

12/20/2021

# Assignment 2 – Data Visualization with Kibana

BDAT 1002 - Data System Architecture

<u>Group 10 Members</u>	<u>Student ID</u>
Amith Heiden	200497933
Karthikeyan Suresh Kumar	200489370
Srilekha Sampath Kumar	200499290

Submitted to : Dr. Saber Amini

# TABLE OF CONTENTS

## Introduction

- ELK Stack.....2
- Background of the Dataset..... 2
- Assignment Objective.....2

## Assignment Approach and Methodology

- My Team..... 3
- Techniques Used.....3

## Answers to Analytical Questions

- Data Visualizations using Kibana.....4

## References.....14

## Appendix.....14

# INTRODUCTION

## ELK Stack

The ELK Stack is the world's most popular log management platform. The ELK Stack was a collection of three open-source products — Elasticsearch, Logstash, and Kibana — all developed, managed and maintained by Elastic.

- Elasticsearch is an open source, full-text search, and analysis engine, based on the Apache Lucene search engine.
- Logstash is a log aggregator that collects data from various input sources, executes different transformations and enhancements and then ships the data to various supported output destinations.
- Kibana is a visualization layer that works on top of Elasticsearch, providing users with the ability to analyze and visualize the data.

## Background of the Dataset

This dataset is taken from [NYC OpenData](#) website and it is associated with remote call taking necessitated by the unprecedented volume 311, handled during the Covid-19 crisis. The main focus of the dataset is all about Service Requests (SR) – When, where, Complaint type, Description, Status of the SR.

## Objective

We have been hired as data analysts by the city of New York to provide valuable insights of their huge data set for 311 service requests. Our task is to work in the ELK stack by installing and configured it in GCP platform.

## Deliverables

- Code for Logstash configuration file
- Geo-point template (for maps) – Screen shot
- Results for the analytical questions (tables, charts, tag clouds, maps, and dashboard) - Screen shots

# ASSIGNMENT APPROACH AND METHODOLOGY

## My Team

Students of Big Data Analytics, Georgian College, Fall 2021 intake



### **Amith Heiden - Data Analyst**

Tasks Performed – Creation of Logstash configuration file, geo-point template, Analytical Questions : 3, 4, 10



### **Karthikeyan suresh kumar – Data Analyst**

Tasks Performed – Creation of Logstash configuration file, geo-point template, Analytical Questions : 1, 2, 6



### **Srilekha Sampath kumar - Data Analyst**

Tasks Performed - Creation of Logstash configuration file, geo-point template, Analytical Question : 5, 7, 8, 9

## Techniques Used

Prepared cloud environment and the below steps are done after setting up ELK on a GCP cluster:

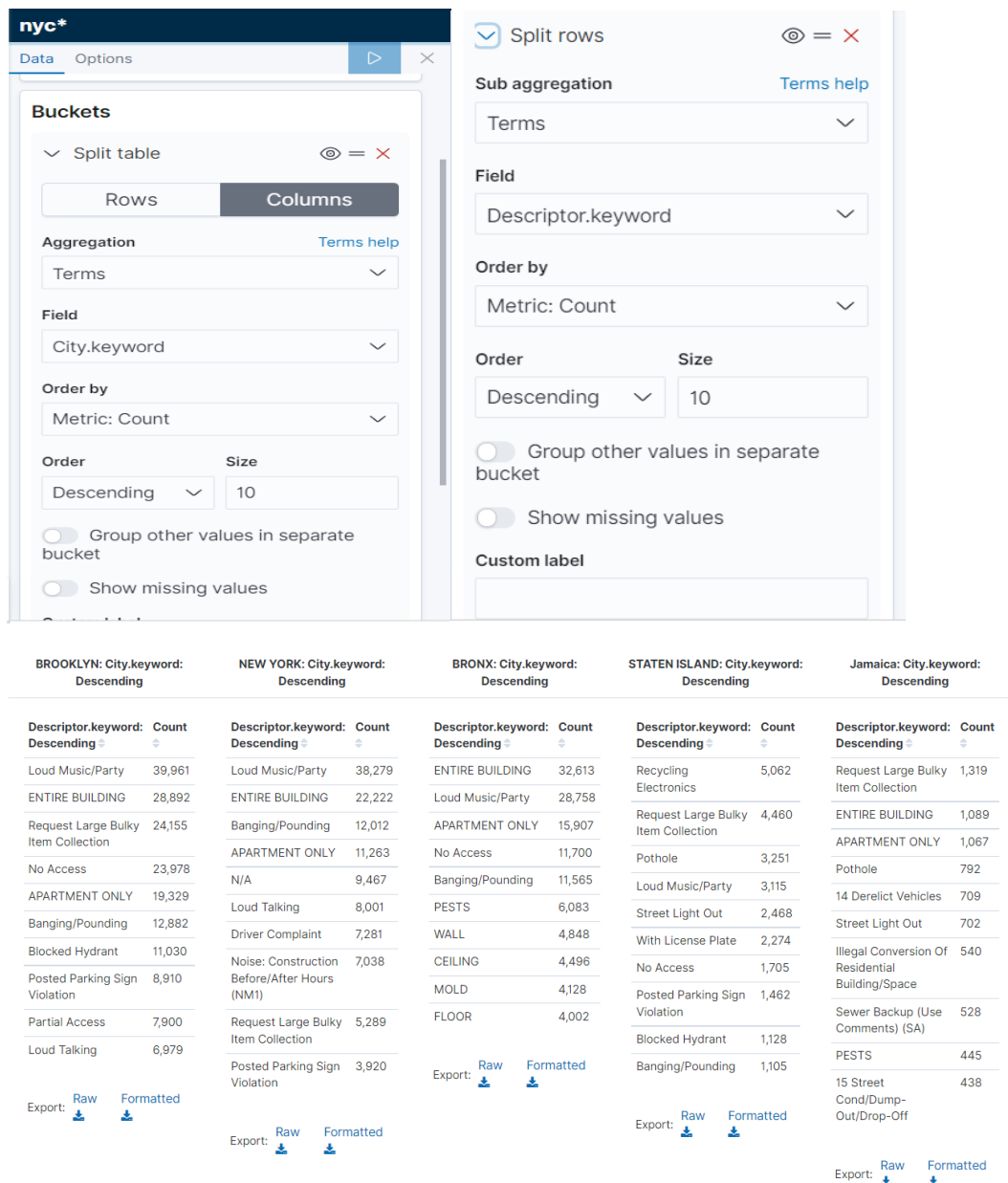
1. Created an index called 'nycinfo' with the sample Logstash configuration file (logstash\_nyc311.config)
2. Created an index pattern called 'nycinfo', which matches the 'nycinfo' index created at step 1. Specify the Time Filter option as "I don't want to use the time filter." This index is used for the analysis of questions 1&2.
3. Created a template called 'geotemplate' to perform analysis of question 4.

# ANSWERS TO ANALYTICAL QUESTIONS

## Data Visualizations are performed using Kibana

1. Create a table showing the top 10 cities with the highest calls alongside the count of top 10 complaint calls (by Descriptor) in each city.

Created the visualizations -data table- in Kibana.



