Citi Bike Ridership Analysis

Jersey City, NJ

Mark DeTiberiis & Jamil Mirabito

Objectives

- Introduction
  - Business case
- Methods and Data
- Findings
- Summary of the Findings
- Limitations and Considerations for future work

Introduction

Citi Bike is a Lyft-owned and operated bike-share service with over 1,000 Citi Bike stations located in Manhattan, Brooklyn, Queens, the Bronx and Jersey City. The service has experienced tremendous growth as people increasingly seek alternative methods for public transportation.

Business case: In determining where to expand their services, Lyft would like to better understand the factors that are most highly correlated with high activity at any given Citi Bike Station.

Methods & Data

For this analysis, we used an **Ordinary Least Squares (OLS) model** to explore the relationship between a number of community, environmental, and spatial variables and ridership in New Jersey.

We decided to use **only New Jersey stations** given the apparent large amounts of missing data from NYC bike stations. Of the nearly 850,000 entries in our dataset, fewer than 100 were from NYC bike stations.

For the sake of this analysis, **ridership** at a given station is defined as **one point of contact with a station** (i.e., if a renter starts or ends at a given location).

Data Sources:

- Citi Bike Trip History Data (08/01/2019 - 08/31/2020)
- Zip-Code-level Population Characteristics
- NOAA Weather Data (08/01/2019 - 08/01/2020)

Ordinary Least Squares (OLS) Regression

OLS Regression results show that the variables that seem to have the largest effect on daily ridership are:

1. **Temperature (+)**
2. **Spring (dummy variable) (-)**
3. **Median income (+)**
4. **Distance to the PATH train (-)**

However, with an R-squared value of 0.293, there is a lot of variance that we cannot account for with the variables we have included in our model.

OLS Regression Results

Dep. Variable:	rider_count	R-squared:	0.293
Model:	OLS	Adj. R-squared:	0.293
Method:	Least Squares	F-statistic:	639.2
Date:	Sun, 13 Sep 2020	Prob (F-statistic):	0.00
Time:	12:54:21	Log-Likelihood:	-24991.
No. Observations:	20064	AIC:	5.001e+04
Df Residuals:	20050	BIC:	5.012e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.082e-17	0.006	3.51e-15	1.000	-0.012	0.012
TAVG	0.2529	0.012	21.051	0.000	0.229	0.276
PRCP	-0.0980	0.006	-16.460	0.000	-0.110	-0.086
dist_PATH	-0.2107	0.008	-25.367	0.000	-0.227	-0.194
dist_landmark	0.1142	0.009	12.160	0.000	0.096	0.133
median_inc	0.2126	0.010	21.209	0.000	0.193	0.232
Sunday	-0.0285	0.006	-4.793	0.000	-0.040	-0.017
07304	-0.1177	0.010	-12.295	0.000	-0.136	-0.099
07305	-0.0609	0.007	-8.340	0.000	-0.075	-0.047
07306	-0.1632	0.010	-17.139	0.000	-0.182	-0.145
07307	-0.0211	0.008	-2.637	0.008	-0.037	-0.005
spring	-0.2190	0.008	-28.819	0.000	-0.234	-0.204
summer	-0.1228	0.010	-12.483	0.000	-0.142	-0.103
winter	-0.0679	0.010	-7.147	0.000	-0.087	-0.049

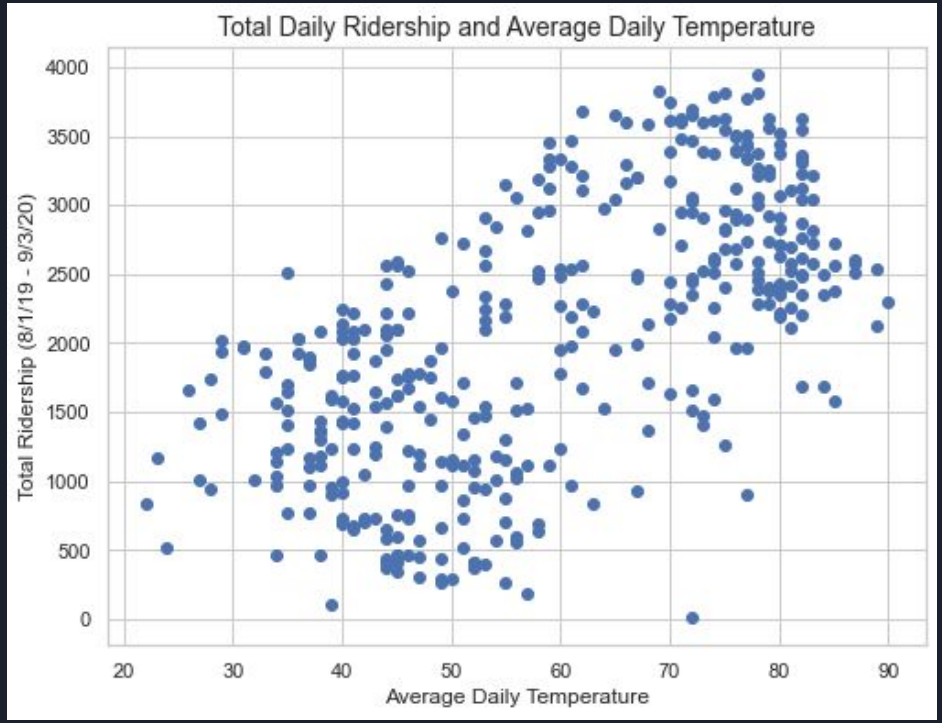
Omnibus:	19522.413	Durbin-Watson:	1.727
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1193679.708
Skew:	4.718	Prob(JB):	0.00
Kurtosis:	39.590	Cond. No.	3.85

