Analysis of SeeClickFix Data Using Tableau

Submitted by – Tazein Fatma

Introduction about SeeClickFix Data:

The data is about 311 issues reported in 4 different cities. People report about issues they care or want city to fix/improve. They can report an issue via their phones, mobile site, new app widget etc. The issues can be described in detail, in a summary, and with a tag type. Longitude and Latitude gives the city location on the map. Each issue gets votes, comments and views and time the issue is created is noted in the data. City or government can act on the most important or urgent issues accordingly.

These are the 10 features/ variables in the data-set and one ID field:
Latitude, Longitude, Summary, Description, Num_votes,
Num_comments, num_views, source, created_time , tag type.

If we look at the **tag_type** details we see lots of issues ranging from trash, graffiti, traffic, potholes, signs and so on. There are 42 total type of tag_type reported. It would be interesting to see what are the top issues that concern the people of a city.

Checking NA values:

There are 169714 out of 223129 "**NA**" entries in the **tag_type**. NAs are almost 76% of total tag_type. If we look at the descriptions and summary for these entries there is detailed description of issues, but it seems that people did not know which tag type to put their issue under. Or certain issues are one-off due to their nature and do not fall in a specific tag_type so the field is left as NA. NAs should be broken down and distributed into their relevant tag types or if needed new tag_type should be created so that all the issues reported as NA are correctly represented in the data.

Used filter to select only one city:

To select one city, I dragged longitude and latitude to the column and rows. I changed them to dimension. This gives us map of US with four cities being shown on the map. To choose only one city, I brought the longitude dimension to the filter area and then edited the filter and in the range placed longitude value of the city I wanted and a closer number to that in the upper range. Specifically, to select San Francisco area, I chose its longitude -122.34 in the lower range and -105 in the upper range to give me an area around the city on the map.

Visualization that is sorted by the count of the number of tags.

