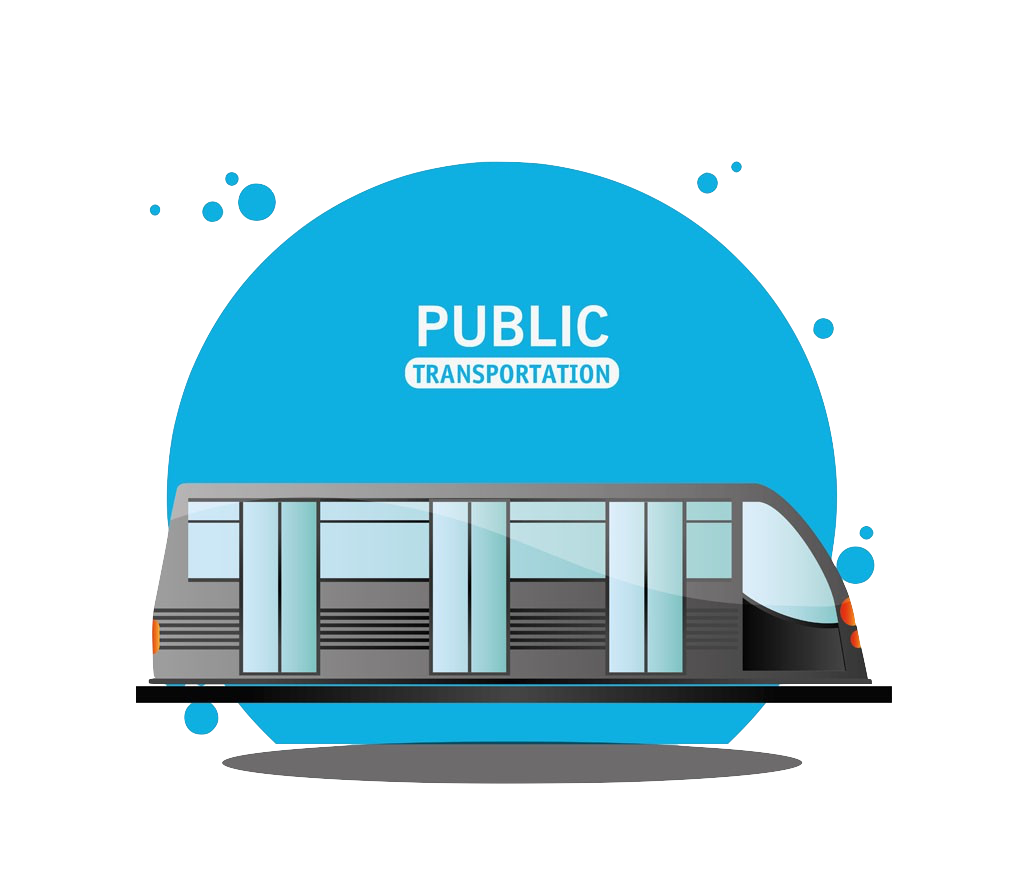
**RER SUCKS**

****

**Big Data Analytics**

**Final Report**

Members:

Dongxue SHAO, Tailai ZHANG

Xiaoman LIU, Xiaofeng XU

2017.12.30

# **- CONTENTS -**

1 Background and Motivation 3

1.1 Project Background 3

1.2 Problem Definition and Motivation 3

2 Previous Efforts 3

3 Data Source 4

3.1 Build the Data Set 4

3.2 Supporting Data Sets 5

4 Model Performance 6

4.1 Feature Engineering 6

4.2 Model Performance 8

4.3 Key Findings 11

5 Limitations and Future Work 15

6 Conclusion 15

Appendix 16

# 1 Background and Motivation

## 1.1 Project Background

The Paris Metro is a rapid transit system, a symbol of the city. It has 303 stations, representing 214 kilometers. There are 16 lines, numbered 1 to 14 with two lines, 3bis and 7bis. Another important rapid transit system in Paris, the Regional Express Network, colloquially called the RER, is a hybrid suburban rapid transit system serving Paris, France and its suburbs. The network consists of five lines: A, B, C, D and E, having 258 stations and representing over 587 kilometers. It has several connections with the Paris Metro within the city of Paris.

The Paris Metro carried 1.698 billion passengers in 2016, 4.65 million passengers a day. The number in 2015 was 1.520 billion passengers a year and 4.16 million passengers a day. Carrying over 20% of the overall traffic in Paris, Paris Metro is one of the densest metro systems in the world. The RER carried 5.65 million passengers a day in 2016. The number of passengers of RER is bigger than the Paris Metro’s.

## 1.2 Problem Definition and Motivation

Having such a large traffic, it is inevitable to have breakdowns or interruptions. As everyone knows, 8 out of 10 Parisians complained about Paris Metro or RER systems. Beijing Metro, carrying 10 million passengers a day in 2014, had less than 50 breakdowns a year. Definitely, Beijing Metro was built 71 years later than Paris Metro. However, except technical reasons, there should be other reasons causing the frequent breakdowns in Paris Metro, such as management level, weather influence, and etc.

Metro and RER are arteries for urban traffic in Paris so they are of fundamental importance. Smooth traffic is a basic guarantee of healthy economic vitality. Frequent breakdowns in Metro and RER systems not only arise complaints among citizens but also lower efficiency and damage productivity.

The main goal of our project is to find out the breakdown details in Paris Metro and RER and use machine learning algorithms to analyze the relation between metro breakdowns and all possible information (weather, traffic volume, arrondissements). We try to answer the question: whether it is possible to predict the reason of breakdown using all these information? We hope that the analytics results could provide inspirations to RATP and SNCF, helping them improve the performance of Paris Metro and RER system, thus saving their operating costs and improve the efficiency of the whole society.

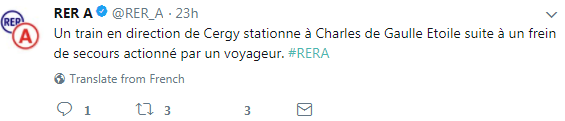
# 2 Previous Efforts

There is some previous work about the fare collection system, quality and security of Paris Metro. For example, Michel Jacoub proposed a computerized automatic fare collection system for Paris Metro in 1975 and Pierre GRIFFE studied how automation will enhance quality and security of meteor 14th line of Paris Metro in 1997. There is few study or project about Paris Metro, let alone about the breakdown of Paris Metro. Consequently, we found nothing we can learn from and there was even no existing data set.

