Hurricane Sandy: 311 open data analysis of pre and post

Final Paper: Foundation Module

TEAM MEMBERS

1. Kim, Tae (thk301@nyu.edu)
2. Putro, Dimas Rinarso (drp354@nyu.edu)
3. Zhuang, Yuzheng (yz2611@nyu.edu)

Table of contents

Table of figure

1. Introduction

New York City has its own story on how it bounced back from one of the biggest disasters ever hit the East Coast since the 20th Century, Hurricane Sandy. With 147 direct deaths reported, in which 72 of m occurred in the mid-Atlantic and Northeastern US, Hurricane Sandy caused the biggest direct fatalities related to tropical cyclones outside the southern areas since 1972 (Eric et al., 2013). This assignment paper examines 311 dataset changes and patterns in the year 2011, 2012, and 2013 respective to the time range close to Hurricane Sandy (between October 1st – December 31st), to see unseen patterns that can be extracted from the event.

In this team I worked together with Tae Kim (thk301@nyu.edu) and Yuzheng Zhuang (yz2611@nyu.edu). My part is to filter and clean the data for other team members. The other tasks were to make visual presentation using Bokeh library in Python and produce correlation analysis using Pandas, Numpy and Statsmodel package in Python.

2. Time series plot

Our first step is to categorize the complaints by agency and plot the time series to find visually see the patterns. From figure 1 to figure 4 we could see that regardless of the scale of number of complaints, NYPD and TLC were showing minimum response to the event of disaster, constantly maintained the trend without noticeable peaks. DPR, on the other hand showed peaks in 2011 and 2012 in figure 2 and figure 3 respectively, with in 2012 it reached maximum value of 4400. This provided us hints that DPR trends in 2011 and 2012 differ compared to 2013. Another clue provided by this time series plot is that the negatives trends NYPD complaints in 2012 when DPR reached it peak value. From this we understood that we wanted to limit our scope of observation to only DPR and the next step to statistically confirm the state of abnormalities for DPR.

3. ACF and PACF test

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) was preferred to auto regression method since we wanted to particularly set a test to see, through the correlogram tools, whether in specific year DPR presents abnormal behavior, specified by the level of randomness between data point in certain time interval. This could be achieved by observing the relation of the data at *t* and *t-h*, where *h* is a time lag and in this case was set in the range of 1-20 days, and then see whether it falls within certain threshold of level of confidence (95%):

$$\rho_h = Corr(y_t, y_{t-h}) = \frac{\gamma_h}{\gamma_0}.$$

The denominator $\gamma 0$ is the lag 0 covariance. The autocorrelation function (ACF) for a time series y_t, t = 1,...,N, is the sequence ρ_h, h = 1, 2,...,N – 1. The partial autocorrelation function (PACF) is the sequence $\phi_{h,h}$, h = 1, 2, ..., N – 1. The lag-*h* autocorrelation would be then obtained by:

$$\hat{\rho}_h = \frac{\sum_{t=h+1}^{T} (y_t - \bar{y})(y_{t-h} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}.$$

To visually observe this, we could see the correlogram presented in figure 5. We could see from ACF (left) figure that in 2011 and 2012, the lag equal to 0 and 1 fell outside the confidence level threshold, while in 2013 lag = 1 was within the boundaries. In addition, we could see that in 2013, the autocorrelation value for DPR fluctuates in smaller periodic in respect to time intervals. This was caused by increased randomness driven by number of complaints scattered. We also figured out that PACF, while showing similar pattern difference 2011-2012 vs 2013, was more difficult to observe see, thus, for setting up threshold of abnormality for natural disaster events, ACF was considered to be simple, relevant model complementing the use of manual time series observation. From the figure none of the correlation value pass the confidence limit test (all within 95% confidence level), although ACF for 2012 when when lag h is equal to 1 passed the confidence limit test. Therefore, in principle we could not fit any model. Since the data does not show any auto correlation, we decided that ACF and PACF is not be the most accurate test since it failed to predict the patterns in the time series data. However, it might be useful to provide visual representation on how the characteristic of the data in 2011, 2012 and 2013 differs.

4. The use of maps and visual representations

Since we know that statistically there were different in patterns of DPR's number of complaints in 2011-2012 versus 2013, we could not tell what is the particular reason why the pattern was different and what variables were correlated. Thus, we did visual representations by plotting the number of DPR complaints per zip code by longitude and latitude given in 311 dataset, and plot it on the map. The result can be seen in figure 6, 7 and 8. From here we could observe that while Sandy's effect is apparent in 2012, there were significant influx of complaints in Staten Island on 2011, while in 2013 is mostly flat. This might be related to the snow storms that in vast majority occurred in Staten Island in October 2011 (Silive, 2011). The next step is to find good predictor for regression analysis and build the model.

5. Bivariate regression analysis

5.1. Number of tree VS number of complaints

First, we obtained the number of trees data in New York City and try to find the correlation by running bivariate regression model with both linear and second order polynomial regression. This is to confirm that the 2 disasters, snowstorm in 2011 and Sandy in 2012 were assumed largely affected by trees. In this paper we use linear regression equation as follows:

$$Y' = ßx + const$$

Where ß is linear coefficient and *const* is the constants. Second order polynomial, on the other hand, was given in standard polynomial equation:

$$\sum_{i=0}^{n} a_i x^i$$

or, equally:

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

where a_2 and a_1 in the statistics represent $ß_0$ and $ß_1$is and a_0 represents the constants. Our finding shows that in 2011 and 2012, in figure 9 and 10, characterized by increased R–squared value at 0.59 and 0.81 for both linear and second order polynomial fit, respectively. This indicates that the model explains most of the variability of the sample data around its mean. R-squared with the value 0.59 means 59% of the number of DPR complaints were affected by the change of by the number of value. Linear fit, represented by black bold line, pointed an increased value from ß = 1.30 to ß = 1.83 from 2011 to 2012, while it decreased to to from ß = 0.77 in 2013, when there was no disaster events. 2011-2012 presented indication of super linearity correlation, meaning that the number of DPR complaints has more increased rate relative to the number of trees. However, 95% confidence level, as shown in figure 12-14 fell between 0.932 - 1.670 and 1.543 - 2.126 in 2011 and 2012 respectively, which tells us in 2011 there under the 95% level of confidence there was a possibility of sub linear characteristics (ß <1).

