NYC OpenData

Integrated Capstone for Information Science

Team: Open Dataset Research, Analysis, & Documentation

November 30, 2024

Diwash Ban, Daniel Gwira, Kamal Patel, William Zhao

University of Maryland, College Park

College of Information Studies

COLLEGE OF
INFORMATION
STUDIES

Table of Contents

Introduction

Purpose of the project

This project, developed for the University of Maryland College of Information Studies (iSchool), focuses on the analysis and documentation of the New York City Transportation Public Data Collection. The objective is to develop a comprehensive user guide that improves the accessibility and usability of these datasets for researchers, students, and analysts. This guide will provide clear instructions on navigating the data platform, extracting relevant datasets, and methods to utilize the information effectively. By making the data more accessible, the guide seeks to facilitate research into trends in NYC transportation over time, with the potential to drive improvements in mobility, transit efficiency, and vehicle safety tracking.

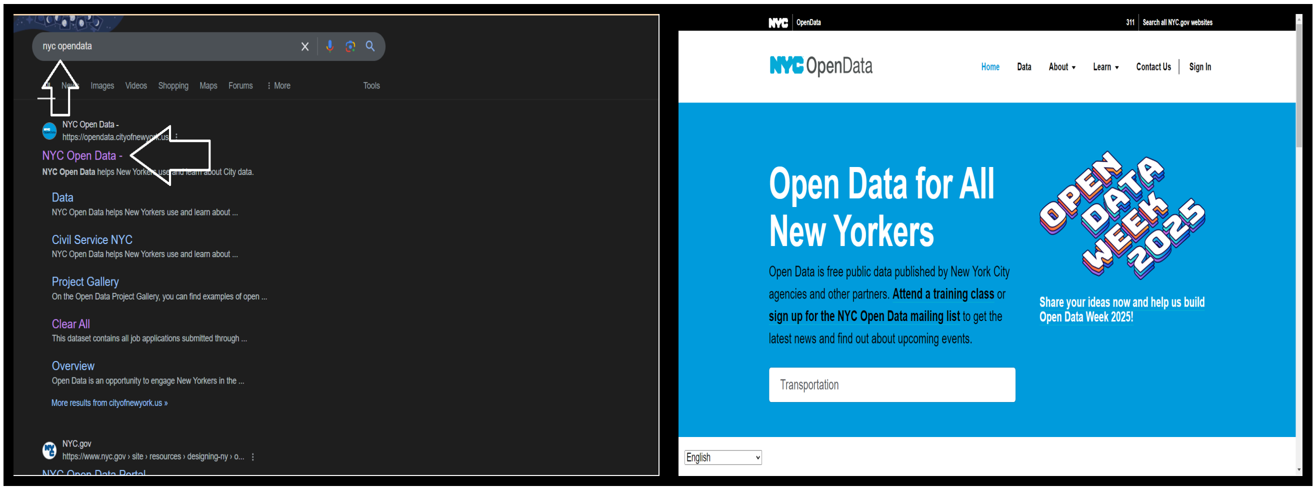
Background

All the datasets used for this project are sourced from the NYC Open Data website, managed by the Open Data Team at the NYC Office of Technology and Innovation (OTI). This team collaborates with various city agencies to identify and release data, oversee platform operations and enhancements, and promote Open Data's use across government entities and the broader NYC community. Our project focuses specifically on the NYC Transportation datasets available on the Open Data website, which range from 2009 to the current date, 2024.

Accessing the Dataset(s)

How to Find Datasets

In order for users to analyze the datasets, users must first access the data. Since this dataset is primarily about New York City transportation, users must enter NYC OpenData on their preferred web browser (Chrome, Safari, Firefox, etc). The web browser would give the user a variety of web applications to choose from but ultimately, the user would choose from the **"NYC Open Data"**. The website has various categories based on the user's needs, such as "Health", "Education", and "Transportation", allowing users to search specific data based on their needs.



NYC Open Data -
https://opendata.cityofnewyork.us
NYC Open Data -
NYC Open Data helps New Yorkers use and learn about City data.

Data
NYC Open Data helps New Yorkers use and learn about ...

Civil Service NYC
NYC Open Data helps New Yorkers use and learn about ...

Project Gallery
On the Open Data Project Gallery, you can find examples of open ...

Clear All
This dataset contains all job applications submitted through ...

Overview
Open Data is an opportunity to engage New Yorkers in the ...

More results from cityofnewyork.us »

NYC.gov
https://www.nyc.gov › site › resources › designing-ny › o...
NYC Open Data Portal

Open Data for All New Yorkers

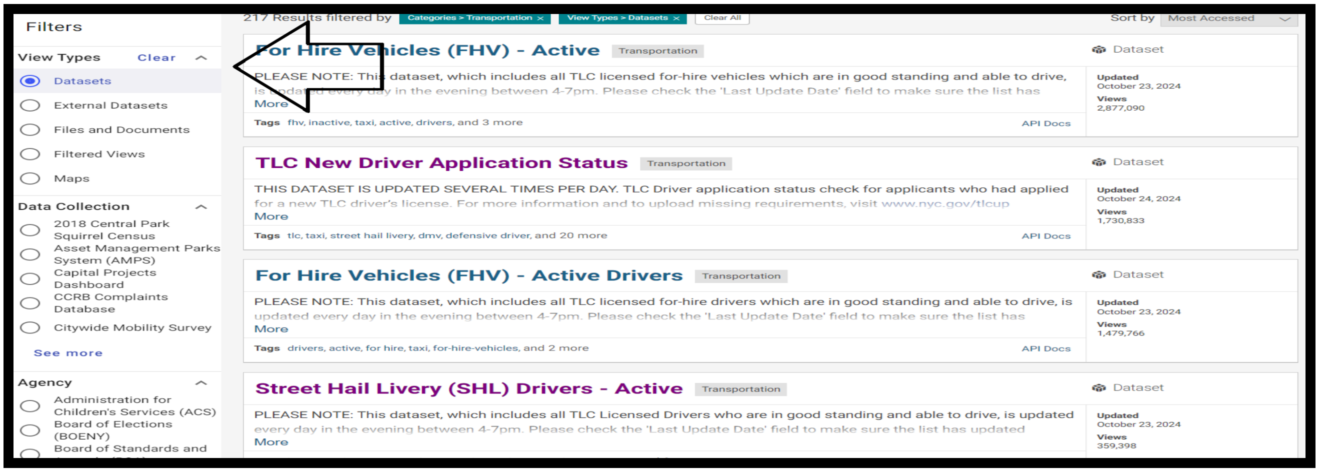
Open Data is free public data published by New York City agencies and other partners. **Attend a training class** or **sign up for the NYC Open Data mailing list** to get the latest news and find out about upcoming events.

Transportation

Share your ideas now and help us build Open Data Week 2025!

When the users gain access to the website, they should be able to locate the search icon in the middle of the page. The user would then type in "transportation and find the available datasets that this user guide is for. This should help the user to narrow down the search to more or the most relevant information about NYC Transportation. In addition to the search bar, the NYC Open Data platform offers categories organized by themes, allowing users to browse by area of focus. By clicking on the "Data" or "Categories" tabs, users can filter results to display datasets relevant to fields like "Health," "Safety," "Environment," and "Transportation." In this case, the user guide focuses primarily on Transportation in NYC, so users will select the Transportation category.

Furthermore, for users who want to narrow down their search further, NYC Open Data offers various filtering options. After entering a keyword or selecting a category like Transportation, users can filter datasets by popularity, date added, or type, such as maps, charts, or tables. When users filter out based on their categories, it will give the users 217 datasets (at the time of this user guide; likely more datasets will be available the a user accesses the website in the future) for their analyses shown in the figure below. To begin to look at and analyze the Transportation datasets under the scope of New York OpenData, it's easy to use the sort-by feature on the top right corner of the page. Click the Alphabetical option for filtering all the datasets in Alphabetical order, this is beneficial because it gives users all the datasets within similar subcategories and headings right below one another to take a look at the change in data over time in years.



217 Results filtered by Categories > Transportation, View Types > Datasets

For Hire Vehicles (FHV) - Active

PLEASE NOTE: This dataset, which includes all TLC licensed for-hire vehicles which are in good standing and able to drive, is updated every day in the evening between 4-7pm. Please check the 'Last Update Date' field to make sure the list has

Tags fhv, inactive, taxi, active, drivers, and 3 more

Updated October 23, 2024
Views 2,877,090

TLC New Driver Application Status

THIS DATASET IS UPDATED SEVERAL TIMES PER DAY. TLC Driver application status check for applicants who had applied for a new TLC driver's license. For more information and to upload missing requirements, visit www.nyc.gov/tlcup

Tags tlc, taxi, street hail livery, dmv, defensive driver, and 20 more

Updated October 24, 2024
Views 1,730,833

For Hire Vehicles (FHV) - Active Drivers

PLEASE NOTE: This dataset, which includes all TLC licensed for-hire drivers which are in good standing and able to drive, is updated every day in the evening between 4-7pm. Please check the 'Last Update Date' field to make sure the list has

Tags drivers, active, for hire, taxi, for-hire-vehicles, and 2 more

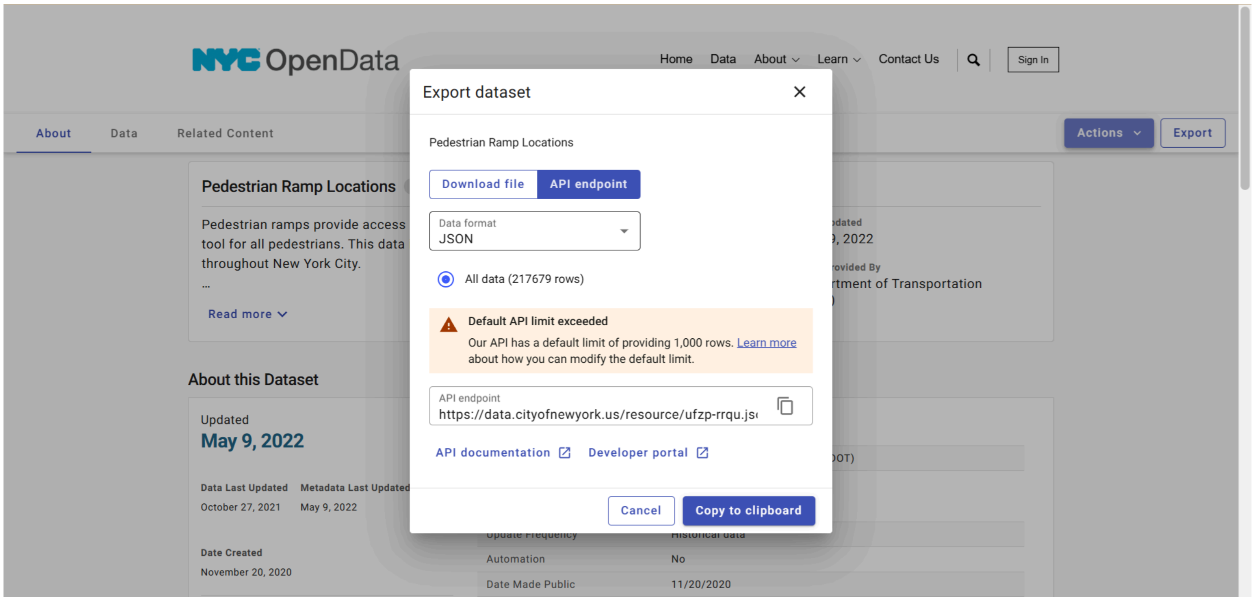
Updated October 23, 2024
Views 1,479,766