# 3 Data Source

## 3.1 Build the Data Set

As we mentioned in ***2. Previous Efforts***, there is no existing data set of metro/RER breakdowns. We have tried Paris Open dataset[[1]](#footnote-1), RATP[[2]](#footnote-2), SNCF[[3]](#footnote-3) Open dataset, but haven’t found any datasets nor research of the delay or annulation of the public transport. Inspired by some machine learning cases based on Twitter, we thought it might be possible to web scratch all useful tweets. We found out that every Metro line and every RER line in Paris has its own twitter account. Each time there is a breakdown in a line, corresponding twitter account will tweet a twitter to state the breakdown detail: when, where and why there is a breakdown. We can scratch these twitters and extract useful information.

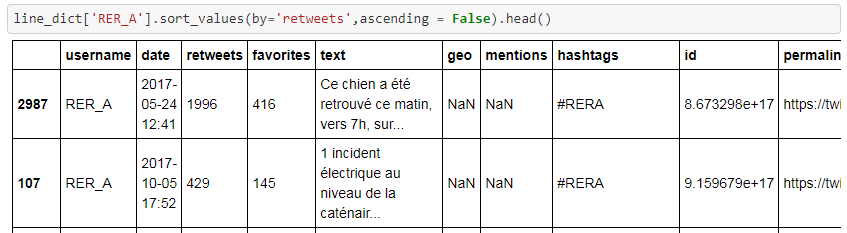


**Figure 1.** RER A twitter official account

As we learnt from the Big Data Analytics class, data scientists spend most of their time cleaning and collecting their data. We spent most of time building our own web scrawler and cleaning (text mining) our data. We wrote several lines of Python code (the code could be found in the html. file) to scratch all the tweeters of 14 Metro lines and 5 RER lines. To make them consistent, we selected the twitter from 1st Jan. 2014 to 21st Oct. 2017. In the end, we got a data frame containing 175,044 rows. The data frame consists of:

* username: which line tweeted the tweeter
* date: when the twitter was tweeted
* retweets: the number of retweets
* favorites: the number of favorites
* text: the content of the twitter
* geo: where the twitter was tweeted
* mentions: the twitter mentions what accounts

A 175,044 x 7 data frame was built and used for following analytics.



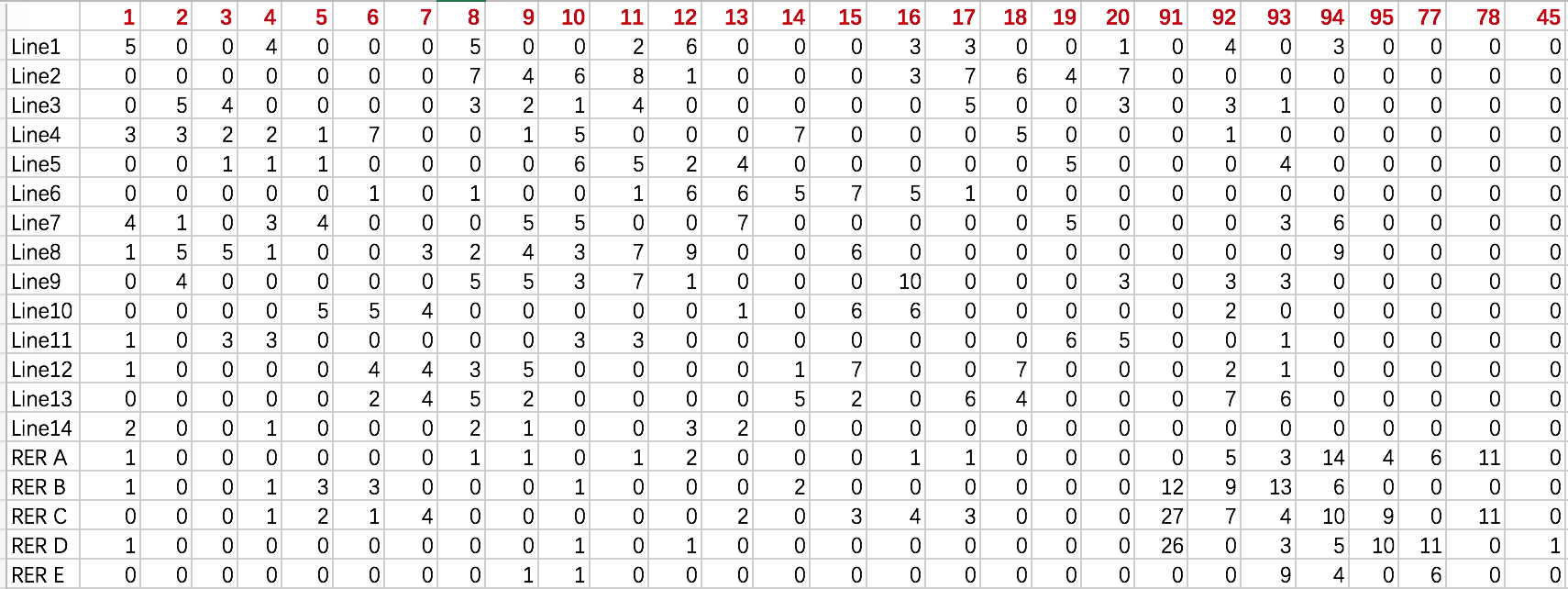
**Figure 2.** Data Frame Preview

## 3.2 Supporting Data Sets

Weather is an important factor that could affect the operation of Metro and RER systems. To see the relation of other environmental factors with breakdowns, we collected the weather information from a website[[4]](#footnote-4). These variables listed below are took into account when predicting whether a breakdown happen or what causes a breakdown in a given day:

* average temperature
* average visibility
* average wind speed
* rain or not
* fog or not
* snow or not
* thunderstorms or not

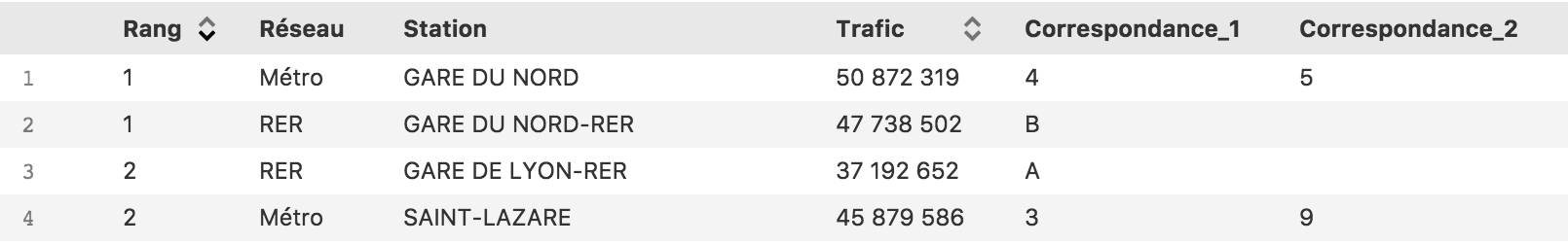
Besides, we also considered that which arrondissements passed through are also linked to the delay/cancellation of the public transportation. Therefore, we calculated the number of the stations of each line in each arrondissement. The number in figure 3. representing the number of stations.



