Citi Bike Analysis Appendices

Big Data Systems (DS 5110)

August 11, 2021

Diana McSpadden (hdm5s)
Nick Daniello (njd9e)
David Fuentes (dmf4ns)
Eric Sarani (es5cj)
Abigail Bernhardt (aeb4rv)

TABLE OF CONTENTS

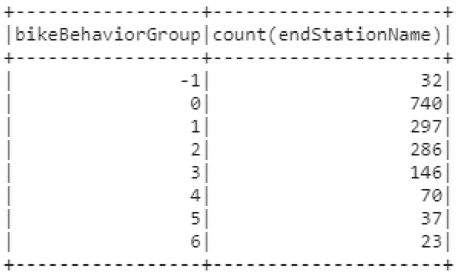
LIST OF TABLES AND FIGURES

1 APPENDIX A: SUPPORTING FIGURES

Table 1-1: “BAD” weather temperature thresholds by month

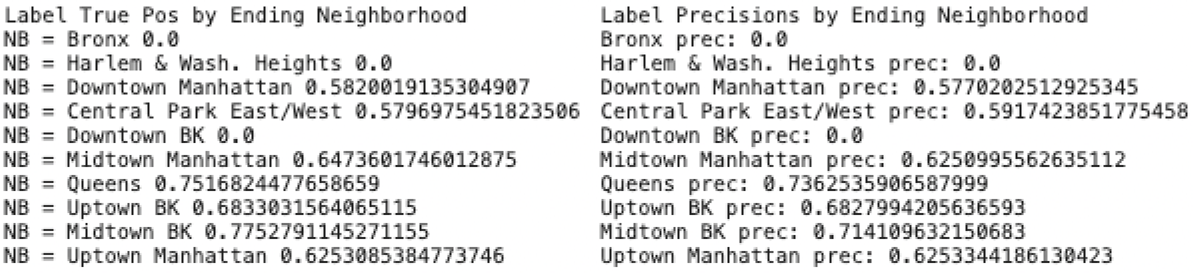
Month	BAD condition LOW temp	BAD condition HIGH temp
Jan	< mean - (1 * std)	90
Feb	< mean - (1 * std)	90
Mar	< mean - (1 * std)	90
Apr	< mean - (1 * std)	90
May	< mean - (2 * std)	90
Jun	45	> mean + (1.5 * std)
Jul	45	> mean + (1.5 * std)
Aug	45	> mean + (1.5 * std)
Sep	< mean - (3 * std)	90
Oct	< mean - (1 * std)	90
Nov	< mean - (1 * std)	90
Dec	< mean - (1 * stds	90

Table 1-2: Stations by Bike Behavior Group



```
+----------------+--------------------+
|bikeBehaviorGroup|count(endStationName)|
+----------------+--------------------+
|              -1|                  32|
|               0|                 740|
|               1|                 297|
|               2|                 286|
|               3|                 146|
|               4|                  70|
|               5|                  37|
|               6|                  23|
+----------------+--------------------+
```

Table 1-3: Ending Neighborhood Model Performance



```
Label True Pos by Ending Neighborhood           Label Precisions by Ending Neighborhood
NB = Bronx 0.0                                  Bronx prec: 0.0
NB = Harlem & Wash. Heights 0.0                 Harlem & Wash. Heights prec: 0.0
NB = Downtown Manhattan 0.5820019135304907      Downtown Manhattan prec: 0.5770202512925345
NB = Central Park East/West 0.5796975451823506  Central Park East/West prec: 0.5917423851775458
NB = Downtown BK 0.0                            Downtown BK prec: 0.0
NB = Midtown Manhattan 0.6473601746012875       Midtown Manhattan prec: 0.6250995562635112
NB = Queens 0.7516824477658659                  Queens prec: 0.7362535906587999
NB = Uptown BK 0.6833031564065115               Uptown BK prec: 0.6827994205636593
NB = Midtown BK 0.7752791145271155              Midtown BK prec: 0.714109632150683
NB = Uptown Manhattan 0.6253085384773746        Uptown Manhattan prec: 0.6253344186130423
```

Table 1-4 Distance Predictor - RF Regression vs GBT Regression vs Linear Regression Features

