**W200 Project 2 Final Report**

**Graffiti Analysis in Saint Paul, Minnesota**

Team Member: Mia Yin, Jude Wentian Zhu

Github repository: https://github.com/UC-Berkeley-I-School/Project2_Yin_Zhu.git

## Overview:

Graffiti is often seen as a sign of disorder in communities, it can open the door to individuals breaking other social norms and are associated with increased crime rates in many cities across the US. In addition to being a sign of destruction, graffiti vandalism in national parks destroys the natural beauty. Graffiti cleanup work involves both labor and material cost, and tends to be recurring. Every year, municipal offices spend a significant amount of funds dedicated to graffiti removal and vandalism prevention.

In this analysis, we will use published data from Saut Paul, MN's Open Data Portal to analyze the city's graffiti removal efforts from 2013 to 2015. We investigate the graffiti cleanup work from four perspectives:

1. **Location 2. Surface 3. Time Series 4. Cost Analysis.**

Below are the main questions we plan to address:

1. How is graffiti distributed across St. Paul?
2. Is graffiti clustering associated with crime rate in St. Paul as suggested in many research articles?
3. How likely does graffiti tend to recur at the same location?
4. Is there any visible trend of what surface do the vandals prefer?
5. What are the general trends of graffiti cases from 2013 to 2015, any seasonality?
6. What insight can we draw from the cost analysis for graffiti cleanup work?

## Dataset:

Parks Graffiti Report - Dataset

This dataset contains graffiti related issues responded to by the Parks and Recreation Department from 2013 to 2015. It includes work done in the City's parks and in non-park locations. It also includes cleanup costs for most incidents. Source:https://information.stpaul.gov/City-Infrastructure/Parks-Graffiti-Report-Dataset/gcu2-spkd

Data Structure

- The dataset has 817 rows and 11 columns. 817 rows represent 817 entries during the years from 2013 to 2015.
- Columns contains information such as date, location name, location latitude/longitude, cost, descriptions and etc.

Crime Incident Report - Dataset

The report contains incidents from Aug 14 2014 through the most recent available, as released by the Saint Paul Police Department. Source: https://information.stpaul.gov/Public-Safety/Crime-Incident-Report-Dataset/gppb-g9cg

Data Structure:

- The dataset has 323173 rows and 14 columns, containing crime reports up till 2021. We don't need all these data so will only use the subset of reports in 2015.
- The columns include time of report and district number. Which will be used to analyze crime rate by district.

District Council - Shapefile - Map

Shapefile for St Paul's 17 District Councils. Source:https://information.stpaul.gov/City-Administration/District-Council-Shapefile-Map/dq4n-yj8b

The shapefile of district council boundaries will be used to plot the St. Paul district map.

**Data Sanity Check and Initial Data Cleaning:**

Parks Graffiti Report - Dataset

| | Location Name | Work Performed Date | Location Address | Issue | Description | Labor Cost | Material Cost | Total Cost | Location |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Kelly's Landing | 05/07/2015 | NaN | Graffiti | Graffiti | 54.20 | 34.75 | 88.95 | (44.9360847, -93.0993576) |

Some issues found through the initial data screening include:

- Column name style
- Missing values
- Location data is stored as string type. Needs to be manipulated to be useful
- Description needs to be manipulated to be useful
- Need to add year and month of the date to analyze trend and seasonality.
- There is no district number in the Graffiti dataset. Need to append district number/location to make a meaningful map plot.

Initial cleanup:

- Remove the entire entry if there's a NaN (34 entries with missing values are removed)
- Rename columns
- Add year and month for additional analysis (work_year, work_month)
- Extract latitude and longitude for plotting geodata
- Identify park vs. non-park

## Analysis:

1. **Graffiti Distribution by Location**

**Q1:** How is graffiti distributed across St. Paul?

**Data wrangling:**

1. Using the location(latitude, longitude) information, append the district number found in the shapefile to each entry in the Graffiti dataset. This step involves mapping the location latitude and longitude to the district polygon, finding the center point of the polygon and converting the centerpoint to a string so it's hashable, and merge the graffiti dataset and the shapefile on the centerpoint to add district number. St. Paul has 17 districts.
2. Group total counts of graffiti cases by district for graffiti cases heatmap generation.
3. Group total counts of crime report by district for crime reports heatmap generation.