**Figure 3.** Data Frame Preview

Finally, we decided to take traffic volume into consideration. We found a data source from RATP Open Data[[5]](#footnote-5):

* Station Name
* 2016 Total Voyagers Number



**Figure 4.** RATP Open Data Preview

We allocated different stations into their related lines and got 2016 Total Voyagers Number by each line.

**Table 1.** Total voyagers number by each line in 2016

|  |  |
| --- | --- |
| Line | Traffic-2016 |
| Line 1 | 217255582 |
| Line 2 | 125819494 |
| Line 3 | 153695350 |
| Line 4 | 215968806 |
| Line 5 | 185094089 |
| Line 6 | 127025812 |
| Line 7 | 83799073 |
| Line 8 | 140779004 |
| Line 9 | 93998611 |
| Line 10 | 69930227 |
| Line 11 | 87001478 |
| Line 12 | 148979688 |
| Line 13 | 94837098 |
| Line 14 | 143760624 |

# 4 Model Performance

## 4.1 Feature Engineering

As most of the breakdown information are extracted from the “text” column (tweets) from the data set, our feature engineering process mainly focus on text mining and eliminating noise.

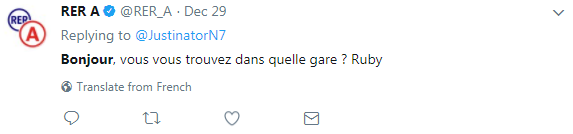
*4.1.1. What are most frequent words for each line*

Firstly, we normalized all abbreviations and accents, such as:

* Pte 🡪 Porte;
* Chateau, chateau 🡪 Château;
* électr. 🡪 électrique;
* tvx. 🡪 travaux.

Secondly, due to different lines are managed by different companies, their twitter accounts have diverse patterns of publishing tweets: SNCF lines always send their traffic related tweets with a hashtag of *“Infotrafic”*.

Thirdly, we also found that all twitter accounts interacted with other twitter accounts and all these discussions are not related to the delay/cancellation of the public transportation:



**Figure 5.** RER A tweets preview

Besides, we standardized all incidents:

* bagage oublié, colis suspect 🡪 bagage abandonné;
* voyageur malade 🡪 malaise voyageur;
* pb signalization 🡪 panne de signalisation ;
* jets de pierre, jets de pierre, jet de projectile 🡪 acte de malveillance.

Lastly, we applied Natural Language Processing techniques to realize text mining:

* add stopwords, such as, “rera”, “ratp”, “ligne”, “entre”, “http”, etc.;
* drop all punctuation;
* lower case all words;
* build a word counter in words\_freq function.

*4.1.2. Who are most frequent stations for all tweets*

We downloaded all RER/metro stations from SNCF Open Data[[6]](#footnote-6) and built a counter in gare\_freq function.

This feature also implied that these stations are most likely to encounter the delay/cancellation issues.

*4.1.3. What are the reasons for the incidents*

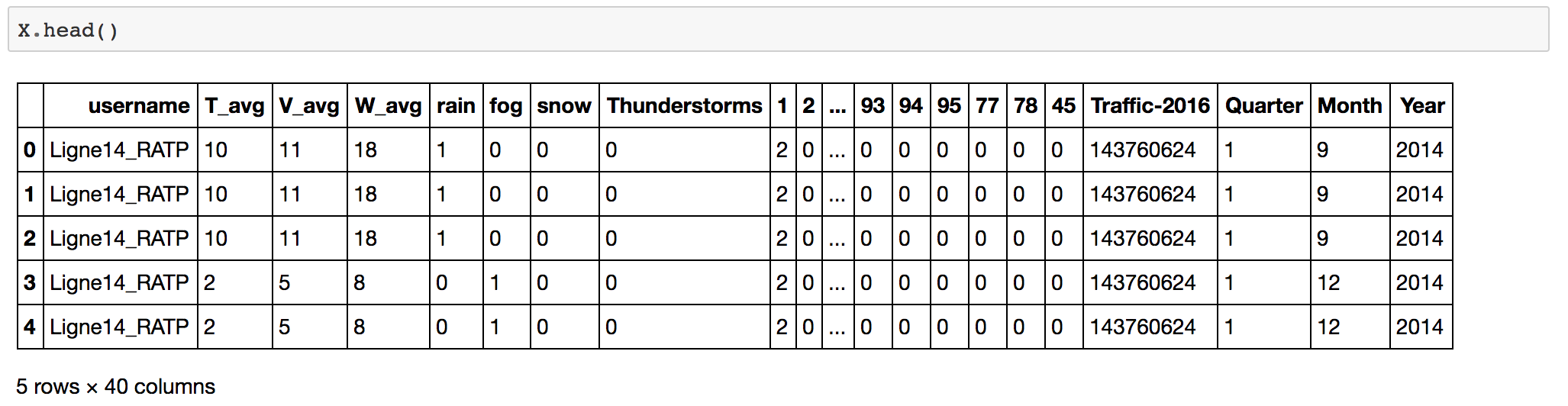
We haven’t come up with a more time efficient and accurate way to find all reasons for different incidents and decided to use an exhaustive method to list all possible reasons:

* 'malaise voyageur'
* "incident d'exploitation"
* "incident technique"
* 'Incident de signalisation', etc.

Then we built a word counter in incident\_freq function.

*4.1.4. Predict the incident reasons for a certain line in a certain date*

In order to get a better performance, we standardize and scale (Min Max Scale) the dataset before runing the Machine Learning algorithms. Thus, we got our final training/test dataset:

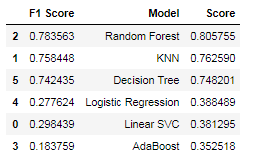
****

**Figure 6.** Final trainging/test dataset

## 4.2 Model Performance

As our primary objective of forecasting the reason of a breakdown of RER/metro lines is a classification problem, we decided to use the following Machine Learning models:

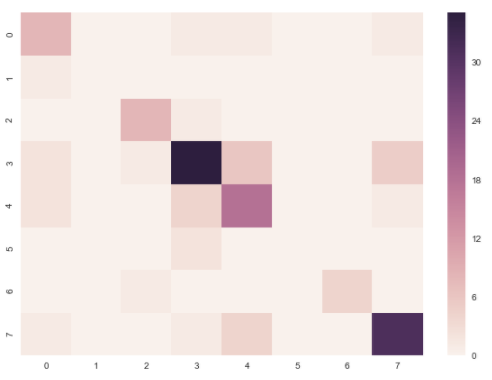
* Decision Tree
* Random Forest
* AdaBoost
* KNN
* Linear SVC
* Logistic Regression



**Figure 7.** Performance Summary

*4.2.1 Decision Tree*

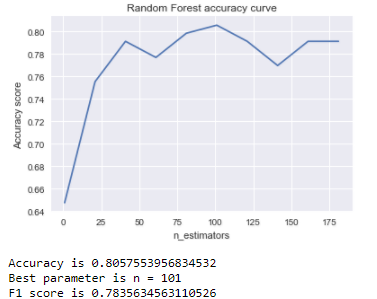
With a F1 score of 0.75, Decision Tree has a modest performance to predict the reason of traffic incident for a certain line in a certain day.



