[CSE590] MINI PROJECT – 2

[ADITI SINGH SBU ID (110285096) ADISINGH@CS.STONYBROOK.EDU NYC OPEN DATA]

1. Introduction

1.1. Data Set:

- The NYC open data site [5] has been used to make this report. I have looked into primarily 3 data sets: **Water Quality, Lead Kit Requests and 311 Service Requests** data sets. [5] These are fixed schema datasets (automated entry) holding transactional data for every complaint made, when, from where and to whom (it is made).

1.2. Known:

- Water Quality and Lead Kit Requests are primarily derived (filtered) subsets of the 311 service requests data sets and thus follow a similar schema. Some primary cleaning of the data was required, wherein the null values were removed. I would be using the some fields like: ID, Created Date, Borough Name, Zip, Complaint Description, Category and so on.

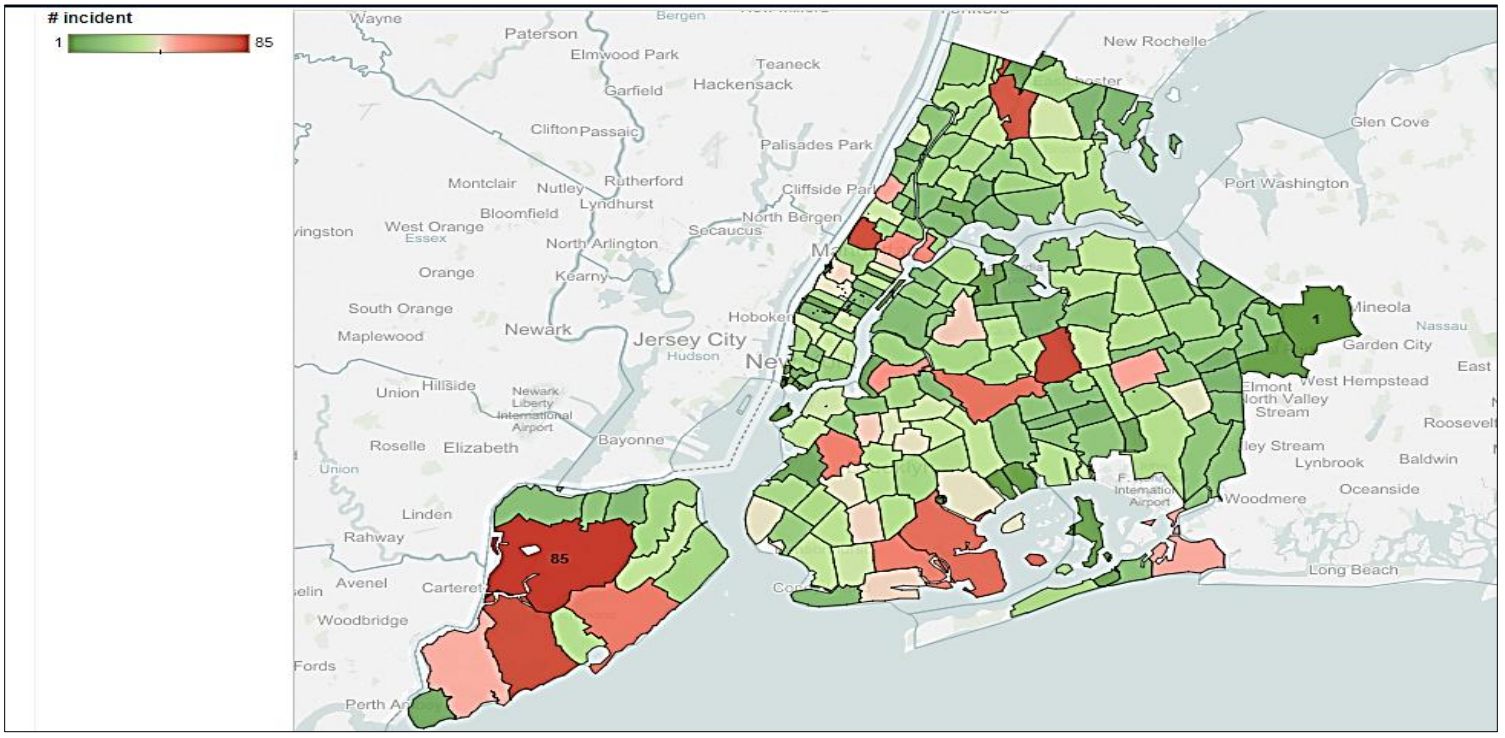
2. Problem (or approach) statement

2.1. Problem:

- There is a huge problem of pollution in this city and therefore a large number of complaints have been filed regarding the contamination of the water. There were some Lead kit requests also raised at the same time.
- **SCOPE:** find how the water quality is related to the lead kit requests. That is, how the trend in both is; are they related and if not, can we have a **smooth function** to define these**. Can the trends spot any anomalies?**

