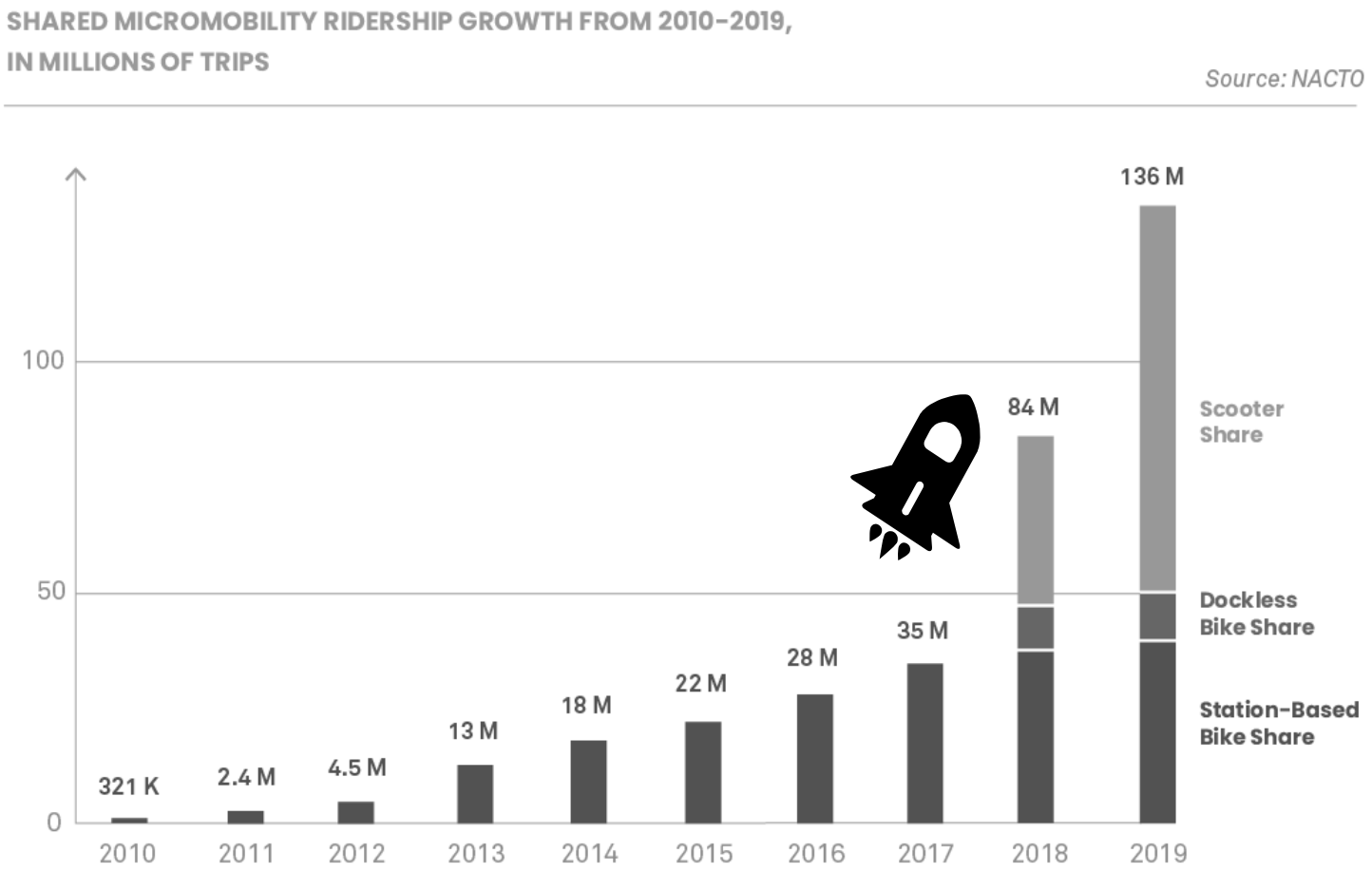
Data Science = Solving Problems = Happiness

Bike Share USA

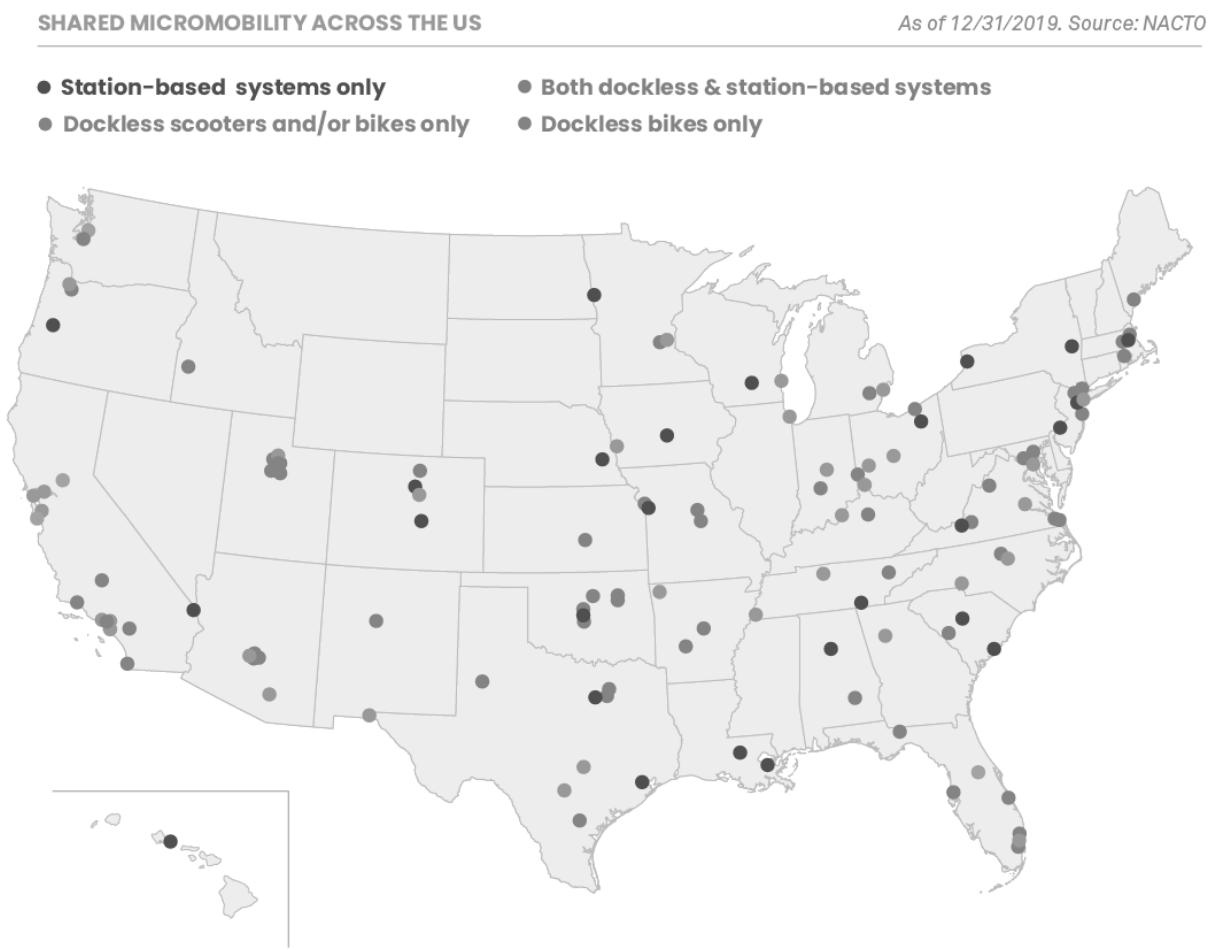
Denzel S. Williams
Springboard DS Career Track '20

What if your commute could be less…**blah**?

Micromobility is the New Black



Over the last decade micromobility ridership growth has been skyrocketing



But, micromobility is only in select locations of the country

About the Project

The What

Expand the station-based bike sharing sector of a State's micromobility services.

The Question

How many bike sharing stations should be built in the no-station zip codes of States that already have bike stations?

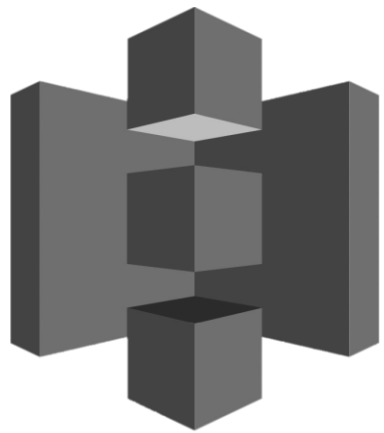
The How

Use zip code and bike sharing data of areas that have stations to build a model that can predict the number of stations that should be built in no-station zip codes.

1.

Data Engineering

Before the science comes the data.



baywheels
BLUEbikes
capital bikeshare
citi bike
DIVVY

Bike Share Trip Datasets

The subset of zip codes that have bike stations are derived from the five largest bike sharing services in the US. Each company hosts their trip data on S3 buckets for public use.