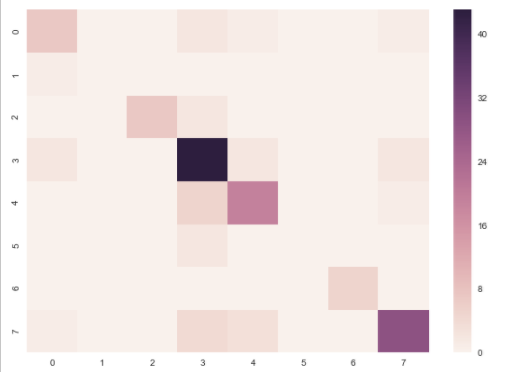
**Figure 8.** Confusion Matrix

*4.2.2 Random Forest*

KNN achieved the best performance of 61.2% (accuracy) when n\_estimator = 101



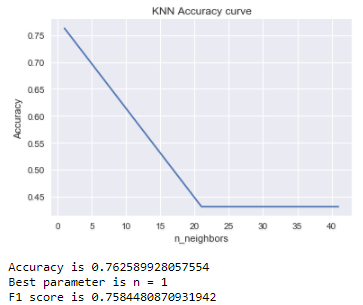
**Figure 9.** Random Forest Accuracy curve



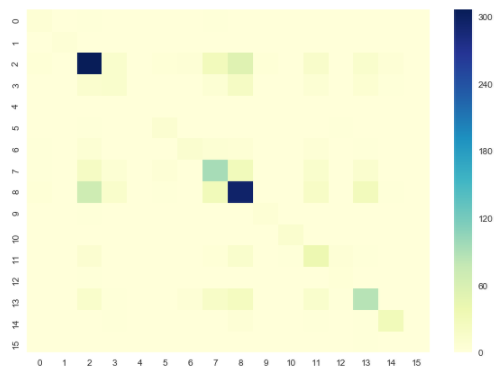
**Figure 10.** Confusion Matrix

*4.2.3 KNN*

Surprisingly, we found out that we got the best performance while n=1.

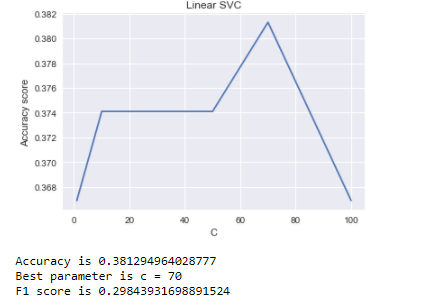


**Figure 11.** KNN Accuracy curve

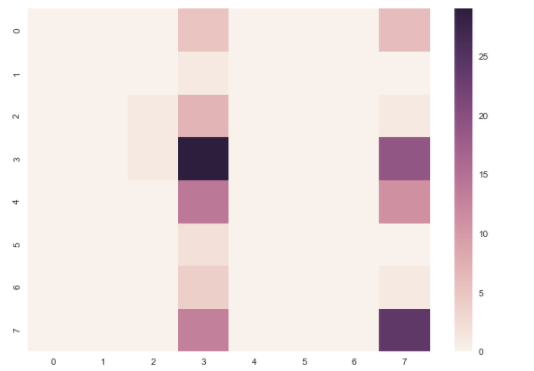


**Figure 12.** Confusion Matrix

*4.2.4 Linear SVC*

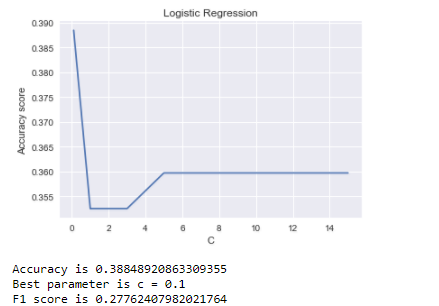


**Figure 13.** Linear SVC Accuracy curve

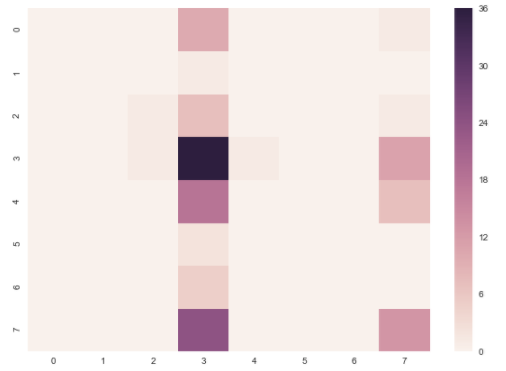


**Figure 14.** Confusion Matrix

*4.2.5 Logistic Regression*



**Figure 15.** Logistic Regression



**Figure 16.** Confusion Matrix

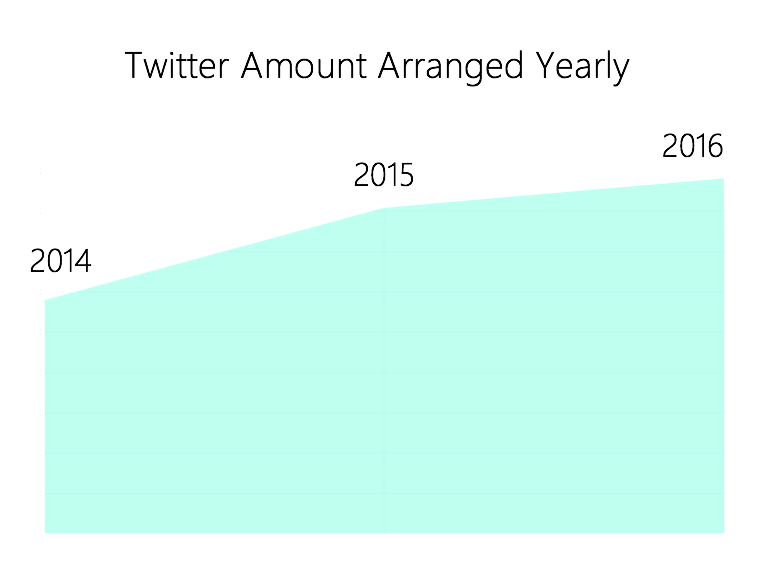
## 4.3 Key Findings

Apart from predicting the breakdown reasons of each line in a given day, there are also many interesting findings by analyzing our data set.

