**Data Analysis of New York 311 Call Center**

Luis Barriere Maurice Tut

Graduate Student Graduate Student

Department of Information Systems Department of Information Systems

California State University California State University

Los Angeles Los Angeles

e-mail: lbarrie9@calstatela.edu e-mail: [mtut@calstatela.edu](mailto:mtut@calstatela.edu)

Narinder Singh Mauricio Franco

Graduate Student Graduate Student

Department of Information Systems Department of Information Systems

California State University California State University

Los Angeles Los Angeles

e-mail: nsingh@calstatela.edu e-mail: mfranco6@calstatela.edu

Jongwook Woo

Department of Information Systems, California State University

Los Angeles

Tel. 323-343-2916, Fax. 323-343--5209

e-mail: jwoo5@caltstatela.edu

**Abstract:** This paper aims at performing data analysis in Hive using Oracle Database for the 311 call center data in New York from 2007 to 2009 along with transferring data with Git Bash and visualization using Power BI. From the data which was available we were able to categorize the calls by different parameters like Agency Name, Complaint Type, Location Type, Incident Address and Street Name etc. The data set is a yearly summary file, containing data for every monitor in the NYC open database.

**1. Introduction**

**3-1-1** is a special [telephone number](https://en.wikipedia.org/wiki/Telephone_number) supported in many communities in [United States](https://en.wikipedia.org/wiki/United_States). The number provides access to non-emergency [municipal](https://en.wikipedia.org/wiki/Municipality) services.This data set is from the NYC Open database tracking calls from 2007 to 2009 [1][2][3]. The data set is a yearly summary file, containing data for every monitor in the NYC Open database. You will know about:

* Complaint types.
* Agencies handling complaints.
* Location type, Incident Address and Street name.
* Transfer data using Git Bash software and scp query.

Visualize using Power BI

2. General Instructions

In this project, we have analyzed and visualized the New York Call Center data. We have downloaded the Dataset from [data.cityofnewyork.us](https://data.cityofnewyork.us/api/views/aiww-p3af/rows.csv?accessType=DOWNLOAD), download to the local system using GIT BASH software. Then you will learn how to upload it to HDFS. You will figure out how to manipulate

and analyze the hazardous air pollutants dataset in HDFS using HiveQL[4]. You will also practice how to visualize the result in Power BI.

**2.1.1 311 Calls data downloaded from** [**data.cityofnewyork.us**](https://data.cityofnewyork.us/api/views/aiww-p3af/rows.csv?accessType=DOWNLOAD)

The first stage of this project was downloading the dataset from [https://data.cityofnewyork.us](https://data.cityofnewyork.us/api/views/aiww-p3af/rows.csv?accessType=DOWNLOAD). To download it, we need to create an account on the website. We won’t be able to download the dataset without the account. The dataset is downloaded and saved in the excel file format.

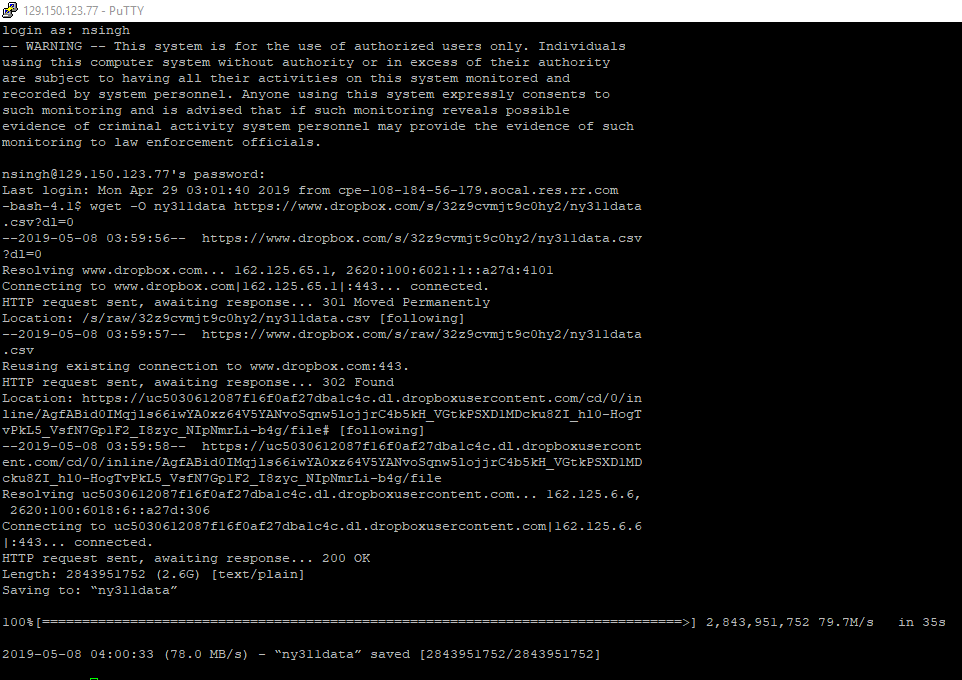


**Figure 1: Downloading the Dataset**

The Figure 1 describes about the dataset which is in .xls format and has total fifty two columns. This data set is from the NYC OpenData repository tracking calls from 2007 to 2009. The data set is a yearly summary file, containing data for every monitor in the NYC database.

**2.1.2 Hazardous Air Pollutants data loaded into Local Directory**

The next stage was uploading the dataset from desktop to local directory using Putty[5]. The dataset has to be in csv file before loading. This can be explained below.



**Figure 2: Downloading the Dataset using shell command**

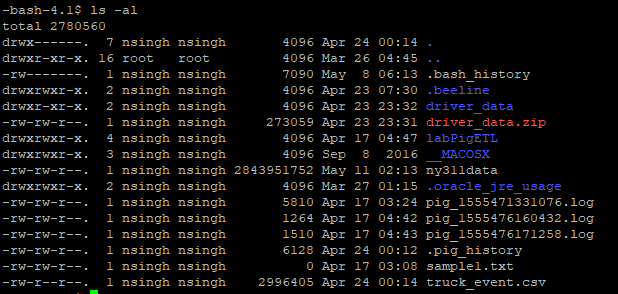
**2.1.3 NYC Data loaded into HDFS File system using Shell commands**

In this stage we have to put the csv file into HDFS. We have to to remotely access the oracle cloud using ssh with the help of a terminal console window like putty, ubuntu or mac terminal.

**$ ls –al**

The above queries will check if the file is downloaded.

If the data is present, we can see it in Figure 3.



**Figure 3: csv Availability**

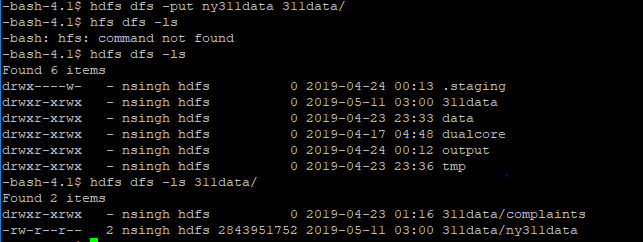
**hdfs dfs -mkdir 311data**

**hdfs dfs -put ny311data.csv 311data/**

**hdfs dfs -ls 311data/**

The above queries will create a directory in hdfs and put the csv file in the hadoop directory[5].