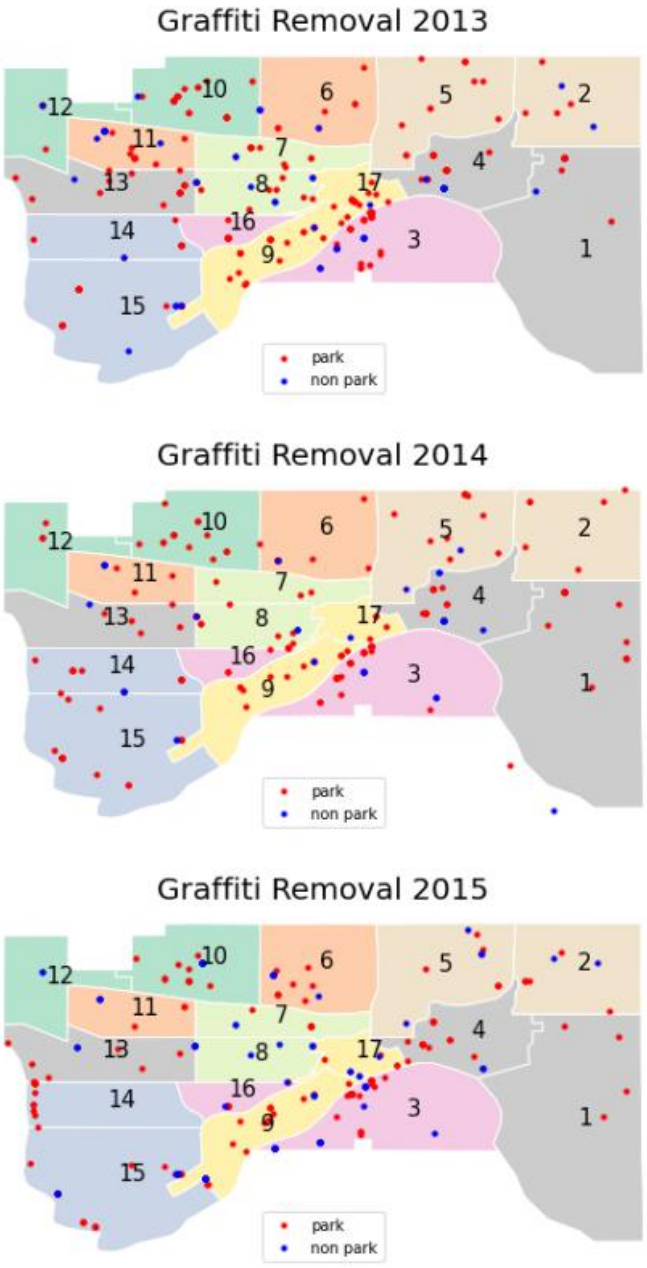


Figure 1: Graffiti Location by District

**Graph Insights(fig1):**

1. Each year, graffiti cases spread across all districts in St. Paul, with some clustering around district 3 and district 17. District 1 has the lowest graffiti density. There is no obvious signs of migration over the years.
2. Two locations in 2014 actually do not belong to St.Pauls' district. These will be left out from the background district map but kept in the Graffiti dataset for further analysis until we have further visibility into the data inconsistency.
3. Graffiti occurs more frequently in parks (red dots) vs. non-parks (blue dots)
4. If multiple graffiti removal occurs at the same location, the dots overlap each other. This scatter plot does not provide insight about repeated graffiti at the same location.