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

There is a moderately strong correlation between ridership and temperature

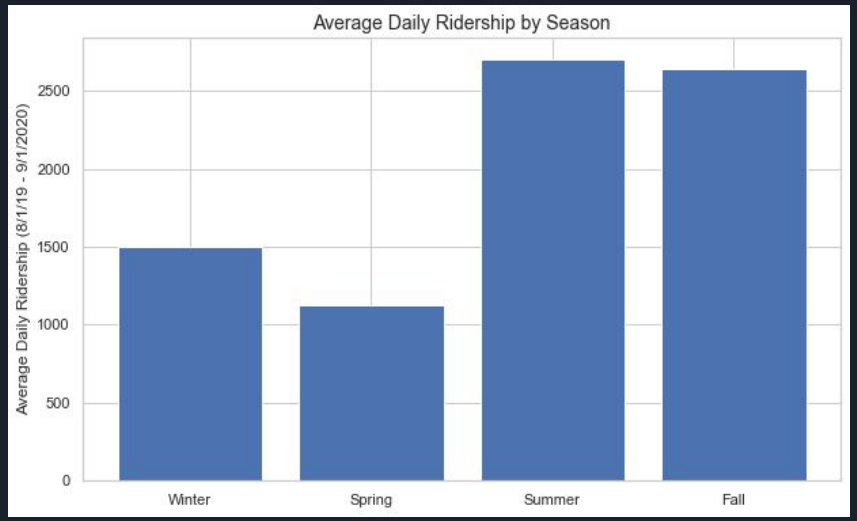
R = 0.629151
p-value = 0.000000

The high correlation coefficient of 0.63 would indicate that there seems to be a fairly strong correlation between average daily temperature and total daily ridership. **The correlation is statistically significant.**



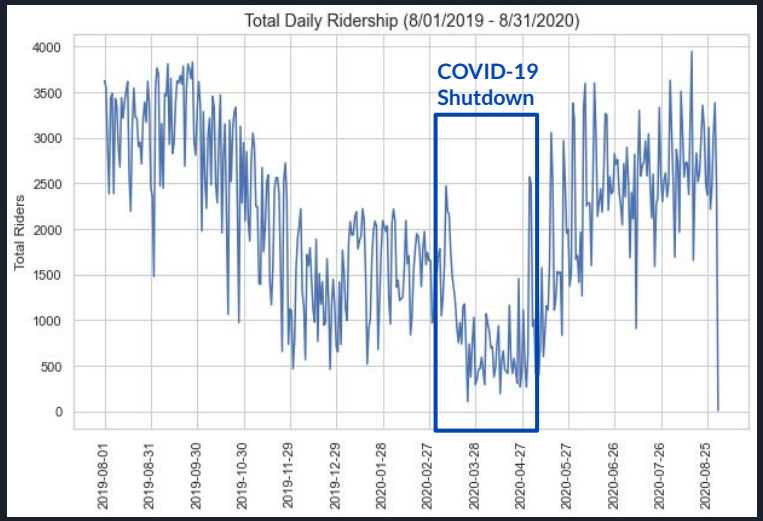
See slide 24 for a comparison of the temperature and daily ridership from 8/1/19 to 9/1/20.

COVID-19 significantly impacted ridership in the spring of 2020



F-statistic: 557.860 p-value: 0.000

We conducted an ANOVA test to determine if the observable differences in average daily ridership were statistically significant. **The ANOVA test confirmed that ridership is statistically different across seasons.**



COVID-19 resulted in a significant decrease in ridership at the start of March before returning to normal ridership near the start of June.

