|  |
| --- |
| **Factors to influence prices and demand on the platform of ride-hailing app - insights from uber deal data in Boston** |

**ABSTRACT / CONCLUSION**

|  |
| --- |
| *Conclusion*  [Comments] |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | |  |  | |
| **Author** | |  | | --- | | **Hao WANG** | | **Huang WEN-TUNG** | | **Lenny BIRACK** | | **Lin Wen-Hsin** | | **Timothée DE GAULEJAC** | | | |
| **Supervisor** | **Serge NYAWA** | | |
| **Date** | **October 22nd, 2021** | | |

**TABLE OF CONTENT**

[1 Version history 3](#_Toc85786912)

[2 Project SCOPE 3](#_Toc85786913)

[3 Project Requirements 4](#_Toc85786914)

[4 Data resources 4](#_Toc85786915)

[5 Data background 5](#_Toc85786916)

[6 our targeting issues 6](#_Toc85786917)

[7 Data dictionary 6](#_Toc85786918)

[8 Data preparation and check 7](#_Toc85786919)

[9 Storyboard 8](#_Toc85786920)

[9.1 STEP1 – Footprint searching 8](#_Toc85786921)

[9.2 Figures 10](#_Toc85786922)

[9.2.1 Unit Price / Price = f(distance, car\_type) 10](#_Toc85786923)

[9.2.2 Price = f(humidity), f(temperaure), f(days), f(hours) 10](#_Toc85786924)

[9.2.3 Unit\_Price = f(humidity), f(temperaure), f(days), f(hours) 10](#_Toc85786925)

[9.2.4 Frequency = f(distance, car\_type, source, destination, temp, humdi, days, hours) 10](#_Toc85786926)

[9.3 Modelling and prediction 10](#_Toc85786927)

[9.3.1 Price = f(car\_type, distance), weather 10](#_Toc85786928)

[9.3.2 N.rides ? = f(rush\_hours) 10](#_Toc85786929)

[10 Actions 10](#_Toc85786930)

[11 Conclusion 11](#_Toc85786931)

[12 Annex 12](#_Toc85786932)

[12.1 Problems and difficulties 12](#_Toc85786933)

[13 Gains and experiences 13](#_Toc85786934)

[13.1 Use sample to improve the speed 13](#_Toc85786935)

[13.2 Modulate all functions 13](#_Toc85786936)

[13.3 Thinking about our code 13](#_Toc85786937)

[14 Codes 13](#_Toc85786938)

**TABLE OF FIGURES**

[Figure 1. Traffic congestion of 2018 in Boston 4](#_Toc85786965)

[Figure 2. Overview of Boston in which red stars indicate 12 districts 4](#_Toc85786966)

[Figure 3. The dictionary of original datasets 6](#_Toc85786967)

[Figure 4.Schema of “merge” procedure 6](#_Toc85786968)

[Figure 5. Validation of “Merge” procedure 6](#_Toc85786969)

[Figure 6. Scatter plot (price, distance, unit\_price, car type, hour, and traffic) 7](#_Toc85786970)

[Figure 7.Scatter plot (price, unit\_price, weather information) 7](#_Toc85786971)

# Version history

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Modified paragraphs and kind of modification** |
| 01 | 07/10/2021 | All | Creation of the document |
| 02 | 15/10/2021 | All | Confirm each chapter |
| 03 | 18/10/2021 | All | Redaction contents |
| 04 | 22/10/2021 | All | Final review |

# Project SCOPE

Nowadays, driven by complicated business issues, big Data analytics are supposed to identify, collect, prepare, check and integrate different types of required data, structured and unstructured, dig out invisible insights, and create a new path to innovation. And through learning from big data analytics course, we start to get familiar with key techniques and common tools applied in the area of Big Data analytics. As we know, the prevalence of ride-hailing caters to increasing urban travel demand of citizens while simultaneously limiting the number of cars on the road. After our 1st group discussion, we hope to choose the data of ride-hailing apps like "Uber" as our topic and through relevant analysis, we hope to seize more insights, such as the main factors affecting the urban travel demand, user behavior etc. We choose, and to import a real-life data set, clean the data, and perform basic exploratory data analysis.

