Data Table Schema

uber\_trips\_2014

Trip data (pickup times, pickup coordinates, etc.) from Uber vehicles in 2014.

*~4.5 million rows & 4 columns.* Size: ~30MB zipped, ~200MB unzipped.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| **pickup\_datetime** | STRING | Time of pickup (format mm/dd/yyyy hh:mm:ss and mm/dd/yy hh:mm) |
| **pickup\_latitude** | FLOAT | Latitude coordinate of pickup location |
| **pickup\_longitude** | FLOAT | Longitude coordinate of pickup location |
| **base** | STRING | Base company affiliated with the Uber ride |

uber\_trips\_2015

Trip data (pickup times, pickup location IDs, etc.) from Uber vehicles in 2015.

*~14 million rows & 4 columns.* Size: ~65MB zipped, ~550MB unzipped.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| **pickup\_datetime** | STRING | Time of pickup (format yyyy-mm-dd hh:mm:ss) |
| **pickup\_location\_id** | INTEGER | Taxi zone ID of pickup location |
| **dispatch\_base** | STRING | Base company that dispatched the Uber ride |
| **affiliate\_base** | STRING | Base company affiliated with the Uber ride |

demographics

Demographic data (population, age, income, etc.) organized alphabetically by NTA.

*188 rows & 33 columns.* Size: ~0.1MB.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| **nta\_name** | STRING | Name of NTA |
| **borough** | STRING | Borough that NTA is located in |
| **nta\_code** | INTEGER | Identifying code for NTA |
| **population** | INTEGER | Total number of people in NTA |
| **age brackets (14 total)** | INTEGER | Number of people in given age bracket |
| **median\_age** | FLOAT | Median age of people in NTA |
| **people\_per\_acre** | INTEGER | Number of people per acre |
| **households** | INTEGER | Total number of households in NTA |
| **income brackets (10 total)** | INTEGER | Number of households in given income bracket |
| **median\_income** | INTEGER | Median household income |
| **mean\_income** | INTEGER | Mean household income |

geographic

Data about the shape of each NTA (latitude and longitude coordinates, in order).

*9,302 rows & 195 columns.* Size: ~4MB.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| **nta\_code sections (195 total)** | FLOAT | Alternating longitude and latitude coordinates, in order, of the vertices of the polygon shape that define the boundaries of the given NTA code |
|  |  |  |

green\_trips

Trip data (pickup/dropoff times, pickup/dropoff locations) from NYC green boro taxis. *Note: in order to keep the dataset size manageable, the provided data is a 20% unbiased sample of the raw data. If using trip count metrics, remember to multiply quantities by 5 to approximate the actual data.*

*~3.5 million rows & 9 columns.* Size: ~140MB zipped, ~400MB unzipped.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| **pickup\_datetime** | STRING | Time of pickup (format yyyy-mm-dd hh:mm:ss) |
| **dropoff\_datetime** | STRING | Time of dropoff (format yyyy-mm-dd hh:mm:ss) |
| **pickup\_longitude** | FLOAT | Longitude coordinate of pickup location |
| **pickup\_latitude** | FLOAT | Latitude coordinate of pickup location |
| **dropoff\_longitude** | FLOAT | Longitude coordinate of dropoff location |
| **dropoff\_latitude** | FLOAT | Latitude coordinate of dropoff location |
| **passenger\_count** | INTEGER | Number of passengers on the ride |
| **trip\_distance** | FLOAT | Miles traveled during ride in miles |
| **total\_amount** | FLOAT | Dollars spent on ride |

mta\_trips

Trip data (time intervals, entries, exits, etc.) from NYC public subway turnstiles.

*~7.5 million rows & 10 columns.* Size: ~50MB zipped, ~700MB unzipped.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| **station** | STRING | Name of station |
| **line\_name** | STRING | Name of subway line |
| **division** | STRING | Transit company that line originally belonged to |
| **audit\_type** | STRING | Measurement type – default is “REGULAR” |
| **unit\_id** | STRING | Unique ID of the turnstile measurement unit/device |
| **datetime** | STRING | Time of measurement (format mm/dd/yyyy hh:mm:ss zzz) |
| **new\_entries** | INTEGER | Turnstile entrances in given four-hour period |
| **new\_exits** | INTEGER | Turnstile exits in given four-hour period |
| **latitude** | FLOAT | Latitude coordinate of turnstile |
| **longitude** | FLOAT | Longitude coordinate of turnstile |

weather

Temperature and precipitation data for three areas in the NYC metropolitan area.

*2,190 rows & 10 columns.* Size: ~0.1MB.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| **date** | STRING | Date of measurement (format mm/dd/yy) |
| **max\_temp** | INTEGER | Maximum temperature in Fahrenheit |
| **min\_temp** | INTEGER | Minimum temperature in Fahrenheit |
| **avg\_temp** | FLOAT | Average temperature in Fahrenheit |
| **precipitation** | FLOAT | Total precipitation in inches when reduced to liquid form |
| **snowfall** | FLOAT | Total snowfall in inches |
| **snow\_depth** | INTEGER | Depth of snow on the ground in inches |
| **location** | STRING | Name of area |
| **latitude** | FLOAT | Latitude of area |
| **longitude** | FLOAT | Longitude of area |

yellow\_trips

Trip data (pickup/dropoff times, pickup/dropoff locations) from NYC yellow medallion taxis. *Note*: *in order to keep the dataset size manageable, the provided data is a 5% unbiased sample of the raw data. If using trip count metrics, remember to multiply quantities by 20 to approximate the actual data.*

*~8 million rows & 9 columns.* Size: ~260MB zipped, ~800MB unzipped.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| **pickup\_datetime** | STRING | Time of pickup (format yyyy-mm-dd hh:mm:ss) |
| **dropoff\_datetime** | STRING | Time of dropoff (format yyyy-mm-dd hh:mm:ss) |
| **pickup\_longitude** | FLOAT | Longitude coordinate of pickup location |
| **pickup\_latitude** | FLOAT | Latitude coordinate of pickup location |
| **dropoff\_longitude** | FLOAT | Longitude coordinate of dropoff location |
| **dropoff\_latitude** | FLOAT | Latitude coordinate of dropoff location |
| **passenger\_count** | INTEGER | Number of passengers on the ride |
| **trip\_distance** | FLOAT | Miles traveled during ride in miles |
| **total\_amount** | FLOAT | Dollars spent on ride |

zones

Information about each ride pickup zone in the NYC metropolitan area.

*263 rows & 5 columns.* Size: ~0.1MB.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| **location\_id** | INTEGER | ID of zone |
| **borough** | STRING | Name of borough zone is located in |
| **zone** | STRING | Name of zone |
| **service\_zone** | STRING | Primary car service in given zone |
| **nta\_code** | STRING | Code of NTA that zone is located in |