The other interesting finding from the tree and DPR number of complaints is that the both of the model, linear and second order polynomial could maintain the similar result, confirming that tree and the number of complaints do scale linearly and therefore no need to use more sophisticated model. This resonates with some emphasis given in class that “in most cases linear approach is the best model to start with”, as linear regression approach was also an interesting subject explored by other urban planners such as in identify the famous “economic of scale” concept proposed in (Luis and Geoffrey, 2007).

5.2. Number of accidents VS number of complaints

Again we ran bivariate regression test with both linear and second order polynomial regression against number of road accidents for each year in 2012 and 2013 due to the limitation that 2011 data is not available. Moreover we were not interested in comparing 2011 because it also refers to other disaster which can state bias. We would like to compare Sandy and non-disaster state, hence 2012-2013 was chosen. Our finding in figure 15 and 16, show correlations characterized by weak R–squared value at 0.01 and 0.04 for 2012 and 2013, respectively. The low R-squared hinted that there relation of accident and DPR complaints were not statistically significant. Moreover, since we know from observation 5.1 that trees are strongly correlated to DPR, we could imply that number of trees does not necessarily correlate with number of accident.

6. Conclusions

In this foundation module final paper we could see how basic applied statistics learned in the class could help us detect abnormality in 311 complaints data in the case of disaster. Moreover, it helped us analyzing some unique phenomenon that have happened during Sandy and be able to provide us hints on which visual representations might be applicable and which new set of predictors could be added before we can evaluate statistically by observing R-squared and so on. For instance, it guided us to the conclusions that tree is the root cause of top 311 complaints during disaster, although, however, it did not cause traffic accident. This capability of manually observe and pick more meaningful model and predictors is essential feature in basic applied statistics when compared to the likes of neural network and other pattern recognition methods.