Categorical Variables	Continuous Variables
Borough (4 levels)	Temperature

Hour (24 levels)	Pressure
User Type (2 levels)	Humidity
Gender (3 levels)	Wind Speed
Day (7 levels)	Rain (previous 3 hours)
Month (12 levels)	Snow (previous 3 hours)
Time-Bin (4 levels)	Cloud Cover
Year (4 levels)	Average Median Real Estate
Precipitation Event (2 levels)	Age
Zip Code (109 levels)	

Table 1-5 Model Results

Model	RMSE
Random Forest	0.955937
Gradient Boosted Tree	0.946306
Linear Regression	0.94544

Feature	GBT Importance	RF Importance	Change in Estimated Miles (beta LinReg)
Night (time-bin category)	0.0736	0.0013	0.0262
Temperature	0.0558	0.1289	0.0016 per degree
Year 2020 (category)	0.0495	0.2406	0.0926
Subscriber vs. Customer (category)	0.0430	0.1578	-0.0739
11237 (Ridgewood)	0.0455	0.0097	-0.1756
11221 (Bushwick - BedStuy)	0.0430	0.0019	0.2181
10023 (Upper West Side)	0.0150	0.0465	0.4781
11222 (Bushwick - BedStuy)	0.0166	0.0421	0.5472
10460 (South Bronx)	0.0107	0.0367	0.3956

Figure 1-1: Station rank for “GOOD” weather trip: rank > 3 sized larger, color by bike behavior group

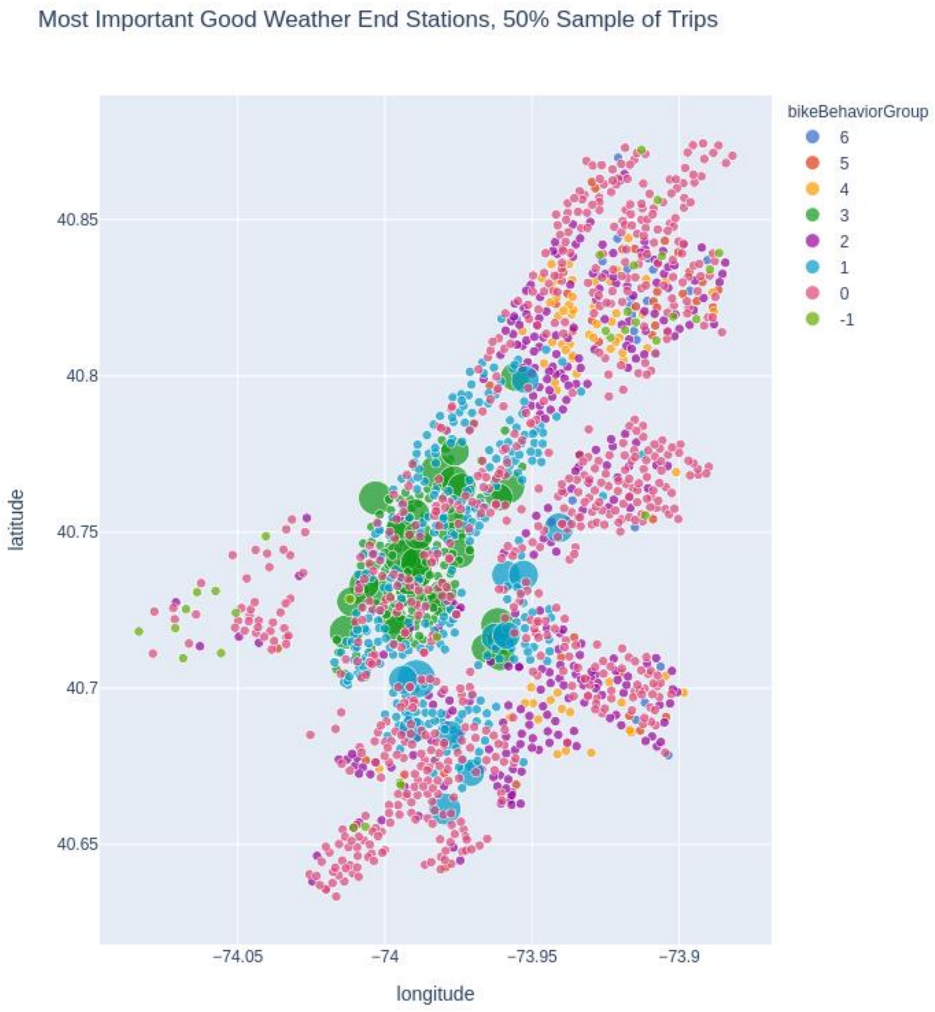


Figure 1-2: Station rank for “BAD” weather trip: rank > 3 sized larger, color by bike behavior group