Station Data

The Trip Datasets were used to derive the station data needed for the prediction.

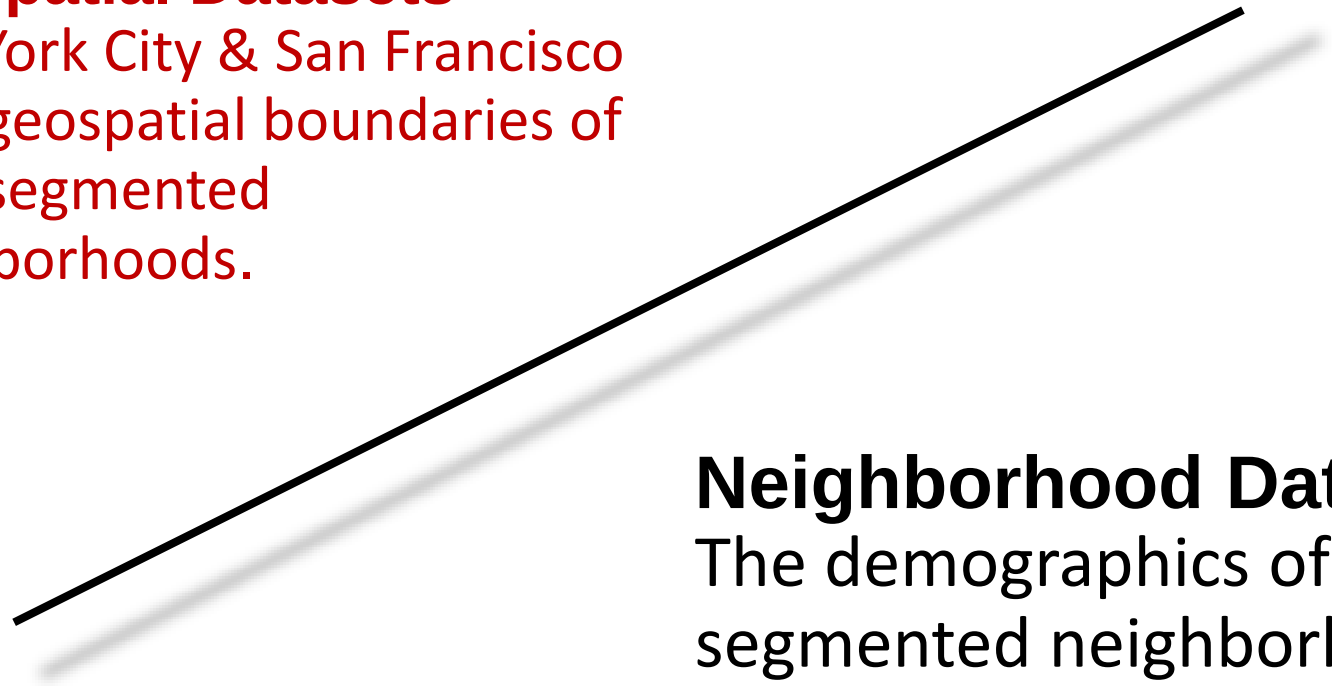

NYU
Furman
Center


SEAL OF THE CITY AND COUNTY OF SAN FRANCISCO
ORO EN PAZ
FIERRO
EN GUERRA


NYC OpenData

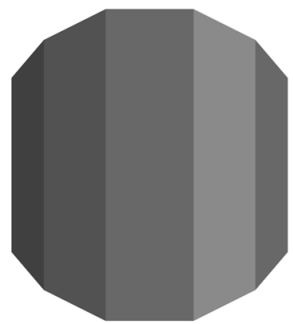
Geospatial Datasets

New York City & San Francisco have geospatial boundaries of their segmented neighborhoods.

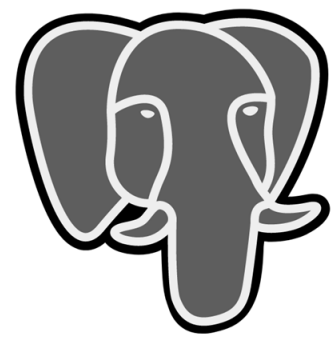


Neighborhood Datasets

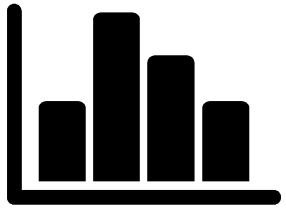
The demographics of those segmented neighborhoods.



Amazon RDS



PostgreSQL



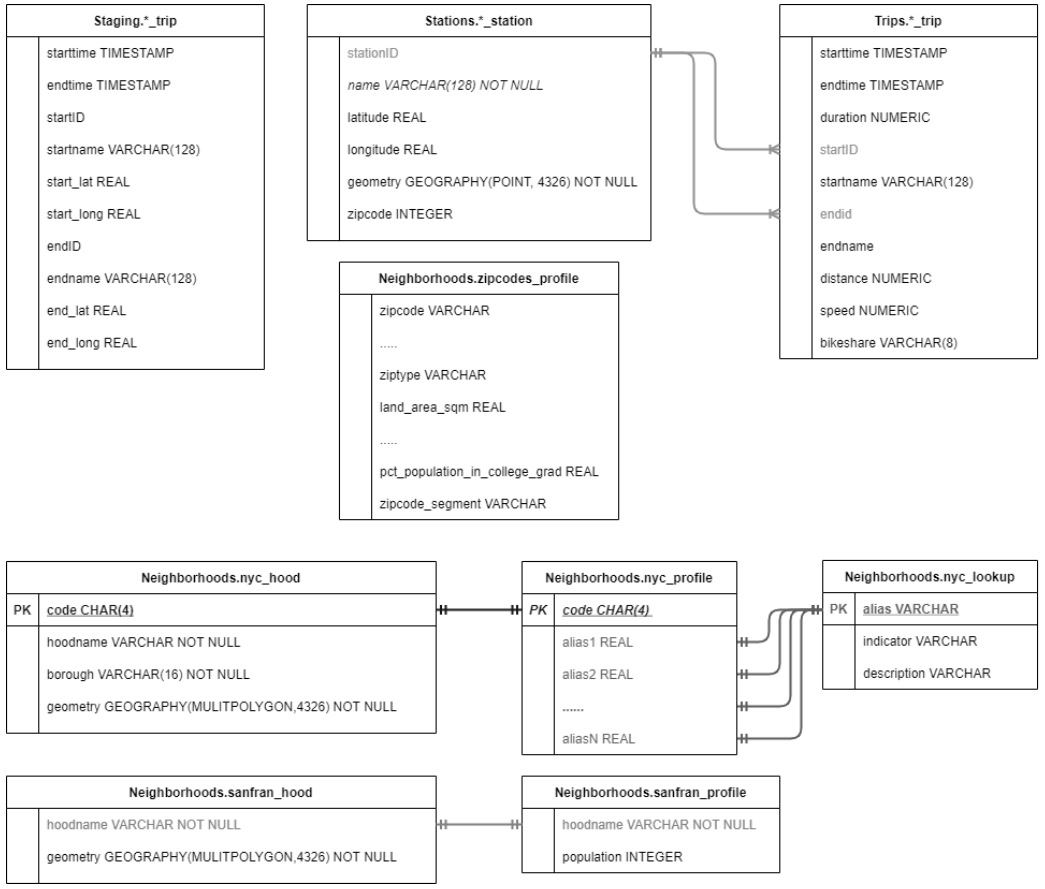
68+ GB of Data

350+ Files

4 File Formats

Leveraging the Cloud

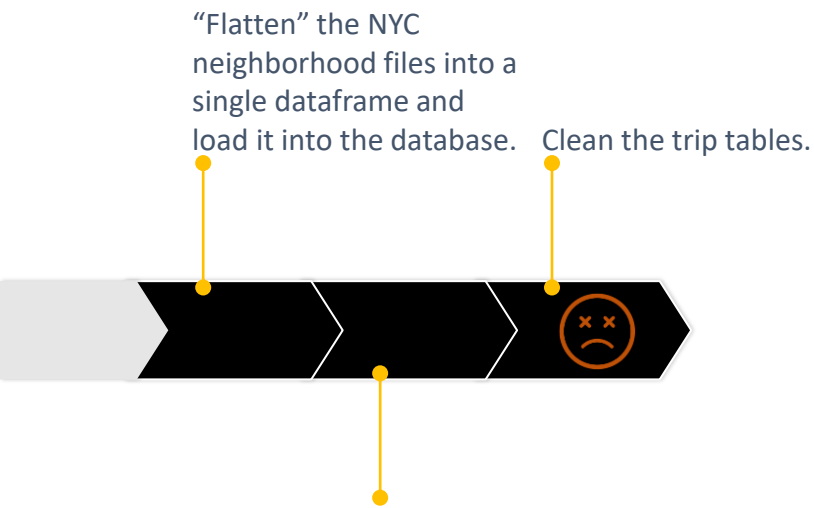
With those statistics, Pandas alone was not cut out for this job. Using RDS, a PostgreSQL database was created. The Entity Relationship Diagram is shown to the right.