Table of figures

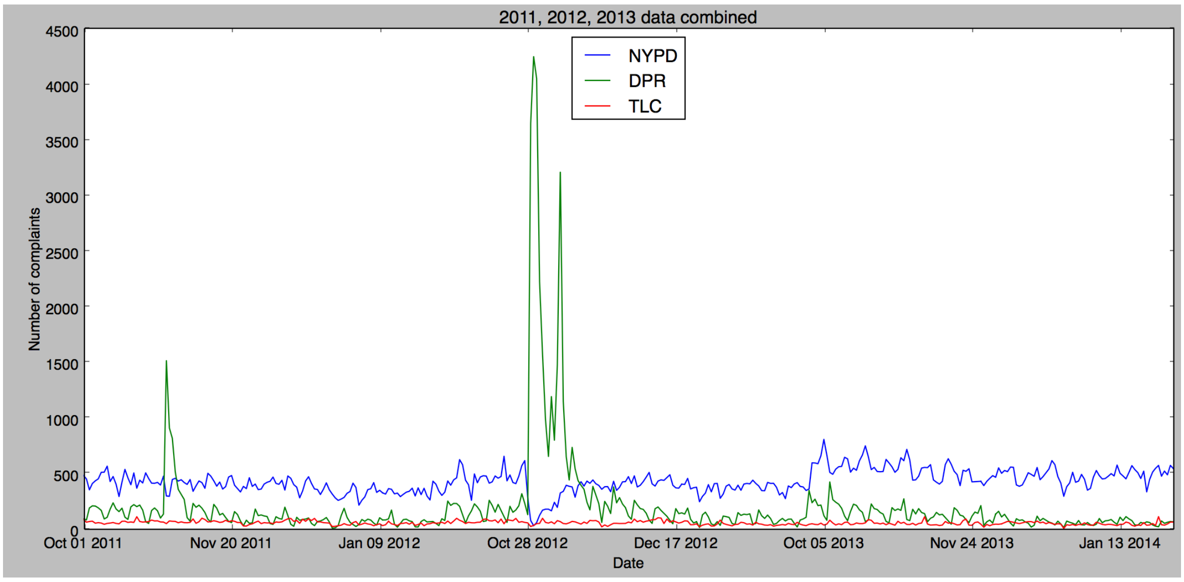

Figure 1 All time series 2011-2014

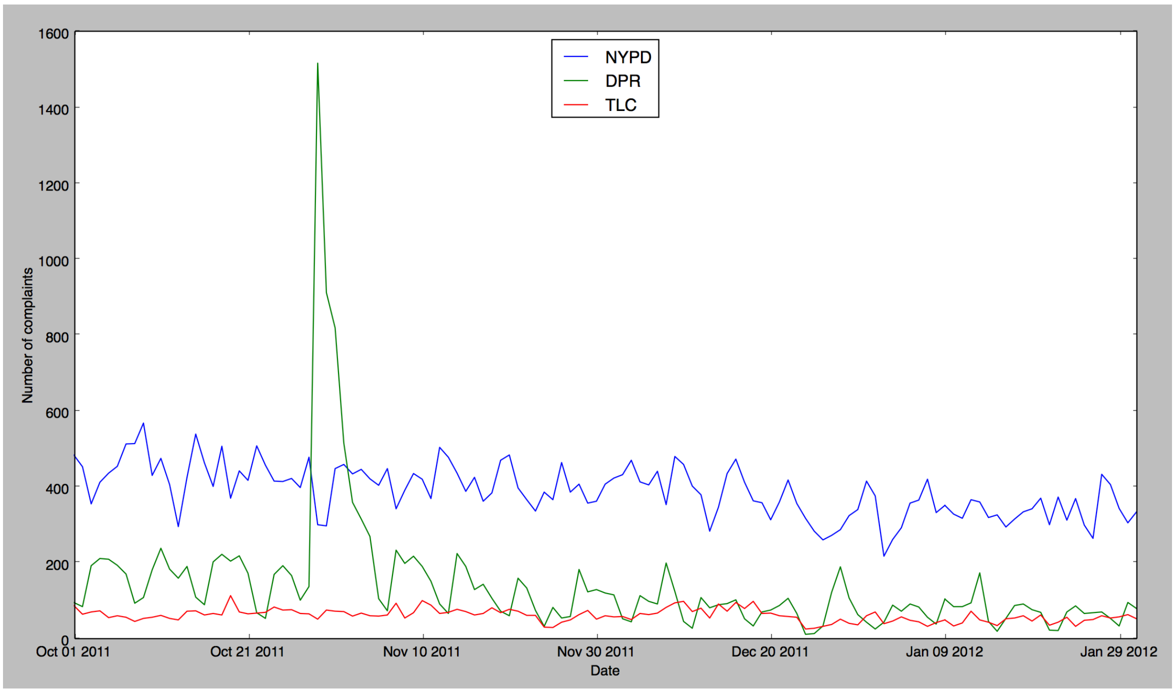

Figure 2 2011 time series plot of top 3 complaints to agency in New York City

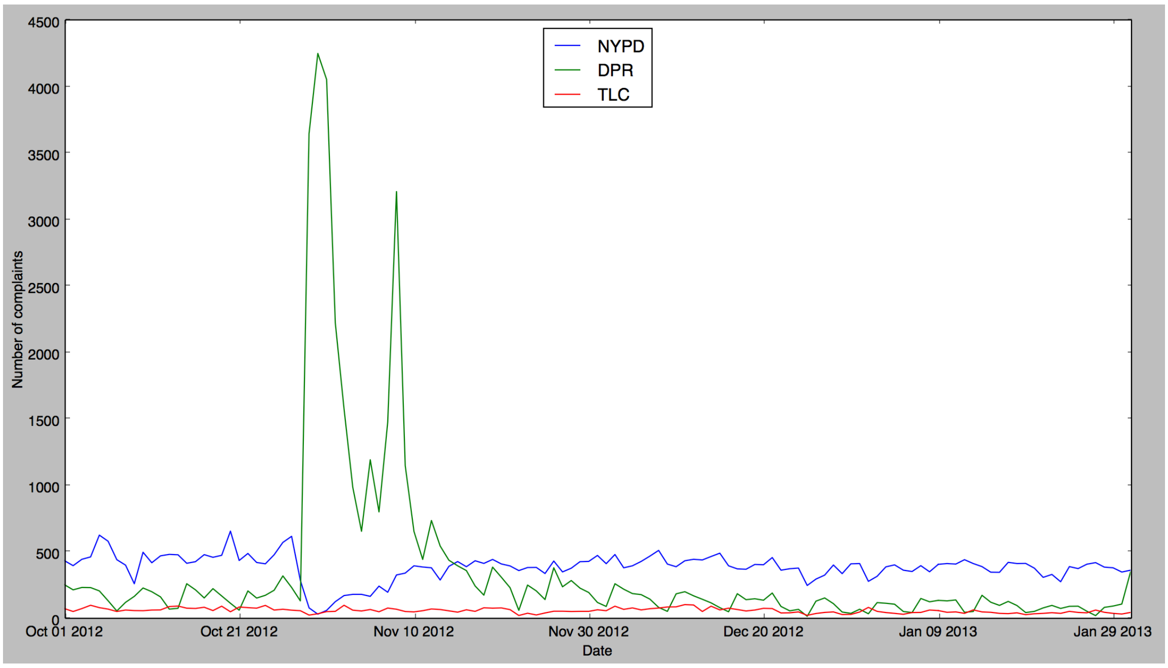

Figure 3 2012 time series plot of top 3 complaints to agency in New York City

