NJ Transit may not have immediately adjusted their procedure in response to the change in character limits in 2017.) A substantial number of tweets also provide no reason, even when not truncated.

Figure 3: Truncated tweet example, accessed at: https://twitter.com/NJTRANSIT_NBUS/status/540322389733556224

Year	**Truncated tweets (containing 't.co' or 'bit.ly' links)**
2013	1
2014	121
2015	56
2016	167
2017	486
2018	484
2019	1
2020	0
Total	1316

Both of these characteristics cause tweets to fall into the "Other" category, since they lack any text to check against the expressions laid out in the table on the previous page. As a result, comparisons of cancellation reasons across the years using this dataset are inevitably harmed, and in unpredictable ways—certain reasons may be more likely to be undercounted than others as a result of truncation or the choice not to report a reason.

[9] Larson, Selena. November 7, 2017. "Welcome to a world with 280-character tweets." CNN. Accessed from: https://money.cnn.com/2017/11/07/technology/twitter-280-character-limit/index.html.

This suspicion is borne out by the breakdowns below of cancellation reasons by year. Notable numbers are bolded:

Cancellation reason	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**All**
Accessibility issue	1.8%	1.4%	0.9%	1.3%	1.5%	0.6%	1.1%	0.4%	0.9%
Accident	0.0%	0.0%	3.6%	1.8%	1.7%	1.2%	2.0%	0.9%	1.2%
Mechanical issue	24.4%	20.6%	15.6%	16.1%	13.7%	6.3%	17.1%	9.2%	12.7%
Medical emergency	1.0%	0.6%	1.3%	0.9%	0.5%	0.9%	0.5%	0.4%	0.6%
Operational issue	33.2%	24.5%	0.4%	7.4%	**60.3%**	**69.9%**	11.8%	1.0%	25.2%
Operator availability	0.0%	0.0%	0.0%	0.5%	0.0%	0.1%	**63.9%**	**87.2%**	39.9%
Other or truncated	27.2%	43.8%	28.4%	14.7%	21.1%	20.1%	1.3%	0.1%	12.5%
Police activity	8.1%	5.8%	0.0%	0.8%	0.6%	0.3%	0.5%	0.6%	1.2%
Schedule change	0.0%	0.1%	**49.3%**	**56.5%**	0.3%	0.3%	0.1%	0.0%	4.9%
Traffic	3.8%	1.2%	0.4%	0.1%	0.1%	0.4%	1.6%	0.1%	0.6%
Weather	0.5%	2.1%	0.0%	0.0%	0.2%	0.0%	0.1%	0.0%	0.3%
All (count)	397	1552	225	1239	2527	2521	2823	5699	16983

Cancellation reason	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**All**
Accessibility issue	7	21	2	16	39	14	32	20	151
Accident	0	0	8	22	44	30	56	52	212
Mechanical issue	97	319	35	199	347	159	482	525	2163
Medical emergency	4	9	3	11	13	22	14	22	98
Operational issue	132	380	1	92	**1523**	**1762**	334	59	4283
Operator availability	0	0	0	6	0	2	**1803**	**4972**	6783
Other or truncated	108	679	64	182	533	507	36	6	2115
Police activity	32	90	0	10	14	7	15	37	205
Schedule change	0	2	**111**	**700**	7	8	3	0	831
Traffic	15	19	1	1	2	10	45	6	99
Weather	2	33	0	0	5	0	3	0	43
All (count)	397	1552	225	1239	2527	2521	2823	5699	16983

Some trends are clear. Overall, many more cancellations are reported as the years go on. In 2020, the two accounts tweeted almost 5700 cancellation announcements, compared with roughly 400 in 2013. Two other trends are apparent: 2015 and especially 2016 show many more cancellations reported due to schedule changes than any other years, and operator availability is reported as a reason more or less only in 2019 and 2020.