JAMAICA: City.keyword: Descending	FLUSHING: City.keyword: Descending	ASTORIA: City.keyword: Descending	Flushing: City.keyword: Descending	Ridgewood: City.keyword: Descending
Descriptor.keyword: Count Descending	Descriptor.keyword: Count Descending	Descriptor.keyword: Count Descending	Descriptor.keyword: Count Descending	Descriptor.keyword: Count Descending
Loud Music/Party 2,398	No Access 2,068	Loud Music/Party 2,643	Request Large Bulky Item Collection 1,744	Request Large Bulky Item Collection 4,055
No Access 2,321	Loud Music/Party 1,111	No Access 1,922	ENTIRE BUILDING 1,463	ENTIRE BUILDING 707
With License Plate 1,095	Partial Access 1,084	Partial Access 740	Pothole 826	Pothole 363
Partial Access 943	Banging/Pounding 830	Banging/Pounding 707	APARTMENT ONLY 627	APARTMENT ONLY 315
Banging/Pounding 795	With License Plate 605	Loud Talking 585	Street Light Out 447	Street Light Out 293
Posted Parking Sign Violation 433	Blocked Hydrant 593	Posted Parking Sign Violation 574	Illegal Conversion Of Residential Building/Space 437	E15 Illegal Postering 260
Driver Complaint 374	Sidewalk Violation 437	Blocked Hydrant 471	PESTS 245	PESTS 189
Blocked Hydrant 365	Posted Parking Sign Violation 377	Blocked Sidewalk 454	Controller 191	15 Street Cond/Dump-Out/Drop-Off 167
Branch or Limb Has Fallen Down 338	Branch or Limb Has Fallen Down 364	With License Plate 395	GARBAGE/RECYCLING STORAGE 163	2 Bulk-Missed Collection 166
Commercial Overnight Parking 298	Loud Talking 353	For One Address 301	1 Missed Collection 161	Leak (Use Comments) (WA2) 162
Export: <a href="#">Raw</a> <a href="#">Formatted</a>	Export: <a href="#">Raw</a> <a href="#">Formatted</a>	Export: <a href="#">Raw</a> <a href="#">Formatted</a>	Export: <a href="#">Raw</a> <a href="#">Formatted</a>	Export: <a href="#">Raw</a> <a href="#">Formatted</a>

2. Create a pie chart showing the top 5 cities with the highest calls alongside the top five calls (Descriptor) in each city

nyc\*

Data Options

Metrics

> Slice size Count

Buckets

Split slices

Aggregation

Field

Order by

Order

Size

Split slices

Sub aggregation

Field

Order by

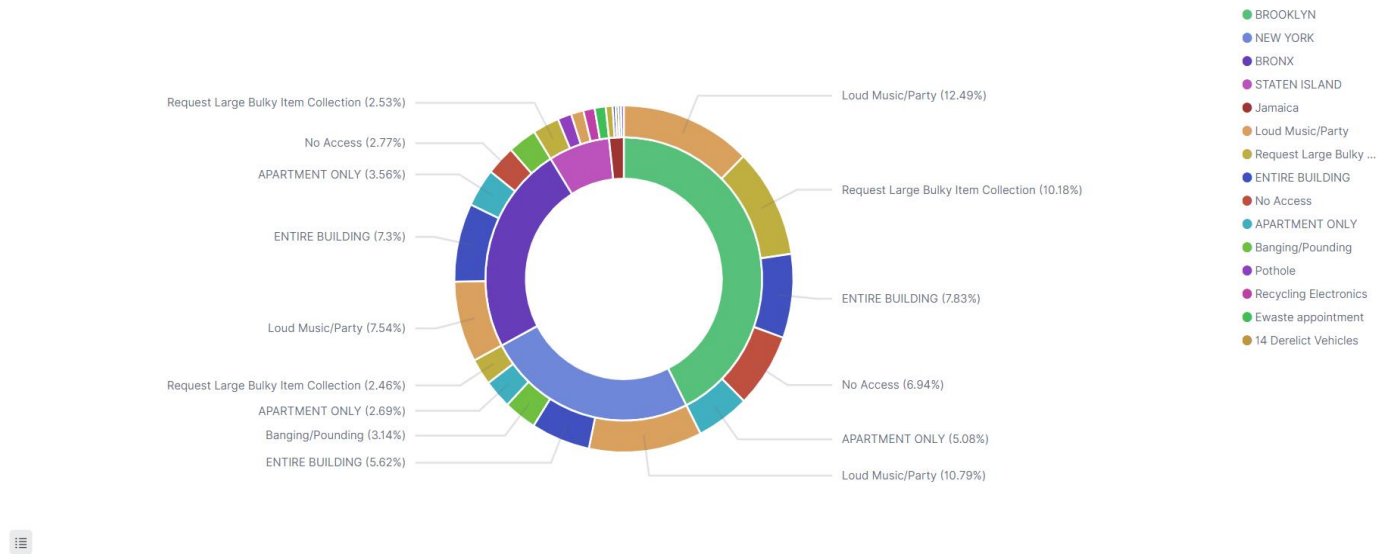
Order

Size

Group other values in separate bucket

Show missing values

Custom label



### 3. Create a tag cloud representing the top 20 call descriptors

**nyc\***

Data Options

Aggregation: Terms

Field: Descriptor.keyword

Order by: Metric: Count

Order: Descending

Size: 20

☐ Group other values in separate bucket

☐ Show missing values

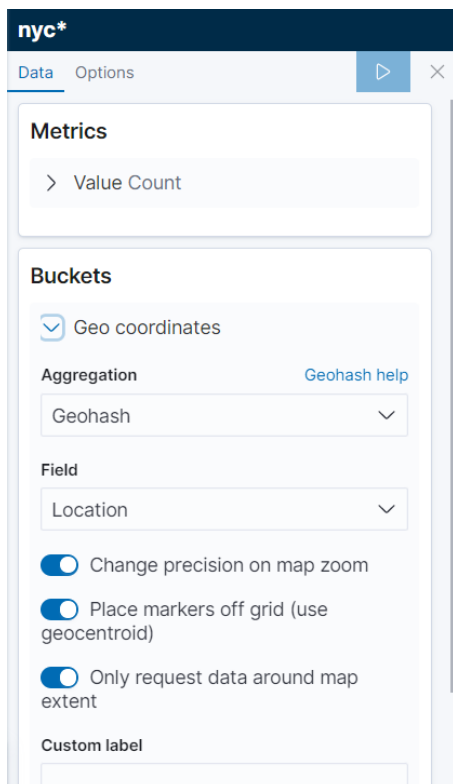
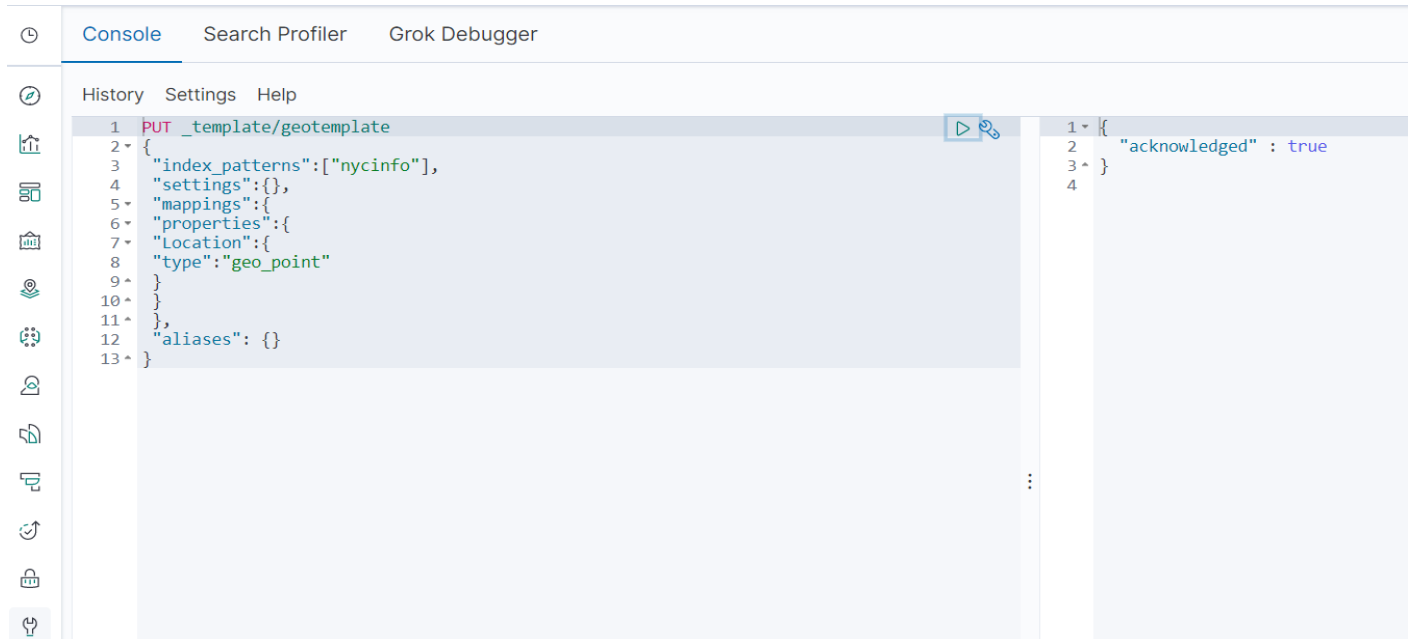
Custom label: Top 20 call Descriptors

