SQL data preparation and cleaning

Data analyst: Abraham Cedeño

Date: 8th October 2022

Version #: 3

Steps:

Date-time: 8/10/22-1:30pm (GMT-5)

1. After I imported the data, I explored the datatypes of each column, I found that I had to change the started time and end time because it was in am-pm format and the data type was string.

Table

Description automatically generated Table

Description automatically generated

Date-time: 8/10/22-1:37pm (GMT-5)

1. I make this query to transform the data to 24h format.

SELECT \* except(started\_time, end\_time),

CAST (FORMAT\_TIMESTAMP(

'%H:%M:%S',

PARSE\_TIMESTAMP('%I:%M:%S %p', started\_time)

) AS time)AS started\_time\_UTC,

CAST (FORMAT\_TIMESTAMP(

'%H:%M:%S',

PARSE\_TIMESTAMP('%I:%M:%S %p', end\_time)

) AS time)AS end\_time\_UTC,

FROM `capstone-1-abraham.Trip\_data.BikeTrips\_2021\_11\_04\_v31`

Date-time: 8/10/22-1:46pm (GMT-5)

1. Then I downloaded the new table and import it again to have permanent tables.

Table

Description automatically generated Table

Description automatically generated

Date-time: 8/10/22-1:55pm (GMT-5)

1. Then I did the same with each table and save it in a new Data Set which I called Trip\_data\_v1

Table

Description automatically generated

Date-time: 8/10/22-12:59pm (GMT-5)

1. Now that I have prepare all my data, I started the cleaning process by making sure I have not duplicated records. I used the next query and compared the result with the number of rows in my data.

SELECT COUNT(DISTINCT ride\_id) AS No\_Duplicados

FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2021\_11\_04\_v31`

Graphical user interface, application, table

Description automatically generated Graphical user interface, text, application, chat or text message

Description automatically generated

Date-time: 8/10/22-2:11pm (GMT-5)

1. In all my data I only found one duplicated value.

A picture containing graphical user interface

Description automatically generated

Application, table

Description automatically generatedGraphical user interface, text, application, chat or text message

Description automatically generated

Date-time: 8/10/22-02:21pm (GMT-5)

1. I Checked the records to make sure it was a duplicate, but I found that the records where different, the problem was that the id was the same.

Graphical user interface, text, application

Description automatically generated

Date-time: 8/10/22-02:30pm (GMT-5)

1. Because I’m using a free BigQuery account, it didn’t allow me to make an update to the table, so I had to change it on Excel and import the file again, I did that. After I followed all the steps before again, I ended up with no duplicates.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated with medium confidence

Date-time: 8/10/22-02:35pm (GMT-5)

1. Then I created a new dataset called Trips\_v2. Inside this dataset, I created a table that resulted from the merge of all the previous tables so that I have all the data in one table. I used the following query multiple times.

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2021\_11\_04\_v31`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2021\_11\_04\_v32`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2021\_12\_08\_v3`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_02\_02\_v3`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_03\_02\_v3`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_04\_06\_v3`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_05\_03\_v3`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_06\_03\_v31`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_06\_03\_v32`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_07\_15\_v31`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_07\_15\_v32`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_08\_05\_v31`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_08\_05\_v32`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_09\_08\_v31`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_09\_08\_v32`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_10\_11\_v31`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_10\_11\_v32`

UNION ALL

SELECT \* FROM `capstone-1-abraham.Trip\_data\_v1.BikeTrips\_2022\_01\_06\_v3`

Graphical user interface, text, application, email

Description automatically generated

Date-time: 8/10/22-02:45pm (GMT-5)

1. Now, with all the data in one table, I’m going to keep cleaning it before I analyze it. The next step I took was to make sure the length of the records where in the range they should be. I used the following queries.

SELECT ride\_id

FROM `capstone-1-abraham.Trip\_data\_v2.BikeTrips\_todos`

WHERE

LENGTH(rideable\_type) > 13

Graphical user interface, text, application

Description automatically generated

SELECT MIN(start\_lat) AS min\_start\_lat, MAX(start\_lat) AS max\_start\_lat

FROM `capstone-1-abraham.Trip\_data\_v2.BikeTrips\_todos`

Graphical user interface, application, table

Description automatically generated

Date-time: 8/10/22-02:49pm (GMT-5)

1. Next, I checked each column to make sure I have no NULL values. I used the following query. I found none.

SELECT \*

FROM `capstone-1-abraham.Trip\_data\_v2.BikeTrips\_todos`

WHERE rideable\_type IS NULL

Graphical user interface, text, application

Description automatically generated