Anomalies in Paris from Dans-Ma Rue Dataset

Presented By:

Kristina Chaikina
Jawad Bin Anwar

Table of contents:

- Short description of the dataset
- Motivation for this project
- Hypothesis & Research Questions
- What we have done so far
- Visualizations
- Our future plans

Short description of the Dataset

- Dans Ma Rue is an app that people use to report an anomaly
- We got the dataset from Kaggle
- It mostly includes categorical data
- The dataset has 17 features, most of them spatial data
- It also has temporal features

Motivation for this project

- This dataset is barely explored
- We were motivated to explore a dataset that allows us to conduct spatial analysis as well as statistical analysis
- It is impactful because if we are able to answer the research questions correctly, we may be able to take preventative measures to curb anomalies in Paris
- The dataset is not very complicated, so for a beginner project it seemed suitable
- It can be easily paired with other spatial datasets

Research Questions

After looking at the data, we came up with questions:

1. Is there a particular time window during the year where most of the anomalies are reported?
2. Which areas of Paris are most affected by anomalies/ vandalisms? What type are they?
3. Can we make a predictive algorithm which may answer which areas might be more susceptible to a particular anomaly?
4. Do tourists and social housing influence the accurance of anomalies? How?

Data Extraction

- The data were sorted year-wise
- We collected the data from 2012 to 2022 (Excluded 2023 since the year has not ended yet)
- The data were in the form if CSV files
- Some of the CSV files had different form of encoding
- Apart from that, the data were very solid and consistent. We easily identified the temporal and spatial data
- Only problem was the dataset did not include any geo-shape

Data Cleaning

- From the 17 features, we kept only 11 features of interest
- The feature name and ordering were different across the CSV files, we fixed that
- We also converted the data type for some of the features
- One of the dataset had a missing feature but it was not vital to our project
- Finally we made sure that the merged dataset was consistent in regards of data-type and feature names