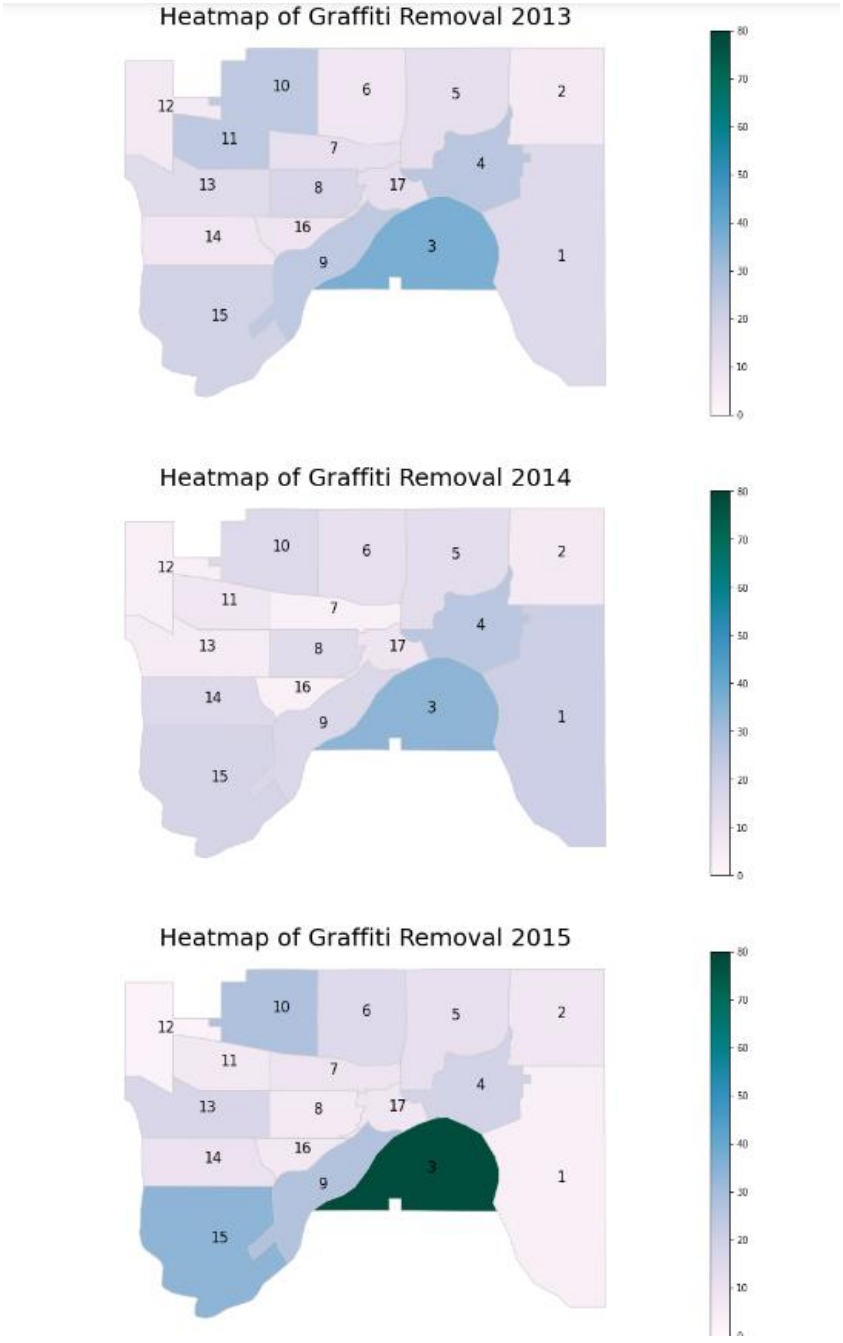


Figure 2: Graffiti Heatmap by District

**Graph Insights (fig2):**

1. The distribution of graffiti in 2013 and 2014 are more or less consistent, where district 3 has the most cases.
2. There is a surge of graffiti cases in 2015 in district 3, where the total number of cases rise to 80.
3. If we check the total number of graffiti removals year over year, 2015 only has 19 more total cases than 2013. We believe the majority of the surge in district 3 is offset by a decrease of cases in other districts, such as decrease in district1 and 8 as shown below:

| district | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | grand_total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| work_year | | | | | | | | | | | | | | | | | | |
| 2013 | 15 | 6 | 37 | 25 | 12 | 8 | 11 | 17 | 24 | 24 | 24 | 6 | 14 | 8 | 19 | 9 | 12 | 271 |
| 2014 | 21 | 6 | 34 | 25 | 13 | 11 | 3 | 14 | 16 | 15 | 8 | 4 | 5 | 15 | 18 | 3 | 9 | 220 |
| 2015 | 4 | 8 | 78 | 19 | 11 | 15 | 9 | 6 | 27 | 28 | 7 | 2 | 17 | 10 | 34 | 7 | 8 | 290 |