3. Results and Discussion

3.1. Water Quality Complaints (Heat Map):



3.2. Lead Kit Requests (Heat Map)

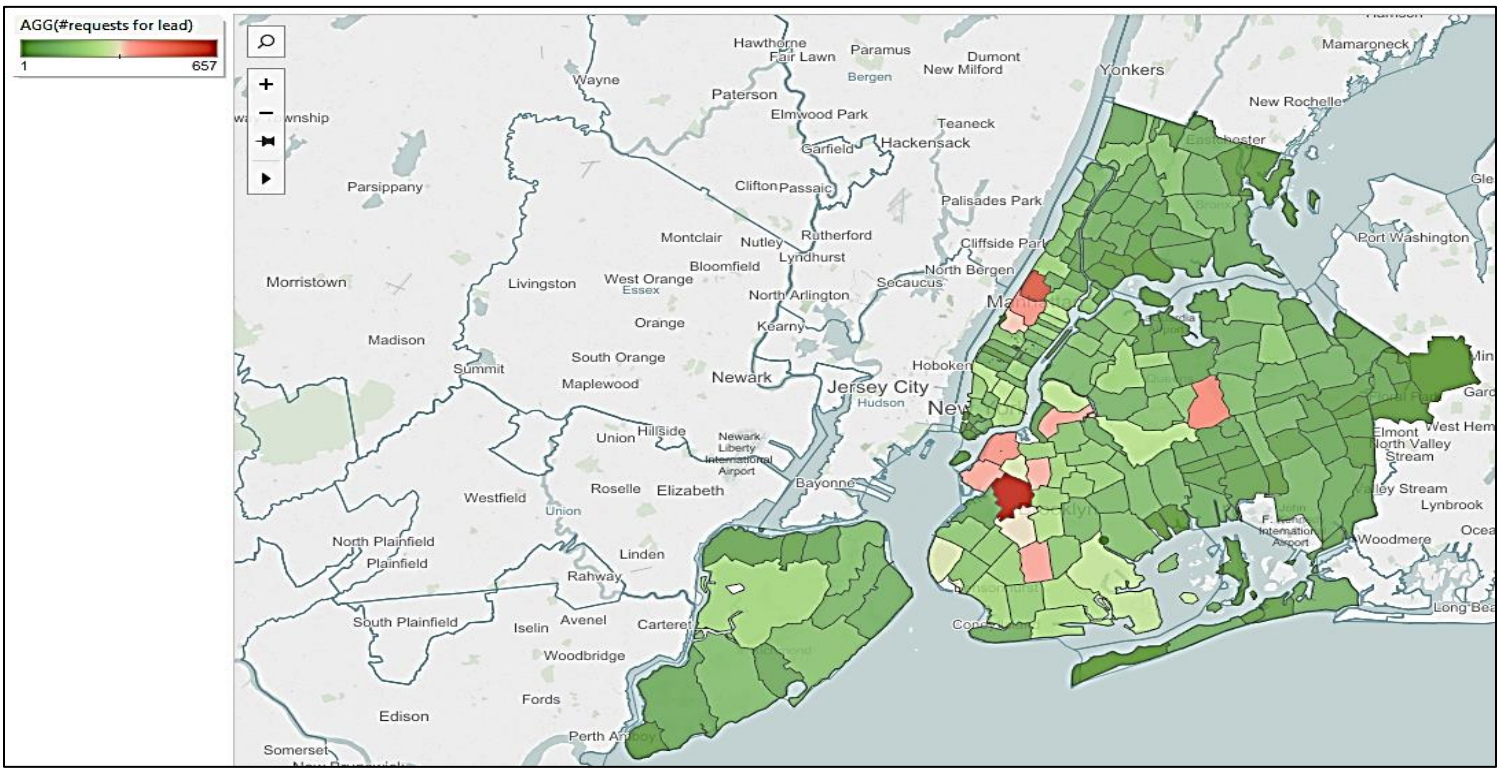


Fig. 1&2 A heat map to plot the # complaints (water) or # requests (Lead kit) per zip code. Red is Highest and Deep Green Lowest

