NYC 311 Service Request Pattern of May 2019 vs. May 2020

Wu, Bi (Vicky)

Data mining final project

CSc 84040 by Prof. Bon Sy

Introduction

311 is a non-emergency phone number for people to call in to find information about services, make complaints, or report problems like noise, graffiti or public infrastructure damages. According to Wired, 311 fields on average more than 50,000 calls a day, offering information about more than 3,600 topics. (Johnson, 2010)

The NYC OpenData website provides a complete record of all complaints reported to 311 from 2010 to present. During the Covid-19 pandemic, 311 also takes calls from the public reporting businesses or facilities violating social distancing or face covering guidelines to help prevent the spread of the novel coronavirus. The purpose of the study presented in this paper is to examine the possible impact of the COVID-19 pandemic on 311 complaints in New York City by comparing the 311 complaints trends and patterns within the time period of a month before and during the pandemic.

Data and Variable Selection

Three datasets were used in this study:

1) The NYC 311 service request data of May 2019 from NYC OpenData. The original dataset has 237224 rows and 41 variables. After variable selection and removing rows with missing values, the dataset is reduced to 198594 rows and 8 columns. A snapshot of the data is given below:

	unique_key	created_date	closed_date	complaint_type	incident_zip	borough	location	resolution_description
1	42526476	2019-05-01T12:00:07.000	2019-05-02T00:00:00.000	Electrical	10032	MANHATTAN	{'latitude': '40.8347374', 'longitude': '-73.9...	The Department of Buildings investigated this ...
2	42527495	2019-05-01T12:00:07.000	2019-07-13T00:00:00.000	Building/Use	10011	MANHATTAN	{'latitude': '40.7444065', 'longitude': '-73.9...	The Department of Buildings investigated this ...
3	42526094	2019-05-01T12:00:14.000	2019-05-08T11:26:00.000	Sidewalk Condition	10312	STATEN ISLAND	{'latitude': '40.531203', 'longitude': '-74.16...	The Department of Transportation is unable to ...

2) 2) The NYC 311 service request data of May 2020 from NYC OpenData. The original dataset has 198679 rows and 41 variables. After variable selection and removing rows with missing values, the dataset is reduced to 184741 rows and 8 columns. A snapshot of the data is given below:

	unique_key	created_date	closed_date	complaint_type	incident_zip	borough	location	resolution_description
0	46099985	2020-05-01T12:00:00.000	2020-04-30T11:30:00.000	Street Light Condition	10461.0	BRONX	{'latitude': '40.8516867', 'longitude': '-73.8...	Service Request status for this request is ava...
1	46105093	2020-05-01T12:00:00.000	2020-04-28T12:00:00.000	Street Light Condition	11694.0	QUEENS	{'latitude': '40.5843873', 'longitude': '-73.8...	Service Request status for this request is ava...
2	46105927	2020-05-01T12:00:00.000	2020-05-01T12:00:00.000	Street Light Condition	10461.0	BRONX	{'latitude': '40.8450793', 'longitude': '-73.8...	Service Request status for this request is ava...

3) The NYC coronavirus data by zip code of May 2020 from NYCHealth coronavirus data. The dataset has 177 rows and 9 columns. A snapshot of the data is shown below:

	MODIFIED_ZCTA	NEIGHBORHOOD_NAME	BOROUGH_GROUP	COVID_CASE_COUNT	COVID_CASE_RATE	POP_DENOMINATOR	COVID_DEATH_COUNT	COVID_DEATH_RATE	PERCENT_POSITIVE
0	10001	Chelsea - Clinton	Manhattan	361	1532.06	23563.03	21	89.12	17.56
1	10002	Union Square - Lower East Side	Manhattan	1035	1348.44	76755.41	145	188.91	24.05
2	10003	Union Square - Lower East Side	Manhattan	445	827.11	53801.62	32	59.48	14.34

The dataset is later joined with the May 2020 311 dataset by zip code to compare covid-19 case counts and 311 complaint counts. A snapshot of the joined dataset is given below:

	MODIFIED_ZCTA	COVID_CASE_COUNT	Latitude	Longitude	number_of_complaints
0	10001	361	40.750742	-73.99653	721
1	10002	1035	40.717040	-73.98700	1883
2	10003	445	40.732509	-73.98935	1102

Research Questions

1) How does the COVID-19 pandemic affect the number and types of 311 complaints in NYC in May 2020 compared to May 2019?
2) How does the COVID-19 pandemic affect the pattern of residential noise complaints versus commercial noise complaints in May 2020 and May 2019?
3) Is there a relationship between reported coronavirus cases and 311 complaints in May 2020?
4) Can we predict the number of days it takes to resolve a 311 complaint by using the information in the resolution description?