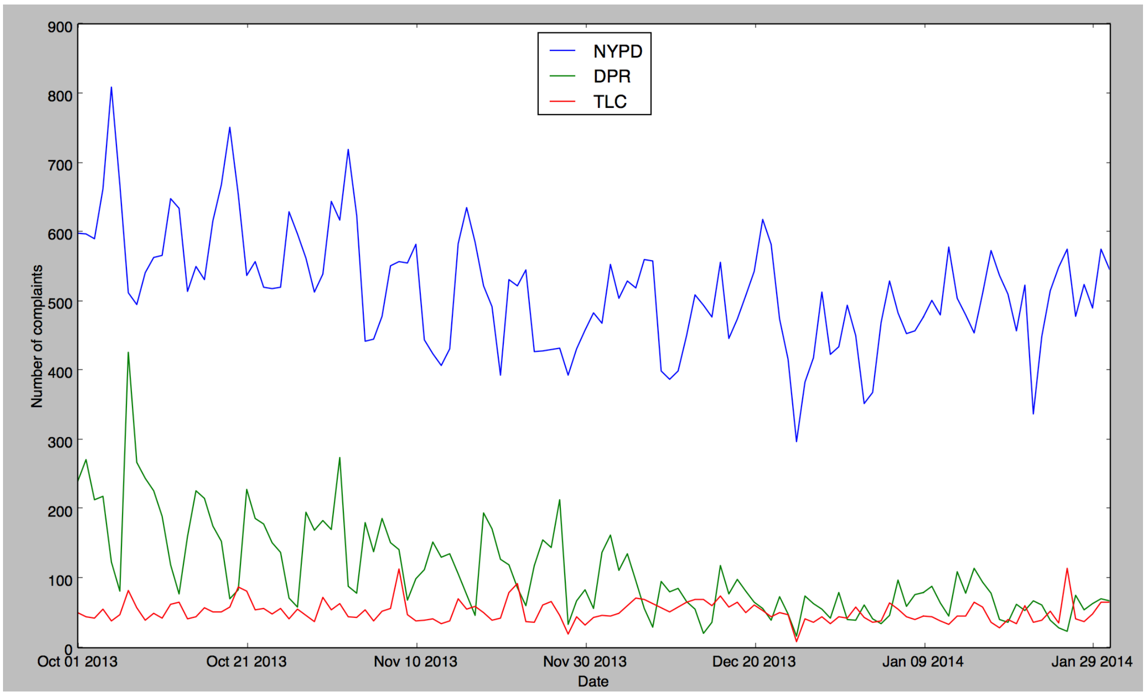

Figure 4 2013 time series plot of top 3 complaints to agency in New York City

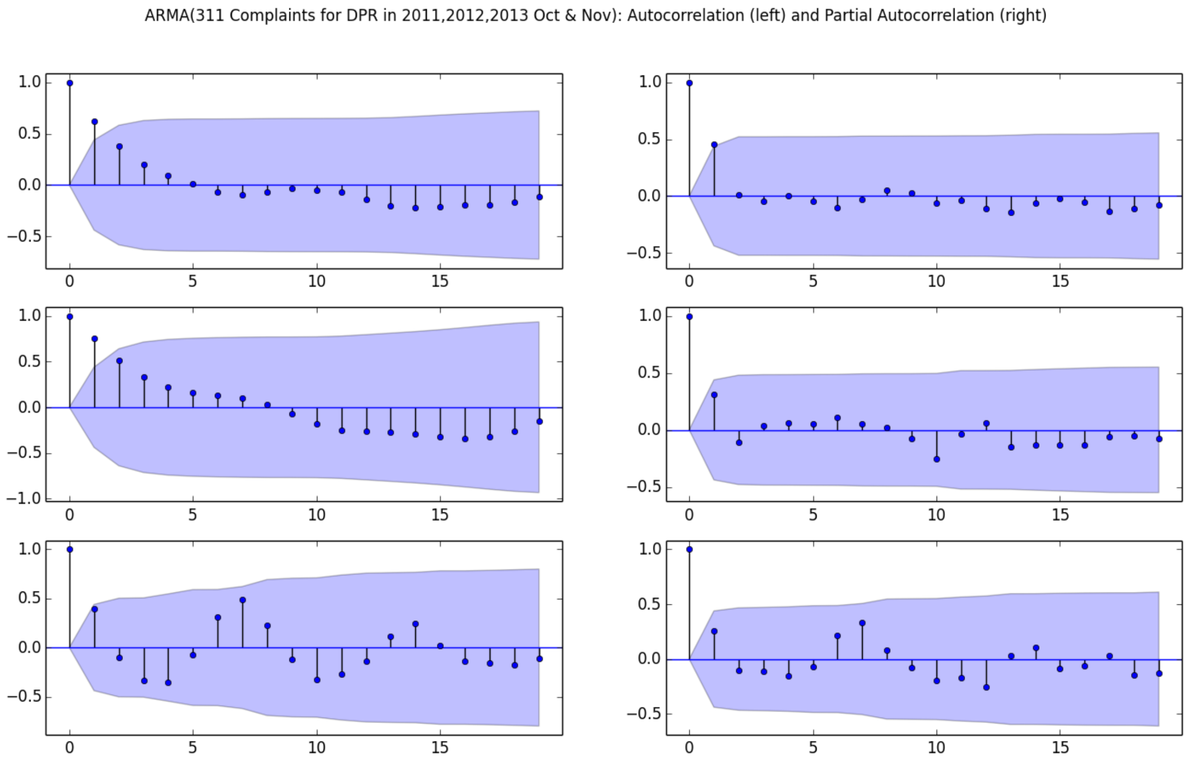

Figure 5 ACF (left) and PACF (Right) of the DPR complaints for Oct and Nov 2011 (top), 2012 (middle), 2013 (bottom)