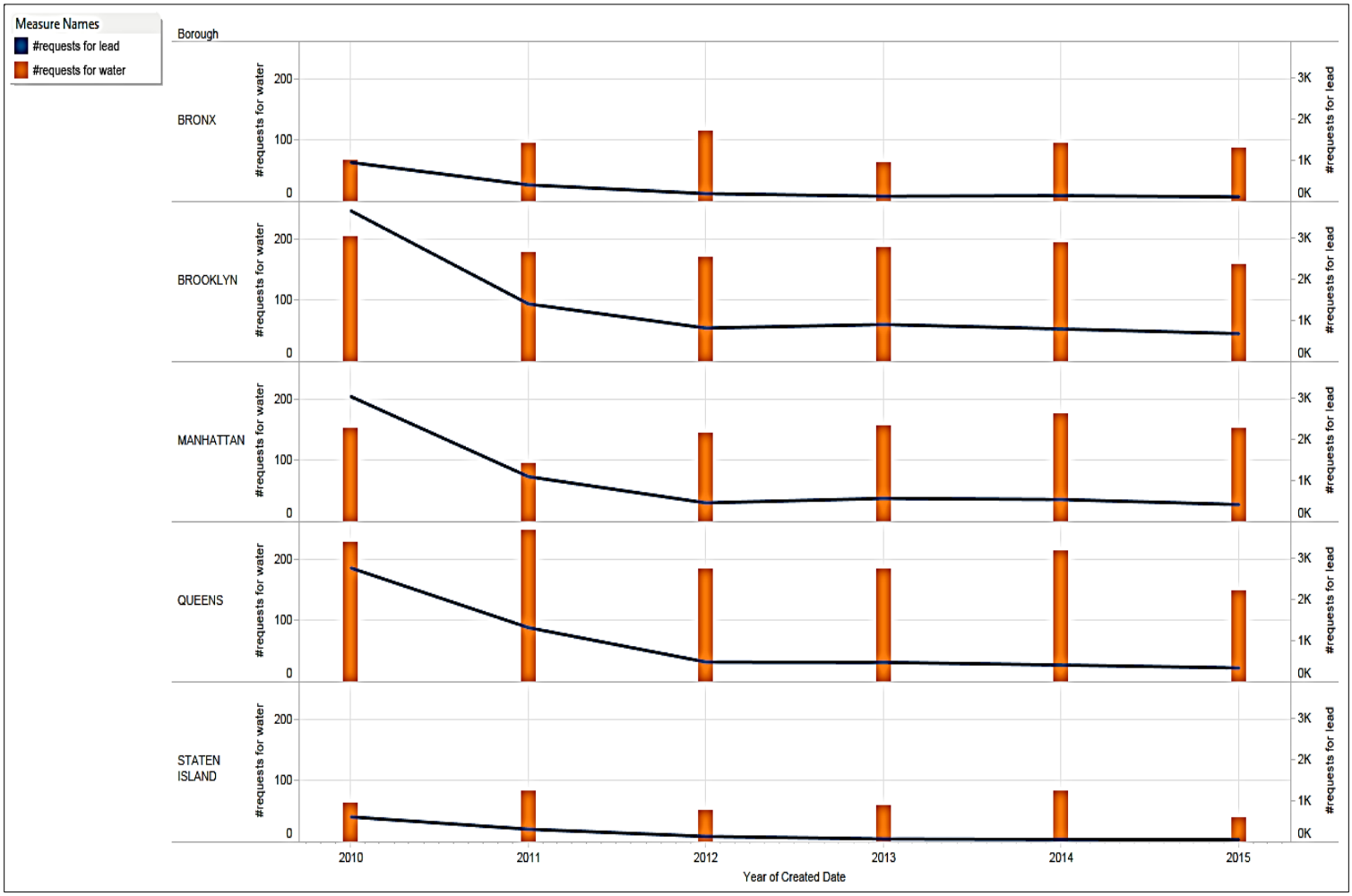


Fig 3: Comparing # Water Quality requests Vs. # Lead Kit Requests (Orange Bar(Water) and Blue Line (Lead).)

3.3. Analyzing 3.1 and 3.2 :

- Looking at both the above heat maps, you can quickly make out that the numbers for "Lead Request kits" and "Water Quality Complaints" seem to be following a **negative correlation** in terms of the ZIP.
- I want to look whether there is a negative correlation in terms of the Year too. For that I am plotting both these **Request Counts in a single plot**. (As Shown Above in Fig.3)
- In fig 4, I have plotted the two curves (# of water complaints and # lead kit requests against Year.)
- Apart from that I wanted to see if a Trend Line which could fit the curves for all the requests (Water Quality and Lead Kit Requests grouped by Borough Name and plotted against Request Created Year.) – Table 1.

