UNIL | Université de Lausanne
HEC Lausanne

ANALYSIS OF THE SAN FRANCISCO AREA BIKE SHARE

BUSINESS INTELLIGENCE AND ANALYZING BIG DATA 2020

12.06.2020

Gaëtan Lovey
Cédric Vuignier
Nicolas Vulliemin

Executive summary

In this study, we aimed to apply business intelligence tools to a real-world research project. In that framework, we made an analysis of the patterns influencing the public bike rentals in the San Francisco bay area. As we consider that bike rentals will play an increasing role in urban mobility, we tried to understand key aspect influencing levels of rentals on various timeframes. Three underlying questions linked to the available data were raised. Does the weather influence the rentals? Are there more rentals on weekends? Do rentals vary by time of day?

To do so, we extracted data from the open platform "kaggle.com" that gave us access to the required data between September 2013 and August 2015. In order to have the best representation of our data and to obtain a clear understanding of the characteristics influencing bike rentals, we decided to build a star schema integrating 4 dimensions tables and 1 fact table.

Dimensions:

- Average bike available (id: ID_Station / ID_Date),
- Station (id: ID Station)
- Weather (id: ID Date)
- Trip (id: ID Date)

All along the ETL process, some modifications were made to assemble the data in the fact table. These modifications were done to build primary keys, reduce the size of the data and match European standards. To conclude the ETL process, the transformed and ready to use data was loaded on the R software that was used for the analytical part of the study.

Our analysis first underlined a clear weekly cyclicality. Opposite to our assumption, we found that bikes rentals were significantly higher during working days. Moreover, we observed peaks of utilization at 8 am and 5 pm, what supports that San Franciscans use bikes to commute. Besides, we could also observe that the rate of bike usage is linked to the season with higher quantity during the summer. Finally, we noted that the average trip duration is 7 minutes what support transport instead of leisure usage. All in all, we identified working day and time of the day as the two key variables of our model.

Interestingly, we could not find any link between the weather and the rate of bike rentals. Based on precipitations and temperature, no significant trend appeared. Thus, we assume that bike rentals and the weather are not correlated. However, it has to be considered that the absence of correlation might be linked to the relatively dry climate San Fransisco enjoys.

Finally, in order to test the importance of the variable in our model, we tested the predictive power of the model with a random forest. It confirmed that the identified variable (working day, time of the day) could significantly predict the amount of bike rentals. Furthermore, we replicated the test without the identified variable and no consistent pattern was found with for example weather or month.

Overall our study allowed to find significant patterns related to bike rentals in the San Fransisco area. However, for many reasons, this study cannot apply to other geographical scope. Furthermore, our analysis opened the door to further investigations focused on how bike rental is combined to other mode of transportation.

Table of contents

List of figures

1. Introduction

1.1 Research questions

Free-access bicycles are taking up more and more space in major western cities. They are used as a form of transport in everyday life but also for leisure activities. This is why it is essential to understand user habits and predict the future in order to guarantee an efficient service. The aim of this study is to analyze the self-service bicycle rentals in the City of San Francisco and its area. We seek patterns characterizing self-service bike rental. Does the weather influence the rentals? Are there more rentals on weekends? Do rentals vary by time of day?

We chose to study bicycle rental because it is a phenomenon that is gaining momentum within our society. This activity, which quickly became essential for sustainable development, has enormous development potential and deserves to be studied more deeply. Moreover, with the appearance of the electric bicycle and electric, non-polluting means of transport market is growing more and more. We consider that transport rental companies will have a fundamental role to play in the future.

In our study, we will focus only on the San Francisco area and it is important to underline that our results could be different in comparison to other cities. Indeed, the use of bicycles varies enormously according to the attributes of a city or a geographical place. Therefore, the results of this study should not be applied to other locations without further investigations

1.2 Methodology

It is important to mention that this study is the result of group work and therefore of a collaboration between Cédric Vuignier who took care of the whole ETL process, Gaëtan Lovey who took care of the data analysis and Nicolas Vulliemin who built the study report.

For this analysis, we used the R software because we are used to work with it during our Business Analytics courses.

To begin, we will start by extracting the data from “kaggle” website, then we will transform and load them in order to perform an exploratory data analysis. Finally, we will build a prediction model in order to predict the number of trips per day and to know which variables are influencing the number rentals in San Francisco.

2. ETL process

In this part, we will explain how we extracted, transformed and loaded our data.

The files were over 70 million lines long. However, we managed to apply the 3 steps of the ETL process as described below. Our information comes from four different tables downloaded from “kaggle” website (*SF Bay Area Bike Share*, s. d.).

2.1 Extraction

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 does not need any transformation. This table will allow us to define the most frequented courses and thus to make the link between a name station and a completed trip.

Figure 1: Station data set

	id	name	lat	long	dock_count	city	installation_date
1	2	San Jose Diridon Caltrain Station	37.32973	-121.9018	27	San Jose	8/6/2013
2	3	San Jose Civic Center	37.33070	-121.8890	15	San Jose	8/5/2013
3	4	Santa Clara at Almaden	37.33399	-121.8949	11	San Jose	8/6/2013
4	5	Adobe on Almaden	37.33141	-121.8932	19	San Jose	8/5/2013
5	6	San Pedro Square	37.33672	-121.8941	15	San Jose	8/7/2013

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. Thus, the data set is way too big. We will then reduce the size during the transformation part.

Figure 2: Status data set

	station_id	bikes_available	docks_available	time
1	2	2	25	2013/08/29 12:06:01
2	2	2	25	2013/08/29 12:07:01
3	2	2	25	2013/08/29 12:08:01
4	2	2	25	2013/08/29 12:09:01
5	2	2	25	2013/08/29 12:10:01

71,984,434 lines and 4 variables

The third data set: trip

The data set contains all the individual bike trips. We also have to reshape this data set. We can already state that there are 669,959 races run during the period. It also contains the beginning and the end of the trip, its duration and the type of users.

Figure 3: Trip data set

id	duration	start_date	start_station_name	start_station_id	end_date	end_station_name	end_station_id	bike_id	subscription_type	zip_code
4576	63	8/29/2013 14:13	South Van Ness at Market	66	8/29/2013 14:14	South Van Ness at Market	66	520	Subscriber	94127
4607	70	8/29/2013 14:42	San Jose City Hall	10	8/29/2013 14:43	San Jose City Hall	10	661	Subscriber	95138
4130	71	8/29/2013 10:16	Mountain View City Hall	27	8/29/2013 10:17	Mountain View City Hall	27	48	Subscriber	97214
4251	77	8/29/2013 11:29	San Jose City Hall	10	8/29/2013 11:30	San Jose City Hall	10	26	Subscriber	95060
4299	83	8/29/2013 12:02	South Van Ness at Market	66	8/29/2013 12:04	Market at 10th	67	319	Subscriber	94103
4927	103	8/29/2013 18:54	Golden Gate at Polk	59	8/29/2013 18:56	Golden Gate at Polk	59	527	Subscriber	94109

669,959 lines and 11 variables

The fourth 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 prediction model.