If the data is present, we can see it in Figure 4.



**Figure 4: Data transferred from Local Directory to HDFS**

**2.1.4 Create Hive table Query to NYC311 data**

The following external table is created using a Hive statement that allows Hive query stored in HDFS. External tables preserve the data in the original file format, while allowing Hive to perform queries against the data within the file.The Hive statement below create a new table named, airpollution, by describing the fields within the files, the delimiter (comma) between fields [6]. This will allow you to create Hive queries over your data.

Open Hive shell environment as follow

**$ beeline**

In the beeline, you need to copy and paste the following beeline link to get into hive and use your database name.

(**Note:** Don’t forget to replace the blue part with your beeline link and red part with your database name)

!connect jdbc:hive2://cis5200spr19-bdcsce-2.compute-608214094.oraclecloud.internal:2181,cis5200spr19-bdcsce-3.compute-608214094.oraclecloud.internal:2181,cis5200spr19-bdcsce-4.compute-608214094.oraclecloud.internal:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive bdcsce\_admin

use nsingh;

In the Hive shell, you need to copy and paste the following Hive QL code to create an external table “airpollution”.

(Note: Don’t forget to replace the red part with your database name)

**CREATE EXTERNAL TABLE IF NOT EXISTS ny311data(**

**unique\_key STRING,**

**created\_date STRING,**

**closed\_date STRING,**

**agency STRING,**

**agency\_name STRING,**

**complaint\_type STRING,**

**descriptor STRING,**

**location\_type STRING,**

**incident\_zip INT,**

**incident\_address STRING,**

**street\_name STRING,**

**cross\_street\_1 STRING,**

**cross\_street\_2 STRING,**

**intersection\_street\_1 STRING,**

**intersection\_street\_2 STRING,**

**address\_type STRING,**

**city STRING,**

**landmark STRING,**

**facility\_type STRING,**

**status STRING,**

**due\_date STRING,**

**resolution\_action\_updated\_date STRING,**

**community\_board STRING,**

**borough STRING,**

**x\_coordinate\_state\_plane BIGINT,**

**y\_coordinate\_state\_plane BIGINT,**

**park\_facility\_name STRING,**

**park\_borough STRING,**

**school\_name STRING,**

**school\_number STRING,**

**school\_region STRING,**

**school\_code STRING,**

**school\_phone\_number STRING,**

**school\_address STRING,**

**school\_city STRING,**

**school\_state STRING,**

**school\_zip STRING,**

**school\_not\_found STRING,**

**school\_or\_citywide\_complaint STRING,**

**vehicle\_type STRING,**

**taxi\_company\_borough STRING,**

**taxi\_pick\_up\_location STRING,**

**bridge\_highway\_name STRING,**

**bridge\_highway\_direction STRING,**

**road\_ramp STRING,**

**bridge\_highway\_segment STRING,**

**garage\_lot\_name STRING,**

**ferry\_direction STRING,**

**ferry\_terminal\_name STRING,**

**latitude BIGINT,**

**longitude BIGINT,**

**location BIGINT**

**)**

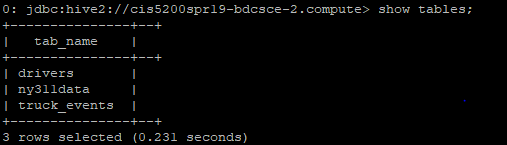
**ROW FORMAT DELIMITED FIELDS TERMINATED BY ','**

**STORED AS TEXTFILE LOCATION '/user/nsingh/311data'**

**TBLPROPERTIES('serialization.null.format'='');**

Then in the Hive shell, you need to check if the table “ny311data” is shown:

**hive>show tables;**



**Figure 5: Displaying the table**

**CREATE EXTERNAL TABLE IF NOT EXISTS complaints(created\_date STRING,**

**complaint\_type STRING,**

**city STRING,**

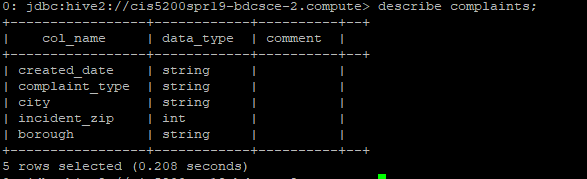
**incident\_zip INT,**

**borough STRING)**

**ROW FORMAT DELIMITED FIELDS TERMINATED BY ','**

**STORED AS TEXTFILE LOCATION '/user/nsingh/311data/complaints';**

**describe complaints;**



**Figure 6: Discribing the complaints**

**--populate table complaints with data from the ny311data table**

**2.1.5 Creating Hive Queries to Analyze Data.**

The following Hive Query will show you the first 10 complaints:

**INSERT OVERWRITE TABLE complaints**

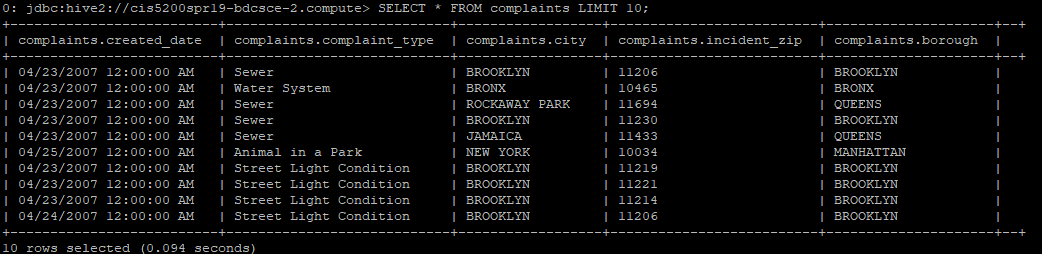
**SELECT created\_date, complaint\_type, city, incident\_zip, borough**

**FROM ny311data**

**WHERE city IS NOT NULL**

**AND incident\_zip IS NOT NULL;**

**SELECT \* FROM complaints LIMIT 10;**



**Figure 7: Displaying the first 10 complaints**

**2.1.6 School complaint data (gather geographic location, complaint & due date to be analyzed)**

The following Hive Query will show you the schools information:

Create a table that stores schools, their location along with corresponding due date for complaints:

**CREATE EXTERNAL TABLE IF NOT EXISTS schools (**

**school\_name STRING,**

**school\_city STRING,**

**school\_state STRING,**

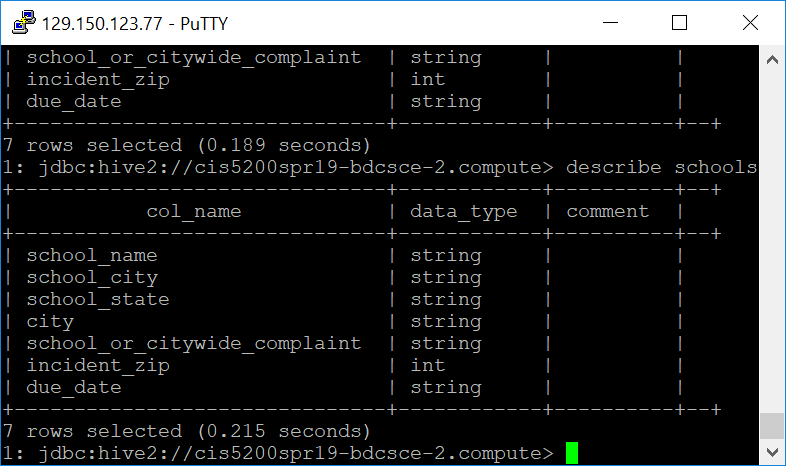
**INSERT OVERWRITE TABLE schools**

**SELECT school\_name, school\_city, school\_state, city, school\_or\_citywide\_complaint, incident\_zip, due\_date**

**FROM ny311data**

**WHERE city IS NOT NULL**

**AND school\_name IS NOT NULL)**

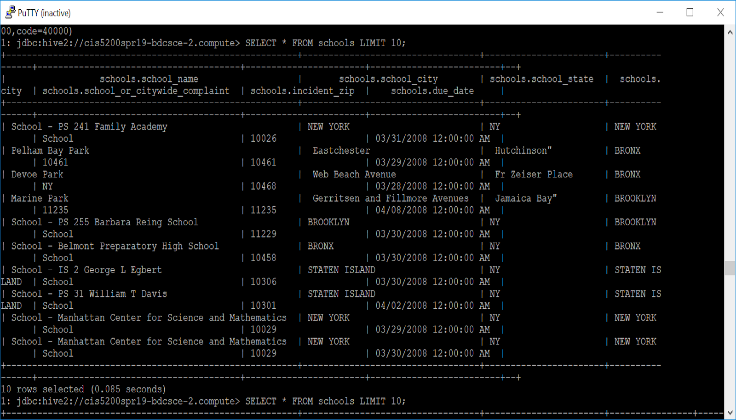


**Figure 8: Describing schools**

**Populating and verifying corresponding values to analyze school complaints, location and due dates based on the fields selected for table ‘**schools’

The following Hive Query will show you the school complaints:

**SELECT \* FROM schools LIMIT 15;**

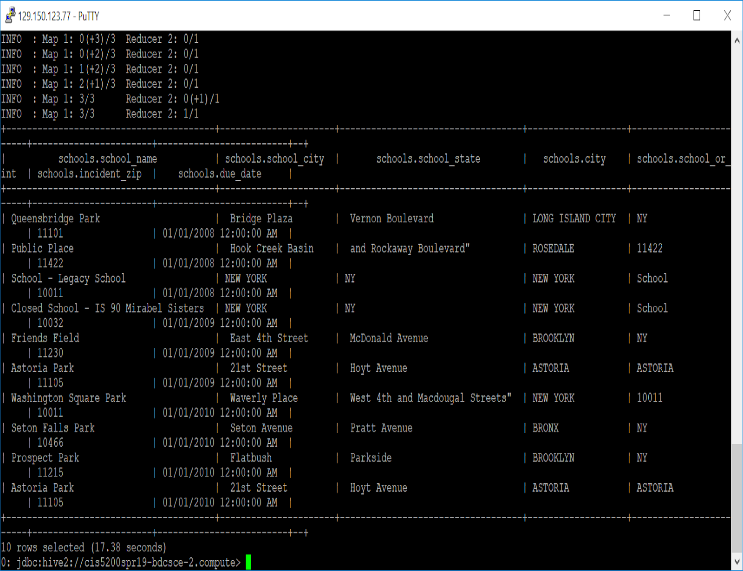
****

**Figure 9: School complaints**

**Lastly, Organize School complaints results in ascending order based on their “due\_date”**

The following Hive Query will show you the top 10 school complaints in ascending order:

**SELECT \* FROM schools ORDER BY due\_date ASC LIMIT 10:**

****

**Figure 10: school complaints in ascending order**

**2.2 Create Tables for Power BI**

In these sections it will described in detail about different tables which are created in Power BI [5]. Since the dataset is too big, you need to create a table for each Hive Queries:

**CREATE TABLE 311data**

**ROW FORMAT DELIMITED FIELDS TERMINATED BY ','**

**STORED AS TEXTFILE LOCATION "/user/nsingh/complaints"**

**AS**

select count (\*) as school\_zip, school\_state

from complaints group by school\_state

order by school\_zip DESC limit 10;

Replace RED TEXT with each query that you need to create a table. Replace YELLOW TEXT with the corresponding table name that you want to name it and Replace GREEN TEXT with the text file storage location.

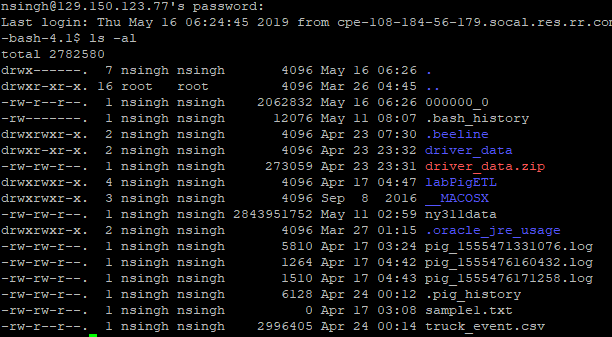
Download the output file 000000\_0 at the HDFS path ‘user/nsingh/complaints’

**hdfs dfs -ls /user/nsingh/311data/complaints**

**hdfs dfs -get /user /nsingh /311data /complaints / 000000\_0**

Use the following command to verify that the data was save to the local file system

**$ ls -al**

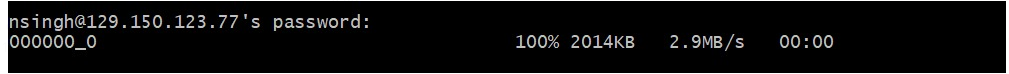


**Figure 11: Displaying the output file**

Replace RED TEXT with each query that you need to download a table.

After the above query is done for each table, open a new terminal, and log in using secure copy protocol query with the help of a terminal console window like putty, ubuntu or mac terminal.

