AXA data science challenge

Christian Nass

Monday 3rd November, 2025

1 Aufgabenstellung

Der Fahrradverleih CitiBike vermietet in New York über 12.000 Fahrräder an 750 Verleihstationen. Somit ist CitiBike eine echte Alternative zu den herkömmlichen Transportmitteln, wie z.B. U-Bahn oder Taxi. CitiBike stellt die durch den Verleih gesammelten Daten der Öffentlichkeit zur Verfügung (s. bspw. „2023-citibike-tripdata.zip“).

Deine Aufgabe als Data Scientist ist es, CitiBike dabei zu helfen diese Daten wertstiftend zu nutzen, beispielsweise indem du für CitiBike Kooperationsmöglichkeiten mit einer Versicherung (und/oder umgekehrt) skizzierst. Dazu kannst du zusätzlich die öffentlich zugängigen Daten des NYPD zu Verkehrsunfällen nutzen.

Versioniere die Ergebnisse deiner Analyse bitte mit Git (bspw. auf Github). Stelle uns deine Unterlagen und deinen Code bitte spätestens am Abend vor unserem gemeinsamen Termin via E-Mail zur Verfügung.

Viel Erfolg!

2 Inspect data

Notebook: Inspect-Data.ipynb

2.1 CitiBike data

- Period of time: June 2013 till September 2025
- Columns:
  Content self-explanatory based on name
  Names vary over the years

Column name	2013 – 2019	2020 – 2025
tripduration	✓	
starttime	✓	✓
stoptime	✓	✓
start station id	✓	✓
start station name	✓	✓
start station latitude	✓	✓
start station longitude	✓	✓
end station id	✓	✓
end station name	✓	✓
end station latitude	✓	✓
end station longitude	✓	✓
bikeid	✓	
usertype	✓	✓
birth year	✓	
gender	✓	
ride_id		✓
rideable_type		✓

- Problem: Running out of storage saving all the data on my laptop
  - Notebook: Prepare-CitiBike-Data.ipynb
  - Slimmed files by removing duplicated information
  - Storing station name, latitude and longitude only once in a separate file together with their id
  - Keep only start and end id in tripdata files

2.2 NYPD Motor Vehicle Collisions - Crashes data

- Period of time: August 2012 till 26th October 2025.
  Most reportings from 2016 onwards (new system enroled in 2016).

- Columns:

Column name	Description
CRASH DATE	Occurrence date of collision
CRASH TIME	Occurrence time of collision
BOROUGH	Borough where collision occurred
ZIP CODE	Postal code of incident occurrence
LATITUDE	Latitude coordinate (EPSG 4326)
LONGITUDE	Longitude coordinate (EPSG 4326)
LOCATION	Latitude, Longitude pair
ON STREET NAME	Street on which the collision occurred
CROSS STREET NAME	Nearest cross street to the collision
OFF STREET NAME	Street address if known
NUMBER OF PERSONS INJURED	Number of persons injured
NUMBER OF PERSONS KILLED	Number of persons killed
NUMBER OF PEDESTRIANS INJURED	Number of pedestrians injured
NUMBER OF PEDESTRIANS KILLED	Number of pedestrians killed
NUMBER OF CYCLIST INJURED	Number of cyclists injured
NUMBER OF CYCLIST KILLED	Number of cyclists killed
NUMBER OF MOTORIST INJURED	Number of vehicle occupants injured
NUMBER OF MOTORIST KILLED	Number of vehicle occupants killed
VEHICLE TYPE CODE	Type of vehicle based on the selected vehicle category: e.g. ATV, bicycle, car, escooter, motorcycle (Up to 5)
CONTRIBUTING FACTOR VEHICLE	Factors contributing to the collision for designated vehicle (Up to 5)
COLLISION_ID	Unique record code generated by system

- 2.2 million collisions recorded

- In about 3% of the collisions a bicycle is involved (67147)
  Most of them with a regular (non-electric) bike (83%)

3 Value-adding opportunities of the data

The focus is set on cooperation of CitiBike with an insurance company. Besides that the data can for sure be used to optimise operations at CitiBike. The data can also be used to evaluate the infrastructure e.g. by analysing the average time to commute between stations and the number of bike crashes.

3.1 Cooperation with an insurance company

- CitiBike company
  - Liabilities for inadequate maintenance or defective equipment —> Business liability insurance, business legal protection insurance
  - Theft and vandalism
- CitiBike customers
  - Bike accident insurance
  - Bike liability insurance
  - Traffic legal protection insurance

It is not possible to provide recommendations for specific insurance plans or their prices based on the available data. Nor is it possible to calculate precise risk levels for CitiBike users, since the crash dataset does not indicate whether the bikes involved were rented or privately owned. However, it can be reasonably assumed that riding a CitiBike involves certain risks, and obtaining insurance coverage against them may be advisable.

3.2 CitiBike operation

Main goal is to improve the operations / service.

CitiBike duties	Data opportunities
Define tariffs; Pricing	Not enough information
Order bikes Hire staff Redistribute bikes Open/close stations	Predict future bike usage in general and for each station
Service of bikes	Identify bikes with low usage And bikes with short usage and the same start/end station
Sell ads on bike screens	Acquire customers with shops at frequently visited stations

If the positions of all bikes are known at a given time, it becomes possible to simulate the number of bikes at each station. This allows for the identification of stations that frequently have no bikes available. However, this is difficult to achieve with the data currently available.

4 Data analysis

This section presents the data analysis with respect to the goals listed in the previous section.

4.1 Cooperation with an insurance company

Notebook: Analyse-crashes.ipynb
Analyse risks for bike riders to be involved in a crash and their causes. There is a significant risk of being involved in a crash as a bike driver. This is also true for experienced drivers due to inattention etc. of other road users. For this reason every driver should have an insurance. CitiBike could over in cooperation with an insurance company a special bike insurance, combining the insurances discussed above. Similarly, casual members may be asked before every drive to get the same insurance just for that drive.
Main takeaway points regarding risks are listed below:

- Number of crashes with a bike increase over the years
- Many people use the bikes to commute to work and the most crashes are reported around 5pm, when most people drive back home
- Crashes mainly in Manhattan at large roads (lots of traffic) –> Most CitiBike rides start there too
- Bikes very vulnerable and take severe injuries in crashes where a bike is involved
- In 1 out of 200 reported bike crashes the cyclist died
- Crashes are caused more than twice as often from other road users than from bikers
- There are many different causes for the crashes
  —> Among the most common are driver inattention / distraction, failure to yield Right-of-Way and traffic control disregarded (this is the same for bikers and cars)
- For fatal crashes, additional factors such as alcohol involvement, unsafe speed, and poor road conditions appear more often
- For new e-bike the risk of a crash is raised due to inexperience

Statements about theft and vanalism cannot be made based on the available data. Crashes due to inadequate maintenance or defective equipment seem to be low. If liabilities do exist, they could be costly (especially in America :D), so it might still be worth CitiBike protecting itself against them.

4.2 CitiBike operation

5 Miscellaneous

- Diffusion map of bikes at different timestamps
  Presumably arrows are pointing towards Manhattan in the morning and outwards in the evening
- Ride insurance price dependent on time, customer history, etc. (ride duration not known at start)
- Many accidents probably not recorded
  - Presumably relative low damage/harm but insurance still helpful
  - Accidents without other people involved are not recorded