*4.3.1 Time series analysis*

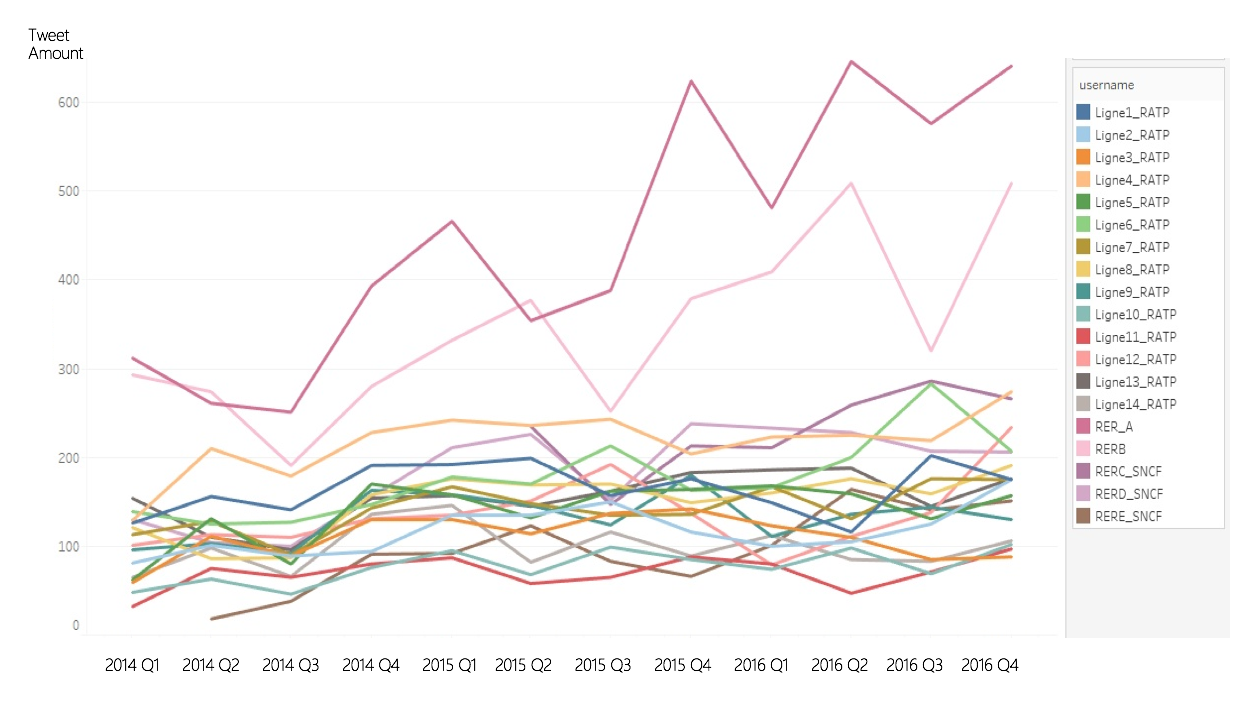
When and only when there is a breakdown or interruption in a line, its official twitter account will tweet a (sometimes two) relevant twitter(s). Although it is impossible to see directly how many breakdowns were happened, what we can make sure is that the number of breakdowns is in direct proportion to the number of twitters. According to this principle, we can carry out more specific exploratory analysis.

We aggregated the twitter amount of all accounts and then arranged them by hour, by month and by year. Figure 15. shows that the total twitter amount of all accounts was increasing from 2014 to 2016, which indicates that there were more and more breakdowns year by year.

****

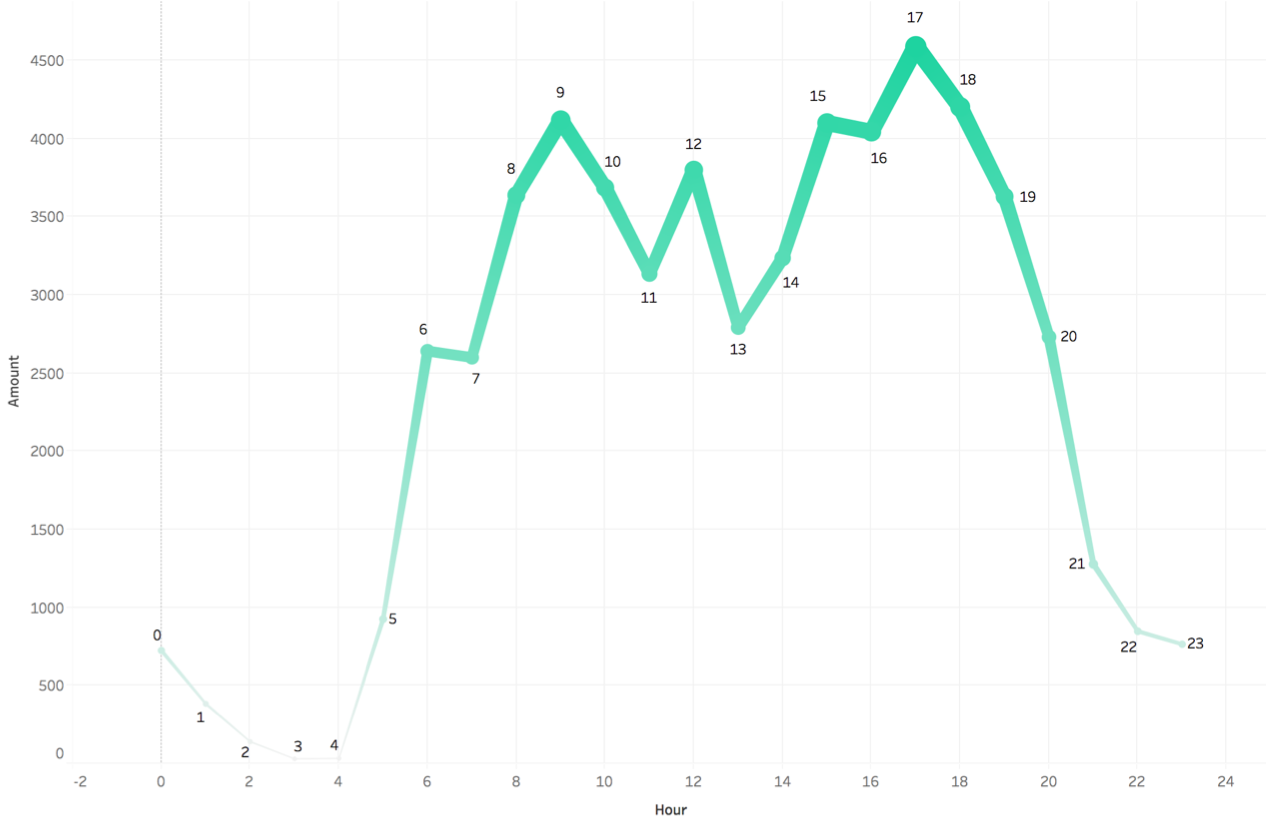
**Figure 17.** Total Twitter amount of all accounts arranged by year

Figure 17. demonstrates the twitter amount of all accounts every month. Obviously, the data has strong seasonality.