Street Hail Livery (SHL) Drivers - Active

PLEASE NOTE: This dataset, which includes all TLC Licensed Drivers who are in good standing and able to drive, is updated every day in the evening between 4-7pm. Please check the 'Last Update Date' field to make sure the list has updated

Updated October 23, 2024
Views 359,398

When users find a dataset that suits their needs, they can preview it directly on the platform. This preview function is helpful to verify the dataset's contents and structure before downloading. For users who wish to integrate real-time data into applications, many NYC Open Data datasets come with API access. This feature allows developers to programmatically access the data, ideal for projects that require continuous updates. The platform provides detailed API documentation to help users retrieve data efficiently and accurately, making NYC Open Data a versatile resource for both research and development purposes.



Export dataset

Pedestrian Ramp Locations

Data format
JSON

● All data (217679 rows)

Default API limit exceeded
Our API has a default limit of providing 1,000 rows. Learn more about how you can modify the default limit.

API endpoint
https://data.cityofnewyork.us/resource/ufzp-rrqu.js

Pedestrian Ramp Locations

Pedestrian ramps provide access tool for all pedestrians. This data throughout New York City.
...

About this Dataset

Updated
May 9, 2022

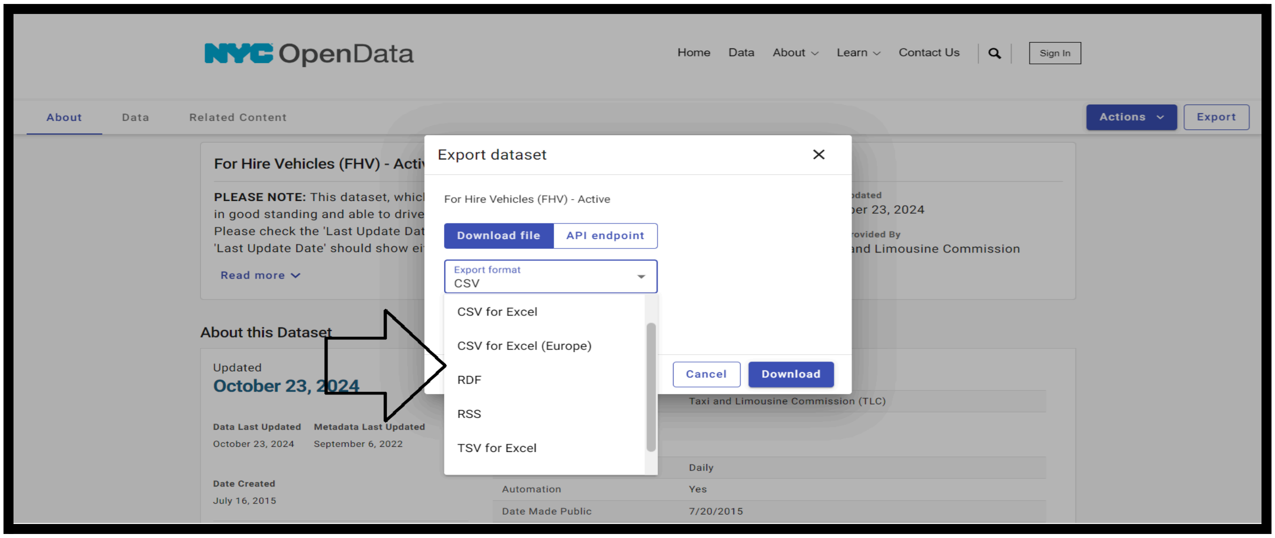
Data Last Updated October 27, 2021
Metadata Last Updated May 9, 2022

Date Created November 20, 2020

Automation No
Date Made Public 11/20/2020

Types of files

Once users have located a specific dataset, they can then access the **Export** icon located at the top right-hand corner of the platform. The icon gives users the option to download in multiple formats, such as CSV or JSON, making them easier for different types of analysis. Furthermore, for any research or project that requires real-time or continuously updated data, the datasets do offer API access. This feature allows an easier transition of data into different applications and also provides documentation to guide users on how to retrieve data program-wise.



For Hire Vehicles (FHV) - Acti

PLEASE NOTE: This dataset, whic
in good standing and able to drive
Please check the 'Last Update Dat
'Last Update Date' should show ei

For Hire Vehicles (FHV) - Active

Export format
CSV

odated
per 23, 2024

rovided By
and Limousine Commission

About this Dataset

Updated
October 23, 2024

Data Last Updated: October 23, 2024
Metadata Last Updated: September 6, 2022

Date Created: July 16, 2015

Taxi and Limousine Commission (TLC)

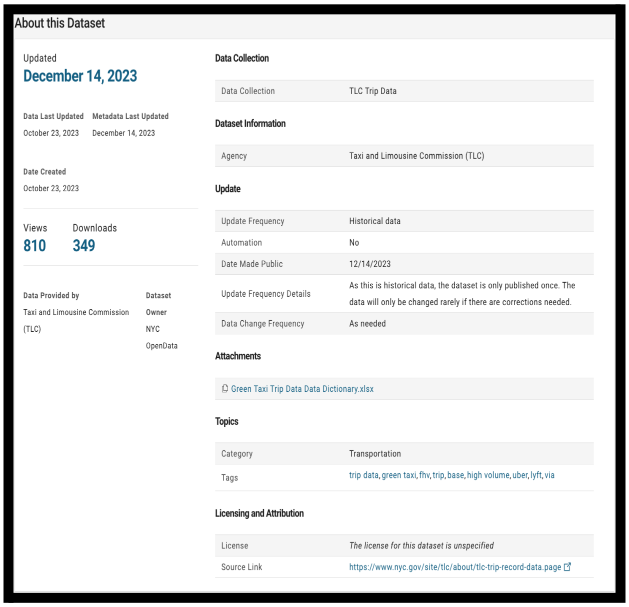
Daily

Automation: Yes

Date Made Public: 7/20/2015

About the Datasets

The NYC Transportation category of the NYC Open Data Portal has exactly 217 dataset submissions at the time this guide is written - all of the datasets are by default accessible as Excel files; other file types are available to be converted to when exporting. After accessing the datasets via the method described in the section above, the metadata of the datasets are made available for viewing. From the metadata, viewers can quickly grasp a general overview of the fields the datasets include. Information about the dataset such as update frequency, dataset agency (provider), file type, tags, source, and total number of rows are all provided.



About this Dataset

Updated
December 14, 2023

Data Last Updated: October 23, 2023
Metadata Last Updated: December 14, 2023

Date Created: October 23, 2023

Views: 810
Downloads: 349

Data Provided by: Taxi and Limousine Commission (TLC)
Dataset Owner: NYC OpenData

Data Collection

Data Collection	TLC Trip Data

Dataset Information

Agency	Taxi and Limousine Commission (TLC)

Update

Update Frequency	Historical data
Automation	No
Date Made Public	12/14/2023
Update Frequency Details	As this is historical data, the dataset is only published once. The data will only be changed rarely if there are corrections needed.
Data Change Frequency	As needed

Attachments

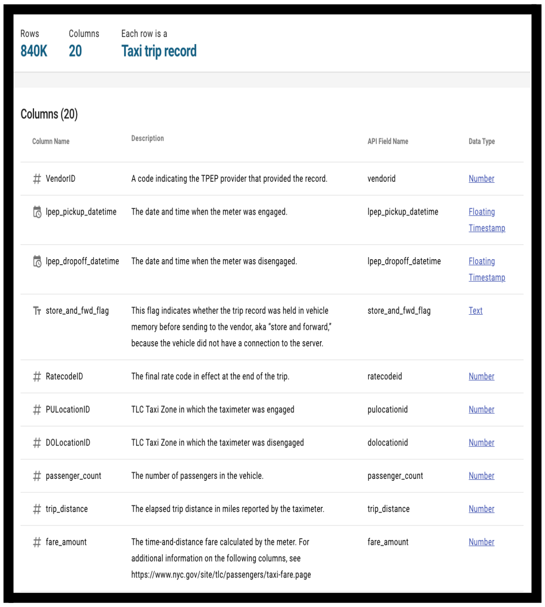
Green Taxi Trip Data Data Dictionary.xlsx

Topics

Category	Transportation
Tags	trip data, green taxi, fhv, trip, base, high volume, uber, lyft, via

Licensing and Attribution

License	*The license for this dataset is unspecified*
Source Link	https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page



Rows: 840K
Columns: 20
Each row is a: Taxi trip record

Columns (20)

Column Name	Description	API Field Name	Data Type
# VendorID	A code indicating the TPEP provider that provided the record.	vendorid	Number
lpep_pickup_datetime	The date and time when the meter was engaged.	lpep_pickup_datetime	Floating Timestamp
lpep_dropoff_datetime	The date and time when the meter was disengaged.	lpep_dropoff_datetime	Floating Timestamp
Tт store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server.	store_and_fwd_flag	Text
# RatecodeID	The final rate code in effect at the end of the trip.	ratecodeid	Number
# PULocationID	TLC Taxi Zone in which the taximeter was engaged	pulocationid	Number
# DOLocationID	TLC Taxi Zone in which the taximeter was disengaged	dolocationid	Number
# passenger_count	The number of passengers in the vehicle.	passenger_count	Number
# trip_distance	The elapsed trip distance in miles reported by the taximeter.	trip_distance	Number
# fare_amount	The time-and-distance fare calculated by the meter. For additional information on the following columns, see https://www.nyc.gov/site/tlc/passengers/taxi-fare.page	fare_amount	Number

Central Google Sheet Repository Overview

Link: Google Sheet Open Dataset Repository
(Users can access this comprehensive file on the Github repository for the user guide)

A Google Sheet was created to extract attribute information for each dataset, categorized under columns such as Year, Title, Link, Columns, Brief Description, File Type, Data Provided By, Tags, and Source Link. This central file is designed to be convenient for end users, including but not limited to, Ischool students, researchers, journalists, and University of Maryland Community members, providing easy access to all the core attributes of the datasets in one place, enabling users to determine if the datasets serve their purpose.