# Project Requirements

**

# Data resources

|  |  |  |
| --- | --- | --- |
| Data resources | Authors / Site | Link |
| Uber & Lyft Cab prices to predict cab prices against weather | Kaggle | [Link](https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices) |
| Full-year historical traffic data | TomTom | [Link](https://www.tomtom.com/en_gb/traffic-index/boston-traffic/) |

# Data background

Based on our requirement, we find three relevant datasets about the city of Boston in America. The first dataset is “car\_rides.csv” which provides the necessary information of each transaction; the second one is “weather.csv” which consists of rain, wind, humidity etc. Both come from “Kaggle” website. And the last one is the annual statistic table of weekly traffic congestion by time of day of 2018 in Boston, as shown below. And we transfer this image to a csv file in order to import it under RStudio.

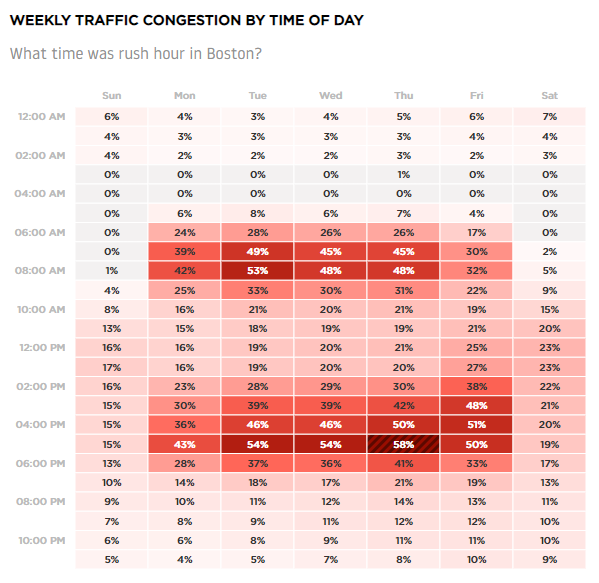


Figure 1. Traffic congestion of 2018 in Boston

Here we list several key features of our datasets:

1. Three .csv files include various information
2. The size of "cab\_rides.cvs” file reaches 89MB
3. “Cab\_rides.csv” contains 690k transaction information.
4. The transaction data covers the period of three weeks, from November 25th to Decembre 18th, 2018 and there was no public holiday during that period.

Finally, here is some information about our researching object – “City of Boston”. Boston is the capital of Massachusetts in US, with a population of 675k in 2020. In Boston, streets and sidewalks make up 56% of the land but traffic congestion in Boston ranked 2nd worst in US in 2020. Therefore, it could be interesting to analyze riders’ behavior in using ride-hailing app during the rush hour. And our dataset “cal\_rides” it mainly consists of the deals within twelve districts as shown in the following figure.



Figure 2. Overview of Boston in which red stars indicate 12 districts

# our targeting issues

The objective is to get a general overview of Uber market in Boston and to look for more insights behind these data

Ride hailing company - surge pricing / dynamic pricing /

From the user’s perspective, they really urge to have clear answers to two questions:

1. Whether I can get the cab once I would like to go, and the fare should also be acceptable to me.
2. Whether the app can tell me the credible fare of the pre-defined trajectory in one future time.

From the driver’s perspective, the most interesting thing is to maximize their efficiency, which means gain the most payment within the shortest time. Therefore, they would like to know in advance the request of cab from users to help them avoid wandering along the streets.