The Data Engineering Timeline

1. Scan the S3 bucket for the relevant links to the trip data and send the request.
2. Unzip, extract, and save the relevant files locally.
3. Scrape the Furman Center site for all the neighborhood excel files and save them locally.
4. Bulk upload all the files to a personal S3 bucket.
5. Transform the trip files and load them into the staging table.
6. Use the staging tables to derive the station data.
7. Use the GCP Geocoding API to find the zip codes of stations and append to the station data.
8. Load the station data into the station tables.
9. Use the staging and station tables to create the trip tables.
10. Append the birthday and death certificates to the station tables.
11. Transform the geospatial datasets and load them into the database.
12. Scrape the San Francisco PDF file for data and load it into the database.

The Data Engineering Timeline

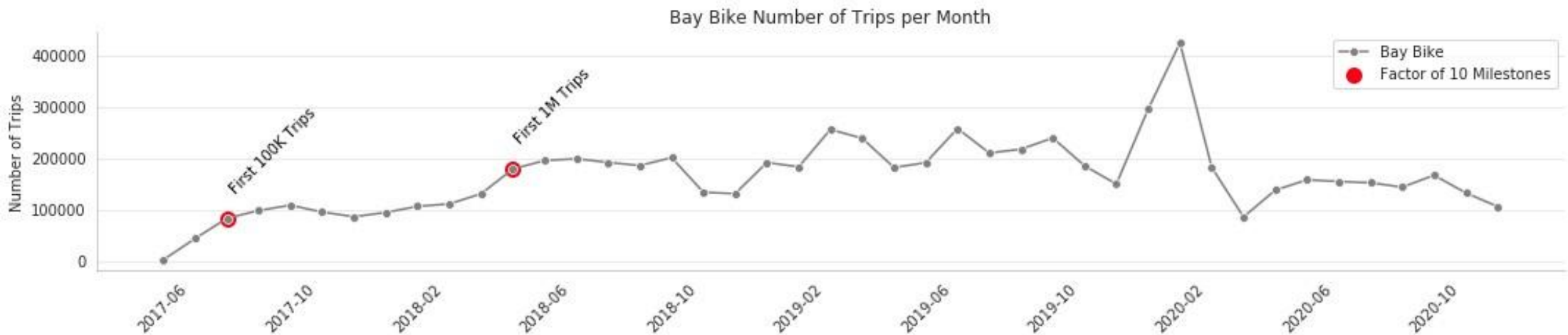


2.

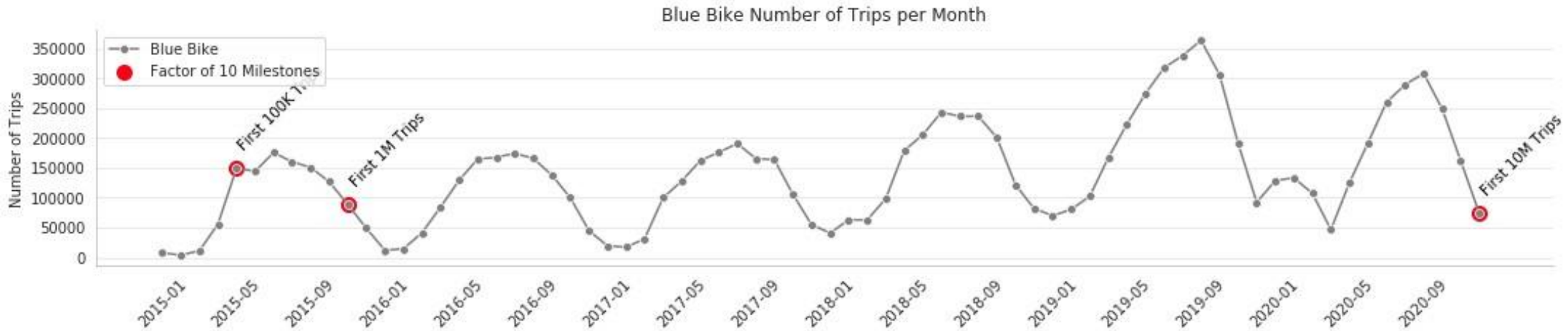
Exploratory Data Analytics

Getting familiar with the data.

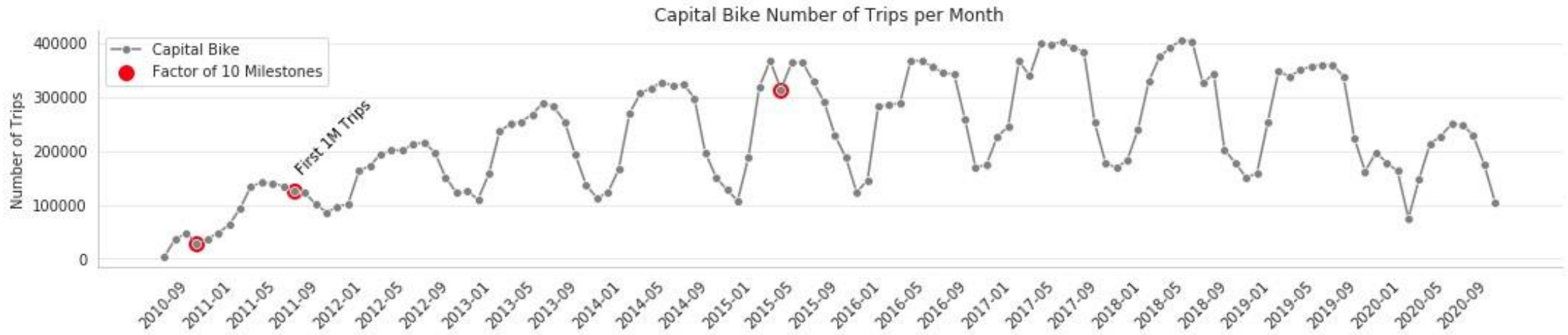
Number of Trips Taken Per Month?



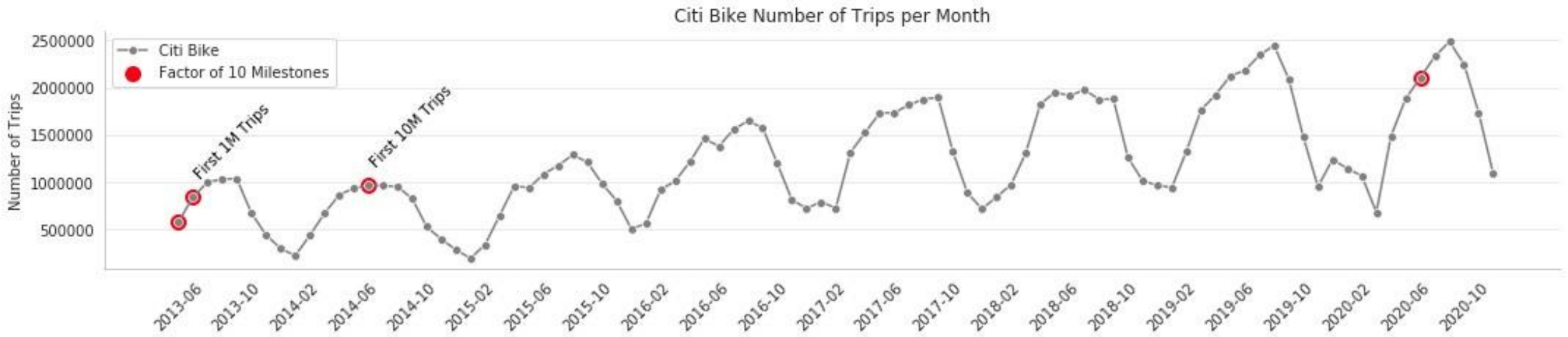
Bay Wheels had a massive spike in demand at the beginning of 2020.



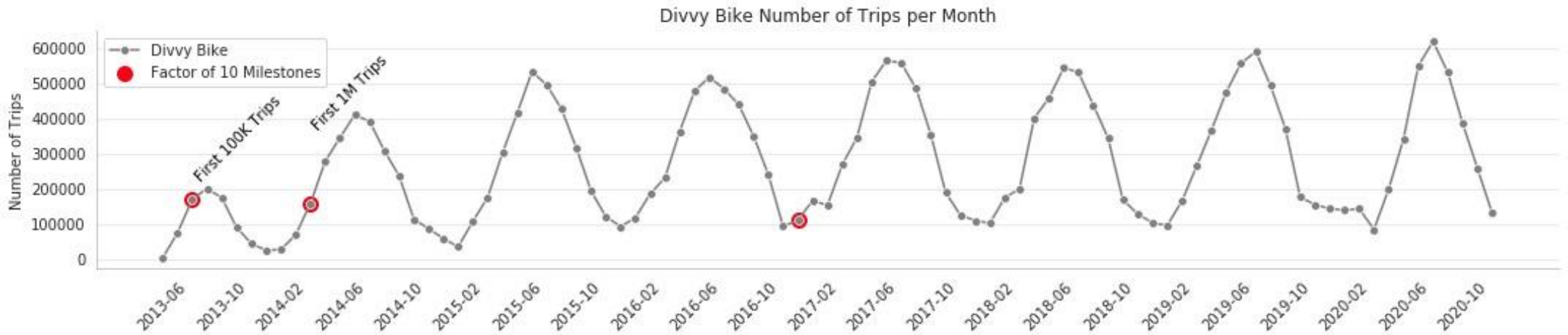
Blue Bike's demand has a predictable seasonality pattern.