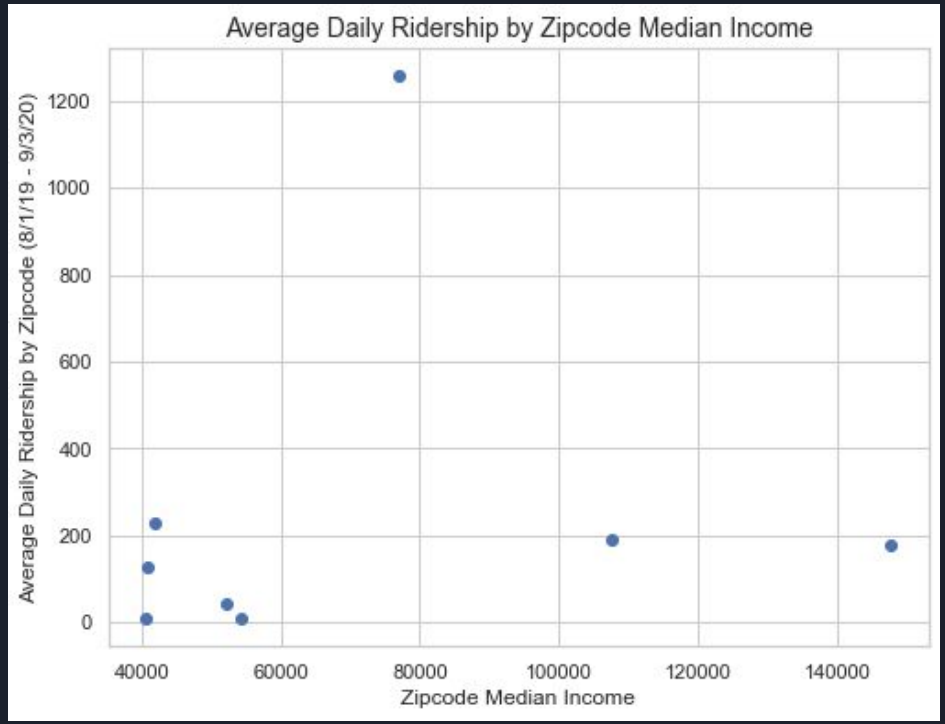
There seems to be a weak correlation between an area's median income and ridership

R: 0.157882
p-value: 0.708853

A correlation analysis between average daily ridership and zip code median income yields a **weak and insignificant correlation**.

Zip codes with lower median incomes also have fewer Citi Bike stations. Those with higher median incomes are located near many stations, but residents may prefer to use their own means of transportation.

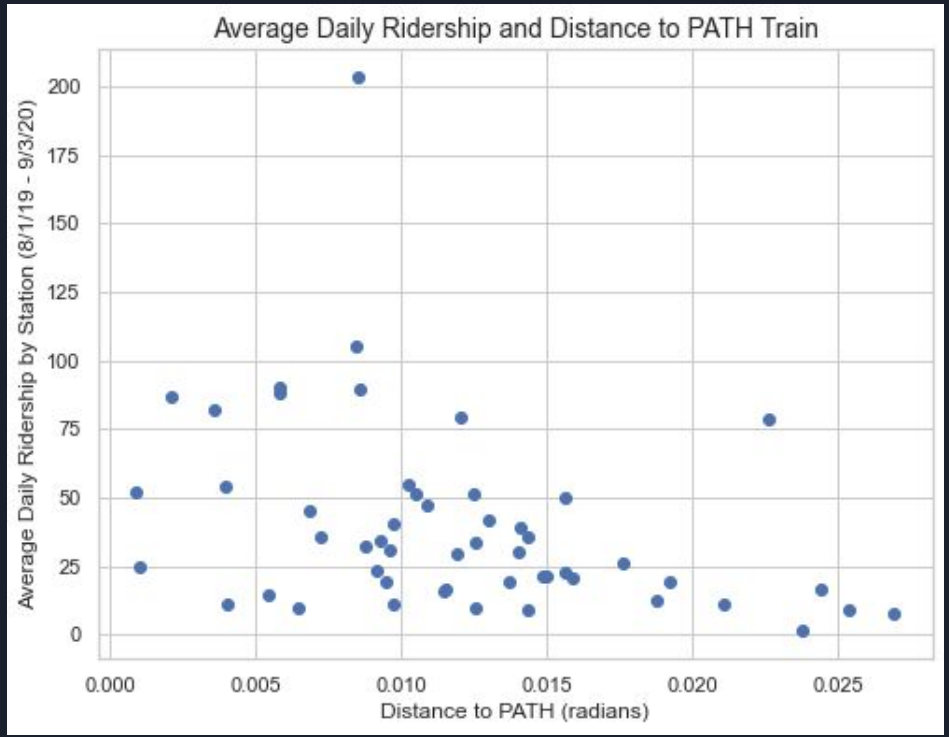
In our model, **it's likely that median income is acting as a proxy for zip code 07302.**



There is a fairly weak negative correlation between distance to NJ PATH and ridership

R: - 0.364822
p-value: 0.007834

Bike stations located closer to PATH train stations experience higher average daily activity than those further away. **This relationship is relatively weak but significant.**



Summary of findings

While our OLS model could not account for much of the variance in Citi Bike ridership, we were able to explore the relationship between the top four most predictive variables and Citi Bike ridership. Our findings from our exploratory analyses can be found below:

- There is a fairly **strong correlation (R = 0.629) between ridership and daily average temperature**. As temperature increases, more people tend to use Citi Bikes.
- **In our model, a spring dummy has a negative impact on ridership.** This is likely due to the effect of the COVID-19 Shelter in Place order beginning in March. An ANOVA test confirms that ridership differs across seasons.
- **There is a weak correlation between zip code median income and ridership.** Although, in the OLS model, it seems that median income is acting as a proxy for 07302 after removal from the model.
- **There seems to be a rather weak negative correlation between ridership and a bike station's proximity to an NJ PATH station.** This relationship, however, is statistically significant.