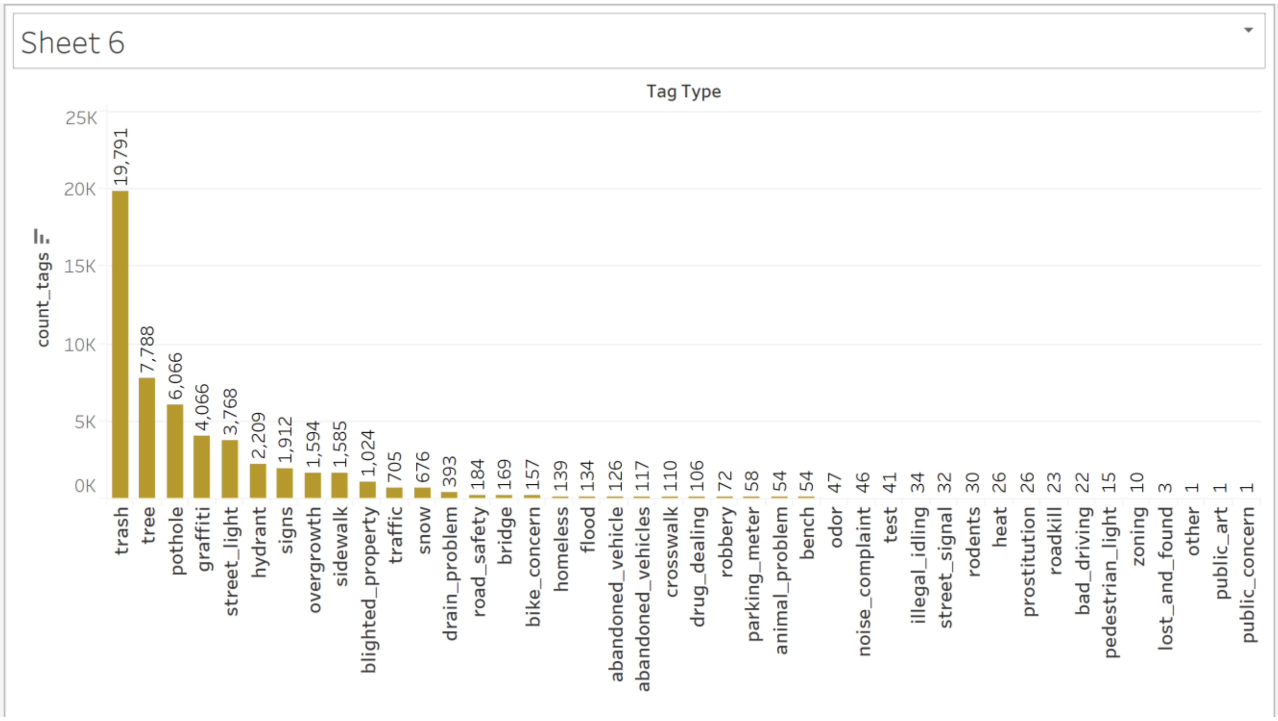
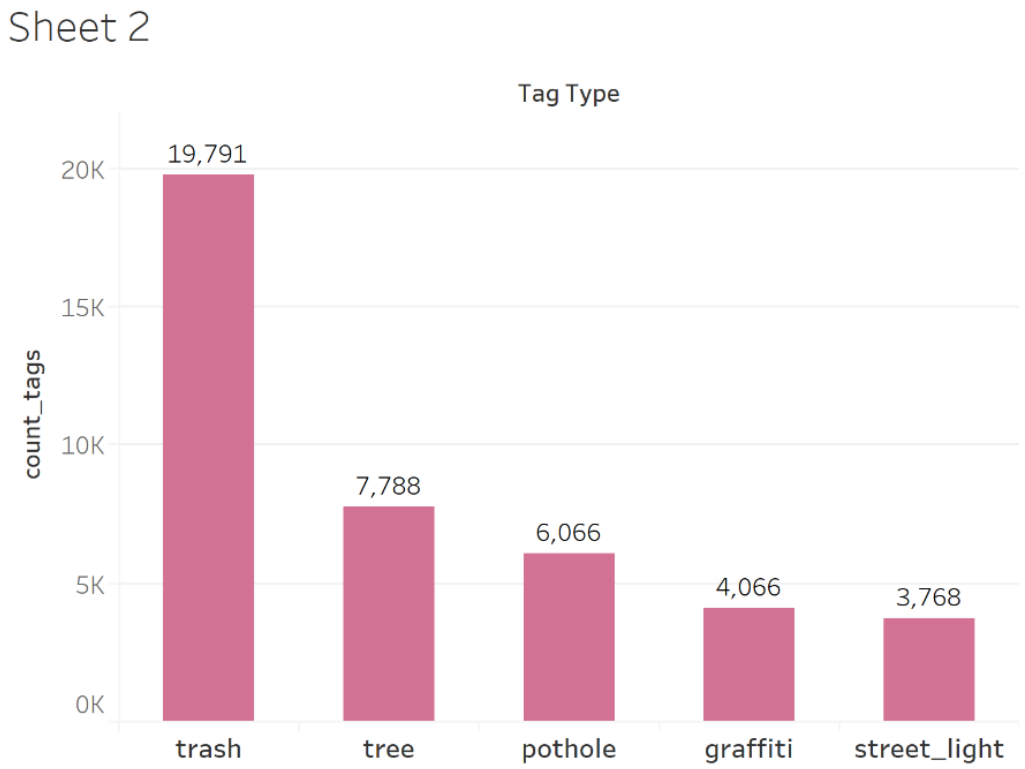
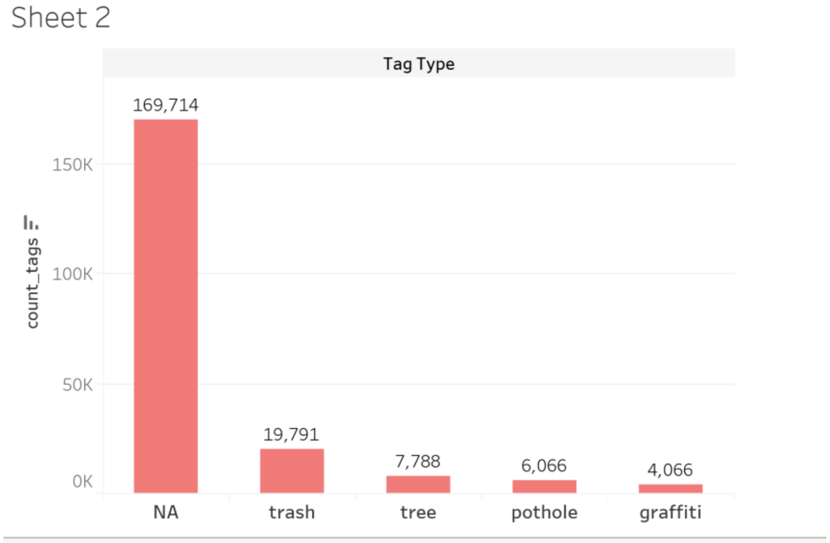


Chart showing the 5 most frequent tag types.