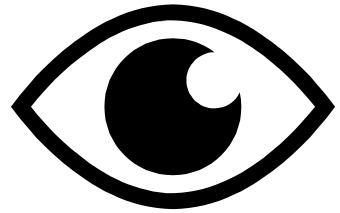
Although Capital Bikeshare does have seasonality it isn't as smooth as Blue Bike's.



Citi Bike accomplished 1M trips in the first months; 10M shortly after the first year.



Unlike the other services, Divvy's peaks and troughs are stable and don't show signs of growth.



Trips Per Month Summary

Seasonality

Except Bay Wheels, the number of trips rises and falls over the year as the seasons change. California doesn't have an intense winter season.

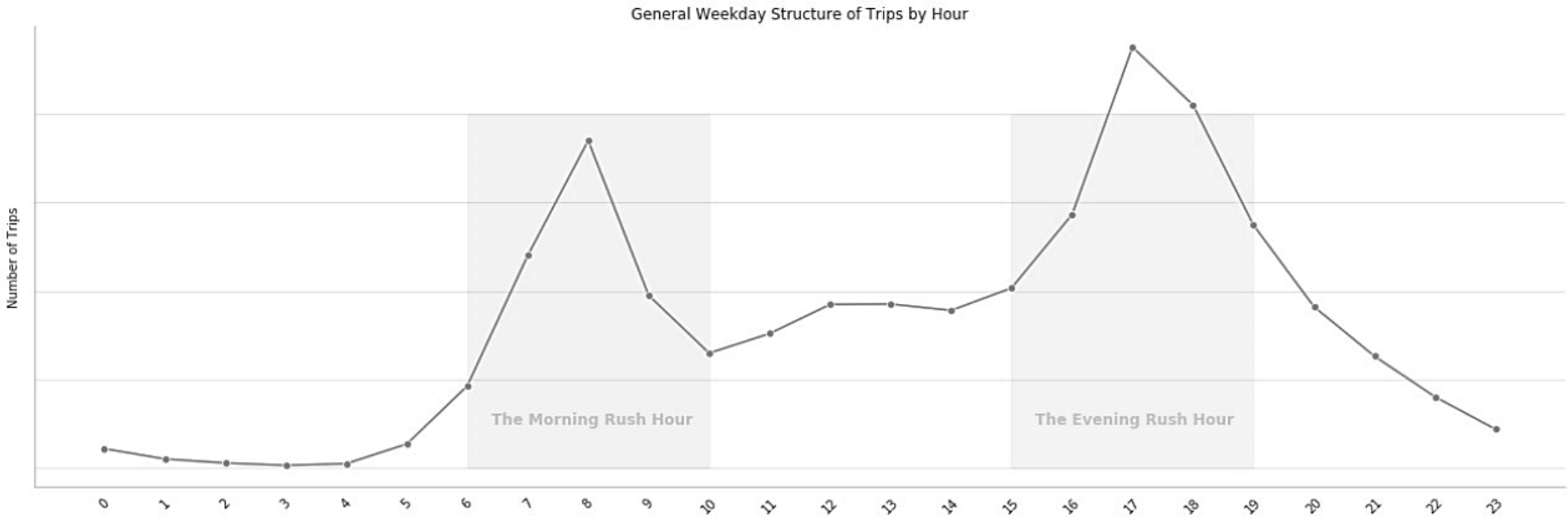
Increased Ridership

In line with the NACTO report, over time both peaks and troughs for the services that have seasonality pattern were gradually getting higher and higher indicating an increase in ridership.

The Lost Year

COVID-19 did influence the number of trips in 2020. In the first months of 2020 the number of trips had a massive drop off going into April instead of following its usual behavior of a steady rise.

Number of Trips Taken Per Hour?



During the weekdays, in general, this is the pattern for the number of trips taken per hour.

The Four Phases

Morning Rush

Trips increase starting at 06:00 and reach its peak during 08:00.

Stability

Trips are stable from 09:00 to 15:00

A secondary volume increase begins at 16:00 and hits its peak around 18:00

Evening Rush

After the evening rush, the trips taper off into the night and early morning

Taper Off

Absent of the sharp rises that are associated with rush hour; on the weekends, there is a gradual rise and fall.

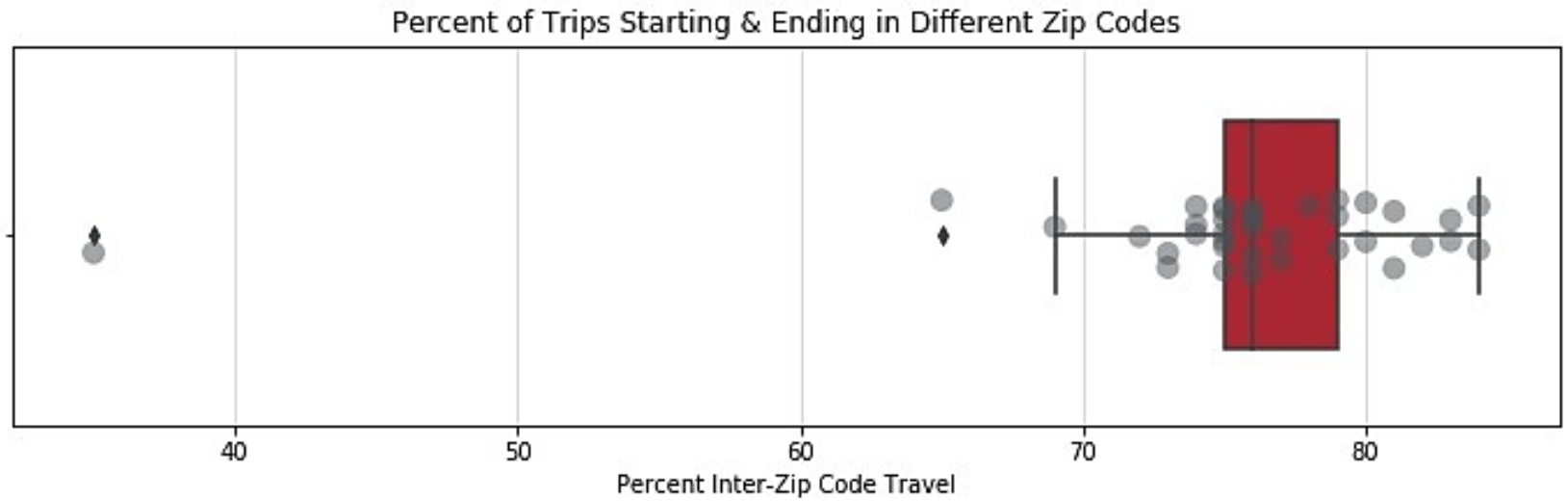
Number of Stations Added Every Year?

	Bay Added	Total	Blue Added	Total	Capital Added	Total	Citi Added	Total	Divvy Added	Total
2010	-	-	-	-	**106**	106	-	-	-	-
2011	-	-	-	-	38	144	-	-	-	-
2012	-	-	-	-	50	194	-	-	-	-
2013	-	-	-	-	111	305	**335**	335	**300**	300
2014	-	-	-	-	41	346	-	335	-	300
2015	-	-	**154**	154	11	357	144	479	175	475
2016	-	-	33	187	77	434	145	624	105	580
2017	**269**	269	11	198	52	486	153	777	5	585
2018	55	324	71	269	42	528	11	788	18	603
2019	100	424	63	332	52	580	111	899	9	612
2020	27	**451**	-11	**321**	30	**610**	275	**1174**	21	**633**

New York City has embraced the bike sharing economy with open arms.

Inter-Zip Code Travel

For An Expansion Into A New Area, How Important Is It To Expand Into Multiple Zip Codes?



How important? **Very**. The overwhelming majority of trips start and end in different zip codes.

	Launch Year	Initial Zip Codes	2020 Zip Codes
Bay Wheels	2017	**34**	47
Blue Bike	2015	31	**46**
Capital Bike	2010	24	**83**
Citi Bike	2013	**34**	78
Divvy Bike	2013	**28**	50

The smallest launch, over a decade ago, was spread across 24 zip codes. In 2020, the ecosystems easily cover 50+ zip codes.

