CODING CHALLENGE

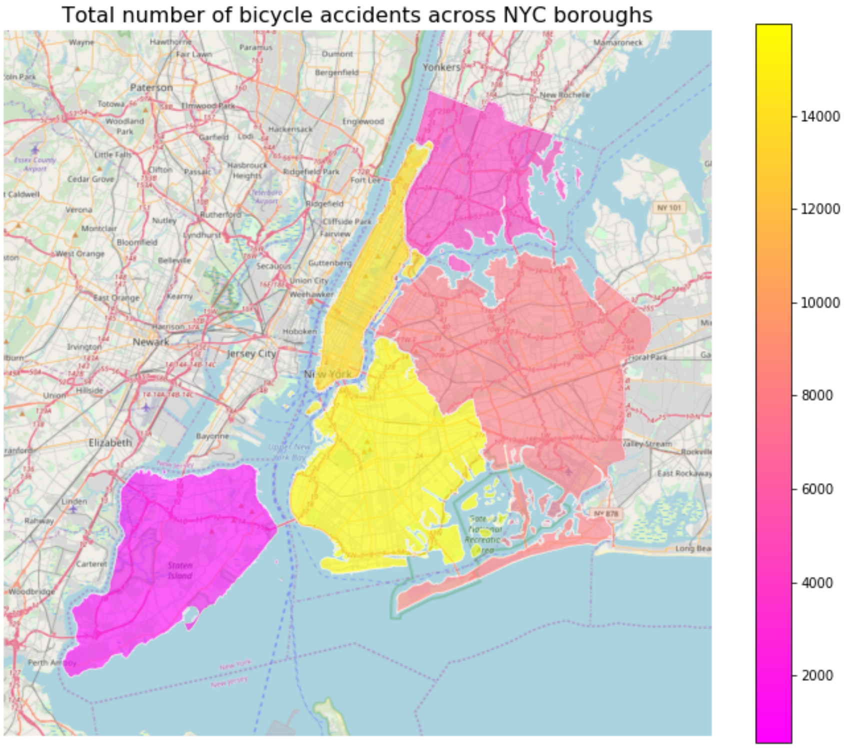
Solution

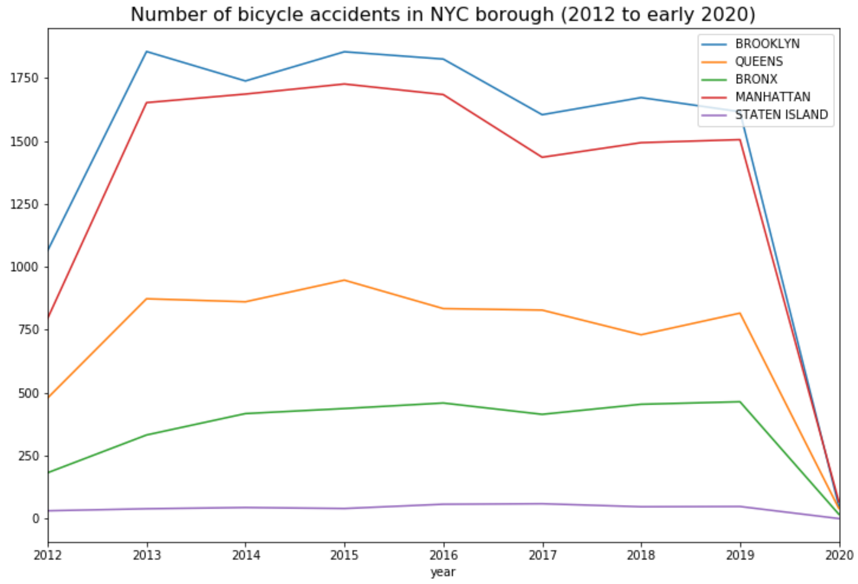
Tamas Barko

+49 14 151 329 325 | tamas.barko@gmail.com

Executive summary

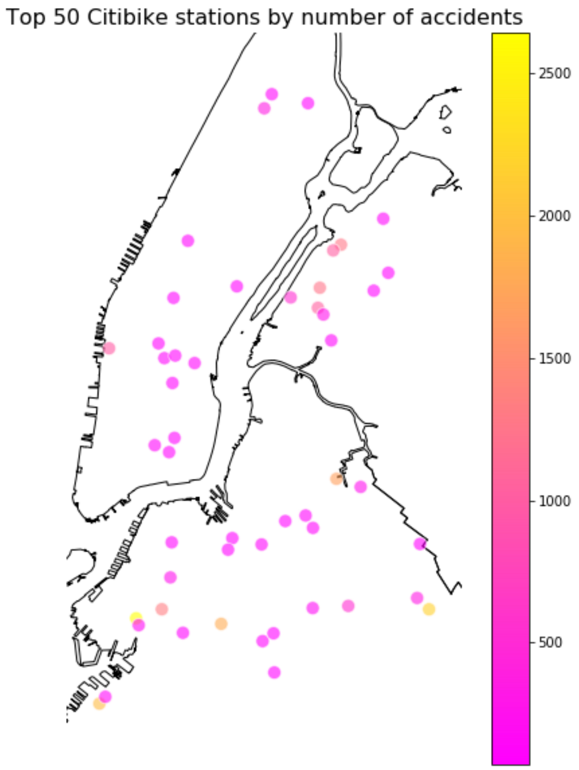
Q1: What is the most dangerous “Borough” as owner of a bicycle in NYC (use visualization)?



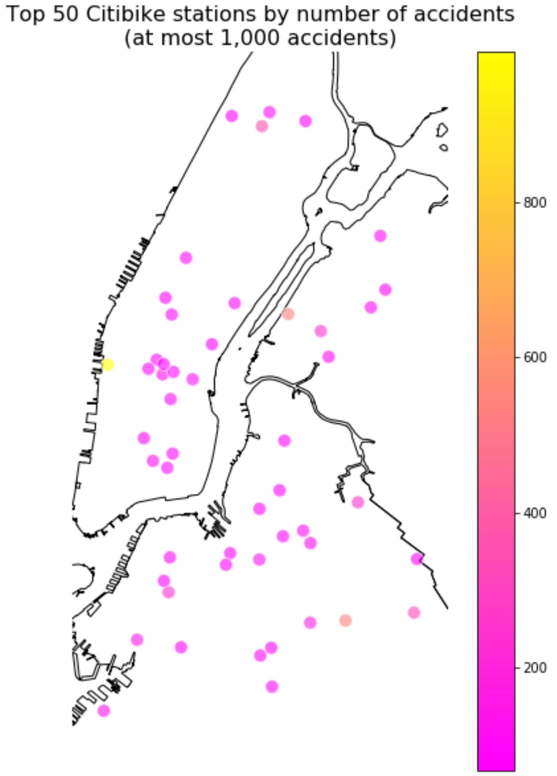


The most dangerous borough in NYC as a bicycle owner is Brooklyn, with over 14,000 accidents documented between 2012 and early 2020.

Q2: What is the worst place to have a CitiBike station (use visualization)?



Top 50 Citibike stations by number of accidents



Top 50 Citibike stations by number of accidents
(at most 1,000 accidents)

The worst place to have a CitiBike station by the number of cycling accidents in the area is at Court St & Nelson St (Brooklyn, Sunset Park) with over 2,400 crashes in the vicinity.

Q3: Create a model to predict where can be an accident and how close it is from the nearest bike station.

Using a stratified sample of 50,000 training observations, and 15,000 test observations, a gradient boosting classifier predicts that an accident involves at least one bicycle with 97.2% accuracy (51.1% AUC score). Accidents predicted to involve at least one bike are on average 725 meters from a CitiBike station.

Data management and cleaning

CitiBike data

1. Several columns contain no information at all, or a singular value – these are dropped ('stAddress2', 'city', 'postalCode', 'location', 'altitude', 'landMark', 'testStation')