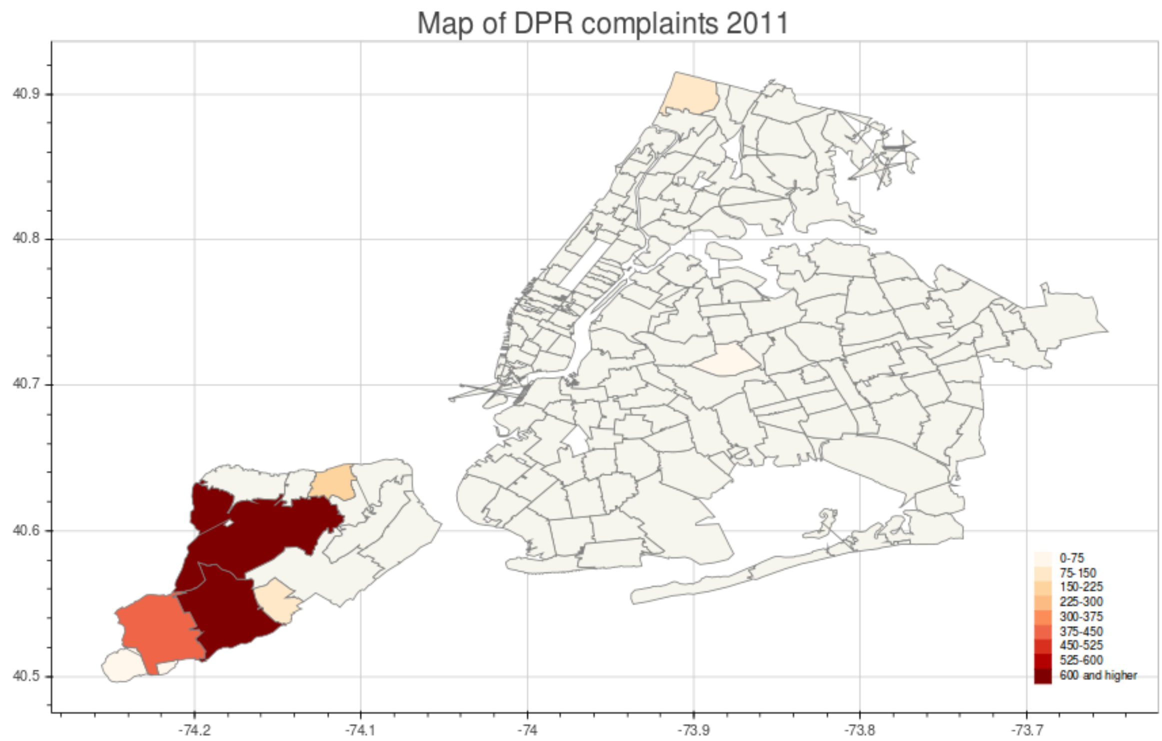

Figure 6 Map of DPR complaints 2011

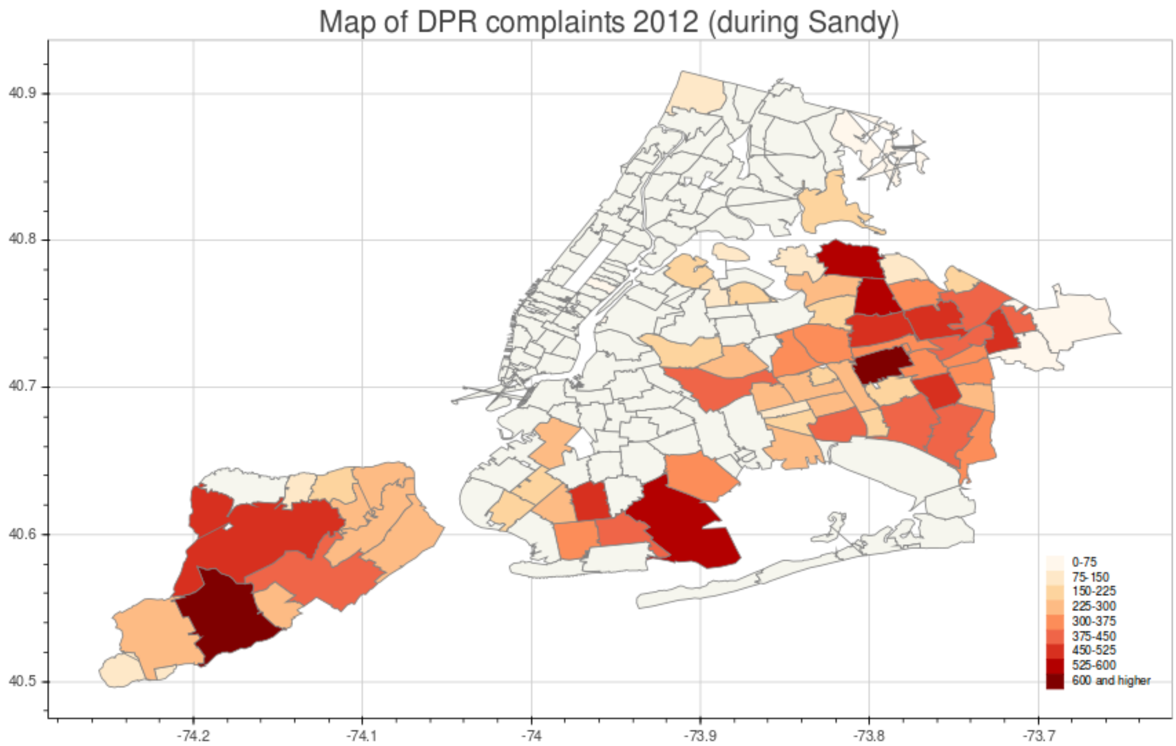

Figure 7 Map of DPR complaints 2012

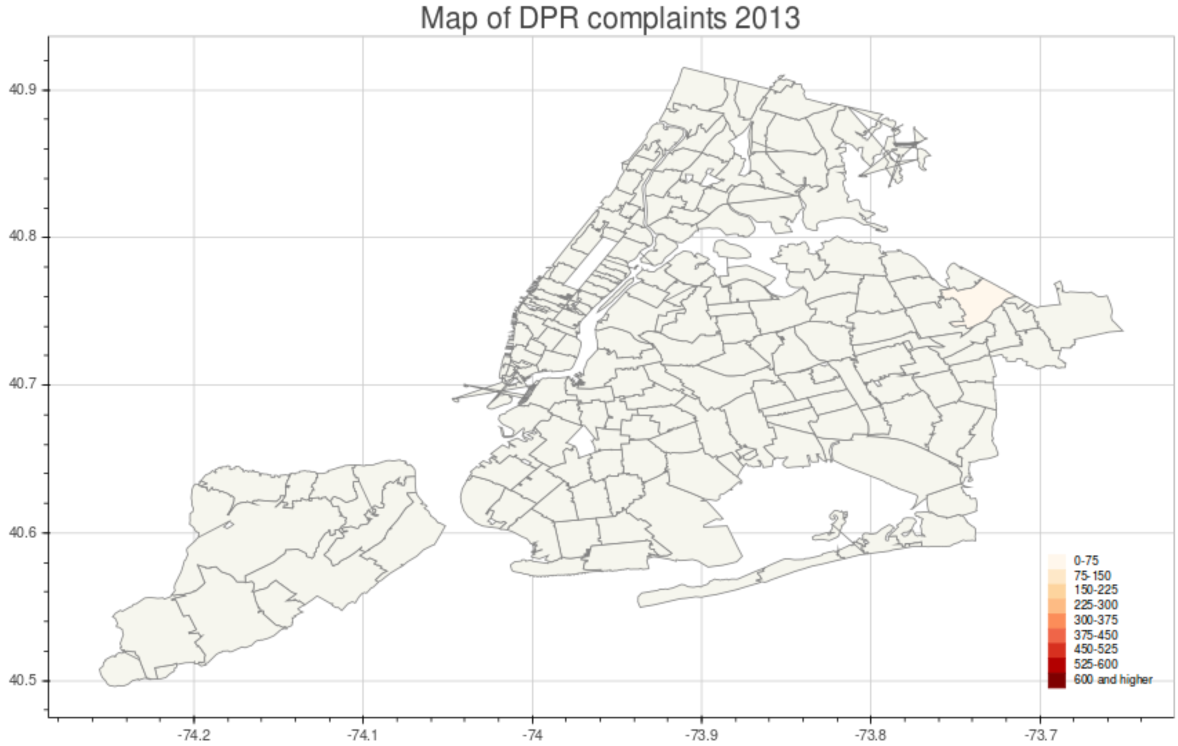

Figure 8 Map of DPR complaints 2013

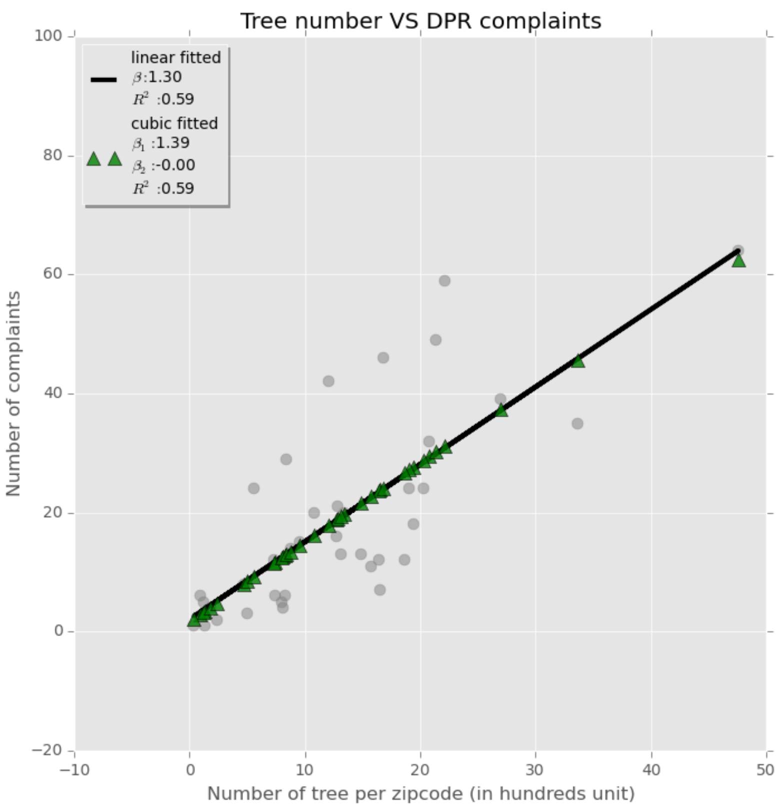

Figure 9 number of trees VS number of DPR complaints 2011

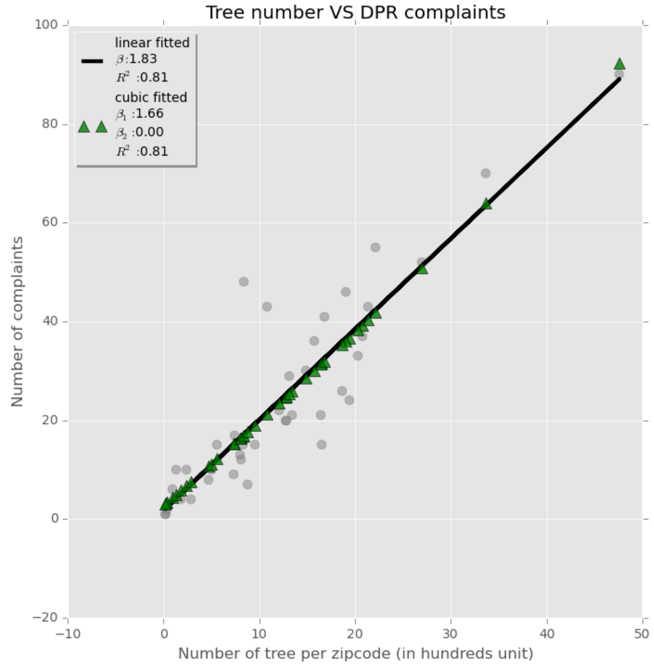

Figure 10 number of trees VS number of DPR complaints 2012

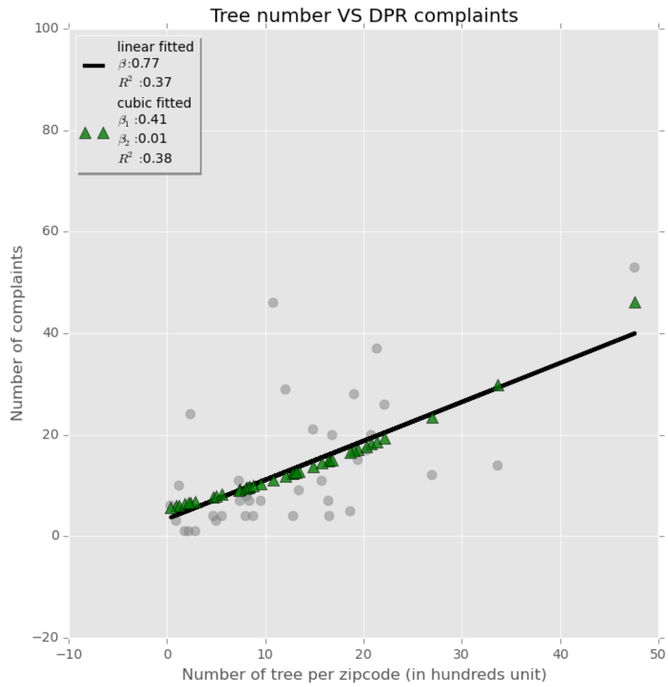

Figure 11 number of trees VS number of DPR complaints 2013

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    num   R-squared:                       0.587
Model:                            OLS   Adj. R-squared:                  0.576
Method:                 Least Squares   F-statistic:                     51.20
Date:                Thu, 27 Nov 2014   Prob (F-statistic):           2.05e-08
Time:                        01:57:23   Log-Likelihood:                -142.86
No. Observations:                  38   AIC:                             289.7
Df Residuals:                      36   BIC:                             293.0
Df Model:                           1
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const          2.0704      2.932      0.706      0.485        -3.875     8.016
counter        1.3011      0.182      7.156      0.000         0.932     1.670
==============================================================================
Omnibus:                        9.272   Durbin-Watson:                   1.961
Prob(Omnibus):                  0.010   Jarque-Bera (JB):                8.296
Skew:                           1.085   Prob(JB):                       0.0158
Kurtosis:                       3.729   Cond. No.                         27.4
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                    num   R-squared:                       0.588
Model:                            OLS   Adj. R-squared:                  0.564
Method:                 Least Squares   F-statistic:                     24.94
Date:                Thu, 27 Nov 2014   Prob (F-statistic):           1.85e-07
Time:                        01:57:23   Log-Likelihood:                -142.84
No. Observations:                  38   AIC:                             291.7
Df Residuals:                      35   BIC:                             296.6
Df Model:                           2
=====================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept             1.5269      4.040      0.378      0.708        -6.676     9.729
counter               1.3873      0.472      2.941      0.006         0.430     2.345
I(counter ** 2.0)    -0.0022      0.011     -0.199      0.844        -0.025     0.021
==============================================================================
Omnibus:                        8.495   Durbin-Watson:                   1.971
Prob(Omnibus):                  0.014   Jarque-Bera (JB):                7.417
Skew:                           1.032   Prob(JB):                       0.0245
Kurtosis:                       3.649   Cond. No.                     1.11e+03
==============================================================================