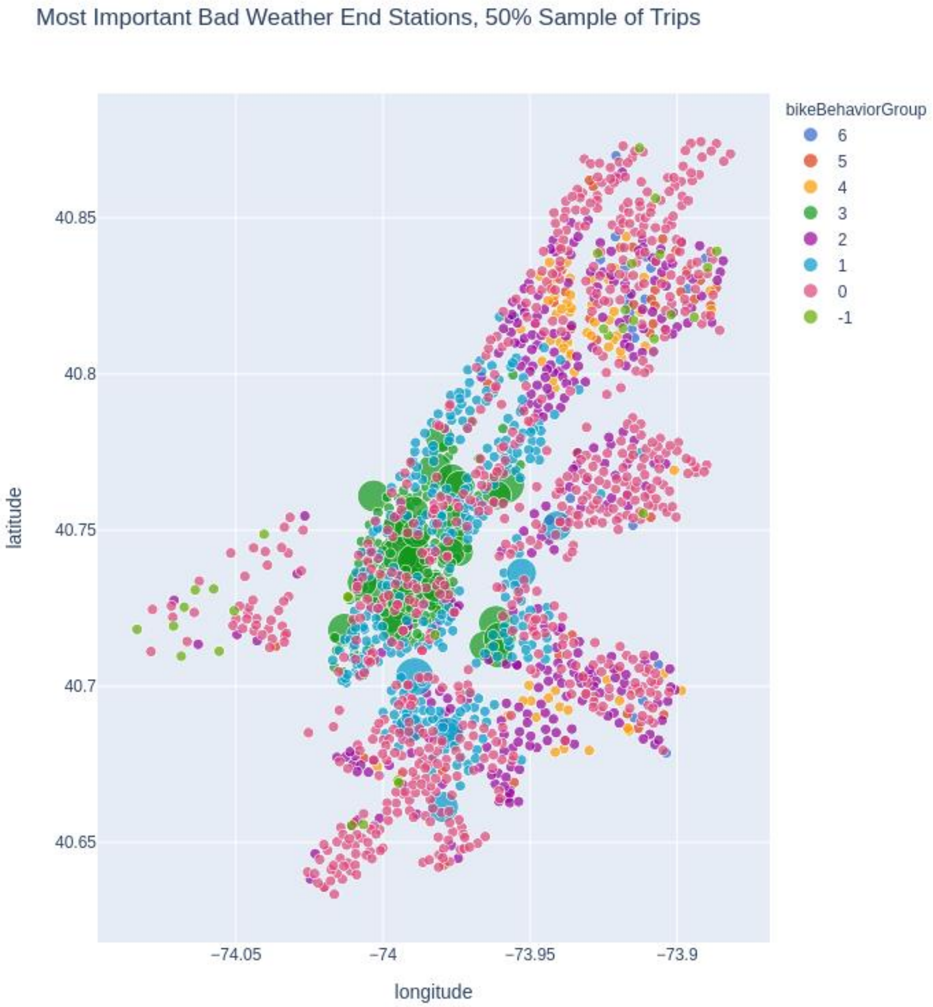


Figure 1-3: Station rank distribution for 50% sample of Citi Bike trips

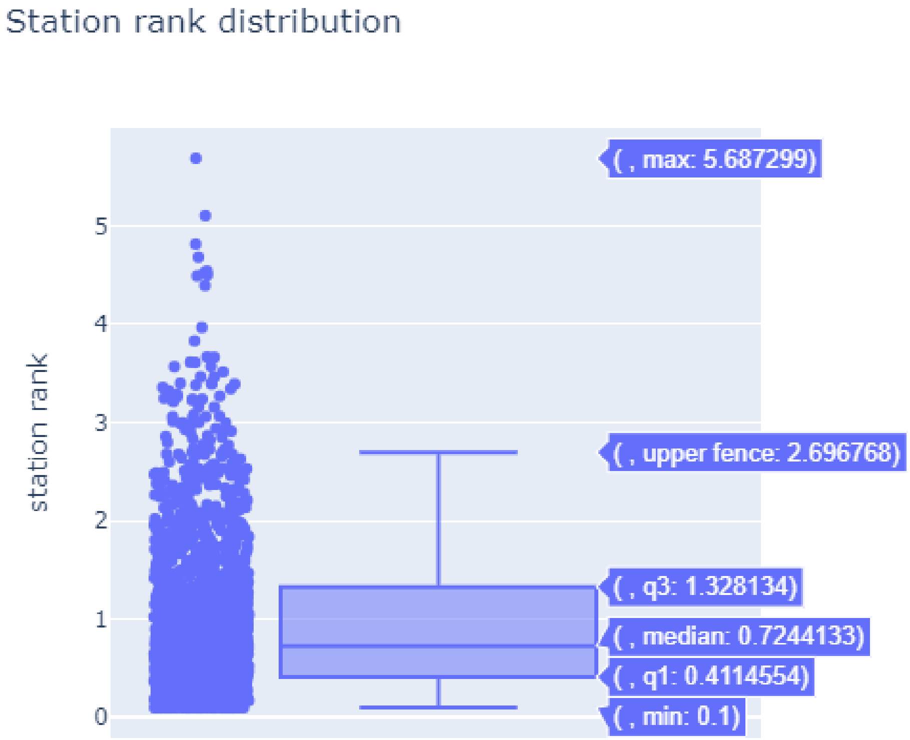