Limitations and Considerations for Future Analyses

Limitations:

- Our data was confined to New Jersey Citi Bike Stations due to a large amount missing data from NYC Citi Bike stations. Gathering more data from NYC could help to better understand how Citi Bike ridership varies in communities that are more diverse than just Jersey City.
- A zip code analysis may have been too large an area to draw any meaningful insights from differences in ridership by community area. A more granular analysis of city blocks could have helped to understand variations in ridership from one station to the next.

Considerations for future work:

- Including neighborhood-level crime statistics
- Creating a feature to capture the number of restaurants and bars that fall within a 100 foot radius of each bike station.

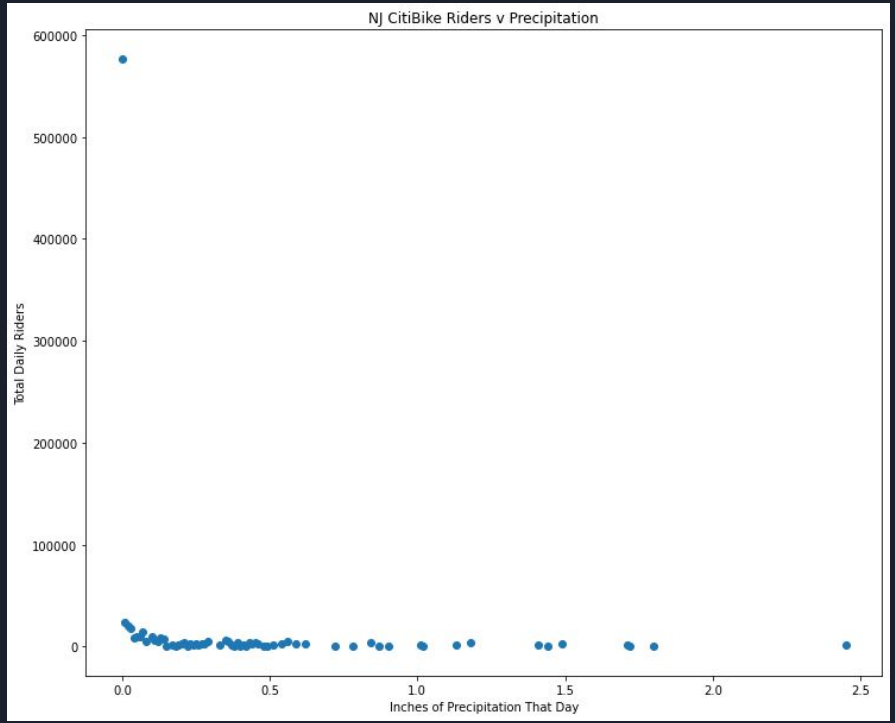
Questions?

Appendix

Citi Bike locations in Jersey City and the surrounding area

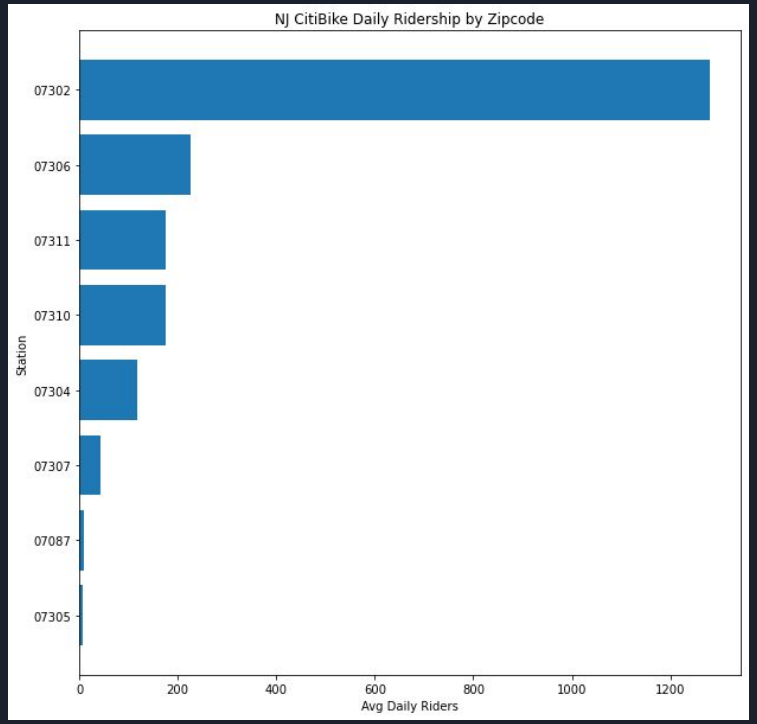
There are very few riders on days where there is any amount of rainfall