Figure 4: Weather data set

ID_day	ID_month	ID_year	max_temperature	min_temperature	mean_temperature	max_visibility	min_visibility	mean_visibility
29	8	2013	23.333333	16.111111	20.000000	16.09340	16.09340	16.09340
30	8	2013	25.555556	15.555556	20.555556	16.09340	11.26538	16.09340
31	8	2013	21.666667	13.888889	17.777778	16.09340	16.09340	16.09340
1	9	2013	23.333333	14.444444	18.888889	16.09340	16.09340	16.09340
2	9	2013	23.888889	16.666667	20.555556	16.09340	9.65604	16.09340
3	9	2013	22.777778	15.555556	19.444444	16.09340	16.09340	16.09340

2.2 Transformation and loading

In this part, we made many modifications on the data in order to assemble them in the Fact_table.

Table “Station”

We only modified the variable "ID_station" to have a unity in our data.

Figure 5: Station table

```
station <- station %>%
  rename(ID_station = id)

write.csv(station, "station.csv")
```

Table “Status”

We only keep the average available bikes per hour. This grain is enough to start our analysis and allow us to reduce the size of our data. We will also make analysis using a finer grain but only for a week for example. We have to keep a reasonable size of data set. We finally reach 1,204,764 rows.

We 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 reduced multiple values down to a single summary by using the function “summarize”.

Figure 6: Status table

```
#wrangling of the data set status
#create a column year
status <- status %>%
  mutate(ID_year = year(anydate(time)))
#add 2 new variables
status <- status %>%
  mutate(ID_day = day(anydate(time))) %>%
  mutate(ID_month = month(anydate(time)))
#add hours
status <- status %>%
  mutate(ID_hour = hour(anytime(time)))

#regroup mean per hour
status <- status %>%
  group_by(ID_station, ID_year, ID_month, ID_hour) %>%
  summarise(n_available = mean(bikes_available))
#Round because we cannot have half bike
status <- status %>%
  mutate_if(is.numeric, round, 0)
#drop X1 column
status <- status %>%
  select(station_id:n_available)
#rename ID
status <- status %>%
  rename(ID_station = station_id)
#create a key ID
date <- with(status, ymd(paste(year, month, day,  sep = ' ')))

status <- status %>%
  mutate(ID_date = date)

status <- status %>%
  rename(ID_hour = hour,
         ID_day = day,
         ID_month = month,
         ID_year = year)
#save new csv
write.csv(status, "avg_bike_available.csv")
```

Table “trip”

Again, we created hour, day, month and year variables in order to be more precise in our model. Then we aggregated our data in order to have the average travel time per hour and the number of trips per hour. The new table “trip_fact” will be used to build our “Fact_table”.

Figure 7: Trip table

```
#keep only 10 variables
trip <- trip %>%
  select(id:subscription_type)
#rename variable
trip <- trip %>%
  rename(ID_trip = id)

#i create a table trip_fact. The value will be aggregated to fit the fact_table
trip <- trip %>%
  mutate(ID_day = day(anydate(start_date))) %>%
  mutate(ID_month = month(anydate(start_date))) %>%
  mutate(ID_year = year(anydate(start_date))) %>%
  mutate(ID_station = start_station_id) %>%
  arrange(start_date)
#create a key ID_date
date <- with(trip, ymd(paste(ID_year, ID_month, ID_day,  sep = ' ')))

trip <- trip %>%
  mutate(ID_date = date)
#add key ID_hour
trip <- trip %>%
  mutate(ID_hour = hour(anytime(start_date)))
write.csv(trip, "trip.csv")

##preparation of the trip in order to match the future Fact_table

trip_mean_duration <- trip %>%
  group_by(ID_station, ID_hour ,ID_date) %>%
  summarise(mean_duration = mean(duration))

trip_count_trip <- trip %>%
  group_by(ID_station, ID_hour ,ID_date) %>%
  summarise(n_trip = n())

trip_fact <- inner_join(trip_mean_duration, trip_count_trip)

write.csv(trip_fact, "trip_fact.csv")
```

Table “weather”

Many modifications were necessary in there. Indeed, we modified the units to have our European standards. However, as these are only linear modifications, it does not create any further issue for our analysis.

Figure 8: Weather table

```
#keep only interesting variables
weather <- weather %>%
  select(
    date:min_temperature_f,
    max_humidity:min_humidity,
    max_visibility_miles:min_visibility_miles,
    max_wind_Speed_mph:mean_wind_speed_mph,
    precipitation_inches,
    events
  )
#drop some noise "T" value for the variable precipitation_inches
precipitation <-
  as.numeric(ifelse(
    str_detect(weather$precipitation_inches, "T"),
    0,
    weather$precipitation_inches
  ))
weather <- weather %>% mutate(precipitation_inche = precipitation)
#transform from miles and fahrenheit to km and celcius
weather <- weather %>%
  transmute(
    date,
    max_temperature = ((max_temperature_f - 32) * (5 / 9)),
    min_temperature = ((min_temperature_f - 32) * (5 / 9)),
    mean_temperature = ((mean_temperature_f - 32) * (5 / 9)),
    max_visibility = (max_visibility_miles * 1.60934),
    min_visibility = (min_visibility_miles * 1.60934),
    mean_visibility = (mean_visibility_miles * 1.60934),
    max_wind_Speed = (max_wind_Speed_mph * 1.60934),
    mean_wind_speed = (mean_wind_speed_mph * 1.60934),
    mean_humidity,
    min_humidity,
    events,
    precipitation_mm = (precipitation_inche * 25.4)
  )
weather <- weather %>%
  mutate(ID_day = day(anydate(date))) %>%
  mutate(ID_month = month(anydate(date))) %>%
  mutate(ID_year = year(anydate(date))) %>%
  select( ID_day, ID_month, ID_year, max_temperature:precipitation_mm)
#create an ID_date
date <- with(weather, ymd(paste(ID_year, ID_month, ID_day,  sep = ' ')))
weather <- weather %>%
  mutate(ID_date = date)
#save new csv
write.csv(weather, "weather_final.csv")
```

Table "Fact_table"

We merged the necessary data into our fact table in order to conduct our analysis. The following two diagrams (Figure 10 and 11) explain our entire process in detail.

Figure 9: Fact_table

```
Fact_table <- full_join(status, trip_fact, by = c("ID_station", "ID_hour",
                                                  "ID_date"))

#Add weather information
weather_fact <- weather %>%
  group_by(ID_date) %>%
  summarise(
    mean_temperature = mean(mean_temperature),
    mean_visibility = mean(mean_visibility),
    mean_wind_speed = mean(mean_wind_speed),
    mean_humidity = mean(mean_humidity),
    mean_precipitation_mm = mean(precipitation_mm)
  ) %>%
  transmute(ID_date,
    mean_temperature,
    mean_visibility,
    mean_wind_speed,
    mean_humidity,
    mean_precipitation_mm,
  )
#add weather to the table fact

Fact_table <- inner_join(Fact_table, weather, by=("ID_date"))

#finally change NA value by zero
Fact_table <- Fact_table %>%
  mutate(mean_duration = ifelse(is.na(mean_duration), 0, mean_duration)) %>%
  mutate(n_trip = ifelse(is.na(n_trip), 0, n_trip))

write.csv(Fact_table, "Fact_table.csv")
```

After having transformed the different previous tables, we loaded them using the R software.

2.3 ETL workflow

Summary:

1. data collection on kaggle.com
2. Transformation process on each of the tables
3. table assembly -> Fact
4. using R to generate our results
5. Creating visualizations.