> Advanced

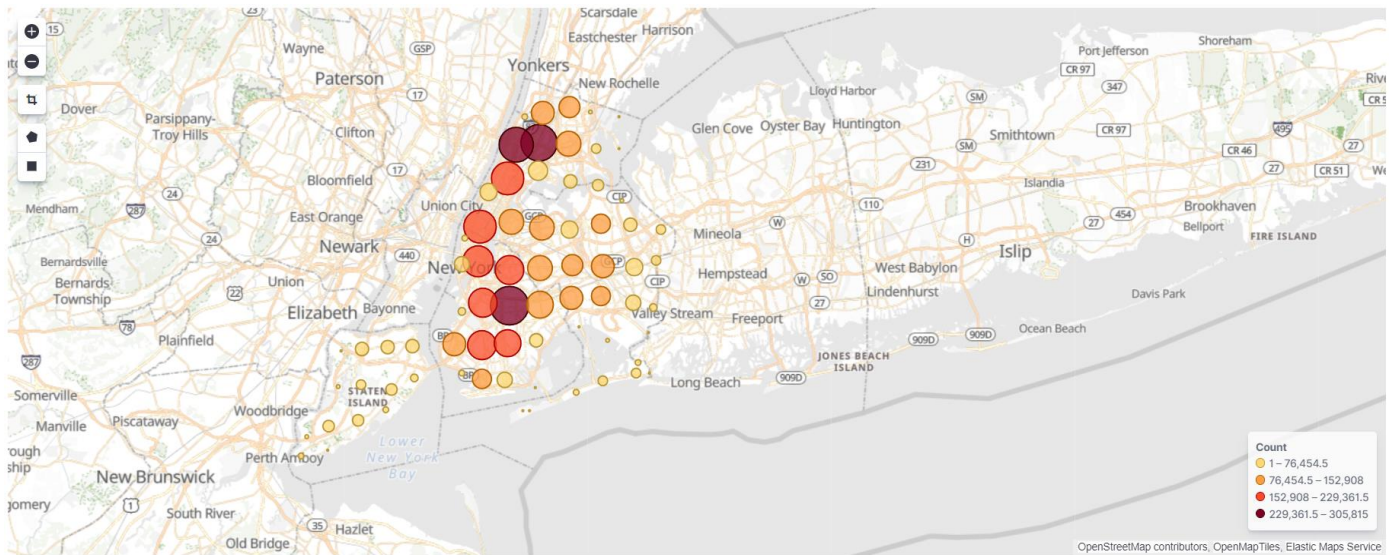


#### 4. Create a coordinated map of all the major call descriptors in each city

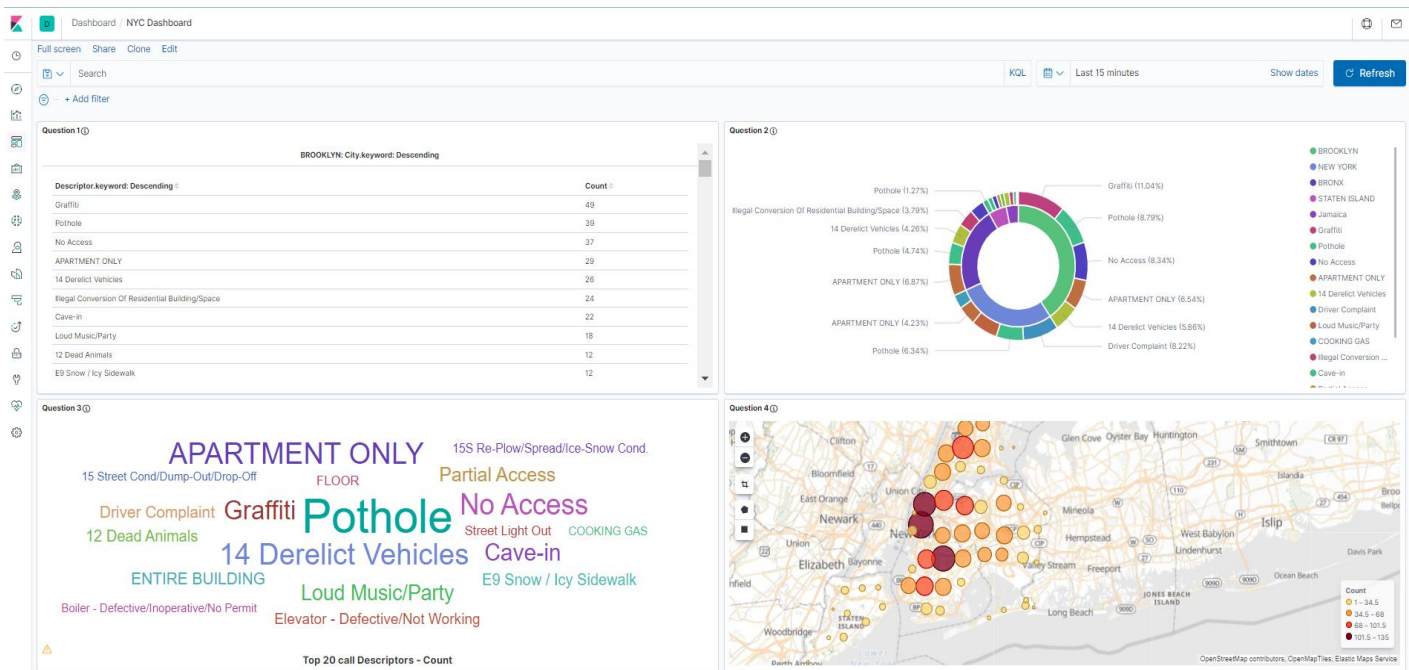
Created a template called 'geotemplate'







