Project: NYC Restaurant Inspection Data

Noé Breton et **Reda Bourssouf**

Installations de Elastic Search

On utilise docker compose et la commande `docker compose up -d` pour lancer elasticsearch et kibana.

```
version: '3.8'

services:
  elasticsearch:
    image: docker.elastic.co/elasticsearch/elasticsearch:8.4.2
    container_name: elasticsearch
    environment:
      - node.name=elasticsearch
      - cluster.name=es-docker-cluster
      - discovery.type=single-node
      - bootstrap.memory_lock=true
      - "ES_JAVA_OPTS=-Xms512m -Xmx512m"
      # Security settings for development
      - xpack.security.enabled=false
      - xpack.security.enrollment.enabled=false
      - xpack.security.http.ssl.enabled=false
      - xpack.security.transport.ssl.enabled=false
    ulimits:
      memlock:
        soft: -1
        hard: -1
    volumes:
      - es_data:/usr/share/elasticsearch/data
    ports:
      - "9200:9200"
      - "9300:9300"
    networks:
      - elastic
    healthcheck:
      test: ["CMD-SHELL", "curl -f http://localhost:9200/_cluster/health
|| exit 1"]
      interval: 30s
      timeout: 10s
      retries: 5

  kibana:
    image: docker.elastic.co/kibana/kibana:8.4.2
    container_name: kibana
    environment:
      - ELASTICSEARCH_HOSTS=http://elasticsearch:9200
```

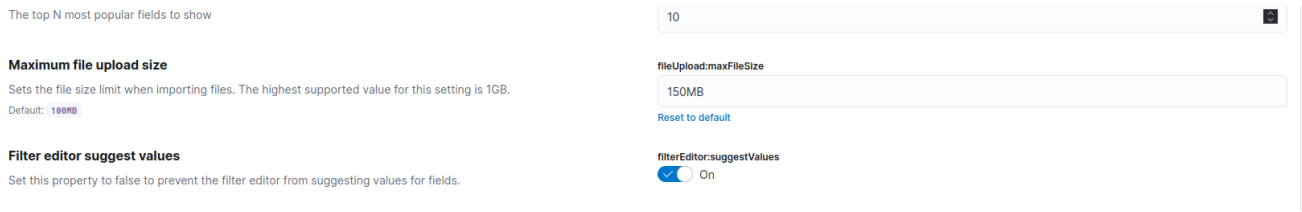
PROF

```
      - ELASTICSEARCH_USERNAME=kibana_system
      - ELASTICSEARCH_PASSWORD=
    ports:
      - "5601:5601"
    networks:
      - elastic
    depends_on:
      elasticsearch:
        condition: service_healthy
    healthcheck:
      test: ["CMD-SHELL", "curl -f http://localhost:5601/api/status ||
exit 1"]
      interval: 30s
      timeout: 10s
      retries: 5

volumes:
  es_data:
    driver: local

networks:
  elastic:
    driver: bridge
```

Ingestion

On cherche a ingerer NYC Restaurant Inspection Results.

La taille limite de fichier à ingerer est par default de 100b, malheuresement le csv a ingerer est de 123Mb, on change la limite a 150mb dans Advanced settings.



PROF

2. Questions

2.1 List all the neighborhoods in New York.

Par une agreggation sur la colonne BORO on recupère les neighborhoods unique en plus de leur nombre d'occurence.

Request:

```
GET /restaurantny/_search
{
  "size": 0,
  "aggs": {
```

```
    "unique_boro": {
      "terms": {
        "field": "BORO",
        "size": 100
      }
    }
  }
}
```

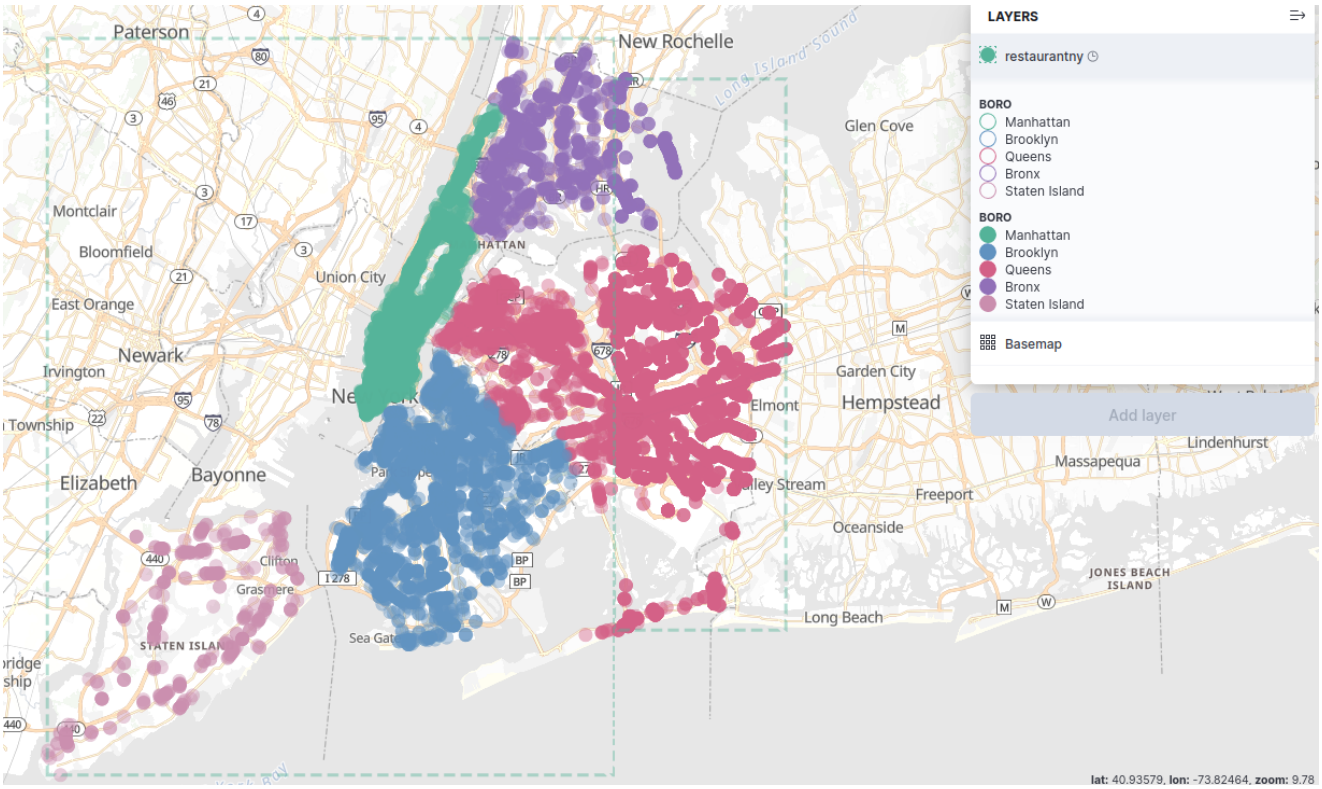
Response:

```
  "aggregations": {
    "unique_boro": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 0,
      "buckets": [
        {
          "key": "Manhattan",
          "doc_count": 106988
        },
        {
          "key": "Brooklyn",
          "doc_count": 74803
        },
        {
          "key": "Queens",
          "doc_count": 71095
        },
        {
          "key": "Bronx",
          "doc_count": 26613
        },
        {
          "key": "Staten Island",
          "doc_count": 10071
        }
      ]
    }
  }
}
```

PROF

Visalisation

Par cette visualisation on peut voir clairement tout les quartier de new york et les restautant present a l'interieur.



2.2 Which neighborhood has the most restaurants ?

La commande precedente peut etre reprise pour recuperer le quartier avec le plus de restaurant en verifiant que l'ordonnement est bien decroissant et en ajoutant size=1 pour obtenir seulemen le premier resultat. On utilise une deuxiemme aggregation de cardinalité sur l'id unique du restaurant pour obtenir le nombre de restaurant.

Request:

PROF

```
GET restaurantny/_search
{
  "size": 0,
  "aggs": {
    "restaurants_by_boro": {
      "terms": {
        "field": "BORO",
        "size": 1,
        "order": { "unique_restaurants.value": "desc" }
      },
      "aggs": {
        "unique_restaurants": {
          "cardinality": {
            "field": "CAMIS"
          }
        }
      }
    }
  }
}
```

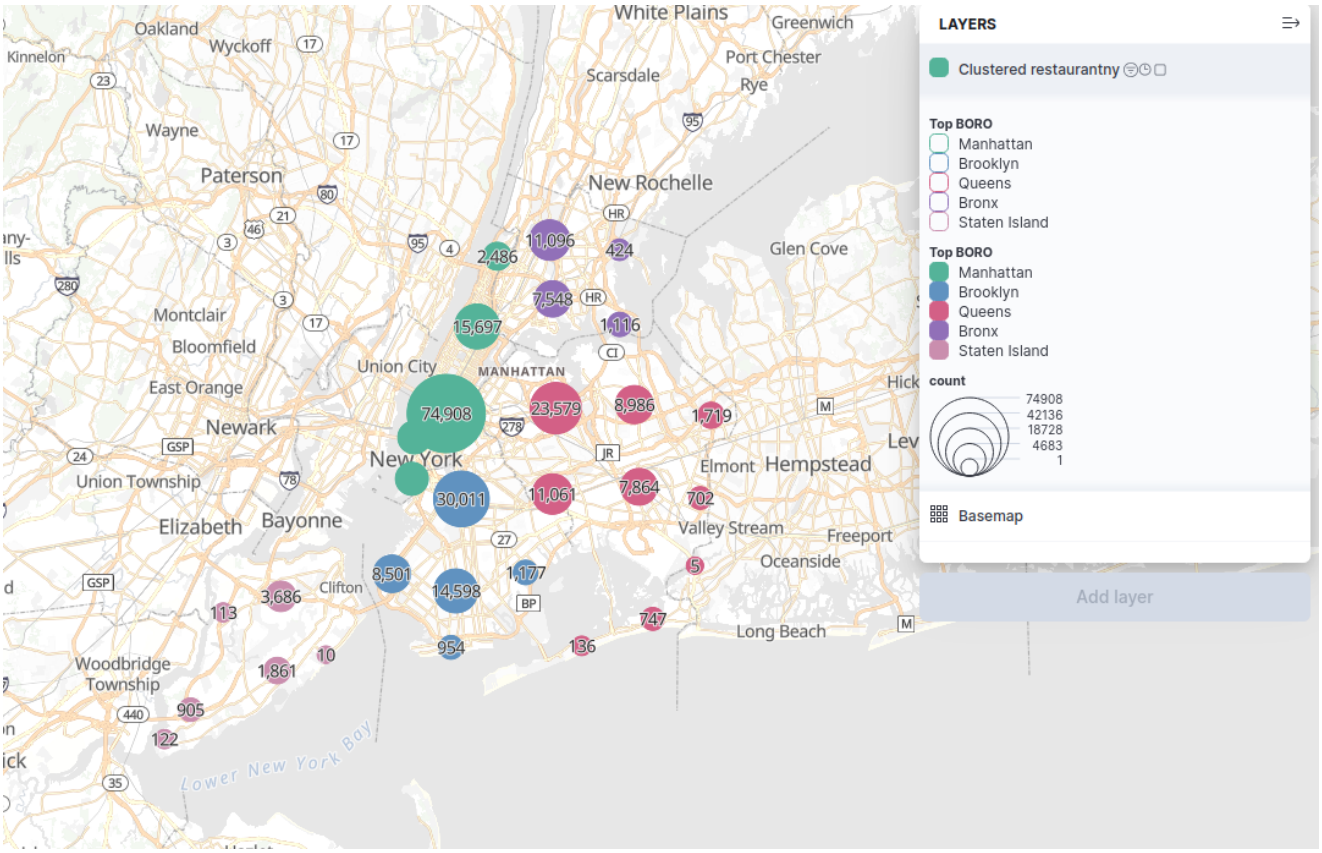
Resulat:


```
"aggregations": {
    "restaurants_by_boro": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 182582,
      "buckets": [
        {
          "key": "Manhattan",
          "doc_count": 106988,
          "unique_restaurants": {
            "value": 12113
          }
        }
      ]
    }
  }
}
```


Manahattan possède 12113 restaurant, faisant de ce quartier celui avec la plus grande offre de restauration.