R = -0.286872

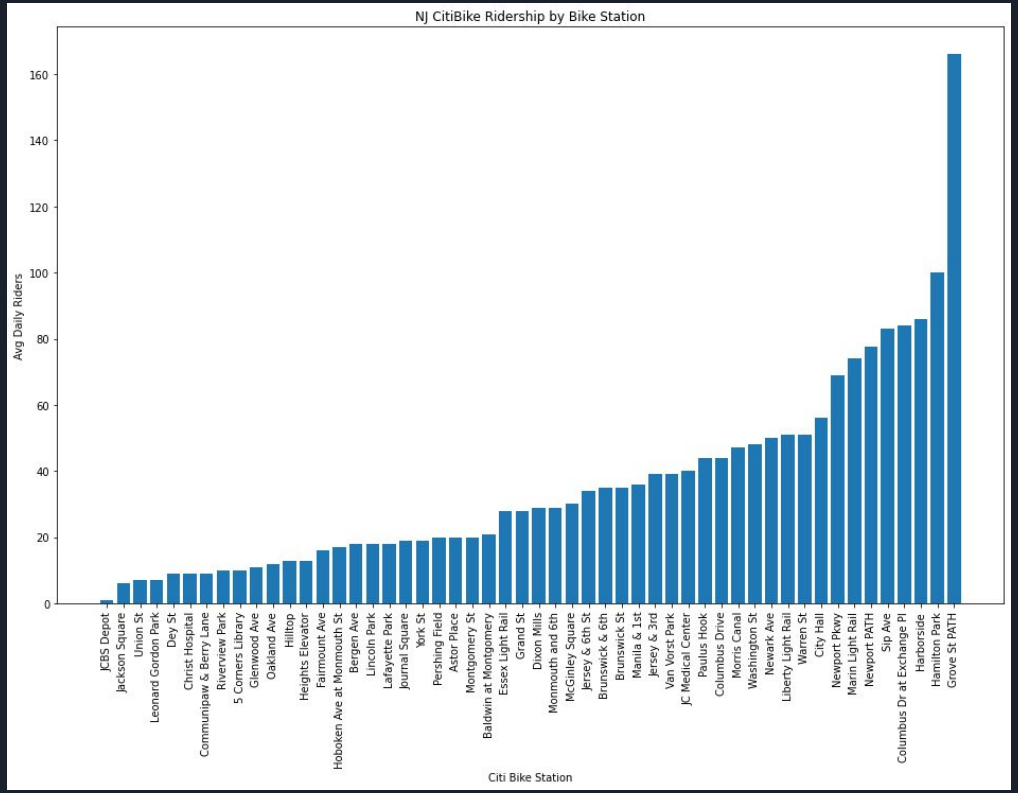


One Jersey City zip code contains the majority of Citi Bike stations

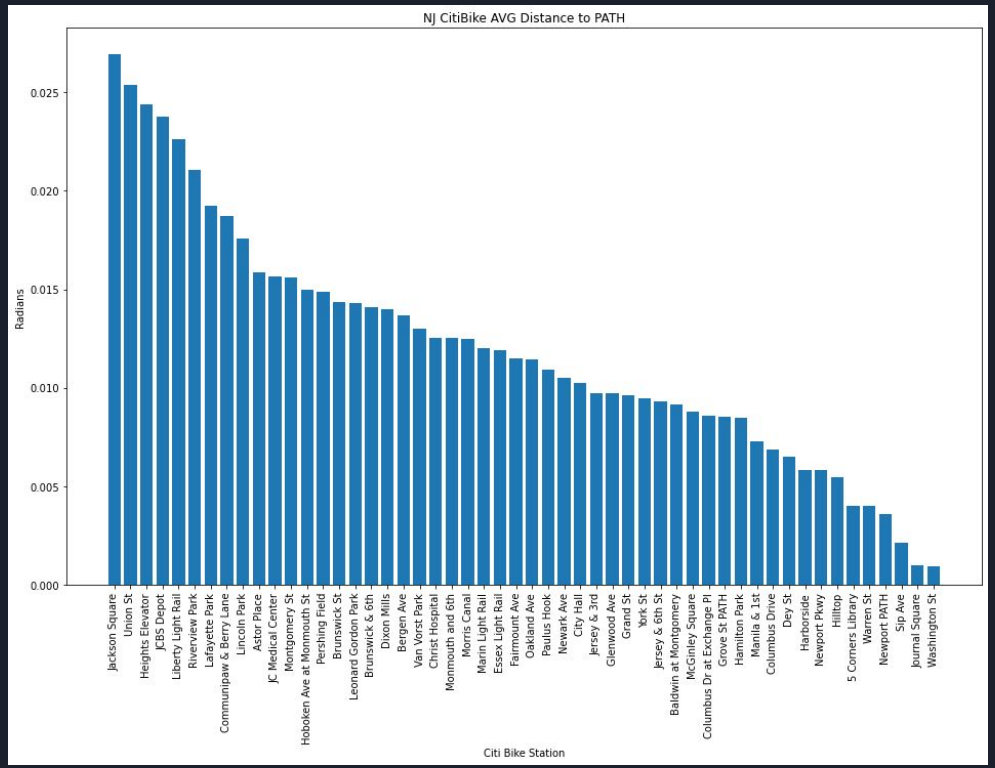
Average daily ridership is highest in the zip code that contains the most citi bike locations. It also contains the majority of landmarks, restaurants, and bars in the area.