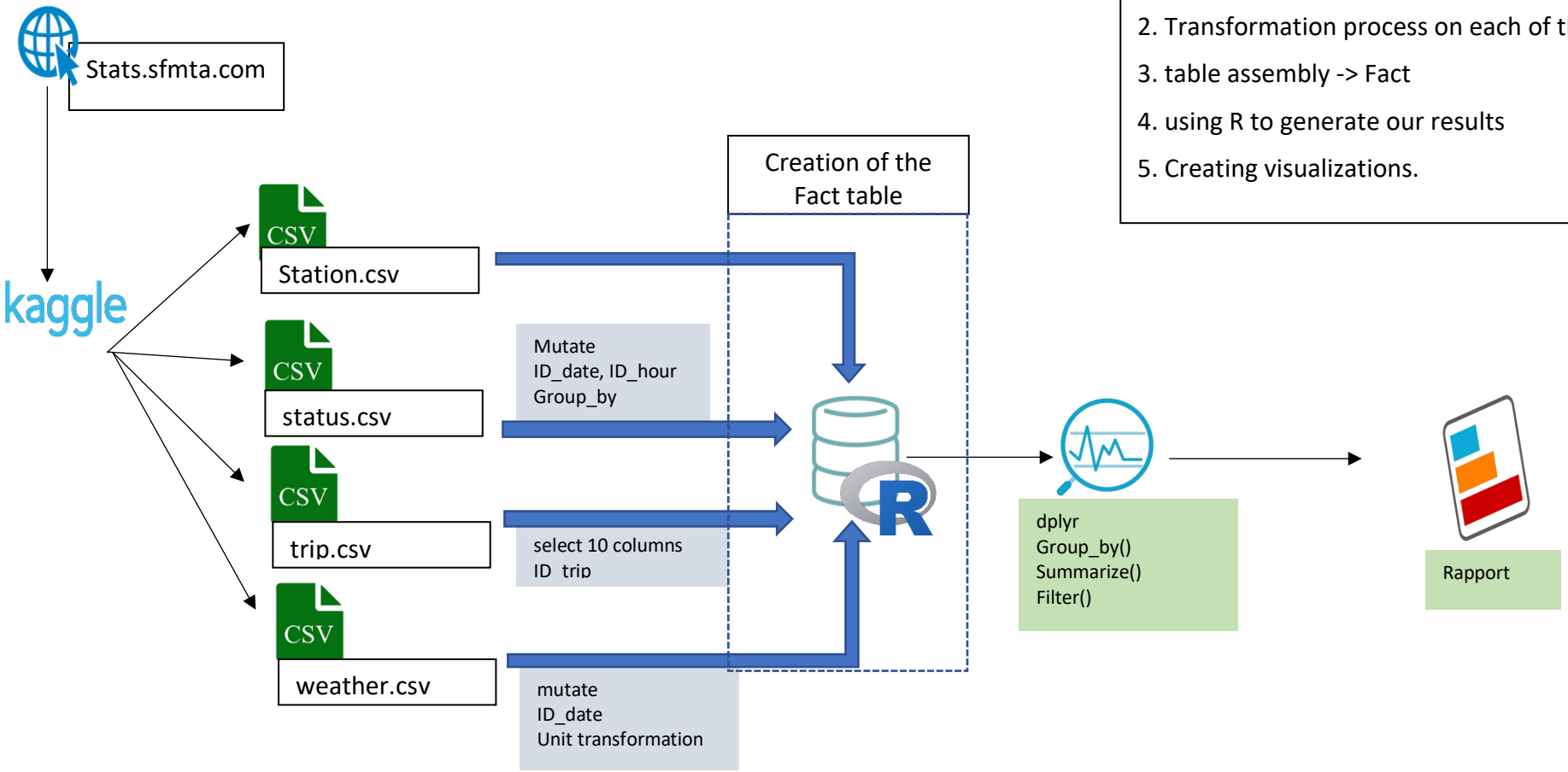


Figure 10: ETL Workflow

2.4 Star Schema

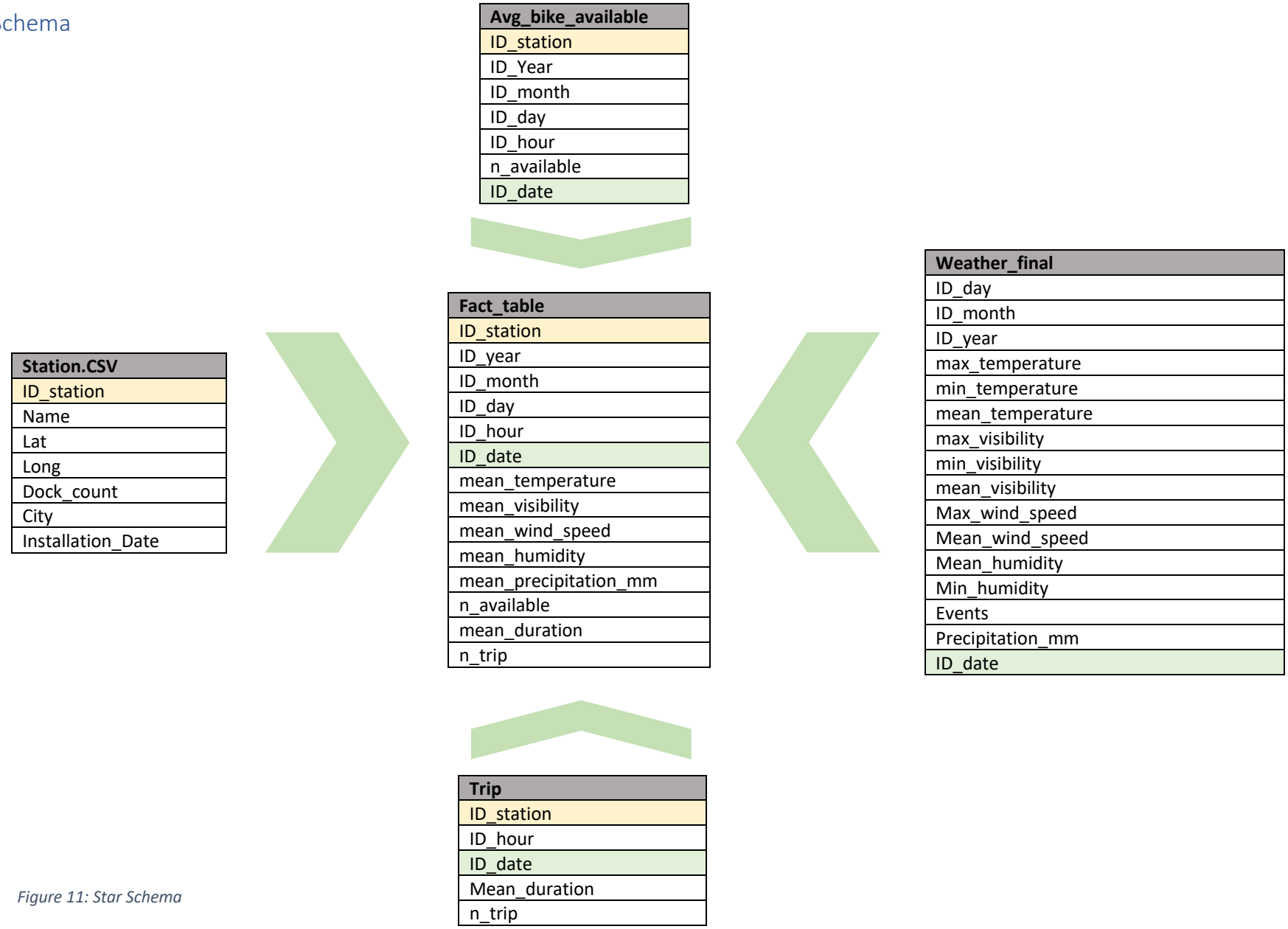


Figure 11: Star Schema

3. Analysis

This chapter is split into two parts: an exploratory data analysis part and a modelling part.