Scp [nsingh@129.150.123.77](mailto:nsingh@129.150.123.77):/home/nsingh/000000\_0 complaints.csv



**Figure 12: Download the CSV files**

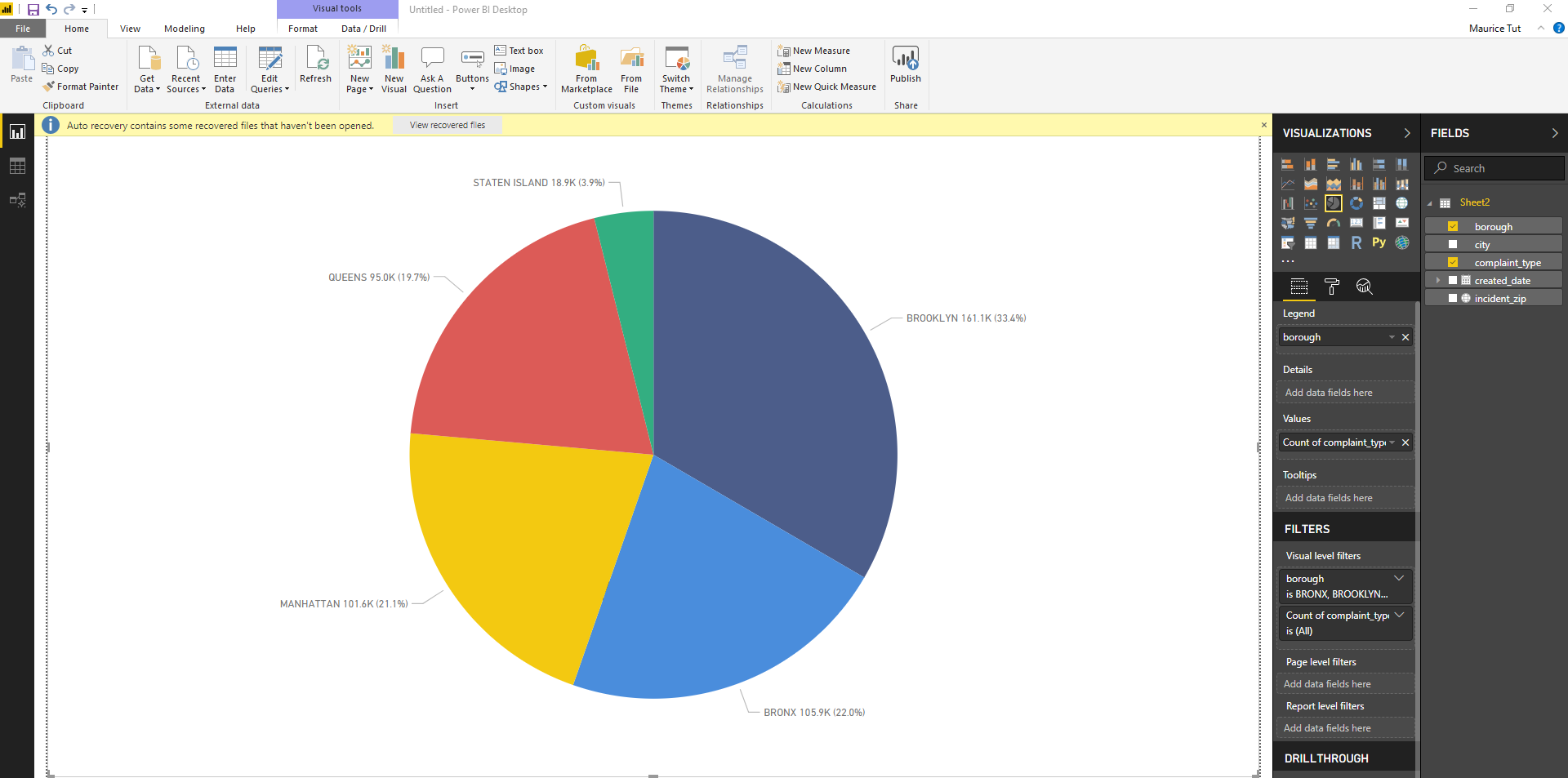
The RED TEXT shows the downloaded data file.

Once the csv files are downloaded, you need to open Power BI on your local computer [7]. Open data file directly from Power BI and Visualise it.

Open Power BI, and open the file according to the following order.

**1.The Boroughs and their count of compliant in numbers and percentage**

File 🡪 Open Excel/CSV 🡪Ny\_311\_Complaint\_Data**.CSV 🡪** OK

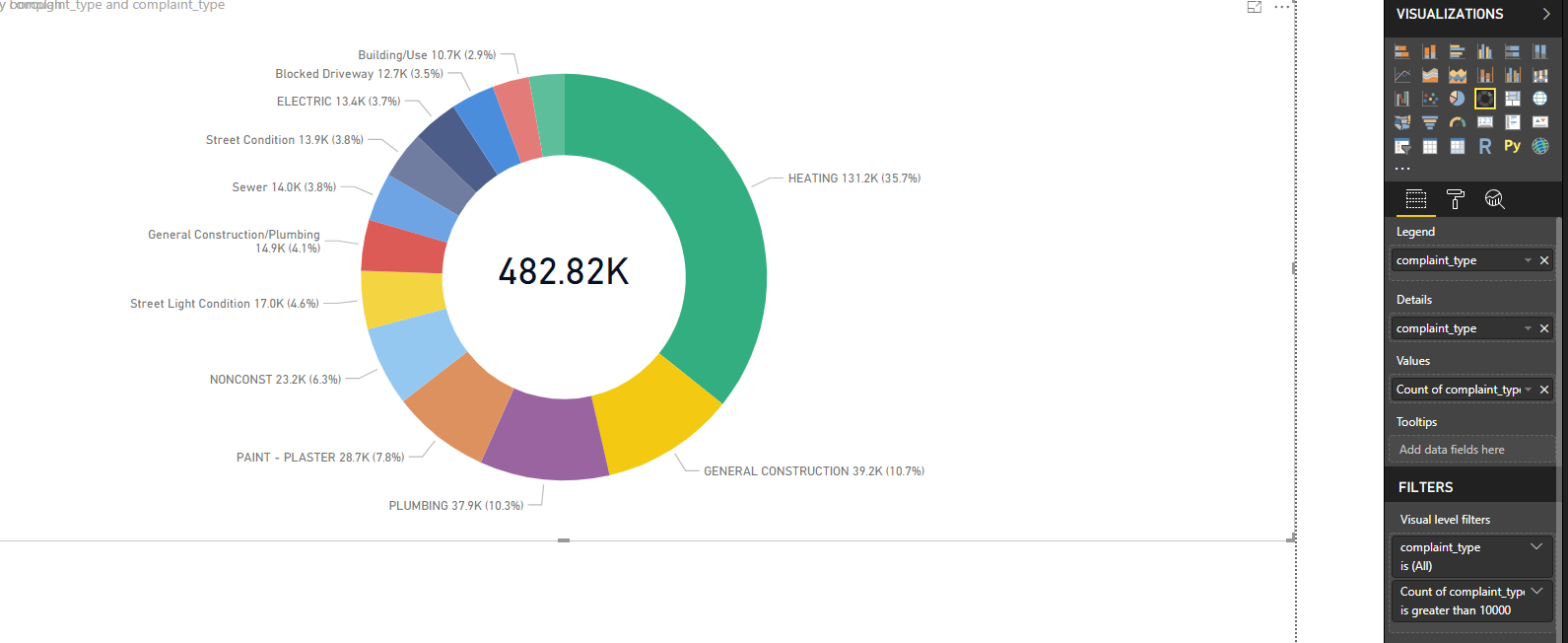


**Figure 13: Boroughs and their compalints in numbers and percentage**

The above visualization shows the boroughs and the corresponding complaints in number and percentage and amongst them BROOKLYN has registered the maximum number of complaints 161.1K(33.4%) and STATEN ISLAND has registered the least number of complaints 18.9K(3.9%).

