Data preparation and pre-cleaning documentation

Data analyst: Abraham Cedeño

Date: 4th October 2022

Version #: 1

Steps:

Date-time: 4/10/22-11:58am (GMT-5)

1. I downloaded the last 12 months of files from the website provided and rename them BikeTrips\_date\_v0 accordingly to each date. I made a copy of all the xlxs files and named them as version 1 (v1).

Date-time: 4/10/22-12:03pm (GMT-5)

1. I opened each file and formatted the data as a table.

Date-time: 4/10/22-12:11pm (GMT-5)

1. Then, I sorted them in ascending order by the date of the started trip (column name: started\_at).

Date-time: 4/10/22-1:20pm (GMT-5)

1. I’ve categorized each column of data-by-data format and type of data, and then change the column type of data to the one below.

Data formats and types

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Data format | Data type before | Data type after |
| ride\_id | - | General | Text |
| rideable\_type | Qualitative (nominal) | General | Text |
| started\_at | Quantitative (continuous) | Custom | Custom (no edit) |
| ended\_at | Quantitative (continuous) | Custom | Custom (no edit) |
| start\_station\_name | Qualitative (nominal) | General | Text |
| start\_station\_id | Qualitative (nominal) | General | Text |
| end\_station\_name | Qualitative (nominal) | General | Text |
| end\_station\_id | Qualitative (nominal) | General | Text |
| start\_lat | Quantitative (continuous) | General | General (no edit) |
| start\_lng | Quantitative (continuous) | General | General (no edit) |
| end\_lat | Quantitative (continuous) | General | General (no edit) |
| end\_lng | Quantitative (continuous) | General | General (no edit) |
| member\_casual | Qualitative (nominal) | General | Text (string) |

Date-time: 5/10/22-12:30pm (GMT-5)

1. I looked up for duplicated values, I didn’t find any.

Date-time: 5/10/22-12:51pm (GMT-5)

1. I’ve proceeded to fill empty values in columns with blank cells.

Empty cells replaced with

|  |  |
| --- | --- |
| Column name | Empty cell replaced with |
| start\_station\_name | start\_station\_none\_given |
| start\_station\_id | start\_station\_id\_none\_given |
| end\_station\_name | end\_station\_none\_given |
| end\_station\_id | end\_station\_id\_none\_given |
| end\_lat | 91.00 |
| end\_lng | 181.00 |

Date-time: 5/10/22-1:02pm (GMT-5)

1. I saved all data as csv and xlsx files.

Date-time: 5/10/22-1:12pm (GMT-5)

1. I protected worksheets so that no one can access them without the password. Also, I protected the whole excel file so that only people who know the password can open it.