3.1 Exploratory data analysis

In this section, we made an exploratory data analysis in order to look for some patterns that characterize self-service bike rental in San Francisco.

It is important to know that we only have one entire year in the data. The periods of our data are the following:

- 29th of August 2013 –> 31 of December 2013
- 1st of January 2014 –> 31 of December 2014
- 1st of January 2015 –> 31 of August 2015

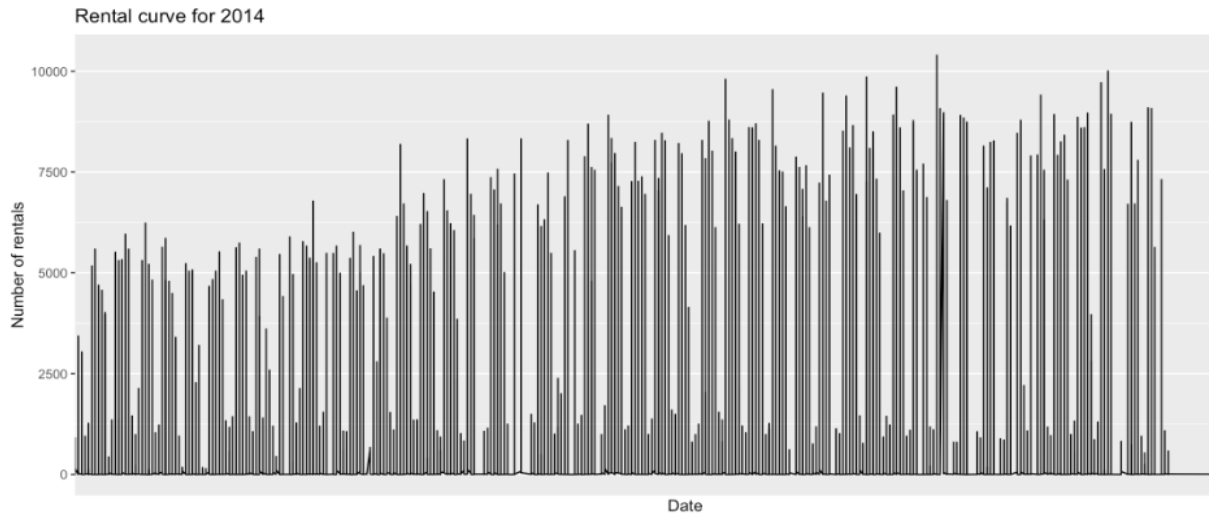
Thus, we work on 2 complete years.

Behavior of the rental curve

In order to know the behavior of the rental bikes, we plotted the rentals of the year 2014 in Figure 12. We can observe cyclicity for each week and an upward trend over the year.

Figure 12: Rental curve for 2014

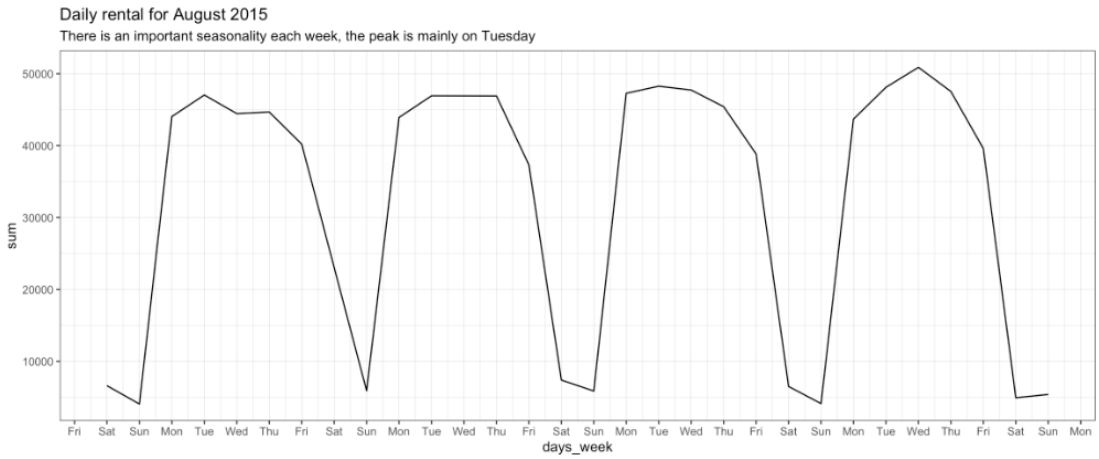
Annual



As we need more details, we will focus on a more specific period such as summer 2015. Thus, in Figure 13, we clearly observe a repeated pattern every week. By analyzing the period from 1st of August to 31th of August 2015, we can see that Sunday is the weakest day in terms of rentals and Tuesday is in majority the best rental day.

Week

Figure 13: Daily rental for August 2015

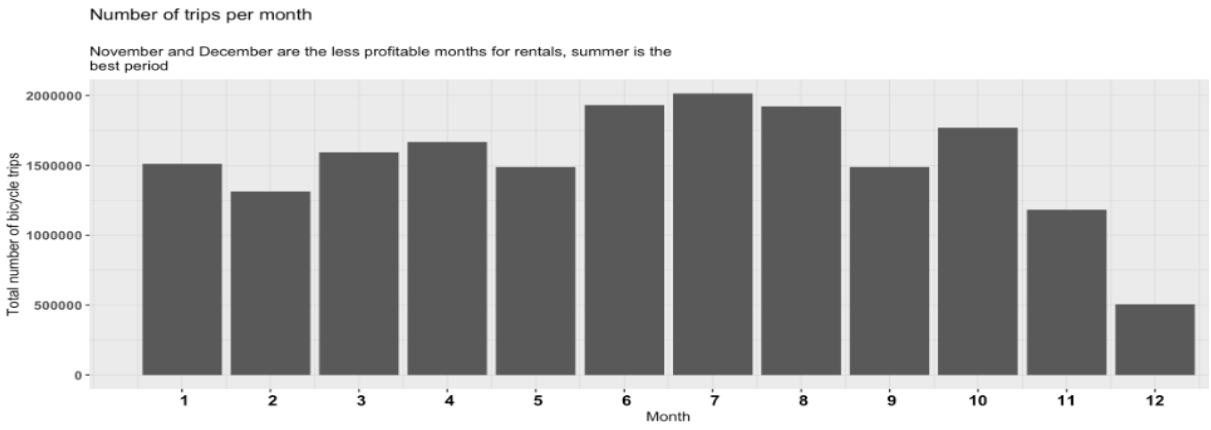


Number of trips

In this study, it is important to find patterns that best characterize the distribution of our predicted variable which will be the number of trips per day. We will do more research using a smaller granulometry.

