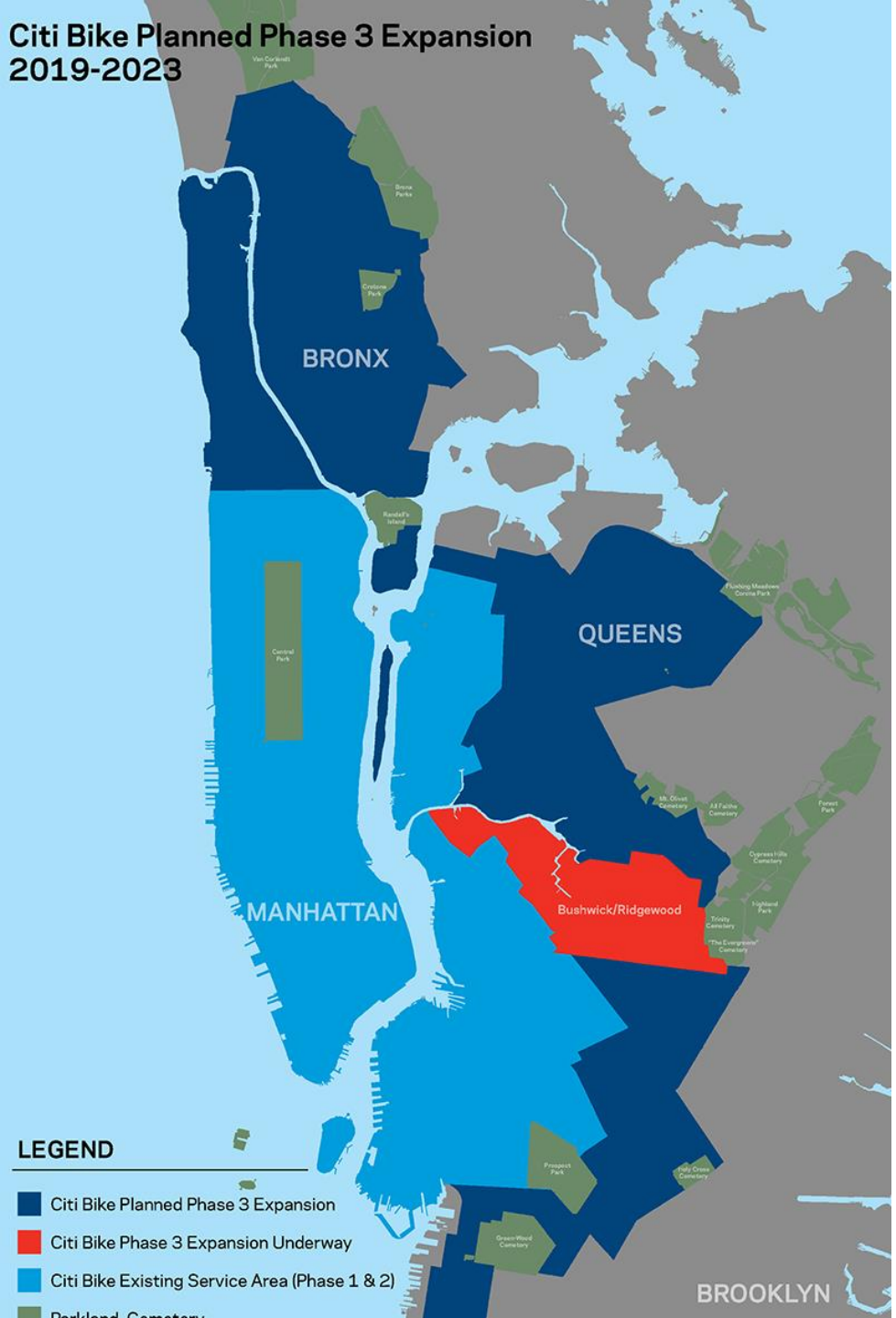
Exploring the Market for Bikeshare in NY



Problem

- The questions of **which areas work best,** and **who is using the bikes** are of paramount importance to this rapid growth, for
  - Marketing
    - Target demographics that use bikes, or expand to demographics that don’t
  - Planning new stations

Agenda:

To answer the questions posed, we present two approaches.