**2. Percentage of complaint types compared to the grand total**

File 🡪 Open Excel/CSV 🡪Ny\_311\_Complaint\_Data.CSV **🡪** OK.

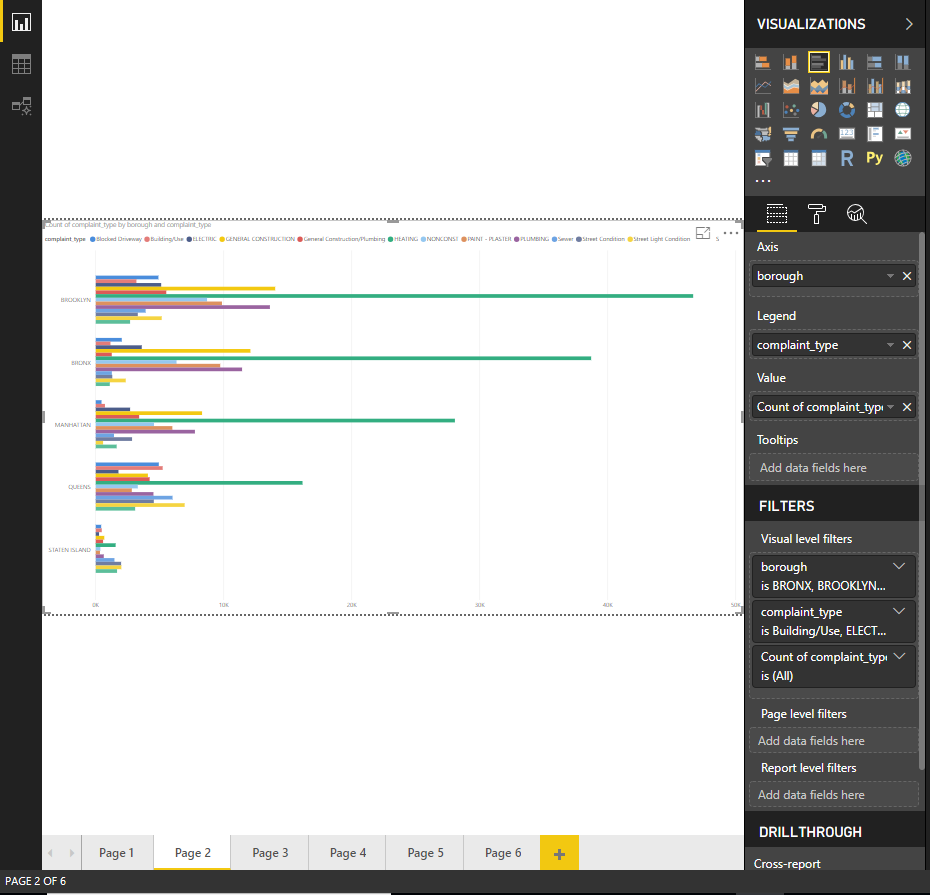


**Figure 14: Complaint types and their numbers and percentage**

The above visualization shows the complaint types and their corresponding numbers and percentage and amongst them Heating complaint has been registered maximum number of times 131.2K (35.7%) and Water System problem has been registered the least number of times 10.2K (2.8%).

**3.Cluster Bar Chart showing the top 13 complaint types and their count by boroughs:**

File🡪OpenExcel/CSV 🡪 Ny\_311\_Complaint\_Data.CSV **🡪** OK.