# Data dictionary

After the data preparation (more details in the following chapiter) of the above three datasets “car\_rides”, “weather” and “rushhour”, we get the final dataset with 17 fields and 356 109 observations. Here is the dictionary to present all necessary information.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **N°** | **Name of fields** | **Type** | **Range** | **Comments** |
| 1 | distance | Numeric | 0.02 – 7.86 |  |
| 2 | cab\_type | Category | Uber |  |
| 3 | destination | Category | North Station / Northeastern University / West End / Haymarket Square / South Station / Fenway / Theatre District / Beacon Hill / Back Bay / North End / Financial District / Boston University |  |
| 4 | source | Category | Haymarket Square / Back Bay / North End / North Station / Beacon Hill / Boston University / Fenway / South Station / Theatre District / West End / Financial District / Northeastern University |  |
| 5 | price | Numeric | 4.5 – 89.5 | Unit: US Dollars |
| 6 | Surge\_multiplier | Category | 1 | We only have one value here. Normally, surging price is one of the most important features for Uber. |
| 7 | name | Category | UberXL / Black / UberX / WAV / Black SUV / UberPool |  |
| 8 | temp | Numeric | 19.62– 55.41 | Unit: Fahrenheit |
| 9 | clouds | Numeric | 0 - 1 | The higher the value means that there’s more clouds |
| 10 | pressure | Numeric | 988.25 - 1035.12 | Unit: mbar |
| 11 | rain | Numeric | 0.0002 - 0.7807 | The higher the value means that there’s more rain |
| 12 | humidity | Numeric | 0.45 – 0.99 | For example, 0.99 corresponds to 99% of humidity |
| 13 | wind | Numeric | 0.29 – 18.18 | Unit: km/h |
| 14 | n\_pick\_up\_time | Date | YYYY-MM-DD HH:MM:SS | Column added – Generated from time\_stamp for the local  time in Boston |
| 15 | n\_days | Category | Days of the week | Column added |
| 16 | n\_hour | Numeric | 0 - 23 |  |
| 17 | traffic | Numeric | 0 – 0.58 | Column added – The higher the value, the worse the traffic condition |

Table 1. Data dictionary after “merge”

Here we attach the dictionary of all original datasets in case of any questions.



Figure 3. The dictionary of original datasets

# Data preparation and check

First, before starting the “merge” procedure, we need to transform Unix timestamp to Boston local time. In addition, in “weather” file, the “rain” field contains many “NA” value and in order to evaluate the influence of weather on prices and deals later, we replace NA by zero. Because zero doesn’t exist before in this “rain” field and we can easily recognize rain = NA later if its value equals to zero.

Then, the main time-consuming part is the “merge” procedure. As shown in Figure 2, we merge the three files one by one and for each “merge” step, we construct different key words.

After the “merge” procedure, we check also the final output through comparing it to the original one. For example, as shown in Figure 3, we can find the numbers of deals per day stay the same as before, which means no valid observation from “cab\_rides” is deleted during this “merge” procedure.

Diagram

Description automatically generated

Figure 4.Schema of “merge” procedure

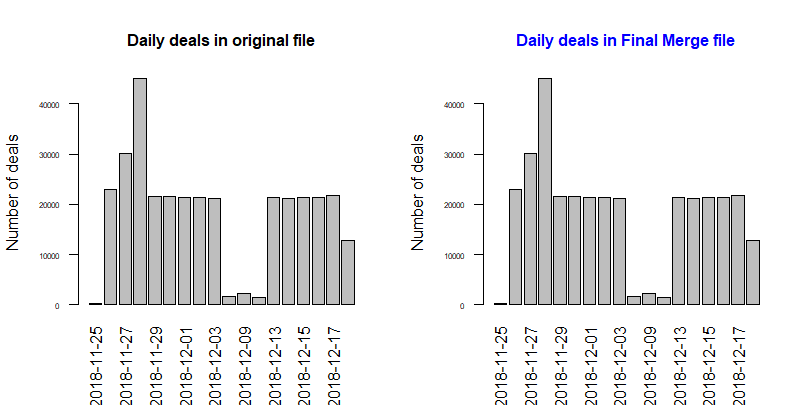


Figure 5. Validation of “Merge” procedure

# Storyboard

## STEP 1 – We search footprints !

To have a global view about which factors could affect the price and unit price of each deal, we demonstrate the relationship between Price, Distance, Unit price, name(car\_type), hours, traffic, and weather conditions as shown in the following two scatter figures. From these plots, we can see:

* Distance and car type has influences on the price. The price increases as the distance increases. Very logical!!
* Unit price decreases as distance increases.
* Price and unit price have no obvious relationship with the weather condition. For us, that phenomenon is not logical. Normally, with bad weather, there are