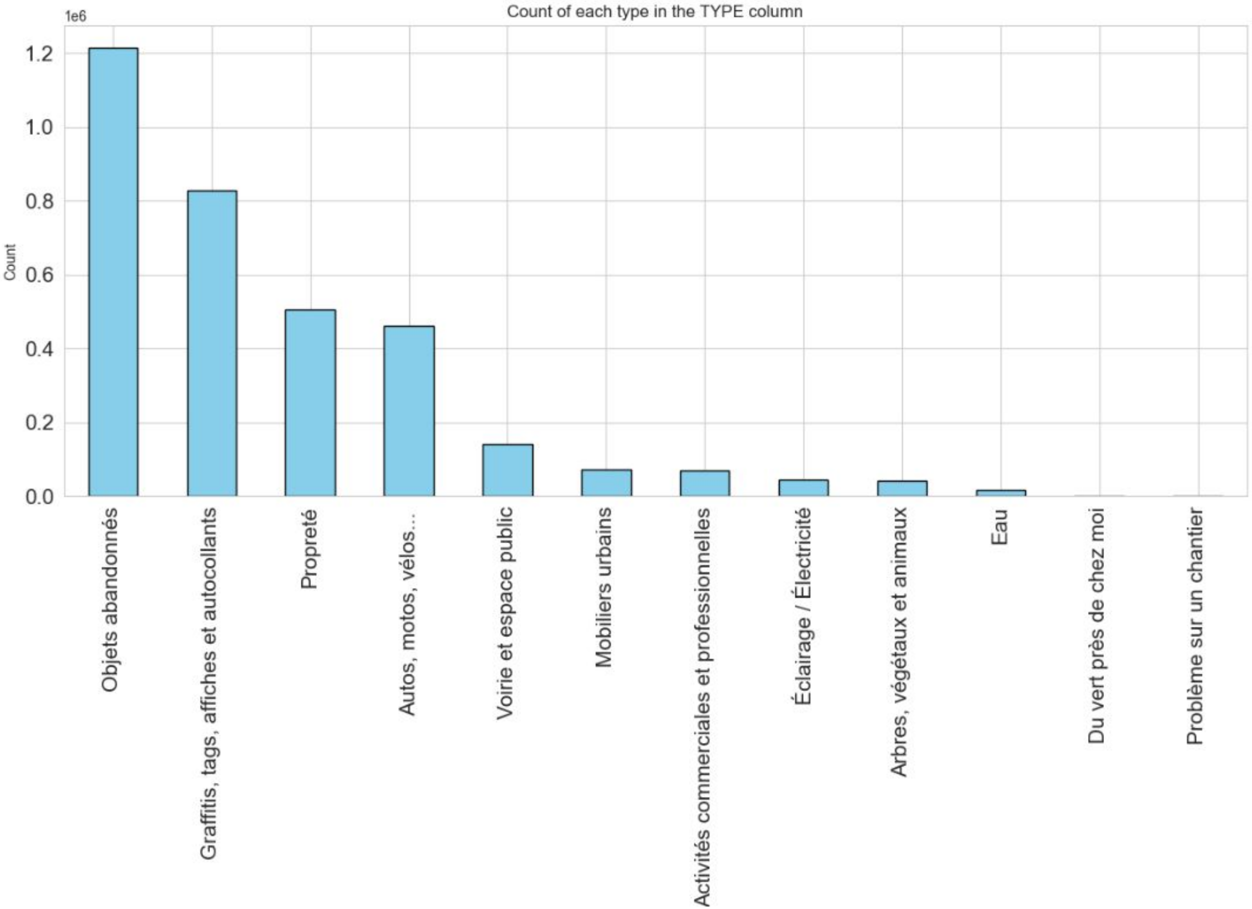
Data Visualization

FIRST AND FOREMOST THING TO DO

LOOKING AT THE DATA!

12 Types of anomalies were present.

Among them, the first 4 contributed to more than 80% of the total anomalies



Cont.

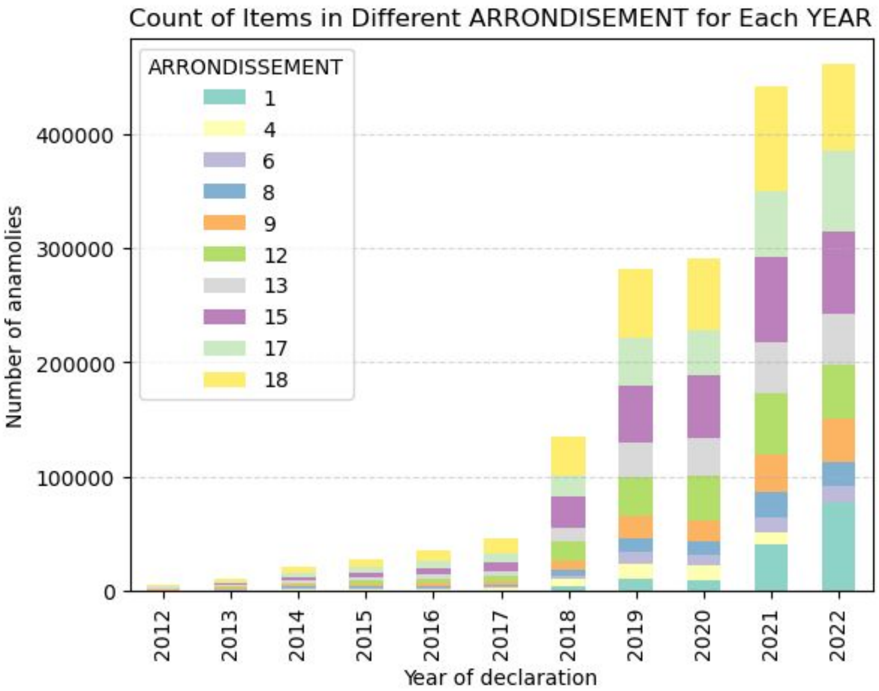
Interesting Fact:

- We thought the number of reported anomalies would be less in 2020 compared to the previous years because of the pandemic and lockdown
- On the contrary, it was not the case. There were more reports in 2020 and the subsequent years despite lockdown

