Analyze the use of self-service bikes in the city of san Francisco.

What influences the use of bikes? IS it the weather? or the days of the week ? Or maybe the location?

# Research question:

1. Introduction

We would like to analyze self-service bicycle rentals in the City of San Francisco. We want to demonstrate if there are any patterns. Does the weather influence the rentals? Are there more rentals on weekends? Do rentals vary by time of day?

For this project we will only use R.

1. Methodology:

….

# ETL process:

## Extraction:

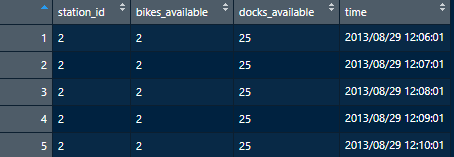
We retrieved our data from the Kaggle site: <https://www.kaggle.com/benhamner/sf-bay-area-bike-share?select=database.sqlite>

The first data set: Station

It contains all the bike rental stations of the bay of san Francisco, with their location, name, city and date of installation. This data set doesn’t need any transformation.

70 lines and 7 variables

The second data set: Status

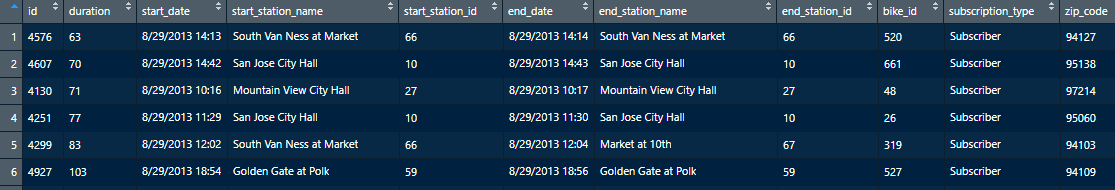
The data set contains all the number of bikes and docks available between 2013 and 2015 for every minute. (71,984,434 lines and 4 variables)

Thus, the data set is way too much big. We must reduce his size. By following these steps:

We keep only the average available bike per hour. This grain is enough to start our analysis. We will make also analyses using a finer grain but only for a week for instances. We must keep a reasonable size for of data set. We finally reach 1,204,764 rows.

The third data set: trip

The data set contains all the individual bike trips. (669,959 lines and 11 variables)

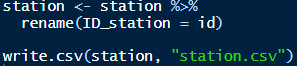
We also must reshape this data set. We drop the variable zip\_code.

The 4th data set: weather

The data set contains all the information about the weather on a specific day. (3665 lines and 24 variables). We decided to reduce the number of variables in order to keep only those that can influence the bike location. This data set will be useful in order to build our future model.

## Transformation

The data transformation part took us a long time. Indeed, we had to make many modifications in order to be able to assemble our data in the Fact\_table.

Table `Station`:

we only modified the variable "ID\_station" to have a unity in our data

Table `Status`:

we have created hour, day, month and year variables in order to be more precise in our model. We also created a primary key "ID\_date". It will later allow us to merge our databases more easily.

Finally, we reduce multiple values down to a single summary by using the function `summarize`.

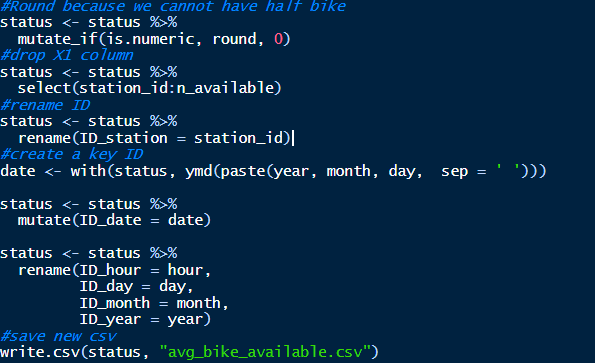
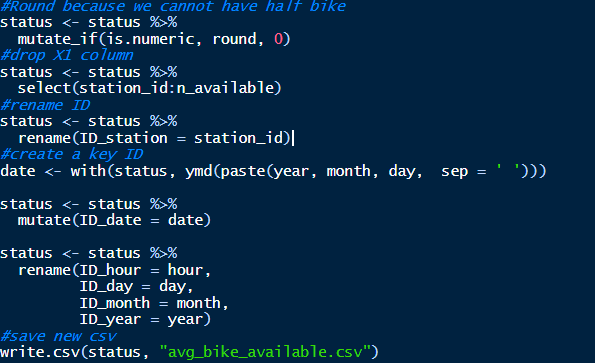
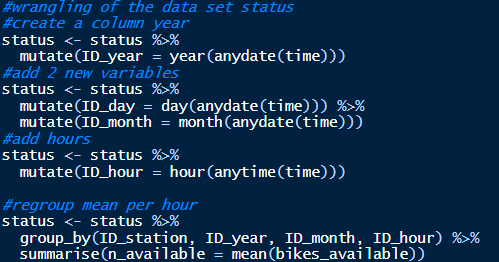


Table `trip`:

we have created hour, day, month and year variables in order to be more precise in our model. Then we have aggregated our data in order to have the average travel time per hour and the number of trips per hour.