The first trend seems plausible, even if it is hard to imagine what would lead to 700 schedule-related cancellations in 2016 and only 7 the next year. The second trend, however, may well be due to similar types of cancellations being reported differently over time. Specifically, phrases like “operational issue” and “operator availability” in practice seem to refer to similar types of cancellations. It seems unlikely that any “operational issues” affecting NJ Transit buses in 2018 would affect them much less strongly in 2019. Even more unlikely is that operator availability caused no cancellations before 2019. Taken together with the roughly similar number of cancellations attributed to these two reasons over several years, it seems likely that some

"operational issue" cancellations began to be reported as "operator availability" cancellations starting in 2019, leaving a smaller set of several hundred cancellations to be attributed to other operational issues. This does not make it impossible to draw conclusions about cancellation causes, but it does mean conclusions can only be tentative.

Even with unclear categories, it is still possible to use the cancellation tweets to suggest when or on which routes certain types of cancellations are most common. The table below takes routes 13 and 18 as examples:

Cancellation reason	**Route 13**	**Route 87**
Accessibility issue	1.2%	0.9%
Accident	0.6%	0.6%
Mechanical issue	**10.5%**	**3.8%**
Medical emergency	0.6%	1.3%
Operational issue	29.9%	40.7%
Operator availability	42.4%	40.8%
Other or truncated	9.0%	8.6%
Police activity	0.6%	0.5%
Schedule change	**4.9%**	**2.4%**
Traffic	0.0%	0.2%
Weather	0.2%	0.3%
Count (all)	488	637

Route 13 is a north-south route that runs from Nutley in the north to Newark's southern border. Route 87 is also a north-south route, running from Hoboken to the southern tip of Jersey City. As the table shows, cancellations for these two routes are roughly similar in some respects. The routes have roughly similar proportions of cancellations reported due to traffic, weather, crashes, and "other" reasons.

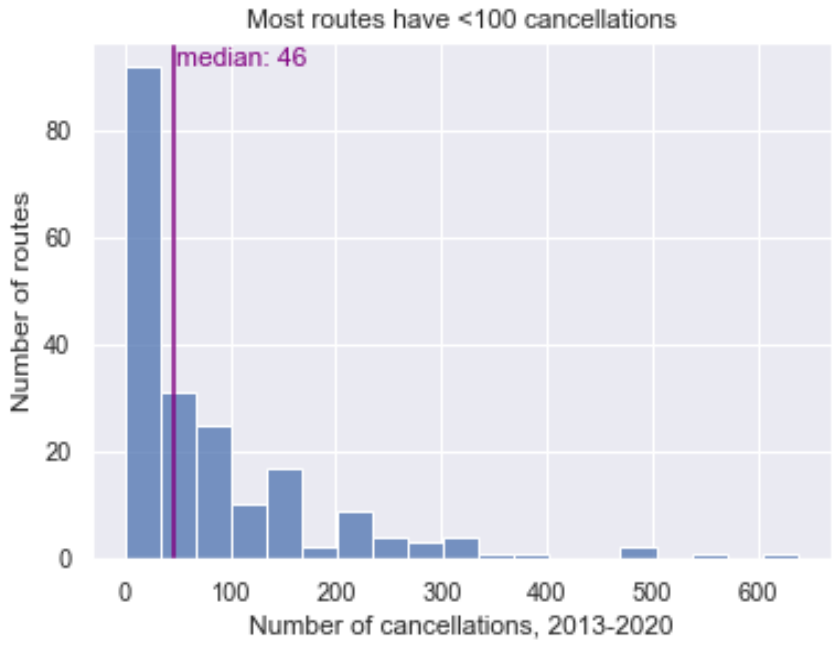
The comparison also reveals some differences. For reasons discussed above, operational issues and operator availability are hard to interpret, but route 13 has fewer cancellations attributed to those two reasons taken together. More schedule change cancellations were reported for route 13, which had 24 such cancellations between 2013 and 2020. Most notably, route 13 had many more cancellations reported due to mechanical issues, at 51 cancellations compared with route 87's 24 cancellations. This dataset does not have information that could explain why some routes might have issues others do not, but identifying the routes that are particularly affected by certain types of cancellations is a necessary first step for further analysis. This kind of route-by-route analysis is possible using a variety of the collected variables, as discussed in the following section.

Cancellations by route

Grouping cancellations by route allows several interesting questions to be asked. Among all the routes mentioned in cancellation tweets, how many cancellations does the typical route see? Do different routes see cancellations at different times of day? Is the ridership of a route associated with the number of cancellation tweets it appears in—in other words, do cancelled trips occur on routes where many riders would be affected? This section combines the tweet data with ridership data by route to try to answer some of these questions.

