Changelog: Preparing and Processing Data Documentation

<All changes are listed chronologically>

|  |  |  |
| --- | --- | --- |
| Reasoning of Need for Actions | Preparing/Processing Actions | Outcomes of Actions Taken |
| Changes made in Excel: | | |
| Missing data in station\_name and station\_id columns | Created new spreadsheet called stations\_692 to compile station\_name, station\_id, latitude, longitude by copy and pasting information into new spreadsheet, removing station\_name and station\_id column from each month’s trip spreadsheet, then using the UNIQUE command in the stations\_692 spreadsheet to remove duplicates | Spreadsheet called stations\_692 that preserved all stations from all raw data and cleaned to list the 692 stations which matches Cyclistic records for number of docking stations, saved all spreadsheets adding v2 to file name |
| Data formatting | Some ride\_id in scientific notation, changed column formatting in each spreadsheet to text | A ride\_id column displaying numbers and letters of raw data alone without scientific notation |
| Invalid data | Using LEN command on ride\_id column in each month’s spreadsheet to ensure correct length of 16 characters, removed rows above or below that length | Cleaned ride\_id column with 131 rows deleted from Jun-21 spreadsheet, 143 from Jul-21, 133 from Aug-21, 117 from Sep-21, 134 from Oct-21, 70 from Nov-21, 49 from Dec-21, 22 from Jan-22, 22 from Feb-22, 38 from Mar-22, 67 from Apr-22, 93 from May-22, saved all spreadsheets adding v3 to file name |
| Metric needed for analysis | Created column in each month’s spreadsheet called ride\_length by subtracting started\_at datetime from ended\_at datetime | Column with the length of time in seconds of each trip |
| Invalid data | Removed rows with negative ride\_length, ended\_at datetime was greater than started\_at datetime in each month’s spreadsheet | Cleaned started\_at, ended\_at and ride\_length columns |
| Invalid data | Removed rows with 0 ride\_length and ended\_at datetime = started\_at datetime in each month’s spreadsheet | Cleaned started\_at, ended\_at and ride\_length columns |
| Metric needed for analysis | Created column in each month’s spreadsheet called day\_of\_week using the WEEKDAY command and the started\_at datetime | Column with 1 = Sunday to 7 = Saturday corresponding to the day of week when each trip started |
| Changes made in RStudio: | | |
| Reasoning of Need for Actions | Preparing/Processing Actions | Outcomes of Actions Taken |
| Analysis of all 12 months of data concurrently | Merged 12 separate months of data using | One data frame trips\_12\_months with 5859111 rows and 13 columns |
| Duplication of information | Removed columns start\_lat\_lng and end\_lat\_lng | One data frame trips\_12\_months with 5859111 rows and 11 columns, start and end latitude and longitude retained |
| Invalid data | Removed rows recorded with latitude and longitude outside of Cyclistic Bikeshare service area by filtering latitude and longitude | 9,188 observations outside of station latitude and longitude filtered out |
| Formatting error introduced during upload to RStudio from Excel introducing miscalculated ride\_length column, ride\_length is formatted with 1899-12-31 if duration <1day followed by HH:MM:SS | Added new column trip\_length using R code, difftime(trips\_12\_months$ended\_at, trips\_12\_months$started\_at) and removed the incorrect column ride\_length | A trip duration column with a single unit of output, seconds |
| Need to analyze a difftime data type | Converted column trip\_length to character then to numeric data type | A trip duration column formatted as one value without units listed after value |
| Needed to analyze trip start times by hour | Added hour column to data frame using started\_at column, trips\_12\_months$hour <-lubridate::hour(trips\_12\_months$started\_at) | Addition of hour column with value between 0 and 23 |
| Needed a column with date with date data type to make columns for month, day, and year | Added date column to data frame as date format, trips\_12\_months$date <- as.Date(trips\_12\_months$started\_at) | Addition of date column to create additional columns for analysis |
| Needed columns for month, day, and year separate to analyze | Created 3 new columns from date column trips\_12\_months$month <- format(as.Date(trips\_12\_months$date), "%m")  trips\_12\_months$day <- format(as.Date(trips\_12\_months$date), "%d")  trips\_12\_months$year <- format(as.Date(trips\_12\_months$date), "%Y") | 3 additional columns for trip timing analysis |