This chart shows that “trash” is the biggest concern for the people of this city. “Tree” problems and “Potholes” are number 2 & 3 followed by “graffiti” and “street_light”. fFor the sake of this question and looking only at the data that is classified correctly, I ignored NA values for this chart.

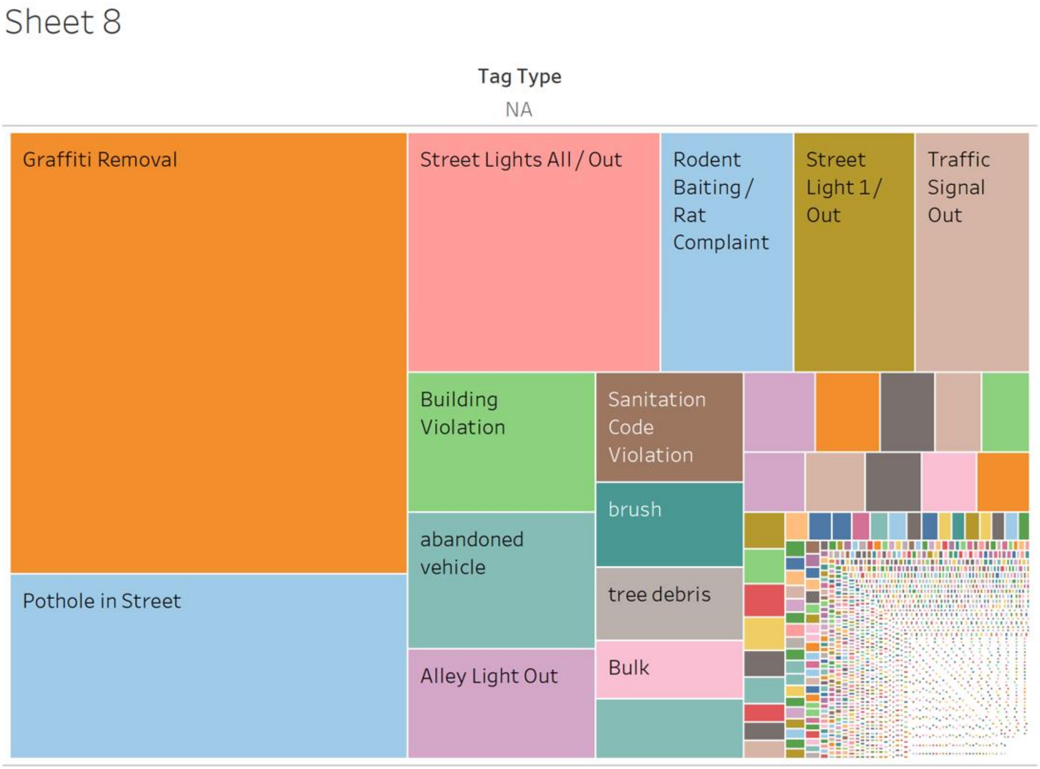


Considering NAs - If we take NA tag_type into account then the top 5 would show up as below chart. Here we see NA tag_type account for highest count followed by trash, tree, pothole & graffiti.