**Figure 18.** Total Twitter amount of all accounts arranged by month

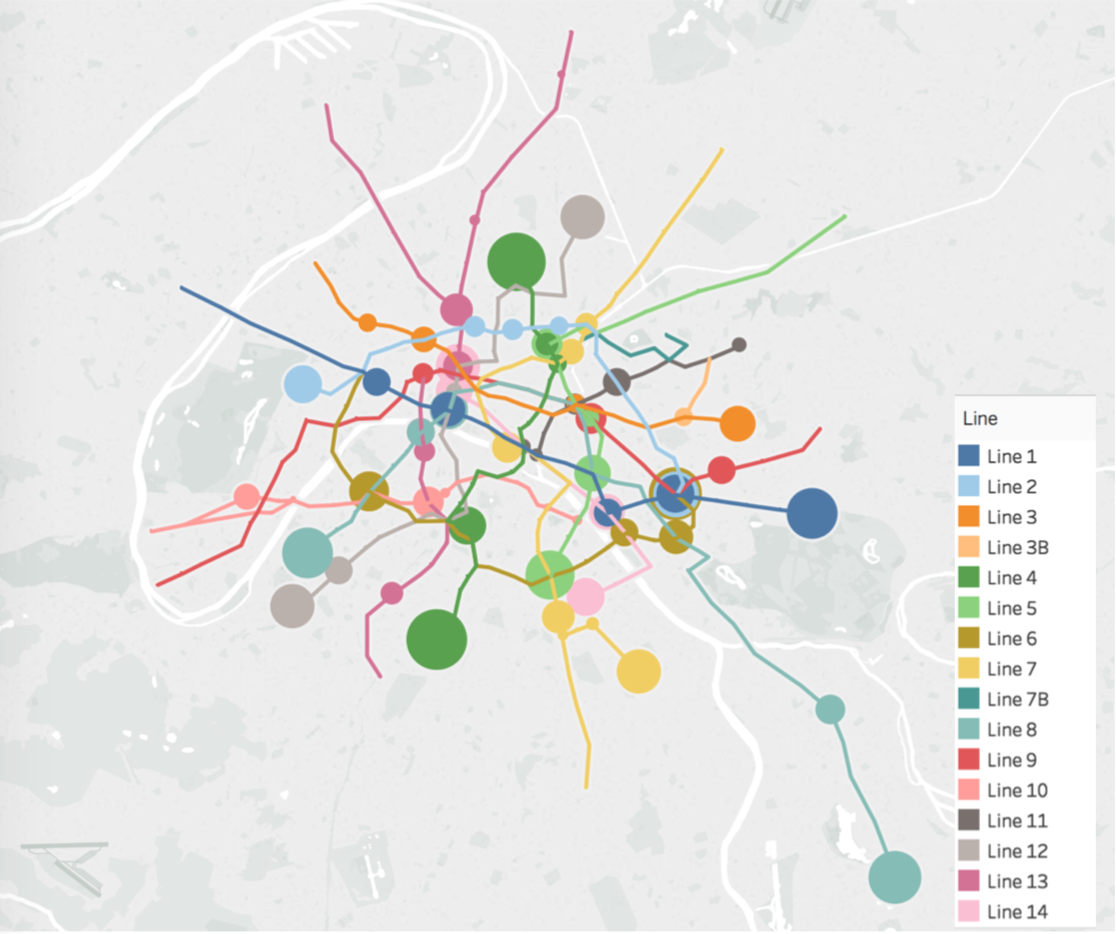
Figure 19. shows the twitter amount of every hour. The same as what we expected, there were more breakdowns in rush hours, such as 9:00 And 17:00.