Identifying Changes Over Time

The approach helps identify changes in specific categories of datasets related to the New York City Transportation collection over the years by examining the column names to see if there have been any modifications in the type of information collected. On the NYC transportation website, end users would need to scroll and navigate through multiple pages to view all 217 datasets. Additionally, through the New York City open data portal website, users must click on each dataset to see its contents, requiring extensive scrolling and navigation to view all column names and captured data.

Improved Accessibility and User Experience

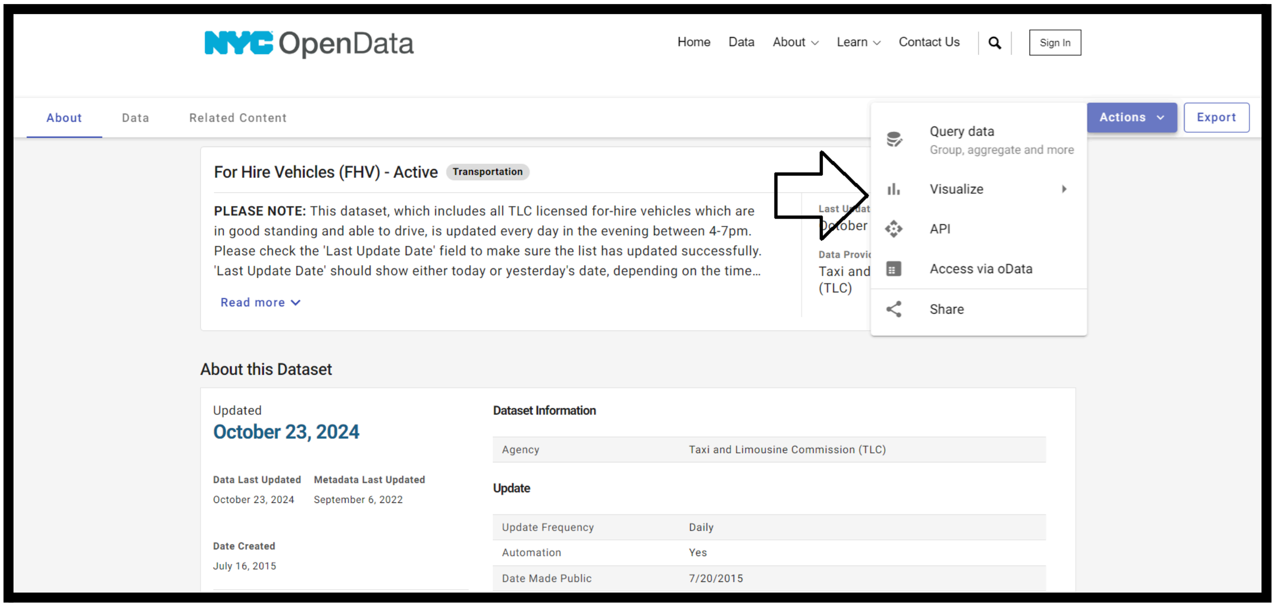
Furthermore, the consolidated sheet ensures that users can quickly identify trends, compare datasets across different years, and make informed decisions without the need for extensive navigation. This initiative not only improves accessibility but also enhances the overall user experience by providing a comprehensive view of the data in an easily digestible format.

NYC Data Visualization

Unlock Insights With NYC Portal Visualization Tools

Using the built-in visualization tools on the New York City transportation data portal can greatly enhance the way you interact with and analyze data. These tools are designed to be user-friendly, providing both professionals and the general public with valuable insights into various transportation metrics. Here's how you can start using these visualization tools:

Users can locate the **Actions** icon on the top right corner of the platform. Within the icon, it gives users the ability to choose either Query data, Visualize, API, or Access via oData and Share. Users will then choose the Visualize tab shown in the figure below. The platform will then send users to different tabs dedicated to Visualization.



For Hire Vehicles (FHV) - Active

PLEASE NOTE: This dataset, which includes all TLC licensed for-hire vehicles which are in good standing and able to drive, is updated every day in the evening between 4-7pm. Please check the 'Last Update Date' field to make sure the list has updated successfully. 'Last Update Date' should show either today or yesterday's date, depending on the time...

Last Updat
October

Data Provi
Taxi and
(TLC)

About this Dataset

Updated
October 23, 2024

Data Last Updated
October 23, 2024

Metadata Last Updated
September 6, 2022

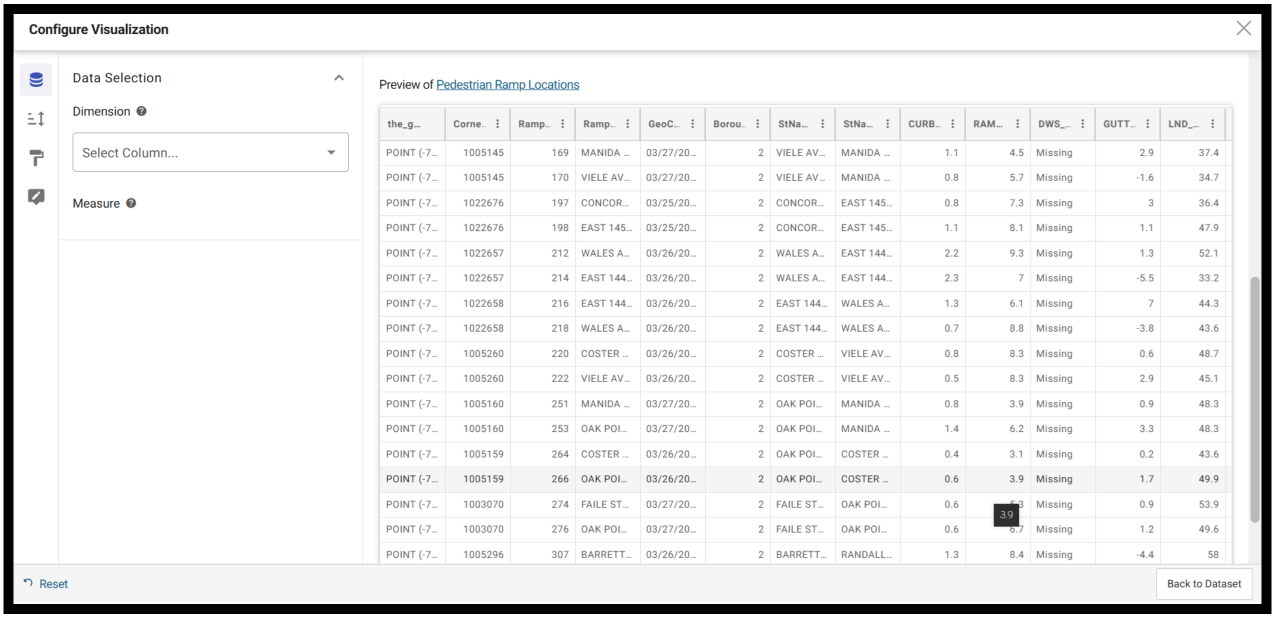
Date Created
July 16, 2015

Dataset Information

Agency	Taxi and Limousine Commission (TLC)

Update

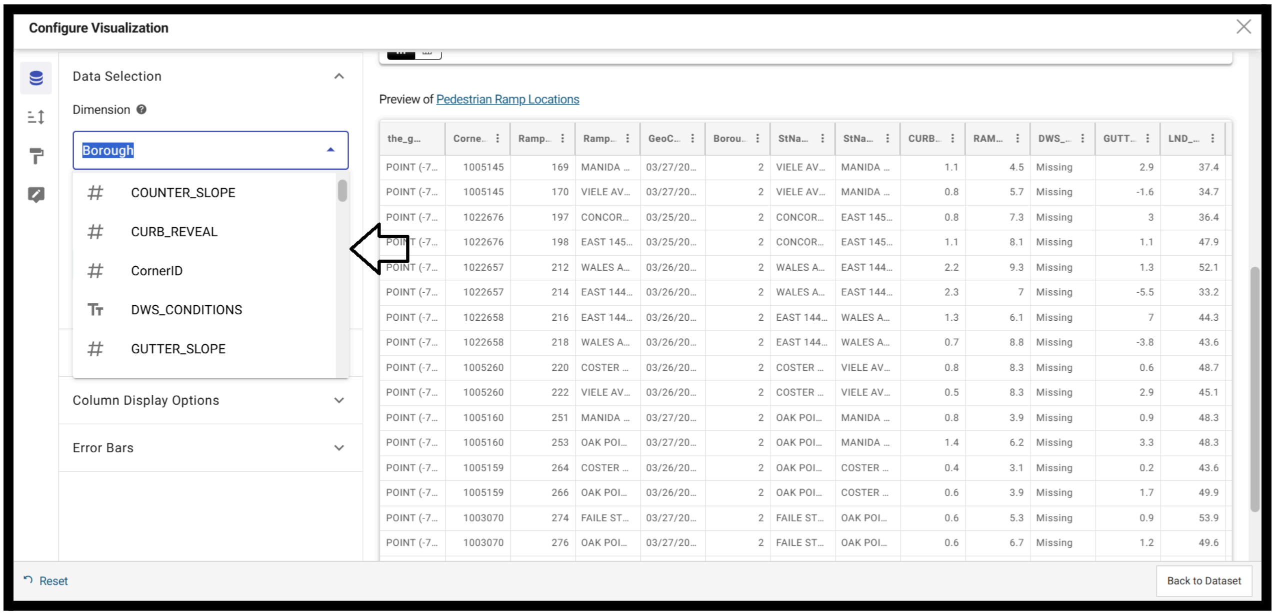
Update Frequency	Daily
Automation	Yes
Date Made Public	7/20/2015