visualitation

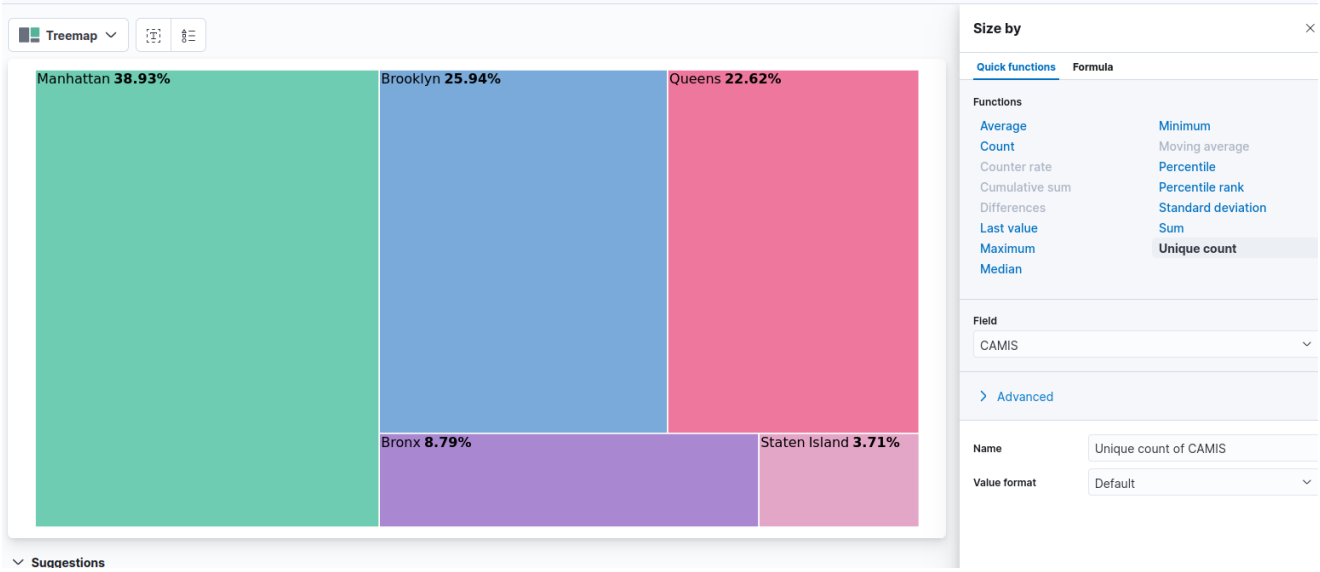
Resprendre la visualition precedente en ajoutant un label permet de se rendre compte approximativement du nombre de restaurant par quartier, mais cela manque de lisibilité



Comme la heatmap, on se rend compte de la plus grande densité de Manahattan mais ça reste moin evident qu'un barchart, on met en unique count CAMIS pour avoir seulement les restaurant.

PROF

On peux appercevoir la repartition des restaurant de new york par BORO



2.3 What does the violation code "04N" correspond to ?

On extrait l'attribut VIOLATION DESCRIPTION (et VIOLATION CODE histoire d'etre sur) d'une ligne dont l'attrivbut VIOLATION CODE a comme valeur 04N.

Requete

```
GET restaurantny/_search
{
  "_source": ["VIOLATION DESCRIPTION","VIOLATION CODE"],
  "query": {
    "term": {
      "VIOLATION CODE": "04N"
    }
  },
  "size": 1
}
```

PROF

Resultat

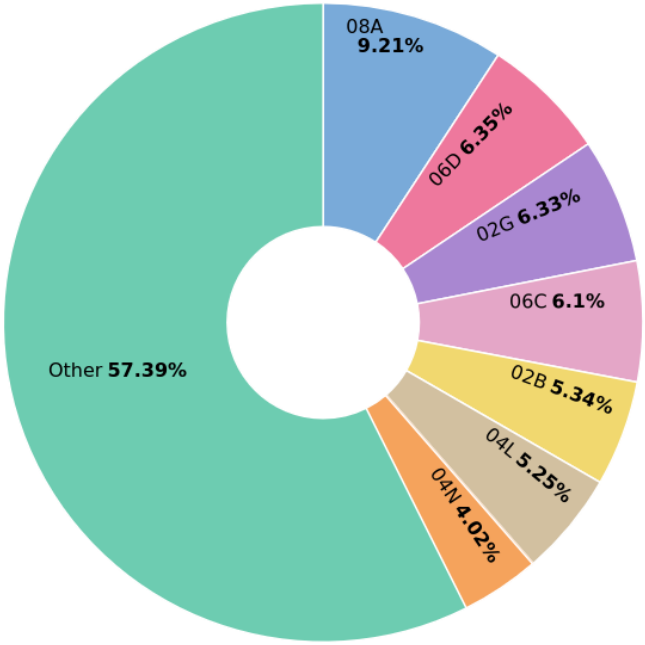

```
    "hits": [
      {
        "_index": "restaurantny",
        "_id": "rBOH3JkByv84jpscZM78",
        "_score": 3.1635146,
        "_source": {
          "VIOLATION CODE": "04N",
          "VIOLATION DESCRIPTION": "Filth flies or food/refuse/sewage
associated with (FRSA) flies or other nuisance pests in establishment’s
food and/or non-food areas. FRSA flies include house flies, blow flies,
bottle flies, flesh flies, drain flies, Phorid flies and fruit flies."
        }
      }
```

```
      ]
    }
}
```

Le code 04N correspond a présence de mouches ou d'autres insectes dans les zones de préparation, de stockage ou de service des aliments.

viusalitation

Voicis la proportion des code de violation present dans les restaurant enregistré, on voit que 10F est le code de violation le plus present, incluant les lignes qui ne presente pas de code.



2.4 Where are the restaurants (name, address, neighborhood) that have a grade of A?

PROF

```
  GET restaurantny/_search
{
  "_source": ["DBA","BUILDING","STREET","ZIPCODE","BORO","GRADE"],
  "query": {
    "term": {
      "GRADE": "A"
    }
  },"size": 100

}
```

response :

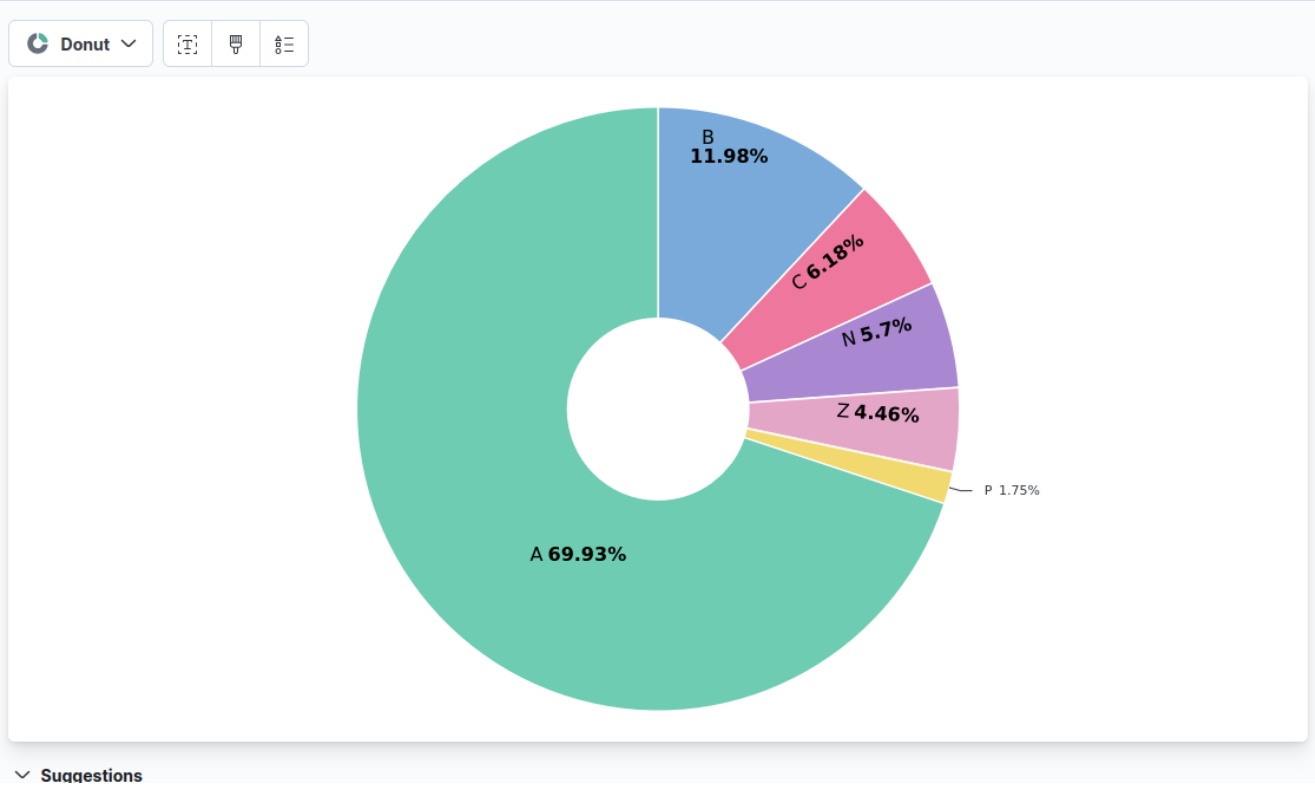
```
 "hits": [
      {
```

```
        "_index": "restaurantny",
        "_id": "rROH3JkByv84jpscZM78",
        "_score": 0.3891273,
        "_source": {
          "DBA": "PRESSED JUICERY",
          "BUILDING": "2857",
          "BORO": "Manhattan",
          "ZIPCODE": 10025,
          "GRADE": "A",
          "STREET": "BROADWAY"
        }
      },
      {
        "_index": "restaurantny",
        "_id": "sBOH3JkByv84jpscZM78",
        "_score": 0.3891273,
        "_source": {
          "DBA": "KPOT KOREAN BBQ & HOT POT (ON 2ND FL, BAY PLAZA
MALL)",
          "BUILDING": "200",
          "BORO": "Bronx",
          "ZIPCODE": 10475,
          "GRADE": "A",
          "STREET": "BAYCHESTER AVENUE"
        }
      },
      {
        "_index": "restaurantny",
        "_id": "shOH3JkByv84jpscZM78",
        "_score": 0.3891273,
        "_source": {
          "DBA": "BLUE MOUNTAIN REST & BAKERY",
          "BUILDING": "959",
          "BORO": "Brooklyn",
          "ZIPCODE": 11226,
          "GRADE": "A",
          "STREET": "FLATBUSH AVENUE"
        }
      },
      {
        "_index": "restaurantny",
        "_id": "tBOH3JkByv84jpscZM78",
        "_score": 0.3891273,
        "_source": {
          "DBA": "Bronx Slice",
          "BUILDING": "37",
          "BORO": "Bronx",
          "ZIPCODE": 10454,
          "GRADE": "A",
          "STREET": "BRUCKNER BOULEVARD"
        }
      },
      {
```

```
        "_index": "restaurantny",
        "_id": "txOH3JkByv84jpscZM78",
        "_score": 0.3891273,
        "_source": {
          "DBA": "CJ DIAMOND CAFE",
          "BUILDING": "4102",
          "BORO": "Queens",
          "ZIPCODE": 11355,
          "GRADE": "A",
          "STREET": "COLLEGE POINT BLVD"
        }
      },...
```

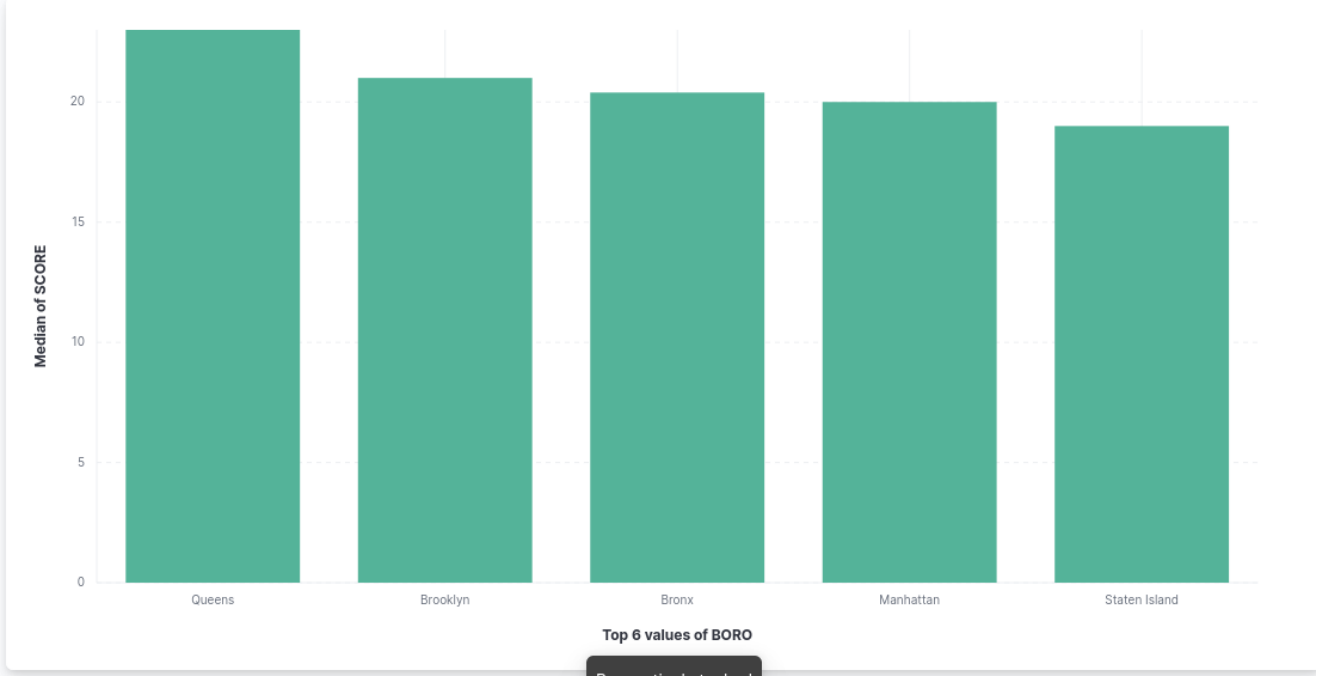

Visualisation

L'outil de visualisation n'autorise pas d'utiliser DBA comme une metrique, je choisis de visualisé la proportion des notes parmis tout les restaurant, les outils dynamique du dashbord permetteron d'ajuster la granulité.