1. Location effect: analysis of the customer demographic

2. Metric design and analysis: measuring "fitness" of each zone.

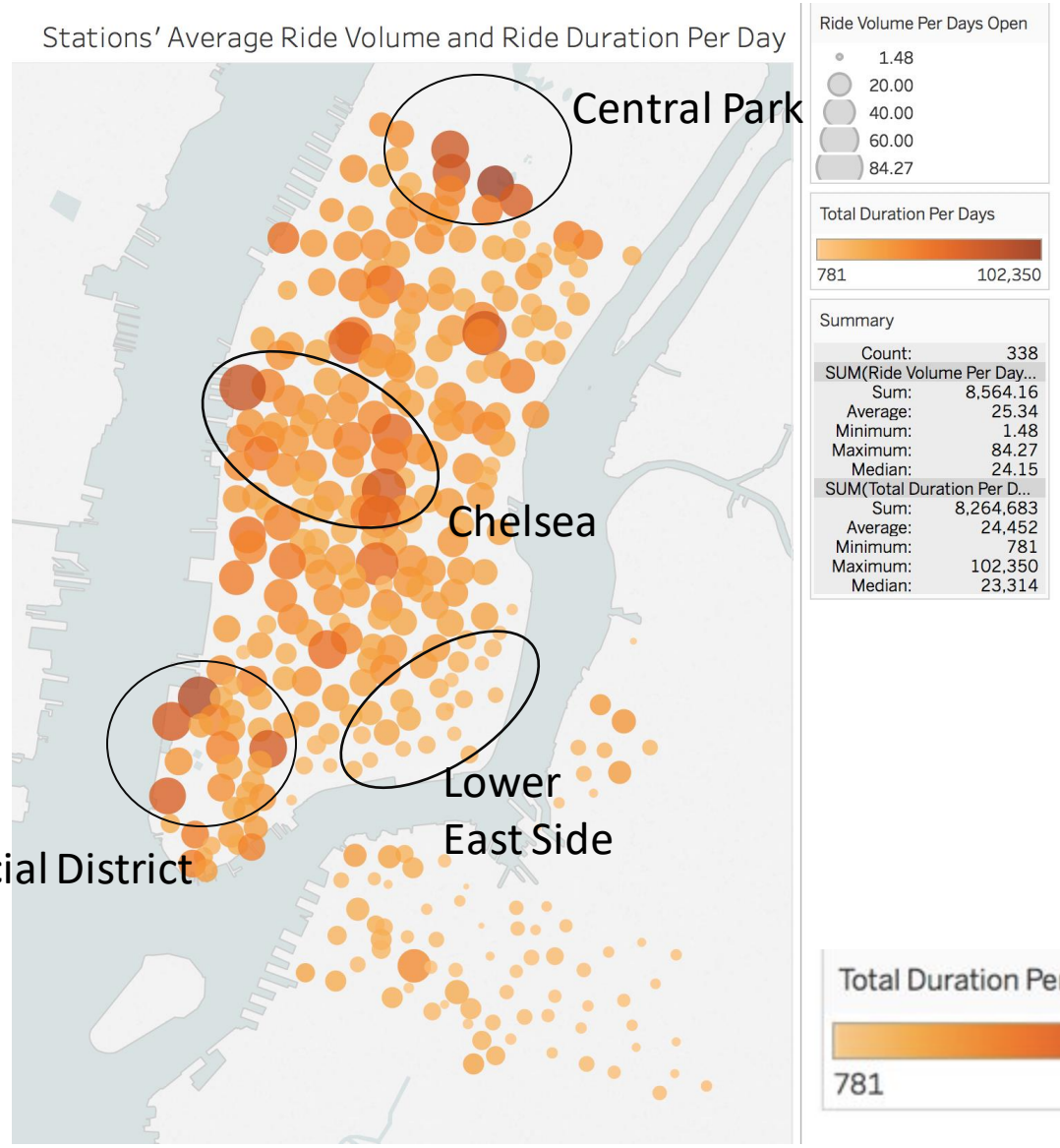
Findings and Significance (TODO)

1. Marketing can be improved drastically based on age and gender.

2. We designed a metric that efficiently measures the "fitness" of a station relative to all other stations, which is uncorrelated to volume

Part I: Correlation Analysis & Findings

An overview of the Citi Bike bikeshare system and demographic of customers.