Two natural first questions are how many cancellations are reported for each route, and how many are reported for the typical route. The histogram in Figure 4 visualizes the distribution of cancellations by route. The distribution is very skewed, with some routes having a large number of cancellations and the large majority of routes having fewer than 100 over the 7-year period covered by the data. Half of routes have fewer than 46 cancellation announcements, while the average cancellation count of 83 reflects the presence of a small number of routes with cancellation counts numbering in the hundreds.

Figure 4: Histogram of cancellations by route (2013-2020)



Assuming that the announced cancellations reflect actual cancellations, this distribution suggests some routes especially suffer. As can be seen in the table below, the 10 most-cancelled routes also include some very heavily ridden routes, with yearly ridership in some cases of more than 4 million. These might be routes to target for service improvements, or at least for analysis to identify the main cancellation causes.

Route	**Cancellations 2013-2020**	**Avg. Yearly Ridership 2015-2018**
87	637	3,224,566
452	549	410,759
400	400	1,440,766
13	488	4,035,659
25	396	3,734,411
126	346	4,015,186
39	329	2,076,422
404	315	479,421
80	304	1,972,852
166	303	4,531,713

Departure time is another important piece of metadata available in the cancellation tweets. As with cancellation reason, the scheduled departure time of cancelled trips can provide information about the nature of cancellations—knowing when cancellations occur can help explain why they occur, as well as give a better sense of how the system as a whole is impacted by cancellations.

The table below shows cancellations by hour of the day for all routes, as well as for each of the ten routes with the highest ridership and at least one cancellation.:

	Percent of all cancellations taking place within each hour										
Time of day:	**13**	**166**	**25**	**39**	**400**	**404**	**407**	**452**	**80**	**87**	**All**
12am - 1am	1%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%
1am - 2am	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%
2am - 3am	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
3am - 4am	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
4am - 5am	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%
5am - 6am	2%	4%	1%	1%	3%	4%	3%	1%	1%	1%	3%
6am - 7am	9%	15%	4%	6%	9%	5%	1%	5%	10%	4%	12%
7am - 8am	9%	29%	7%	4%	3%	4%	6%	4%	43%	32%	17%
8am - 9am	4%	5%	4%	3%	4%	3%	2%	2%	9%	13%	5%
9am - 10am	1%	1%	1%	1%	2%	1%	1%	1%	0%	2%	2%
10 am - 11am	1%	1%	1%	2%	1%	0%	2%	1%	1%	1%	1%
11am - 12pm	1%	3%	1%	0%	2%	1%	1%	1%	1%	1%	2%
12pm - 1pm	2%	0%	0%	2%	5%	1%	1%	1%	2%	0%	2%
1pm - 2pm	3%	1%	1%	1%	6%	6%	5%	2%	3%	3%	3%
2pm - 3pm	6%	2%	3%	3%	8%	9%	5%	9%	5%	5%	6%
3pm - 4pm	11%	5%	7%	4%	13%	19%	7%	9%	6%	9%	8%
4pm - 5pm	13%	6%	22%	16%	14%	19%	20%	14%	6%	9%	14%
5pm - 6pm	12%	9%	32%	16%	7%	10%	8%	15%	4%	8%	10%
6pm - 7pm	9%	10%	10%	15%	9%	5%	15%	14%	5%	6%	7%
7pm - 8pm	5%	7%	5%	12%	4%	4%	12%	7%	3%	3%	5%
8pm - 9pm	4%	3%	2%	12%	5%	3%	6%	4%	0%	2%	3%
9pm - 10pm	2%	0%	1%	2%	1%	3%	1%	4%	0%	1%	2%
10pm - 11pm	3%	1%	0%	1%	1%	2%	2%	3%	0%	0%	1%
11pm - 12am	0%	0%	1%	1%	2%	0%	1%	2%	0%	0%	1%
All cancellations	488	303	396	329	493	315	286	549	304	637	16983

In general, each of the routes and the routes as a whole show the same pattern, with cancellations concentrated during the morning and afternoon peaks. This is expected, given that many of the cancellation reasons described above might be most impactful during peak periods, due to the scheduling of more service. There is a higher need for operators during the peaks, for example, and traffic is especially bad.