**Q2:** Is graffiti clustering associated with crime rate in St. Paul?

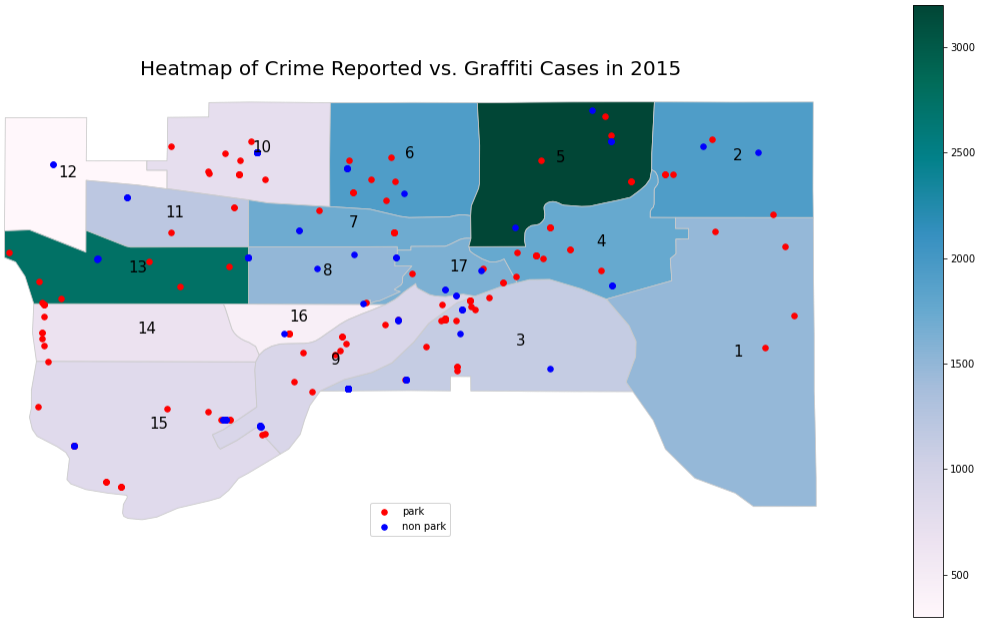


Figure 3: Crime Report Heatmap by District vs. Graffiti Clustering

**Graph Insights(fig3):**

1. The assumption is that higher graffiti density leads to increased crime rates in one area, as signs of disorders in communities are often associated with increased crime rates.
2. The heatmap background represents the density of crime reports (number of crimes reported in the district), and the scatters represent locations of park (or non park) graffiti removal work.
3. Unfortunately, the graph does not support my assumption. District 5 has the highest crime reports but does not show graffiti clustering. This might due to many reasons such as type of

crime reports, the possibility of one or two factors (non-graffiti) contributing to a surge in crime rates, the limited data points (only 2015), or simply a weak association between graffiti and crime reports in St. Paul.

4. It is still helpful to visualize the relationship even though it contradicts our initial assumption.

**Q3**: How likely do graffiti tend to recur at the same location?

**Data wrangling:**

1. Group the Graffiti dataset by location (latitude, longitude) which is the most granular level of location data.
2. Select locations with repeated cleanups by filtering total count of graffiti cases by >1