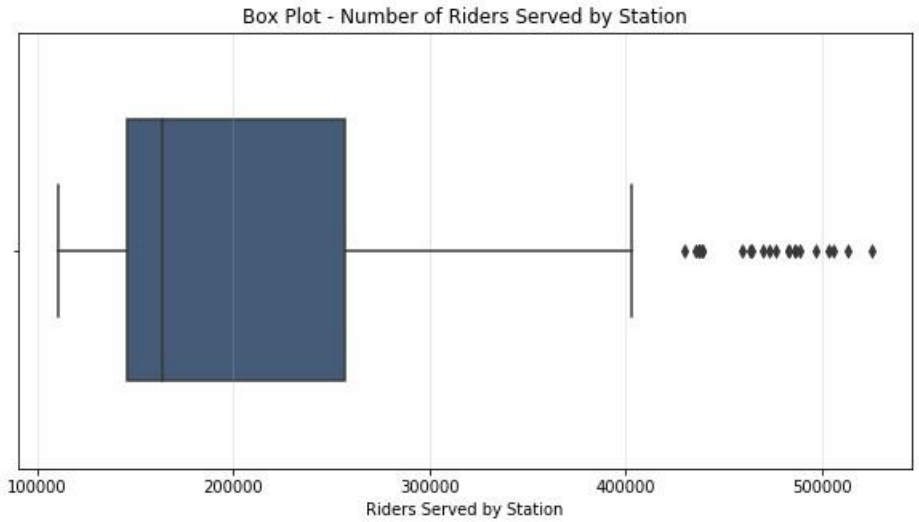
How Many People Does Each Station Serve?

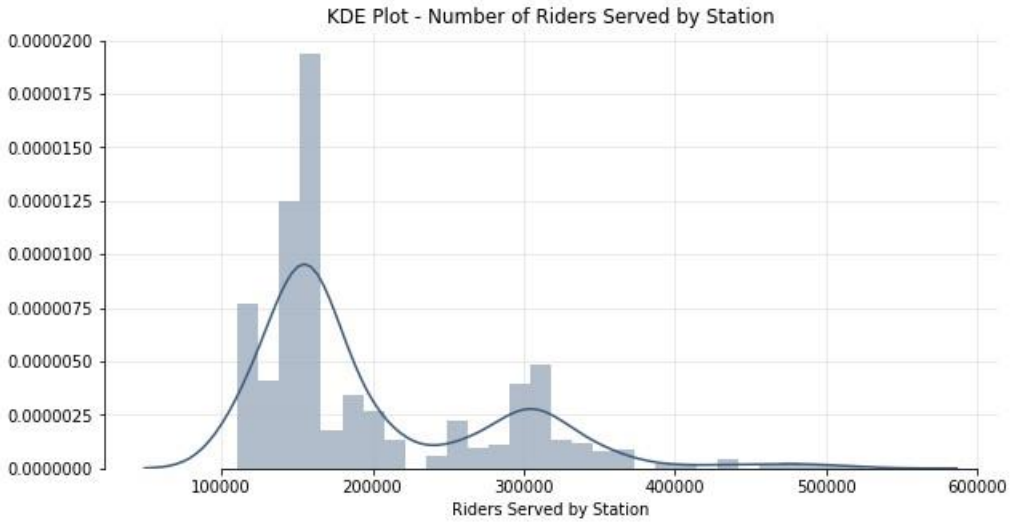
A Station Serves the People that are Closer to It than to Any Other Station

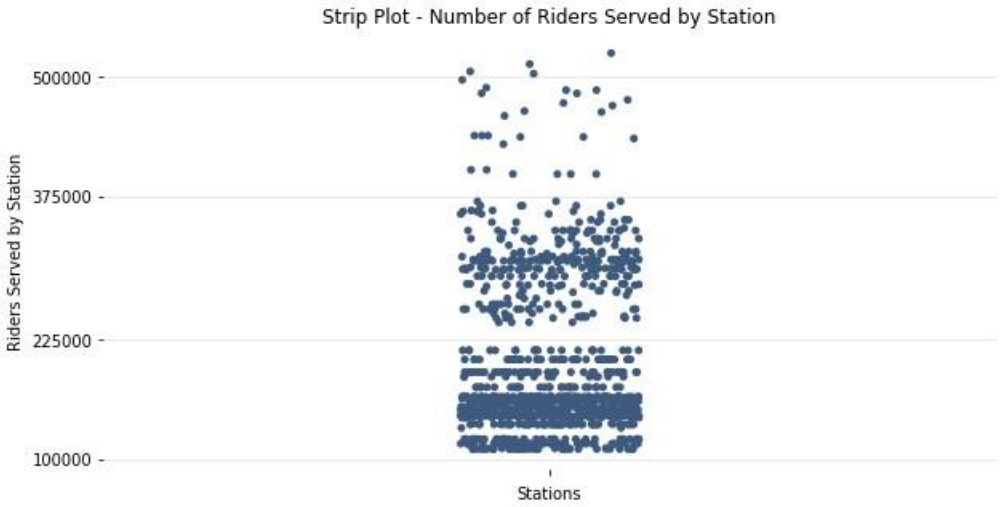
By using Voronoi Diagrams all the points (people) closer to one station than any other station can be determined **for all the stations.**

By **multiplying** the **portion** of a station's Voronoi polygon that is within a neighborhood by the **population density** of the neighborhood you can get the number of people that the station serves in the neighborhood.

Repeat For Every Neighborhood = Total People Served





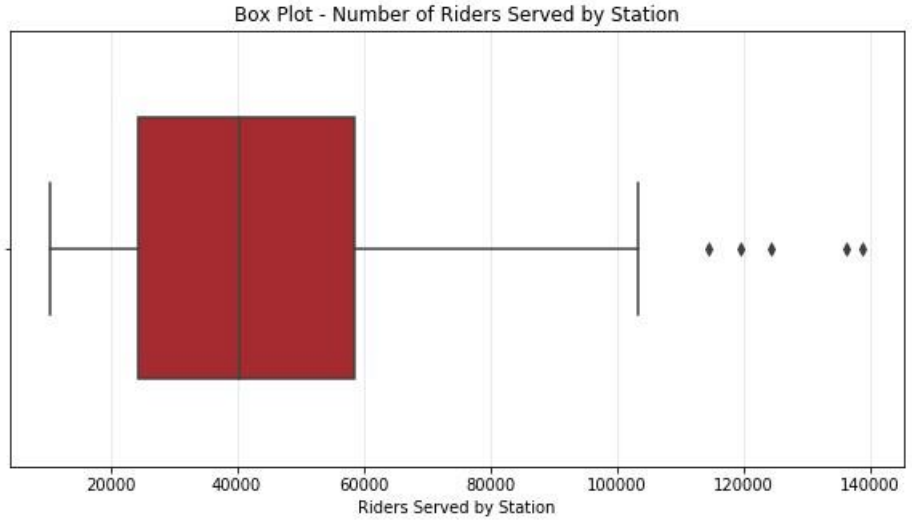


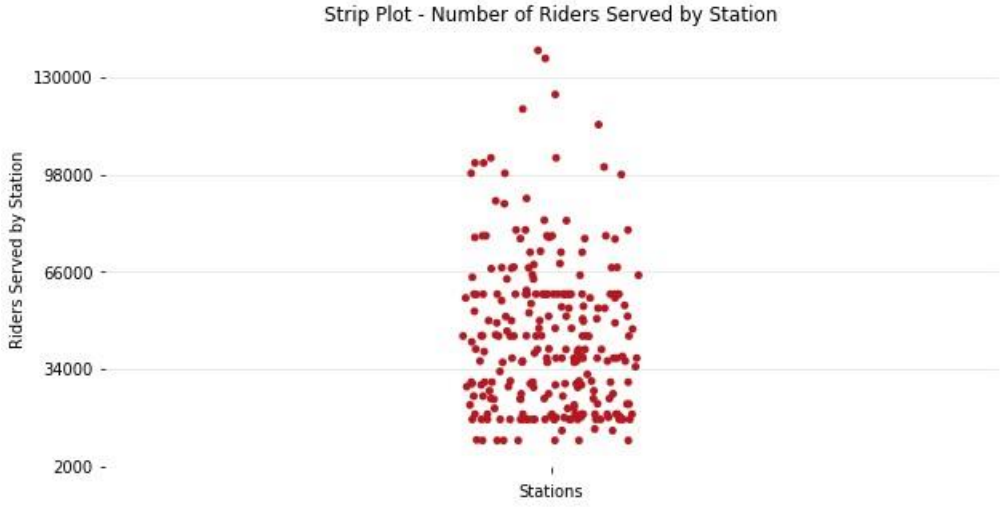
NYC People Served

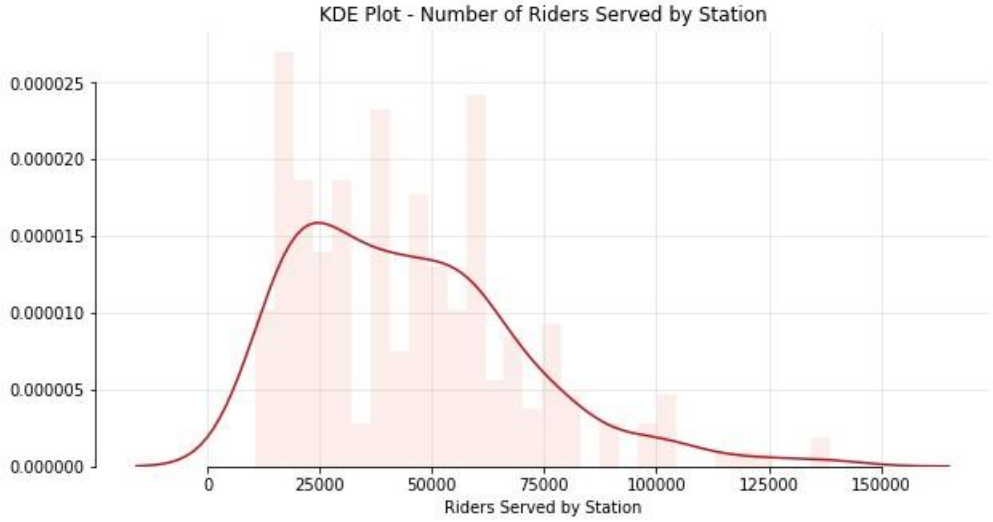
About 75% of stations serve between 100K and 225K people. There is a smaller group that serves 225K to 350K people.