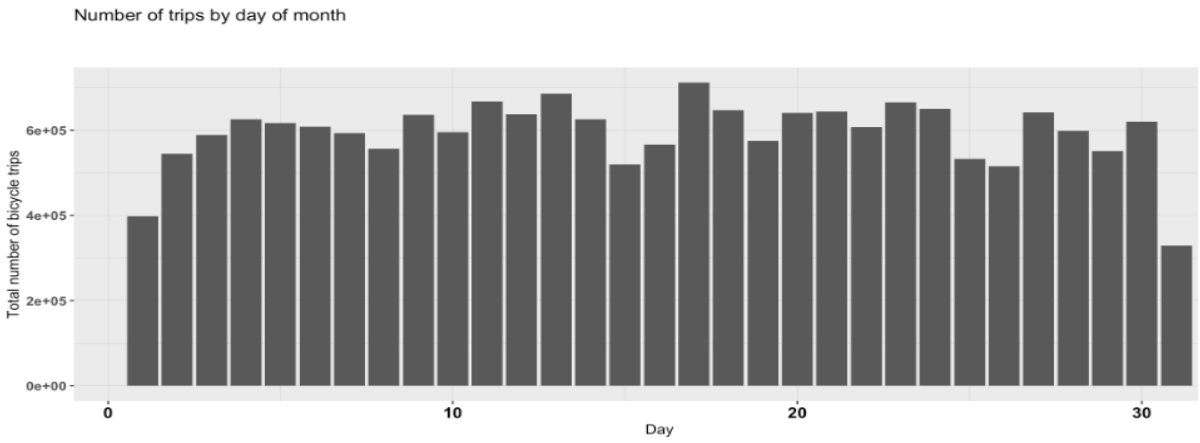
We see that the time of the year influences the number of bike rentals (Figure 14). Summer is the best period for rentals and winter is the worst one. It will therefore be interesting to include this variable in our future prediction model.

Figure 14: Number of trips per month during the whole data period



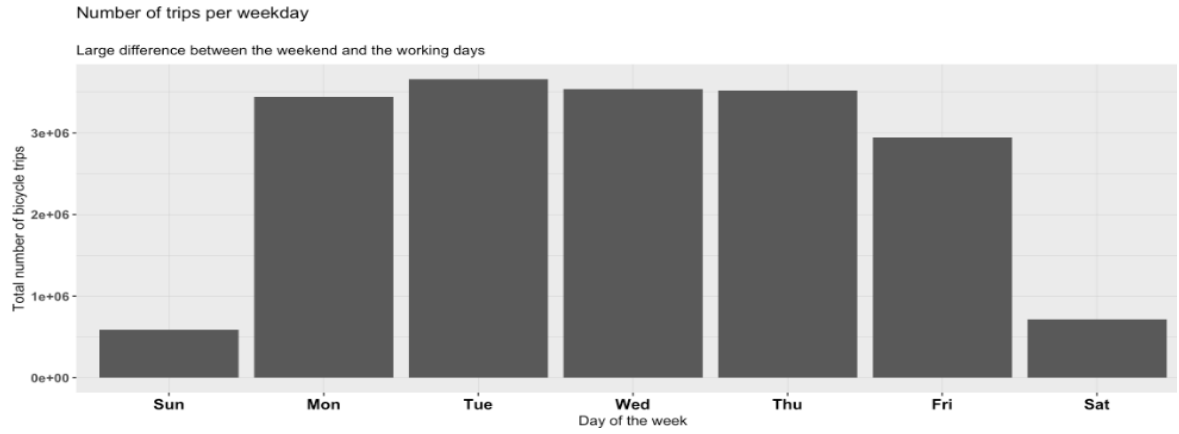
However, in Figure 15, we can see that the day has no influence on the number of bike rentals. This result is logical because this variable does not express the characteristic of the day (work or holiday). Without any surprise, the day 31th has less rentals because this day does not occur each month.

Figure 15: Number of trips per day of month



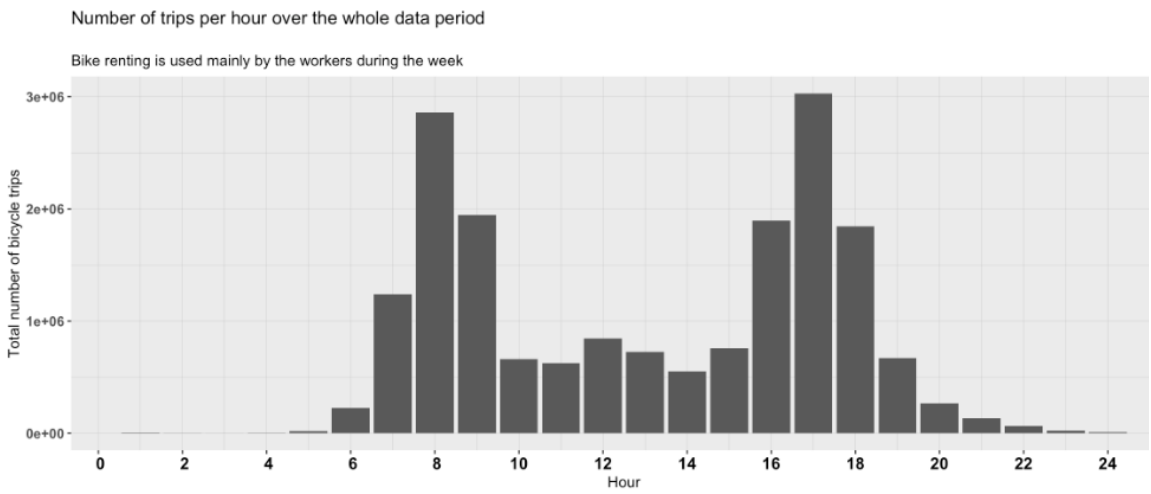
Finally, by defining the day of the week, we get a significant result in Figure 16, which confirms the results obtained in Figure 13. Indeed, the rentals are much lower on weekends. This means that the inhabitants of the bay of San Francisco use bike rentals primarily to get to work or for various activities during the week. It will therefore be an interesting variable to consider for our model.

Figure 16: Number of trips per weekday



In average, location peaks during the day are at 8 o'clock in the morning and at 5 o'clock in the evening as we can see in Figure 17.

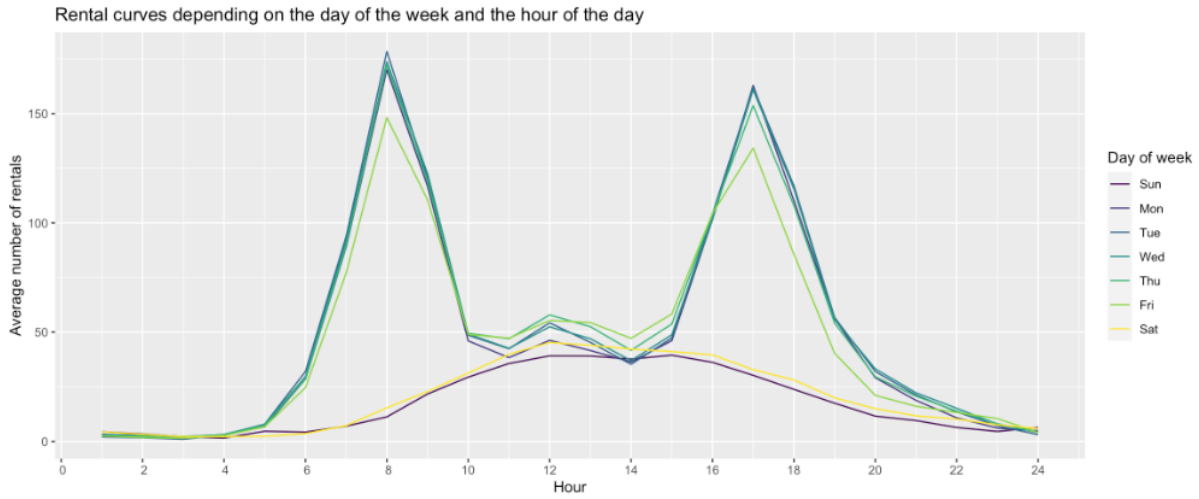
Figure 17: Number of trips per hour over the whole data period



These hours correspond to people commuting to their workplace in the morning and returning home in the evening. This confirms our hypothesis that bike rental rate is highly linked to journeys between the house and the workplace.