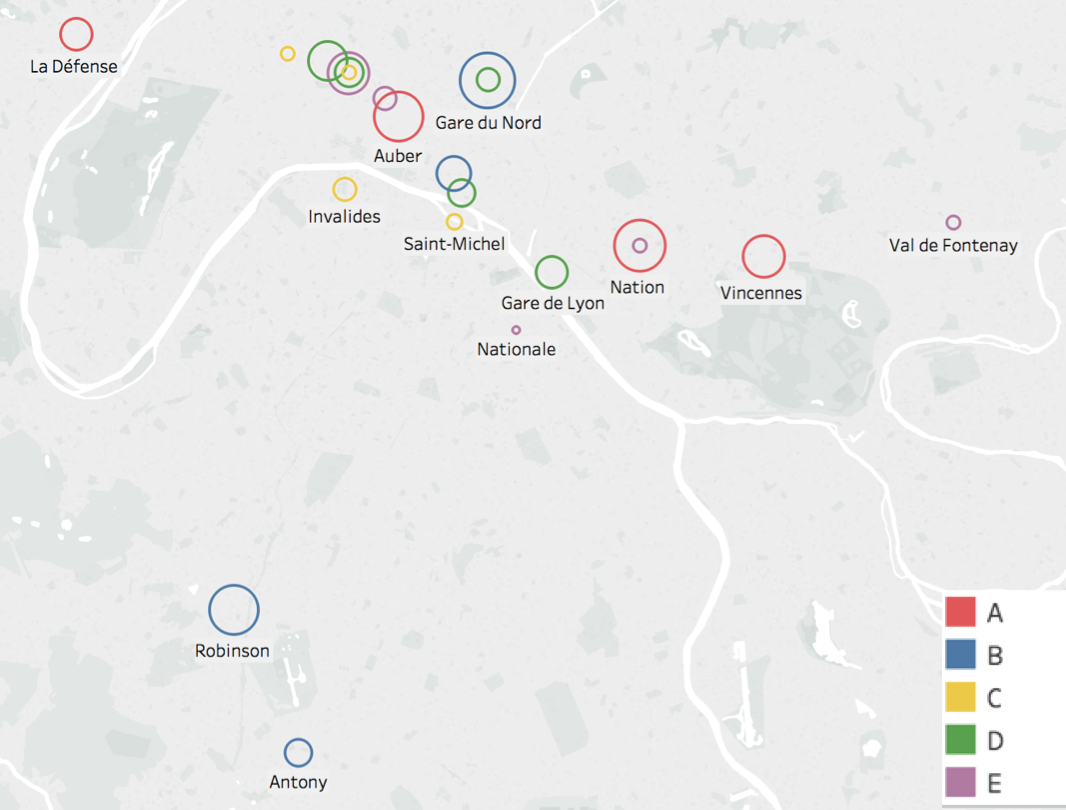


**Figure 19.** Total Twitter amount of all accounts arranged by hour

*4.3.2 What are those stations having most breakdowns?*

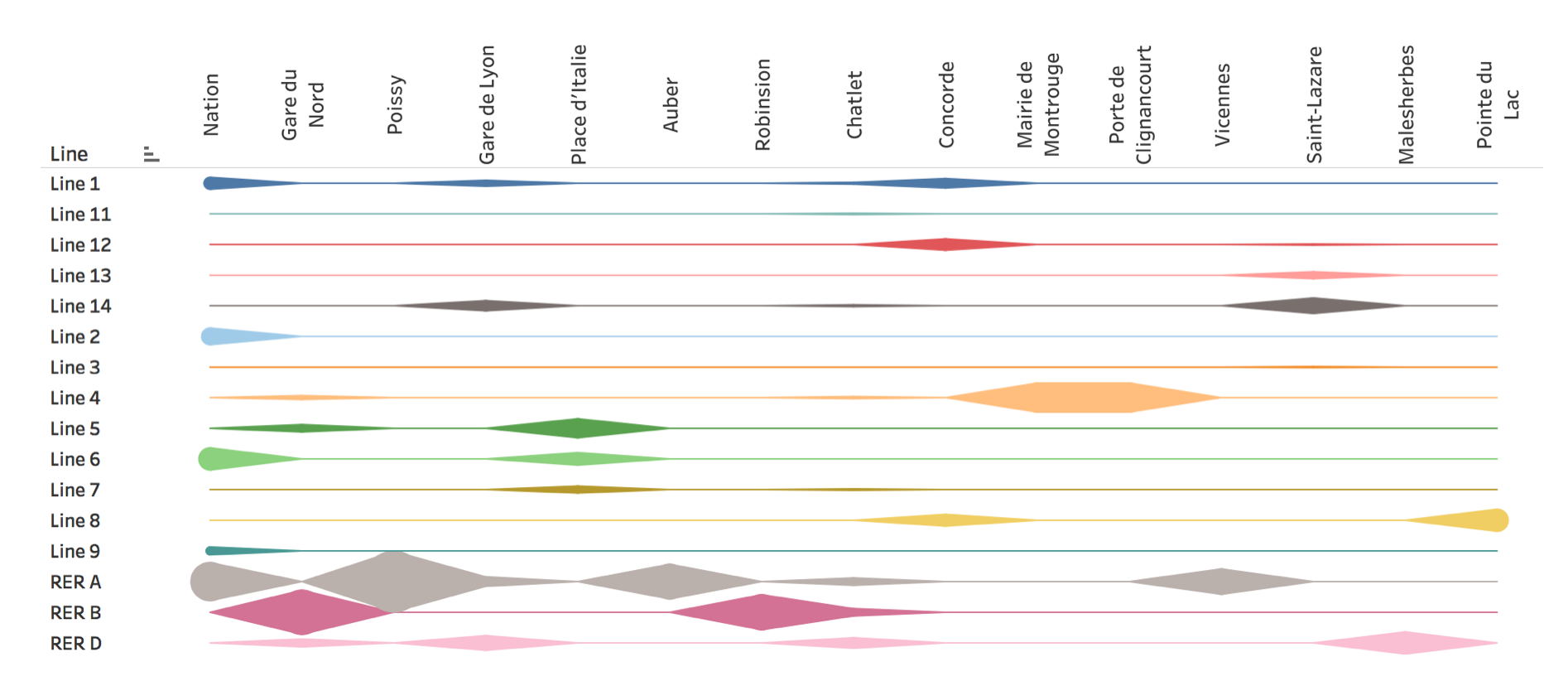
We made maps to show the most frequent mentioned stations of every metro and RER. Most mentioned could be regarded as most breakdowns. More details could be found in appendix 1 in the last part of the report.

  
**Figure 20.** The 5 most frequently mentioned stations of every metro



**Figure 21.** The 5 most frequently mentioned stations of every RER

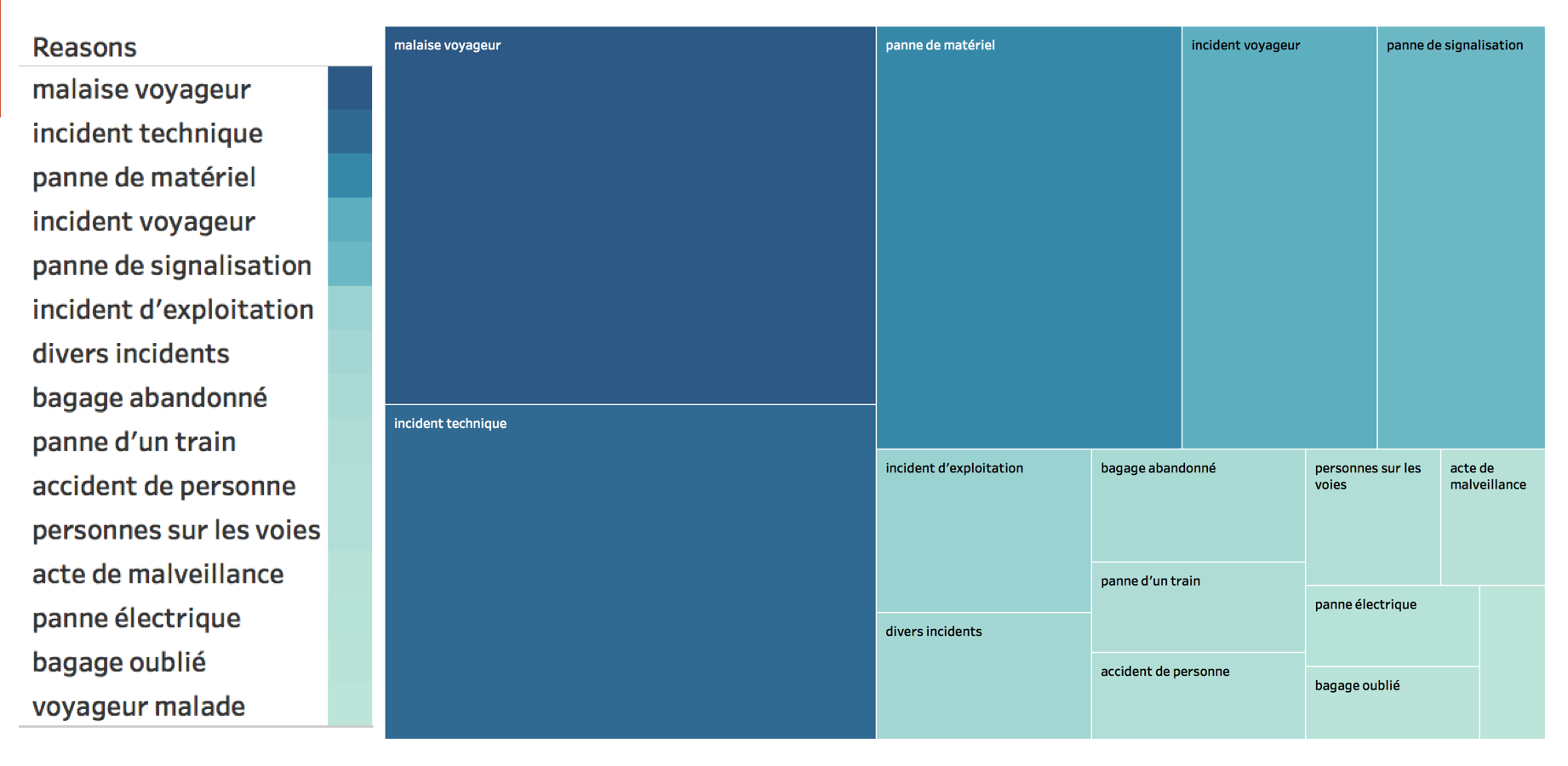
From figure 20. and figure 21. we found that breakdowns are more likely to occur in the terminals and the joint stations.