PROF

On peux aussi ajouter la medianne des score par BORO, on decouvre que le Queens a la meilleure medianne parmis les BORO



2.5 What is the most popular cuisine? And by neighborhood?

```
GET restaurantny_final/_search
{
  "size": 0,

  "aggs": {
    "top3_Description": {
      "terms": {
        "field": "CUISINE DESCRIPTION",
        "size": 3,
        "order": { "unique_restaurants.value": "desc" }
      },
      "aggs": {
        "unique_restaurants": {
          "cardinality": {
            "field": "CAMIS"
                      }
        }
      }
    }
  }
}
```

PROF

response:

```
  "aggregations": {
    "top3_Description": {
      "doc_count_error_upper_bound": -1,
      "sum_other_doc_count": 192398,
```

```
      "buckets": [
        {
          "key": "American",
          "doc_count": 45070,
          "unique_restaurants": {
            "value": 4929
          }
        },
        {
          "key": "Chinese",
          "doc_count": 28142,
          "unique_restaurants": {
            "value": 2154
          }
        },
        {
          "key": "Coffee/Tea",
          "doc_count": 20167,
          "unique_restaurants": {
            "value": 2100
          }
        }
      ]
    }
  }
```

La cuisine la plus populaire de new york par le nombre de restautant est la cuisine americaine, avec 4964 restaurant suivit par la cuisine chinoise et les coffee/tea shop.

Par BORO:

PROF

```
GET restaurantny_final/_search
{
  "size": 0,

  "aggs": {
    "by_boro": {
      "terms": {
        "field": "BORO",
        "size": 10
      },
  "aggs": {
    "top3_Description": {
      "terms": {
        "field": "CUISINE DESCRIPTION",
        "size": 1,
        "order": { "unique_restaurants.value": "desc" }
      },
      "aggs": {
        "unique_restaurants": {
```

```
            "cardinality": {
              "field": "CAMIS"
                       }
          }
        }
      }
      }
      }
    }
  }
```

reponse:

```
"aggregations": {
    "by_boro": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 0,
      "buckets": [
        {
          "key": "Manhattan",
          "doc_count": 106988,
          "top3_Description": {
            "doc_count_error_upper_bound": -1,
            "sum_other_doc_count": 82973,
            "buckets": [
              {
                "key": "American",
                "doc_count": 22371,
                "unique_restaurants": {
                  "value": 2476
                }
              }
            ]
          }
        },
        {
          "key": "Brooklyn",
          "doc_count": 74803,
          "top3_Description": {
            "doc_count_error_upper_bound": -1,
            "sum_other_doc_count": 63742,
            "buckets": [
              {
                "key": "American",
                "doc_count": 10057,
                "unique_restaurants": {
                  "value": 1094
                }
              }
            ]
          }
```

PROF

```
      },
      {
        "key": "Queens",
        "doc_count": 71095,
        "top3_Description": {
          "doc_count_error_upper_bound": -1,
          "sum_other_doc_count": 62322,
          "buckets": [
            {
              "key": "American",
              "doc_count": 7988,
              "unique_restaurants": {
                "value": 854
              }
            }
          ]
        }
      },
      {
        "key": "Bronx",
        "doc_count": 26613,
        "top3_Description": {
          "doc_count_error_upper_bound": -1,
          "sum_other_doc_count": 23433,
          "buckets": [
            {
              "key": "American",
              "doc_count": 2948,
              "unique_restaurants": {
                "value": 317
              }
            }
          ]
        }
      },
      {
        "key": "Staten Island",
        "doc_count": 10071,
        "top3_Description": {
          "doc_count_error_upper_bound": -1,
          "sum_other_doc_count": 8237,
          "buckets": [
            {
              "key": "American",
              "doc_count": 1706,
              "unique_restaurants": {
                "value": 164
              }
            }
          ]
```

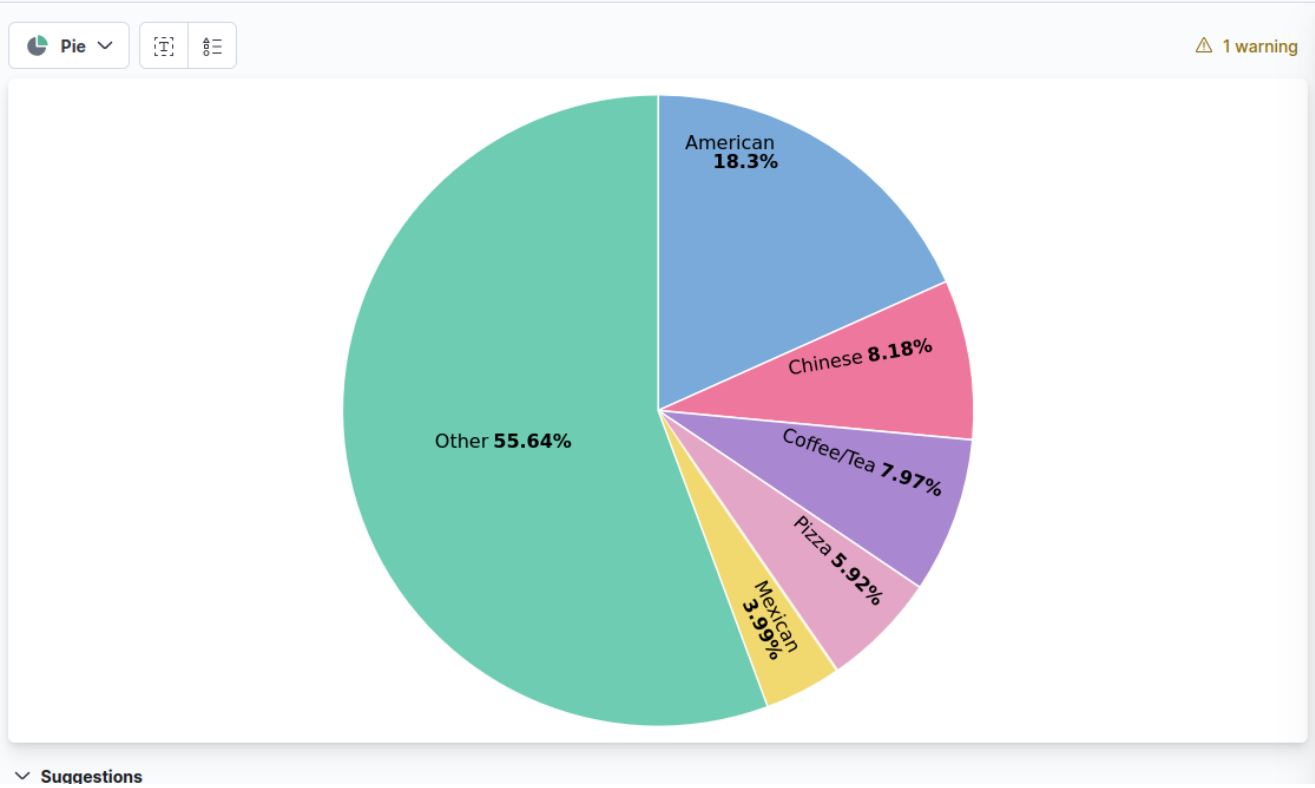

PROF

par BORO, la cuisine la plus populaire est

BORO	Cuisine la plus populaire	Nombre de restaurants
Manhattan	American	2476
Brooklyn	American	1094
Queens	American	854
Bronx	American	317
Staten Island	American	164