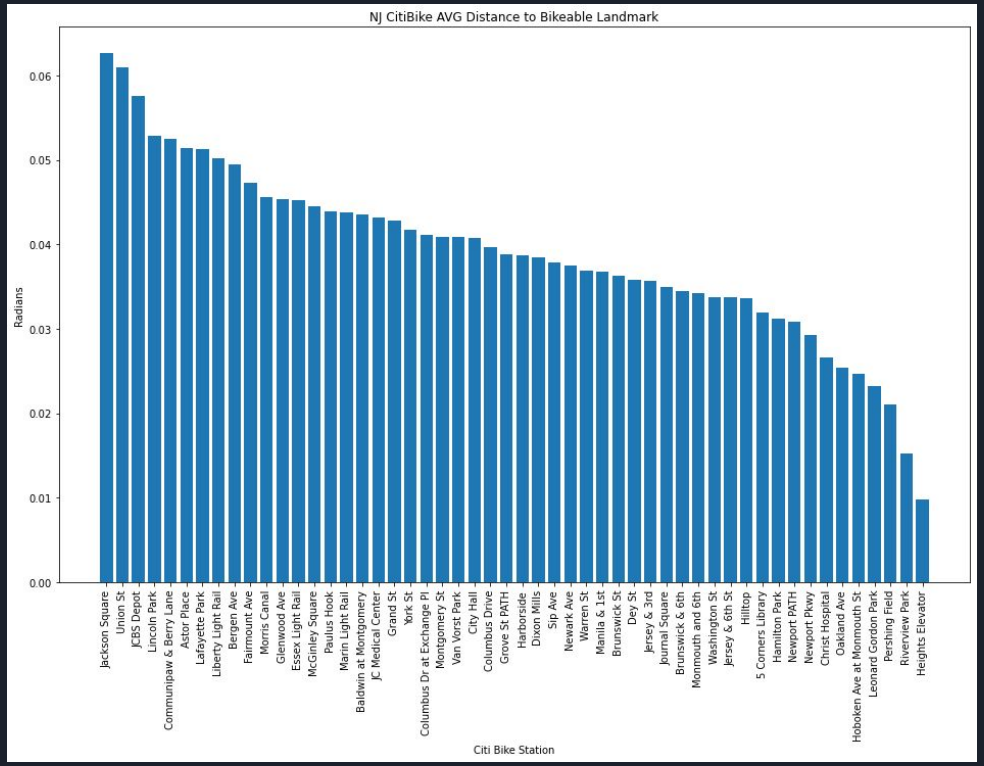
Bike stations located near PATH trains or the waterfront seem to attract more activity