5. Create a dashboard for all visualizations of 1 to 4 above



## 6. What are the Top 10 responding city government agency?

**nyc\***

Data Metrics & axes Panel settings

Add

**Buckets**

▼ X-axis

**Aggregation** [Terms help](#)

Terms ▼

**Field**

Agency Name.keyword ▼

**Order by**

Metric: Response ▼

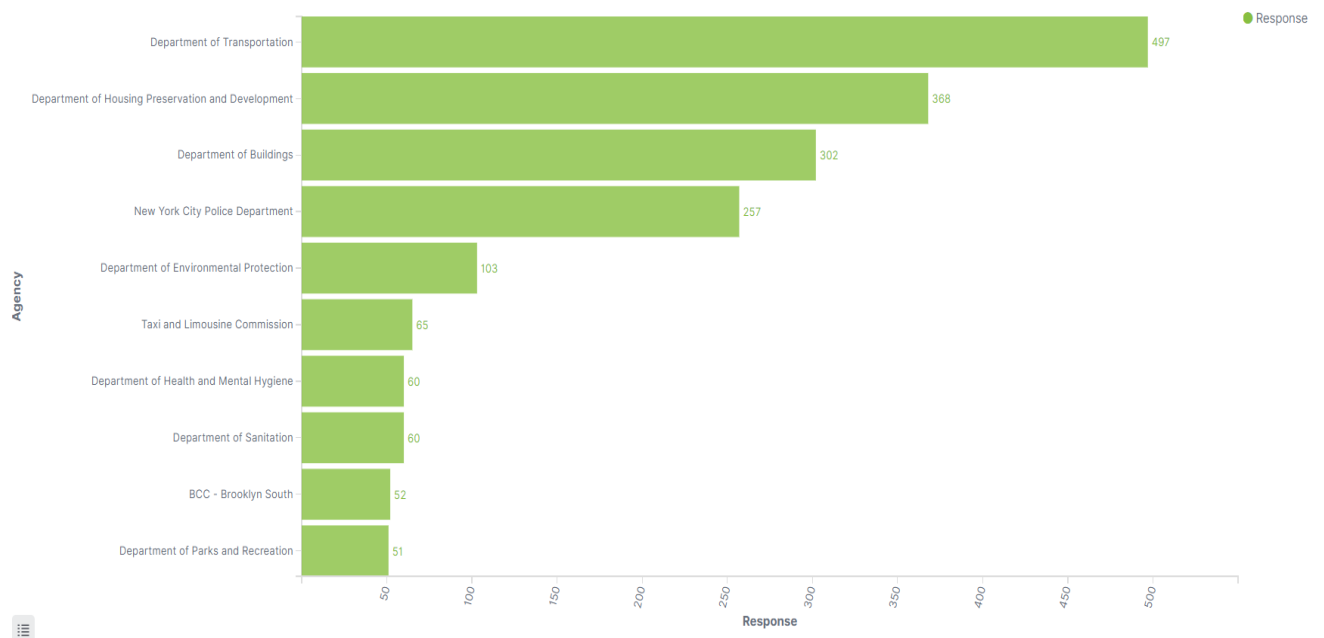
**Order** **Size**

Descending ▼ 15

☐ Group other values in separate bucket

☐ Show missing values

**Custom label**



## 7. What is the Total count of each complaint status of Top 5 cities?

nyc\*

Data Options

▶

×

**Buckets**

Split slices

👁 = ✕

Aggregation

Terms

▼

Field

City.keyword

▼

Order by

Metric: Count

▼

Order

Descending

▼

Size

5

☐

Group other values in separate bucket

☐

Show missing values

Custom label

nyc\*

Data Options

▶

×

**Sub aggregation**

Terms help

Terms

▼

Field

Status.keyword

▼

Order by

Metric: Count

▼

Order

Descending

▼

Size

5

☐

Group other values in separate bucket

☐

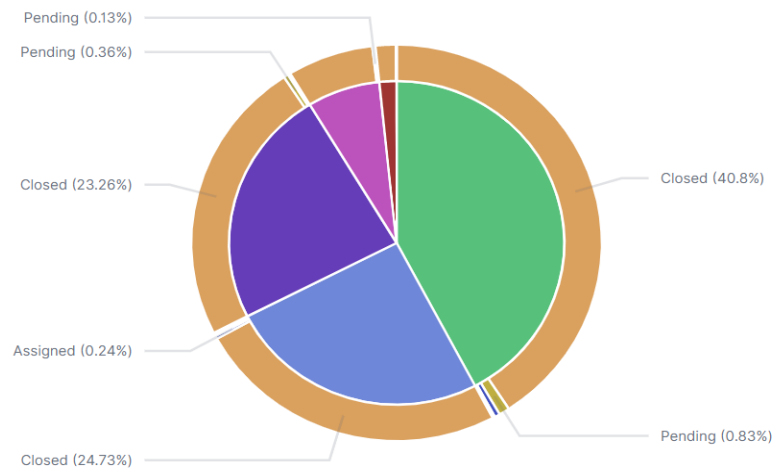
Show missing values

Custom label

> Advanced

+ Add

● BROOKLYN
 ● NEW YORK
 ● BRONX
 ● STATEN ISLAND
 ● JAMAICA
 ● Closed
 ● Pending
 ● Assigned
 ● Open
 ● Started
 ● In Progress



## 8. What is the total count of Top 5 Complaint types in Top 5 Cities?

nyc\*

Data

Metrics & axes

Panel settings

▶

×

▼ X-axis

Aggregation

Terms

▼

Field

City.keyword

▼

Order by

Metric: Count

▼

Order

Descending

▼

Size

5

☐ Group other values in separate bucket

☐ Show missing values

Custom label

City

Sub aggregation

Terms help

Terms

▼

Field

Complaint Type.keyword

▼

Order by

Metric: Count

▼

Order

Descending

▼

Size

5

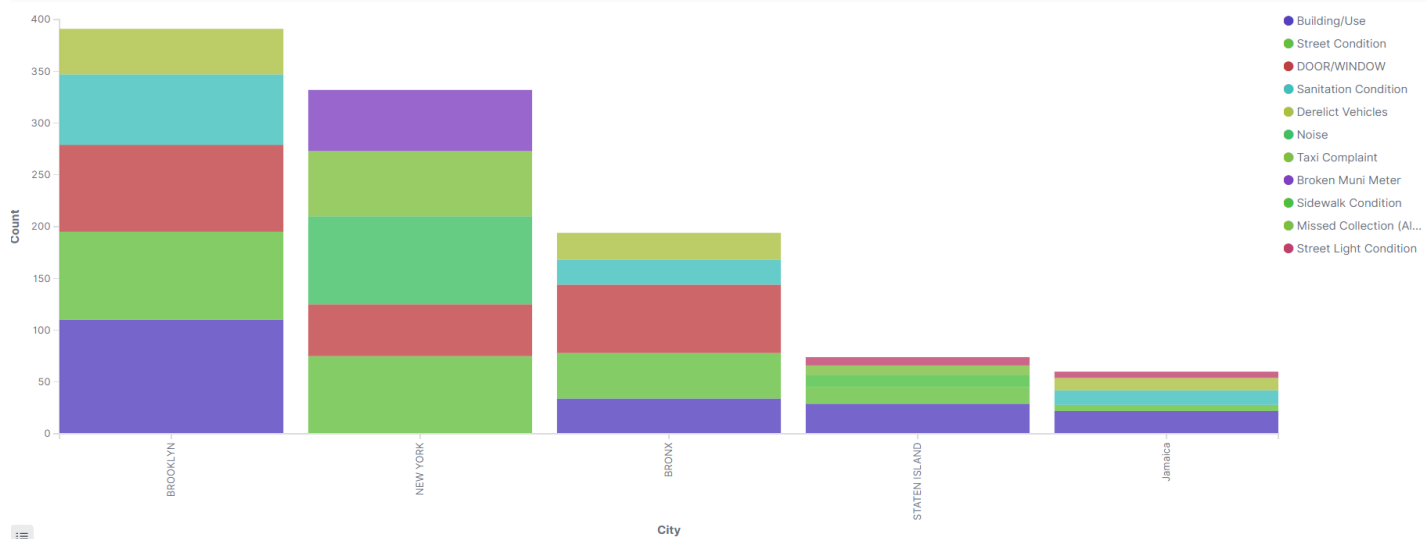
☐ Group other values in separate bucket

☐ Show missing values

Custom label

> Advanced

+ Add



## 9. What are the Top 5 Service Request Status of Top 5 Government Agencies?

nyc\*

Data

Options

▶

×

Field

Agency Name.keyword

▼

Order by

Metric: Count

▼

Order

Descending

▼

Size

5

☐ Group other values in separate bucket

☐ Show missing values

Custom label

> Advanced

> X-axis Status.keyword: D...