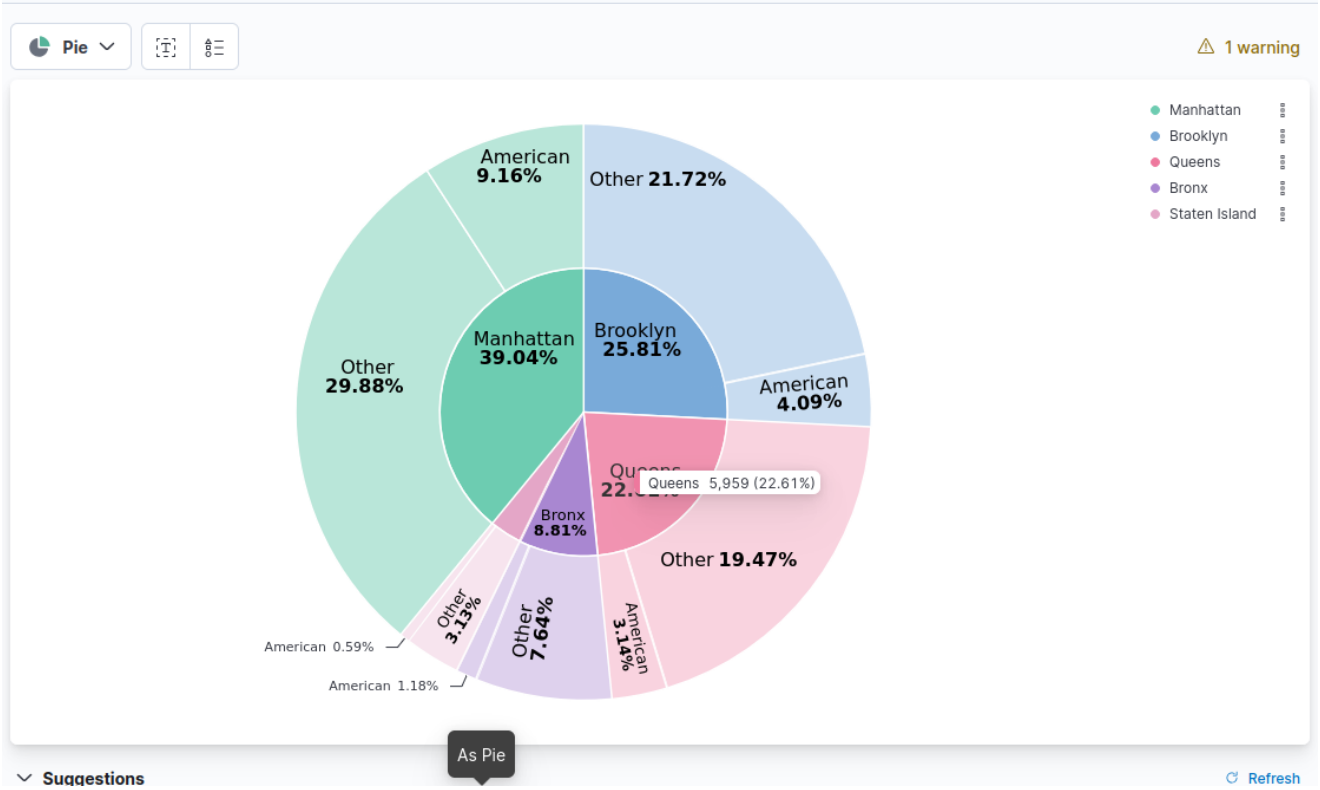
visualisation

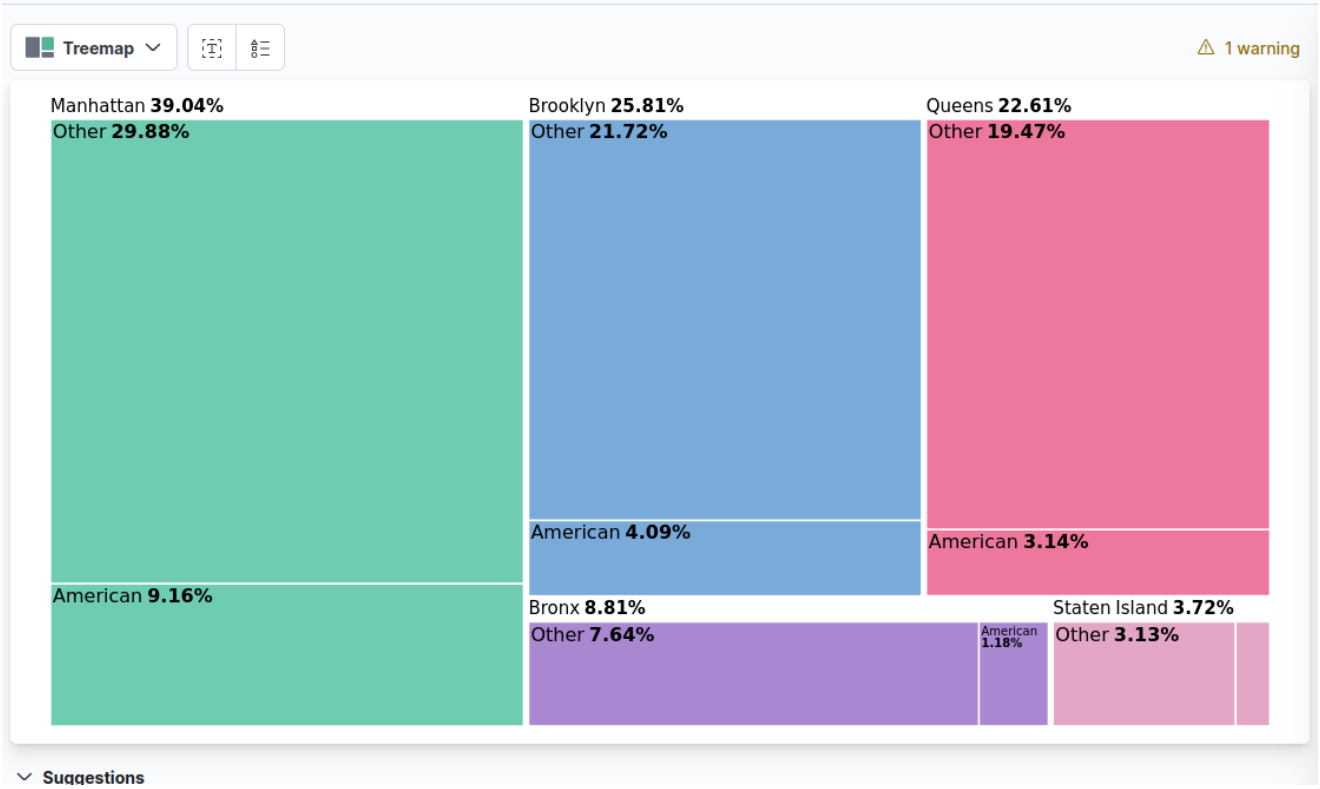
Pour New-york, on peux utiliser un pie chart classique, ou un waffle chart (la visibilité reste moin bonne)



PROF

On peux utiliser un double pie chart ou une mosaique pour voir la cuisine la plus populaire par BORO





PROF

2.6 What is the date of the last inspection?

```
GET restaurantny/_search
{
  "_source": ["INSPECTION DATE","DBA"],
  "sort": [
    { "INSPECTION DATE": { "order": "desc" } }
```

```
    ],
    "size": 1
  }
```

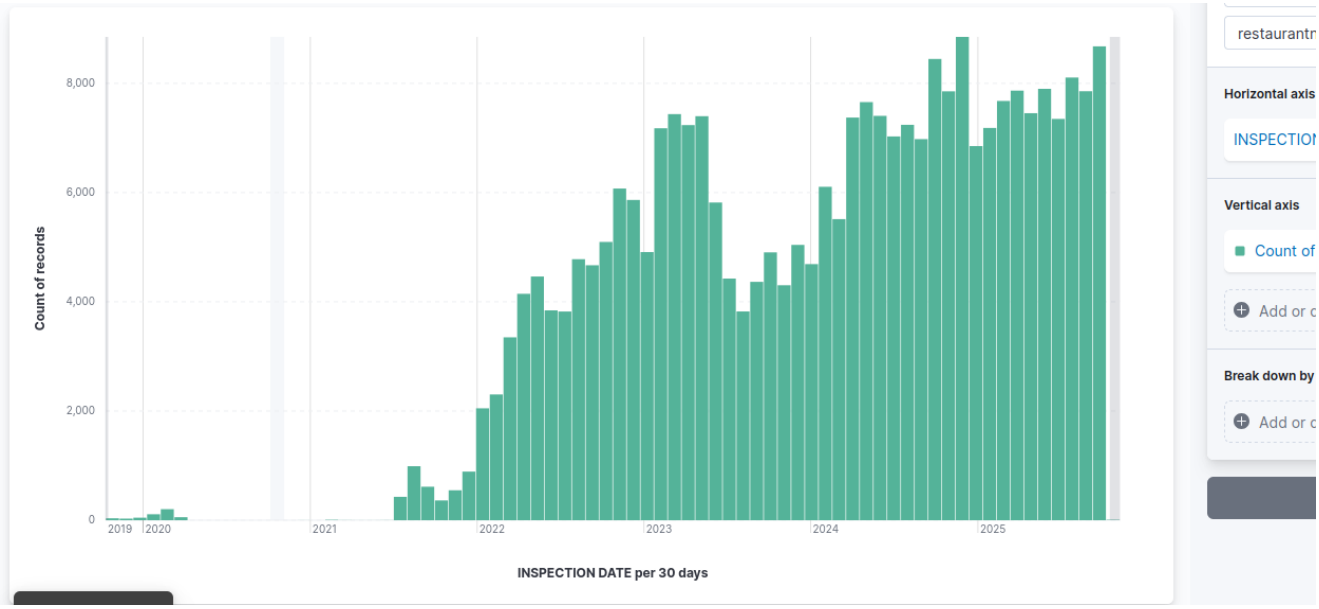
reponse :


```
    "hits": [
      {
        "_index": "restaurantny",
        "_id": "UBOH3JkByv84jpscW10l",
        "_score": null,
        "_source": {
          "DBA": "LE PETIT MONSTRE",
          "INSPECTION DATE": "10/09/2025"
        },
        "sort": [
          1759968000000
        ]
      }
```


la derniere inspection date du 10/09/2025 et concerne LE PETIT MONTRE.

visulation

On peux utiliser un diagramme temporel en barre ou en ligne pour connaire l'evolution du nombre d'inspection en fonction du temps



PROF

2.7 Provide a list of Chinese restaurants with an A grade in Brooklyn.

On utilise la commande bool pour trouver les resaurant qui doivent (must) remplir les condition "CUISINE DESCRIPTION": "Chinese", "GRADE": "A" et "BORO": "Brooklyn".

```
GET restaurantny/_search
{
  "_source": ["DBA"],
  "query": {
    "bool": {
      "must": [
        { "term": { "CUISINE DESCRIPTION": "Chinese" } },
        { "term": { "GRADE": "A" } },
        { "term": { "BORO": "Brooklyn" } }
      ]
    }
  }
}
```

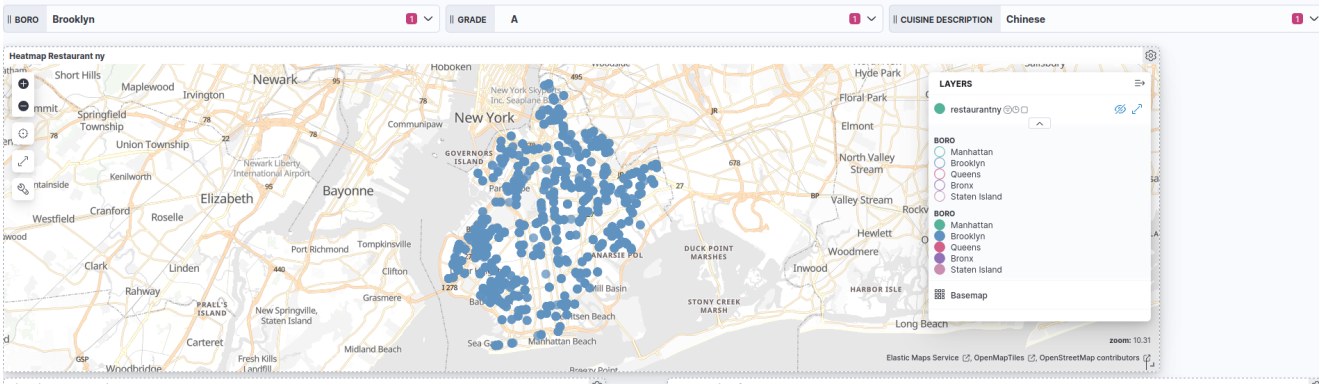
resultat:

```
 "hits": [
      {
        "_index": "restaurantny",
        "_id": "rBOH3JkByv84jpscZM_8",
        "_score": 4.0605974,
        "_source": {
          "DBA": "NEW CENTURY CHINESE RESTAURANT"
        }
      },
      {
        "_index": "restaurantny",
        "_id": "HhOH3JkByv84jpscZND8",
        "_score": 4.0605974,
        "_source": {
          "DBA": "NEW GREAT WALL 1419"
        }
      },
      {
        "_index": "restaurantny",
        "_id": "nhOH3JkByv84jpscZND8",
        "_score": 4.0605974,
        "_source": {
          "DBA": "FOOD LOVER BAKERY"
        }
      },
      {
        "_index": "restaurantny",
        "_id": "UROH3JkByv84jpscZNH8",
        "_score": 4.0605974,
        "_source": {
          "DBA": "SUN GARDEN"
        }
      },
      {
```

```
        "_index": "restaurantny",
        "_id": "WROH3JkByv84jpscZNH8",
        "_score": 4.0605974,
        "_source": {
          "DBA": "GRAND PANDA"
        }
      },
      {...
```

Visualisation

On peux utiliser les controles pour obtenir une liste des restaurant avec ces critères



2.8 Provide the address of the restaurant LADUREE

en faisant un match, on decouvre qu'il y a plusieur restaurant LADUREE (laduree, LADUREE soho et LADUREE)

PROF

On essaye une requete pour avoir le terme exact LADUREE. mais cela ne marche toujours pas, comme si DBA ne marchais pas avec une recherche exacte. En inspectant le mapping on s'apperçoit que DBA n'a pas d'option "keyword" qui permet une recherche exacte (contrairement a GRADE par exemple). On indexe keyword a DBA dans un nouveau index.

mapping de GRADE

```
{
  "restaurantny": {
    "mappings": {
      "GRADE": {
        "full_name": "GRADE",
        "mapping": {
          "GRADE": {
            "type": "keyword"
          }
        }
      }
    }
  }
}
```

mapping de DBA

```
{
  "restaurantny": {
    "mappings": {
      "DBA": {
        "full_name": "DBA",
        "mapping": {
          "DBA": {
            "type": "text"
          }
        }
      }
    }
  }
}
```

PROF

ajout de keyword a DBA a restaurant NY v1

```
PUT restaurantny_v1
{
  "mappings": {
    "properties": {
      "DBA": {
        "type": "text",
        "fields": {
          "keyword": { "type": "keyword" }
        }
      }    }
  }
}
```

reindexation

```
POST _reindex
{
  "source": { "index": "restaurantny" },
  "dest":   { "index": "restaurantny_v1" }
}
```

On ressaye la requete, on aggress sur le numero CAMIS pour être sur d'avoir la liste des restaurant unique en fonction de leur ID appelé LADUREE, on met la size de la seconde agreggation a 1 pour ne pas avoir des doublons en cas de multiple inspections.

```
{
  "size": 0,
  "query": {
    "term": {
      "DBA.keyword": "LADUREE"
    }
  },
  "aggs": {
    "unique_CAMIS": {
      "terms": {
        "field": "CAMIS",
        "size": 100
      },
      "aggs": {
        "sample_doc": {
          "top_hits": {
            "_source": ["DBA", "BUILDING", "STREET", "BORO"],
            "size": 1
          }
        }
      }
    }
  }
}
```

PROF

resultat

```
{
  "took": 0,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
```

```
      "failed": 0
    },
    "hits": {
      "total": {
        "value": 6,
        "relation": "eq"
      },
      "max_score": null,
      "hits": []
    },
    "aggregations": {
      "unique_CAMIS": {
        "doc_count_error_upper_bound": 0,
        "sum_other_doc_count": 0,
        "buckets": [
          {
            "key": 50054046,
            "doc_count": 6,
            "sample_doc": {
              "hits": {
                "total": {
                  "value": 6,
                  "relation": "eq"
                },
                "max_score": 10.704353,
                "hits": [
                  {
                    "_index": "restaurantny_v1",
                    "_id": "jhWH3JkByv84jpsclOJK",
                    "_score": 10.704353,
                    "_source": {
                      "DBA": "LADUREE",
                      "BUILDING": "864",
                      "BORO": "Manhattan",
                      "STREET": "MADISON AVENUE"
                    }
                  }
                ]
              }
            }
          }
        ]
      }
    }
  }
```


PROF

l'adresse du restautant LADUREE est 864 MADISON AVENUE,Manhattan.

visulisation

pour visualiser avec les nouveau filtre, on update notre dataview pour utiliser le nouvel index, mais je me suis rendu compte que le reindexage n'a pas pris en compte les dates ainsi que "location", je corrige le reindexage.