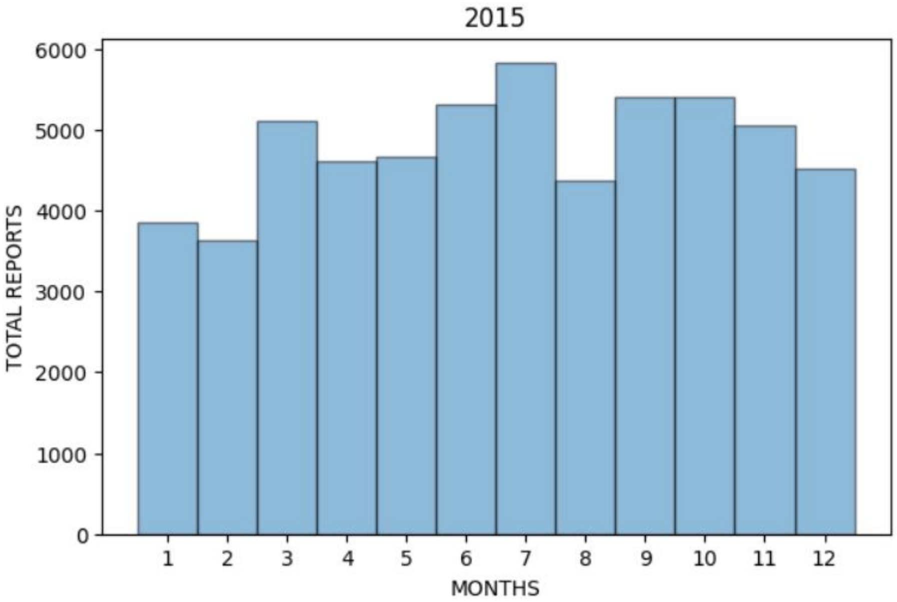
Cont.

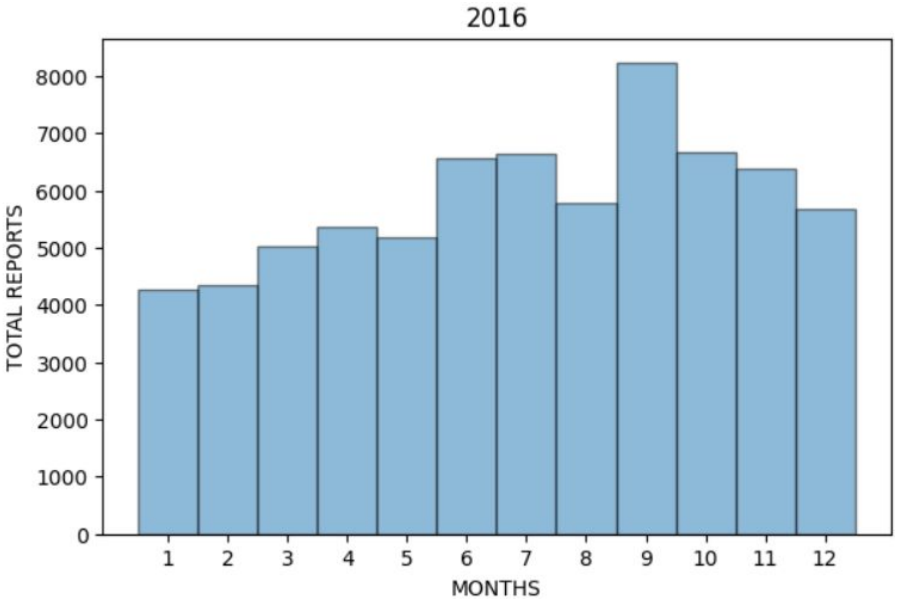
Each year, the overall number of reports increased