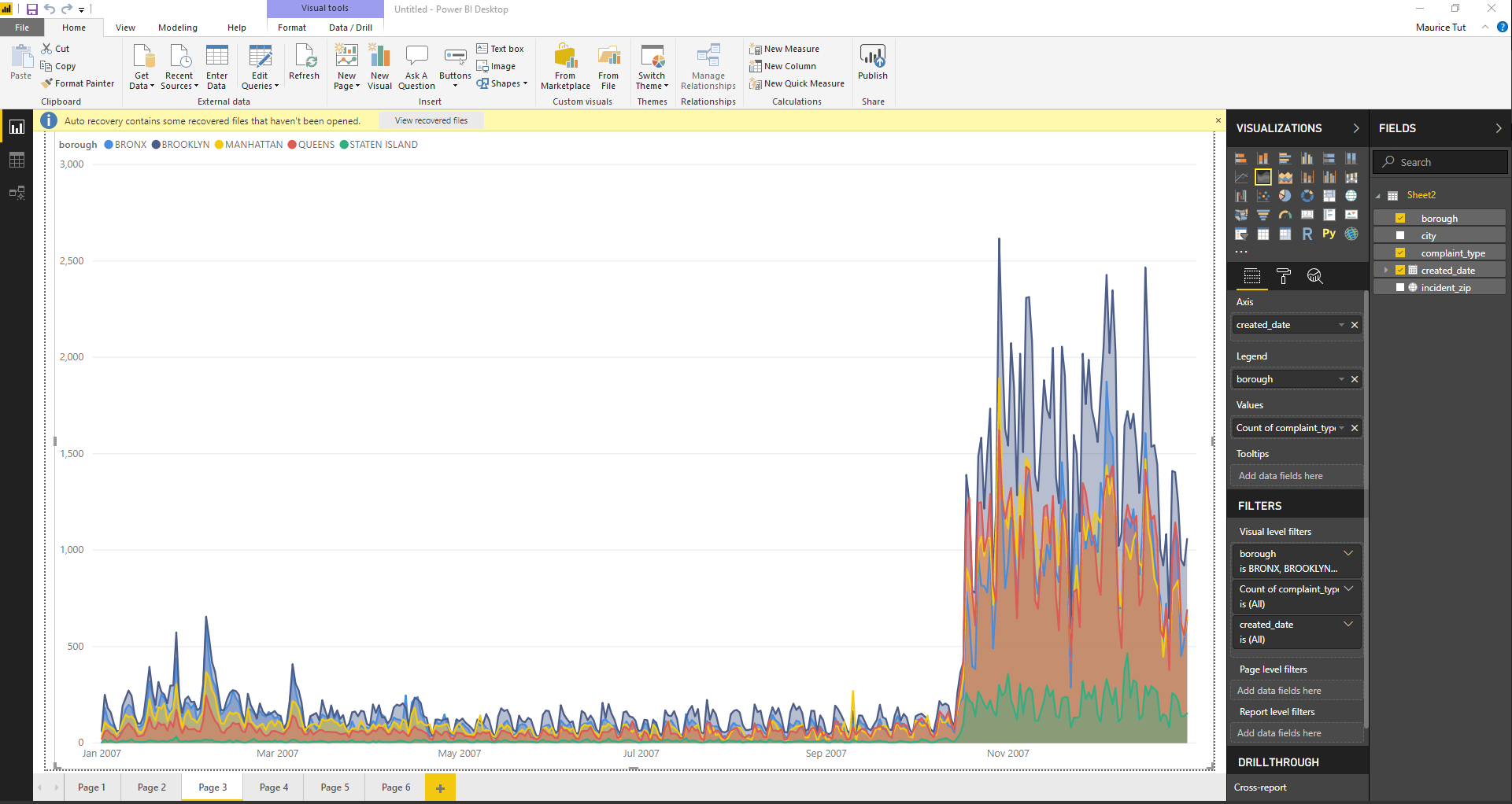


**Figure 15: 13 Complaint types and their count wrt to borough.**

The above Cluster bar chart shows the top 13 complaints and their count w.r.t boroughs. BROOKLYN registered the maximum number of Electric, General Construction, General Construction/Plumbing, Heating, Non-const., Paint-Plaster, Plumbing. BRONYX registered the least number of Street condition and water system complaints. MANHATAN registered least number of Street light condition problems. QUEENS registered maximum number of Blocked Driveway, Building use and water system complaints. STATE ISLAND registered least number of Blocked Driveway, building use, Electric, General construction, General Construction/Plumbing, Heating, Non-const., Paint-Plaster, Plumbing and sewer complaints.

**4.Complaints per day for each borough**

File🡪Open Excel/CSV🡪 Ny\_311\_Complaint\_Data.CSV 🡪OK.

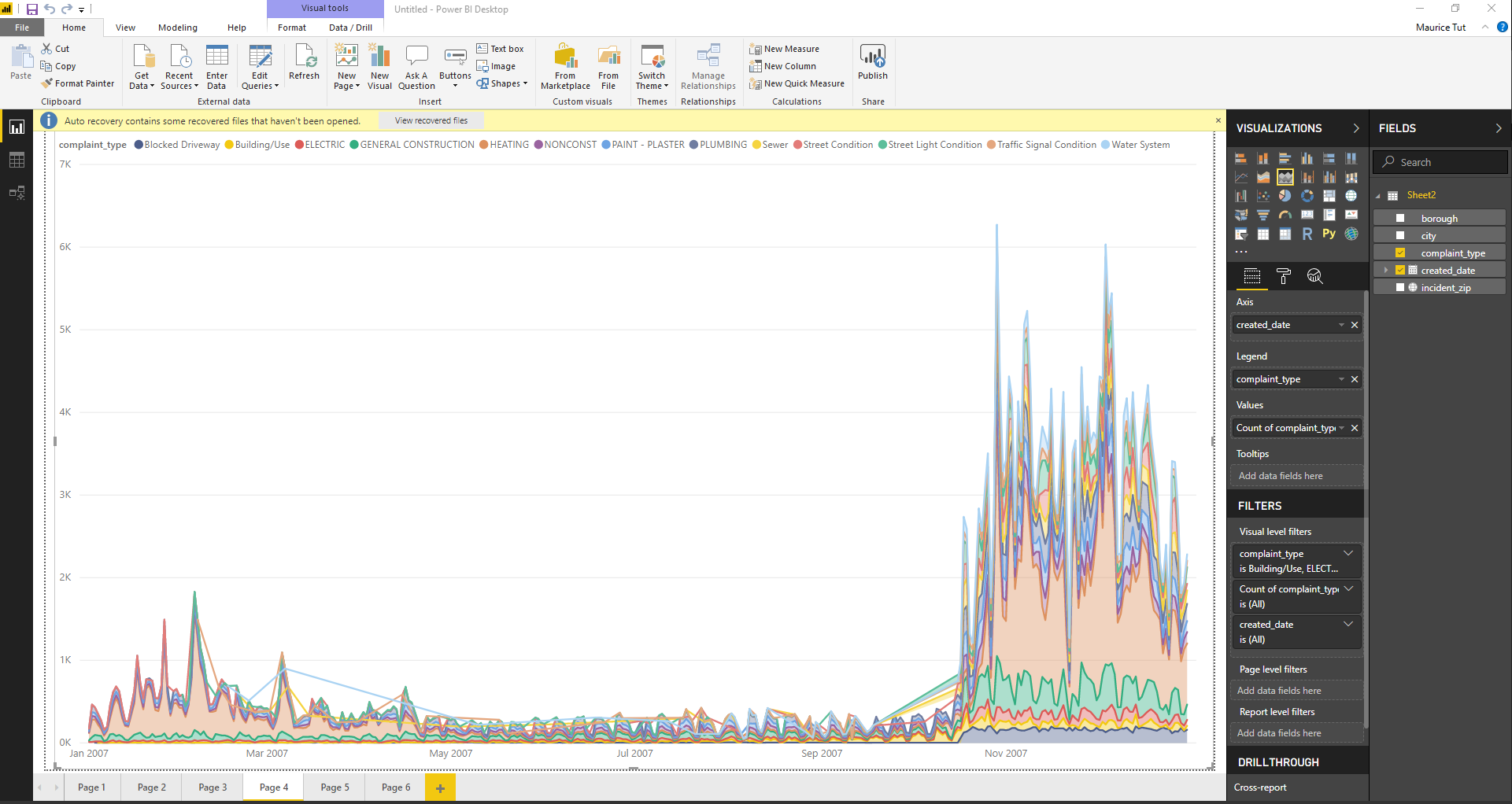


**Figure 16: Complaints per day by each borough**

The above visualization shows the complains registered on daily basis by each borough. Maximum complaints per day are registered in the months of November and December and minimum are registered from March to April.

**4.Top 13 Complaints per day.**

File🡪Open Excel/CSV🡪Ny\_311\_Complaint\_Data.CSV 🡪OK.