Methodology and Results

1) Exploratory Data Analysis

Exploratory data analysis was performed on the 311 datasets to discover similarities and differences between the May 2019 311 complaints and May 2020 311 complaints. In total, there are 19.4% more complaints in 2019 than 2020. When compared by borough, Bronx and Manhattan saw an increase in number of complaints in 2020 with the rest of the neighborhood seeing a drop in the number of complaints in 2020.

BRONX	34653
BROOKLYN	64113
MANHATTAN	38101
QUEENS	50336
STATEN ISLAND	11045
Unspecified	346

Number of complaints by borough 2019

BRONX	45232
BROOKLYN	52458
MANHATTAN	39279
QUEENS	40455
STATEN ISLAND	7305
Unspecified	12

Number of complaints by borough 2020

When looking at the number of complaints by day in May 2019 and May 2020, we can see a surge in the complaints made in the beginning, middle and end of the month in 2020 while in 2019 the number of complaints stayed relatively flat over the month. (Figure 1)

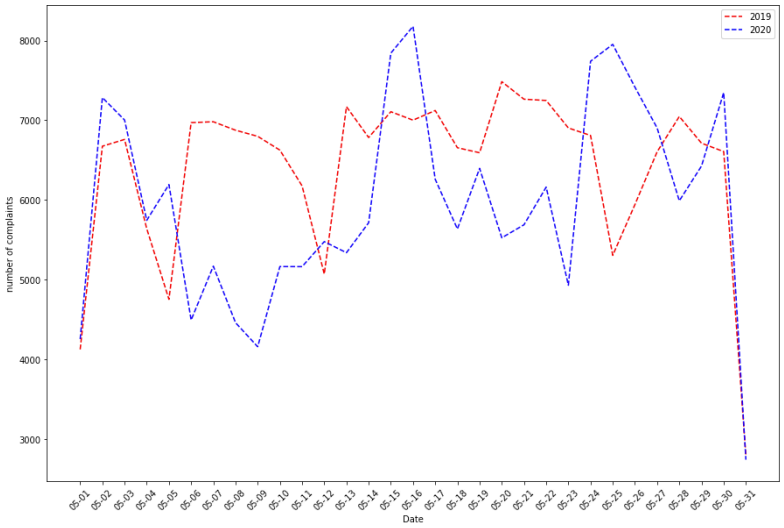

Figure 1: Number of complaints by day

Next, I looked at the top 20 types of complaints by count made in May 2019 and 2020 respectively. Notice that the "Noise – Residential" complaints stayed on the top of the chart in both years, while the number almost doubled in 2020. Correspondingly, the number of "Noise – Commercial" complaints in 2020 seemed to have decreased to its half in 2019. The assumption is that the Covid-19 pandemic forced a lot of businesses to close down in 2020, thus producing less commercial noises. However, people are more likely to stay and gather indoor, thus producing more residential noises. (Figure 2)

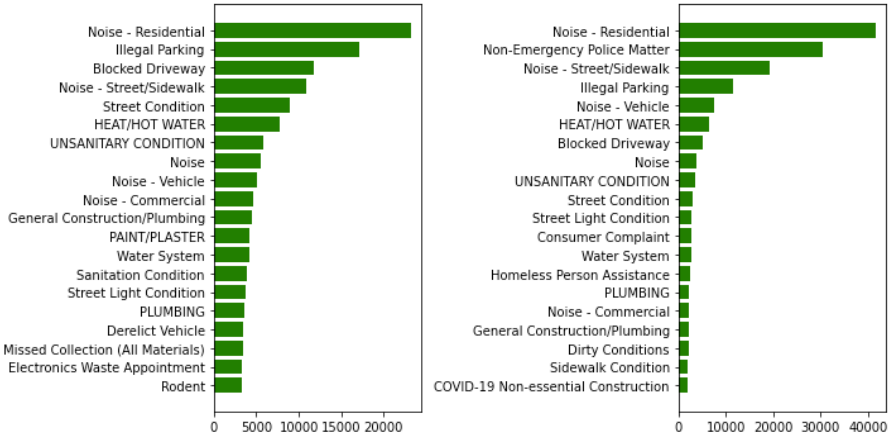

Figure 2: Top 20 types of complaints of May 2019 (left) and May 2020 (right)