👁 = ✕

+ Add

nyc\*

Data

Options

▶

×

> Advanced

☒ X-axis
 

👁 = ✕

Sub aggregation

Terms

▼

Field

Status.keyword

▼

Order by

Metric: Count

▼

Order

Descending

▼

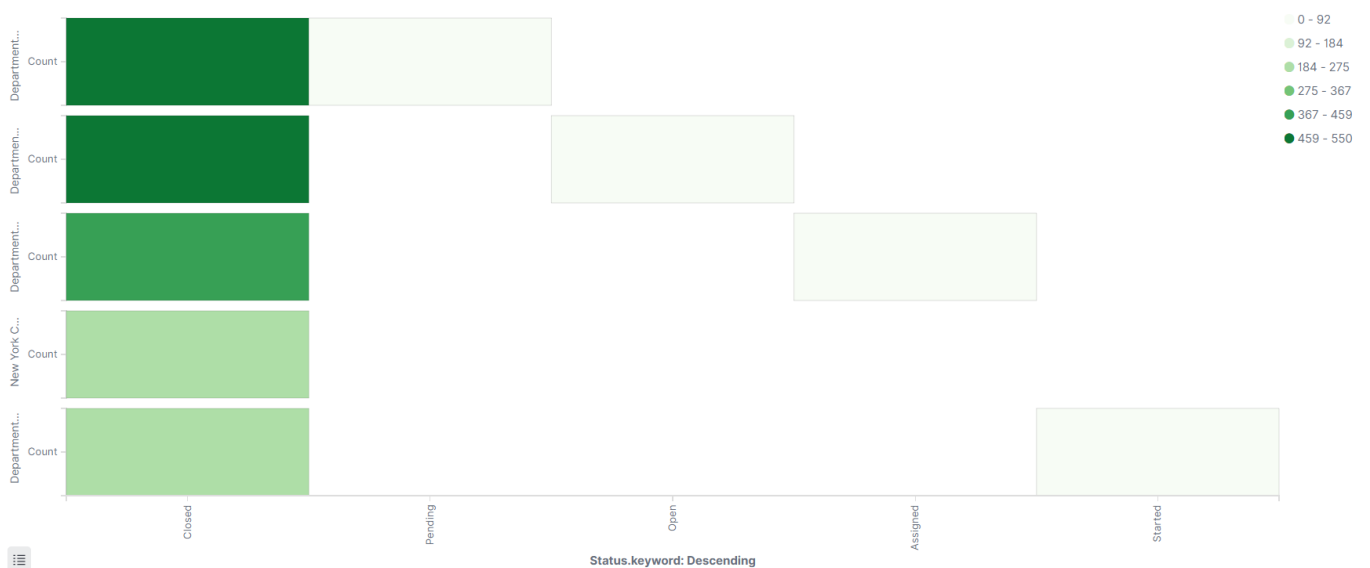
Size

5

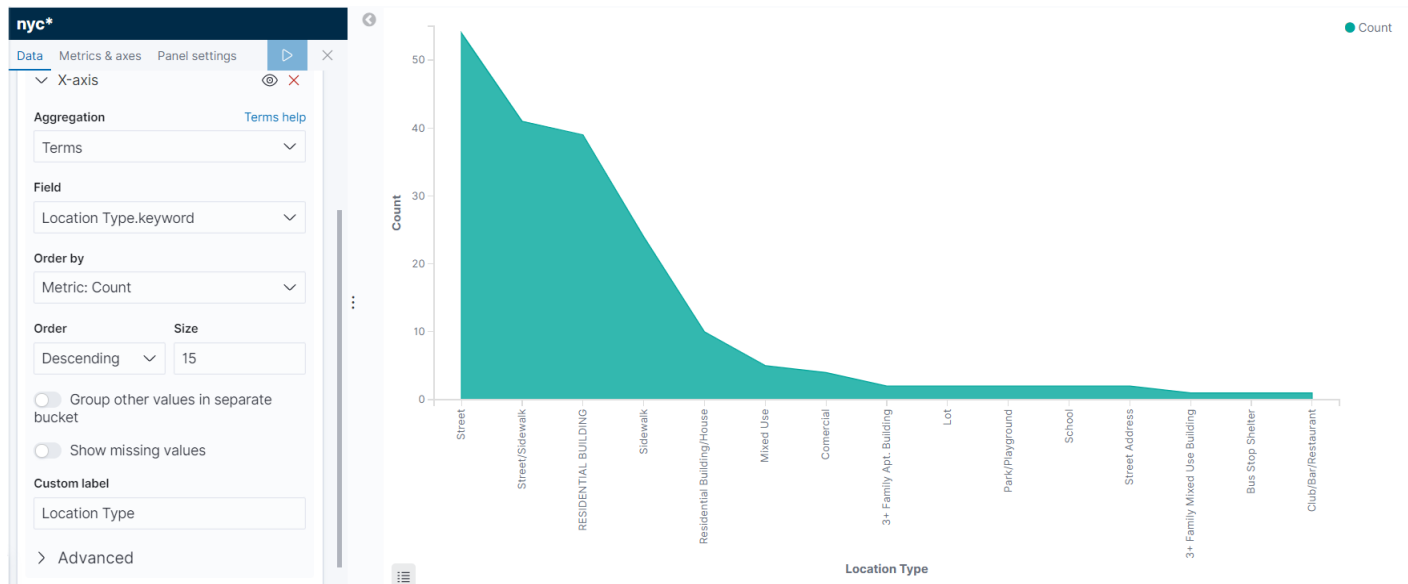
☐ Group other values in separate bucket

☐ Show missing values

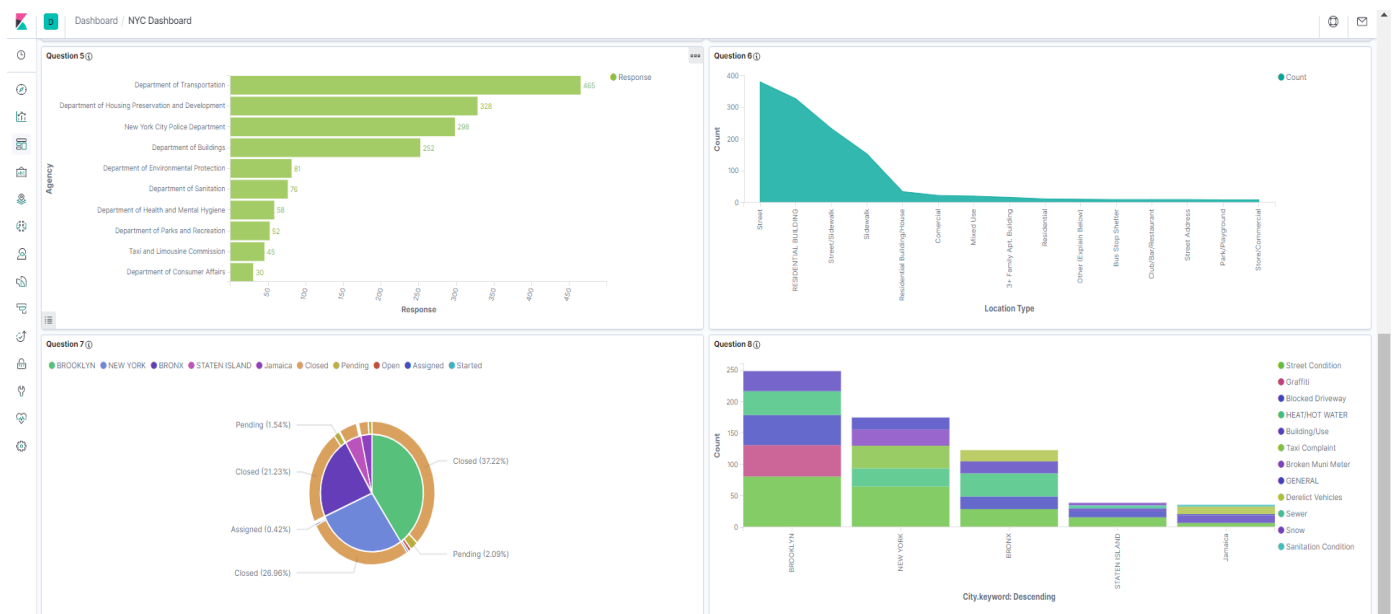
Custom label

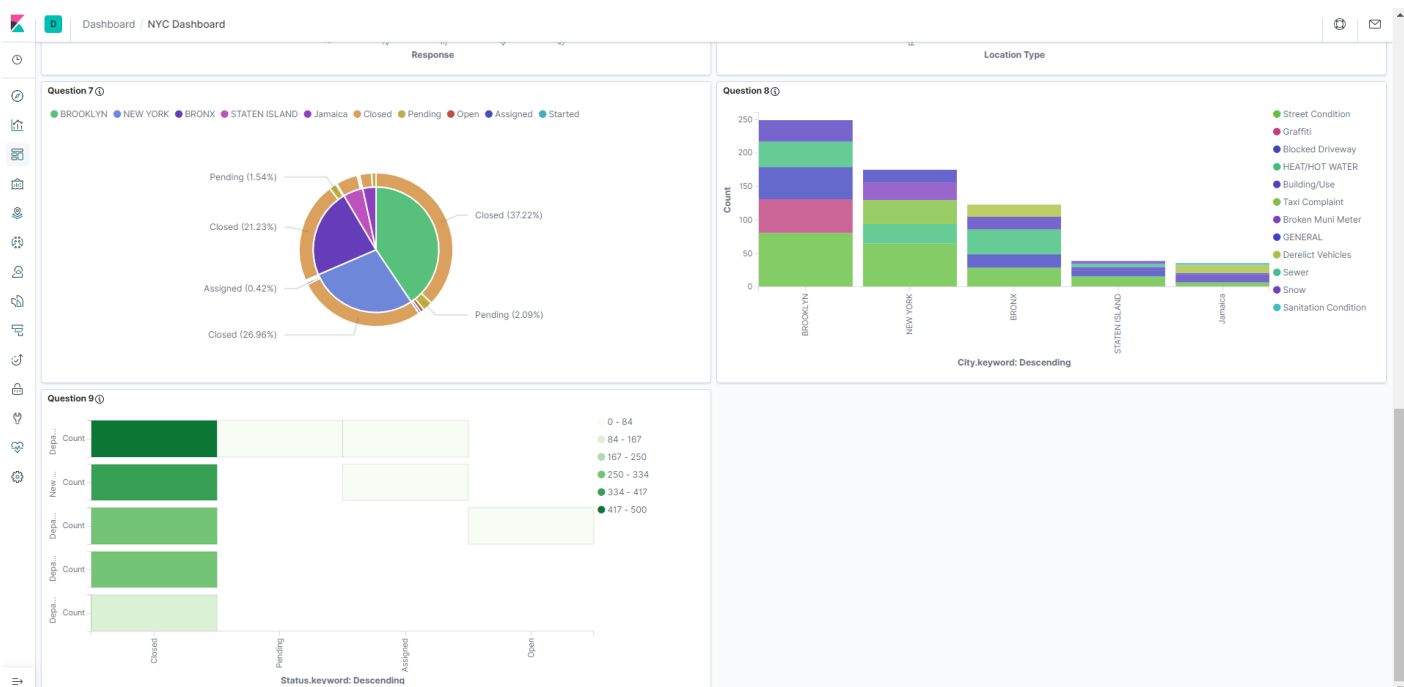


## 10. What is the top 15 location type based on frequency of SR?



## Dashboard for Questions 6 – 10





## References

College, C. (2021, 12 13). blackboard . Retrieved from blackboard:

[https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course\\_id= 317687\\_1&content\\_id= 6464709\\_1](https://gc.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id= 317687_1&content_id= 6464709_1)

## Appendix

### Setting up ELK on a GCP cluster

```

sriekhasampath992@dsa-ssignment2-m: ~ - Google Chrome
ssh.cloud.google.com/projects/elite-replica-335523/zones/us-central1-a/instances/dsa-ssignment2-m?authuser=1&hl=en_GB&projectNu...

the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
sriekhasampath992@dsa-ssignment2-m:~$ wget https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-7.5.1
-linux-x86_64.tar.gz
--2021-12-18 00:55:59-- https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-7.5.1-linux-x86_64.tar.g
z
Resolving artifacts.elastic.co (artifacts.elastic.co)... 34.120.127.130, 2600:1901:01d7::
Connecting to artifacts.elastic.co (artifacts.elastic.co)|34.120.127.130|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 290094012 (277M) [application/x-gzip]