The number of reports increases exponentially, and 2020 does not fit the pattern



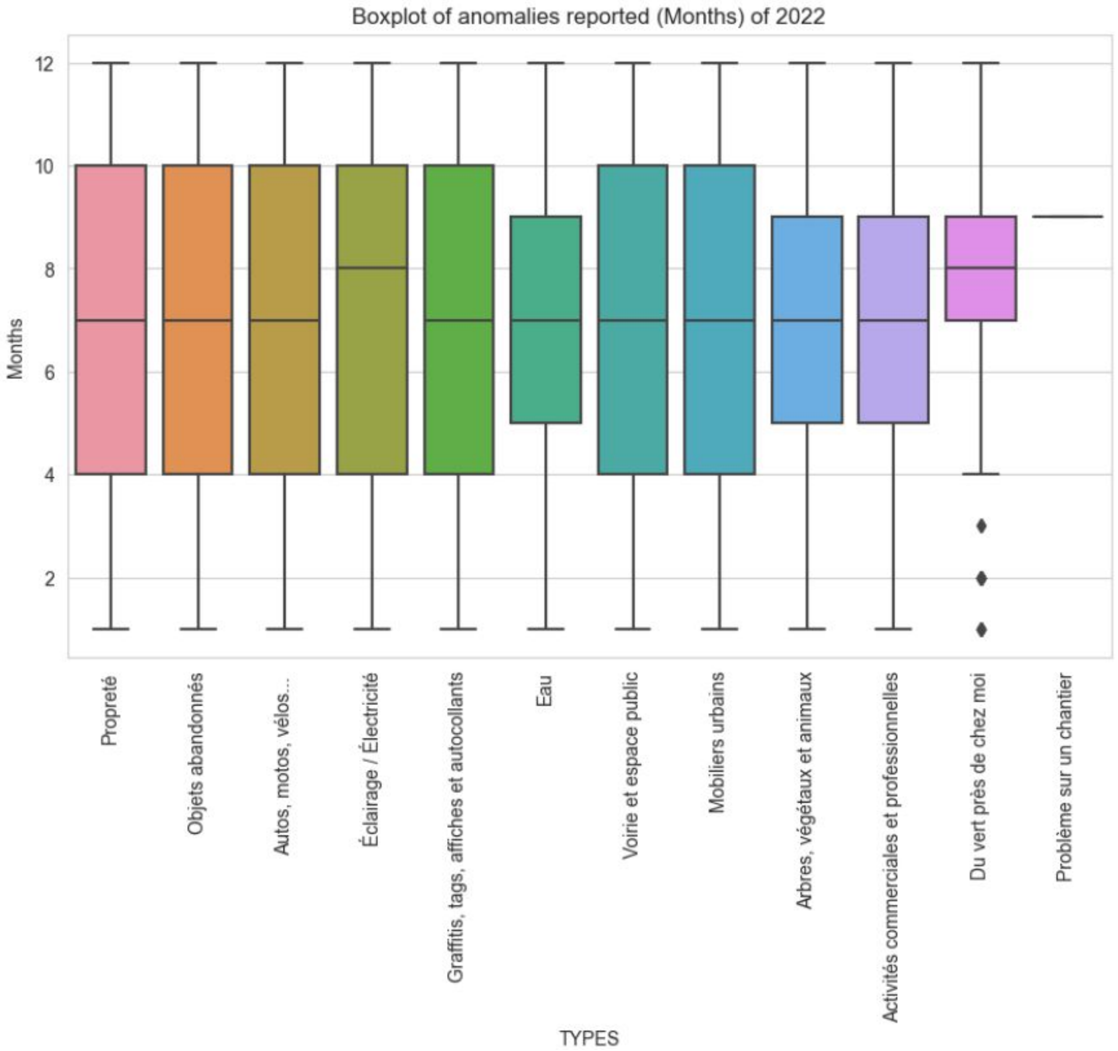
Cont.