A ranking of Citi Bike Stations by distance to PATH Trains

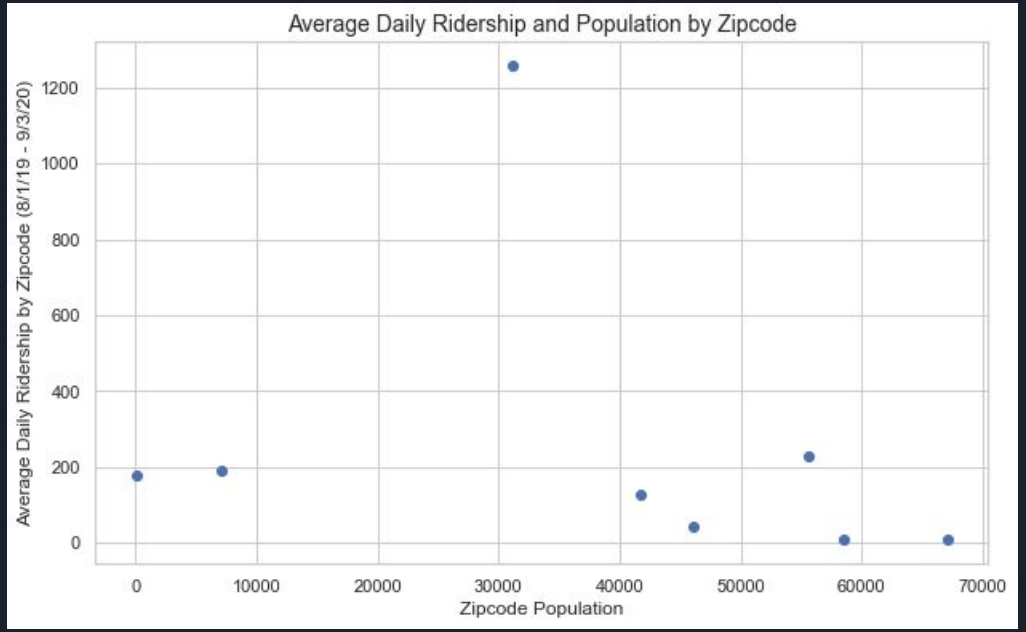


A ranking of Citi Bike stations by average distance to landmarks



There is a weak and insignificant relationship between area population and daily ridership

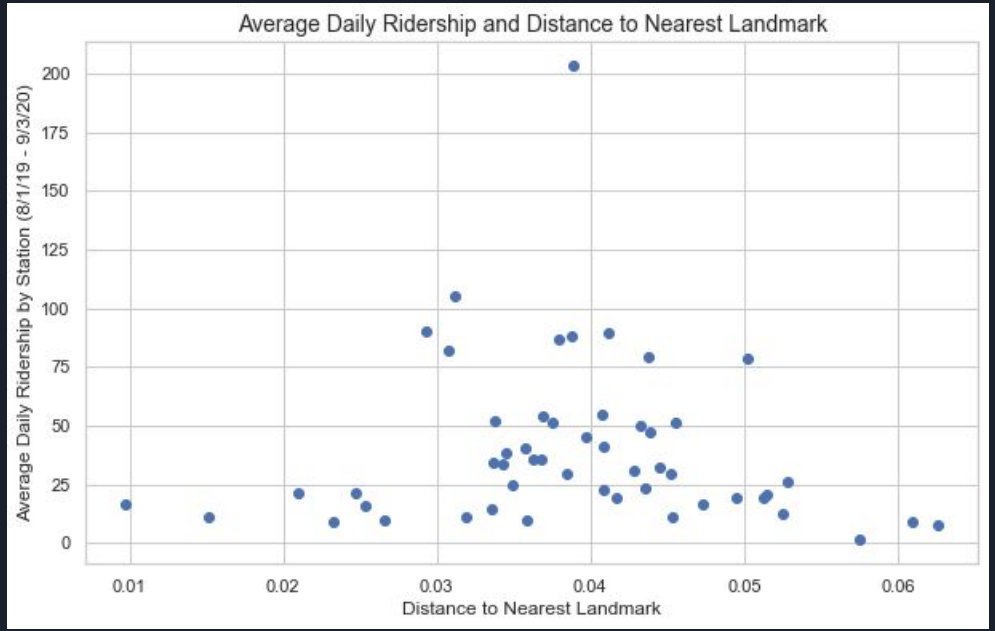
R: - 0.240424
p-value: 0.566276



There seems to be no relationship between average daily ridership and distance to a landmark