Preview of Pedestrian Ramp Locations

the_g...	Corne...	Ramp...	Ramp...	GeoC...	Borou...	StNa...	StNa...	CURB...	RAM...	DWS_...	GUTT...	LND_...
POINT (-7...	1005145	169	MANIDA ...	03/27/20...	2	VIELE AV...	MANIDA ...	1.1	4.5	Missing	2.9	37.4
POINT (-7...	1005145	170	VIELE AV...	03/27/20...	2	VIELE AV...	MANIDA ...	0.8	5.7	Missing	-1.6	34.7
POINT (-7...	1022676	197	CONCOR...	03/25/20...	2	CONCOR...	EAST 145...	0.8	7.3	Missing	3	36.4
POINT (-7...	1022676	198	EAST 145...	03/25/20...	2	CONCOR...	EAST 145...	1.1	8.1	Missing	1.1	47.9
POINT (-7...	1022657	212	WALES A...	03/26/20...	2	WALES A...	EAST 144...	2.2	9.3	Missing	1.3	52.1
POINT (-7...	1022657	214	EAST 144...	03/26/20...	2	WALES A...	EAST 144...	2.3	7	Missing	-5.5	33.2
POINT (-7...	1022658	216	EAST 144...	03/26/20...	2	EAST 144...	WALES A...	1.3	6.1	Missing	7	44.3
POINT (-7...	1022658	218	WALES A...	03/26/20...	2	EAST 144...	WALES A...	0.7	8.8	Missing	-3.8	43.6
POINT (-7...	1005260	220	COSTER ...	03/26/20...	2	COSTER ...	VIELE AV...	0.8	8.3	Missing	0.6	48.7
POINT (-7...	1005260	222	VIELE AV...	03/26/20...	2	COSTER ...	VIELE AV...	0.5	8.3	Missing	2.9	45.1
POINT (-7...	1005160	251	MANIDA ...	03/27/20...	2	OAK POI...	MANIDA ...	0.8	3.9	Missing	0.9	48.3
POINT (-7...	1005160	253	OAK POI...	03/27/20...	2	OAK POI...	MANIDA ...	1.4	6.2	Missing	3.3	48.3
POINT (-7...	1005159	264	COSTER ...	03/26/20...	2	OAK POI...	COSTER ...	0.4	3.1	Missing	0.2	43.6
POINT (-7...	1005159	266	OAK POI...	03/26/20...	2	OAK POI...	COSTER ...	0.6	3.9	Missing	1.7	49.9
POINT (-7...	1003070	274	FAILE ST...	03/27/20...	2	FAILE ST...	OAK POI...	0.6	5.3	Missing	0.9	53.9
POINT (-7...	1003070	276	OAK POI...	03/27/20...	2	FAILE ST...	OAK POI...	0.6	6.7	Missing	1.2	49.6
POINT (-7...	1005296	307	BARRETT...	03/26/20...	2	BARRETT...	RANDALL...	1.3	8.4	Missing	-4.4	58

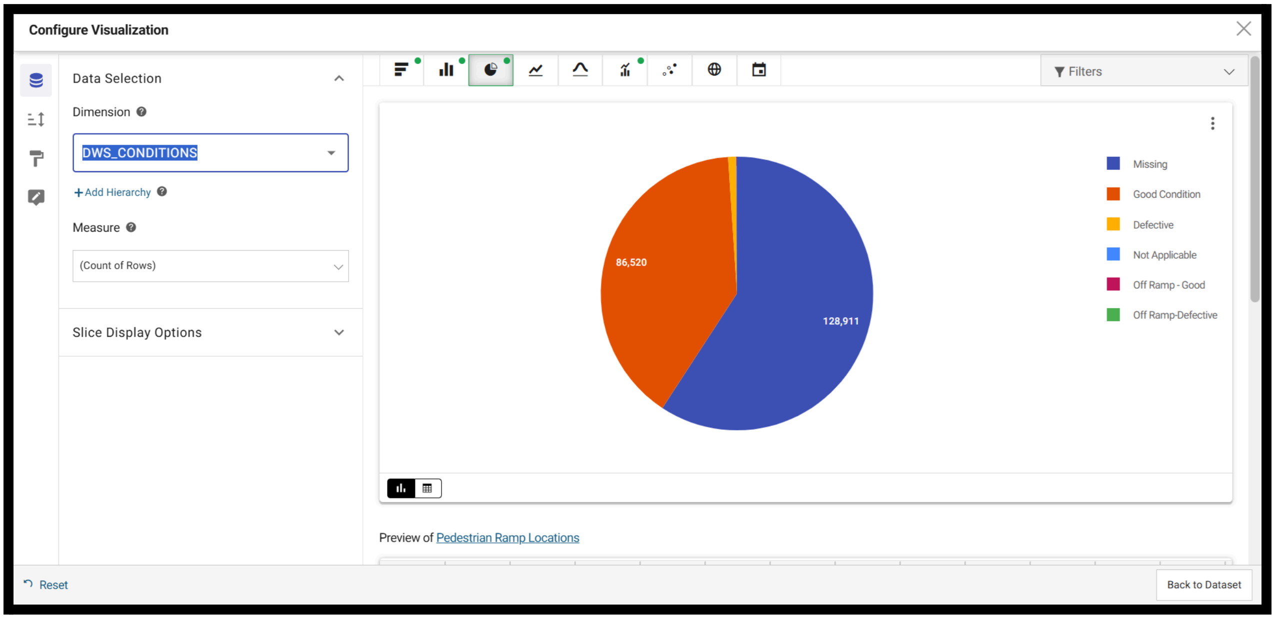
Once you choose a visualization type, the next step is to customize it to fit your specific analysis needs. The portal allows you to filter data based on various parameters such as date ranges, geographic areas, or other dataset-specific metrics. This can be done through simple dropdown menus and sliders that refine the data dynamically (*see figure below*). Additionally, if the dataset includes multiple variables, you can layer these variables in a single visualization to provide a more comprehensive view. For example, a map could show both traffic incident hotspots and road closures, giving a fuller picture of transportation challenges in a particular area.



Preview of Pedestrian Ramp Locations

the_g...	Corne...	Ramp...	Ramp...	GeoC...	Borou...	StNa...	StNa...	CURB...	RAM...	DWS_...	GUTT...	LND_...
POINT (-7...	1005145	169	MANIDA ...	03/27/20...	2	VIELE AV...	MANIDA ...	1.1	4.5	Missing	2.9	37.4
POINT (-7...	1005145	170	VIELE AV...	03/27/20...	2	VIELE AV...	MANIDA ...	0.8	5.7	Missing	-1.6	34.7
POINT (-7...	1022676	197	CONCOR...	03/25/20...	2	CONCOR...	EAST 145...	0.8	7.3	Missing	3	36.4
POINT (-7...	1022676	198	EAST 145...	03/25/20...	2	CONCOR...	EAST 145...	1.1	8.1	Missing	1.1	47.9
POINT (-7...	1022657	212	WALES A...	03/26/20...	2	WALES A...	EAST 144...	2.2	9.3	Missing	1.3	52.1
POINT (-7...	1022657	214	EAST 144...	03/26/20...	2	WALES A...	EAST 144...	2.3	7	Missing	-5.5	33.2
POINT (-7...	1022658	216	EAST 144...	03/26/20...	2	EAST 144...	WALES A...	1.3	6.1	Missing	7	44.3
POINT (-7...	1022658	218	WALES A...	03/26/20...	2	EAST 144...	WALES A...	0.7	8.8	Missing	-3.8	43.6
POINT (-7...	1005260	220	COSTER ...	03/26/20...	2	COSTER ...	VIELE AV...	0.8	8.3	Missing	0.6	48.7
POINT (-7...	1005260	222	VIELE AV...	03/26/20...	2	COSTER ...	VIELE AV...	0.5	8.3	Missing	2.9	45.1
POINT (-7...	1005160	251	MANIDA ...	03/27/20...	2	OAK POI...	MANIDA ...	0.8	3.9	Missing	0.9	48.3
POINT (-7...	1005160	253	OAK POI...	03/27/20...	2	OAK POI...	MANIDA ...	1.4	6.2	Missing	3.3	48.3
POINT (-7...	1005159	264	COSTER ...	03/26/20...	2	OAK POI...	COSTER ...	0.4	3.1	Missing	0.2	43.6
POINT (-7...	1005159	266	OAK POI...	03/26/20...	2	OAK POI...	COSTER ...	0.6	3.9	Missing	1.7	49.9
POINT (-7...	1003070	274	FAILE ST...	03/27/20...	2	FAILE ST...	OAK POI...	0.6	5.3	Missing	0.9	53.9
POINT (-7...	1003070	276	OAK POI...	03/27/20...	2	FAILE ST...	OAK POI...	0.6	6.7	Missing	1.2	49.6

Users can then start by selecting the type of visualization they wish to create, like maps, line charts, bar graphs, and pie charts. The choice depends on what the users find most suitable for the kind of data they are analyzing. For instance, spatial data like traffic volumes or accident locations are most effectively represented on maps, while trends over time are conveyed through line charts.