2015
TOTAL REPORTS
0
1000
2000
3000
4000
5000
6000
1
2
3
4
5
6
7
8
9
10
11
12
MONTHS


2016
TOTAL REPORTS
0
1000
2000
3000
4000
5000
6000
7000
8000
1
2
3
4
5
6
7
8
9
10
11
12
MONTHS

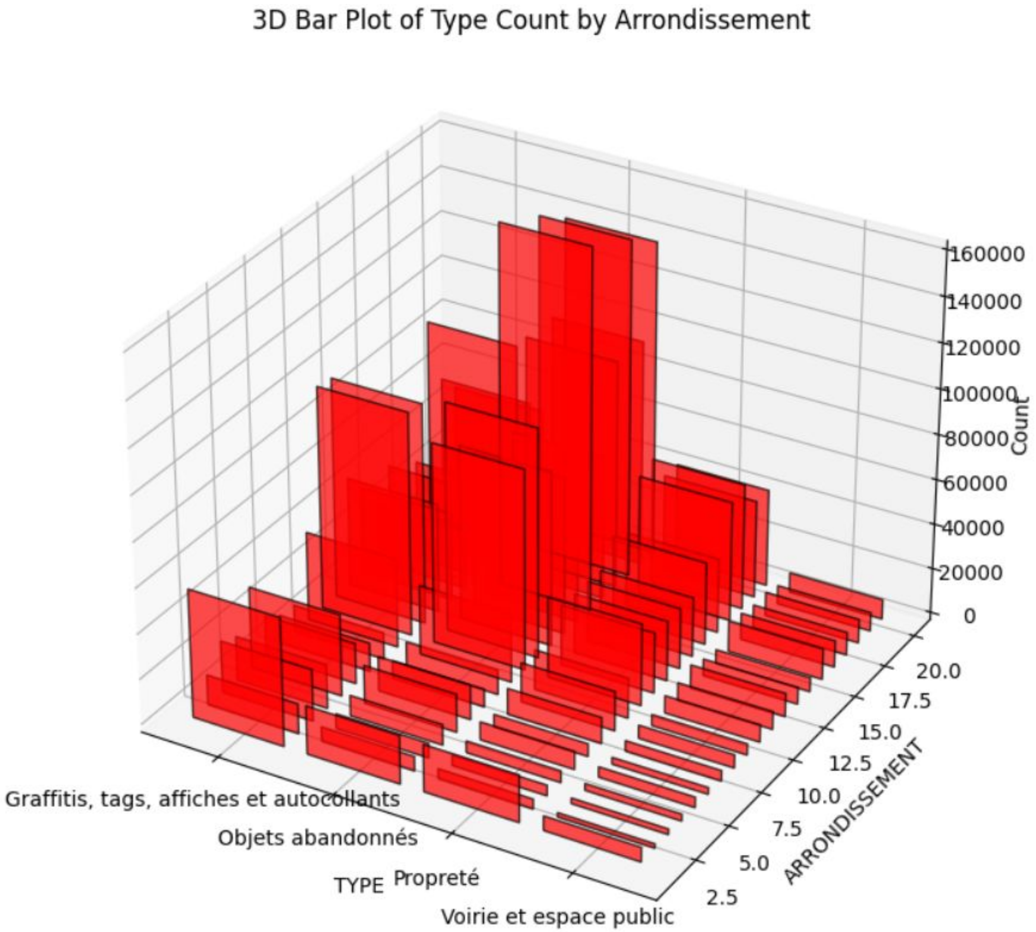
Cont.

Most of the reports are made during summer, which is also the peak tourist season in Paris