Warnings:
[1] The condition number is large, 1.11e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 12 Statistics summary for tree VS DPR complaints 2011

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    num   R-squared:                       0.810
Model:                            OLS   Adj. R-squared:                  0.805
Method:                 Least Squares   F-statistic:                     162.5
Date:                Thu, 27 Nov 2014   Prob (F-statistic):           2.68e-15
Time:                        02:00:45   Log-Likelihood:                -142.63
No. Observations:                  40   AIC:                             289.3
Df Residuals:                      38   BIC:                             292.6
Df Model:                           1
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const          1.7568      2.262      0.777      0.442        -2.822     6.336
tree           1.8343      0.144     12.748      0.000         1.543     2.126
==============================================================================
Omnibus:                       17.896   Durbin-Watson:                   2.111
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.930
Skew:                           1.247   Prob(JB):                     1.42e-06
Kurtosis:                       6.152   Cond. No.                         25.7
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                    num   R-squared:                       0.812
Model:                            OLS   Adj. R-squared:                  0.802
Method:                 Least Squares   F-statistic:                     79.88
Date:                Thu, 27 Nov 2014   Prob (F-statistic):           3.75e-14
Time:                        02:00:45   Log-Likelihood:                -142.47
No. Observations:                  40   AIC:                             290.9
Df Residuals:                      37   BIC:                             296.0
Df Model:                           2
==================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept          2.7921      2.980      0.937      0.355        -3.246     8.830
tree               1.6562      0.360      4.599      0.000         0.927     2.386
I(tree ** 2.0)     0.0048      0.009      0.541      0.592        -0.013     0.023
==============================================================================
Omnibus:                       19.755   Durbin-Watson:                   2.080
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               30.815
Skew:                           1.374   Prob(JB):                     2.04e-07
Kurtosis:                       6.308   Cond. No.                         996.
==============================================================================
```

Figure 13 Statistics summary for tree VS DPR complaints 2012

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    num   R-squared:                       0.367
Model:                            OLS   Adj. R-squared:                  0.350
Method:                 Least Squares   F-statistic:                     21.42
Date:                Thu, 27 Nov 2014   Prob (F-statistic):           4.43e-05
Time:                        02:02:34   Log-Likelihood:                -143.53
No. Observations:                  39   AIC:                             291.1
Df Residuals:                      37   BIC:                             294.4
Df Model:                           1
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const          3.4246      2.642      1.296      0.203        -1.928     8.777
tree           0.7677      0.166      4.628      0.000         0.432     1.104
==============================================================================
Omnibus:                       18.019   Durbin-Watson:                   2.272
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               23.554
Skew:                           1.399   Prob(JB):                     7.68e-06
Kurtosis:                       5.582   Cond. No.                         26.7
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                    num   R-squared:                       0.381
Model:                            OLS   Adj. R-squared:                  0.346
Method:                 Least Squares   F-statistic:                     11.06
Date:                Thu, 27 Nov 2014   Prob (F-statistic):           0.000180
Time:                        02:02:34   Log-Likelihood:                -143.09
No. Observations:                  39   AIC:                             292.2
Df Residuals:                      36   BIC:                             297.2
Df Model:                           2
==================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept          5.6142      3.590      1.564      0.127        -1.667    12.896
tree               0.4137      0.426      0.972      0.338        -0.450     1.277
I(tree ** 2.0)     0.0092      0.010      0.903      0.372        -0.011     0.030
==============================================================================
Omnibus:                       20.469   Durbin-Watson:                   2.234
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.947
Skew:                           1.509   Prob(JB):                     3.14e-07
Kurtosis:                       6.053   Cond. No.                     1.08e+03
==============================================================================

Warnings:
[1] The condition number is large, 1.08e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 14 Statistics summary for tree VS DPR complaints 2012

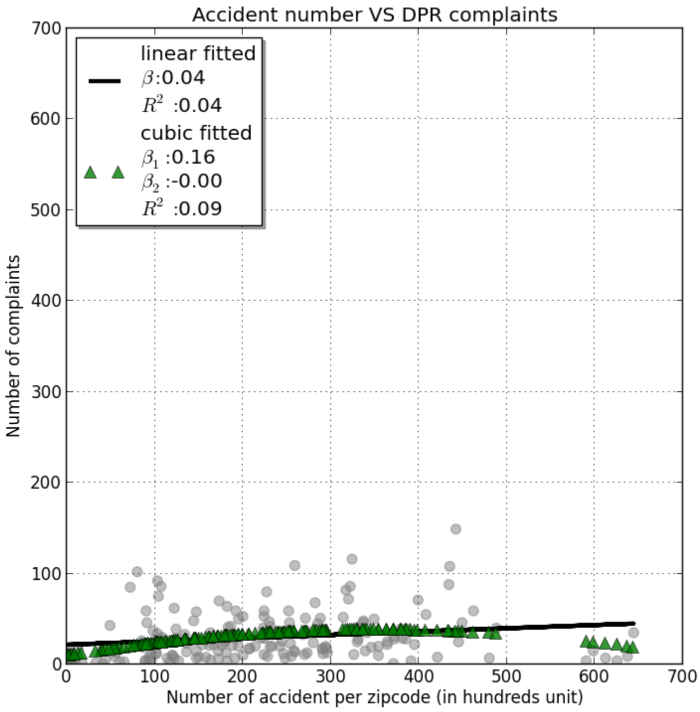

Figure 15 Accident VS DPR complaints 2013

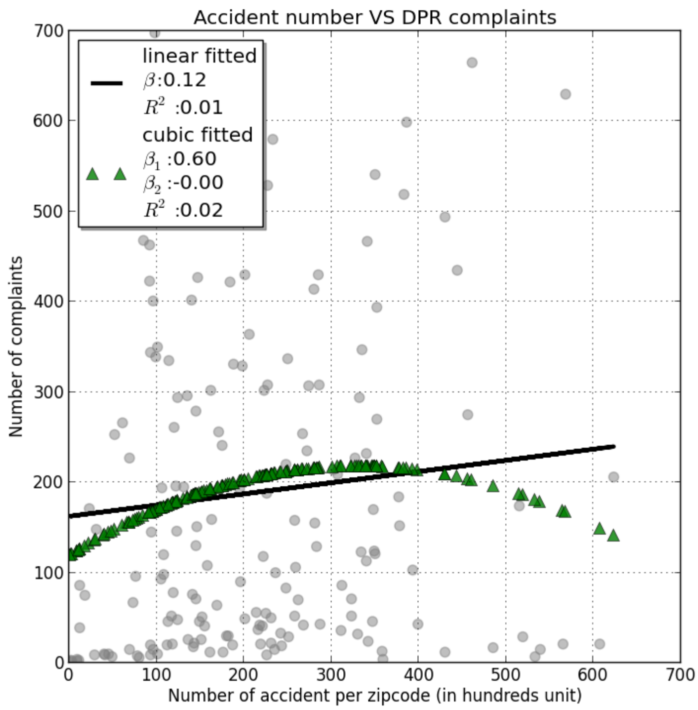

Figure 16 Accident VS DPR complaints 2012

