Cyclitic Bike-Sharing: Data Cleaning Changelog

Date: 2020/06/05

Analyst: Sandisiwe Buthelezi

Project: BikeShare Data Cleaning and Manipulation

A strong analysis depends on the integrity of the data. The following steps were done to ensure the accuracy, completeness, consistency, and trustworthiness of the data before conducting the analysis. The step below will be performed using Structured Query Language (SQL).

## 2023/06/05 - Verify that data types are appropriate and consistent

**Decision:**

* Verifying if all field data type matches across all 12 tables

**Steps Taken:**

* Firstly, I retrieved table information from information\_schema.columns to learn about the tables, especially field data type.
* Then performed queries that include table name, column name, and corresponding data type to check and confirm the consistency in fields data type across all tables.
* Lastly, I updated the fields with appropriate data types and inconsistencies in data types.

**Reasoning:**

* To ensure no error arises during analysis, due to inconsistent data type.

**Validation:**

* I verified again if all field data types matched across all 12 tables.

## 2023/06/05 - Adding calculated field

**Decision:**

* Adding “ride\_length.” field
* Adding “day\_of\_week,” field

**Steps Taken:**

* I altered the tables by adding new columns on each table by calculating ride length using started \_at field and ended\_at for “ride\_length” and calculated day\_of\_week by using Datepart() function to get a day of the week.

**Reasoning:**

* The newly added calculated fields will be helpful during the analysis phase as provide a better opportunity for aggregations and visualization.

**Validation:**

* I perform a query to verify if all values are positive in the ride\_length column.
* I performed a query to verify all the numbers representing days ranging between 1 and 7

## 2023/06/06 - Handling inconsistent or incorrect values

**Decision:**

Identify inconsistent or incorrect values in the data.

**Steps Taken:**

* During the data exploration phase, I discovered the column in started\_at has floating point values instead of date and time like the rest of the other tables.
* I examined the floating-point values in the started\_at column of tripdata\_202210 to understand their structure and potential patterns, and to confirm if it was a formatting error.
* However, the data did not a pattern to convert to a date or time, which is a field data type.
* The next option was to treat a column data as missing values or invalid, in that way I would invalidate the column since the scenario is hypothetical and I have no stakeholders to assist in handling such data or further investigate the reason.

**Reasoning:**

* The inconsistent or incorrect values in data would have negative consequences on the analysis as it would not present the true picture of the company.
* Additionally, other analyses would not happen with incorrect values.

## 2023/05/31 - Splittings Date and Time columns into two columns

**Decision:**

* Splitting the date and time of started\_at and ended\_at columns to make data into a workable dataset.

**Steps Taken:**

* Used the convert() function to extract only the date on the started\_at and ended\_at columns.

**Reasoning:**

* The newly added calculated fields will be helpful during the analysis phase as provide a better opportunity for aggregations and visualization.

**Validation:**

* Performed a query to verify the field data type is the date and in the correct and consistent format all across 12 datasets.

## 2023/05/31 - Handling null values

**Decision:**

* excluding rows with null values with condition query.

**Steps Taken:**

* Used ISNULL() to exclude null values.

**Reasoning:**

* Null values can disrupt data integrity and consistency. Handling null values helps maintain the integrity of the dataset and ensures accurate and reliable analyses.

**Validation:**

* I calculated the sample size that would ensure the statistical power, precision, and validity of a dataset.
* Then, made sure the results are within the sample size.