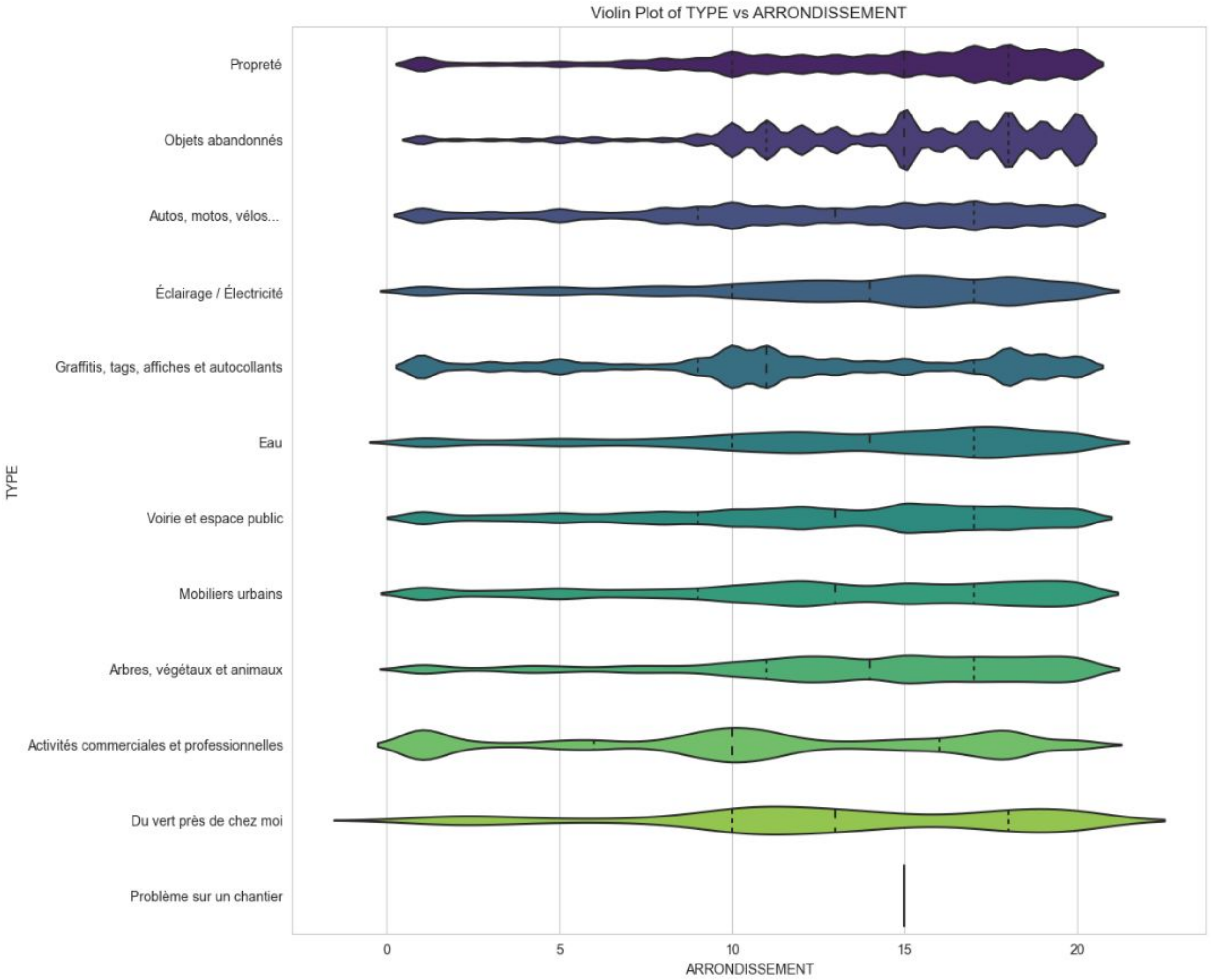
Cont.