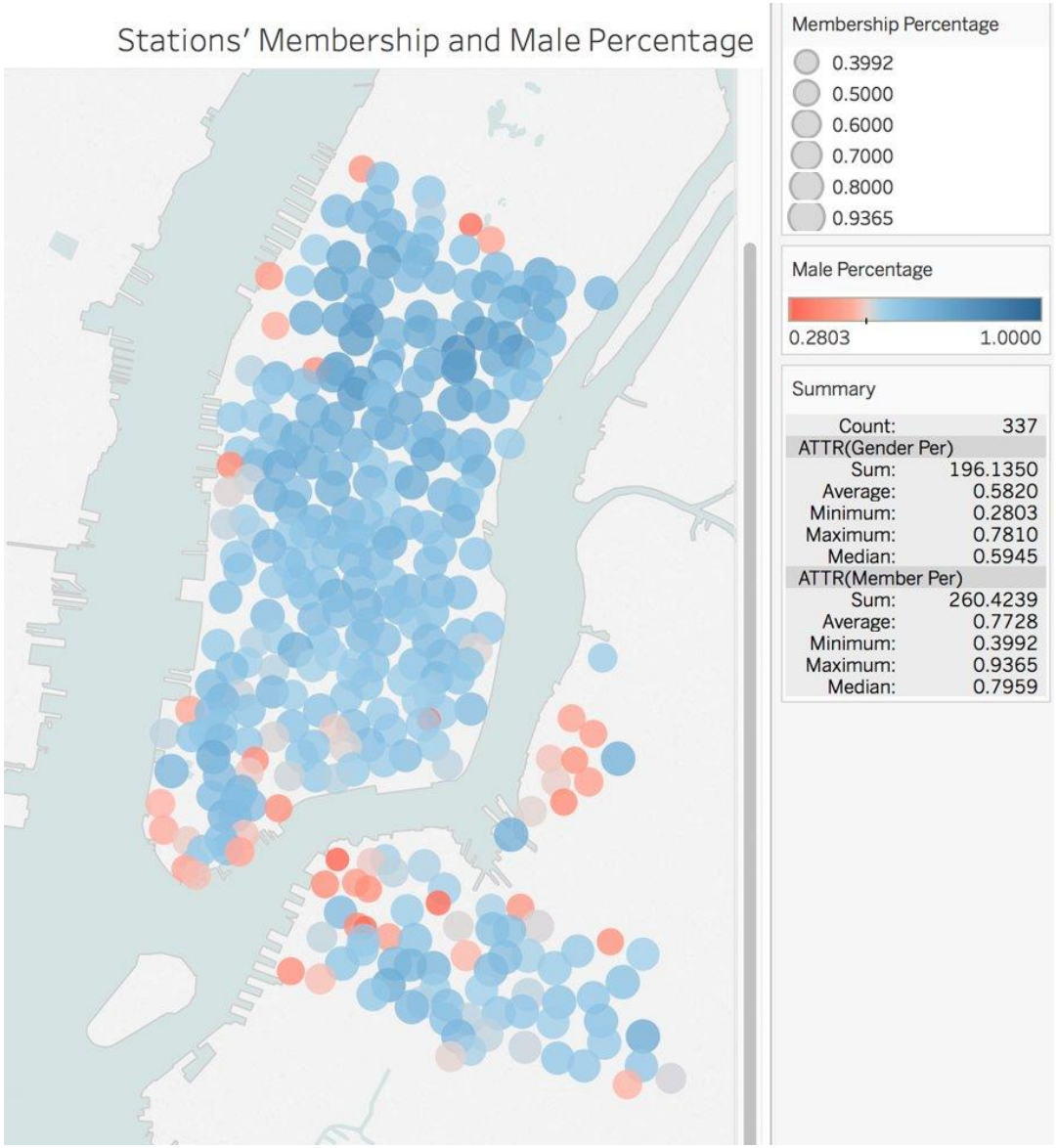


Usage of Citi Bike shows strong geographic patterns

- Certain districts have higher Citi Bike ride volumes:
  - Financial District
  - Chelsea
  - Central Park
- Density of bike stations does not imply the volume
  - Comparing Financial district (low density) with Lower East Side (high density)

The Gender of members shows geographic patterns

- Membership rate is relatively constant spatially.
- Inland bikeshare stations have more male members
- Some stations next to the sea are have more female members



Findings & Significance

- Citi Bike usages show geographic patterns
  - Should further investigation on geographic factors to evaluate the causality