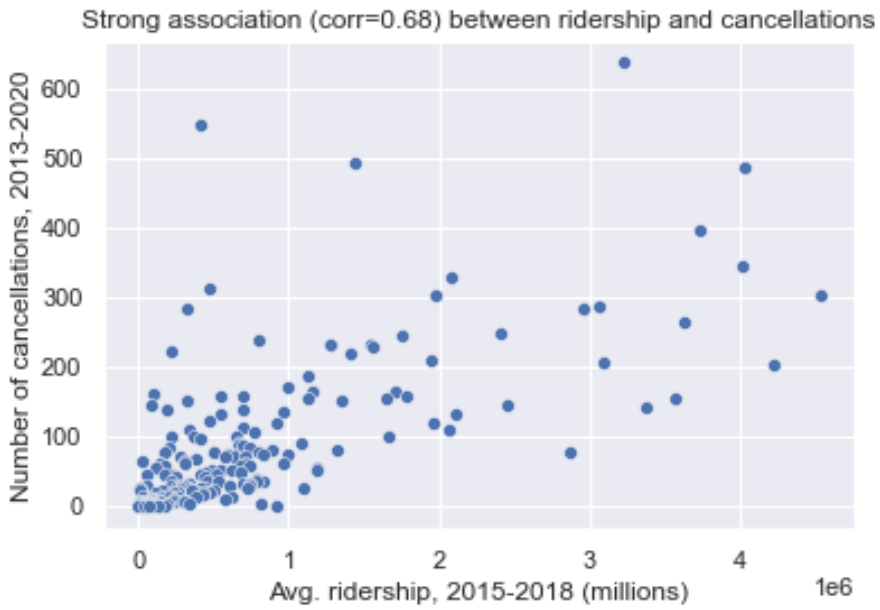
However, the table also shows that this broad pattern varies significantly by route. Although all ten of the most-ridden routes show more cancellations around the morning and afternoon peaks, some show reported cancellations very prominently at one or the other peak period. The

80 and 87, both heavily used local routes serving Jersey City and Hoboken, show more than 30% of their cancellations taking place between 7am and 8am. The 25, a route serving Newark but not otherwise obviously different from the 80 and 87, shows a similar proportion of cancellations occurring between 5pm and 6pm. There are no immediate explanations for the difference, which might be related to anything from varying traffic patterns to varying peak period strains on different bus garages. As with the breakdown by route and cancellation reason, the breakdown by route and hour shows that routes with similar numbers of cancellations can experience those cancellations in quite different ways.

Cancellations and ridership

The last major issue this dataset can help describe is the relationship between cancellations and ridership. As with bus cancellations, NJ Transit does not make ridership data readily available in aggregate or by route. However, NJ Transit has submitted ridership data to the NJ state legislature several times since 2013 during budget hearings. The most recent such submission occurred in 2019, and the document included ridership for FY 2015, 2016, 2017, and 2018, as well as projected annual ridership for FY 2019.[10] These data help answer two other important questions about cancellations: how many people do they affect, and is there a relationship between ridership and the number of cancellations? The scatterplot in Figure 5 suggests cancellations affect many people. In other words, there appears to be a strong association between the number of yearly riders on a route and the number of cancellations on the route since 2013:

Figure 5: Number of cancellations vs. route ridership



[10] New Jersey Transit. 2019. "Discussion Points." Accessed from: https://www.njleg.state.nj.us/legislativepub/budget_2020/NJT_response_2020.pdf. Pp. 33-38.

The correlation table below reinforces this impression. Routes' average yearly ridership for the years 2015 through 2018 has a strong linear correlation with their number of cancellations between 2013 and 2020, at almost 0.7. Such a strong, positive correlation suggests a very close association between a route's ridership and the number of cancellations it sees. In other words, the more riders on a route, the more cancellations it can be expected to have. (Farebox recovery, also available from the NJT ridership tables, has only a weak positive association with cancellations.)