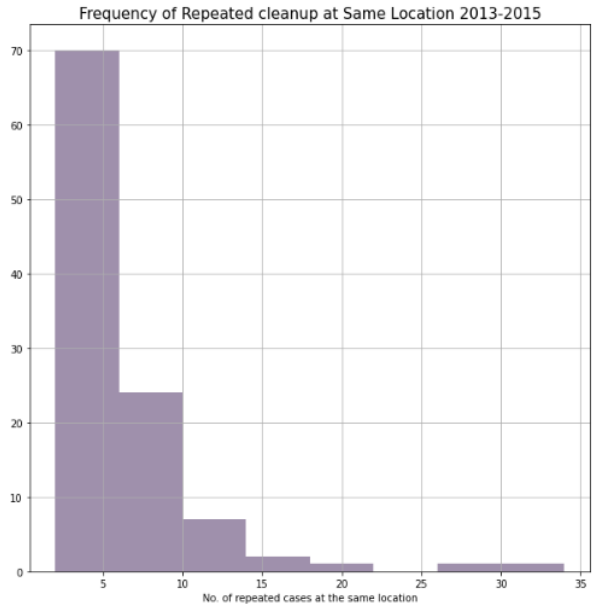


Figure 4: Crime Report Heatmap by District vs. Graffiti Clustering

**Graph Insights(fig4):**

1. Many locations incur repeated graffiti cases. Most locations has around 5 repeated graffiti cleanups, some locations undergo 30+ cleanups but such case is rare.

## 2. <u>Graffiti Distribution by Surface</u>

**Q4:** Is there any visible trend of what surface do the vandals prefer?

**Data wrangling:** The 'Description' column in the graffiti dataset tells what surface each graffiti was on.


Figure 5: Graffiti surface word frequency

**Graph Insight(fig5)**

1. A WordCloud was developed to show the frequency of graffiti surface, as identified by the case description. Size of the word represents the proportionate frequency of keyword appearance.
2. It is interesting to notice that building, wall, bridge, park, sidewalk, tot lot, parking lot, picnic table, electrical box, etc. are among the most frequency surface the graffities may appear on. Among these, building appears to be vandals' favorite spots.
3. A more detailed comparison is shown below. Figure 5 ranks the places/words in graffiti surface information in detail. 'Building' tops the list and have far more counts than the second graffiti surface 'Park', implying that building is the best drawing board for graffiti creators.