mapping de restaurantny


```
{
  "restaurantny": {
    "mappings": {
      "_meta": {
        "created_by": "file-data-visualizer"
      },
      "properties": {
        "@timestamp": {
          "type": "date"
        },
        "ACTION": {
          "type": "text"
        },
        "BBL": {
          "type": "long"
        },
        "BIN": {
          "type": "long"
        },
        "BORO": {
          "type": "keyword"
        },
        "BUILDING": {
          "type": "keyword"
        },
        "CAMIS": {
          "type": "long"
        },
        "CRITICAL FLAG": {
          "type": "keyword"
        },
        "CUISINE DESCRIPTION": {
          "type": "keyword"
        },
        "Census Tract": {
          "type": "long"
        },
        "Community Board": {
          "type": "long"
        },
        "Council District": {
          "type": "long"
        },
        "DBA": {
          "type": "text"
```


PROF

```
        },
        "GRADE": {
          "type": "keyword"
        },
        "GRADE DATE": {
          "type": "date",
          "format": "MM/dd/yyyy"
        },
        "INSPECTION DATE": {
          "type": "date",
          "format": "MM/dd/yyyy"
        },
        "INSPECTION TYPE": {
          "type": "keyword"
        },
        "Latitude": {
          "type": "double"
        },
        "Longitude": {
          "type": "double"
        },
        "NTA": {
          "type": "keyword"
        },
        "PHONE": {
          "type": "keyword"
        },
        "RECORD DATE": {
          "type": "date",
          "format": "MM/dd/yyyy"
        },
        "SCORE": {
          "type": "long"
        },
        "STREET": {
          "type": "keyword"
        },
        "VIOLATION CODE": {
          "type": "keyword"
        },
        "VIOLATION DESCRIPTION": {
          "type": "text"
        },
        "ZIPCODE": {
          "type": "long"
        },
        "location": {
          "type": "geo_point"
        }
      }
    }
  }
}
```

PROF

Mapping de restaurantny_v1


```
{
  "restaurantny_v1": {
    "mappings": {
      "properties": {
        "@timestamp": {
          "type": "date"
        },
        "ACTION": {
          "type": "text",
          "fields": {
            "keyword": {
              "type": "keyword",
              "ignore_above": 256
            }
          }
        },
        "BBL": {
          "type": "long"
        },
        "BIN": {
          "type": "long"
        },
        "BORO": {
          "type": "text",
          "fields": {
            "keyword": {
              "type": "keyword",
              "ignore_above": 256
            }
          }
        },
        "BUILDING": {
          "type": "text",
          "fields": {
            "keyword": {
              "type": "keyword",
              "ignore_above": 256
            }
          }
        },
        "CAMIS": {
          "type": "long"
        },
        "CRITICAL FLAG": {
          "type": "text",
          "fields": {
            "keyword": {
              "type": "keyword",
```

```
            "ignore_above": 256
          }
        }
      },
      "CUISINE DESCRIPTION": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword",
            "ignore_above": 256
          }
        }
      },
      "Census Tract": {
        "type": "long"
      },
      "Community Board": {
        "type": "long"
      },
      "Council District": {
        "type": "long"
      },
      "DBA": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword"
          }
        }
      },
      "GRADE": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword",
            "ignore_above": 256
          }
        }
      },
      "GRADE DATE": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword",
            "ignore_above": 256
          }
        }
      },
      "INSPECTION DATE": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword",
```

```
            "ignore_above": 256
          }
        }
      },
      "INSPECTION TYPE": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword",
            "ignore_above": 256
          }
        }
      },
      "Latitude": {
        "type": "float"
      },
      "Longitude": {
        "type": "float"
      },
      "NTA": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword",
            "ignore_above": 256
          }
        }
      },
      "PHONE": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword",
            "ignore_above": 256
          }
        }
      },
      "RECORD DATE": {
        "type": "text",
        "fields": {
          "keyword": {
            "type": "keyword",
            "ignore_above": 256
          }
        }
      },
      "SCORE": {
        "type": "long"
      },
      "STREET": {
        "type": "text",
        "fields": {
          "keyword": {
```

```
              "type": "keyword",
              "ignore_above": 256
            }
          }
        },
        "VIOLATION CODE": {
          "type": "text",
          "fields": {
            "keyword": {
              "type": "keyword",
              "ignore_above": 256
            }
          }
        },
        "VIOLATION DESCRIPTION": {
          "type": "text",
          "fields": {
            "keyword": {
              "type": "keyword",
              "ignore_above": 256
            }
          }
        },
        "ZIPCODE": {
          "type": "long"
        },
        "location": {
          "type": "text",
          "fields": {
            "keyword": {
              "type": "keyword",
              "ignore_above": 256
            }
          }
        }
      }
    }
  }
}
```

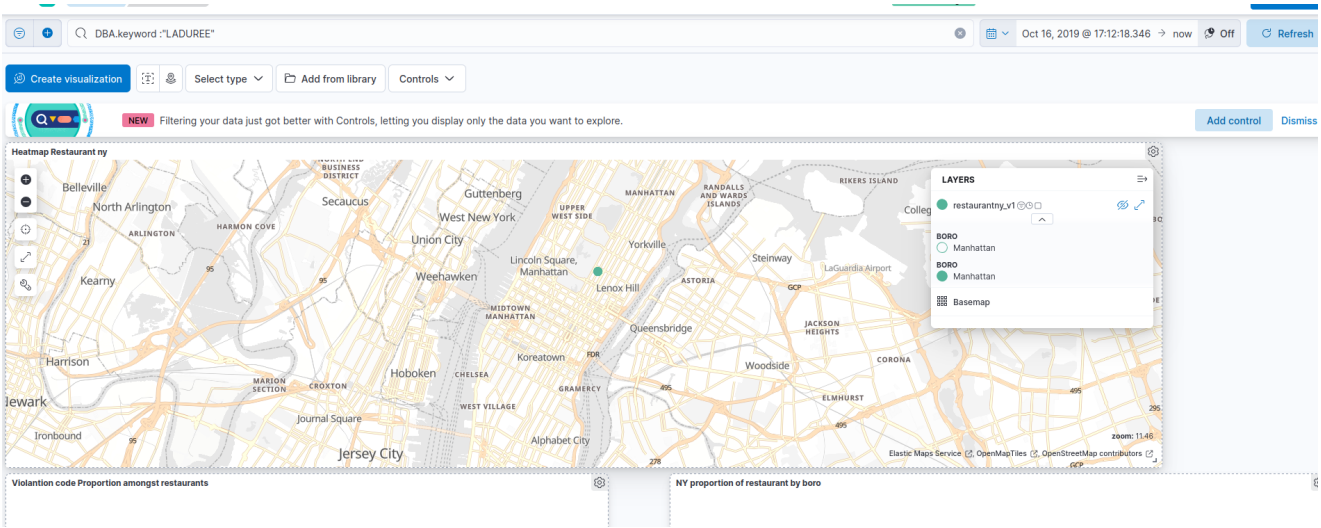
PROF

On recommance l'indexation

```
PUT restaurantny_v1
{
  "mappings": {
    "properties": {
      "DBA": {
        "type": "text",
        "fields": {
          "keyword": { "type": "keyword" }
        }
```

```
        },
                "INSPECTION DATE": {
          "type": "date",
          "format": "MM/dd/yyyy"
        },
             "GRADE DATE": {
          "type": "date",
          "format": "MM/dd/yyyy"
        },
                "RECORD DATE": {
          "type": "date",
          "format": "MM/dd/yyyy"
        },         "location": {
          "type": "geo_point"
        }
      }

    }
}
```

On peux finallement visualiser la localisation du restaurant LADUREE



PROF

Cette methode, bien que optimisé, est assez longue et provoque des erreurs si l'on oublie des field dans le mapping ou si l'on ne reecrit pas le meme mapping a la main (par exemple, tout mes type keyword ont été transformé en text avec une propriété keyword.

Une methode plus rapide :

- pour les petit dataset, reupload la data et ajoutant keyword au field DBA
- ajouter un field DBA_keyword a notre index de base, mais ce serais moins optimisé car les deux champs sont stocké et indexé differement.

2.9 Identify the cuisine most affected by the violation “Hot food item not held at or above 140° F”

Pour cette requete, on peux utiliser match phrase pour trouver l'exacte phrase (non case insensitive) ou match, mais cela requière d'ajuster la fuzziness pour ne pas avoir des resultat avec une description completement differente .

requète:

```
GET restaurantny/_search
{
  "size": 0,
  "query": {
    "match_phrase": {
      "VIOLATION DESCRIPTION": "Hot food item not held at or above 140°
F"
    }
  },
  "aggs": {
    "most_affected_cuisine": {
      "terms": {
        "field": "CUISINE DESCRIPTION",
        "size": 10,
        "order": { "_count": "desc" }
      }
    }
  }
}
```

resultat

```
 "aggregations": {
    "most_affected_cuisine": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 401,
      "buckets": [
        {
          "key": "Pizza",
          "doc_count": 187
        },
        {
          "key": "American",
          "doc_count": 179
        },
        {
          "key": "Chinese",
          "doc_count": 174
        },
        {
          "key": "Latin American",
          "doc_count": 84
        },
```

PROF

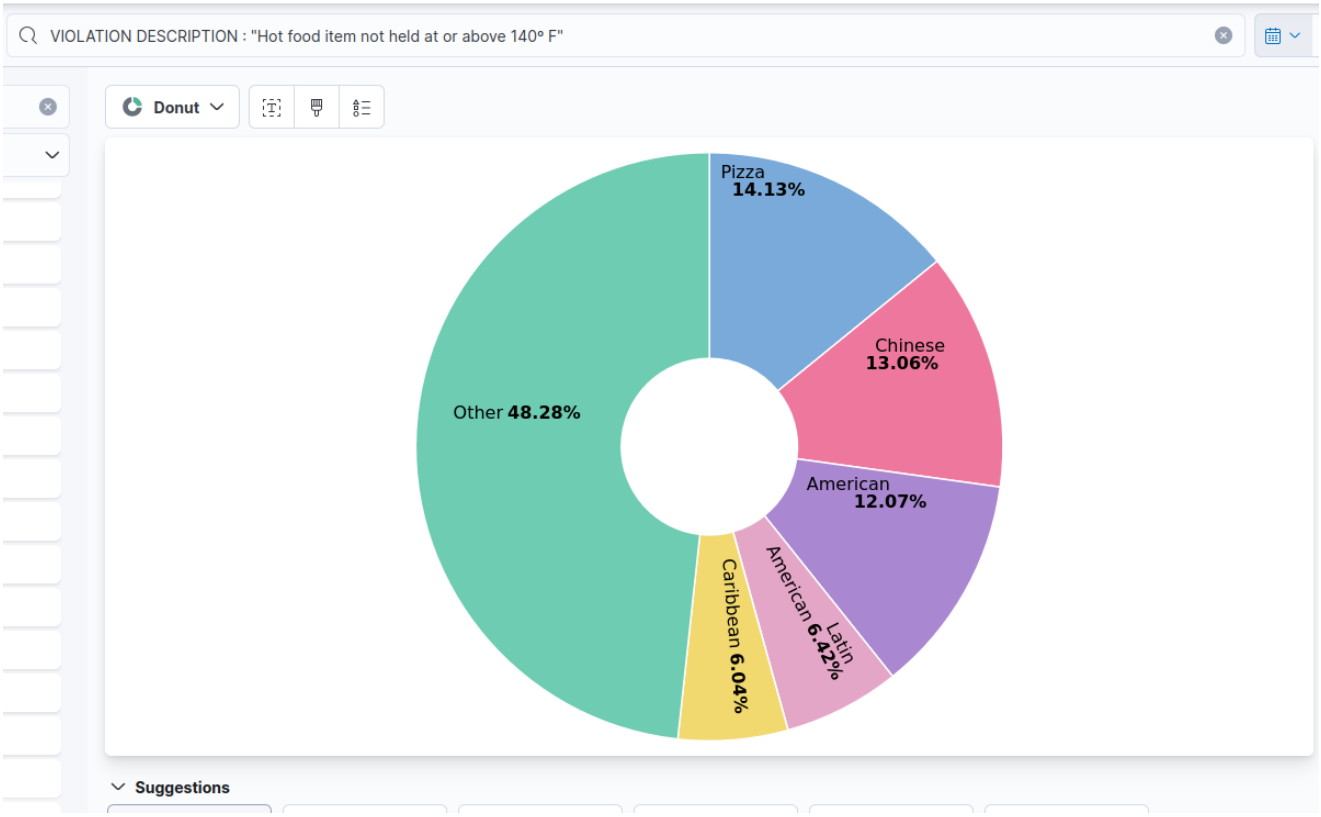
```
          {
            "key": "Caribbean",
            "doc_count": 80
          },
          {
            "key": "Japanese",
            "doc_count": 60
          },
          {
            "key": "Mexican",
            "doc_count": 49
          },
          {
            "key": "Bakery Products/Desserts",
            "doc_count": 48
          },
          {
            "key": "Italian",
            "doc_count": 47
          },
          {
            "key": "Coffee/Tea",
            "doc_count": 37
          }
        ]
      }
    }
```