- Some bicycle stations are not used efficiently
  - Need better metric to determine the location of future expansion
- Gender ratio shows geographic pattern
  - More men inland, more women close to the sea

Suggestions:

- Set flexible pricing based on location (price differentiation base on volume)
- Better membership marketing to encourage customer diversity
  - Marketing that can appeal to women should be considered, inland.

Part II: Metric Design & Findings

Define the metric to measure the performance of bikeshare stations

What defines a good metric to measure the “fitness” of a station?

- Meaningful
  - To increase customers
  - To cut the cost for unnecessary stations
  - To grow business efficiently
- Measurable
  - Provide **quantitative scores** to measure the performance/efficiency of each Citi Bike station
- Movable
  - Evaluate current projects
  - Provide predictive suggestions on future expansions
- Trips count is not a good measure on itself

Model description (Regression)

We define “fitness" as the predictive power of a zone's start station frequency on the total volume of bikes per day.

- Using temperature and volume of taxis as controls

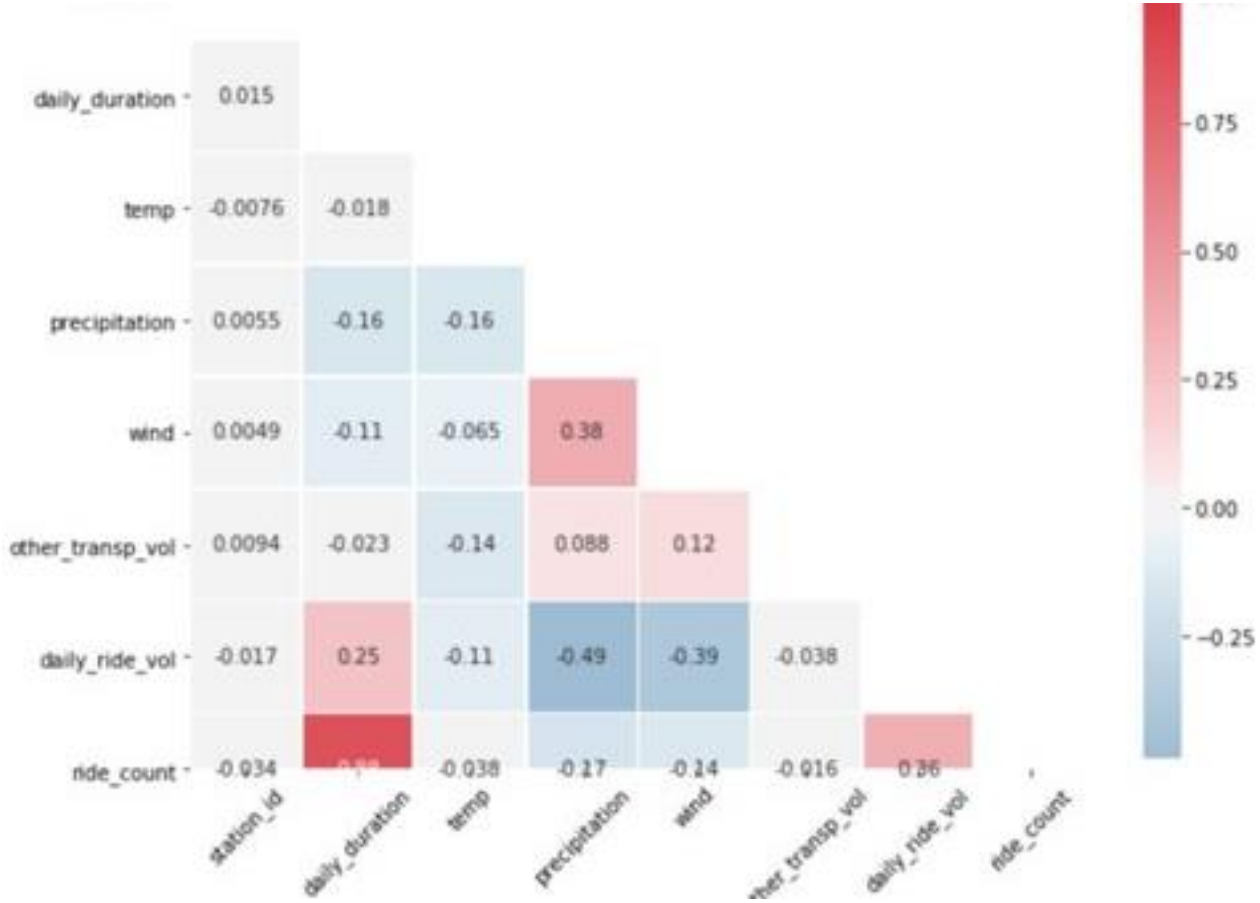
$$V_{tot}^{(t)} = \beta_{controls}^{T} X_{controls}^{(t)} + \beta_{station} v_{station}^{(t)} + \epsilon^{(t)}$$