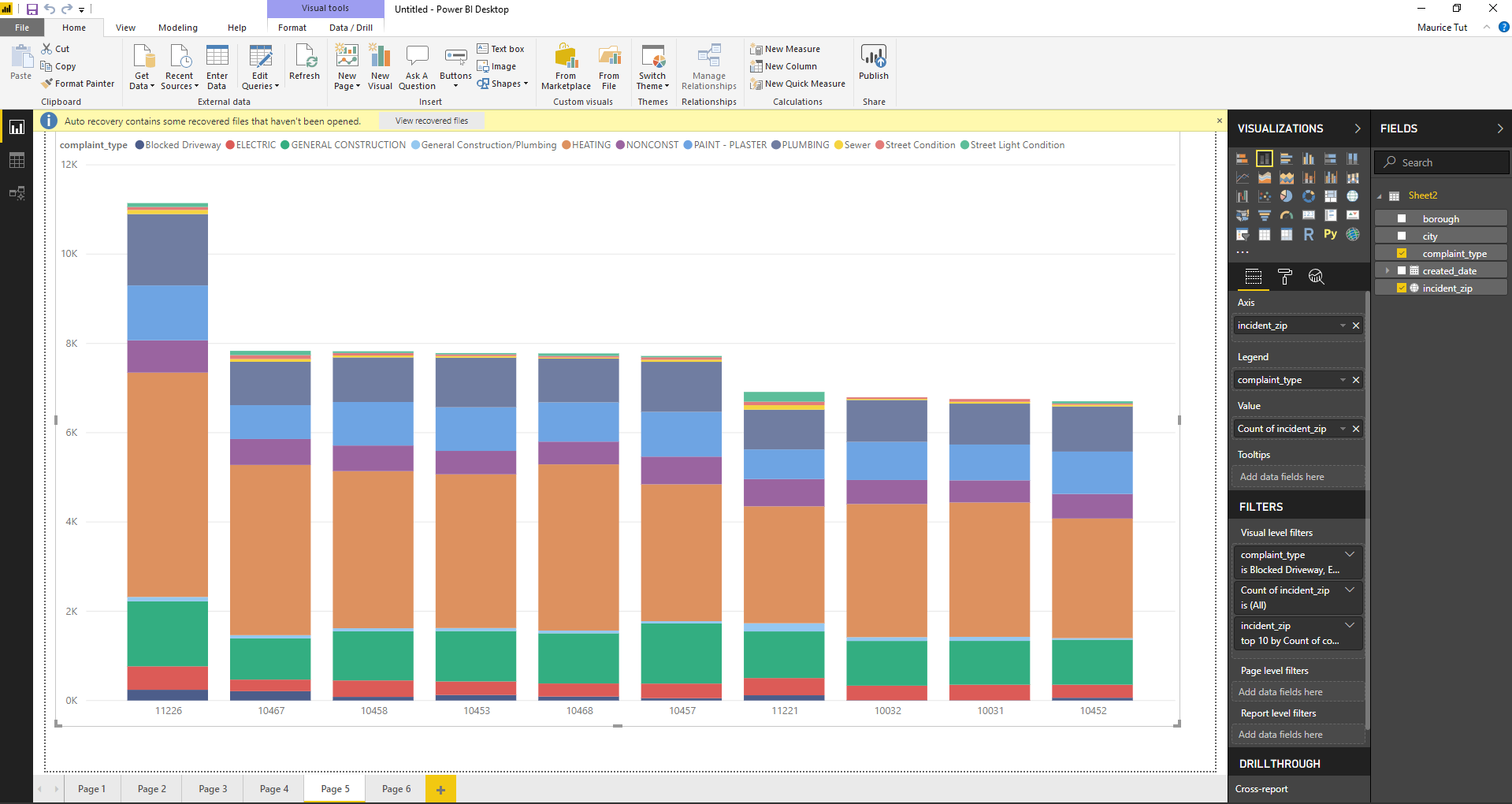


**Figure 17: Top 13 Complaints by count in 2007**

The above visualization shows the count of complaints per day for top 13 complaints in year 2007.

**5. Top** **11 complaint types based on 10 zip codes.**

File🡪Open Excel/CSV🡪 Ny\_311\_Complaint\_Data.CSV 🡪OK.

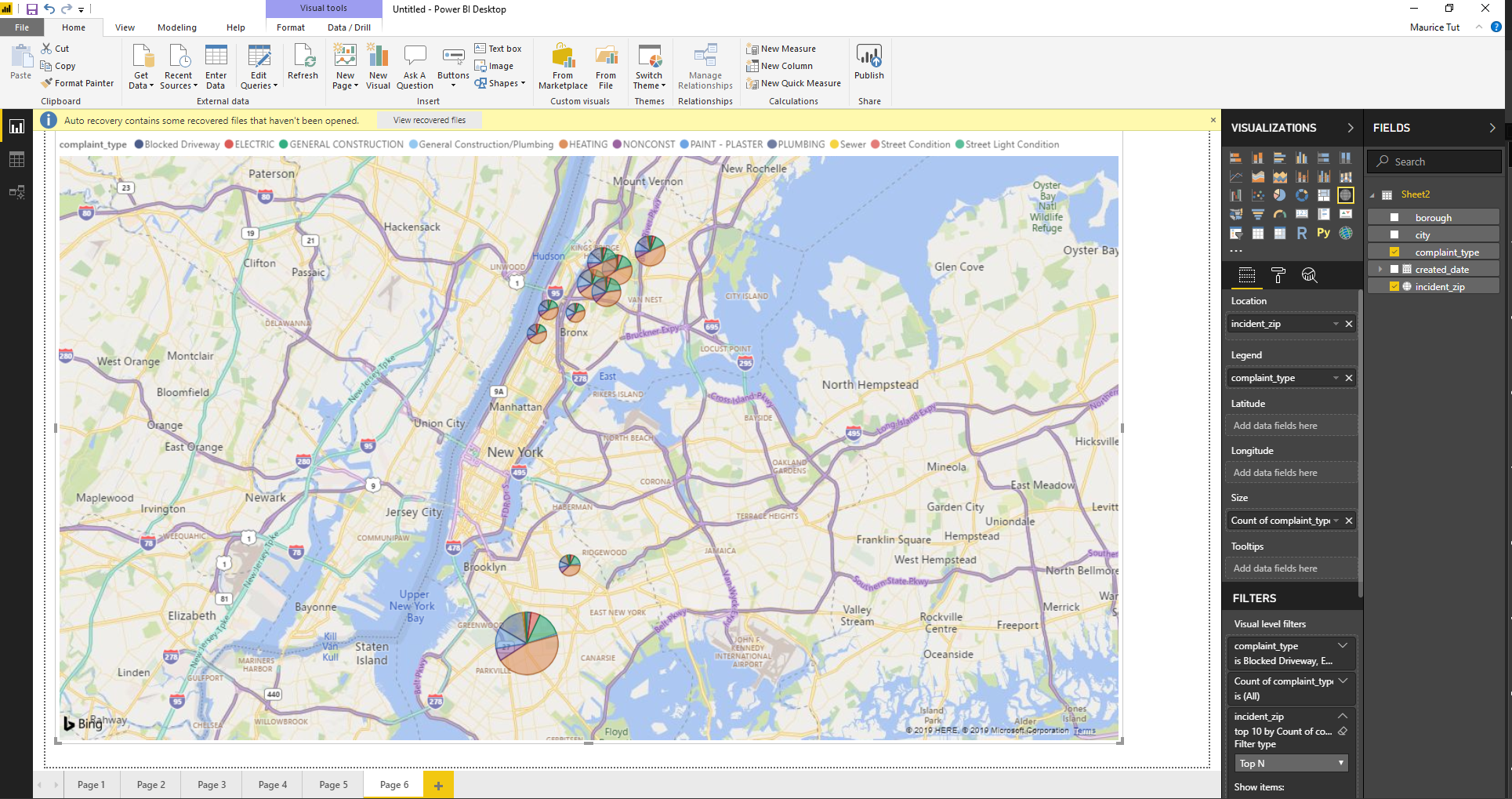


**Figure 18: Top 11 Complaints based on zip codes**

The above visualization shows top 11 complaints based on 10 zip codes. Zip code 11226 registered maximum number of blocked-delivery, electric, general construction, heating, non-construction, Paint-plaster and plumbing complaints.11221 registered maximum number of General construction/plumbing, sewer and Street Light Condition complaints and 10467 registered maximum number of Street condition complaints.