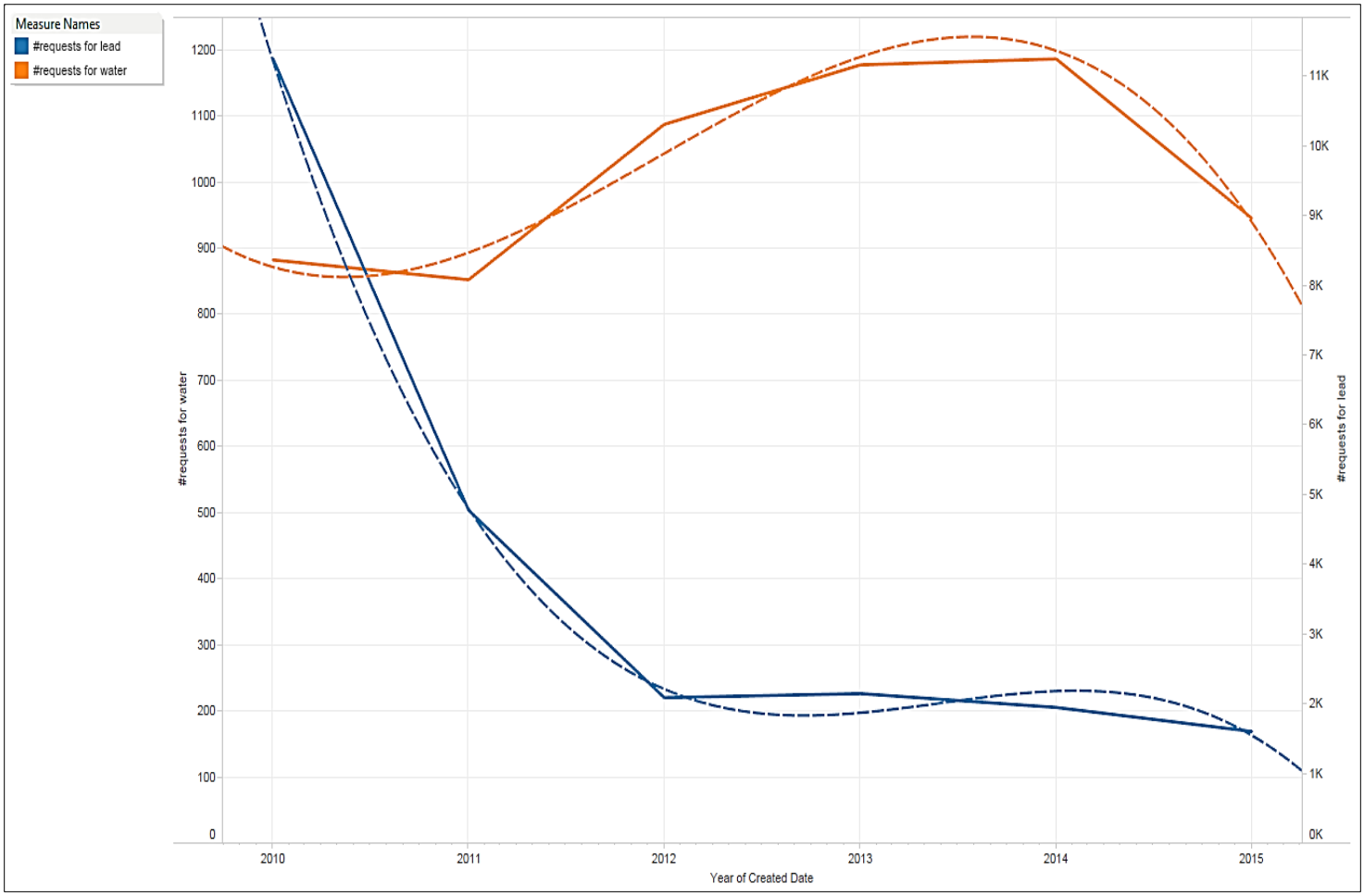


Fig 4: Comparing # Water Quality requests Vs. # Lead Kit Requests (Orange Bar (Water) and Blue Line (Lead).)

[X= Year of Request Created.; Y = # Requests Created]

1. **For the trend curve (Y= #requests for water quality - complaints)**

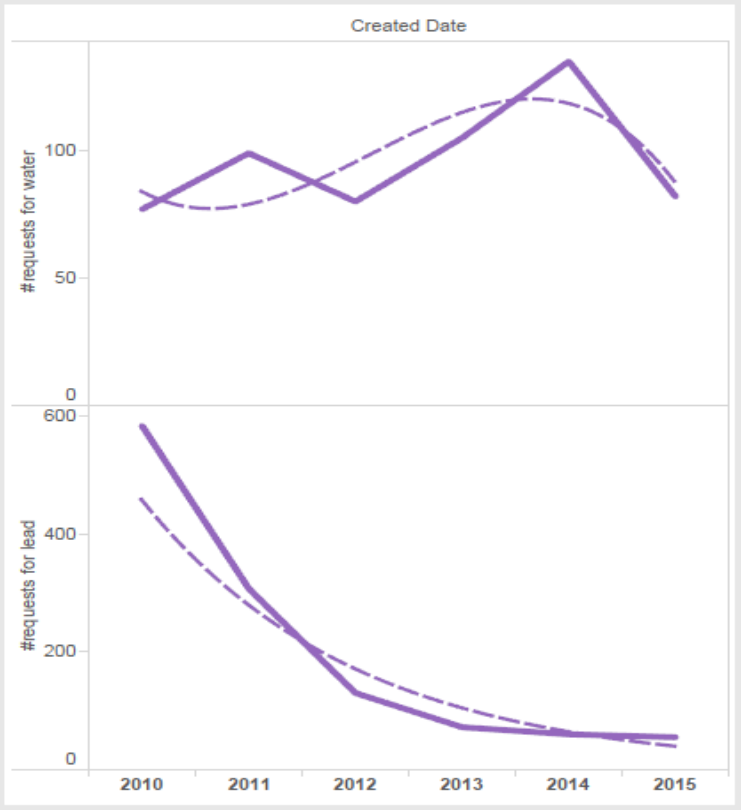
P-value: 0.0552101 **[P value is quite small – so a definite linear trend exists between X and Y]**
Equation: Y = -4.52689e-07*X^3 + 0.0555433*X^2 + -2271.19*X + 3.09513e+07