SF People Served

About 75% of stations serve between 10K and 60K people.

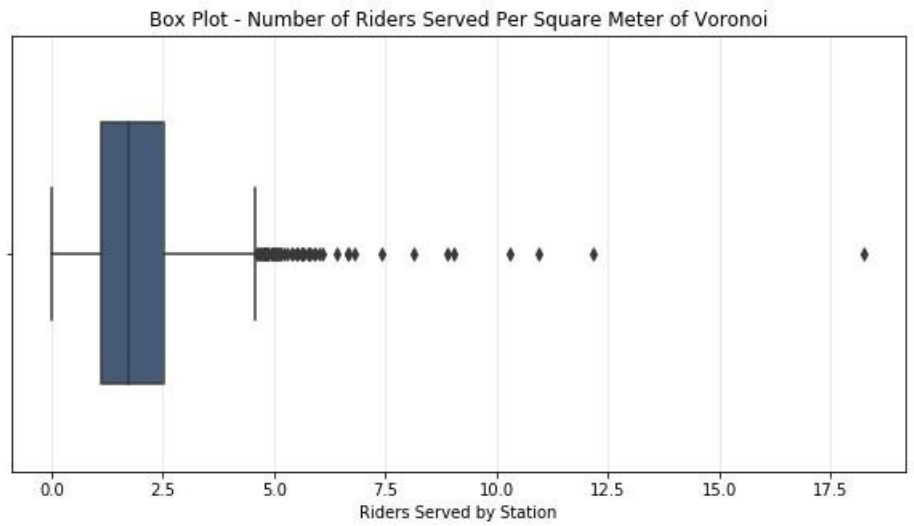


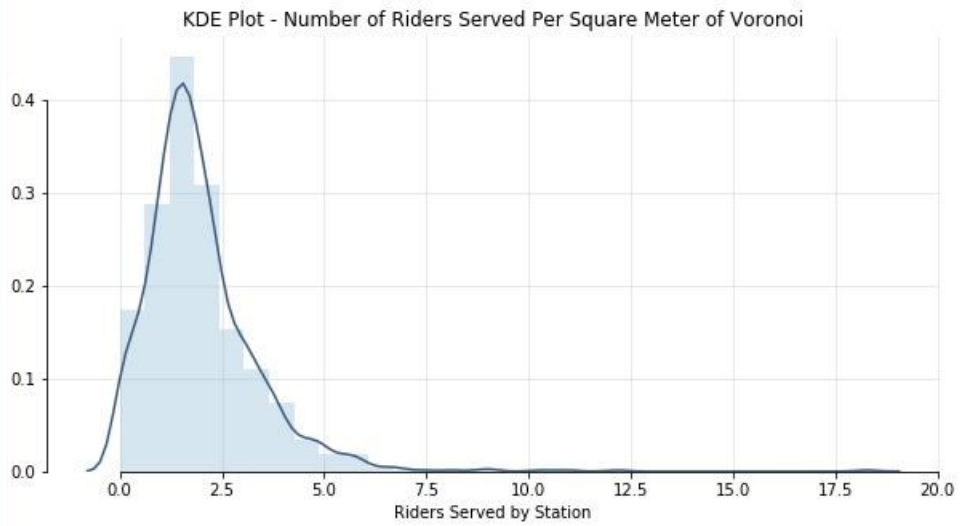


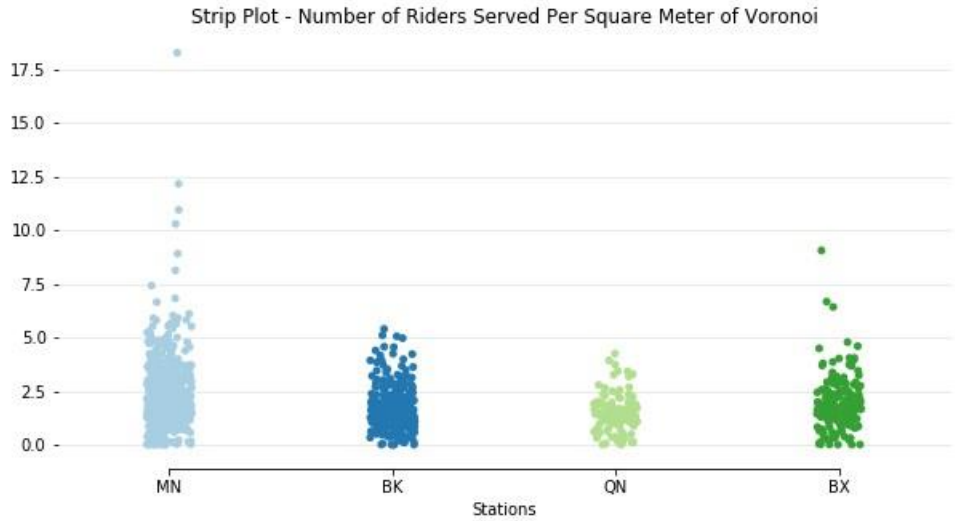


Do stations with higher people served metrics have a bigger Voronoi Area or are they in denser parts of the city?

It’s impossible to tell with just the metric alone. Let’s look at the **ratio** between the people served and the Voronoi Area.