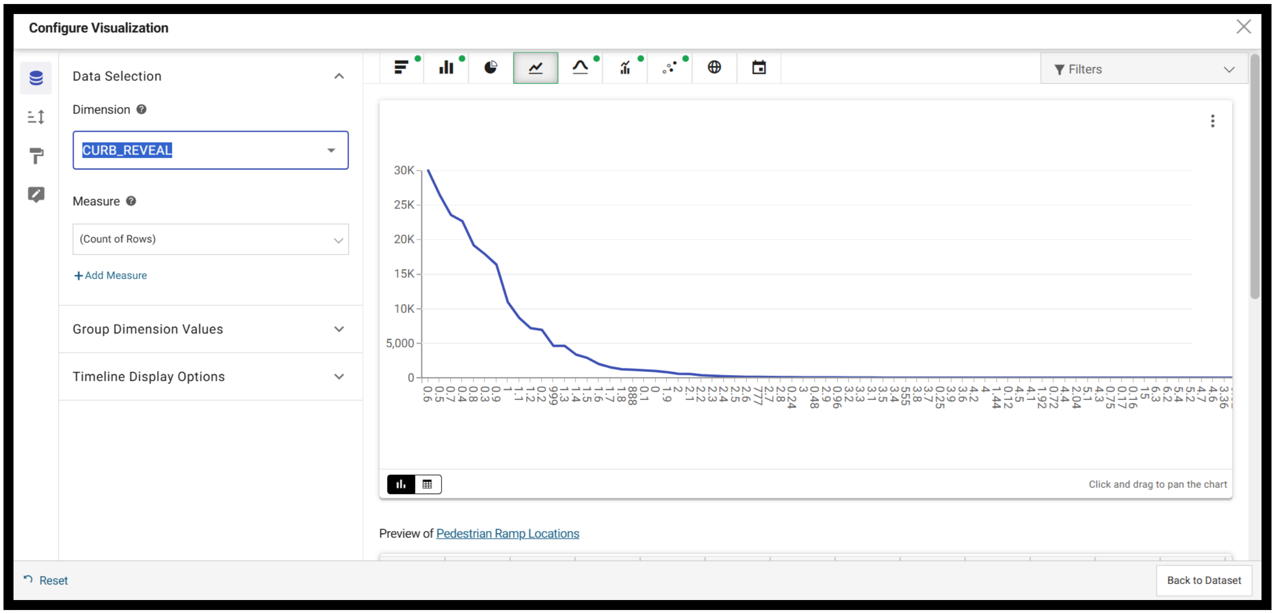
Configure Visualization

Data Selection

Dimension: DWS_CONDITIONS

Measure: (Count of Rows)

Preview of Pedestrian Ramp Locations



Configure Visualization

Data Selection

Dimension: CURB_REVEAL

Measure: (Count of Rows)

Preview of Pedestrian Ramp Locations

Furthermore, these visualization tools also offer advanced features such as the ability to export your visualizations. Users can download the visualized data in common formats such as PNG or JPEG for images, and CSV or Excel for raw data tables. This feature is particularly useful for including visuals in reports, and presentations, or sharing them through email or social media to reach a broader audience or to use in further detailed analysis using external software.

By mastering these visualization tools, users can unlock the potential to turn raw data into actionable insights. Whether it's identifying trends, making comparisons, or pinpointing issues, the visual representation of data can be a powerful tool for anyone working in or interested in urban planning, policy-making, or simply looking to understand more about the transportation landscape of New York City. The insights can also be applied elsewhere such as new or smart city developments where transportation must be designed to be efficient.

Data Analysis

Dataset Background:

To illustrate how users can perform data analysis on all NYC Transportation datasets, this guide uses the "Pedestrian Ramp Locations" dataset as an example to show users how to better perform analysis on any datasets they choose to do. The "Pedestrian Ramp Locations" dataset has a collection of data that ranges from 200,000 rows to 250,000 rows. The data has a list of pedestrian ramps that are organized across 24 columns. Within each row, the dataset describes a specific pedestrian ramp, detailing various things including geographical coordinates, ramp dimensions, conditions, etc. Additionally, the dataset is organized to contain both numerical and letter data types, as well as categorical data types. These include descriptive information such as the state of visible warning areas, which are displayed in column **DWS_CONDITIONS**, and geometrical points, which are displayed in column **the_geom** and indicate the exact position of the ramp.

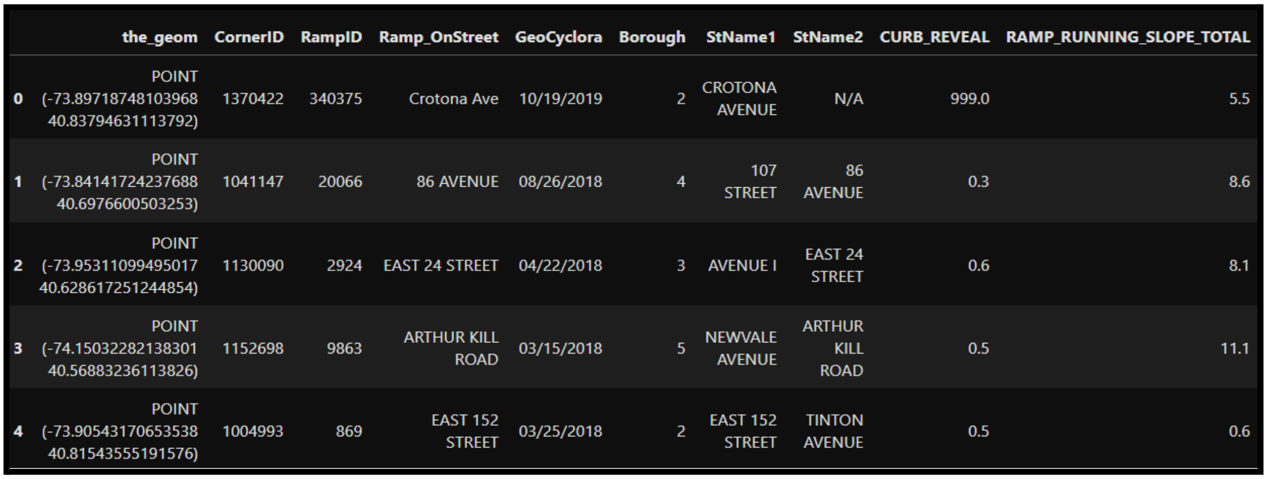
This dataset's objective is to make urban planning and maintenance easier by offering detailed information about pedestrian ramp infrastructure. This data is essential for ensuring that pedestrian ramps in various areas are kept up to code and meet accessibility requirements. This information can be used by city planners, accessibility coordinators, and maintenance teams to evaluate legal compliance, schedule new construction or improvements, and prioritize operations according to the data highlighted, based on issues and particular needs.

Let's take columns like **CURB_REVEAL** and **RAMP_RUNNING_SLOPE_TOTAL**, for example, those columns are directly related to determining whether ramps comply with legal accessibility requirements. Moreover, columns like **OBSTACLES_RAMP** and **PONDING** provide insight into possible safety concerns that might affect the ramps' usability. City authorities can pinpoint areas of concern within the columns and deploy resources as quickly as possible for repairs or improvements, and ultimately boost the overall efficiency and security of public pedestrian walkways by examining this dataset.

Data cleaning:

This data cleaning process was mainly used in Jupyter Notebook for a detailed analysis and well-documented approach. Users can use alternative programs like VS Code, and Google Colab for their cleaning processes and data visualization, enhancing the accuracy and effectiveness of the data cleaning.

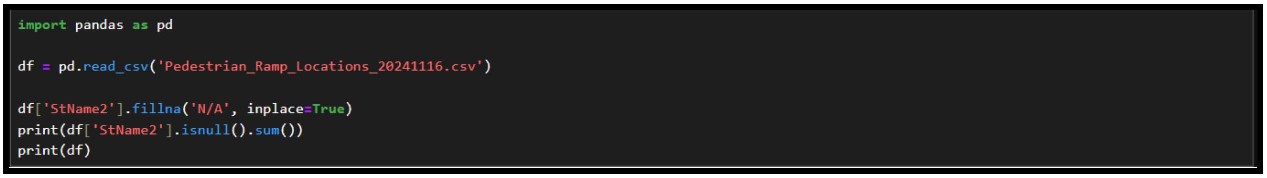
To clean and prepare the data for analysis, it is crucial to first understand the purpose of the dataset. Since our dataset "Pedestrian Ramp Locations" consists of ramp conditions in the NYC area, let's dive into the steps and Python code needed to clean this data. The "Pedestrian Ramp Locations" dataset consists of geographical and operational details about pedestrian ramps in a given area. This dataset also includes columns such as ramp location (latitude and longitude), installation date, condition, compliance status with accessibility standards, and possibly maintenance records shown in the table below.



	the_geom	CornerID	RampID	Ramp_OnStreet	GeoCyclora	Borough	StName1	StName2	CURB_REVEAL	RAMP_RUNNING_SLOPE_TOTAL
0	POINT (-73.89718748103968 40.83794631113792)	1370422	340375	Crotona Ave	10/19/2019	2	CROTONA AVENUE	N/A	999.0	5.5
1	POINT (-73.84141724237688 40.6976600503253)	1041147	20066	86 AVENUE	08/26/2018	4	107 STREET	86 AVENUE	0.3	8.6
2	POINT (-73.95311099495017 40.628617251244854)	1130090	2924	EAST 24 STREET	04/22/2018	3	AVENUE I	EAST 24 STREET	0.6	8.1
3	POINT (-74.15032282138301 40.56883236113826)	1152698	9863	ARTHUR KILL ROAD	03/15/2018	5	NEWVALE AVENUE	ARTHUR KILL ROAD	0.5	11.1
4	POINT (-73.90543170653538 40.81543555191576)	1004993	869	EAST 152 STREET	03/25/2018	2	EAST 152 STREET	TINTON AVENUE	0.5	0.6

Missing Values:

Handling missing values is an important step in the data-cleaning process because incomplete data can lead to incorrect results. Depending on the data's characteristics and the planned analysis, there are many approaches users may utilize. When the loss of data has little impact on the overall quality of the dataset, users may use common techniques involving removing data values or whole columns that have a large number of missing values. Use this method **.fillna** to fill in any blanks or missing values, as shown in the figure, which enables users to replace a textual or numeric value for any empty values in a column. Furthermore, users may include **.isnull** and **.sum** to replace empty numerical columns with NULL.