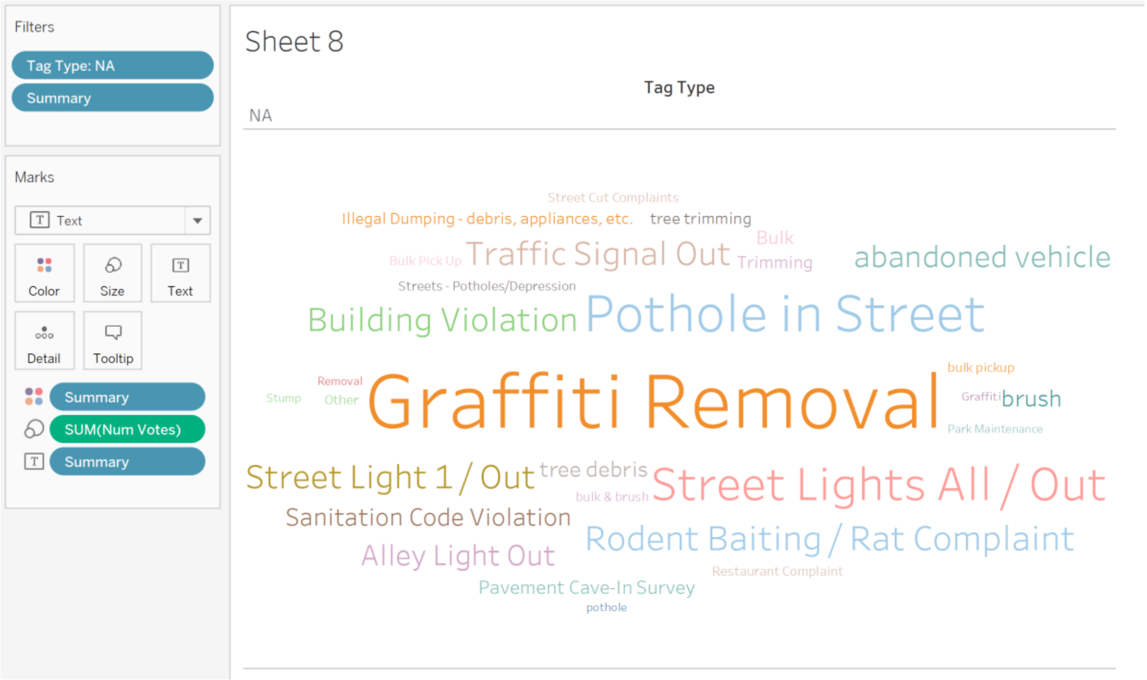
What is hiding in NA?

1. As I mentioned above there are almost 76% of **tag_types** that are entered as NA values in the data set. I wanted to see which issues are the most important within the set that is reported as NA. I tried to get an idea of which words come up the most in the "summary" field of the tag_types which have NA values in them. I created below visualization for NA tag_type by "summary", to try to get a word cloud and see which words show up most within the NA category. I came up with below visualizations, a Tree map first.



I converted the above Tree map to a word cloud to give an idea of which words are used more within the summary of the issues which are reported as NA tag_types. The list of words is way too long to generate a proper word cloud in tableau, so I filtered top 50 words. We can see that within the NA category, Graffiti Removal is the biggest concern followed by Pothole, street lights, building violation etc. In this set some of the issues like Rodents, abandoned vehicle also features as bigger issues which we do not see in the data if we ignore the NA.

Ideally these NA tag_types should be converted to some meaningful tags to get true picture of top issues in a city.



Filters
- Tag Type: NA
- Summary

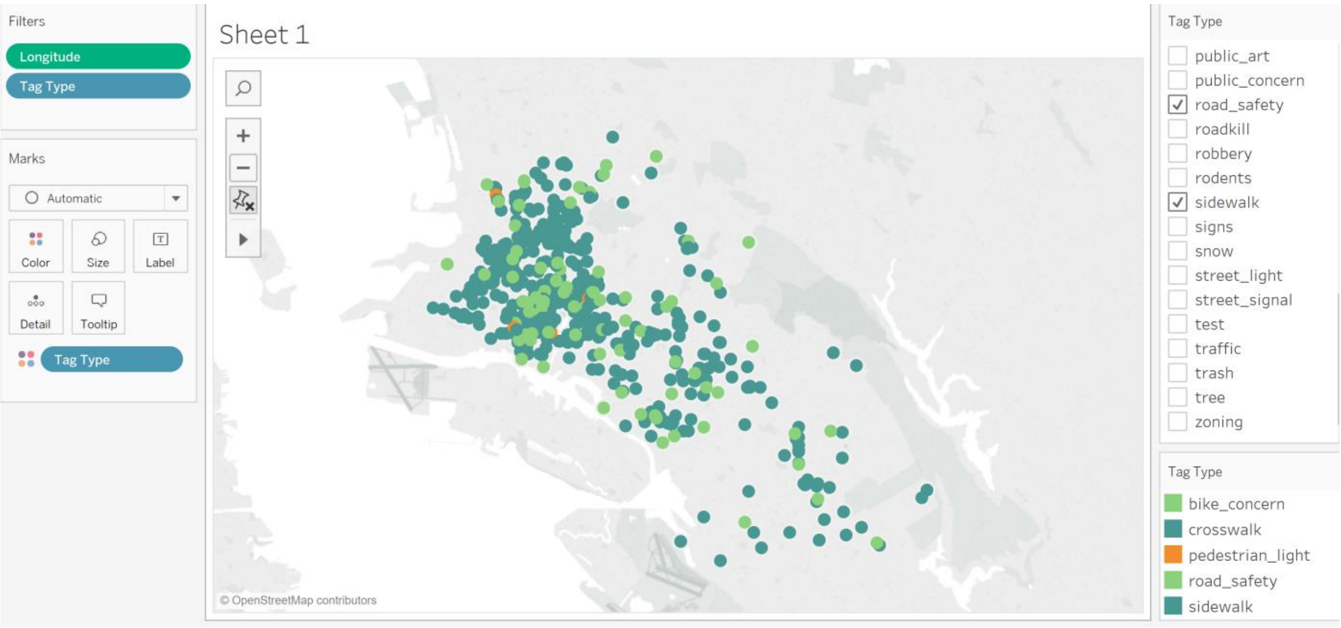
Marks
- Text
- Color: Summary
- Size: SUM(Num Votes)
- Text: Summary