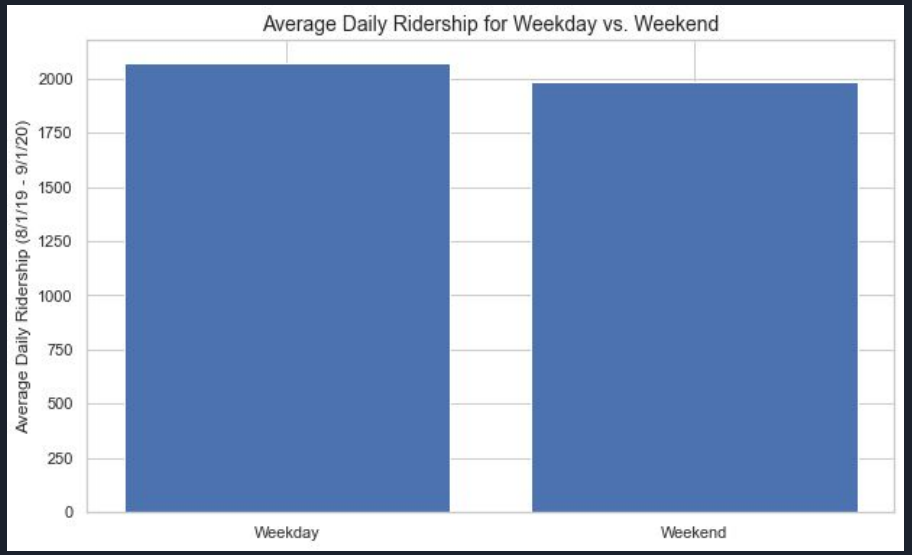
R: - 0.053237

p-value: 0.707786



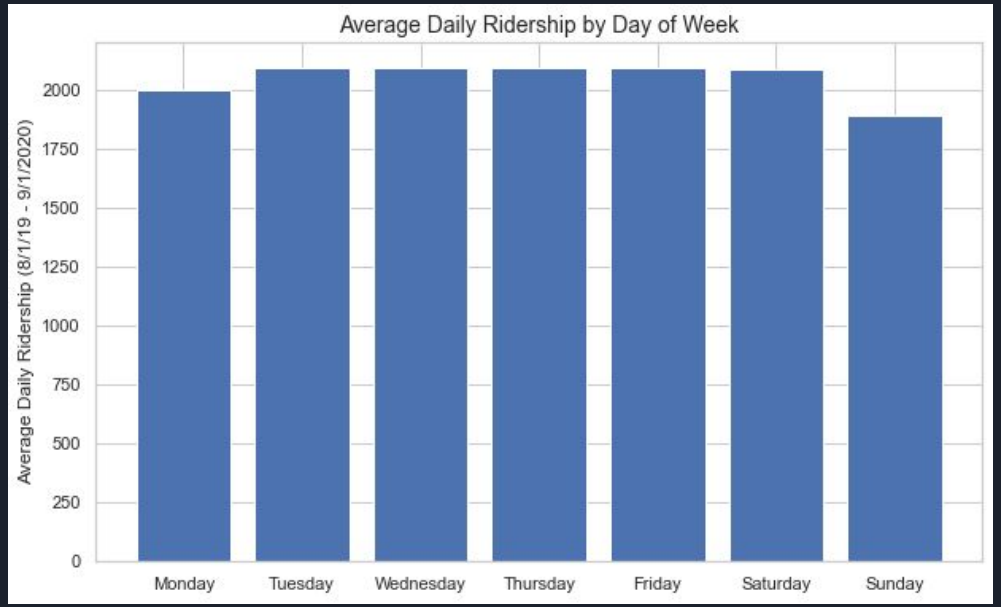
There is no statistically significant difference in ridership between weekdays and weekends

t-stat: 0.802
p-value: 0.423

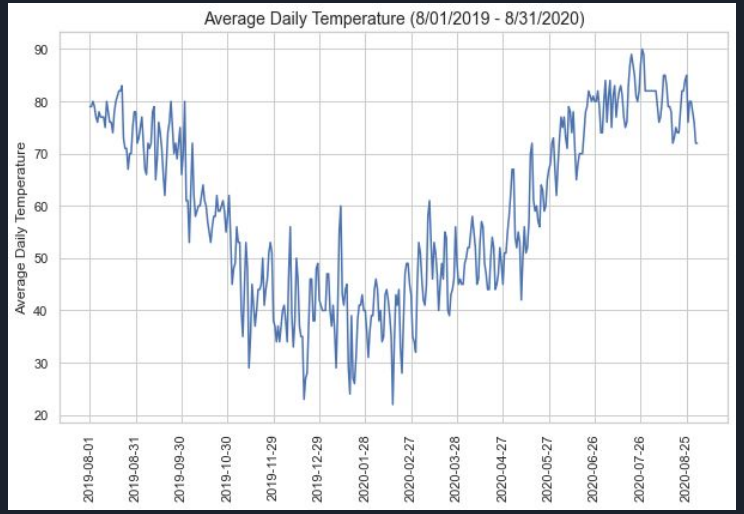


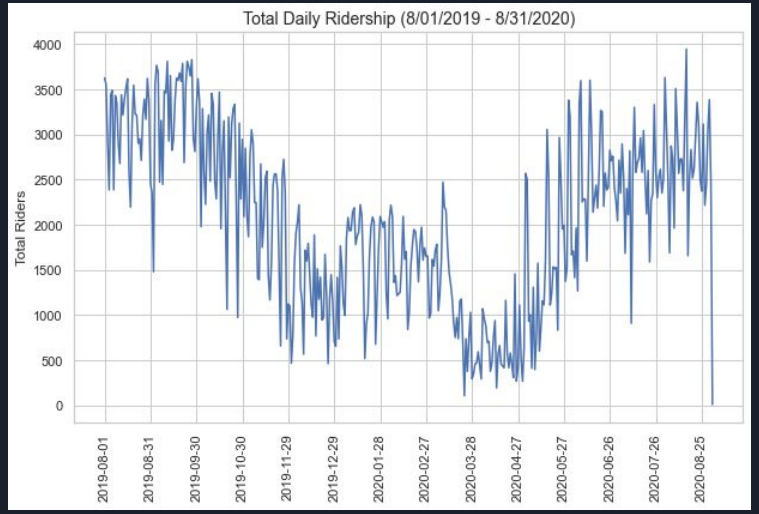
Similarly, there is no significant difference in ridership across each day in the week

F-stat: 0.369
p-value: 0.898



Ridership seems to map onto daily average temperatures fairly well





The trend in ridership seems to fluctuate with the seasons with the unusual anomaly that was COVID-10 in the spring of 2020.

Average daily precipitation in the past year