Saving to: 'elasticsearch-7.5.1-linux-x86_64.tar.gz'

elasticsearch-7.5.1-linux-x86_64.tar.gz 100%[=====>] 276.65M 28.9MB/s in 11s

2021-12-18 00:56:10 (24.8 MB/s) - 'elasticsearch-7.5.1-linux-x86_64.tar.gz' saved [290094012/290094012]

sriekhasampath992@dsa-ssignment2-m:~$ wget https://artifacts.elastic.co/downloads/kibana/kibana-7.5.1-linux-x86_64.
tar.gz
--2021-12-18 00:57:07-- https://artifacts.elastic.co/downloads/kibana/kibana-7.5.1-linux-x86_64.tar.gz
Resolving artifacts.elastic.co (artifacts.elastic.co)... 34.120.127.130, 2600:1901:01d7::
Connecting to artifacts.elastic.co (artifacts.elastic.co)|34.120.127.130|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 238481011 (227M) [application/x-gzip]
Saving to: 'kibana-7.5.1-linux-x86_64.tar.gz'

kibana-7.5.1-linux-x86_64.tar.gz 100%[=====>] 227.43M 20.4MB/s in 13s

2021-12-18 00:57:20 (17.0 MB/s) - 'kibana-7.5.1-linux-x86_64.tar.gz' saved [238481011/238481011]

sriekhasampath992@dsa-ssignment2-m:~$ wget https://artifacts.elastic.co/downloads/logstash/logstash-7.5.1.tar.gz
--2021-12-18 00:57:31-- https://artifacts.elastic.co/downloads/logstash/logstash-7.5.1.tar.gz
Resolving artifacts.elastic.co (artifacts.elastic.co)... 34.120.127.130, 2600:1901:01d7::
Connecting to artifacts.elastic.co (artifacts.elastic.co)|34.120.127.130|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 165760774 (158M) [application/x-gzip]
Saving to: 'logstash-7.5.1.tar.gz'

logstash-7.5.1.tar.gz 100%[=====>] 158.08M 24.2MB/s in 7.6s

2021-12-18 00:57:39 (20.9 MB/s) - 'logstash-7.5.1.tar.gz' saved [165760774/165760774]

sriekhasampath992@dsa-ssignment2-m:~$