A bad visualization



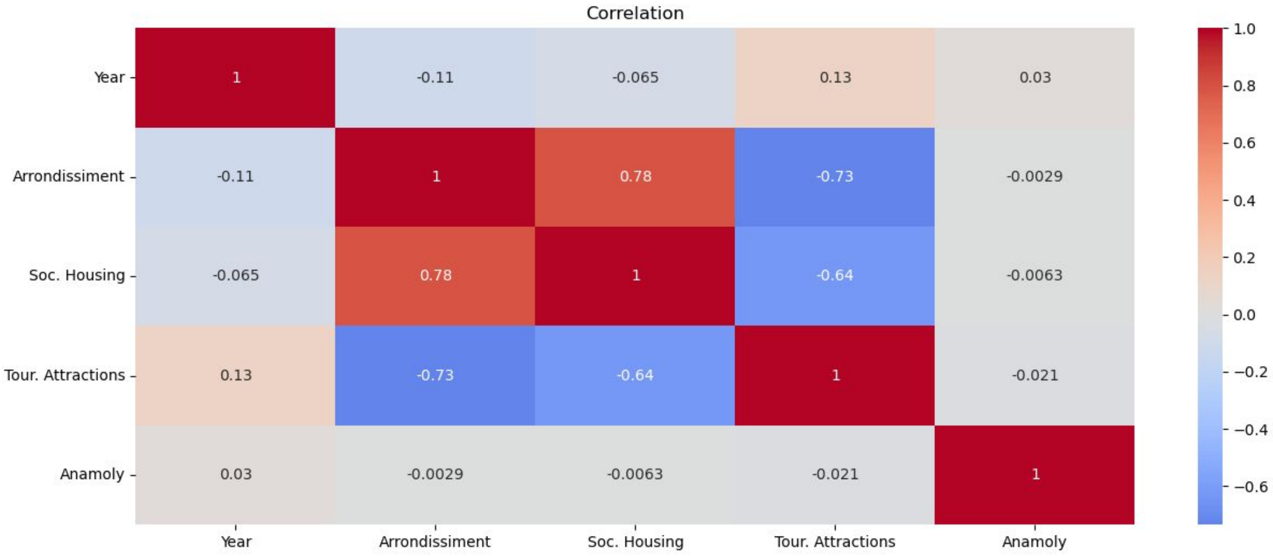
Cont.

Visualizing which type of anomalies are more prevalent in which areas of Paris



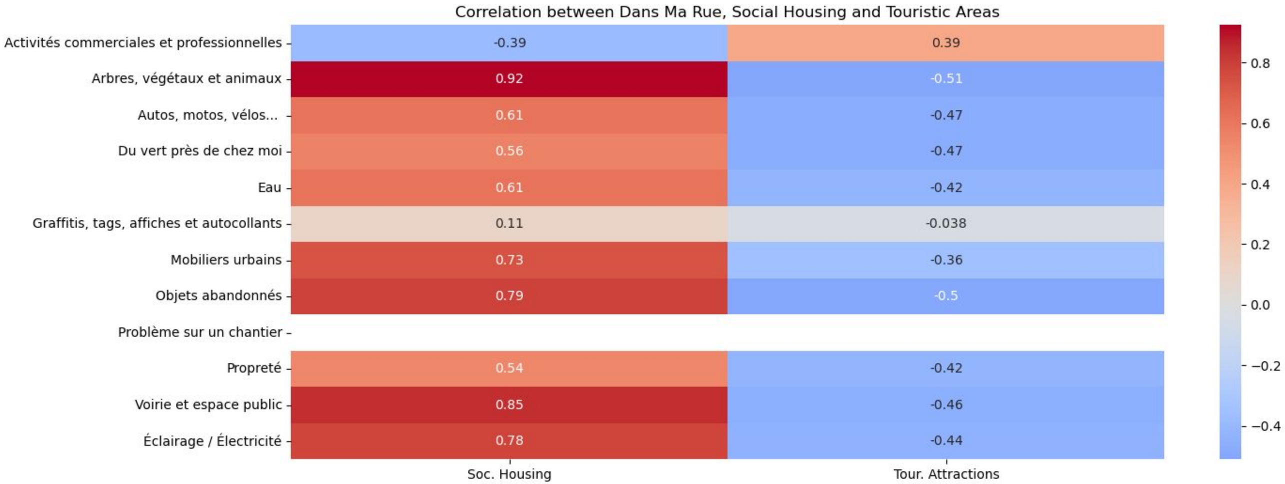
Cont.

Overall, it is hard to see the correlation



Cont.

If seperate the anomalies by type, we can see the influence of touristic attractions and Social housing on the anomalies



Future works

- Utilizing a tourist dataset and social housing dataset with current dataset
- Statistical analysis
- Making a predictive model

THE END