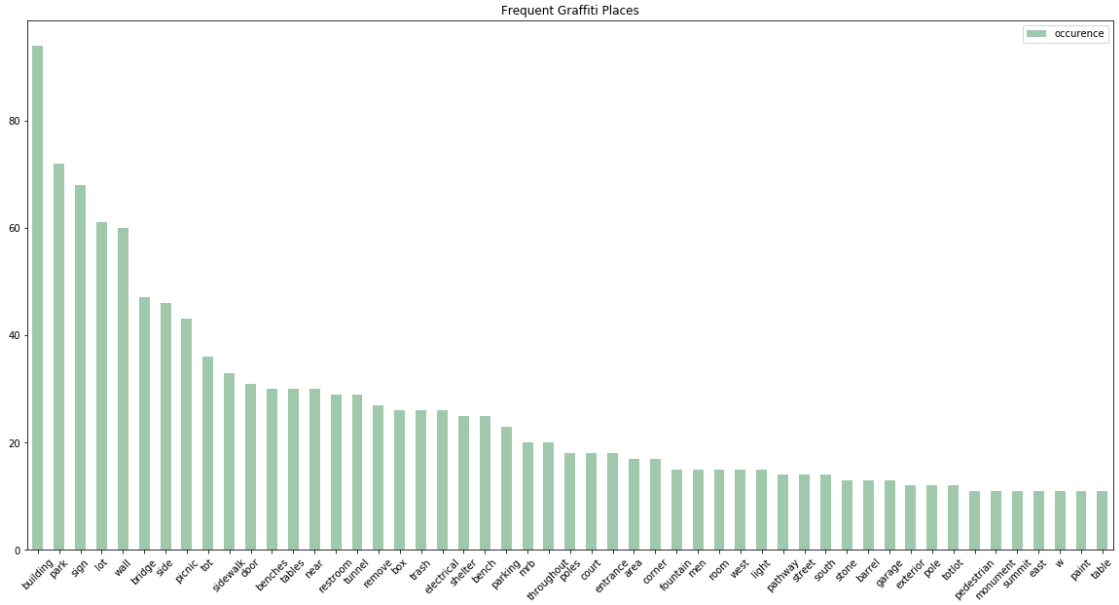

Figure 6: Most Frequent Graffiti Places

## 3. Time Series

**Q5:** What are the general trends of graffiti cases from 2013 to 2015, any seasonality?

**Data Wrangling:**

1. Group graffiti case count by year-month combined, year, and month, respectively.

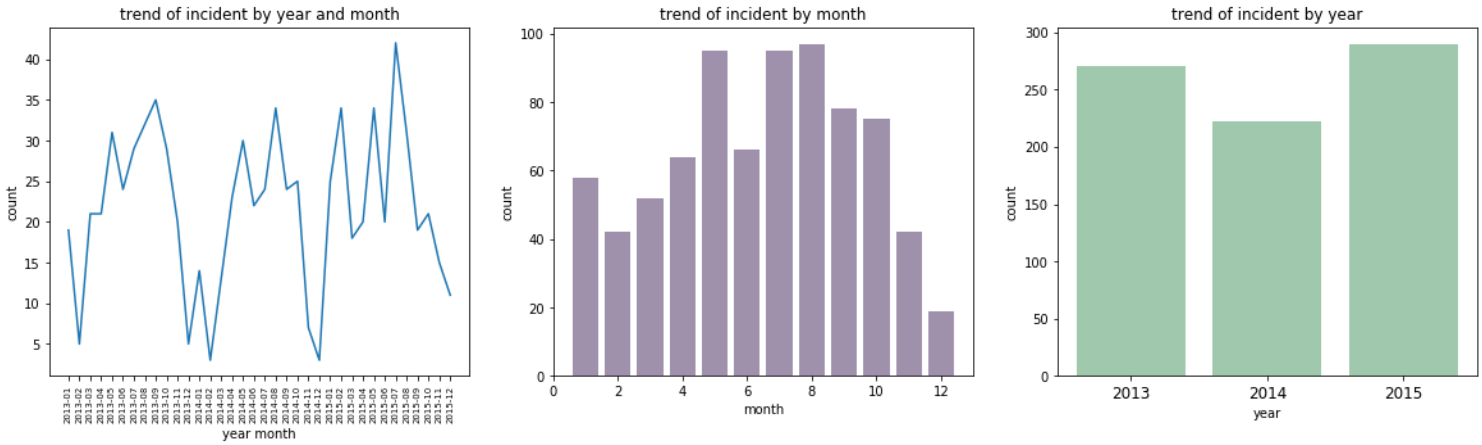


Figure 7: Trend of Incident Counts – Time series

**Graph Insight(fig7):**

1. The leftmost line chart shows noticeable fluctuation in monthly total graffiti cases during the 3 years. Graffiti cleanup slightly decreased in 2014 but bounced back up in 2015. This chart also demonstrates seasonality, where graffiti cleanup seems to drop in December and peak around August.
2. By just looking at the monthly count in the middle bar chart, we notice that May, July, August are the months when graffities counts are the highest. Accordingly, we can suggest local government to focus on graffiti prevention during these periods. In addition, December always has the lowest graffiti counts which seems reasonable, because December is a holiday season.