```

Modifying relevant information in elasticsearch configuration file using vi text editor,

```
srilekhasampath992@dsa-ssignment2-m: ~/elasticsearch-7.5.1/config - Google Chrome
ssh.cloud.google.com/projects/elite-replica-335523/zones/us-central1-a/instances/dsa-ssignment2-m?authuser=1&hl=en_GB&projectNu...
# Path to log files:
#path.logs: /path/to/logs

----- Memory -----
# Lock the memory on startup:
#bootstrap.memory_lock: true
# Make sure that the heap size is set to about half the memory available
# on the system and that the owner of the process is allowed to use this
# limit.
# Elasticsearch performs poorly when the system is swapping the memory.

----- Network -----
# Set the bind address to a specific IP (IPv4 or IPv6):
#network.host: 0.0.0.0
# Set a custom port for HTTP:
#http.port: 9200
# For more information, consult the network module documentation.

----- Discovery -----
# Pass an initial list of hosts to perform discovery when this node is started:
# The default list of hosts is ["127.0.0.1", "::1"]
#discovery.seed_hosts: ["10.128.0.2:9300"]
# Bootstrap the cluster using an initial set of master-eligible nodes:
#cluster.initial_master_nodes: ["dsa-ssignment2-m"]
# For more information, consult the discovery and cluster formation module documentation.

----- Gateway -----
# Block initial recovery after a full cluster restart until N nodes are started:

-- INSERT --
72,50 79%
```

Modifying relevant information in Kibana configuration file using vi text editor,

```
srilekhasampath992@dsa-ssignment2-m: ~/kibana-7.5.1-linux-x86_64/config - Google Chrome
ssh.cloud.google.com/projects/elite-replica-335523/zones/us-central1-a/instances/dsa-ssignment2-m?authuser=1&hl=en_GB&projectNu...
# Kibana is served by a back end server. This setting specifies the port to use.
#server.port: 5601

# Specifies the address to which the Kibana server will bind. IP addresses and host names are both valid values.
# The default is 'localhost', which usually means remote machines will not be able to connect.
# To allow connections from remote users, set this parameter to a non-loopback address.
#server.host: "0.0.0.0"

# Enables you to specify a path to mount Kibana at if you are running behind a proxy.
# Use the 'server.rewriteBasePath' setting to tell Kibana if it should remove the basePath
# from requests it receives, and to prevent a deprecation warning at startup.
# This setting cannot end in a slash.
#server.basePath: ""

# Specifies whether Kibana should rewrite requests that are prefixed with
# 'server.basePath' or require that they are rewritten by your reverse proxy.
# This setting was effectively always 'false' before Kibana 6.3 and will
# default to 'true' starting in Kibana 7.0.
#server.rewriteBasePath: false

# The maximum payload size in bytes for incoming server requests.
#server.maxPayloadBytes: 1048576

# The Kibana server's name. This is used for display purposes.
#server.name: "your-hostname"

# The URLs of the Elasticsearch instances to use for all your queries.
#elasticsearch.hosts: ["http://localhost:9200"]

# When this setting's value is true Kibana uses the hostname specified in the server.host
# setting. When the value of this setting is false, Kibana uses the hostname of the host
# that connects to this Kibana instance.
#elasticsearch.preserveHost: true
```



## Created new Firewall rules for elasticsearch and Kibana

Get real-time analytics with Network Intelligence Center

Use Network Intelligence Center for comprehensive monitoring and troubleshooting. [Learn more](#)

- ✓ Visualise your network resources
- ✓ Diagnose and prevent connectivity issues
- ✓ View packet loss and latency metrics
- ✓ Keep your firewall rules strict and efficient

[GO TO NETWORK INTELLIGENCE CENTER](#) [REMIND ME LATER](#)

Firewall rules control incoming or outgoing traffic to an instance. By default, incoming traffic from outside your network is blocked. [Learn more](#)

Note: App Engine firewalls are managed in the [App Engine firewall rules section](#).

Filter: Enter property name or value

Name	Type	Targets	Filters	Protocols/ports	Action	Priority	Network	Logs	Hit count	Last hit	Insights
<input checked="" type="checkbox"/> elasticsearch	Ingress	Apply to all	IP ranges: 0.0.0.0/0	tcp:9200	Allow	1000	default	Off	—	—	
<input checked="" type="checkbox"/> kibana	Ingress	Apply to all	IP ranges: 0.0.0.0/0	tcp:5601	Allow	1000	default	Off	—	—	
<input type="checkbox"/> default-allow-icmp	Ingress	Apply to all	IP ranges: 0.0.0.0/0	icmp	Allow	65534	default	Off	—	—	
<input type="checkbox"/> default-allow-internal	Ingress	Apply to all	IP ranges: 10.128.0.0/9	tcp:0-65535 udp:0-65535 icmp	Allow	65534	default	Off	—	—	
<input type="checkbox"/> default-allow-rdp	Ingress	Apply to all	IP ranges: 0.0.0.0/0	tcp:3389	Allow	65534	default	Off	—	—	
<input type="checkbox"/> default-allow-ssh	Ingress	Apply to all	IP ranges: 0.0.0.0/0	tcp:22	Allow	65534	default	Off	—	—	