La cuisine la plus affecté par cette violation sont les pizzerias

visualisation

Pie Chart avec le filtre

PROF



2.10 Determine the most common violations (Top 5)

On aggrege sur les violation code et on affiche la violation description avec une size de 5

```
  GET restaurantny/_search
{
  "size": 0,
  "aggs": {
    "most_affected_cuisine": {
      "terms": {
        "field": "VIOLATION CODE",
        "size": 5,
        "order": { "_count": "desc" }
      },
      "aggs": {
        "sample_description": {
          "top_hits": {
            "_source": ["VIOLATION DESCRIPTION"],
            "size": 1
          }
        }
      }
    }
  }
}
```

PROF

resultat

```
"aggregations": {
    "most_affected_cuisine": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 161778,
      "buckets": [
        {
          "key": "10F",
          "doc_count": 40108,
          "sample_description": {
            "hits": {
              "total": {
                "value": 40108,
                "relation": "eq"
              },
              "max_score": 1,
              "hits": [
                {
                  "_index": "restaurantny",
                  "_id": "qROH3JkByv84jpscZM78",
                  "_score": 1,
                  "_source": {
                    "VIOLATION DESCRIPTION": "Non-food contact surface
or equipment made of unacceptable material, not kept clean, or not
properly sealed, raised, spaced or movable to allow accessibility for
cleaning on all sides, above and underneath the unit."
                  }
                }
              ]
            }
          }
        },
        {
          "key": "08A",
          "doc_count": 27330,
          "sample_description": {
            "hits": {
              "total": {
                "value": 27330,
                "relation": "eq"
              },
              "max_score": 1,
              "hits": [
                {
                  "_index": "restaurantny",
                  "_id": "sROH3JkByv84jpscZM78",
                  "_score": 1,
                  "_source": {
                    "VIOLATION DESCRIPTION": "Establishment is not free
of harborage or conditions conducive to rodents, insects or other
pests."
                  }
                }
```

```
              ]
            }
          }
        },
        {
          "key": "06D",
          "doc_count": 18552,
          "sample_description": {
            "hits": {
              "total": {
                "value": 18552,
                "relation": "eq"
              },
              "max_score": 1,
              "hits": [
                {
                  "_index": "restaurantny",
                  "_id": "qxOH3JkByv84jpscZM78",
                  "_score": 1,
                  "_source": {
                    "VIOLATION DESCRIPTION": "Food contact surface not
properly washed, rinsed and sanitized after each use and following any
activity when contamination may have occurred."
                  }
                }
              ]
            }
          }
        },
        {
          "key": "02G",
          "doc_count": 18056,
          "sample_description": {
            "hits": {
              "total": {
                "value": 18056,
                "relation": "eq"
              },
              "max_score": 1,
              "hits": [
                {
                  "_index": "restaurantny",
                  "_id": "xhOH3JkByv84jpscZM78",
                  "_score": 1,
                  "_source": {
                    "VIOLATION DESCRIPTION": "Cold TCS food item held
above 41 ˚F; smoked or processed fish held above 38 ˚F; intact raw eggs
held above 45 ˚F; or reduced oxygen packaged (ROP) TCS foods held above
required temperatures except during active necessary preparation."
                  }
                }
              ]
            }
```

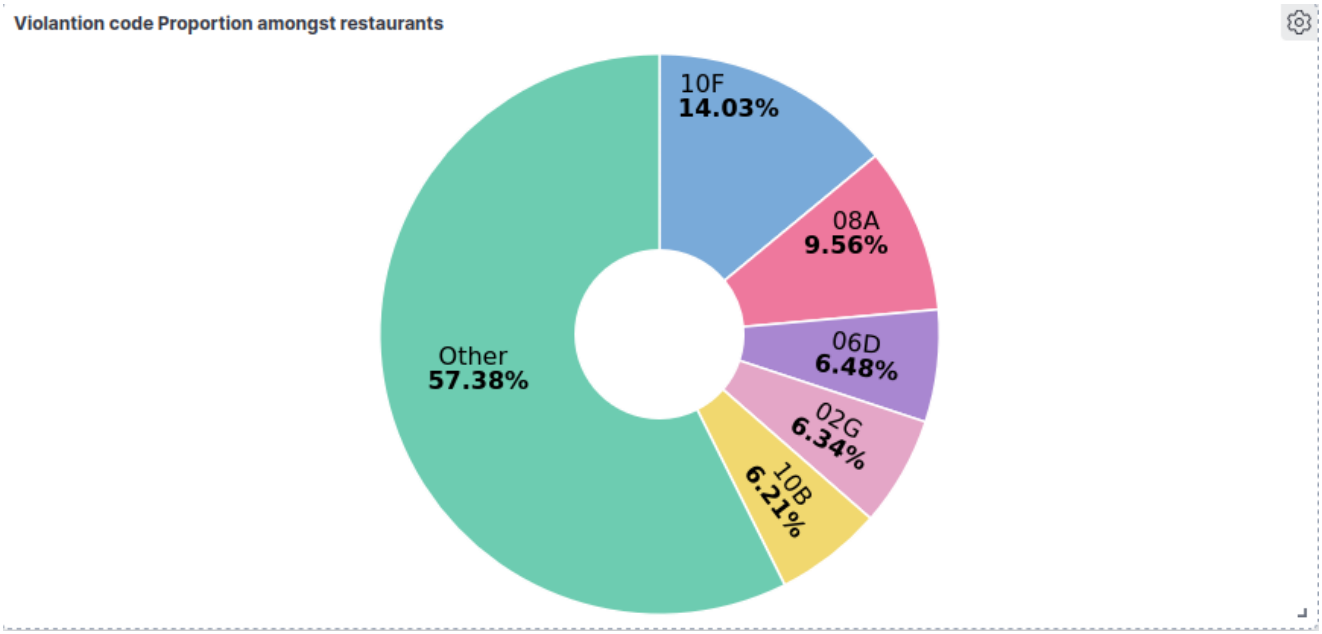
```
          }
        },
        {
          "key": "10B",
          "doc_count": 17746,
          "sample_description": {
            "hits": {
              "total": {
                "value": 17746,
                "relation": "eq"
              },
              "max_score": 1,
              "hits": [
                {
                  "_index": "restaurantny",
                  "_id": "shOH3JkByv84jpscZM78",
                  "_score": 1,
                  "_source": {
                    "VIOLATION DESCRIPTION": "Anti-siphonage or back-
flow prevention device not provided where required; equipment or floor
not properly drained; sewage disposal system in disrepair or not
functioning properly. Condensation or liquid waste improperly disposed
of."
                  }
                }
              ]
            }
          }
        }
```


le top 5 des violations :

Rank	Violation Code	Description	Number of Occurrences
1	**10F**	Non-food contact surface or equipment made of unacceptable material, not kept clean, or not properly sealed, raised, spaced, or movable to allow accessibility for cleaning on all sides, above and underneath the unit.	**40 108**
2	**08A**	Establishment is not free of harborage or conditions conducive to rodents, insects, or other pests.	**27 330**
3	**06D**	Food contact surface not properly washed, rinsed, and sanitized after each use and following any activity when contamination may have occurred.	**18 552**
4	**02G**	Cold TCS food item held above 41 °F; smoked or processed fish held above 38 °F; intact raw eggs held above 45 °F; or reduced-oxygen packaged (ROP) TCS foods held above required temperatures except during active necessary preparation.	**18 056**

Rank	Violation Code	Description	Number of Occurrences
5	**10B**	Anti-siphonage or back-flow prevention device not provided where required; equipment or floor not properly drained; sewage disposal system in disrepair or not functioning properly; condensation or liquid waste improperly disposed of.	**17 746**

visualisation



2.11 Identify the most popular restaurant chain

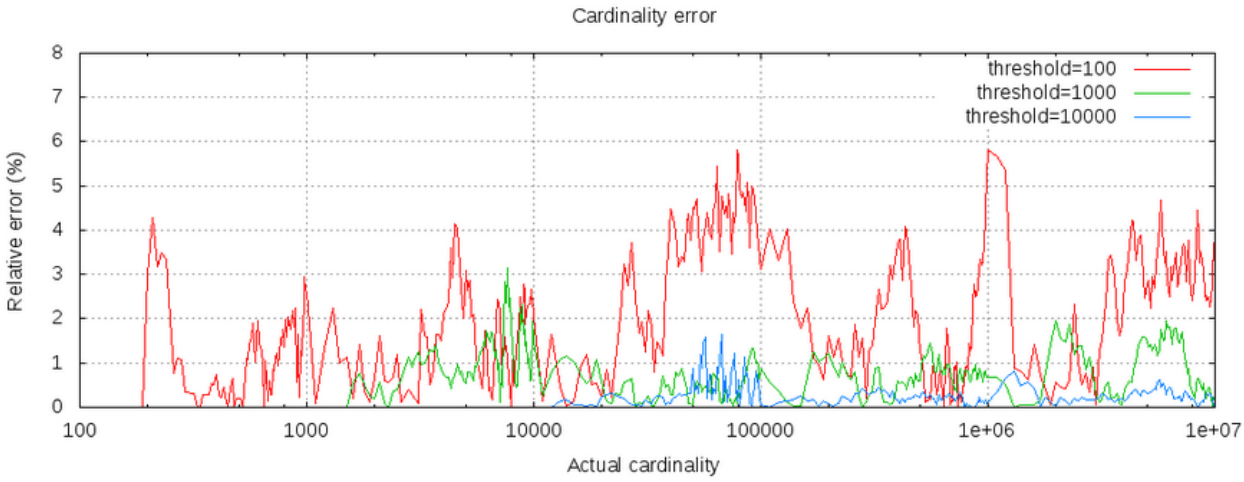
Je doit faire une aggregation sur les DBA des restaurant unique (cardinalité sur CAMIS).

pour la derniere requète, ma memoire n'est pas suffisante, je doit optimiser ma requete, plusieur possibilité s'ouvre a moi

PROF

- Effectuer la requete sur un sample
- reduire la precision de Cardinality
- Effectuer la requete en batch (a la main ou via un script) pour avoir l'exact resultat.

Pour le projet, j'ai choisis de reduire le seuil de precision de la cardinnalité de 3000 a 1000, en effet la cardinnalité dans elastic search est basé sur l'algorithme Hyperloglog++ le comptage est approximatif. On peux reduire la precision pour recuperer de la puissance de calcul (https://www.elastic.co/docs/reference/aggregations/search-aggregations-metrics-cardinality-aggregation)



Pour verifier la pertinence du resultat, On verifie le nombre de document compté avec une aggregation classique

```
GET restaurantny_final/_search
{
  "size": 0,
"aggs": {
    "restaurants_by_DBA": {
      "terms": {
        "field": "DBA.keyword",
        "size": 3,
        "order": { "_count": "desc" }
      }}}

}
```

reponse

PROF

```
"aggregations": {
    "restaurants_by_DBA": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 282779,
      "buckets": [
        {
          "key": "DUNKIN",
          "doc_count": 3241
        },
        {
          "key": "SUBWAY",
          "doc_count": 2003
        },
        {
          "key": "STARBUCKS",
          "doc_count": 1547
        }
```

```
        ]
      }
    }
```

On compte 3241 document pour DUNKIN, 2003 pour SUBWAY et 1547 pour STARBUCKS. Il est probable qu'on retrouve le même top 3 pour les restaurant unique en prenant en compte la correlation entre le nombre d'inspection et le nombre de restaurant.

requete avec cardinalité sur CAMIS et seuil de precision a 1500:

```
GET restaurantny_final/_search
{
  "size": 0,

  "aggs": {
    "restaurants_by_DBA": {
      "terms": {
        "field": "DBA.keyword",
        "size": 3,
        "order": { "unique_restaurants.value": "desc" }
      },
      "aggs": {
        "unique_restaurants": {
          "cardinality": {
            "field": "CAMIS",
              "precision_threshold": 1500
          }
        }
      }
    }
  }
}
```

PROF

reponse

```
"aggregations": {
    "restaurants_by_DBA": {
      "doc_count_error_upper_bound": -1,
      "sum_other_doc_count": 282779,
      "buckets": [
        {
          "key": "DUNKIN",
          "doc_count": 3241,
          "unique_restaurants": {
            "value": 348
          }
        },
        {
```

```
            "key": "STARBUCKS",
            "doc_count": 1547,
            "unique_restaurants": {
              "value": 209
            }
          },
          {
            "key": "SUBWAY",
            "doc_count": 2003,
            "unique_restaurants": {
              "value": 176
            }
          }
        ]
      }
    }
```


On obtient le même nombre de document, le seuil de precision a 1500 est donc assez precis, avec en top 3.