Figure 1-4 Station rank distribution by bike behavior group for 50% sample of Citi Bike trips

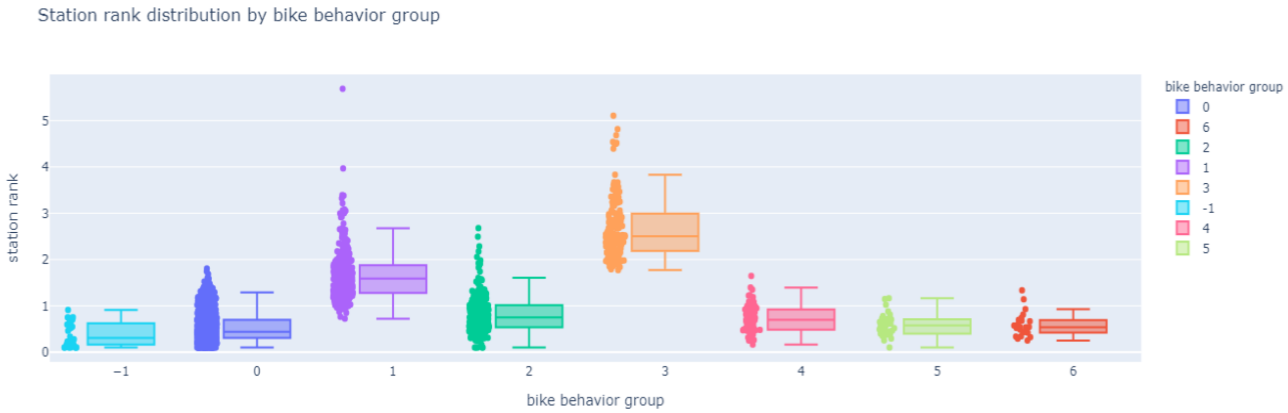
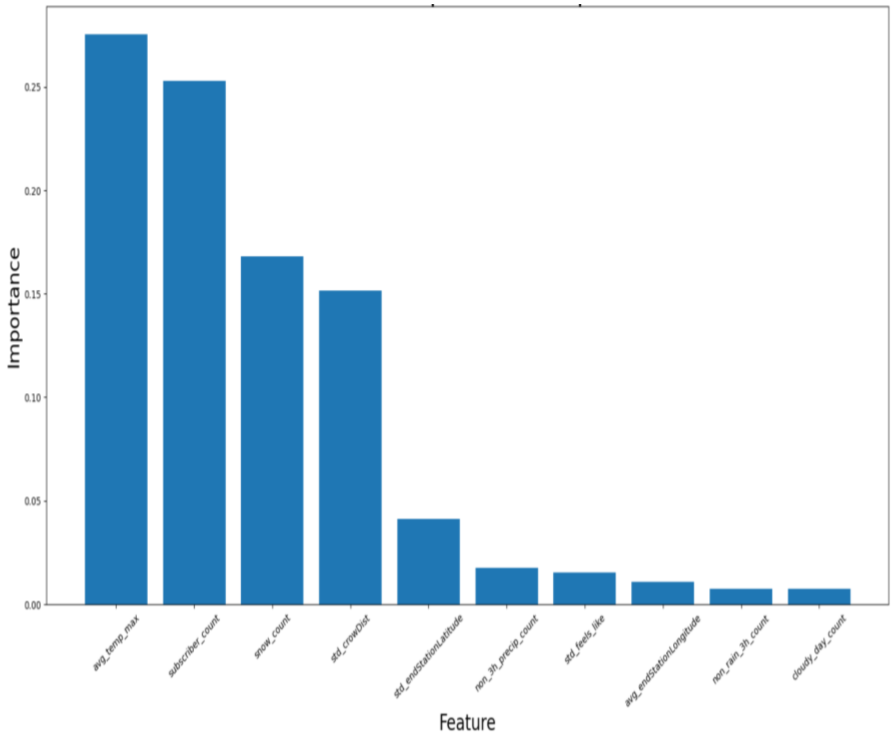


