**Data Cleaning Documentation**

**Objective**

To clean and standardize the dataset containing daily Metrobus flow information, originally in Spanish, for further analysis and usability.

**Steps Taken**

**1. Translate Column Names**

* Original column names in Spanish:
  + fecha → Date
  + anio → Year
  + mes → Month
  + linea → Line
  + afluencia → Influx
* Added a new column Stations in the cleaned dataset.

**2. Translate Data**

* Translated the mes (Month) column values from Spanish to English, but retained the lowercase format for consistency.
  + Example: Julio → julio

**3. Format Data Consistently**

* Standardized capitalization in the Line column:
  + Changed Línea to line (lowercase).
  + Example: Línea 1 → line 1.

**4. Add New Column**

* Introduced the Stations column in the cleaned dataset:
  + Populated some rows with station names (e.g., "General Hospital, Miguel Hidalgo").
  + Retained NaN (missing values) for other rows.

**5. Preserve Missing Data**

* Retained missing values (NaN) in the Influx column and for rows without corresponding station information.

**Validation**

After cleaning, the dataset structure is as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Date** | **Year** | **Month** | **Line** | **Stations** | **Influx** |
| 2005-07-26 | 2005 | julio | line 1 | NaN | 3032667.0 |
| 2005-07-26 | 2005 | julio | line 5 | NaN | NaN |
| 2005-07-26 | 2005 | julio | line 2 | NaN | NaN |
| 2005-07-26 | 2005 | julio | line 3 | General Hospital, Miguel Hidalgo | NaN |
| 2005-07-26 | 2005 | julio | line 6 | NaN | NaN |

**Data Cleaning Documentation – Covid 19**

**Objective**

To clean and standardize the COVID-19 Metrobus Traffic Index dataset, originally in Spanish, for further analysis and usability.

**Steps Taken**

**1. Translate Column Names**

* Original column names in Spanish:
  + fecha → covid\_date
  + hora → covid\_hour
  + linea → covid\_line
  + indice → covid\_influx
* Added a new column covid\_station in the cleaned dataset.

**2. Format Column Values**

* Standardized the naming convention for lines in the covid\_line column:
  + Example: linea → line (lowercase).
  + Retained the original structure of the line numbers (e.g., line 1, line 2).

**3. Add New Column**

* Introduced the covid\_station column in the cleaned dataset:
  + Populated rows with known station names where applicable.
  + Example: line 3 mapped to "General Hospital, Miguel Hidalgo."
  + For some rows, added multiple station names where relevant (e.g., "Merced, Benito Juarez, Miguel Hidalgo").
  + Retained NaN (missing values) for rows where station information was unavailable.

**4. Preserve Original Data**

* Retained the numeric values in the covid\_influx column without transformation.
* Ensured that missing data (NaN) in the original dataset was preserved in the cleaned version.

**Validation**

After cleaning, the dataset structure is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **covid\_date** | **covid\_hour** | **covid\_line** | **covid\_station** | **covid\_influx** |
| 2020-03-11 | 00:00:00 | line 1 | NaN | 1.652340 |
| 2020-03-11 | 00:00:00 | line 2 | NaN | 0.542022 |
| 2020-03-11 | 00:00:00 | line 3 | General Hospital, Miguel Hidalgo | 0.481213 |
| 2020-03-11 | 00:00:00 | line 4 | Merced, Benito Juarez, Miguel Hidalgo | 0.140385 |
| 2020-03-11 | 00:00:00 | line 5 | NaN | 0.530010 |