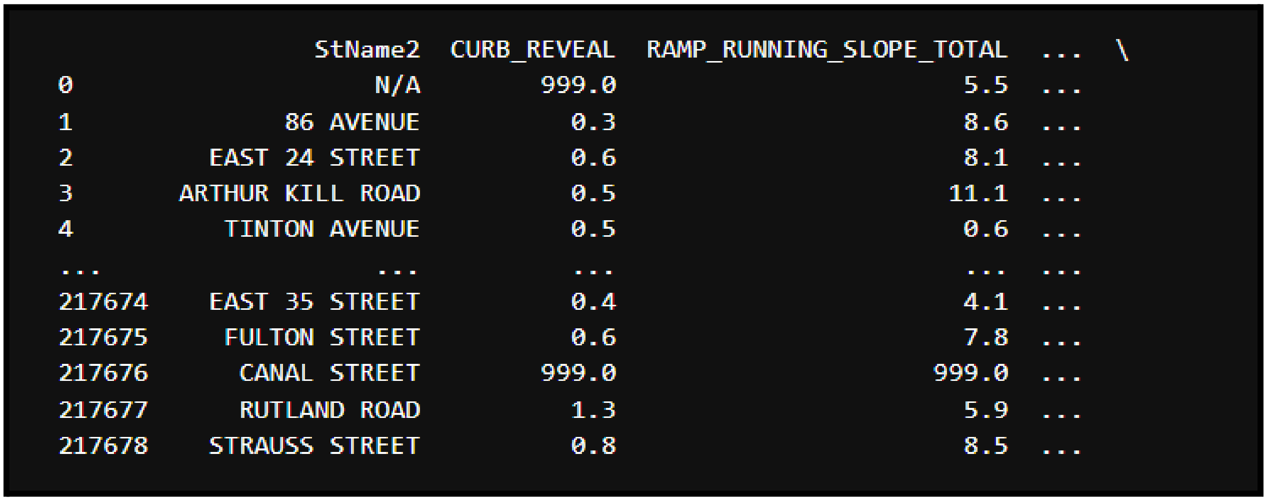


```
import pandas as pd

df = pd.read_csv('Pedestrian_Ramp_Locations_20241116.csv')

df['StName2'].fillna('N/A', inplace=True)
print(df['StName2'].isnull().sum())
print(df)
```

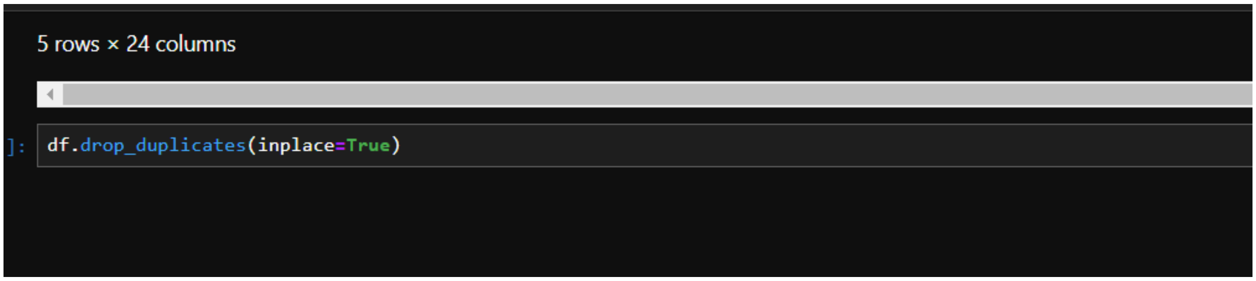
As you can see in the figure below, the function goes through each column to find missing entries like in column StName2, and replaces it with N/A if the column is a string. If the column is numerical it then replaces empty values with NULL, making it easier to perform visually without any syntax error.



```
                 StName2  CURB_REVEAL  RAMP_RUNNING_SLOPE_TOTAL  ...  \
0                    N/A        999.0                       5.5  ...
1              86 AVENUE          0.3                       8.6  ...
2         EAST 24 STREET          0.6                       8.1  ...
3       ARTHUR KILL ROAD          0.5                      11.1  ...
4         TINTON AVENUE          0.5                       0.6  ...
...                  ...          ...                       ...  ...
217674    EAST 35 STREET          0.4                       4.1  ...
217675     FULTON STREET          0.6                       7.8  ...
217676      CANAL STREET        999.0                     999.0  ...
217677      RUTLAND ROAD          1.3                       5.9  ...
217678    STRAUSS STREET          0.8                       8.5  ...
```

Removing Duplicate:

Removing duplicates from a dataset is a very important data-cleaning method that helps improve the quality of every analysis. Data analysis can be skewed by duplicate entries, causing results that are incorrect and misleading. When users use the **.drop_duplicates** function, it removes any rows of information that are repeated in the dataset. Below is a coded function on how to remove duplicates within a dataset, users can use the method if they choose to.

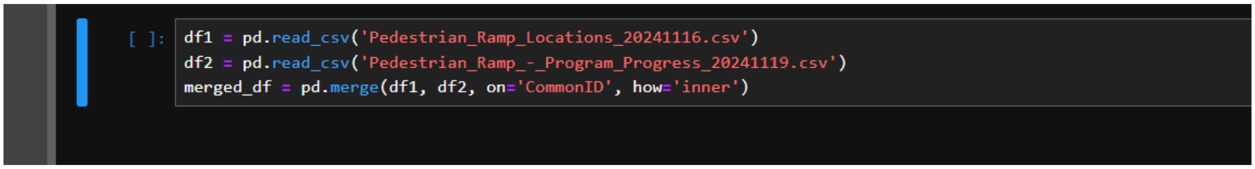


```
5 rows × 24 columns

]: df.drop_duplicates(inplace=True)
```

Merging Multiple Datasets:

In data analysis, merging several datasets is a common thing, especially when working with huge data sets from various places. Using a common identity or related columns like the primary key, this technique matches different datasets into a single, logical dataset. For example, within the 217 dataset for NYC Transportation, there are titles like Yellow Taxi that have datasets that date back to 2012. Users can merge all the Yellow Taxi datasets into one to make it easier to create a visualization. Users must keep in mind that one single dataset consists of 200,00 rows to 250,000 rows, merging these datasets could cause issues with the user's devices and software program. When merging datasets, the Panda's package within Python offers a robust function for combining datasets, including .merge, which is highly adaptable and allows inner, outer, left, and right joins.



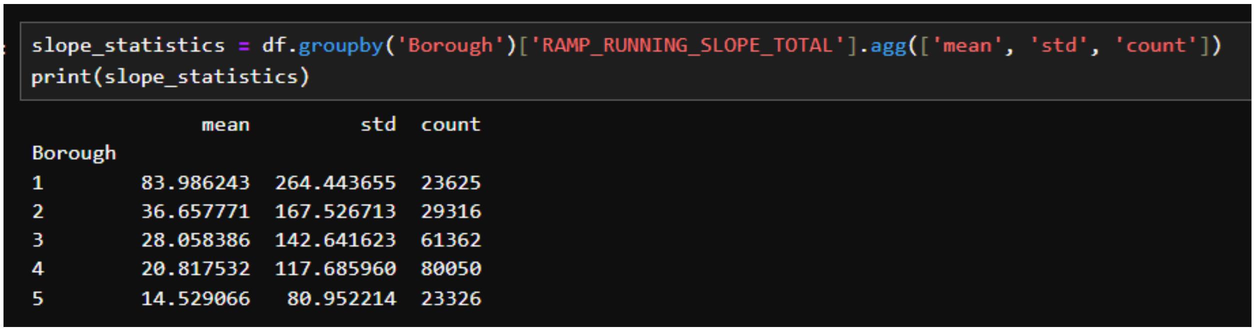
```
df1 = pd.read_csv('Pedestrian_Ramp_Locations_20241116.csv')
df2 = pd.read_csv('Pedestrian_Ramp_-_Program_Progress_20241119.csv')
merged_df = pd.merge(df1, df2, on='CommonID', how='inner')
```

Data Filtering:

Jupyter also allows users to filter data based on specific criteria, making it a useful tool for focusing on specific portions of big datasets. For example, when analyzing the "Pedestrian Ramp Locations" dataset, users can use filters to identify ramps with certain features, such as those with specific obstacles, situated in specific boroughs, or falling within defined width and length dimensions.

Aggregate:

The aggregate data frame allows users to summarize specific data by grouping them into rows based on a particular column. For instance, using pandas and Matplotlib, users can compute and plot aggregate statistics like the mean, standard deviation, and count of ramp slopes for each borough. This method allows users to have a visual representation of their data, which can indicate locations where the ramp design may not satisfy the accessibility needs or where there are multiple errors within the slope design, which could potentially cause issues or mistakes within a planning section of the upgrade or change on those areas of improvement. Below is an example of how to perform an aggregate for your dataset, as you can see users must first choose the columns they want to focus on and what aggregate the user suggests to do.

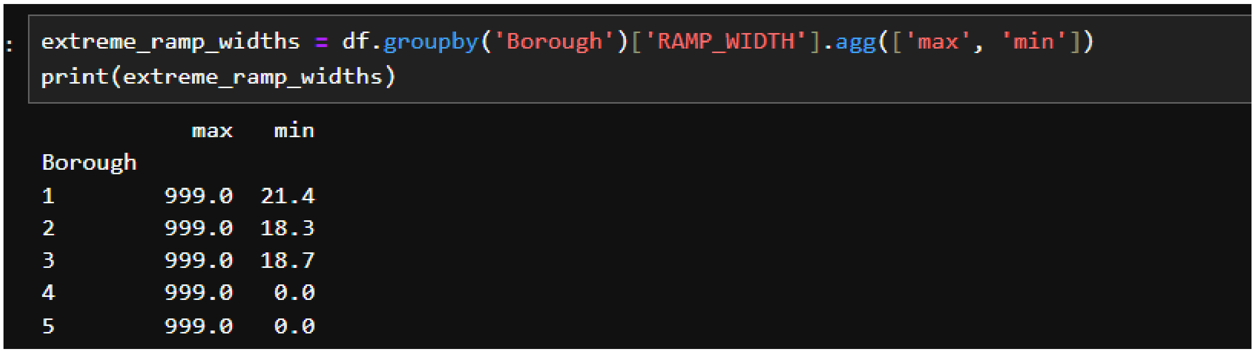


```
slope_statistics = df.groupby('Borough')['RAMP_RUNNING_SLOPE_TOTAL'].agg(['mean', 'std', 'count'])
print(slope_statistics)
```

```
              mean         std  count
Borough
1        83.986243  264.443655  23625
2        36.657771  167.526713  29316
3        28.058386  142.641623  61362
4        20.817532  117.685960  80050
5        14.529066   80.952214  23326
```

Min and Max:

Users can also perform accurate analyses and visualizations to improve the efficiency of exploring the maximum and minimum of the NYC Transportation ramp widths across different Boroughs. By performing these tasks, it allows researchers or individuals to evaluate the compliance with the accessibility regulations which is critical in assessing compliance with accessibility regulations and any similarity in urban infrastructure all across NYC. Below is an example of how to perform a min and max for your dataset, finding the min and max of a specific column follows the same steps in the aggregate section however there is a difference in what you put in the **.agg** method.