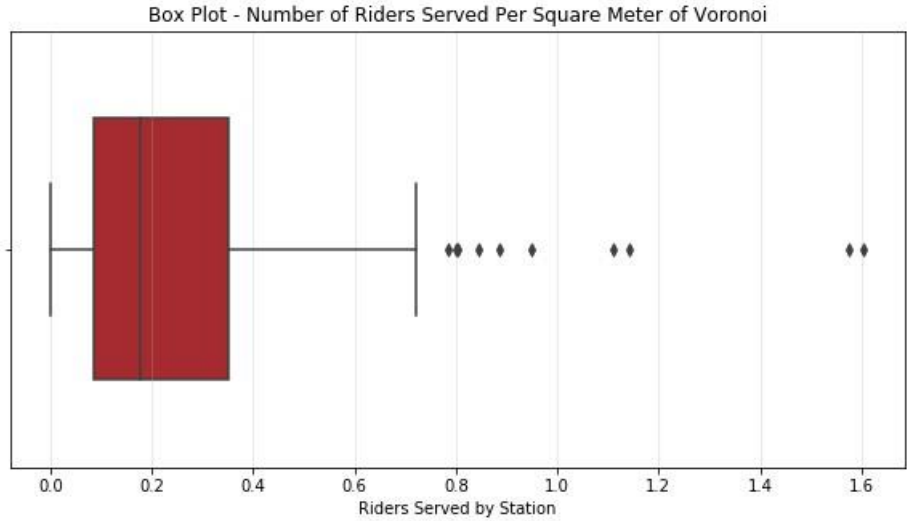


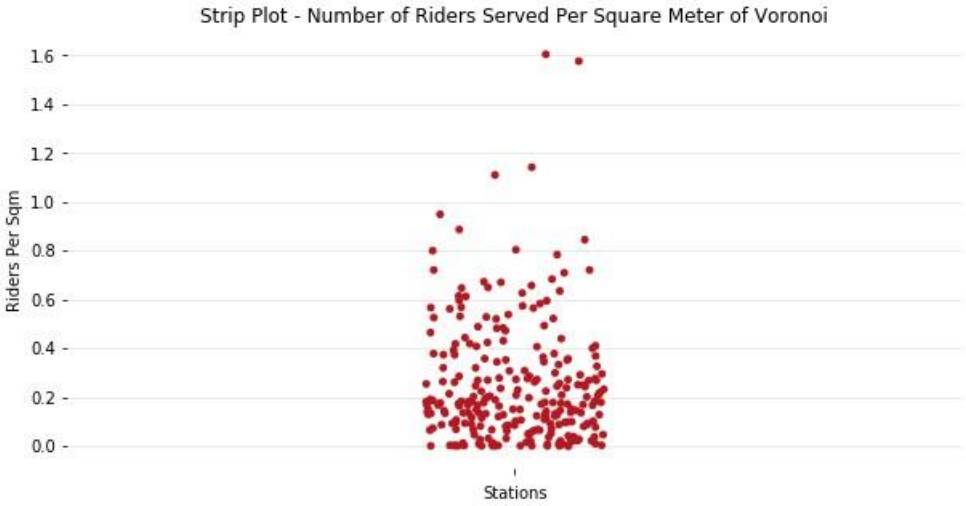
NYC People Served Ratio

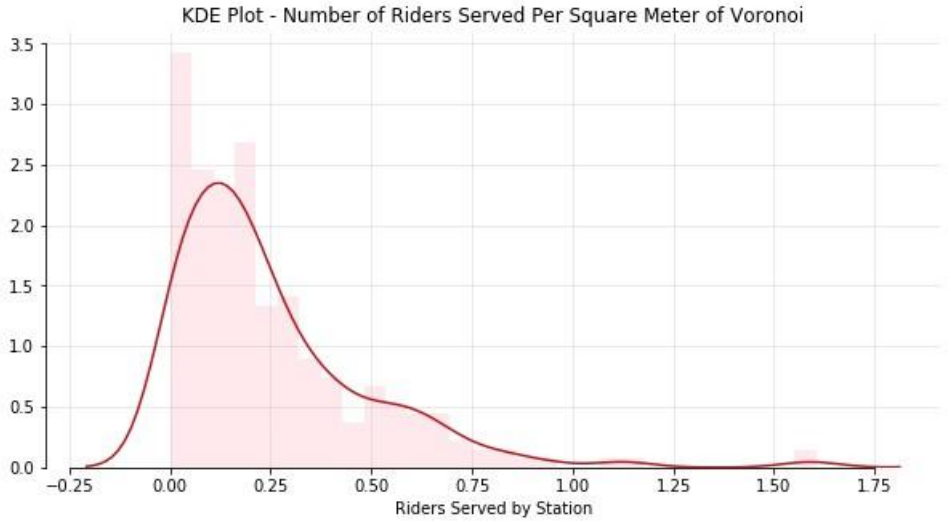
The data is much tighter when looking at the ratio between the people served and the area of the Voronoi. Regardless of the borough, regardless of the location, the number of people served by a station is rarely over 3.5 people per square meter of it’s Voronoi area.

SF People Served Ratio

The number of people that a station serves is rarely over 0.5 people per square meter of it's Voronoi. NYC is bigger and denser than SF, so it makes sense that the ratios are smaller.







Would You Have Bike Access?

Randomly Pick 100 People in NYC or San Francisco.

Determine The Percentage of People that Live Within X Miles of **Any** Bike Station.

Repeat 100 Times – Take the Average Percentage.

		New York City				San Francisco	
		CitiBike		MTA Subway Stations		Bay Wheels	
		All Hoods	Hoods w/ Stations	All Hoods	Hoods w/ Stations	All Hoods	Hoods w/ Stations
0.5 (805m)	10-Min Walk	**41%**	**83%**	67%	85%	**73%**	**88%**
0.375 (604m)	7.5	39%	80%	58%	76%	65%	80%
0.25 (402m)	5	36%	76%	43%	59%	53%	65%
0.125 (201m)	2.5	28%	59%	17%	26%	28%	36%
0.05 (80m)	1	6%	15%	4%	6%	6%	7%

→ The Set of Stations Used in the Search

→ People Selected From...

CitiBike vs. MTA

When it comes to the subset of neighborhoods that have stations, CitiBike outperforms the subway stations. That isn't the case when all neighborhoods are used.

CitiBike Concentration

When all neighborhoods are included, the percent of people that have access to CitiBike is cut in half for all distances. This shows that the distribution of stations is concentrated in a small number of neighborhoods.

BayWheels Concentration

For BayWheels the drop when all neighborhoods is included isn't as big as Citi's. They are always ~10% away from each other.

		New York City				San Francisco	
		CitiBike		MTA Subway Stations		Bay Wheels	
		All Hoods	Hoods w/ Stations	All Hoods	Hoods w/ Stations	All Hoods	Hoods w/ Stations
0.5 (805m)	10-Min Walk	**41%**	**83%**	67%	85%	**73%**	**88%**
0.375 (604m)	7.5	39%	80%	58%	76%	65%	80%
0.25 (402m)	5	36%	76%	43%	59%	53%	65%
0.125 (201m)	2.5	28%	59%	17%	26%	28%	36%
0.05 (80m)	1	6%	15%	4%	6%	6%	7%

→ The Set of Stations Used in the Search

→ People Selected From...

Lower Walking Distances

For neighborhoods that have stations, at the lower distances CitiBike is the most accessible service.

Goal to Aim For

When looking to expand into a new area a good goal to shoot for is to distribute stations in a way that allows ~80% of people to reach a station within 10-minutes.

Close Quarters

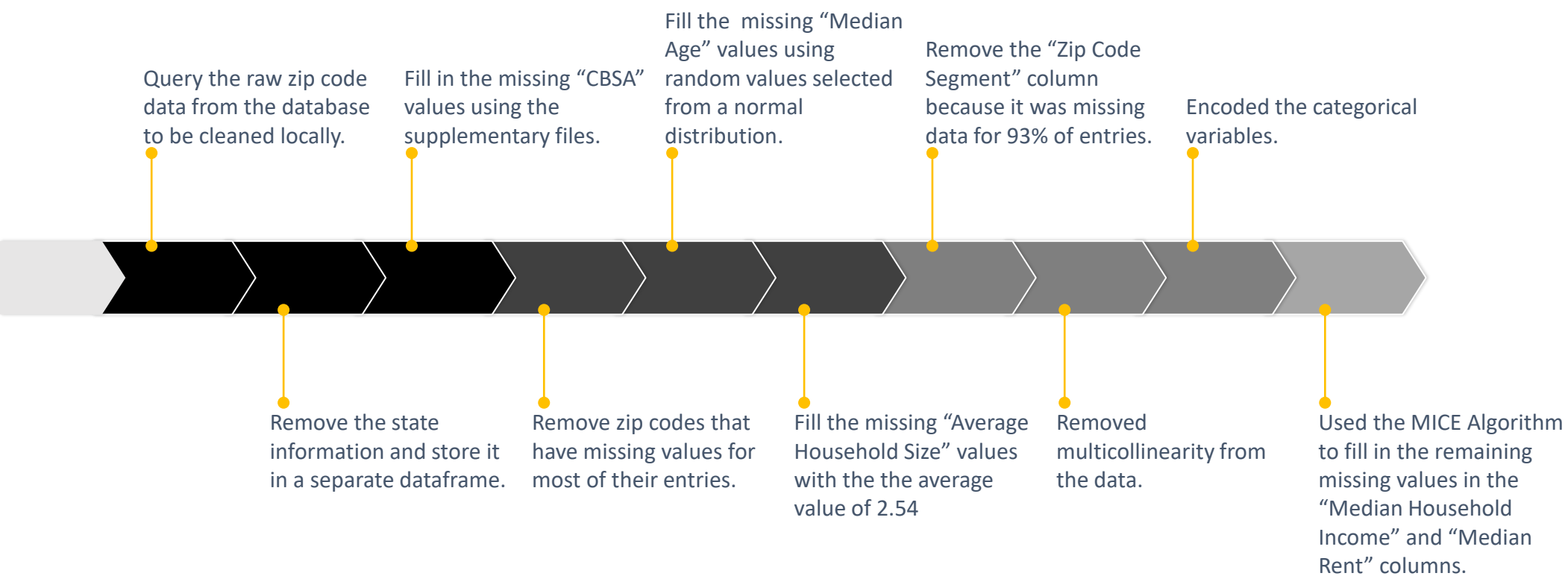
It's very rare to live within a 1-minute walk from any transit service.

3.