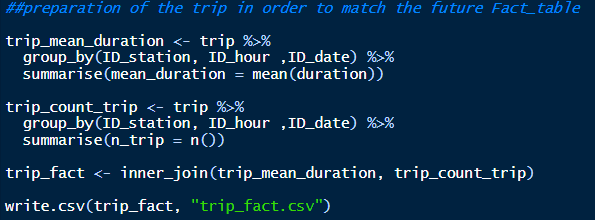
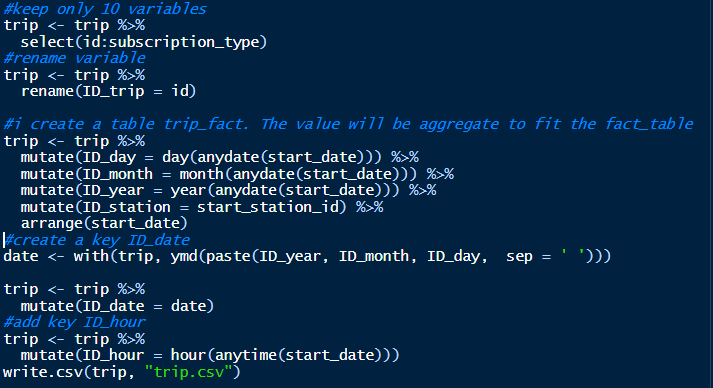


Table `weather`:

many modifications were necessary. Indeed, we modified the units to have our European standards.

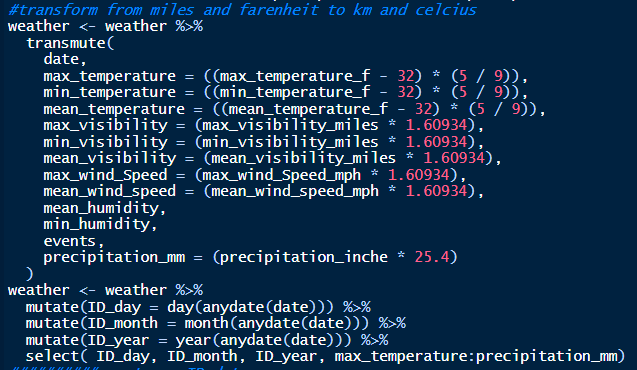
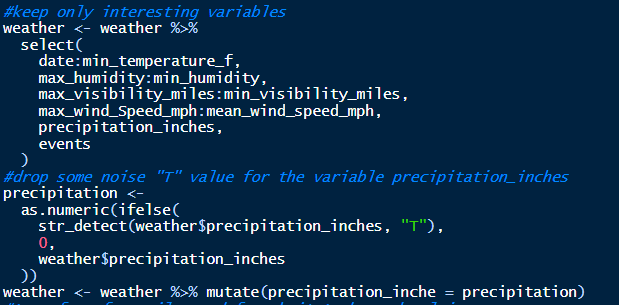
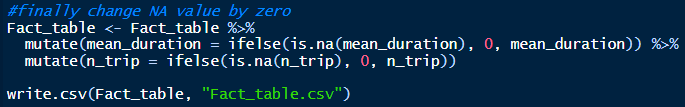


Table `Fact\_table`:

we merged the necessary data into our fact table in order to conduct our analysis.



## C:\Users\cevui\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\5FDEEE4C.tmpAllan PorterETL workflow

Rapport

dplyr

Group\_by()

Summarize()

Filter()



Mutate

ID\_date, ID\_hour

Group\_by

select 10 columns

ID\_trip

mutate

ID\_date

Unit transformation



Station.csv

status.csv

trip.csv

weather.csv

Stats.sfmta.com



|  |
| --- |
| **Avg\_bike\_available** |
| ID\_station |
| ID\_Year |
| ID\_month |
| ID\_day |
| ID\_hour |
| n\_available |
| ID\_date |

|  |
| --- |
| **Trip** |
| ID\_station |
| ID\_hour |
| ID\_date |
| Mean\_duration |
| n\_trip |

|  |
| --- |
| **Station.CSV** |
| ID\_station |
| Name |
| Lat |
| Long |
| Dock\_count |
| City |
| Installation\_Date |

|  |
| --- |
| **Weather\_final** |
| ID\_day |
| ID\_month |
| ID\_year |
| max\_temperature |
| min\_temperature |
| mean\_temperature |
| max\_visibility |
| min\_visibility |
| mean\_visibility |
| Max\_wind\_speed |
| Mean\_wind\_speed |
| Mean\_humidity |
| Min\_humidity |
| Events |
| Precipitation\_mm |
| ID\_date |

|  |
| --- |
| **Fact\_table** |
| ID\_station |
| ID\_year |
| ID\_month |
| ID\_day |
| ID\_hour |
| ID\_date |
| mean\_temperature |
| mean\_visibility |
| mean\_wind\_speed |
| mean\_humidity |
| mean\_precipitation\_mm |
| n\_available |
| mean\_duration |
| n\_trip |

## Star Schema:

# Analysis