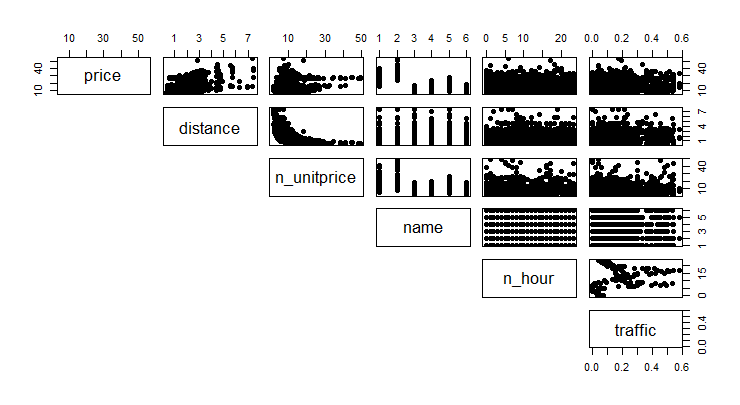


Figure 6. Scatter plot (price, distance, unit\_price, car type, hour, and traffic)

Diagram

Description automatically generated

Figure 7.Scatter plot (price, unit\_price, weather information)

Besides, we also check the influence of district and days on uber cases. As shown in the following figures,

|  |  |
| --- | --- |
|  |  |

## STEP 2 – We want to dig out more!

In order to understand the relation in price and unit price as function of distance, here we demonstrate the two figures down below.

|  |  |
| --- | --- |
|  |  |

Figure 8. Price, Unit Price, and distribution of cases as function of distance

From the above figure, we can find:

* The unit price decreases dramatically when distance increases from 0 to 3 miles, and finally it keeps stable at around 5 dollars per mile once the distance is higher than 3 miles.
* Most of cases of distance are between 1 mile and 3 miles.
* The reason why we don’t have lots of cases at 2 miles is because short distance is 1.5 miles and long distance is more than 2 miles mostly, that’s why we don’t have exact same distance.
* Fe

|  |  |
| --- | --- |
|  |  |

Figure 9.

As we can see above in the first picture, we have the relationship between the number of cases and the distance based on the car types.

When the distance is fewer than 1 mile

* The "Black SUV" Uber is not used for small rides. Most number of cases of "Black SUV" Uber are between 1 and 5 miles.
* The cases of "Black" Uber are minimal comparing to others for long distances.

When the distance is between 1-2 miles

* The "Black SUV" doesn't reach its maximum value in terms of number of cases. "Black SUV "is, still, the less used comparing to others.

When the distance is higher than 2 miles

* The number of cases by car type is approximately divided into equal groups.

Moreover, the usage of Uber become less wanted when the distance is higher than 5 miles.

As we can see above in the second picture, we have the relationship between the price and distance based on the car type.

* Comparing to other car type, "UberPool" remains the cheapest Uber even when the distance increases.
* "Black SUV" is the most expensive Uber car type, it starts with 23$ and with a distance of 1.08 miles.

## STEP 3 – Now let’s predict!

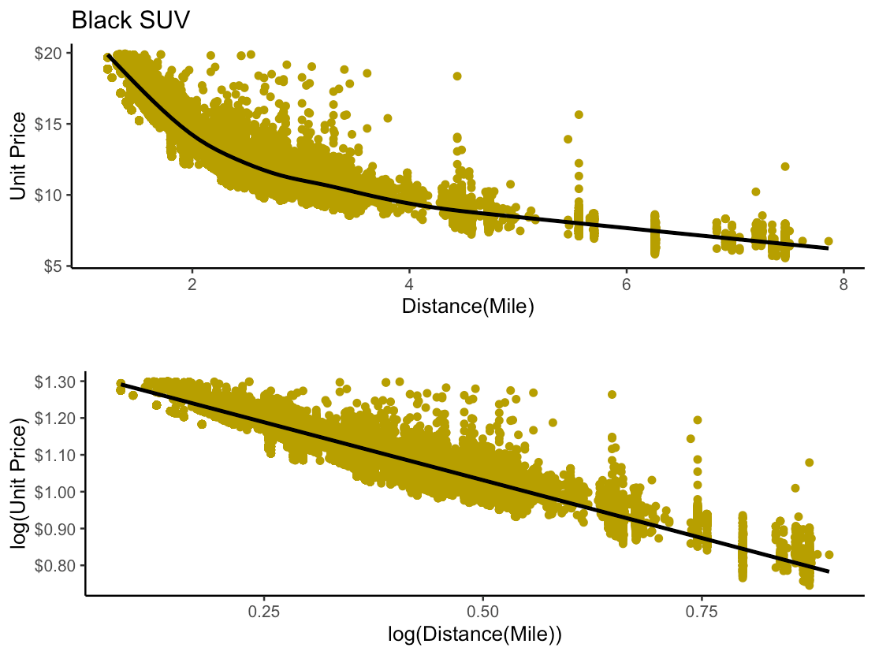
### Modelling of Price

|  |  |
| --- | --- |
|  |  |

From the above two figures, we find that the unit price seems like an inverse function of distance. Therefore, we try to build the model for unit price.

