Introduction

The purpose of this project is to find out whether the COVID-19 epidemic has any impact of the number of emergencies in different classifications happening in NYC so that we may find out in which way the epidemic affects people life most in the city.

Dataset we are using is Emergency Notifications data posted from OEM. We may choose data from year 2019 to 2020, check out whether there's fluctuations during the virus breakout.

Data Cleaning

1. Dataset Overview

There are two major dataset we decide to employed. Although both of these datasets updating in realtime, we adopt data until 26/04/2020 for convenience.

1.**OEM Emergency Notifications**(emc) is a dataset includes messages sent with information about emergency events and important City services offered by OEM (https://data.cityofnewyork.us/Public-Safety/OEM-Emergency-Notifications/8vv7-7wx3)

Column name	Description	Type
Record_ID	record_id	object
Date and Time	Date and time that notification was sent	object
NotificationType	Notification type	object
Notification Title	Notification titile	object, blank 1728
Email Body	text of notification	object, null 523

Totally 16872 entries

2. **NYC-Covid** is a dataset includes daily counts of new confirmed cases, hospitalizations, and deaths. https://github.com/nychealth/coronavirus-data/blob/master/case-hosp-death.csv

Column name	Description	Type
DATE_OF_INSERT	Case fount by the date	date
NEW_COVID_CASE_COUNT	Cases are by date of diagnosis	int
HOSPITALIZED_CASE_COUNT	Hospitalizations are by date of admission	int

Column name	Description	Type
DEATH_COUNT	Deaths are by date of death	int

Totally 54 entries without null values

2. Data Quality Discussion

1. data cleaning

During exploring the two above datasets, we found some problems such as missing values (e.g. NULL, N/A, Blank, UNSPECIFIED etc) and duplicate entires and outlier that should be dropped.

OEM has 523 of them have a null value in the email_body column. [blank] occurs 1728 times in the Notification title.

Check date range and find min value is 1900-01-01, which seems like a outlier.

NYC-coivd has no missing values or outlier.

2. data labeling

Since we do not need the whole raw data, it should be remarked and renamed.

For instance, 'Date and Time' is renamed as 'Date' and was kept only the date part because this project do not plan to use time.

The result re-arrange for OEM is below:

New Name	Type	Explanation
Date	datetime	kept only the date part because this project do not plan to use time.
Notification Type	category	one type has occurred at least 16 times
Count	int	count each notification type occurred times by date

'Record ID' and 'Notification Title' are dropped because no significant effect.

Check duplicated entries and keep the first one. And did not find duplicated entry.

Check date range and find min value is 1900-01-01, which seems like a wrong value.

Count the number of occurrences of different notification types. Found there has no wrong value, and the least type has occurred 16 times.

Count the number of occurrences of different notification title. [blank] occurs 1728 times. Convert [blank] to np.NaN

Count the number of occurrences of different notification types. Found there has no wrong value, and the least type has occurred 16 times.

Count the number of occurrences of different notification title. [blank] occurs 1728 times. Convert [blank] to np.NaN

The result re-arrange for NYC-Covid is below:

New Name	Type	Explanation
Date	datetime	reformat the date the same as OEM.date
Newcase	int	count each new covid cases by date

Drop HOSPITALIZED_CASE_COUNT and DEATH_COUNT

Rename DATE_OF_INSERT as Date and rename NEW_COVID_CASE_COUNT as Newcase.

Check the range of date, min is 2020-03-03, max is 2020-04-25.

2. data integration

To do integration, we select the entries after 2020-03-02 and join these two datasets together. Part of the joined table are shown below

	Date	Newcase	notificationtype	count
0	2020-03-03	2	Aerial (Fly-Over)	1
1	2020-03-03	2	Mass Transit Disruption	1
2	2020-03-03	2	Mass Transit Restoration	1
3	2020-03-03	2	Road Closure	5
4	2020-03-03	2	Weather	1
5	2020-03-04	5	Environmental	1
6	2020-03-04	5	Mass Transit Disruption	4
7	2020-03-04	5	Mass Transit Restoration	3
8	2020-03-04	5	Missing Adult	1
9	2020-03-04	5	Public Health	1
10	2020-03-04	5	Road Closure	3
11	2020-03-04	5	Utility	1
12	2020-03-04	5	Weather	2
13	2020-03-05	3	Public Health	1
14	2020-03-05	3	Road Closure	6
15	2020-03-06	7	Mass Transit Disruption	2
16	2020-03-06	7	Public Awareness	2
17	2020-03-06	7	Road Closure	2
18	2020-03-06	7	Utility	1
19	2020-03-06	7	Weather	2
20	2020-03-07	7	Aerial (Fly-Over)	2