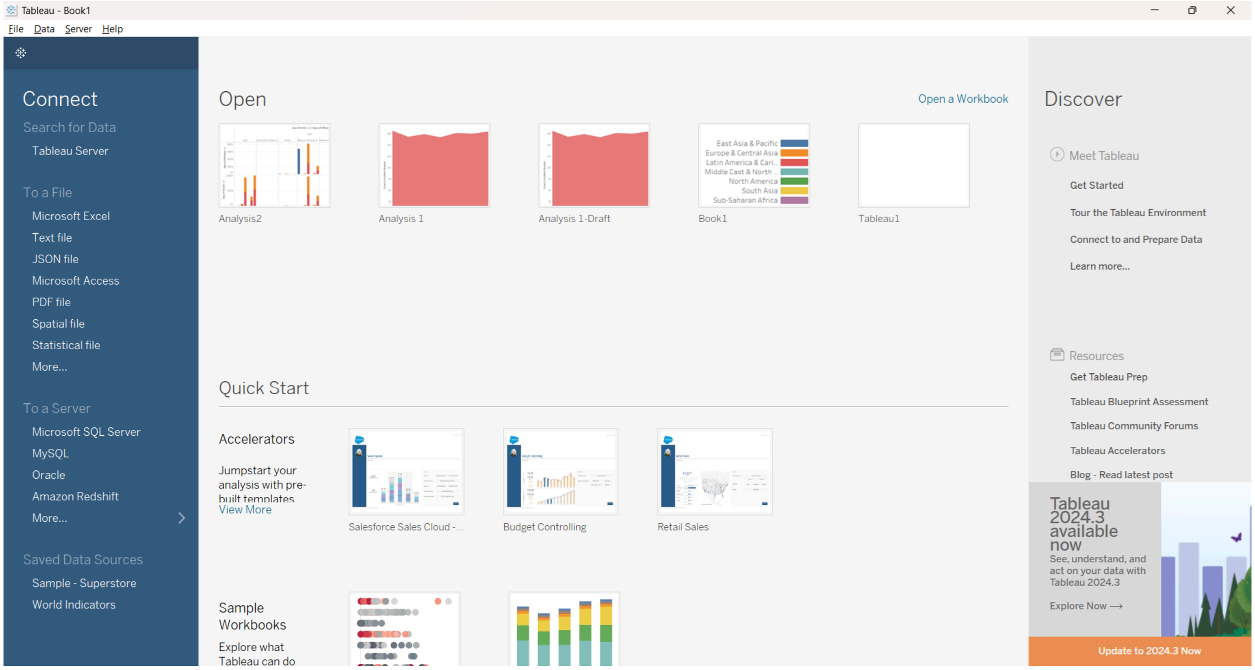
```
extreme_ramp_widths = df.groupby('Borough')['RAMP_WIDTH'].agg(['max', 'min'])
print(extreme_ramp_widths)
```

```
           max   min
Borough
1        999.0  21.4
2        999.0  18.3
3        999.0  18.7
4        999.0   0.0
5        999.0   0.0
```

Data Visualization Tools:

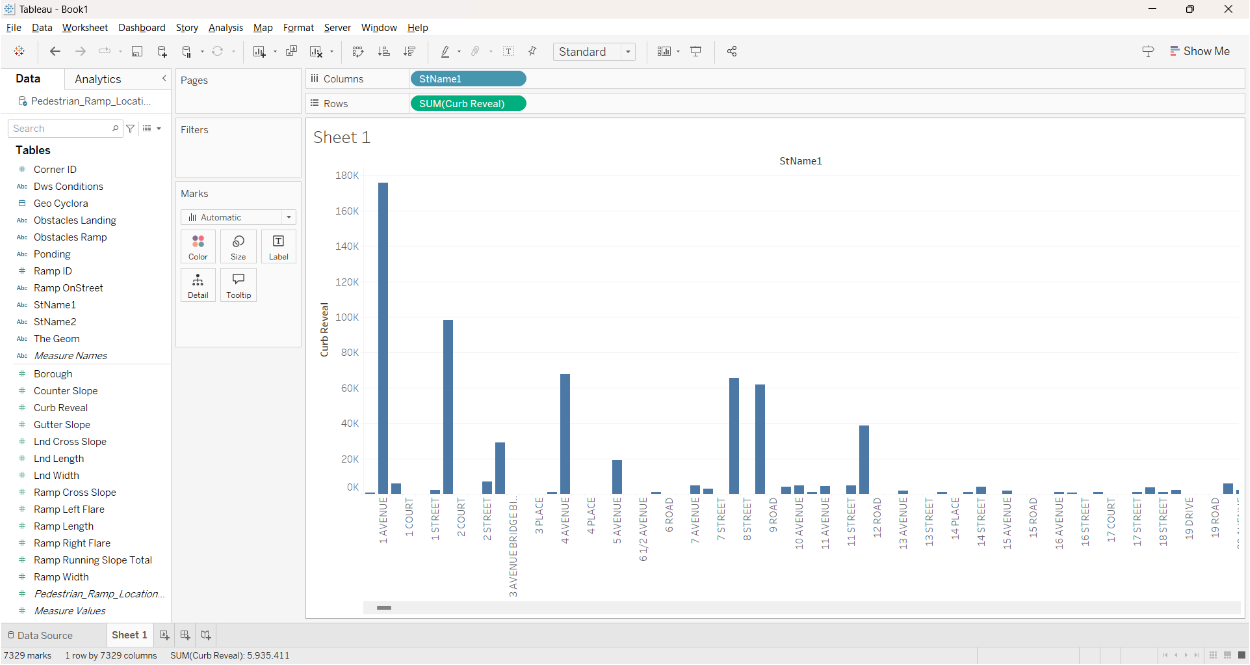
Tableau:

Users can use Tableau to improve the visualization of the "Pedestrian Ramp Locations" dataset or any other dataset, revealing deeper insights and presenting data in a more digestible and visually appealing format. Tableau offers users a variety of options of what format of dataset they want to import from shown on the left-hand side as well as what servers the user wants to use. Furthermore, if the user finds it hard to use Tableau, the program offers a tutorial guide for the user to learn the various techniques within Tableau shown on the right-hand side of the home page.



Furthermore, Tableau allows for flexible interaction with data by utilizing filters and accessibility capabilities. This allows users, such as researchers or students, to adjust the view to certain settings, allowing for particular analysis, as demonstrated below. The tool also allows for the automatic creation of calculated fields, which enhances the functionality of the dataset by allowing additional metrics to be generated immediately within the visualization tool. These visualizations can be used to create dashboards, which provide users with a full and interactive report to examine.

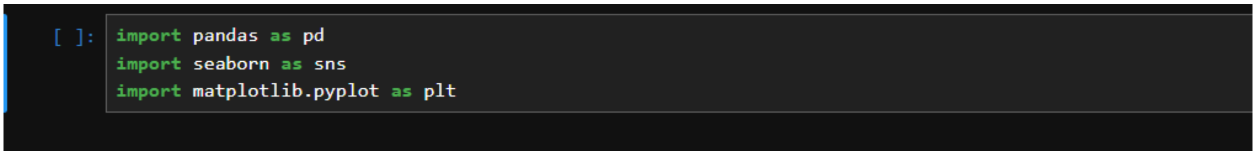
Let's take the analysis we worked on and import it to Tableau. Once the user imports the file, they will visually see a table of contents on the left-hand side. These contents are already formatted within Tableau based on the program suggestion, allowing users to choose what table/title goes on columns or rows. Once the user selects their column and rows shown in the figure below, Tableau will give the user a list of various visualizations. Moreover, the program already calculates the sum of Curb Reveal so users do not have to do it.



7329 marks 1 row by 7329 columns SUM(Curb Reveal): 5,935,411

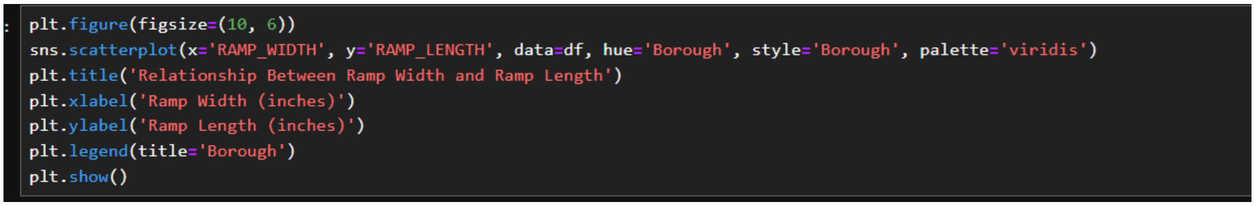
Jupyter:

Jupyter Notebook is another great tool for data visualization, it allows users to run their Python code in an interactive environment and produces visual graphs based on the code written. It is more of an adaptive method of data analysis that uses a combination of coding and graphics. Users would mainly use libraries like Matplotlib, Seaborn, Pandas, or Plotly for their visualization within Jupyter Notebook. These libraries are excellent for examining data patterns and spotting trends.



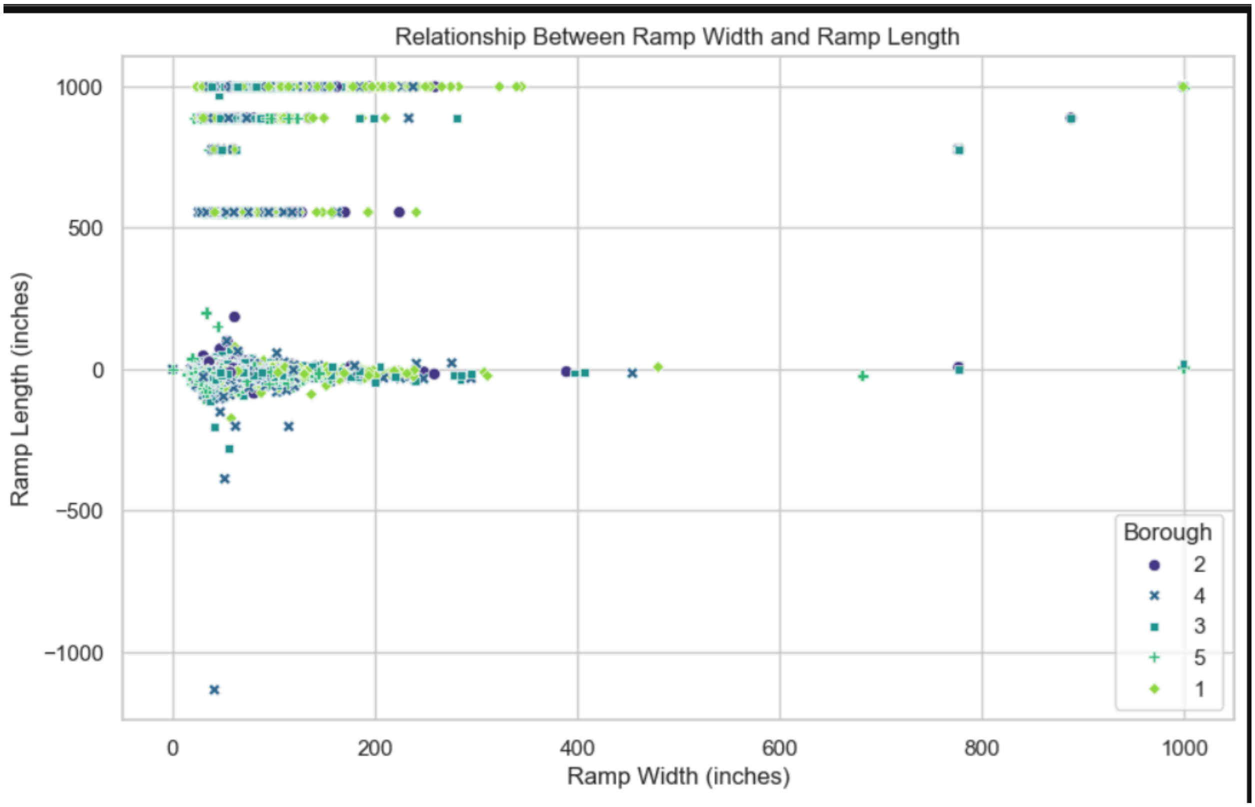
```
[ ]: import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
```