| Needed column with the name of the day of the week to analyze | Created a new column with day of the week, trips\_12\_months$day\_of\_week <- base::weekdays(trips\_12\_months$started\_at) | 1 additional column for trip timing analysis |
| Unable to perform calculations on month, day, and year columns as they are in character format | Used as.numeric() on these 3 columns to convert to numeric | Month, day, and year columns in numeric format available for calculation and analysis |
| rideable\_type captured as “docked\_bike” when the bike was a “classic\_bike” to enable analysis of bike choice | Changed all “docked\_bike” to “classic\_bike” in rideable\_type column, mutate(rideable\_type = recode(rideable\_type, "docked\_bike" = "classic\_bike") | In rideable\_type column only 2 valid variables of electric and classic to analyze |
| Add meaning and organization to hour column to aid analysis, interpretation, and visualization | Created column ride\_hour\_category with 5 categories based on hour column to group related hours | Ride\_hour\_category column with 5 variables: morning commute, late morning, afternoon, evening commute, and night life |
| Data formatting of ride\_hour\_category is not defined as categorical data | Created factor with 5 levels to convert the ride\_hour\_category column to discrete data | Ride\_hour\_category column as a factor with 5 levels of predefined, finite number of values that group based on level |
| Need to organize day of the week for logical grouping | Created factor with 7 levels corresponding to each day of the week beginning with Monday and ending with Sunday | Day\_of\_week column as a factor with 7 levels of predefined finite number of values than group based on level |
| Need a month name column to aid visuals of trip timing analysis | Renamed month column as month number, converted month number to numeric data type, created column of month abbreviations based on month number column, ordered month abbreviations from Jan to Dec | A month\_number column with variables 1 to 12 as numeric and a month column with month abbreviations that group based on order |
| Need weekday as a number for some calculations | Created column using format(trips\_12\_months$started\_at,"%u”) and formatted as numeric | New column week\_day\_number with days of the week coded as Monday = 1 to Sunday = 7 |
| Add meaning and organization to hour column to aid analysis, interpretation, and visualization | Created column to categorize trip lengths into 5 trip length ranges: mini rides, short rides, medium rides, long rides, longest rides | New column trip\_length\_category with 5 trip length ranges |
| Data formatting of trip\_length\_category is not defined as categorical data | Converted column into factor with 5 levels | Trip\_length\_category with 5 levels ordered based on trip length |
| Add meaning and organization to latitude and longitude data to aid analysis, interpretation, and visualization | Created column location\_category with categories of 9 equal squares encompassing Cyclistic service area based on latitude and longitude data | New location\_category column with 9 values: south west, south central, south east, west central, central, east central, north west, north central, north east |
| Data formatting of location\_category not defined as categorical data | Converted column into factor with 9 levels | Location\_category with 9 levels of ordered based on south to north and west to east |
| Trip\_length skewness 188.884 with means and medians differing greatly and numerous outliers that could decrease confidence in interpretation of trip\_length analysis | Removed outliers by filtering out trips with trip length less than the 1st quartile (317 seconds) and greater than the 3rd quartile (1668 seconds) | no\_outs\_trips\_12\_months data frame with 3849679 rows and skewness of 0.6903038, new now reflects the population rather than outlier trips skewing analysis results |
| Need subsets of member and casual data for some analysis steps | Created 2 new data frames from no\_outs\_trips\_12\_months data frame, one with only casual riders, one with only members | 2 new data frames: member\_no\_outs\_trips\_12\_months and casual\_no\_outs\_trips\_12\_months |