**Data Cleaning Changelog for Bike Share Data Analysis Project for Google Data Analytics Professional Certificate Capstone.**

**Author:** William Melahouris

**Total Excel files cleaned:** 12

202012-divvy-tripdata.xlsx (December 2020 Bike Share Ride data)

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* In 202012-divvy-tripdata.xlsx, there are rows where started\_at is in December while ended\_at is in November, making for a very long negative ride duration in minutes (i.e. -28000 minutes), these rows were removed
* In 202012-divvy-tripdata.xlsx, some rides have ended\_at before started\_at, these were switched
* In 202012-divvy-tripdata.xlsx, some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202101-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* Some rides have ended\_at before started\_at, these were switched
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202102-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202103-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202104-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* One ride had ended\_at before started\_at, these were switched
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202105-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* One ride had ended\_at before started\_at, these were switched
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202106-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* Two rides had ended\_at before started\_at, these were switched
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202107-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202108-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* Several rides had ended\_at before started\_at, these were switched
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202109-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* Several rides had ended\_at before started\_at, these were switched
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202110-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order

202111-divvy-tripdata.xlsx

* Created a ride length column that calculates the length of the ride in minutes called ride\_length\_mins
* Quite a number of rides had ended\_at before started\_at, these were switched
* Some rides have a duration of 0, kept these as they might have only been ridden for less than a minute (i.e. maybe 30 seconds) but all the started\_at and ended\_at values only have times in HH:MM, so there is no way to tell exactly how many seconds they were ridden for.
* Created year, month, and day\_of\_week columns to represent the year, month, and day of week the ride started on.
* Note that there are rides that did not start at any start station and/or did not end at any end station.
* Noted that there are rides with no end\_lat or end\_lng
* Checked for duplicate values in the ride\_id column, none were found
* Sorted data by the started\_at column in ascending order