Furthermore, by utilizing Python code in our visualization, we can highlight any potential trends in ramp design as well as uncover variances in ramp widths among boroughs. Let's make a visualization based on the relationship of the ramp width and the ramp length. Within the visualization code users can use different graphs based on their preferences but for this demonstration, we would be using .scatterplot for our visualization. Furthermore, Jupyter allows users to write a program for the format of the graph like the title for the x-axis and y-axis shown in the figure below.



```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='RAMP_WIDTH', y='RAMP_LENGTH', data=df, hue='Borough', style='Borough', palette='viridis')
plt.title('Relationship Between Ramp Width and Ramp Length')
plt.xlabel('Ramp Width (inches)')
plt.ylabel('Ramp Length (inches)')
plt.legend(title='Borough')
plt.show()
```

As a result, the scatter plot of ramp width against ramp length provides plenty of important findings. To start with, the plot displays clusters of data points indicating common sizes for ramps, showing consistency in their design. On the other hand, if the data points are scattered, it may indicate that there isn't enough consistency in ramp sizes, limiting usability. Also, the use of colors by borough allows for geographical examination, for example, one borough may have more larger ramps than another, showing the differences in local laws.



Categorization of NYC Transportation Datasets by Data Types and Applications

Evaluating Transportation Patterns and Their Implications

1. Taxi and For-Hire Vehicle Data

- This category includes datasets about Yellow Taxi, Green Taxi and For-Hire Vehicle (FHV) trips, and high volume FHV trip data from ridesharing services. The datasets include the following elements of trips: where and when trips started and ended, how many passengers were in the vehicle, the distance traveled by the vehicle, and how much the passenger paid.
- It will also allow researchers to analyze the demand trend of transportation in NYC-from traditional taxis to ride-hailing services like Uber and Lyft. These datasets expose peak usages, help in evaluating the effects of FHV services on city traffic, and offer a geographic study of popular pickup and drop-off points.

Exploring Research Opportunities with Subsets of Correlated Datasets

Yellow Taxi Trip Data

- ❖ Traffic Flow Analysis: Understanding peak hours for taxi usage to optimize traffic management and reduce congestion.
- ❖ Economic Impact Study: Analyzing the contribution of taxi services to the local economy.
- ❖ Environmental Research: Estimating emissions from taxi services and exploring green alternatives.
- ❖ Urban Planning: Assessing the need for taxi stands and improving public transport routes.

Green Taxi Trip Data

- ❖ Accessibility Research: Studying the availability of transportation in underserved areas.
- ❖ Public Health Studies: Examining the impact of transportation on community health and mobility.
- ❖ Infrastructure Development: Planning for new green taxi routes and service improvements.

For-Hire Vehicle Trip Data:

- ❖ Market Analysis: Assessing the demand for for-hire vehicles and forecasting future trends. (ex: tesla taxi future)
- ❖ Service Quality Evaluation: Evaluating customer satisfaction and service quality of for-hire vehicles.
- ❖ Policy Development: Formulating policies to regulate and improve for-hire vehicle services.

2. Bus and Transit Data

- This dataset contains bus performance measures, breakdowns, and delays, bus lanes, and transportation arrangements of PreK students. Each dataset provides information on particular aspects of the service and transit infrastructure.
- This would consequently underscore the efficiency and reliability of public transport for the researchers. Breakdown and delay analysis, together with the performance of bus lanes, is therefore allowed to be carried out in focused ways that allow improvements to be made in transit services. Moreover, comprehension of PreK transportation arrangements does make assessments of transit needs for different age groups possible, thus optimizing school bus routes.

Exploring Research Opportunities with Subsets of Correlated Datasets

Bus Breakdowns and Delays

- ❖ Operational Efficiency Studies: Identifying patterns in bus breakdowns to improve maintenance schedules.
- ❖ Service Reliability Research: Evaluating the impact of delays on passenger satisfaction and punctuality.
- ❖ Transit Policy Making: Developing policies to mitigate delays and enhance bus service reliability.

3. Bicycle Data

- Datasets in this category include bicycle counters, specific location counts, and bicycle parking infrastructure. They let the general public learn about usage intensity, the count in one or another borough, and other interesting subjects, such as the availability of bike parking.
- This category outlines the benefit to the researcher using data to estimate cycling popularity, determine points of high bicycle traffic, and assess the sufficiency of bike parking and infrastructure. This informs sustainable urban mobility, as the result of the

analysis will point to deficiencies in cycling infrastructure that require planning and policy measures promoting safe cycling.

Exploring Research Opportunities with Subsets of Correlated Datasets

Bicycle Counts

- ❖ Sustainability Studies: Promoting cycling as an eco-friendly mode of transportation. (electric bicycles/scooters)
- ❖ Safety Research: Analyzing accident rates and safety measures for cyclists.
- ❖ Urban Mobility Planning: Planning for bike lanes and cycling infrastructure development.

4. Pedestrian and Walkability Data

- The data includes accessible pedestrian signal locations, pedestrian demand zones, pedestrian ramp locations, and the status of improvements to pedestrian infrastructure. It also identifies complaints about pedestrian ramps.
- The category will help analyze walkability and accessibility in NYC. Researchers can look at pedestrian data to see where infrastructure improvements should be provided that include accessible signals or ramps. The analysis using this set of information helps in making the city more friendly for pedestrians and helps in improving accessibility to people with disabilities.

5. Traffic and Roadway Data

- This category adds greater depth to the traffic flow and roadway infrastructure by adding traffic volume counts, bridge conditions, bridge incident data, and requests for traffic signals and electronic signage.
- Researchers can get an idea from traffic volumes and patterns. Determine hotspots where congestion is commonplace and roadway conditions. This data identifies system bottlenecks, accesses the efficiency of traffic control measures, and develops strategies to decrease congestion and improve roadway safety.

Exploring Research Opportunities with Subsets of Correlated Datasets

EZ Pass Readers

- ❖ Traffic Pattern Analysis: Studying toll usage to manage traffic flow and reduce bottlenecks.
- ❖ Revenue Forecasting: Projecting toll revenue for budgeting and financial planning.

- ❖ Infrastructure Planning: Improving toll collection systems and road maintenance schedules.

6. Parking Data

- Included in this category are datasets such as but not limited to permits, parking violations, and meter locations. The sets range in topics from annual permits to disability parking permits, clergy parking, temporary permits, and agencies' issuance of parking permits.
- Parking data informs studies on parking patterns across cities in terms of supply, control, and access. This would allow researchers to gain an understanding of parking demand across town, assess the adequacy of parking restrictions, and achieve equitable distribution in accessing parking facilities, including accessible parking for persons with disabilities.

7. Surveys and Mobility Studies

- The datasets in this category result from annual mobility surveys capturing travel behaviors, mode choices, and demographic characteristics across NYC. They are day-based, household-based, trip-based, and vehicle datasets emanating from different years.
- These are the behavioral insights for analysts into how and why NYC residents travel the way they do. It shall, therefore, enable researchers to assess the adoption of new modes of travel, the socio-economic influence on travel patterns, and the responses to changes in transit or citywide events like the COVID-19 pandemic.

Exploring Research Opportunities with Subsets of Correlated Datasets

Citywide Mobility Surveys

- ❖ Demographic Studies: Understanding transportation habits across different demographics.
- ❖ Behavioral Analysis: Examining how different factors influence transportation choices.
- ❖ Policy Impact Assessment: Evaluating the effectiveness of transportation policies on mobility.

8. Environmental and Public Health Data

- The data herein, inclusive of the Vision Zero Base Report, outlines safety and public health aspects relative to transportation. This provides data on traffic injuries, fatalities, and locations with high accident rates.

- This will, in turn, help the public health and safety researchers understand where high-risk areas are, what forms of safety interventions work, and so on, and develop policies to have safer roads and better quality of life for the NYC residents.

9. Miscellaneous Data

- This category includes, but is not limited to street sign orders, truck routes, structure changes, and lost property contact information. It also ranges to historical information such as driver training records and vehicle permits issued.
- The miscellaneous data serves as background information on the long-term trends and minor yet deeply influential features in managing metropolitan infrastructure. These datasets may be studied by researchers for their regularity of compliance, historical comparison, and continuity in transportation services that offer an all-inclusive perspective of NYC transportation at large.

Overall Benefits for Researchers

Taken together, these nine categories provide a rich foundation for a wide range of transportation-related studies. These datasets can be combined by researchers for comprehensive analyses, such as evaluating multi-modal transportation options, studying environmental impacts, and formulating policies for sustainable urban mobility. The layered data will allow for the study of complex relationships among the many transportation modes, infrastructure, and public behavior in NYC to foster data-informed urban planning.

User Guide Update Potential

Limitations and Updates

Please note that the NYC transportation data website is consistently updated; new datasets are being contributed and existing ones relevant to the dates are updated on a daily basis, even as this user guide is produced. As a result, the Google sheet may not reflect the latest changes or additions. Viewers are encouraged to proceed with this awareness in mind. However, this user guide remains relevant for navigating the website, accessing and understanding the core attributes behind the datasets.

Version History

Version	Time
1.0	December 2024