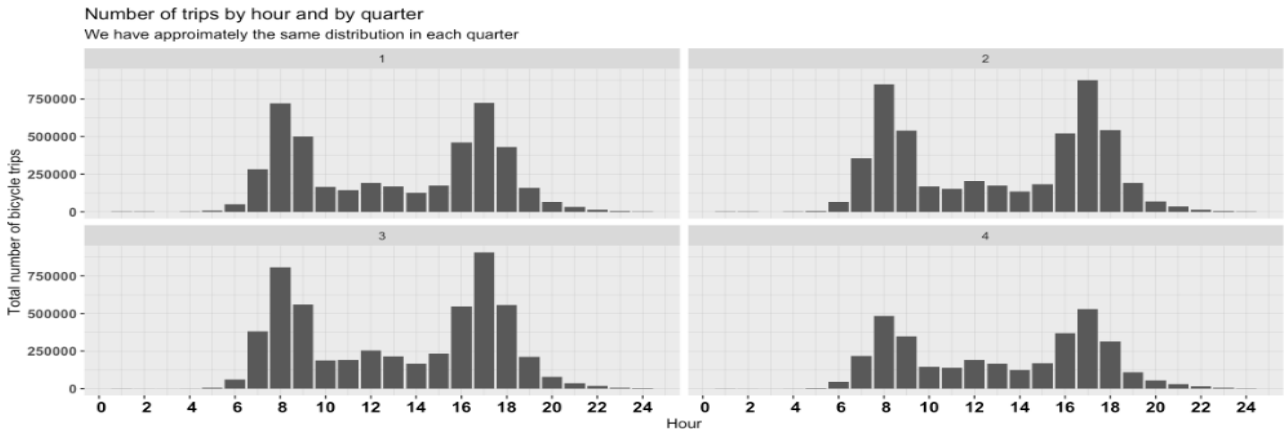
The Figure 18 confirms our hypothesis. Users use it mainly during the week. However, on weekends the curve is much smoother and therefore has no peak as rentals are supposedly matched to leisure activities. The time and day of the week therefore directly influence the number of final rentals.

Figure 18: Rental curves depending on the day of the week and the hour of the day



Finally, we looked to see if the quarter influenced bike rentals (Figure 19). We observe that the differences are very small and are essentially due to the effect of the months.

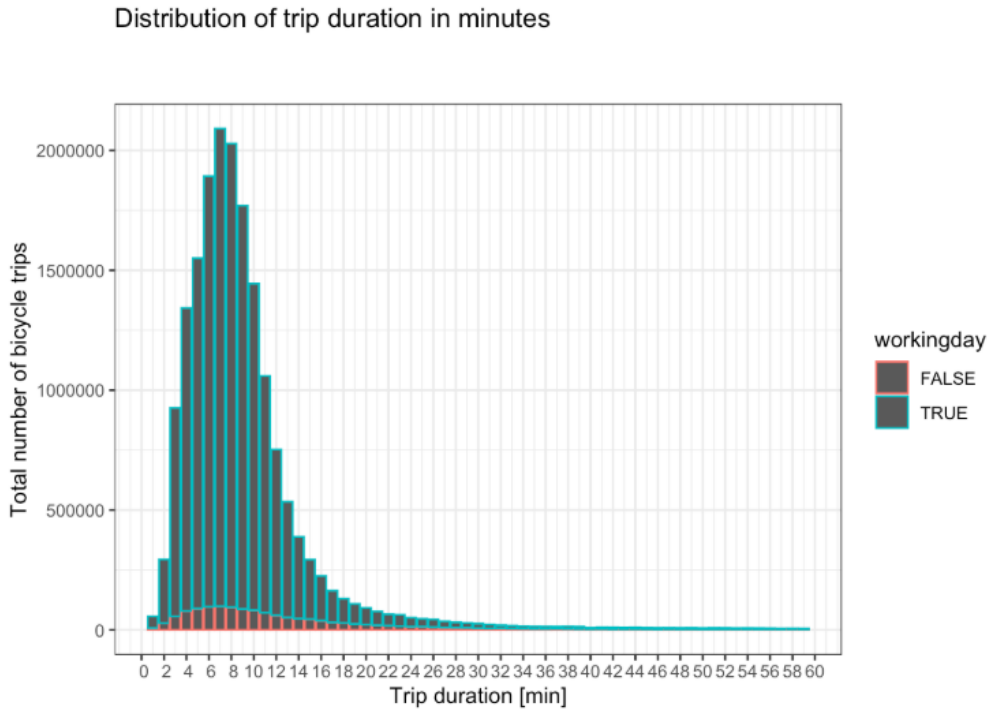
Figure 19: Number of trips by hour and by quarter



Trip duration

The distribution is left skewed in Figure 20. We can note that during the working days, the most frequent duration is the same than during the weekend (about 7 minutes).

Figure 20: Distribution of trip duration in minutes



We have seen that it is essential to provide several grains. This allows us to multivariate the explanatory variables and to know very precisely the influences on our predicted variable.

Weather

Here, we will look for linear relationships between the number of rentals per day and the weather characteristics. It will allow us to eventually admit that weather can affect the number of bike trips or their duration. First, we compared weather features to the number of trips per day and then, we reiterated the method for the trip duration. We chose to display, in this report, only the average temperature and the amount of precipitation in millimeters as it is the two most important weather features.