Figure 1-5 Top 10 Feature Importance for RF - Price by zip code



2 APPENDIX B: STATION RANK AND COVID-19

This analysis was based on a 30% sample of data pre-COVID-19: March 2019 - February 2020, and a 30% sample of data during COVID-19: March 2020 - February 2021.The most important stations were defined by a threshold of rank greater than three.

For the pre-COVID-19 timeframe, fourteen stations had a rank greater than three. Six of the fourteen high-ranked stations were unique to the pre-COVID-19 timeframe. The *max pre_COVID* rank was 6.97. The top five pre-COVID stations were:

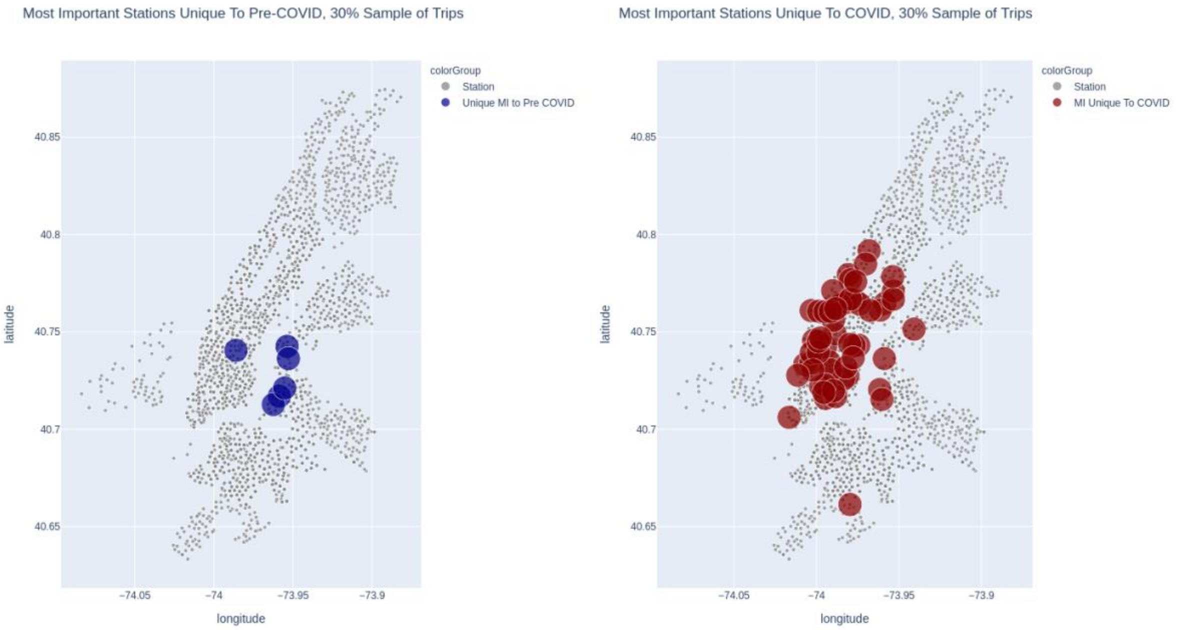
1. Front St & Washington St
2. N 6 St & Bedford Ave
3. S 3 St & Bedford Ave
4. N 12 St & Bedford Ave
5. E 17 St & Broadway