### Modelling of unit price

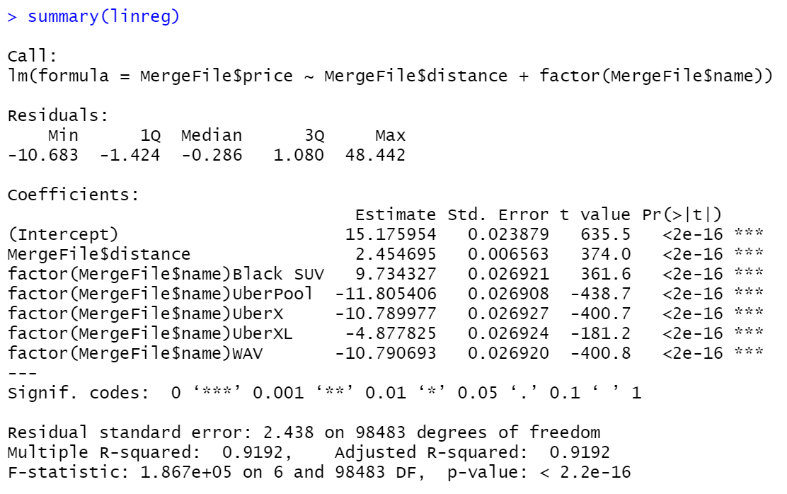
Since we can not use linear regression for an inverse function, we suppose



Question

Not date between 5 and 6 miles

### Model prediction



We can study the linear regression of our variable price with the other significant variables which are the distance and the car type.

The equation is:

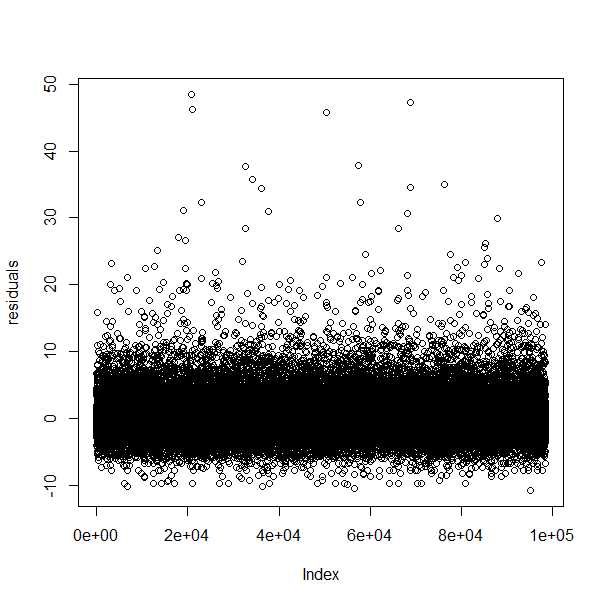
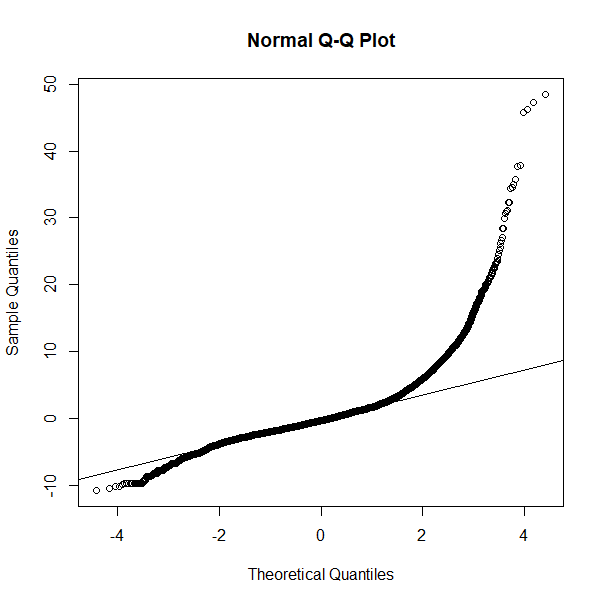
Price = 15,18 + 2,45\*Distance + 9,73\*Black SUV -11,81\*UberPool - 10,79\*UberX - 4,88\*UberXL - 10,79\*WAV

The adjusted R-Squared is 0,92, which means that 92% of the variation of the price is explained by the distance and the type of car in our model.

