**AUTOLIB DATASET DATA INTEGRITY REPORT:**

Ensuring data integrity of the data that one is working with is vital because the validity and accuracy of decisions/ analyses made is tied to that of the data used. How you intend to of all of the data you generate.

**ALCOA TEST:**

We shall subject the data generated to the ALCOA TEST to ensure its reliability and completeness. The ALCOA test measures various characteristics of data which are as follows:

* **Attributable:**

Information should be captured in the records so that it is uniquely identified as having been executed by the originator of the data (e.g. a person or computer system).

Our data was downloaded from Dalberg insights, who in turn had extracted the data from opendataparis.com, where the Autolib availability information was available in real-time.

During the collection periods, on some occasions the download failed, which created dome gaps in the data. These gaps are represented on Figures 1 and 2 . Except for an important gap from November 11 to November 14, 2017, missing files were rare.

The accessed database was the following:

*Name:* Stations Autolib: Disponibilité en temps réel

*Producer* : Autolib

*Date*: Provided

*License* : Open Database License

* **Legible, traceable, and permanent:**

Data should be readable, understandable, and should allow a clear picture of the sequencing of steps or events in the record so that all activities conducted can be fully reconstructed by the people reviewing these records at any point during the records retention period

* **Contemporaneous:**

Data should be recorded at the time when the data are generated or observed.

* Data was extracted from opendataparis.com, where the Autolib availability information was available in real-time.
* Data was collected every minute from October 30, 16:59 PM to July 31, 2018, 23:59 PM
* **Original (or “True Copy”):**

Data should be in the format in which it was originally generated, preserving the integrity (accuracy, completeness, content and meaning) of the record. Original data include the first or source capture of data or information and all subsequent data required to fully reconstruct the conduct of the activity.

* **Accurate:**

Data should be correct, truthful, complete, valid and reliable.

* During the collection periods, on some occasions the download failed, which created some gaps in the data. Except for an important gap from November 11 to November 14, 2017, missing files were rare- The period with the missing information does not affect the integrity of our data since it does not relate to the period which we are concerned with- April 2018)
* There were 549 missing records in the month of April 2018. Compared with the millions of entries per minute) this is a negligible amount (*percentage 1.27%)* which would not affect the validity of our results
* There were no missing values in our data save for two columns, which were dropped in our data cleaning, therefore data passed the completeness test.
* Other data cleaning procedures conducted to ensure the validity and reliability included removing of outliers and duplicated entries.

**Limitation**: There was official description of the fields was published together with the data and Dalberg (our source) had to make assumptions as to the meaning of the fields collected.

**CONCLUSION:**

**The Dataset passes the ALCOA test and is consequently reliable and complete**