Sheet 8

Tag Type

NA

2. In San Francisco main city area people tend to walk and use public transport more and use less cars, so I wanted to see what are the issues for pedestrians and bikers. Below visualization shows that lot of Sidewalk issues, crosswalk issues, pedestrian lights, bike concern are rampant within the main city area. City could do a lot to increase road safety for pedestrians and bikers.



Filters
- Longitude
- Tag Type

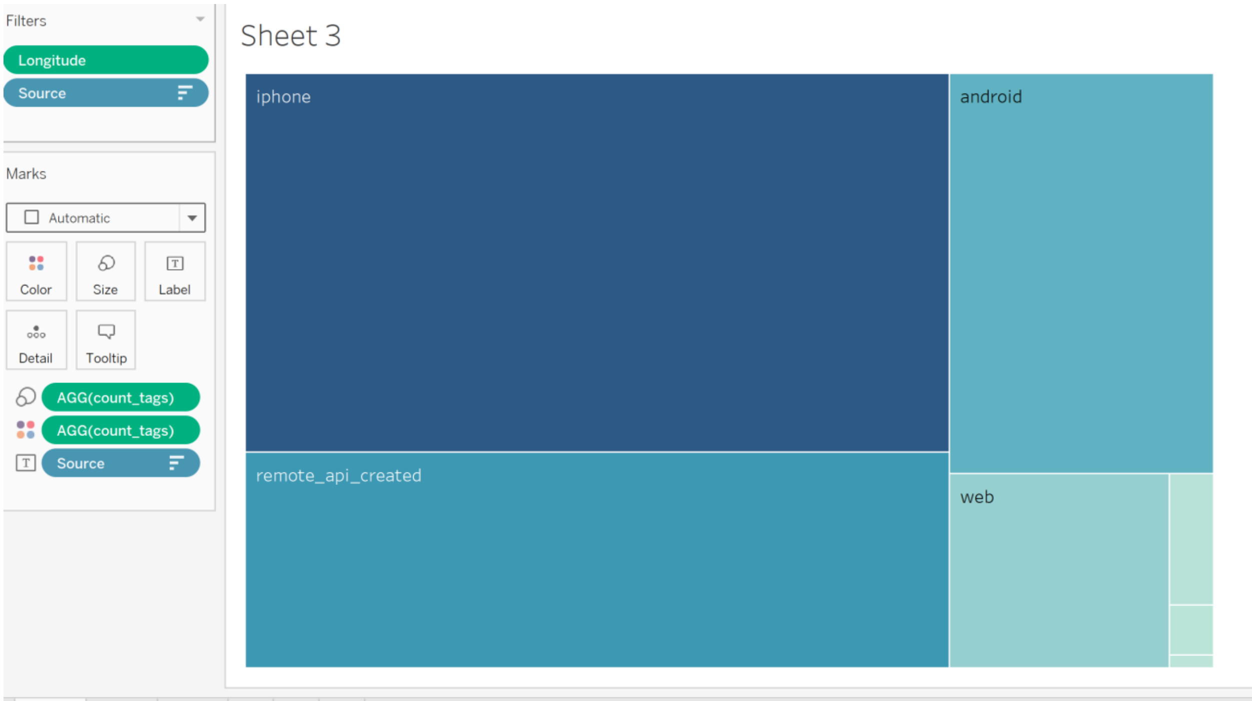
Marks
- Automatic
- Color: Tag Type

Sheet 1

Tag Type
- ☐ public_art
- ☐ public_concern
- ☑ road_safety
- ☐ roadkill
- ☐ robbery
- ☐ rodents
- ☑ sidewalk
- ☐ signs
- ☐ snow
- ☐ street_light
- ☐ street_signal
- ☐ test
- ☐ traffic
- ☐ trash
- ☐ tree
- ☐ zoning