## 4. Cost Analysis

**Q6**: What insight can we draw from the cost analysis for graffiti cleanup work?

**Data Wrangling:** Group cost components by year-month, month, year.

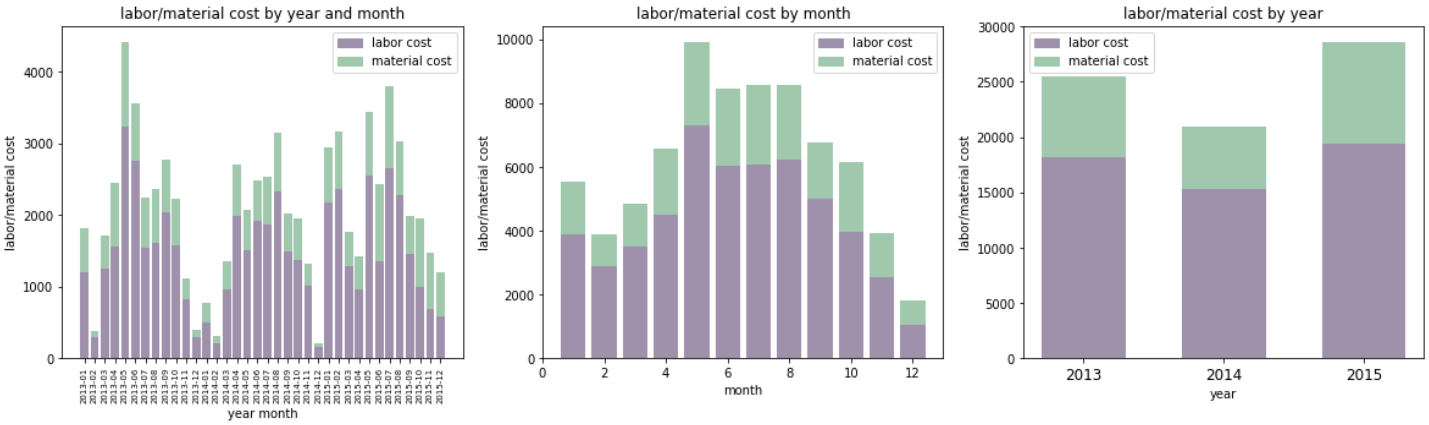


Figure 8: Trend of Incident Counts – Time series

**Graph Insight(fig8):**

1. Total cost trend is consistent with the total graffiti case trend, indicating a consistent unit cost for graffiti cleanup.
2. Labor cost is consistently higher than material cost. Graffiti removal is indeed heavy labor work!

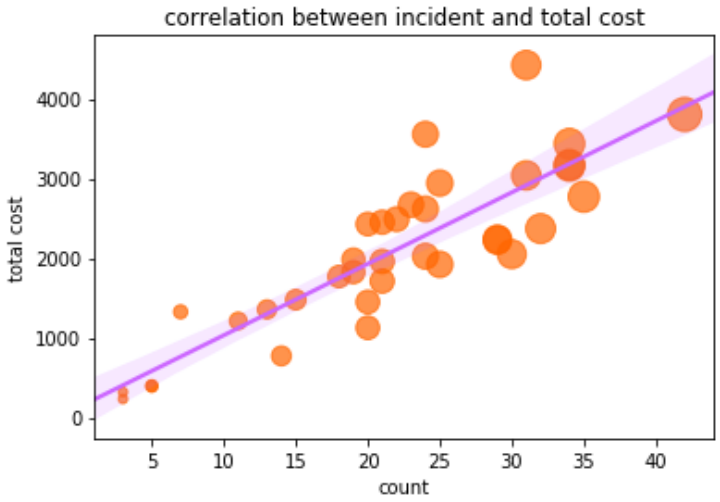


Figure 9: Correlation between Incident counts and total graffiti removal cost

**Graph Insight(fig9):**

1. Fig8 and fig9 together seems to imply a high correlation between total number of graffiti cases and total removal cost during the same period. In order to validate this observation, we conducted a regression analysis by fitting a regression line to a bubble chart to show the correlation between these two variables.

2. The bubble chart in fig9 also conveys information about the magnitude of cost – larger bubbles correspond to higher total cost, and vice versa. In general, bubble size grows in proportion with total graffiti cases, which further validates the close association between case count and total cost.

**Conclusion:**

1. Graffiti occurs more frequently in parks (identified by red dots) vs. non-parks

2. Each year, St. Paul has graffiti clustering around district 3, district 17. District 1 has the lowest graffiti density.

3. The association between graffiti and crime rate in St. Paul is weak, as demonstrated by 2015 data.

4. Graffiti does come back. Many locations experienced repeated cleanups, some as high as 30+ cases.

4. Building, wall, bridge, park, sidewalk, tot lot, parking lot, bench, table, electrical box are the favorite graffiti places for vandals.

6. Graffiti cleanup does display time trend and seasonality. There is a slight dip in total graffiti cleanup cases in 2014, but activities rose back up in 2015. Cleanup activities tend to peak in the summary and became stale in winter times.

5. Graffiti cleanup is labor intensive work, as demonstrated by high labor cost as proportion of the total.

6. There seems to exit a linear relationship between cleanup cases and total costs, indicating a stable unit cleanup price.