During the COVID-19 timeframe sixty-three stations had a PageRank greater than three. Fifty-five of the high-rank COVID-19 stations were unique to the COVID time frame. The max COVID-19 rank was 10.71. The top five COVID stations were:

1. 1 Ave & E 68 St
2. E 13 St & Avenue A
3. Front St & Washington St
4. Broadway & W 60 St
5. W 21 St & 6 Ave

It is interesting that in all analysis of station rank the Front St & Washington St station is consistently in the top 5 stations by station rank. Also interesting is that high ranking stations were ranked significantly higher during COVID-19, indicating that effective rebalancing of bikes was more important during this timeframe.

Figure 2-1: Unique pre-COVID-19 high ranking stations (left); unique COVID-19 high ranking stations (right)



Our hypothesis that the cause of more stations with a rank greater than three, and the increased importance of lower Manhattan neighborhoods is that more people were using Citi Bikes instead of public transportation, and more trips were ending at a natural outlet for quarantine weary folks: Central Park. We also understand that Citi Bike employees were cleaning bikes more often, and this increased attention to bikes may have led to more rebalancing .

By adding subway stations to the plots, we also found a lower percentage of high-rank stations were directly at a subway station during the COVID-19 timeframe when compared to pre_COVID-19: 33% pre-COVID-19, 25% during COVID (Figured above).

Figure 2-2: Pre-COVID-19 high rank stations with subway stations identified.