The cancellation and ridership data on their own do not provide much insight into the causes behind this relationship, and there could be many. The most obvious is that higher ridership routes are likely to be those with more service (i.e. more trips), and routes with more trips would almost always have more cancellations. This can be accounted for by dividing a route's cancellations by its ridership, creating a ratio that reflects the number of cancellations "per capita" on a given route. As can be seen in the correlation table, this variable has a weak and negative association with ridership. That is, the more riders on a route, the lower its ratio of cancellations to riders.

	Cancellations 2013-2020	**Average ridership 2015-2018**	**Average farebox recovery 2015-2018**	**Cancellations per 1 mill. riders**
Cancellations 2013-2020	1	0.69	0.02	0.17
Average ridership 2015-2018	0.69	1	0.25	-0.19
Average farebox recovery 2015-2018	0.02	0.25	1	-0.24
Cancellations per 1 mill. riders	0.17	-0.19	-0.24	1

A negative relationship between ridership and cancellations per rider will always be expected just because of the definition of the ratio—more riders means a larger denominator, which means a smaller ratio. Still, the ratio is useful for thinking about the relative impact of a cancellation on different routes. The table below shows the ten routes with the largest cancellation per rider ratios:

Route number	**Cancellations 2013-2020**	**Average ridership 2015-2018**	**Cancellations per 1 mill. riders**
417	66	23062	2861.851
555	28	10166	2754.279
82	146	79586	1834.494
453	163	105415	1546.27
452	549	410759	1336.55
414	24	19053	1259.644
405	224	222927	1004.813
418	8	8759	913.3463
407	286	321890	888.5023
451	48	59516	806.5058

The routes with the highest ratios are generally, though not entirely, those with very low ridership. No route in the table has an average yearly ridership of more than 500,000. The ratio is far from a perfect measure of the relative impact of cancellations. In the future, the tweet dataset could be combined with a count of trips per route (e.g. from GTFS). Knowing the number of trips per route would directly account for the amount of service by describing the relationship between ridership and a cancellation rate, rather than a cancellation count

A measure like that might also clarify the factors that could cause routes with higher ridership to be associated with more cancellations. Routes with higher ridership might be located in places where they are more likely to be affected by the cancellation reasons discussed previously, for example.[11] High-ridership local routes in northern New Jersey like the 13 and 87 may run out of garages that are at or beyond capacity. High-ridership "commuter" routes like the 126 between Hoboken and Manhattan or the 400 into Philadelphia might be particularly affected by traffic conditions or operator availability. These and other explanations could be behind the close association between ridership and cancellations.