2. **For the trend curve (Y= #requests for water quality - complaints)**

P-value: 0.0031986 **[P value is quite small – so a definite linear trend exists between X and Y]**
Equation: Y = -5.45903e-06*X^3 + 0.678409*X^2 + -28101.5*X + 3.88001e+08

TABLE 1.
CREATED DATE vs #COMPLAINTS (WATER & LEAD KIT REQUESTS)

STATEN ISLAND



TREND LINE FITTING : (POLYNOMIAL or EXPONENTIAL)

X= Year of Request Created; Y = # Requests Created

1. TREND LINE : (Y= # requests for water (Complaints))

P-value: 0.592693

Equation: Y = -3.12048*X^3 + 18836.7*X^2 + -3.79024e+07*X + 2.54219e+10

2. TREND LINE : (# requests for Lead Kit)

P-value: 0.0028612

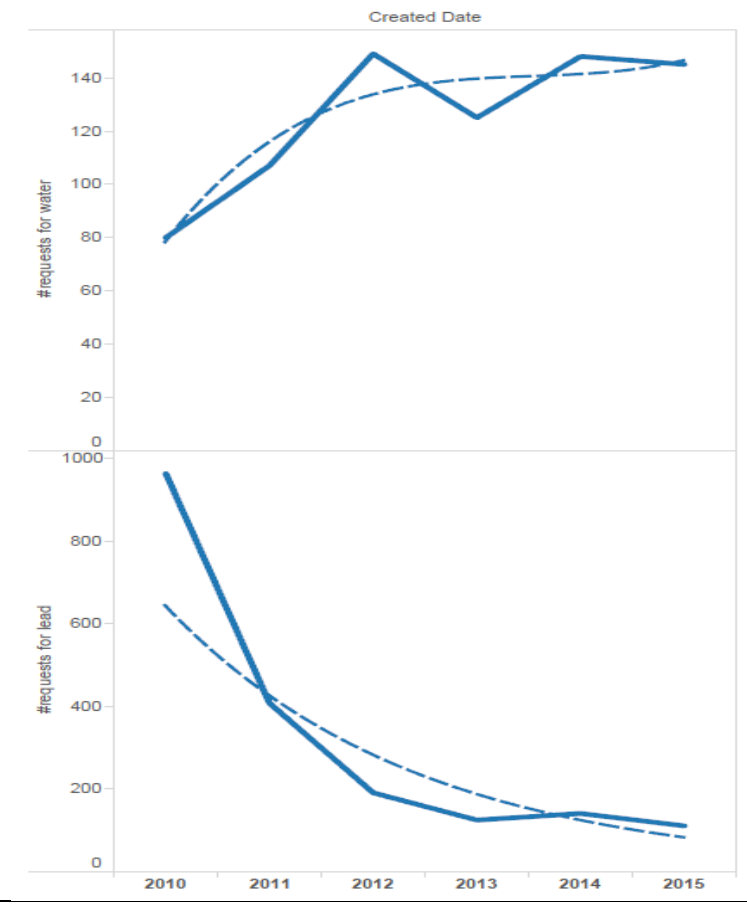
Equation: ln(Y) = -0.494076*X + 999.217

Coefficients

Term	Value	StdErr	t-value	p-value
Year of Created Date	-0.494076	0.0758066	-6.51758	0.0028612
intercept	999.217	152.561	6.54963	0.0028095

[P value is quite small for 2 (but not for 1)– so a definite linear trend exists between X and Y]

BRONX



1. TREND LINE : (# requests for water (Complaints))

(Polynomial Curve Fitting (n=3))