**Figure 22.** The 15 most frequently mentioned stations overall

The size of those lines in figure 22. represent the number of mentioned times. From the graph, it is easy to find that RER A is the worst line among all lines in metro and RER. Station Nation is the worst station that has the most breakdowns.

*4.3.3 What are those most frequent breakdown reasons?*



**Figure 23.** The 15 most frequently mentioned reasons of breakdowns

The most frequent breakdown reason is about passengers. The second most breakdown reason is technical problem. Reasons could be grouped into three categories: passenger problem, technical problem and suspected packages.

# 5 Limitations and Future Work

Due to lack of existing data set and the difficulty to build a specific data set by ourselves, we could just use the twitter account to represent the breakdown times. Although the analogy is indeed reasonable, it is hard to make specific analysis. For example, those twitters we collected would just tell there was a breakdown from station A to station B. As Fig 22. Indicated, we can only know that the traffic is interrupted between Cergy-Le-Haut and Maisons-Laffite, but cannot count all stations in between, (say, Cergy-Prefecture) in our model. Thus, our most frequent stations are biased to Origination station and Terminals.



**Figure 24.** The 15 most frequently mentioned reasons of breakdowns

There are many other factors could also cause or increase the probability of breakdown, such as the security situation in the region metro station situated (a region with poorer security situation would has more suspected packages, or passengers’ incidents would be more likely to occur), traffic volume, the operation years of metro systems, etc. Limited by the data source, we could not make predictions and analytics.

As our aim is to help improve RER and metro in Paris Region, we can share our model and analysis with SNCF and RATP agencies. Through a possible cooperation with them, we can get a more reliable dataset and include more relevant features.

# 6 Conclusion

Social media is a huge coal mine, from which we could extract many interesting and useful information. In our project, in the case that there is no available data set and previous work, we came up with a new way to show the breakdowns times of Paris metro/RER. 80% of our time and effort in this project was made to build and clean the data set.

Abandoned baggage is the primary reason for most delay/cancellation of all metro/RER lines. Besides, with a 80.5% accuracy of predicting the incident reason, our model can help RATP and SNCF prevent incidents and improve their service in Paris Region. However, due to the limited time and limited data source, our analysis can be enhanced by a possible cooperation with RATP and SNCF agencies.

# Appendix

Appendix 1:

**Table 2.** The 5 most frequently mentioned stations of every metro and RER

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Line | 1st frequent stations | 2nd frequent stations | 3th frequent stations | 4th frequent stations | 5th frequents stations |
| Line 1 | Château de Vincennes | Nation | Concorde | Gare de Lyon | George V |
| Line 2 | Nation | Port Dauphine | Anvers | Blanche | Blanche |
| Line 3 | Gallieni (Parc de Bagnolet) | Villiers | Péreire | République | Gambetta |
| Line 4 | Mairie de Montrouge | Porte de Clignancourt | Vavin | Gare du Nord | Gare de l'Est (Verdun) |
| Line 5 | Place d'Italie | Bastille | Gare du Nord | République | Oberkampf |
| Line 6 | Nation | La Motte-Picquet-Grenelle | Place d'Italie | Daumesnil (Félix Eboué) | Bercy |
| Line 7 | Pont Neuf | Maison Blanche | Place d'Italie | Château Landon | Stalingrad |
| Line 8 | Pointe du Lac | Blalard | Concorde | Maisons-Alfort-Les Juilliottes | Invalides |
| Line 9 | Nation | Oberkampf | Porte de Montreuil | Miromesnil | République |
| Line 10 | Duroc | La Motte-Picquet-Grenelle | Gare d'Austerlitz | Vaneau | Jussieu |
| Line 11 | Belleville | République | Châtelet | Mairie des Lilas | Hôtel de Ville |
| Line 12 | Front Populaire | Mairie d'Issy | Concorde | Porte de Versailles | Madeleine |
| Line 13 | La Fourche | Saint-Lazare | Invalides | Porte de Vanves | Varenne |
| Line 14 | Saint-Lazare | Olympiades | Madeleine | Gare de Lyon | Châtelet |
| RER A | Poissy | Nation | Auber | Vincennes | La Défense |
| RER B | Gare du Nord | Robinson | Palaiseau | Les Halles | Antony |
| RER C | Palaiseau | Invalides | Saint-Michel | Pereire | Villiers |
| RER D | Malesherbes | Gare de Lyon | Villiers | Chatelet | Gare du Nord |
| RER E | Villiers | Saint-Lazare | Val de Fontenay | Nation | Nationale |

1. https://opendata.paris.fr [↑](#footnote-ref-1)
2. https://www.ratp.fr/en/ratp-and-open-data [↑](#footnote-ref-2)
3. https://data.sncf.com [↑](#footnote-ref-3)
4. <https://www.wunderground.com/> [↑](#footnote-ref-4)
5. https://data.ratp.fr/explore/dataset/trafic-annuel-entrant-par-station-du-reseau-ferre-2016/ [↑](#footnote-ref-5)
6. https://ressources.data.sncf.com/explore/dataset/ [↑](#footnote-ref-6)