2) KMeans Clustering

To compare the noise complaints geographically in May 2019 and May 2020, I used KMeans clustering to cluster both types of complaints based on their longitudes and latitudes on the map. (Figure 3) The light blue dots represent all the complaints made in the month. The red circles are the 20 cluster centroids of residential noise complaints and the black circles are the 20 cluster centroids of commercial noise complaints. The circle sizes correspond to the number of complaints that are clustered into each cluster centroid. The larger the circle, the more complaints the cluster centroid is representing.

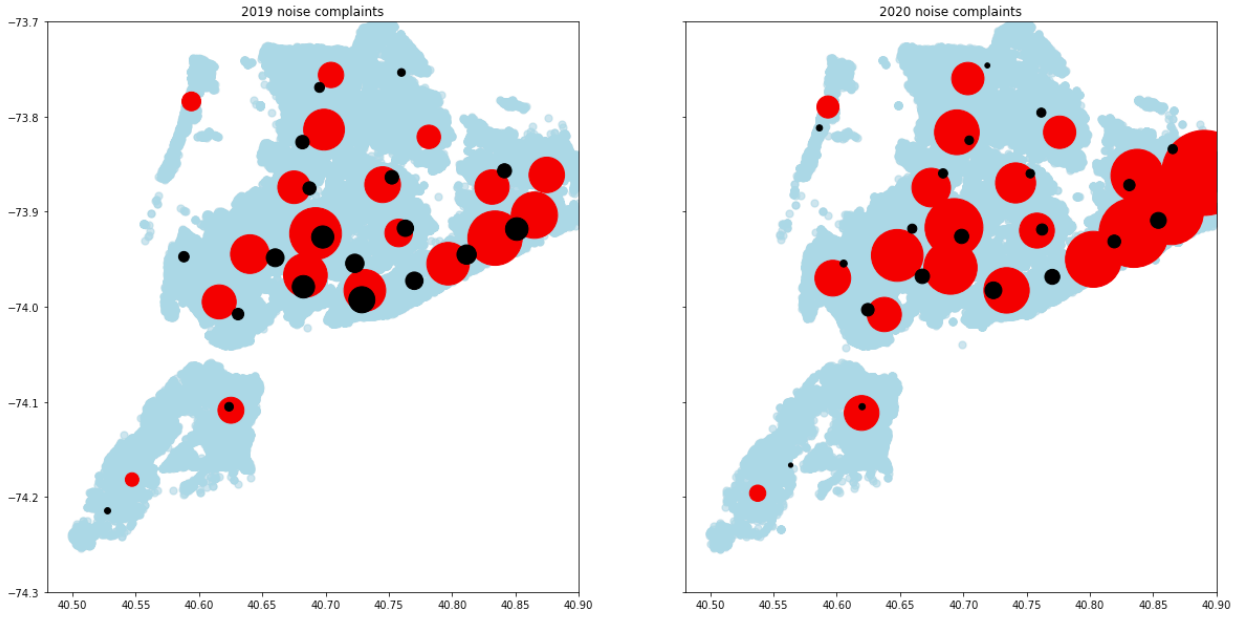

Figure 4: residential noise complaints vs. commercial noise complaints of May 2019 (left) and May 2020 (right)

As seen from the two figures, residential complaints increased in 2020 as the size of the red circles expanded. Commercial complaints decreased in 2020 as the size of the black circles shrank. Geographically speaking, residential noise complaints increased dramatically in the area of center Queens and Brooklyn as the red circles become bigger in those areas. Commercial noise complaints appeared to be decreasing in Brooklyn in 2020. However, a cluster centroid of commercial noise complaints appeared in Manhattan in 2020.