2. Some stations (4) are out of service at the time of the query. However, I keep these as the query is only a snapshot of the CitiBike service
3. CitiBike also operates in Jersey City, keep only stations in Boroughs

Crash data

1. Check if cyclist was involved in a crash alone, e.g. hit by something during parking the bike or tipped over alone
2. Check potential representations of bike in the data and create list (e.g., "bike" vs. "byke")
3. Check if any vehicle involved in crash was a bike using the list from 2)
4. Create indicator variable for bike crash using 1) and 3)
5. Some coordinates are out of New York City, e.g., on equator or completely out of bounds. Use map data to filter these out
6. About 200k crashes do not have GPS/borough information
7. For instances in 5) and 6), I use the street information (on and off street) and calculate average coordinates from non-missing values using the entire sample
8. Classify points from 7) into boroughs, drop out-of-bounds coordinates, e.g. coordinates ending up on the East River

Variable description

Dependent variable, class (Y):

- Bike crashed: bike was involved in a crash

Independent variables, features (X):

- Year: capture time fixed effect for a given year, e.g., in later years, there can be more designated bicycle paths/roads reducing accident risk
- Month: capture time fixed effect for a calendar month, some months may be riskier for cyclists, e.g., winter months
- Weekend: there may be more accidents when commuters are in the boroughs
- Time of the day:

    - Morning (6 AM to 12 PM): includes morning rush, drivers may be more attentive in the morning
    - Afternoon (12 PM to 6PM): includes afternoon rush, drivers could be less cautious/tired
    - Night (6 PM to 6 AM – undefined to avoid collinearity): night hours may influence risk differently, e.g., drunk driving
- Borough: capture area fixed effects, as some boroughs may be more dangerous
- Number of casualties:
    - Number of people killed or injured
    - Number of pedestrians killed or injured
    - Not using number of motorists or cyclists to avoid data leakage
- Distance from bike station: the presence of a CitiBike station may increase bicycle traffic and hence accident risk

Discussion

Q1: What is the most dangerous "Borough" as owner of a bicycle in NYC (use visualization)?

I used all available accident information to calculate the total number of bicycle crashes for each borough.

There are two underlying assumptions I had to make to arrive at the results. First, my solution assumes that there is no significant difference between the accident rate of bike owners and bike renters. I believe that this assumption is reasonable, as people who regularly rent a bike are probably as experienced riders as those who use their own bikes. Second, I assumed that accidents happen in the borough where riders live. This is a somewhat strong assumption given that bicycle riders likely ride because the distance they need to cover is too long to walk, which likely also means that they cross boroughs.

Nonetheless, I have no information to relax these assumptions. A potential extension of the analysis would be acquiring the home address of those who had a bike accident and whether they were riding a rental.

Q2: What is the worst place to have a CitiBike station (use visualization)?

I used all available bicycle crash information to calculate the distance between the crash site and the closest CitiBike station.

This approach assumes that all crashes are with rental bikes. I also assumed that any distance is relevant for the calculation of nearby crashes for each station. Finally, I assumed that a crash of a rental bike happens close to the rental station. These are rather strong assumptions that resulted in some likely distortions of the true distribution of crashes near bicycle stations.

As CitiBike stations are only in Manhattan and the parts of Queens and Brooklyn that are across Manhattan, all accidents beyond the boundaries of the CitiBike station area are pooled to a few stations. For example, even though Staten Island does not have many bicycle accidents among the boroughs, all crashes happening there are closest to the CitiBike station at Court St & Nelson St in Brooklyn, Sunset Park.

Assuming that a crash should be attributed to the nearest station also disregards the fact that riders rent a bike to get farther away from the point of renting, and it is unclear why they would be more likely to crash right after starting the ride and not further down the road, for example in an unfamiliar neighborhood.

A worthwhile extension of this analysis would be acquiring information whether the reported accident is with a rental bike and how far the accident is from the original rental location.

Q3: Create a model to predict where can be an accident and how close it is from the nearest bike station.

With the data I have available, the only feasible analysis is to predict whether a crash involves a bike, and how far it is from the nearest CitiBike station. I first standardized continuous measures to avoid variable ranges affecting my results, then used several models to obtain a robust estimate from the training sample. Specifically, I used SVC, logistic regression, random forest and gradient boosting classifiers.

All models have a great accuracy, close to 100%. However, this is due to the fact that most accidents do not involve any bikes (about 97%), so even a classifier predicting every event not to involve a bike would do a great job. Looking at AUM scores, it is apparent that the SVC and the logistic regression do no better than a coin toss (AUM is 50% for both). In fact, the SVC classifies all accidents *not* involving a bike. While the random forest and gradient boosting classifier do somewhat better, they still only slightly outperform random class assignment.

The model specification also relies on the previous assumptions in Q1 and Q2. An immediately feasible extension of this analysis would be using spatial regression including information on the spatial distribution of events. This would help with identifying locations where many riders suffer an accident, but would yield little in figuring out whether a bike picked up at a CitiBike station would actually be a participant of an accident.

Overall, the analysis would benefit from acquiring information on CitiBike rides. Specifically, where a bike was rented, where it was left, and whether it suffered an accident. Without these pieces of information, any model is highly speculative because of the underlying strong assumptions and does not have a strong predictive power for real life situations.

```
--------------------------------------------------
Model: SVC
--------------------------------------------------

Accuracy: 0.972

AUC: 0.5

Distance from CitiBike station (mean): nankm

--------------------------------------------------
Model: Logistic
--------------------------------------------------

Accuracy: 0.972

AUC: 0.5

Distance from CitiBike station (mean): 2.956km

--------------------------------------------------
Model: Random forest classifier
--------------------------------------------------

Accuracy: 0.968

AUC: 0.541

Distance from CitiBike station (mean): 1.601km

--------------------------------------------------
Model: Gradient boosting classifier
--------------------------------------------------

Accuracy: 0.972

AUC: 0.511

Distance from CitiBike station (mean): 0.725km
```