**6. Map showing top 11 complaint types based on 10 zip codes.**

File🡪Open Excel/CSV🡪 Ny\_311\_Complaint\_Data.CSV 🡪OK.

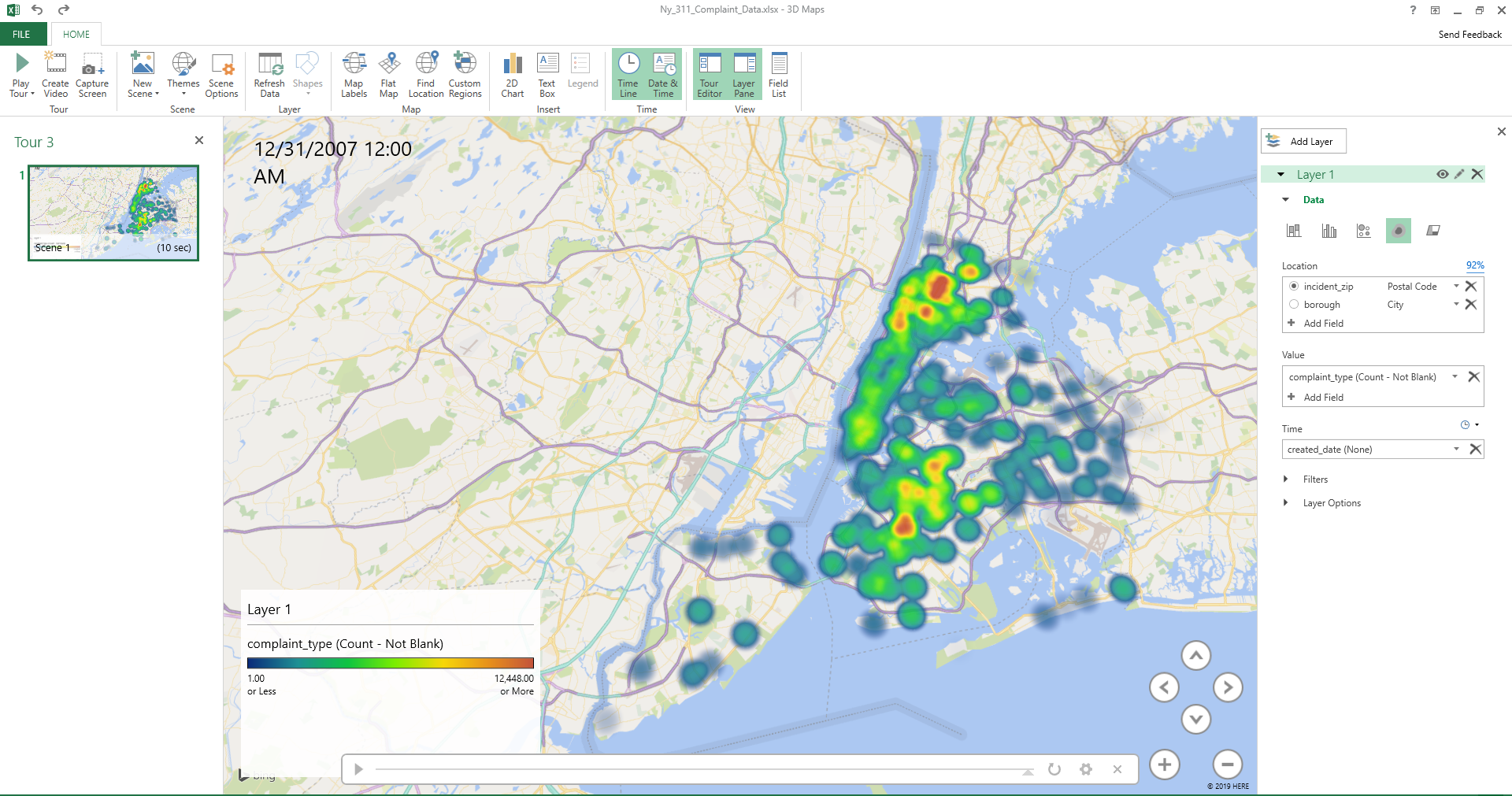


**Figure 19: Top 11 Complaints based on zip codes (Map view).**

The above map shows zip code 11226 registered maximum number of blocked-delivery, electric, general construction, heating, non-construction, Paint-plaster and plumbing complaints. Zip code11221 registered maximum number of General construction/plumbing, sewer and Street Light Condition complaints and zip code 10467 registered maximum number of Street condition complaints.

**7.Heat Map showing complaints for each zip code.**

Open Excel🡪 Get data🡪 Ny\_311\_Complaint\_Data.CSV🡪3D Map🡪Tour1🡪Play



**Figure 20: Top 11 Complaints based on zip codes and date (Heat Map).**

**3. Conclusion**

From the analysis, we can get a better understanding of data with the interesting set of trends and patterns in large data sets.

The advancement in cloud technologies helps us to harness the power of parallel processing of a cluster of computers with little investment and almost no maintenance of the underlying computer hardware.

3-1-1 is a special telephone number supported in many communities in United States. The number provides access to non-emergency municipal services.

Heating, General construction, Plumbing, Sewer etc. are some of the major complaints registered in 2007.

The above complains are mostly registered in BROOKLYN, BRONX, MANHATTAN and QUEENS.

Highest 161.1K complains are registered in BROOKLYN and lowest 18.9 K complains are registered in STATEN ISLAND.

Around 482.2K complains are registered in 2007.

Among them, maximum 131.2K (35.7%) are registered for heating and minimum 10.2 K (2.8%) are registered for water system.

**References**

[1] NYC311 dataset for 2007:

[https://data.cityofnewyork.us/api/views/aiww- p3af/rows.csv?accessType=DOWNLOAD](https://data.cityofnewyork.us/api/views/aiww-%20p3af/rows.csv?accessType=DOWNLOAD)

[2] NYC311 dataset for 2008:

<https://data.cityofnewyork.us/api/views/uzcy-9puk/rows.csv?accessType=DOWNLOAD>

[3] NYC311 dataset for 2009:

<https://data.cityofnewyork.us/api/views/3rfa-3xsf/rows.csv?accessType=DOWNLOAD>

[4] Apache Hive http://hive.apache.org/

[5] Putty <https://the.earth.li/~sgtatham/putty/latest/w64/putty.exe>

[6] Oracle Big Data cloud.

<https://www.oracle.com/big-data/big-data-cloud-service/>

[7] CIS 5200 Lab – Hive Sensor Data Analysis https://www.dropbox.com/home/cis4560\_5200Fall2018/labs?preview=hiveSensorDataAnalysisLab\_oracle.pdf

[8] Power BI <https://app.powerbi.com>

[9] GIThub <https://github.com/lab1213/cis5200>