- Dukin : 348 restaurant
- STARBUCKS : 209 restaurant
- SUBWAY : 176 restaurant

donc DUKIN est le restaurant le plus populaire.

visualisation

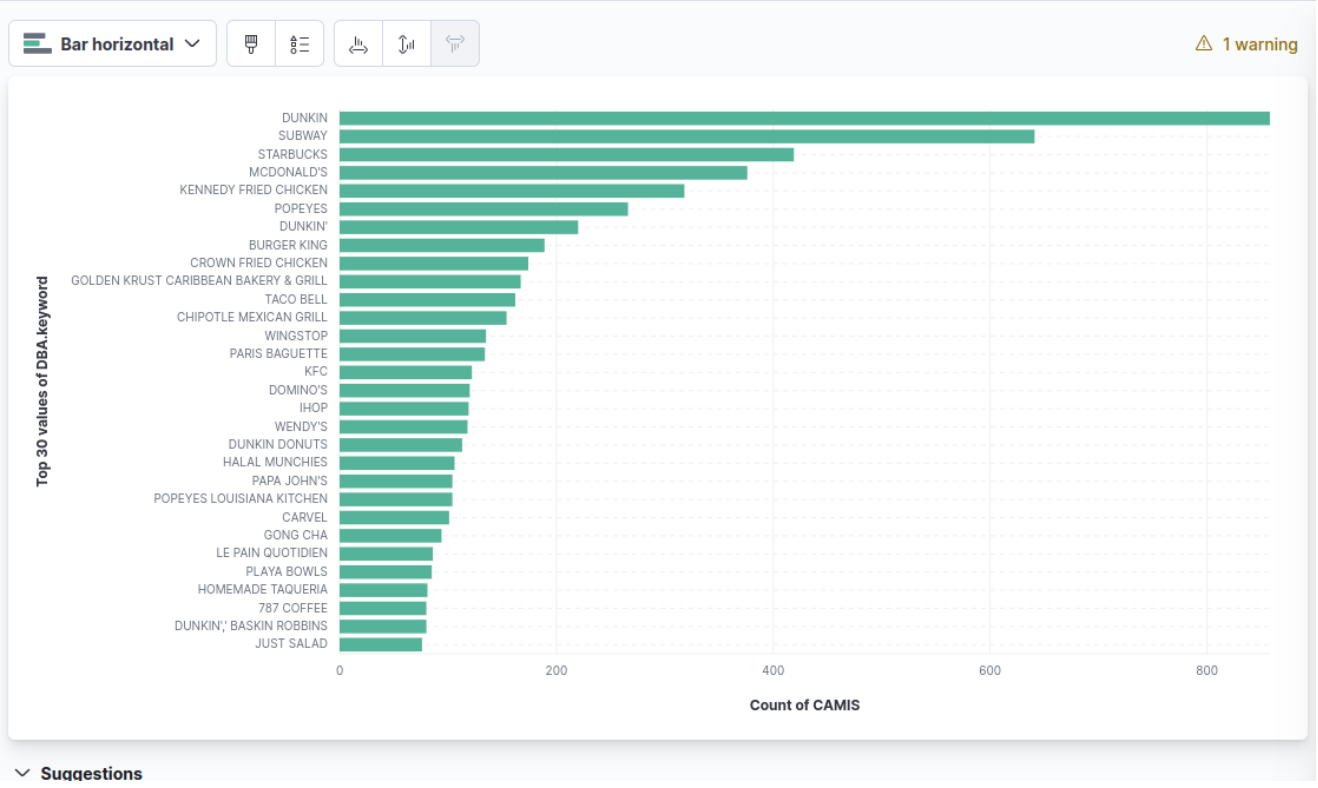
On peux faire une table

Table ⚠ 1 warning

Top 100 values of DBA.keyword	Count of CAMIS
DUNKIN	8
SUBWAY	641
STARBUCKS	419
MCDONALD'S	376
KENNEDY FRIED CHICKEN	318
POPEYES	266
DUNKIN'	220
BURGER KING	189
CROWN FRIED CHICKEN	174
GOLDEN KRUST CARIBBEAN BAKERY & GRILL	167
TACO BELL	162
CHIPOTLE MEXICAN GRILL	154
WINGSTOP	135
PARIS BAGUETTE	134

Suggestions

ou un bar chart (en enlevant "other" pour la lisibilité)



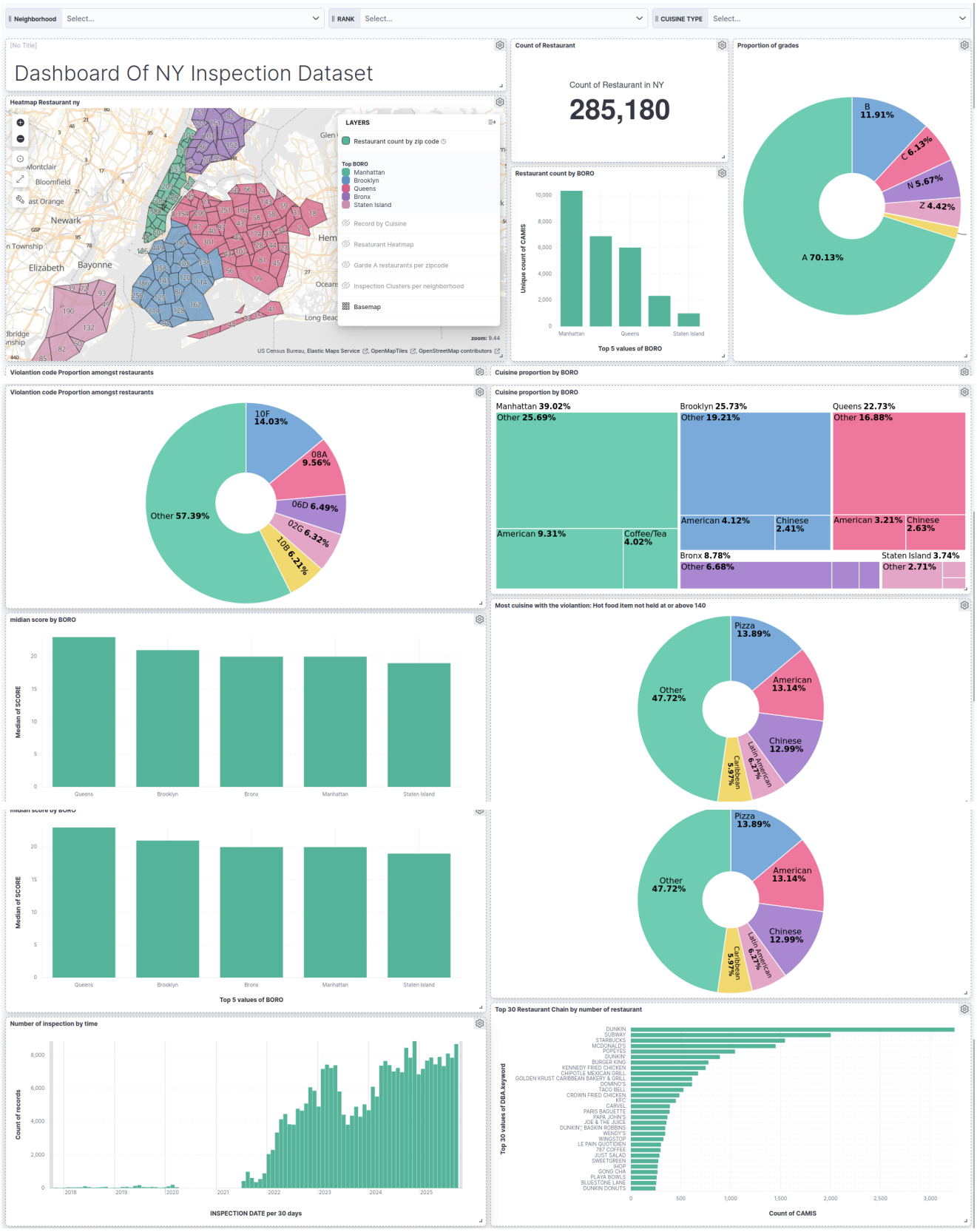
On remarque DUNKIN et DUNKIN', on pourrait refaire l'index pour enlever les caractère speciaux, mais cela pourrais aussi aggreger des restaurant different, nous estimont que notre degres de precision est suffisant.

PROF

Dashboard

Visualisez le dashboard regroupant les visualisation ici :