Conclusion and Recommendations

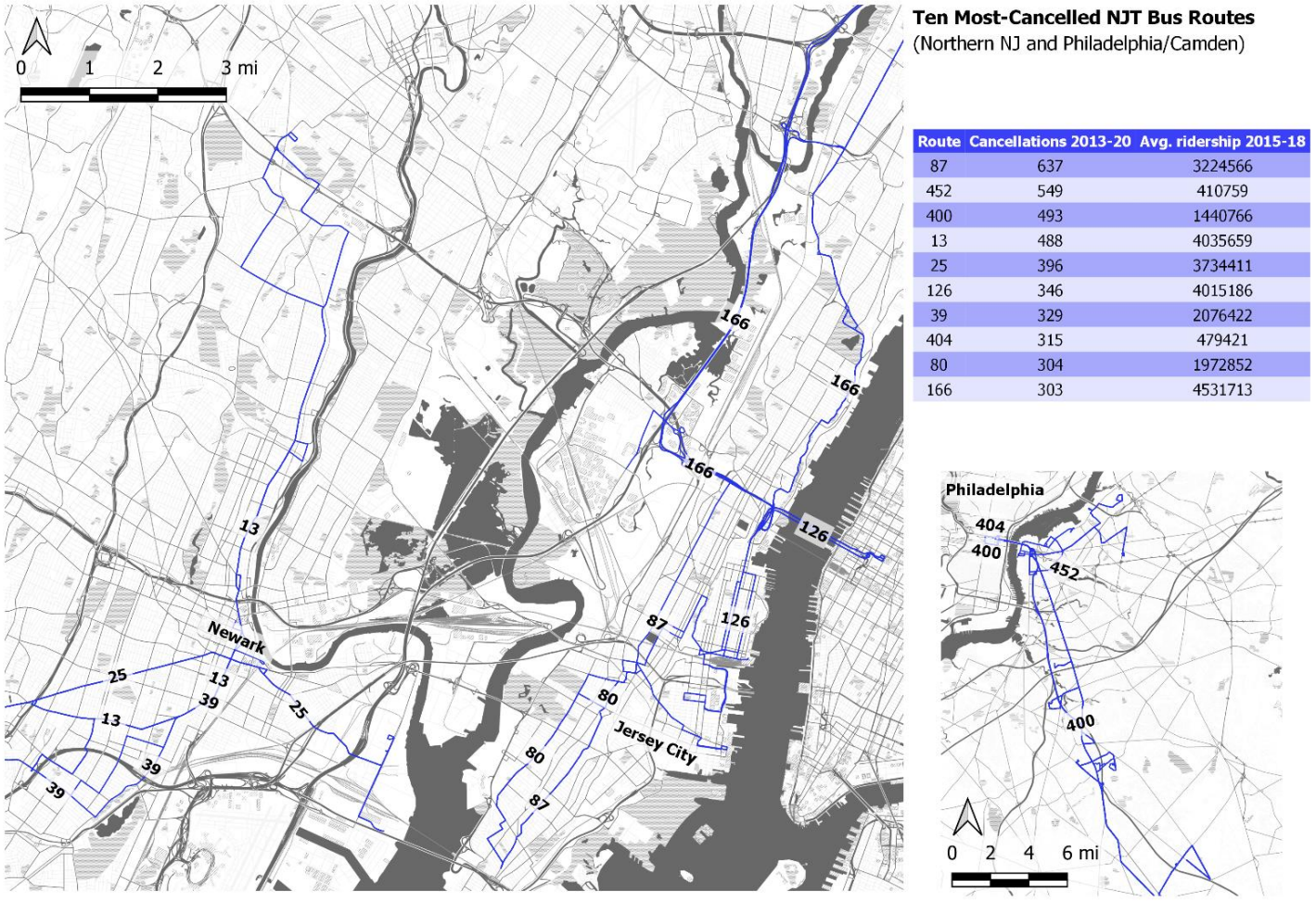
Although relatively easy to collect, the tweet cancellation dataset is not simple to process or interpret. The main issues include unclear categories for cancellation reasons and the uncertainty that recent increases in cancellation tweets reflect an increase in actual cancellations. The cancellation announcement tweets are not intended for analytical use, so difficulty in using them for that purpose is understandable.

However, if NJ Transit intends to create a public bus cancellation dashboard like its rail cancellation dashboard, these observations lead to two main recommendations: that NJ Transit create and use clearly defined categories for recording and announcing bus trip cancellations, and that NJ Transit ensure that all bus trip cancellations are accounted for. The analysis in this paper shows that cancellation announcements vary considerably by cancellation reason, by time of day, and by route. With improved bus cancellation data, factors leading to trip cancellation could be more confidently identified, and the state of NJ Transit's bus service could be more transparently conveyed to transit users in New Jersey.

[11] Maps of the ten most-cancelled routes, the ten most-ridden routes with at least one cancellation, and the ten most-ridden routes without any cancellations are provided in the appendix.

Appendix

Figure 6: Map of ten most-cancelled NJT bus routes



Ten Most-Cancelled NJT Bus Routes
(Northern NJ and Philadelphia/Camden)

Route	Cancellations 2013-20	Avg. ridership 2015-18
87	637	3224566
452	549	410759
400	493	1440766
13	488	4035659
25	396	3734411
126	346	4015186
39	329	2076422
404	315	479421
80	304	1972852
166	303	4531713

Figure 7: Ten most-ridden NJT bus routes

Route	Cancellations 2013-20	Avg. ridership 2015-18
166	303	4531713
1	204	4231856
13	488	4035659
126	346	4015186
25	396	3734411
165	265	3636338
94	156	3566796
27	144	3373530
87	637	3224566
139	206	3090160

Ten Most Ridden NJT Bus Routes
(All in Northern NJ)