```
OLS Regression Results
==============================================================================
Dep. Variable:                    num   R-squared:                       0.041
Model:                            OLS   Adj. R-squared:                  0.035
Method:                 Least Squares   F-statistic:                     7.354
Date:                Sat, 06 Dec 2014   Prob (F-statistic):            0.00736
Time:                        13:16:53   Log-Likelihood:                -819.77
No. Observations:                 176   AIC:                             1644.
Df Residuals:                     174   BIC:                             1650.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         21.3198      3.713      5.742      0.000        13.992    28.648
accident       0.0288      0.011      2.712      0.007         0.008     0.050
==============================================================================
Omnibus:                       55.258   Durbin-Watson:                   1.237
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              110.711
Skew:                           1.479   Prob(JB):                     9.11e-25
Kurtosis:                       5.519   Cond. No.                         671.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    num   R-squared:                       0.089
Model:                            OLS   Adj. R-squared:                  0.079
Method:                 Least Squares   F-statistic:                     8.499
Date:                Sat, 06 Dec 2014   Prob (F-statistic):           0.000301
Time:                        13:16:53   Log-Likelihood:                -815.16
No. Observations:                 176   AIC:                             1636.
Df Residuals:                     173   BIC:                             1646.
Df Model:                           2
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [95.0% Conf. Int
--------------------------------------------------------------------------------------
Intercept              9.4318      5.326      1.771      0.078        -1.080    19.9
accident               0.1199      0.032      3.791      0.000         0.057     0.1
I(accident ** 2.0)    -0.0001   4.11e-05     -3.049      0.003        -0.000 -4.42e-
==============================================================================
Omnibus:                       63.173   Durbin-Watson:                   1.255
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              144.648
Skew:                           1.615   Prob(JB):                     3.89e-32
Kurtosis:                       6.048   Cond. No.                     5.24e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 5.24e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 17 Statistics summary accident VS DPR complaints 2013

```
OLS Regression Results
==============================================================================
Dep. Variable:                    num   R-squared:                       0.007
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     1.313
Date:                Sat, 06 Dec 2014   Prob (F-statistic):              0.253
Time:                        13:22:43   Log-Likelihood:                -1188.8
No. Observations:                 178   AIC:                             2382.
Df Residuals:                     176   BIC:                             2388.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        162.4972     27.187      5.977      0.000       108.843   216.152
accident       0.1240      0.108      1.146      0.253        -0.090     0.338
==============================================================================
Omnibus:                       32.153   Durbin-Watson:                   0.848
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               43.202
Skew:                           1.165   Prob(JB):                     4.16e-10
Kurtosis:                       3.626   Cond. No.                         471.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    num   R-squared:                       0.019
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     1.733
Date:                Sat, 06 Dec 2014   Prob (F-statistic):              0.180
Time:                        13:22:43   Log-Likelihood:                -1187.7
No. Observations:                 178   AIC:                             2381.
Df Residuals:                     175   BIC:                             2391.
Df Model:                           2
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------------
Intercept              118.7172     40.354      2.942      0.004        39.074   198.361
accident                 0.5989      0.342      1.752      0.082        -0.076     1.273
I(accident ** 2.0)      -0.0009      0.001     -1.464      0.145        -0.002     0.000
==============================================================================
Omnibus:                       33.539   Durbin-Watson:                   0.831
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               45.758
Skew:                           1.195   Prob(JB):                     1.16e-10
Kurtosis:                       3.675   Cond. No.                     2.72e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.72e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 18 Statistics summary accident VS DPR complaints 2013

7. References

1. Blake, Eric S., Kimberlain, Todd B., J. Berg, Robery, P. Cagnialosi, John, and L. Beven II, John, 2013. *"Tropical Cyclone Report: Hurricane Sandy,"* National Hurricane Center*, National Hurricane Center* February 12, 2013.
2. Bloomberg, Michael, 2013. *A Stronger, More Resislient New York.* NYC Special initiative for Rebuilding and Resiliency, 2013, pp 1-438.
3. Silive 2011, *October snow storm pounds Staten Island with 2 inches*. Last updated on November 03 2011. http://www.silive.com/news/index.ssf/2011/10/october_surprise_2_inches_of_s.html.
4. Luís M. A. Bettencourt and Geoffrey B. West, "Bigger Cities Do More with Less", U.S. Patent And Trademark Office (data on patents filed between 2000–2005 for U.S. metropolitan areas).