## Loaded 311 service request csv file

```
srilekhasampath992@dsa-ssignment2-w-0:~/logstash-7.5.1$ wget https://www.dropbox.com/sh/smx7s2f32y4izkk/AADhiDbPkwjL
MYfrVDu76PvXa/311_service.csv?dl=0
--2021-12-18 20:19:08-- https://www.dropbox.com/sh/smx7s2f32y4izkk/AADhiDbPkwjLMYfrVDu76PvXa/311_service.csv?dl=0
Resolving www.dropbox.com (www.dropbox.com)... 162.125.3.18, 2620:100:601b:18::a27d:812
Connecting to www.dropbox.com (www.dropbox.com)|162.125.3.18|:443... connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: /sh/raw/smx7s2f32y4izkk/AADhiDbPkwjLMYfrVDu76PvXa/311_service.csv [following]
--2021-12-18 20:19:08-- https://www.dropbox.com/sh/raw/smx7s2f32y4izkk/AADhiDbPkwjLMYfrVDu76PvXa/311_service.csv
Reusing existing connection to www.dropbox.com:443.
HTTP request sent, awaiting response... 302 Found
Location: https://uce654e85ef7cef4c42c9249924f.dl.dropboxusercontent.com/cd/0/inline/BcHLcnRlL3qXZBTfQ_2QdEJu7TSz6G
V-iOFKnJ6aCjJ4btEoJvVFalDgLW2c1Q6jg5fGCva7Iaf87QmifPAk3mI15H98f-DOevA_wWU90Hy7Mt10VGCqet99StP6-sfK5nyd3pVGds2NCVArM-
Hi6du/file# [following]
--2021-12-18 20:19:09-- https://uce654e85ef7cef4c42c9249924f.dl.dropboxusercontent.com/cd/0/inline/BcHLcnRlL3qXZBTf
Q_2QdEJu7TSz6GV-iOFKnJ6aCjJ4btEoJvVFalDgLW2c1Q6jg5fGCva7Iaf87QmifPAk3mI15H98f-DOevA_wWU90Hy7Mt10VGCqet99StP6-sfK5ny
d3pVGds2NCVArM-Hi6du/file
Resolving uce654e85ef7cef4c42c9249924f.dl.dropboxusercontent.com (uce654e85ef7cef4c42c9249924f.dl.dropboxusercontent
.com)... 162.125.3.15, 2620:100:601b:15::a27d:80f
Connecting to uce654e85ef7cef4c42c9249924f.dl.dropboxusercontent.com (uce654e85ef7cef4c42c9249924f.dl.dropboxusercontent
.com)|162.125.3.15|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 13247207051 (12G) [text/plain]
Saving to: '311_service.csv?dl=0'

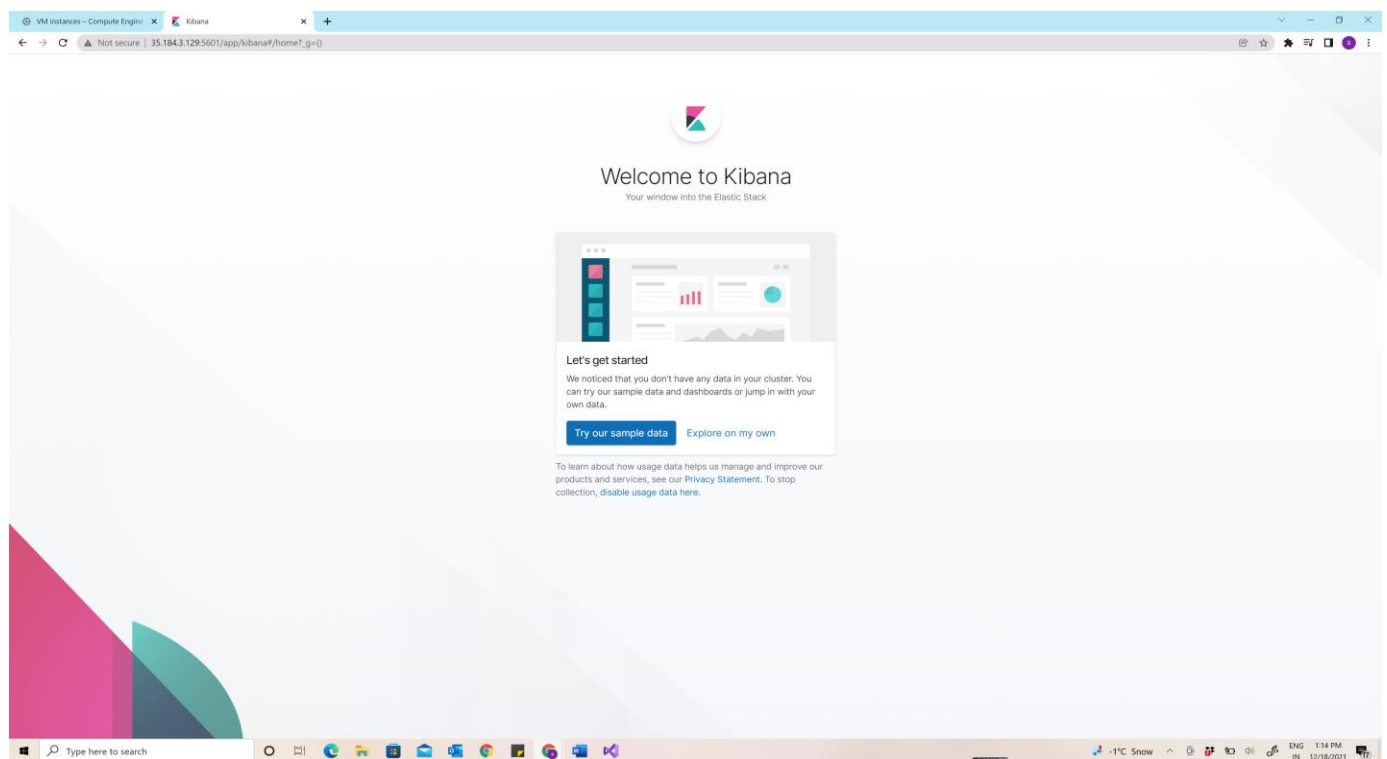
311_service.csv?dl=0      100%[=====>] 12.34G  107MB/s   in 2m 28s

2021-12-18 20:21:37 (85.4 MB/s) - '311_service.csv?dl=0' saved [13247207051/13247207051]
```

Modifying relevant information in Logstash configuration file using vi text editor,

```
srilekhasampath992@dsa-ssignment2-w-0:~$ ls
elasticsearch-7.5.1      kibana-7.5.1-linux-x86_64.tar.gz  logstash-7.5.1.tar.gz.1
elasticsearch-7.5.1-linux-x86_64.tar.gz  logstash-7.5.1                    logstash-7.5.1.tar.gz.2
kibana-7.5.1-linux-x86_64  logstash-7.5.1.tar.gz
srilekhasampath992@dsa-ssignment2-w-0:~$ cd logstash-7.5.1/
srilekhasampath992@dsa-ssignment2-w-0:~/logstash-7.5.1$ ls
'311_service.csv?dl=0'  Gemfile.lock  bin      lib      logstash-core-plugin-api  vendor
CONTRIBUTORS           LICENSE.txt   config   logs     modules                  x-pack
Gemfile                 NOTICE.TXT  data     logstash-core  tools
srilekhasampath992@dsa-ssignment2-w-0:~/logstash-7.5.1$ mv '311_service.csv?dl=0' newyork.csv
srilekhasampath992@dsa-ssignment2-w-0:~/logstash-7.5.1$ ls
CONTRIBUTORS  Gemfile.lock  NOTICE.TXT  config  lib  logstash-core  modules  tools  x-pack
Gemfile        LICENSE.txt   bin          data    logs logstash-core-plugin-api  newyork.csv  vendor
srilekhasampath992@dsa-ssignment2-w-0:~/logstash-7.5.1$ vi newyork.config
srilekhasampath992@dsa-ssignment2-w-0:~/logstash-7.5.1$ vi newyork.config
```

Starting Kibana using External IP of the cluster and Port number



## Creating Index pattern

### Create index pattern

Kibana uses index patterns to retrieve data from Elasticsearch indices for things like visualizations.

☐ Include system indices

#### Step 1 of 2: Define index pattern

Index pattern

You can use a \* as a wildcard in your index pattern.  
You can't use spaces or the characters \, /, ?, \*, <, >, |.

✓ **Success!** Your index pattern matches **1 index**.

nycinfo\_geo

Rows per page: 10

> Next step

### Create index pattern

Kibana uses index patterns to retrieve data from Elasticsearch indices for things like visualizations.

☐ Include system indices

#### Step 2 of 2: Configure settings

You've defined **nycinfo\_geo\*** as your index pattern. Now you can specify some settings before we create it.

Time Filter field name [Refresh](#)

I don't want to use the Time Filter

The Time Filter will use this field to filter your data by time.  
You can choose not to have a time field, but you will not be able to narrow down your data by a time range.

> Show advanced options

< Back **Create index pattern**

### ★ nycinfo\_geo\*

Default

This page lists every field in the **nycinfo\_geo\*** index and the field's associated core type as recorded by Elasticsearch. To change a field type, use the Elasticsearch [Mapping API](#).

Fields (90) Scripted fields (0) Source filters (0)

Q Filter All field types

Name	Type	Format	Searchable	Aggregatable	Excluded
@timestamp	date		•	•	
@version	string		•		
@version.keyword	string		•	•	
Address Type	string		•		
Address Type.keyword	string		•	•	
Agency	string		•		
Agency Name	string		•		
Agency Name.keyword	string		•	•	
Agency.keyword	string		•	•	
BBL	number		•	•	

Rows per page: 10

< 1 2 3 4 5 ... 9 >