P-value: 0.215511

Equation: Y = 1.27877*X^3 + -7724.58*X^2 + 1.55538e+07*X + -1.04395e+10

2. TREND LINE : (# requests for Lead Kits)

(Exponential curve fitting)

P-value: 0.0108759

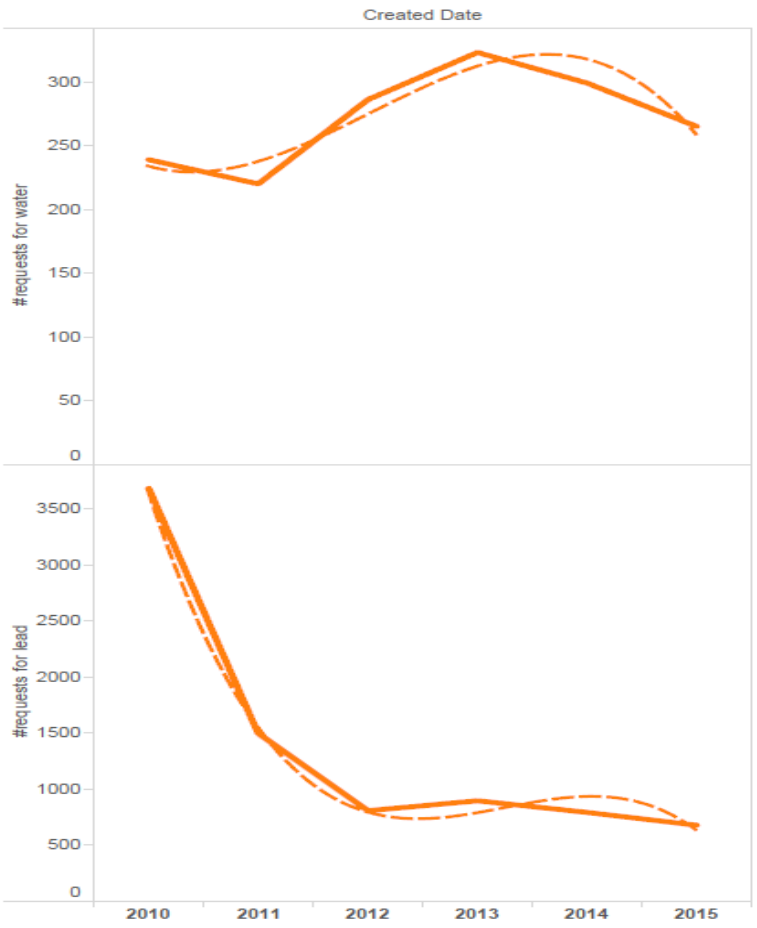
Equation: lnY = -0.41198*X + 834.546

Coefficients

Term	Value	StdErr	t-value	p-value
Year of Created Date	-0.41198	0.0916817	-4.49358	0.0108759
Intercept	834.546	184.51	4.52305	0.0106336

[P value is quite small for 2 (but not for 1)– so a definite linear trend exists between X and Y]

BROOKLYN



1. **TREND LINE : (# requests for water (Complaints))**
   (Polynomial curve fitting (n=3))

P-value: 0.191589

Equation: Y = -5.4489*X^3 + 32889.9*X^2 + -6.61753e+07*X + 4.43821e+10

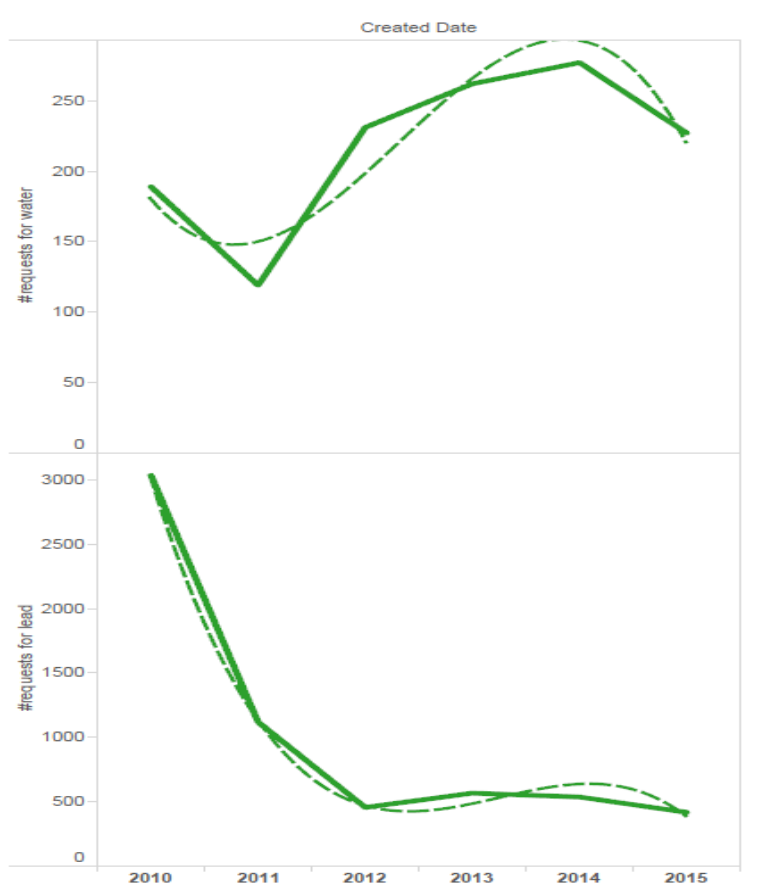
2. **TREND LINE : (# requests for Lead Kits)**
   (Exponential Curve Fitting)