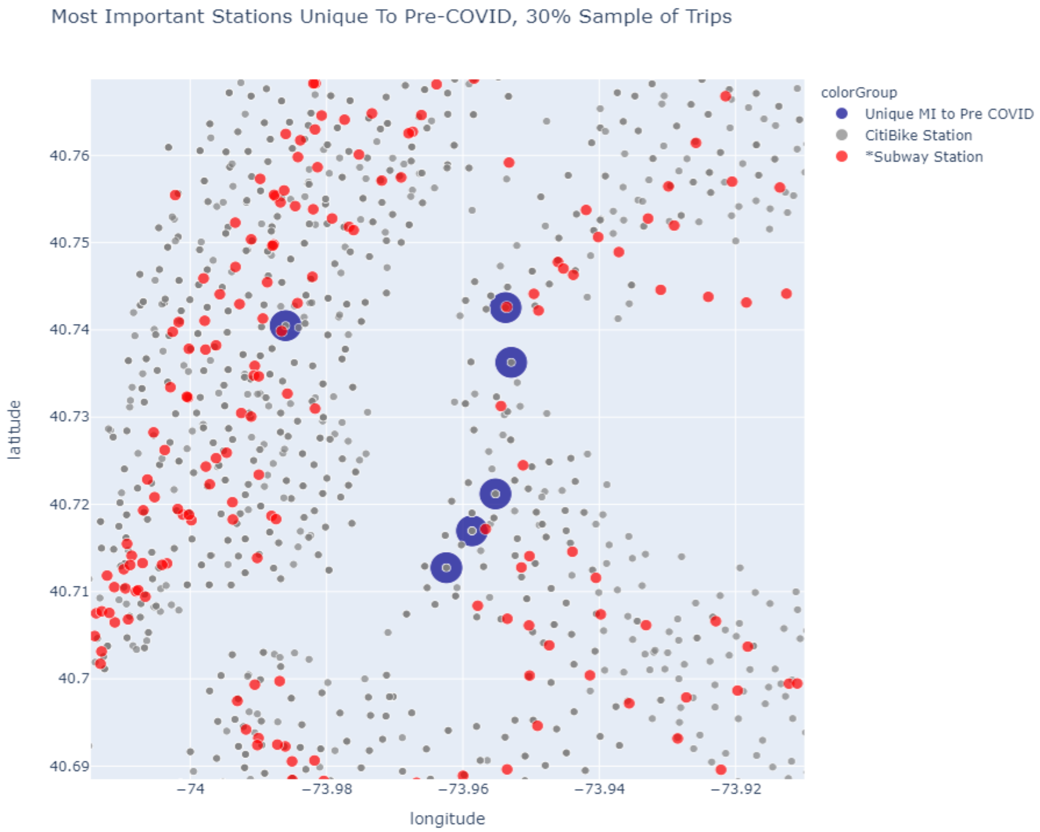
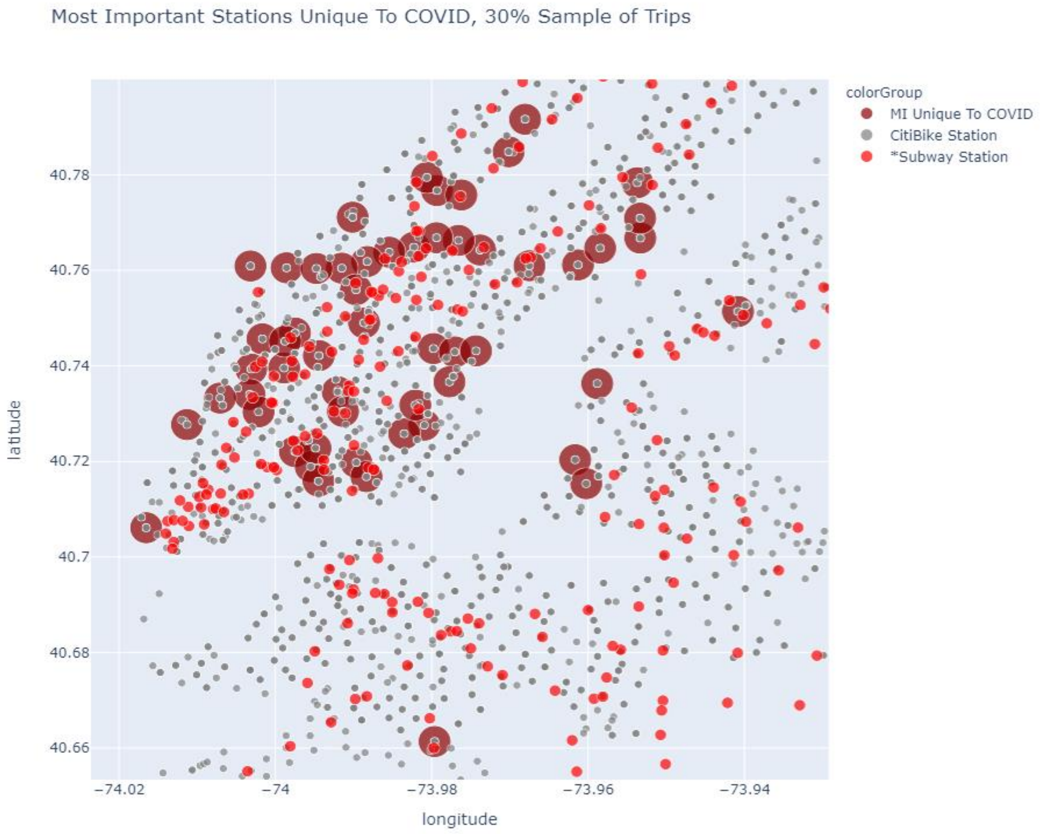


Figure 2-3: BAD weather temperature thresholds by month



3 APPENDIX C: UNIQUELY MOST IMPORTANT STATIONS IN BAD WEATHER

Top five uniquely Most Important during BAD weather trips:

1. 1 Ave & E 68 St

Major hospitals and colleges were located near this Citi Bike station. In this region, subway stations are not as dense as in other NYC locations. Regardless of weather, people need to access these locations.

2. Front St & Washington St

This station is the heart of the Directly Under Manhattan Bridge Overpass (DUMBO), which is a vibrant location within Manhattan. A hypothesis is that during *BAD* weather more people may want a bike to access this area instead of casually walking.

3. Pershing Square North

This station is next to Grand Central and is most likely related to commuting.

4. E 17 St & Broadway

This station is the heart of Union Square and a large park, with many bars, restaurants, and residences. The reasoning to why this is unique to the Most Important during bad weather, is unclear, except for what else is there to do in bad weather except go to bars, restaurants, or stay home.

5. W 21 St & 6 Ave

Trader Joe’s and a liquor store are located at this location. Individuals need food, and most need alcohol. Maybe during bad weather people prefer a speedier bike ride to other alternatives.