Zip Code Station Predictions

Making predictions from the data

The Zip Code Data Cleaning Timeline



Mean Shift Clustering

Clustering I

After cleaning, Scikit's Mean Shift clustering algorithm was used to cluster ALL the zip codes into groups. This clustering was then used as a feature to train the model. The clustering produced 156 clusters with a silhouette score of 0.57

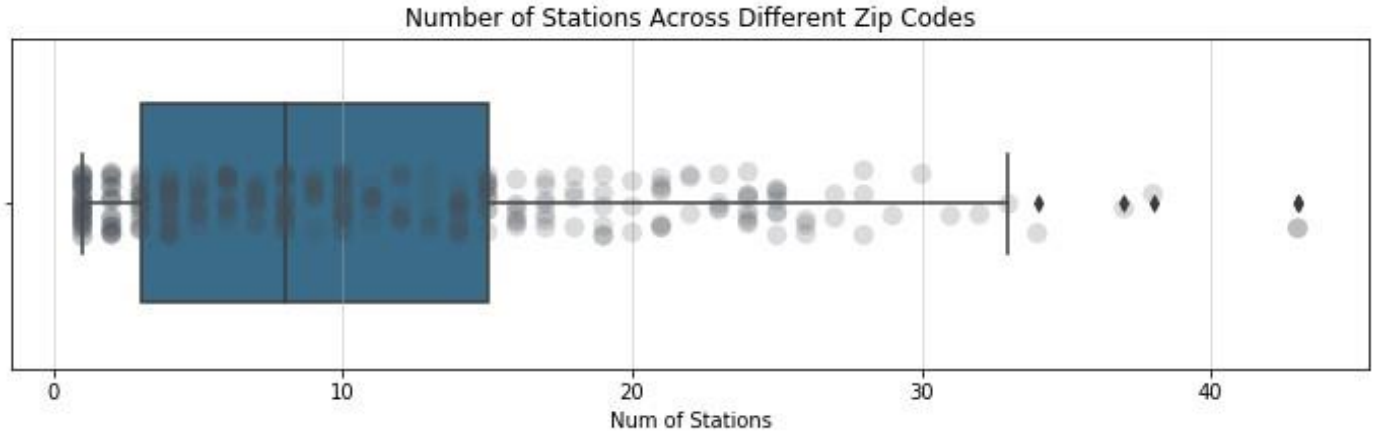
Training Data Partition

The number of bike stations in each zip code was queried from the database and merged to the zip code data. The zip codes that **have** bike stations were filtered out.

That partition is what the model is trained on, the no-station zip codes is what gets fed into the pipeline.

Clustering II

After the partition, another Mean Shift was run only on the partition and added as a feature. Unlike the 1st one, this Mean Shift is part of the pipeline and not an already known feature.



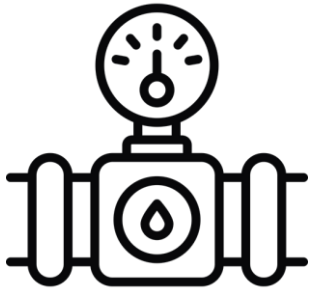
Most zip codes that do have stations, have between 3-15 stations

Regression Models

Leveraging AWS Sagemaker, all five models were tuned using TuneSearchCV w/ Hyperopt Search Optimization and RepeatedKFold

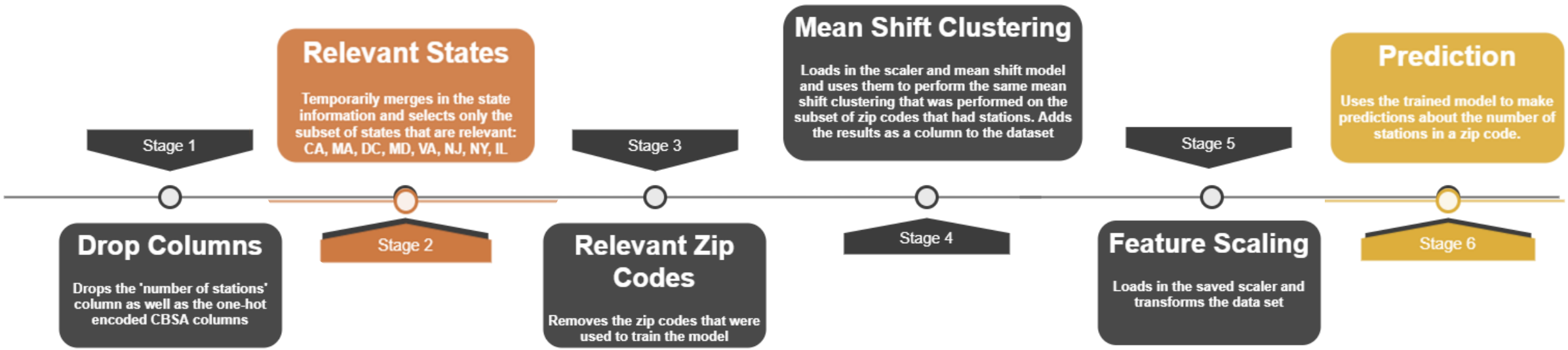
	Training Set			Testing Set		
	Root Mean Square Error	Maximum Error	**Median Absolute Error**	**Root Mean Square Error**	Maximum Error	**Median Absolute Error**
Baseline Dummy	-	-	-	**9.64**	**35.0**	**6.0**
Linear Regression	5.5	18.96	3.39	7.58	29.75	4.08
Ridge Regression	5.5	19.09	3.4	**7.56**	**29.8**	**4.06**
Support Vector Machines	5.7	20.6	3.03	**7.62**	**31.39**	**4.14**
Random Forest	7.03	24.18	5.24	8.29	31.85	5.43
Gradient Boosting	4.42	20.07	0.24	8.89	32.35	5.03

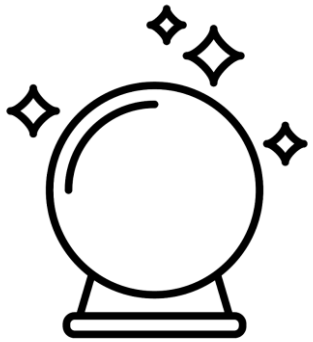
The model selected was the Ridge Regression Model. The support vector model was also used as a secondary prediction.



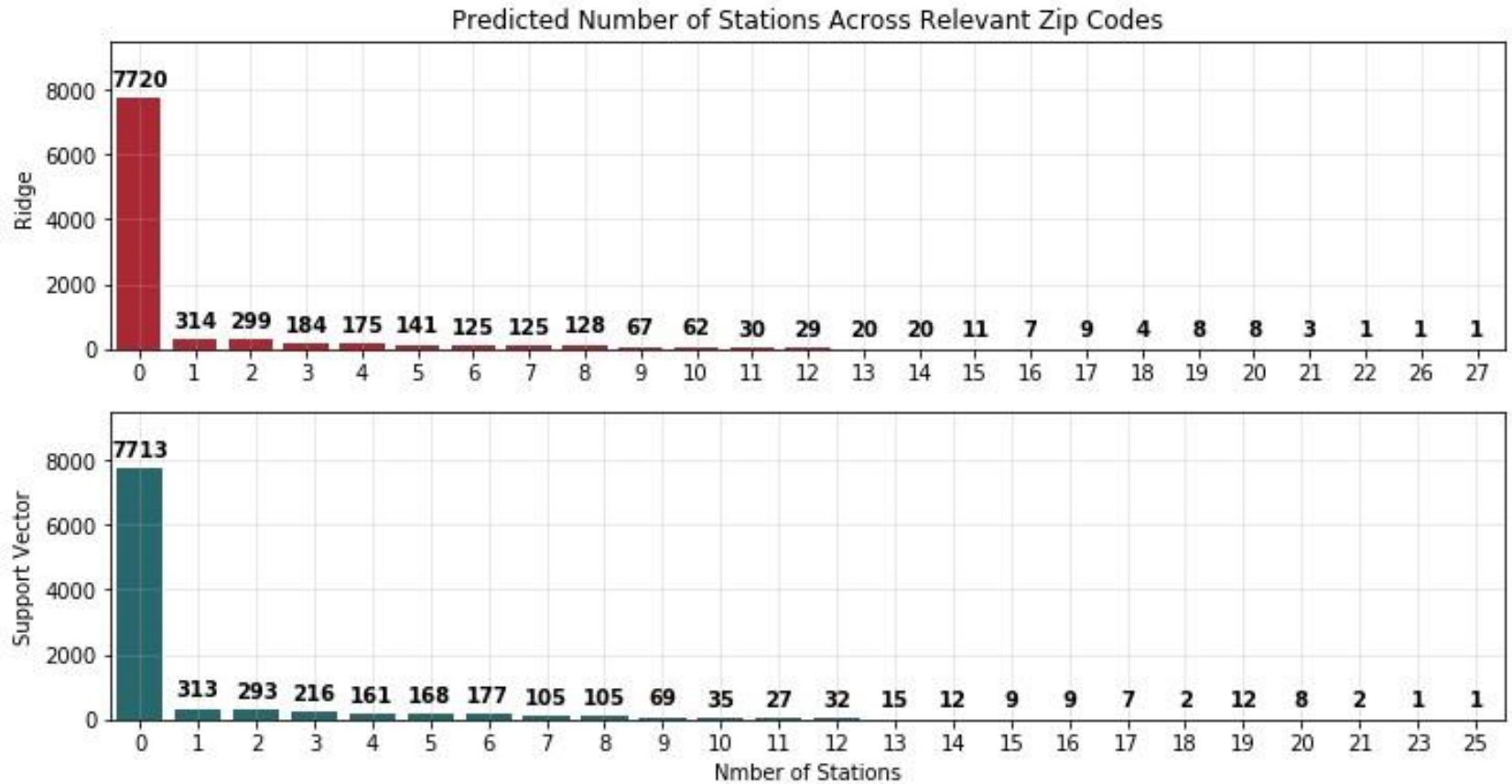
The Pipeline

The data that gets fed into the pipeline is the original zip code data with all the “all-zip code” mean shift clustering *(clustering I).*





The Prediction



The models predicted very similar results. Both mainly predicting 0 stations for most zip codes and had maximum predictions of about 25.