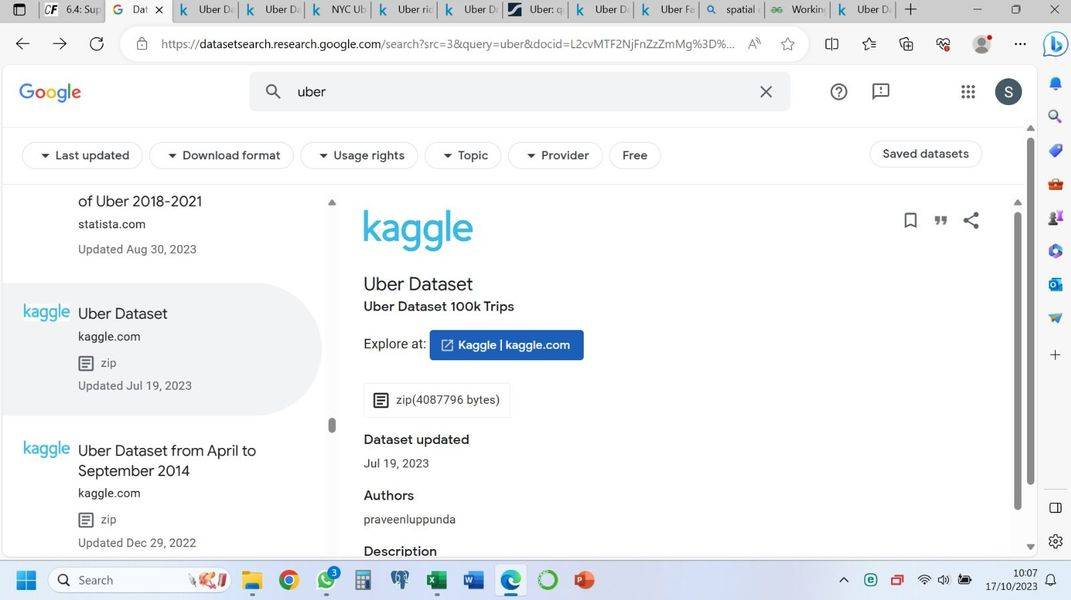
6.1: Sourcing Open Data

**Data Source:**

Data link: [www.kaggle.com/datasets/praveenluppunda/uber-dataset](http://www.kaggle.com/datasets/praveenluppunda/uber-dataset)



Last updated: 19/07/2023

Data source: this data set is owned by Uber, which collected it from its trip records of NYC's Taxi and Limousine Commission (TLC). It can be considered as internal data, both administrative and usage. Reliable, although there are some issues to be looked at in the data.

Data collection: it was collected by both driver and passengers apps/API, as well as other Uber administrative data.

Data contents (overview): this data set includes data about trips, both logistical (pick up and drop off date and time, length of trip, driver id) and economical (fare of trip, form of payment, tips, tolls, total amount and whether there was a dispute about the fare).

Data limitations: this is a subset of the original data set, containing data about 100k trips, limiting the scope of vendors and dates to be examined.

Data relevance: this data is relevant as it includes enough information about Uber service use by the people of NY city. It has been recently updated, so it is also timely.

Why I chose this data set: living in a big city, I sometimes need to use private transportation (especially for short trips), as it is faster and more convenient than public transportation. However, using my own car comes with the problem of finding parking and formal/regulated taxis are expensive to use. Private hire is a better option, and amongst the options available Uber is by far the easiest, fastest and more convenient to use. Additionally, it is normally also cheaper. Therefore, I do have a queen interest into looking into Uber’s data: understanding how its is normally used by other people, how fares work and whether they depend on time or length of trip, are points worth analysing.

**Data profile**

Data types

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column name** | **Time variant?** | **Structured?** | **Qualitative/quantitative?** | **Type data (nominal)?** |
| VendorID | Invariant | Structured | Qualitative | Nonimal |
| Pick up date | Variant | Structured | Quantitative | Continuous |
| Pick up time | Variant | Structured | Quantitative | Continuous |
| Dropoff date | Variant | Structured | Quantitative | Continuous |
| Dropoff time | Variant | Structured | Quantitative | Continuous |
| Passenger count | Invariant | Structured | Quantitative | Discrete |
| Trip distance | Invariant | Structured | Quantitative | Continuous |
| Pickup longitude | Invariant | Structured | Quantitative | Continuous |
| Pickup latitude | Invariant | Structured | Quantitative | Continuous |
| RatecodeID | Invariant | Structured | Qualitative | Nominal |
| Store and fwd flag | Invariant | Structured | Qualitative | Nominal |
| Dropoff longitude | Invariant | Structured | Quantitative | Continuous |
| Dropoff latitude | Invariant | Structured | Quantitative | Continuous |
| Payment type | Invariant | Structured | Qualitative | Nominal |
| Fare amount | Variant | Structured | Quantitative | Continuous |
| Extra | Invariant | Structured | Quantitative | Continuous |
| Mta tax | Invariant | Structured | Quantitative | Continuous |
| Tip amount | Variant | Structured | Quantitative | Continuous |
| Tolls amount | Invariant | Structured | Quantitative | Continuous |
| Improvement surcharge | Invariant | Structured | Quantitative | Continuous |
| Total amount | Variant | Structured | Quantitative | Continuous |

Initial statistical figures:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column name** | **Min** | **Max** | **Mean** | **Median** | **St. Dev.** |
| Trip distance | 0 | 184.4 | 3.03 | 1.67 | 3.80 |
| Fare amount | 0 | 200 | 13.23 | 9.5 | 11.09 |
| Extra | 0 | 4.5 | 0.1 | 0 | 0.2 |
| Mta Tax | 0 | 0.5 | 0.49 | 0.5 | 0.03 |
| Tip amount | 0 | 125.88 | 1.87 | 1.36 | 2.6 |
| Tolls amount | 0 | 25.54 | 0.36 | 0 | 1.52 |
| Improvement surcharge | 0 | 0.3 | 0.299 | 0.3 | 0.005 |
| Total amount | 0 | 246.11 | 16.37 | 11.8 | 13.92 |

NOTE: while I removed the biggest values from the dataset (10 in total) as they greatly skew the data (to the left), the mean and median values show that there are still outliers on the right side of the data. However, these seem to be genuine, so – at the moment – I decided not to remove them from the dataset. During further analysis, I plant to divide these values into two datasets, analysing them separately.

**Limitations and ethics:**

Ethics: this data is obtained from NYC taxy and limousine services, and owned by Uber. This dataset is a subset from the bigger dataset. There is no PII information on this dataset, and I can observe no ethical issues involved in this data collection (which is done by users and drivers apps/APIs, and Uber’s administrative and usage data).

Limitations: as this is a subset, it contains information collected from 2 vendors (1 and 2), and from 10 days (from 1st to the 10th of October 2016). While it is great for an initial analysis, I would have looked into the bigger dataset for further and deeper conclusions.

Clarifying questions:

1. How much is the average length trip on Uber?
2. How much is the average spent on Uber trips?

Adjoining questions:

1. Is there a difference in spending or trip length between day and night hours?
2. Does the pattern for spending and length of trip differ depending on the date and time in the month?

Funnelling questions:

1. What is the busiest time of the day for trips?
2. What are the busiest days on the week for trips?
3. What are the more profitable days and hours?

Elevating questions (for further study, in a bigger dataset):

* How does NYC differ from other big American cities in performance?
* How do Vendors in this dataset differ from other uber drivers in performance for both trip length and revenue?