Using the same approach, I clustered the Covid-19 cases of May 2020 and the entire 311 complaints of May 2020 on the same map. The yellow circles represent the covid-19 clusters while the red circles represent the 311 complaint clusters. (Figure 5)

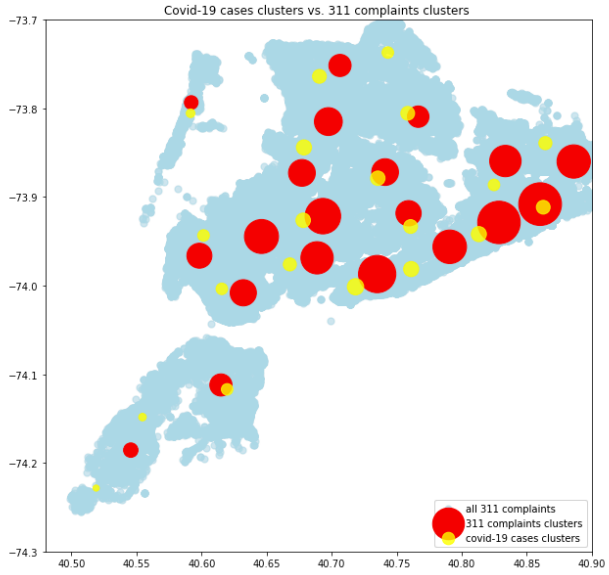

Figure 5

The covid-19 cases clusters and 311 complaints clusters are surprisingly close to each other with high density in the Queens and Brooklyn areas. To examine if there is a correlation between covid-19 case counts by zip code and of 311 complaint counts by zip code, I calculated the covariance between the two variables, which is 400856. The number is positive, suggesting the variables change in the same direction. In addition, the Spearman's correlation coefficient between the two variables is 0.708, suggesting a monotonic relationship between the two variables

```
[[607912.77234071 400856.66978938]
 [400856.66978938 711486.40369411]]
```

Figure 6 covariance between 311 complaints and covid-19 cases

3) Multinomial Naïve Bayes

Multinomial Naïve Bayes is used to predict the number of days it takes to close a 311 complaint by looking at its resolution description. The problem was translated into a classification problem using two feature variables: "days_to_resolve" and "resolution_description".

- Days_to_resolve: categorical variable of 10 levels. It is computed by subtracting "created_date" from "closed_date" and quantized to 10 equal-sized levels. Each level represents a class and is denoted by integers [0, 1, 2, … , 9]
- Resolution_description: String variable describing the resolution to each 311 complaint.

To be able to use multinomial naïve bayes model for the classification, each "resolution_description" is transformed into a bag of words and vectorized using the word count. Punctuation, numbers, and word with less semantic meanings were removed from the descriptions and grammar, word order is disregarded. Vectorized description each has 382 features, with each feature representing a word. (Figure 6) In order for the classification to run on memory, 1000 entries were randomly selected from the May 2019 dataset with 70% of it used for training and 30% used for testing. The multinomial naïve bayes model from the scikit-learn package was used to train the data and an 88.7% test accuracy was obtained.

	able	accept	acceptable	access	accord	act	action	additional	address	adequate	adjoin	administrative	administratively	advise
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0	0	0	0	0	0	0
4	0	0	0	0	0	0	1	0	0	0	0	0	0	0

Figure 7 Bag of words

Limitations

This study's conclusions have a strong regional and periodical limitation, given the data used in the study is solely representing NYC in May 2019 and May 2020.

There is a statistical limitation from the COVID-19 case count data, due it is preliminary and subject to changes. (NYC Health, 2020)

The 311 calls dataset does not represent a full picture of the service, in part because of operational and system complexities associated with remote talk talking necessitated by the unprecedented volume of calls handled during the COVID-19 pandemic. (NYC OpenData, 2020)

Bibliography

Johnson, S. (2010, 11 01). *What a Hundred Million Calls to 311 Reveal About New York*. Retrieved from WIRED: https://www.wired.com/2010/11/ff_311_new_york/

NYC Health. (2020, May 31). *NYC Coronavirus Disease 2019 (COVID-19) Data*. Retrieved from Github: https://github.com/nychealth/coronavirus-data

NYC OpenData. (2020, 5). *https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/7ahn-ypff*. Retrieved from NYC OpenData: https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/7ahn-ypff