Figure 21: Number of rentals and amount of precipitation

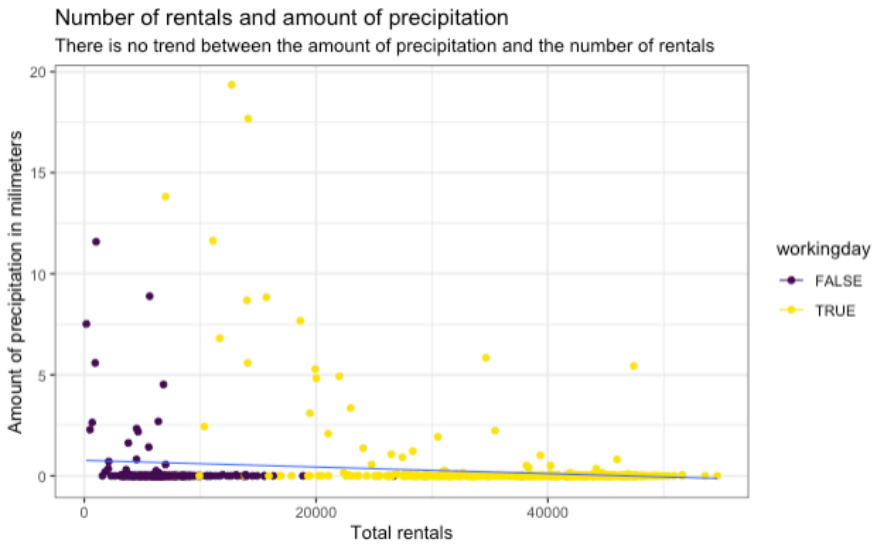
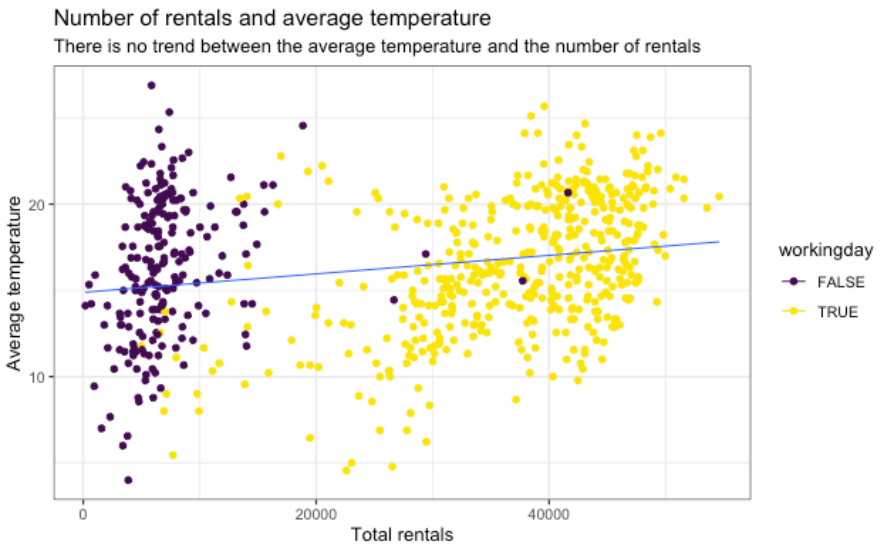


Figure 22: Number of rentals and average temperature



As we can notice in those previous graphs (Figure 21 and 22), there is absolutely no relationship between the chosen variables. It is also the case for all the other weather characteristics. One source of explanation might be that there are few rainy days in San Francisco. Actually, there were only 83 rainy days from September 2013 to August 2015.

As an indication, Switzerland experienced 124 rainy days only in 2014. It is therefore difficult to know if the amount of precipitation decreases the number of bicycle trips in San Francisco since this city is favorable for the use of bicycles due to its relatively dry climate.

Regarding the trip duration, weather have no influence either.

For further information, we suggest to refer to additional material and more specifically to the files “Report.Rmd” or “Report.html”.

Most popular routes and stations

Figure 23: Most popular routes over the whole data period

Most popular routes

ID_station_start	name.x	city.x	ID_station_end	name.y	city.y	sum
69	San Francisco Caltrain 2 (330 Townsend)	San Francisco	65	Townsend at 7th	San Francisco	6216
50	Harry Bridges Plaza (Ferry Building)	San Francisco	60	Embarcadero at Sansome	San Francisco	6164
65	Townsend at 7th	San Francisco	70	San Francisco Caltrain (Townsend at 4th)	San Francisco	5041
61	2nd at Townsend	San Francisco	50	Harry Bridges Plaza (Ferry Building)	San Francisco	4839
50	Harry Bridges Plaza (Ferry Building)	San Francisco	61	2nd at Townsend	San Francisco	4357

Figure 24: Most popular stations over the whole data period

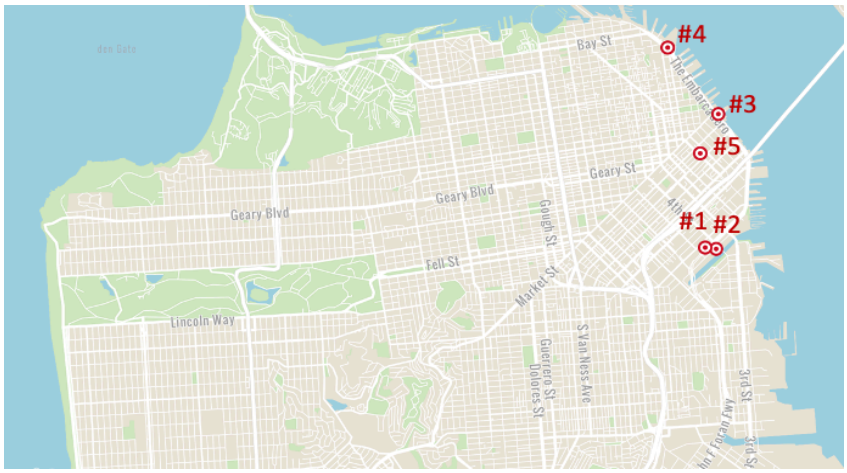
Most popular stations

ID_station_start	name	sum
70	San Francisco Caltrain (Townsend at 4th)	49092
69	San Francisco Caltrain 2 (330 Townsend)	33742
50	Harry Bridges Plaza (Ferry Building)	32934
60	Embarcadero at Sansome	27713
55	Temporary Transbay Terminal (Howard at Beale)	26089

Without any surprise, stations belonging to the most popular routes (Figure 23) are also the most popular stations (Figure 24).

Our analysis suggests that bikes may be used as a complementary mean of transport associated to public transport (Figure 25). Further research could be undertaken to determine whether bikes are frequently used in association to other modes of transportation.

Figure 25: Map of San Francisco with the most popular stations



#1 Next to the main railway station

#2 Next to the main railway station

#3 Next to an important boat station

#4 Next to an important boat station

#5 Next to the bus central terminal

3.2 Modelling

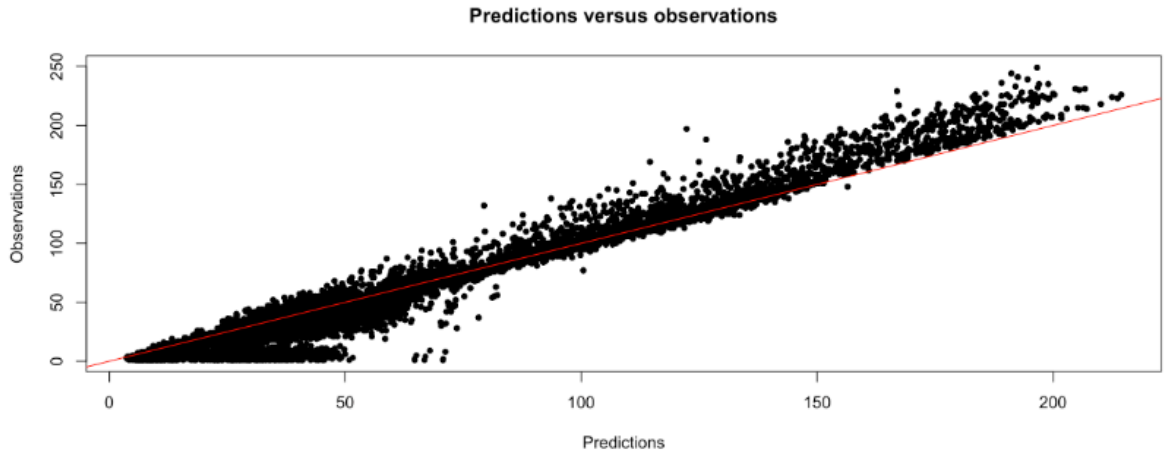
In this part, the aim was to build a model allowing us to predict the number of trips per day and to know which variables are contributing the most.