Tag Type
- 🟩 bike_concern
- 🟩 crosswalk
- 🟧 pedestrian_light
- 🟩 road_safety
- 🟩 sidewalk

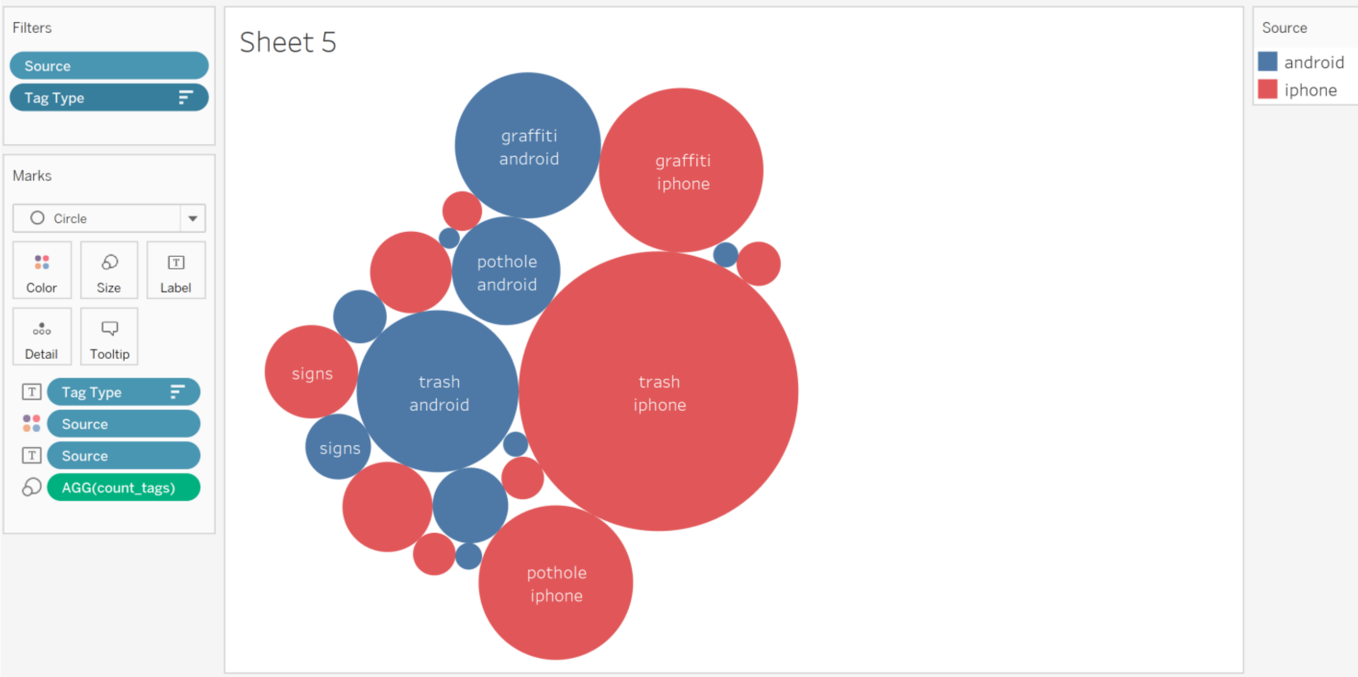
3. This tree-map shows that “iPhone” was used by most to report an issue followed by “remote_api_created” and then “Android” phones, web, mobile site, map widget and “new map widget”. This visualization also ignores “NA” values in the source.



Filters
Longitude
Source

Marks
Automatic
AGG(count_tags)
AGG(count_tags)
Source

4. This packed bubble shows that iphone users are most bothered by trash compared to android users. Graffiti issues are reported almost in similar numbers by iphone and android users. But we can see red bubbles representing iphone users are bigger compared to the blue bubbles for android users.