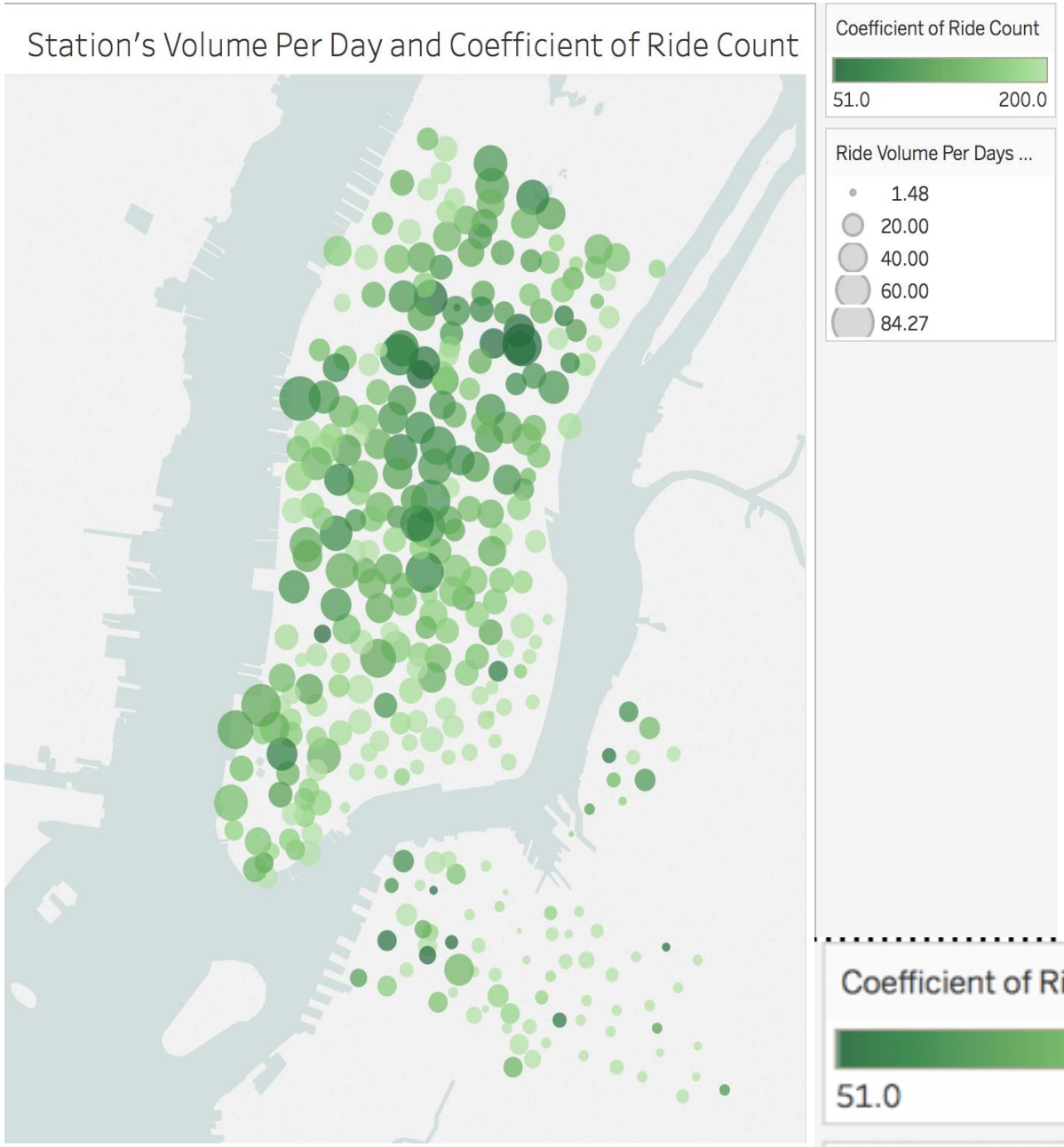
$$metric = \beta_{station}$$

Correlation Matrix

- Relatively low correlations:

Prevent the collinearity of inputs





Station Fitness

- Fitness and ride volume appear to be uncorrelated.
- This means the metric is meaningful and measure the impact of the station on the overall flow of bikes.
- This metric has a plethora of uses.

Metric model

```
Call:
lm(formula = daily_ride_vol ~ . - station_id, data = final_final[,
    -c(1, 9)])

Residuals:
    Min      1Q  Median      3Q     Max
-7344.9 -1189.0   296.1  1399.0  5618.5

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.557e+04  1.384e+02 112.477   <2e-16 ***
daily_duration    1.805e-02  4.651e-04  38.799   <2e-16 ***
temp             -6.053e+01  1.397e+00 -43.318   <2e-16 ***
precipitation    -2.400e+03  2.669e+01 -89.924   <2e-16 ***
wind             -2.232e+02  4.217e+00 -52.931   <2e-16 ***
other_transp_vol  9.939e-03  8.542e-03   1.163    0.245
```