Our database is very large and has many missing values. We therefore chose to use the random forest modelling technique. It handles the missing values and gives more chance for extrapolation outside the training set.

First, we split our data into two parts: a training set and a testing set. Then, we used the random forest in order to make the predictions of the variable “count” which is the number of trips per day and per hour across all stations.

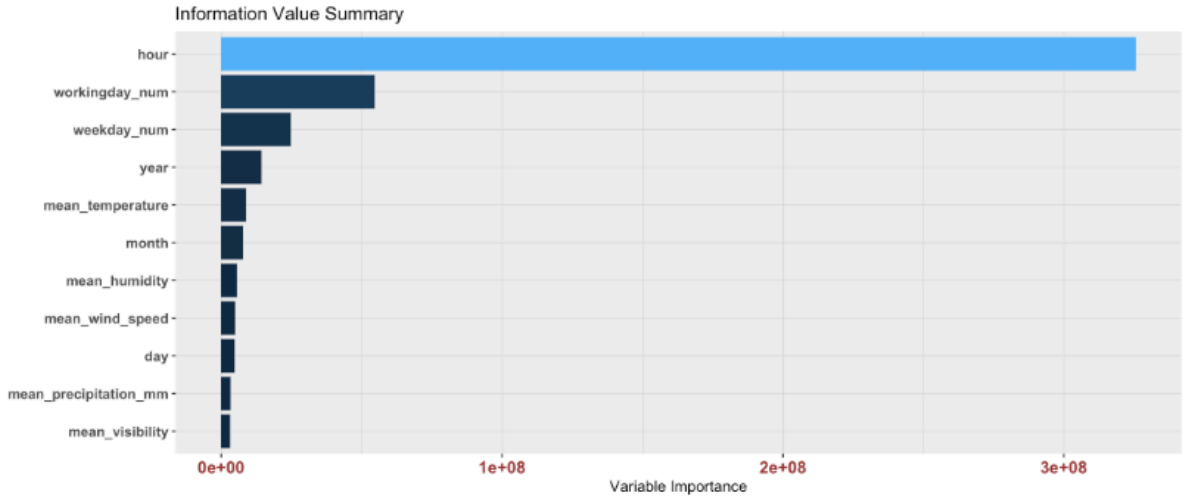
With Figure 26, we remark that our prediction is quite accurate.

Figure 26: Predictions of the random forest versus observations



Below, we plotted the importance of our variables in our random forest and it has been found that the variable “hour” is by far most essential one. It is followed by the variable “workingday_num” which designates whether the day is a working day or not and by the variable “weekday_num” corresponding to number allocated to each day of the week. In Figure 27, we observe that weather characteristics do not contribute to the number of trips per day.

Figure 27: Variable importance



To go further and to check how our random forest will react by removing the most important variables, we built a new model without the variables “hour” and ““workingday_num”. It resulted in an inaccurate prediction model (Figure 28) and confirmed the results from our exploratory data analysis highlighting that weather conditions have strictly no influence on the number of trips per day in San Francisco (Figure 29).

Figure 28: Predictions of our second random forest model versus observations

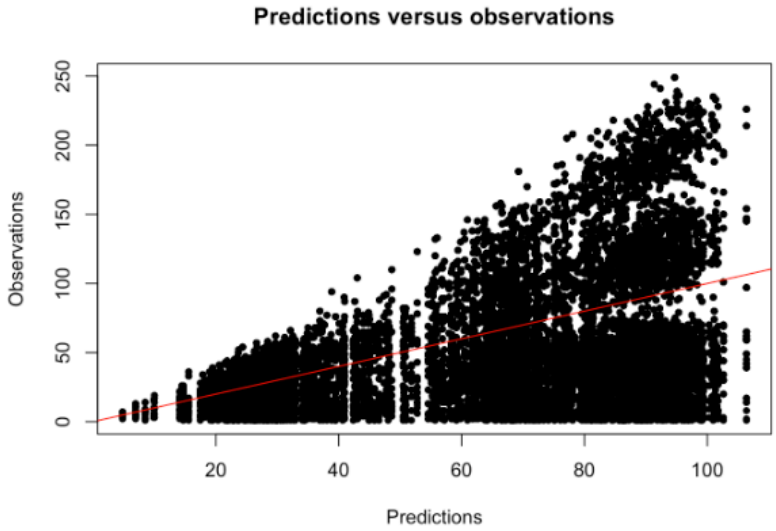
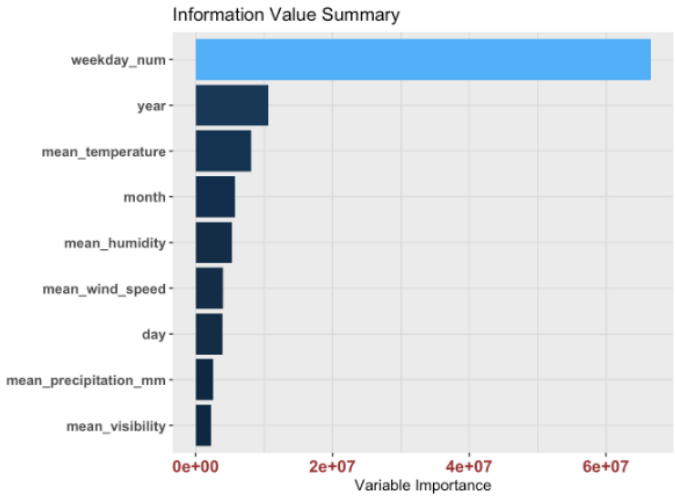


Figure 29: Variable importance of the second random forest model



4. Conclusion

From this study, we now have a better vision of bike rental in San Francisco. Actually, we have seen that rented bikes are mostly used by people who want to go to their workplace in the morning and return to their home in the evening. As a result, rentals are underexploited during the weekend and brings weekly cyclicality to the data.

Concerning the trip duration, we have observed that the most frequent duration is about 7 minutes and the average trip is about 6 to 10 minutes.

By testing the correlation between the weather characteristics and the number of journeys per day as well as its duration, we concluded that there was not necessarily a link. It has been confirmed by our modelling part and especially by the importance of the variables computed from the random forest. We also observed that the variable “hour” is the most important one. Thus, the number of trips across all stations for each day depends on the hour of the day, the weekday and whether or not it is a workday.

Also, by aggregating the data, we were able to find the stations and routes with the highest frequentation and the study showed that bikes could be used as a complementary mean of transport associated to public transport.

Further research could be undertaken to determine whether bikes are frequently used in association to other modes of transportation. Finally, the Covid-19 crisis offers an experimental framework to validate our underlying assumption. Indeed, people stayed two months in home office due to the sanitary urgency so, by analyzing the data during the crisis, we should then see a flattening of the rentals during the week as the number of people moving to their office was significantly reduced for weeks. In a reversed reasoning, we could use the bike rentals results to assess the widespread of home-office during the crisis.

5. References

SF Bay Area Bike Share. (s. d.). Consulté 5 juin 2020, à l'adresse https://kaggle.com/benhamner/sf-bay-area-bike-share

6. Additional material

- R files
  - big_data_analysis.Rproj
  - Report.Rmd
  - wrangling.R

- Excel files
  - avg_bike_available.csv
  - Fact_table.csv
  - station.csv
  - trip_fact.csv
  - weather_final.csv

- Html files
  - Report.html