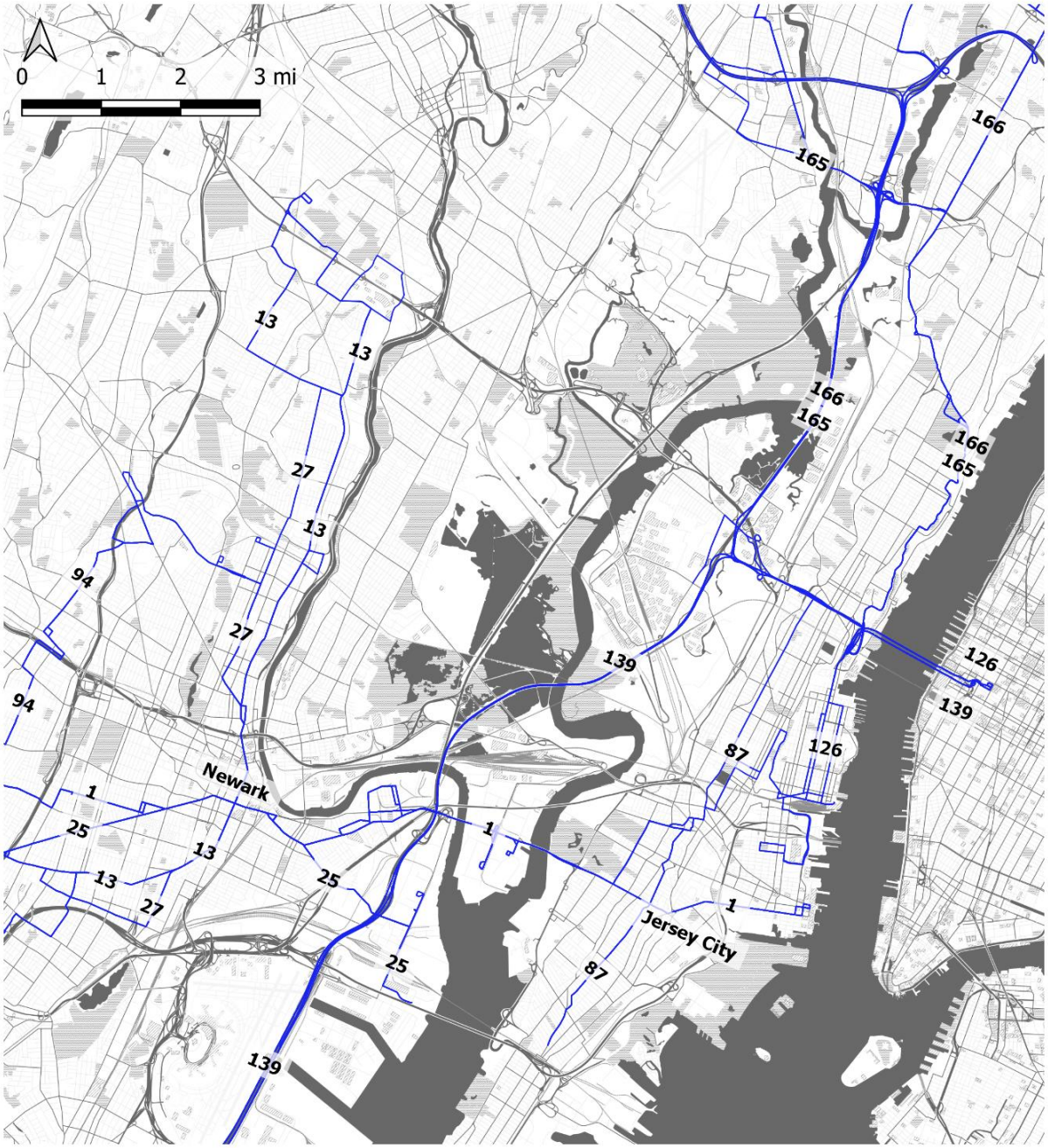


Figure 8: Ten most-ridden NJT bus routes with no cancellations

Route	Avg. ridership 2015-20
10	1323707
119	1036758
88	893747
22	666779
709	404559
744	396704
815	350396
780	292036
814	269087
832	251542

Ten Most Ridden NJT Bus Routes with No Cancellations
(All in Northern and Central NJ)