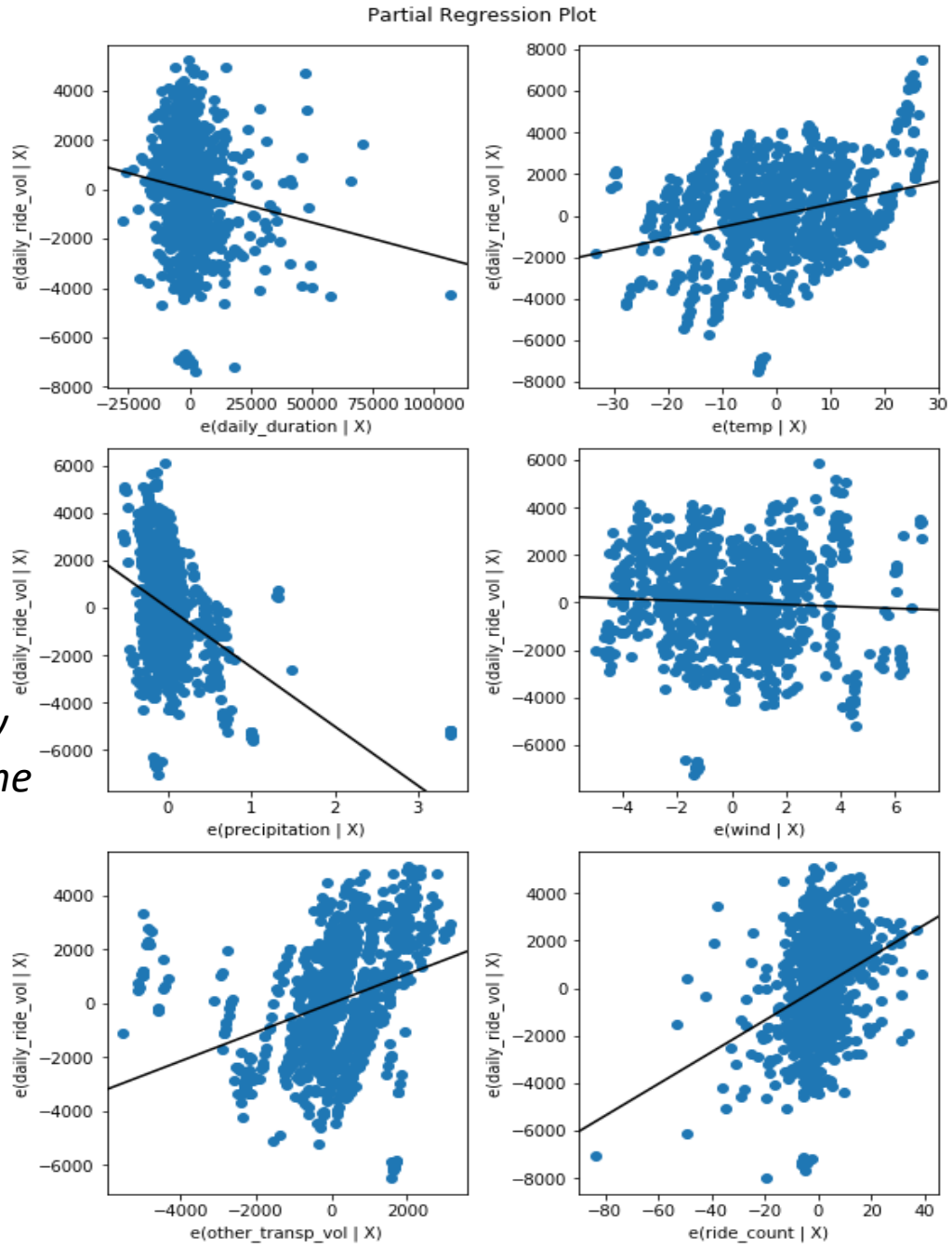
The summary table shows how we overall regression model: Daily_bike_ride_volume is related to *trip_duration_by_station*, *daily temperature, wind*, *precipitation,* and less correlated with the *volume of other transportation types, i*n combination with *ride_count* per station, we design the metric to evaluate each station.

```
```{r}
vif(model)
## to check the collinearity

```
```

```
daily_duration     temp  precipitation     wind other_transp_vol
      1.031949 1.043291       1.216708 1.182806        1.033169
```

We checked the collinearity using variance inflation factor and all of them << 5, pass the test



Impacts

- Retrospective
  - Improve the evaluation of the projects/decisions of expansion
  - Evaluate current efficiency of Citi Bike share system
- Prospective
  - Assessing the potential impact of a future expansion plan
  - Provide insights for choosing next Citi Bike stations
  - Invest money/resources on stations have higher potential; reduce the cost of building and managing unnecessary stations

PIP – Point in Polygon

- We use shapely Python library to draw polygon of NTAs
- To determine the NTA regions each station belongs
- In progress:
  - Regression to investigate
  - Response: metric score
  - Features: Demographic factors of each NTAs
  - Next steps:
    - To find the relationship of demographic factors and how good the station is
    - To determine the specific region of expansion