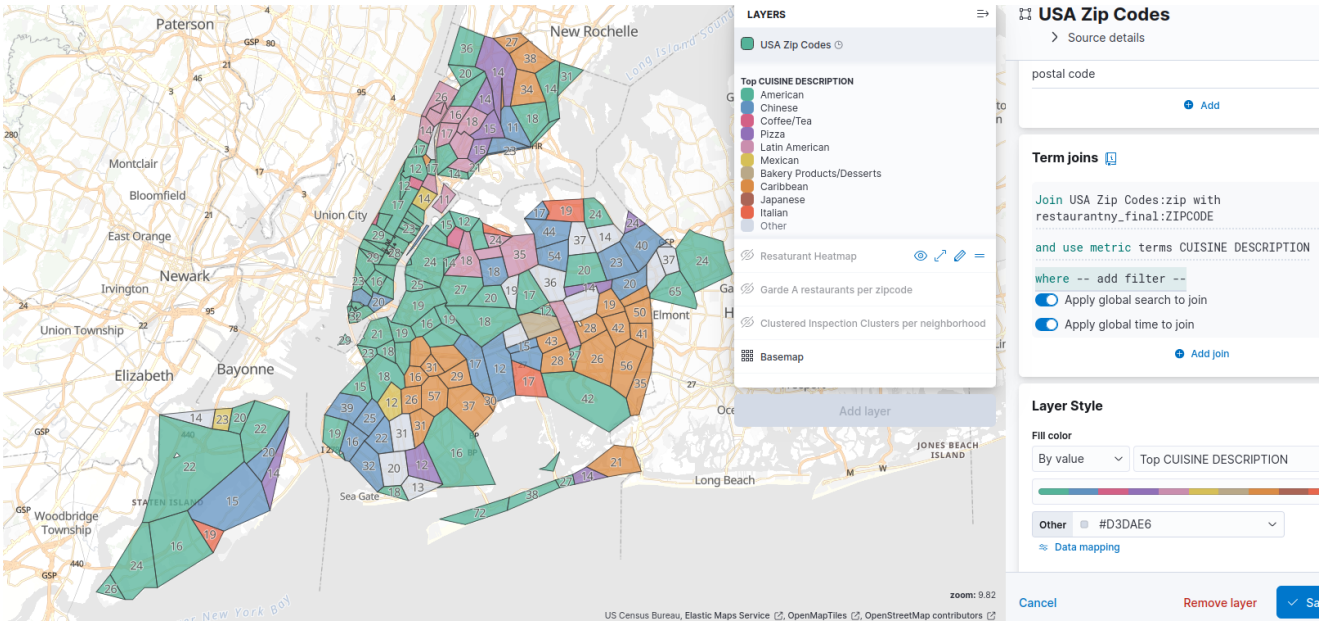
http://localhost:5601/app/r/s/6ns45



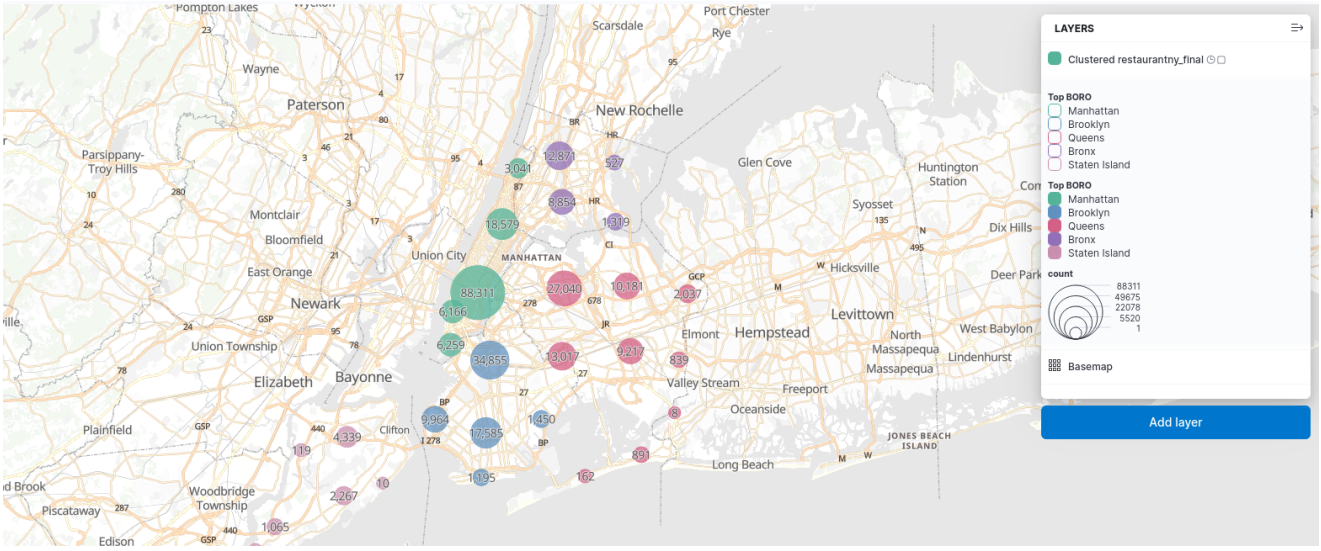
PROF

Visualisation with MAP

Nombre de document par zipcode en fonction du type de cuisine

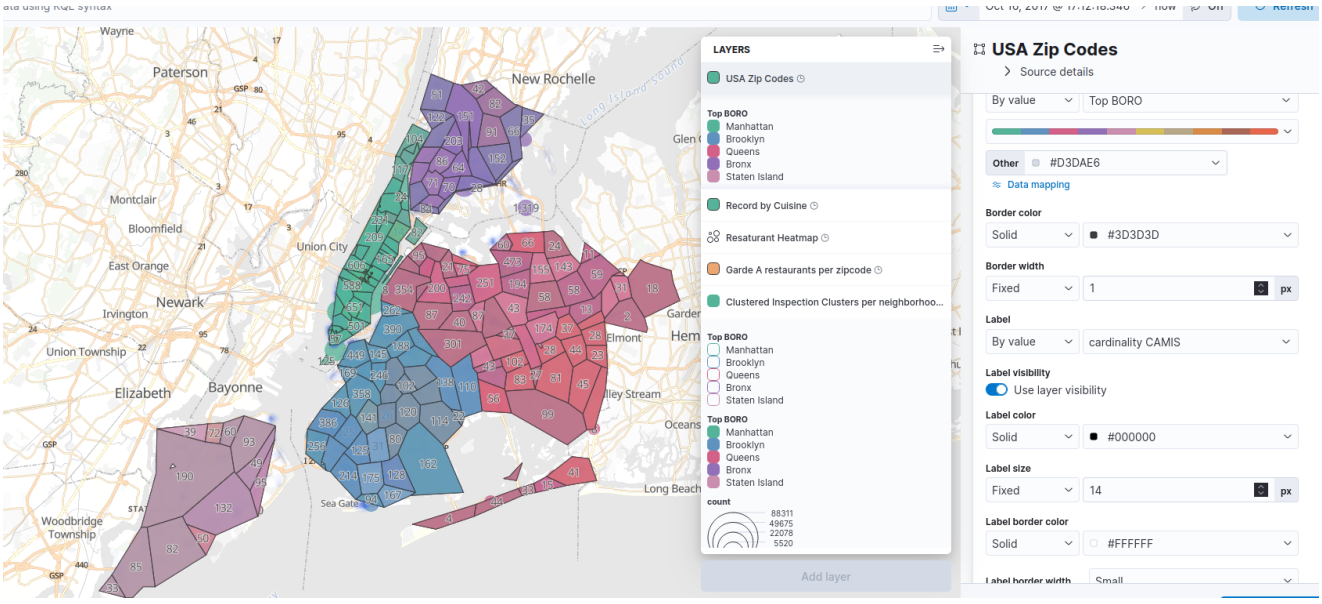


Cluster des inspection par Voisinage

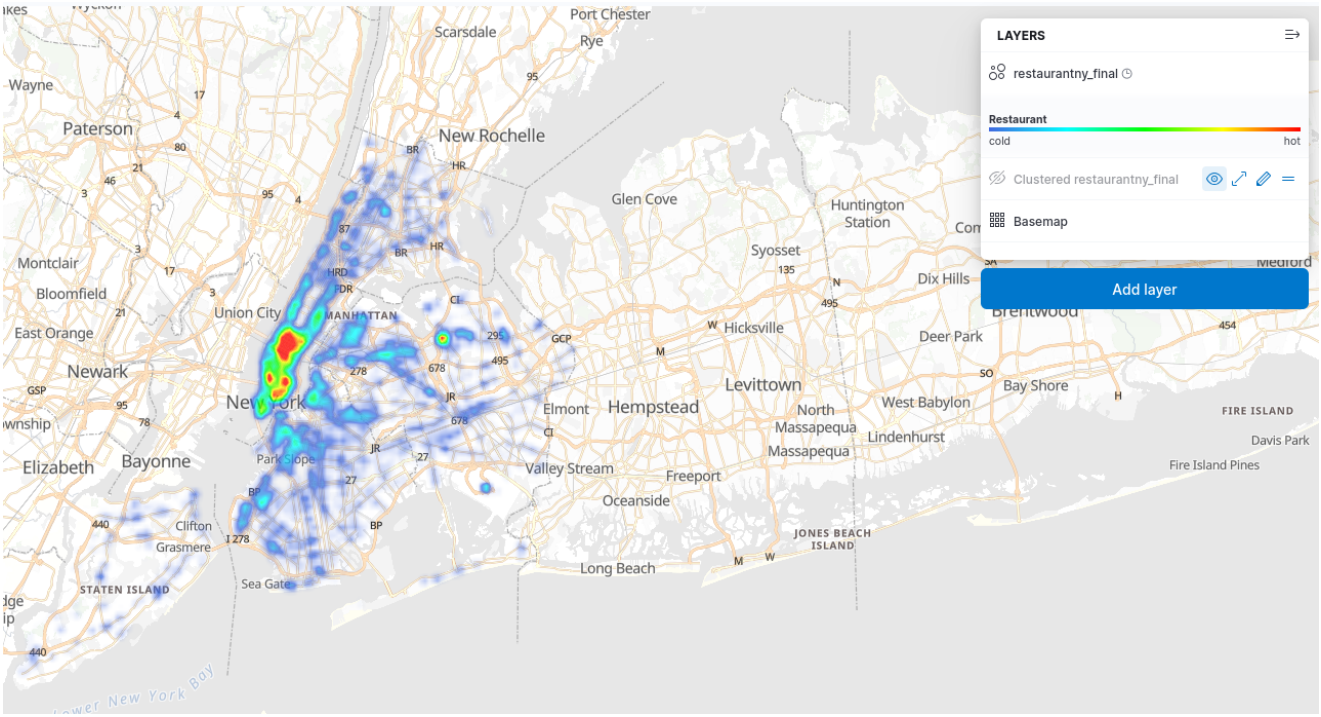


PROF

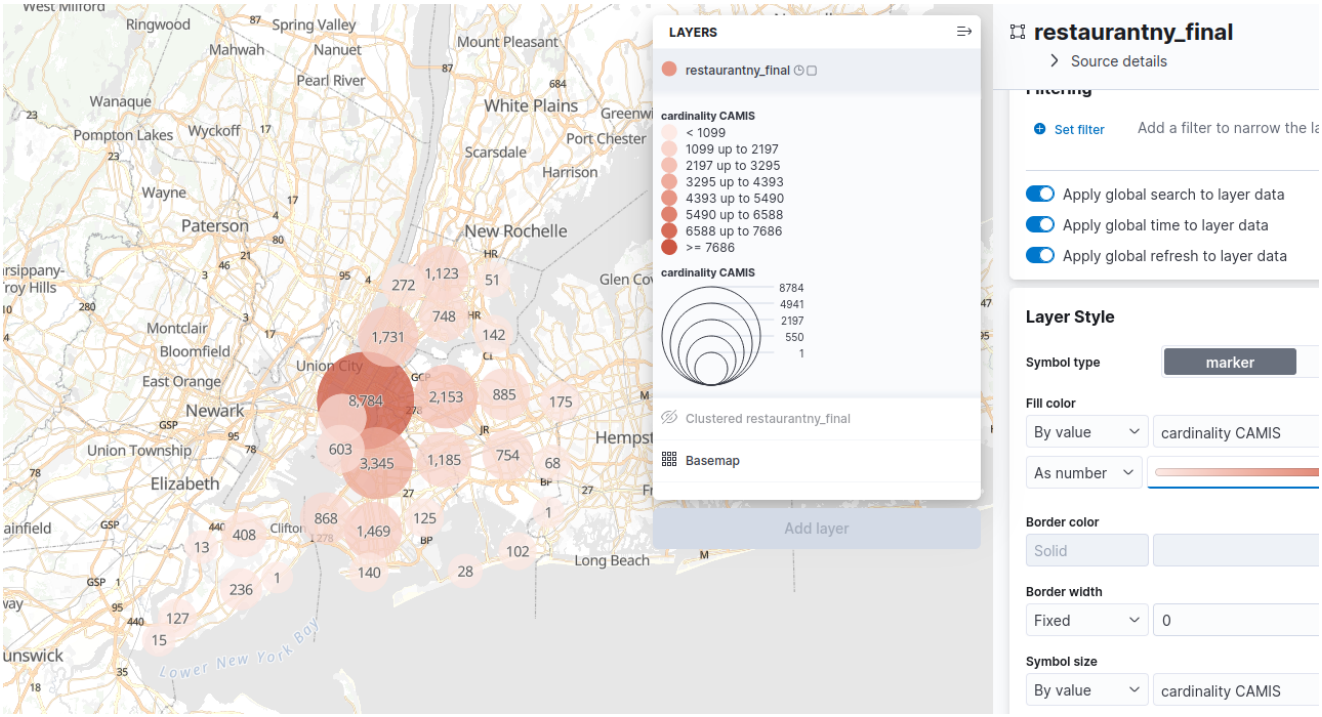
restaurant par voisinage



Heatmap of Unique Restaurant

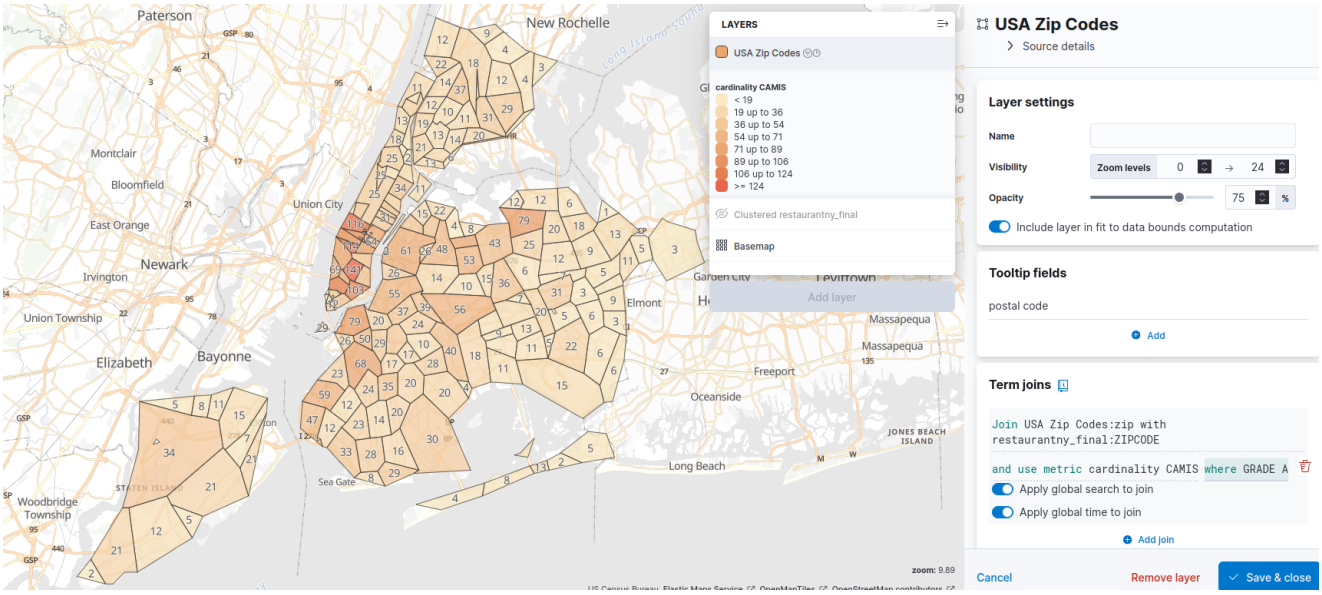


Cluster of restaurant



PROF

Restaurant A la note de A par zipcode


Paterson
New Rochelle
Montclair
Bloomfield
Union City
East Orange
Newark
Irvington
Union Township
Elizabeth
Bayonne
Woodbridge Township
Sea Gate
Elmont
Massapequa
Massapequa
Freeport
Oceanside
Long Beach
JONES BEACH ISLAND
LAYERS
USA Zip Codes
cardinality CAMIS
< 19
19 up to 36
36 up to 54
54 up to 71
71 up to 89
89 up to 106
106 up to 124
>= 124
Clustered restaurantny_final
Basemap
Add layer
USA Zip Codes
Source details
Layer settings
Name
Visibility
Zoom levels
0
24
Opacity
75
%
Include layer in fit to data bounds computation
Tooltip fields
postal code
Add
Term joins
Join USA Zip Codes:zip with restaurantny_final:ZIPCODE
and use metric cardinality CAMIS where GRADE A
Apply global search to join
Apply global time to join
Add join
Cancel
Remove layer
Save & close
zoom: 9.89