P-value: 0.0083738

Equation: Y = -99.8003*X^3 + 602772*X^2 + -1.21353e+09*X + 8.14386e+11

[P value is quite small for 2 (but not for 1)– so a definite linear trend exists between X and Y]

MANHATTAN



1. **TREND LINE : (# requests for water (Complaints))**
   (Polynomial Curve fitting (n=3))

P-value: 0.213822

Equation: Y = -9.9245*X^3 + 59914*X^2 + -1.20567e+08*X + 8.0873e+10

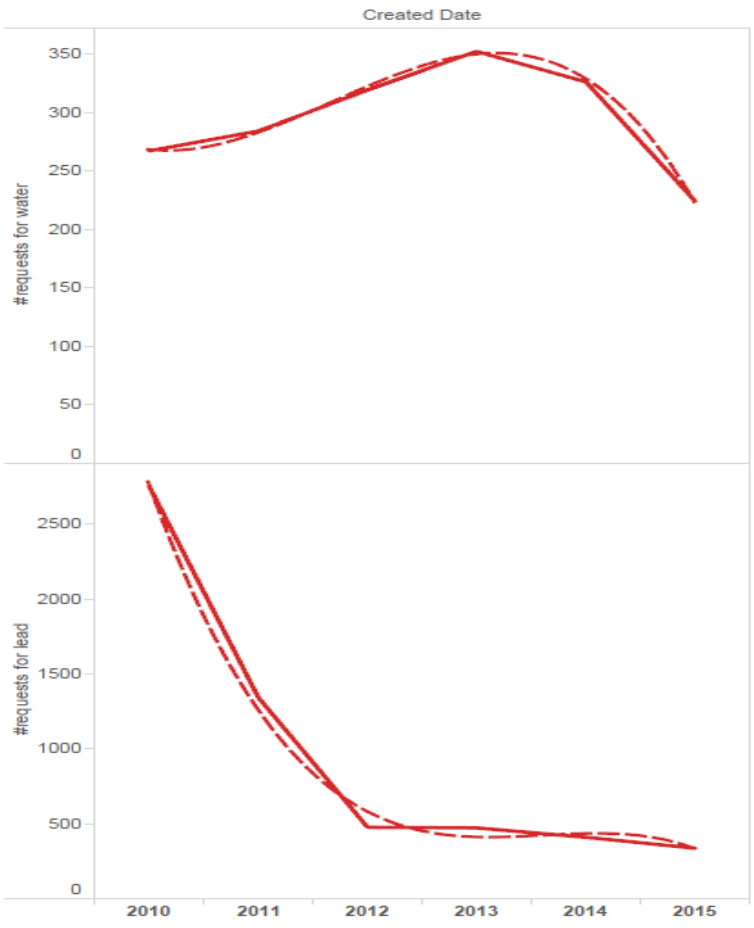
2. **TREND LINE : (# requests for Lead Kit)**
   (Polynomial Curve fitting (n=3))

P-value: 0.0052991

Equation: Y = -90.2287*X^3 + 544962*X^2 + -1.09715e+09*X + 7.36282e+11

[P value is quite small for 2 (but not for 1)– so a definite linear trend exists between X and Y]

QUEENS



1. **TREND LINE : (# requests for water (Complaints))**
   (Polynomial Curve fitting (n=3))