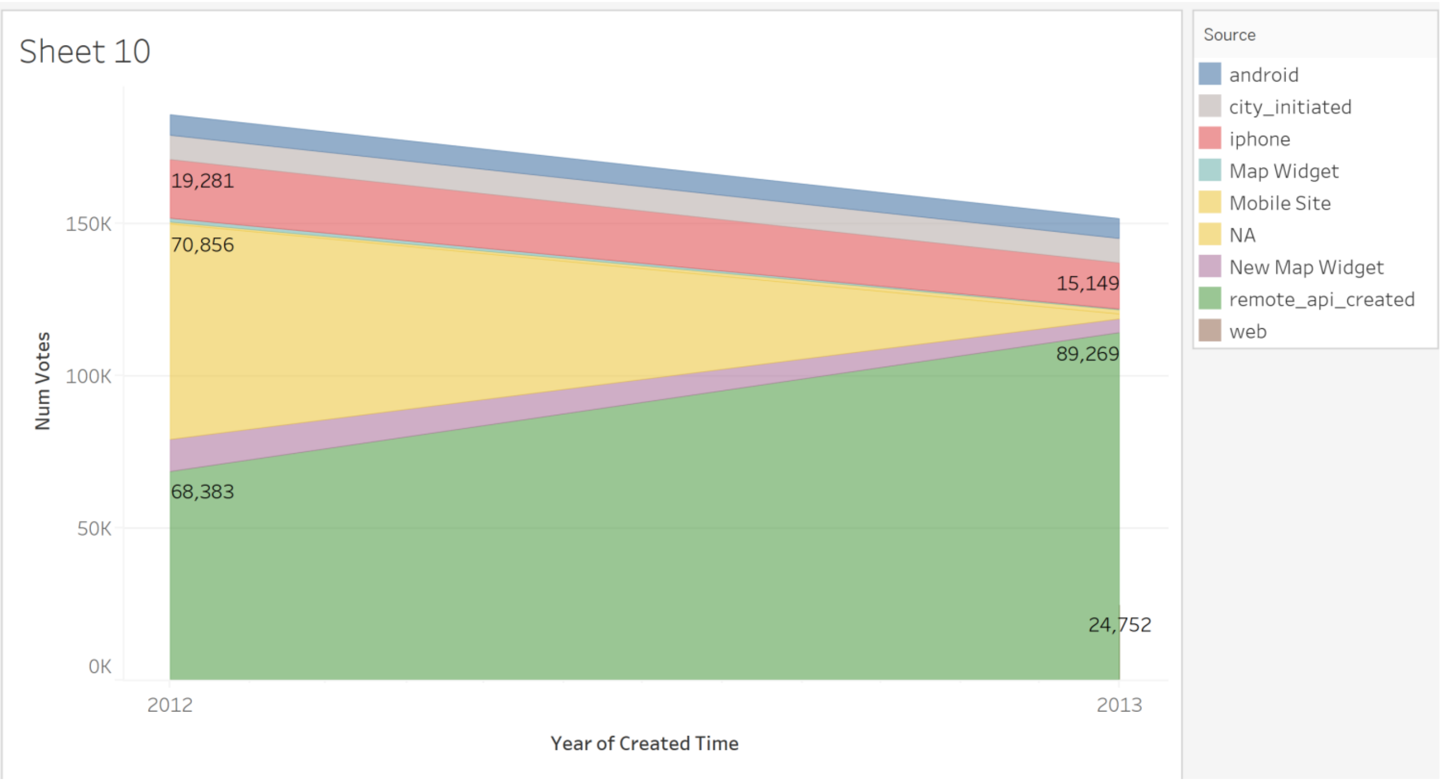
Filters
Source
Tag Type

Marks
○ Circle
Tag Type
Source
Source
AGG(count_tags)

Source
🟦 android
🟥 iphone

5. I wanted to see how different "sources" used this service over a period with respect to num_votes. So, we have data from Jan 1, 2012 to Apr 1, 2013. It is interesting to note that overall num_votes have decreased over time. May be the initial enthusiasm of the service has faded over the year.

-"Remote_api_created" is being used more to vote.
-"NA"sources have reduced.
-votes coming from "Iphone", "city_initiated" & "Andriod" sources have seen a decline.



---End---