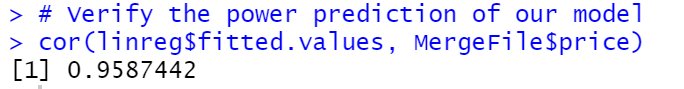
We can say that our model is representative.

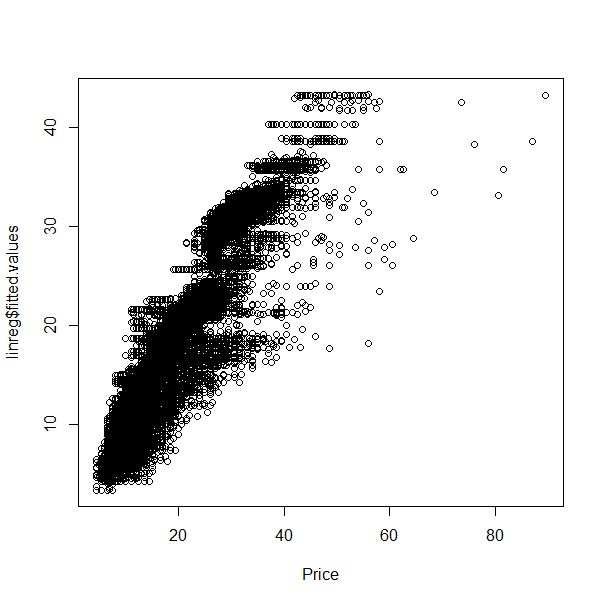
Now, we analyse the residuals of our model.

We check that the residuals of our regression are normally distributed and not autocorrelated.



The first graph shows that the residuals are normally distributed, and the second graph shows that they are not autocorrelated.





With the function cor, we can check the power prediction of our model.

Our model is powerful because it could give 96% of good predictions.

We can start making predictions! We make a simulation of our predictive model. Let’s take an exemple.

In Boston, taking an Uber for a 3,2 mile-trip with a Black SUV would cost:

Price = 15,18 + 2,45\*Distance + 9,73\*Black SUV -11,81\*UberPool - 10,79\*UberX - 4,88\*UberXL - 10,79\*WAV

Price = 15,18 + 2,45\*3,2 + 9,73\*Black SUV

Price = 32,75 $

# Actions

Based on our refined analysis, in Boston, Uber should take into consideration the recommendations below:

- Uber should make great deals for a distance less than or equal to 1 mile in "Black SUV" and "Black" Uber; This will encourage people to use a "Black SUV" for short rides and increase the demand for "Black" Uber.

- People usually think of public transportation because Uber is expensive for long trips, For the distance longer than 5 miles, Uber should make reasonable discounts to motivate people to benefit from uber (even if it's a long drive).

# Conclusion

# Annex

Problems and difficulties

* + - 1. The first difficulty was to start the analysis without the clean dataset. To gain some time, we started the analysis before the cleaning of the dataset. Once we had the clean dataset, it was easier to test our codes and to find all the analysis we wanted to do.
      2. At first glance, we did not know the meaning of the variable surge multiplier. We did some research to have more information and define if this variable is relevant in our analysis.
      3. There is a surge multiplier when the demand of Uber is too high compared to the number of cab drivers available at that time. The original price is multiplied by a coefficient and the price gets higher. <https://www.uber.com/us/en/drive/driver-app/how-surge-works/>
      4. Some analyses seemed interesting with the variable surge multiplier but we realized that once the dataset was cleaned, this variable was not relevant anymore. When we removed Lyft from the cab types, we realized that there was no surge multiplier. The surge multiplier was always equal to 1.
      5. Absence of rain information in weather file and incoherence between rain and humidity

# Gains and experiences

## Use sample to improve the speed

## Modulate all functions

Thinking about our code

# Codes