P-value: 0.0032793

Equation: Y = -1.21842e-07*X^3 + 0.0149078*X^2 + -607.897*X + 8.26154e+06

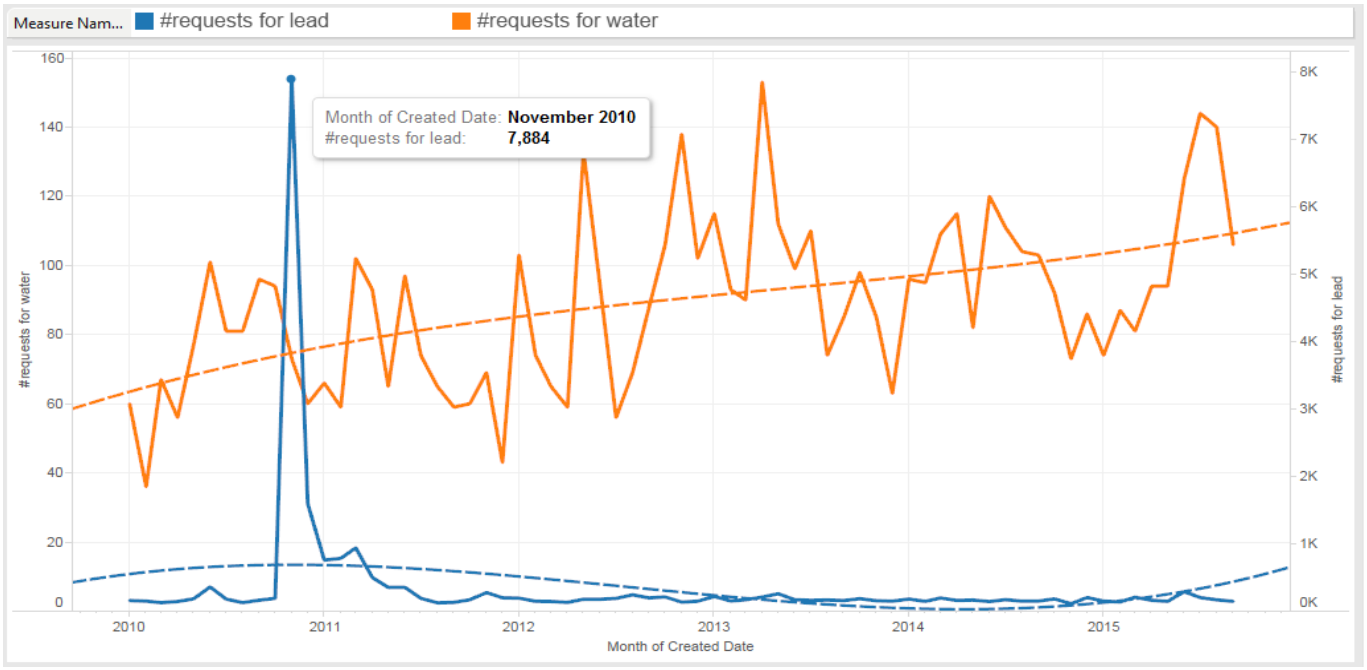
2. **TREND LINE : (# requests for Lead Kits)**
   (Polynomial Curve fitting (n=3))

P-value: 0.0072856

Equation: Y = -1.05377e-06*X^3 + 0.131232*X^2 + -5447.54*X + 7.53764e+07

[P value is quite small for both 1 & 2- so a definite linear trend exists between X and Y]

3.4 Dual Line Chart:



The figure above shows a plot with both the curves (# lead requests and # Water complaints (monthly trend)).

For the curve # lead kit requests – there exists a spike in the month of **November 2010 (**count=7884). The rest of the curve is relatively smooth**. This can be an anomaly**.

For the next curve - # of water complaints per month – there is a lot of unevenness, but looking at the curve fit we can see a definitely a non-decreasing curve. This shows that the quality of water decreases over time, while the lead kit requests are relatively non- increasing.

4. Conclusion:

- From the heat map (3.1) we can see that Staten Island has the highest number of Water complaints in NYC. Most of the areas in this borough are in red, showing a definite bad water quality existence or awareness of its residence. So are some areas in Queens , Brooklyn and Manhattan.
- From the heat map (3.2) we can see that some ZIPs in Manhattan and Brooklyn have the highest number of Lead Kit requests in NYC. Some of the areas in these boroughs are in red, showing lead content presence or awareness of its residence. Manhattan and Brooklyn have both highs in 3.1 and 3.2.
- Figure 4. Shows how both the curves for water complaints and lead kit requests are negatively correlated (while one is decreasing, the other is decreasing).
- The P value calculated shows that - **P value is quite small – so a definite linear trend exists between X and Y for lead while not for all water complaints.**
- For the curve # lead kit requests – there exists a spike in the month of **November 2010 (**count=7884). The rest of the curve is relatively smooth**. This can be an anomaly**.
- While the **Water Quality complaints** are overall **increasing,** the **Lead Requests are mostly stable and non-increasing**., except for a spike in November 2010.

5. References:

1. NYC Environmental Data - *https://nycopendata.socrata.com/*

2. Water Quality Complaints - *https://data.cityofnewyork.us/Environment/Water-Quality-complaints/qfe3-6dkn*

3. Lead Kit Requests - *https://data.cityofnewyork.us/Environment/Lead-Kit-Requests/myrj-umam*

4. 311 Service Requests - *https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9*