Datasheets for Datasets

# Motivation for Dataset Creation

## **Why was the dataset created?** (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

The dataset was created to enable research on grading restaurants in NYC based on the compliance in food handling, food temperature, personal hygiene and vermin control.

## **What (other) tasks could the dataset be used for?** Are there obvious tasks for which it should *not* be used?

The dataset can be used for restaurants recommendation based on consumers’ preferences. For example, if a customer is recently addicted to Thai cuisine, then some of the top graded Thai restaurants could be pushed and recommended to him/her.

It can also be used to provide suggestions for restaurants regarding its violation description, which contains the details about the violation content.

It can also be used to optimize the inspection resources based on the score in each cuisine type and borough.

## **Has the dataset been used for any tasks already?** If so, where are the results so others can compare (e.g., links to published papers)?

In such a big data time, food industry is adopting big data technology and big data analytics to stay competitive by understanding customer preferences and tastes.

[How Food Delivery Apps are leveraging Big Data Analytics?](https://www.dezyre.com/article/how-food-delivery-apps-are-leveraging-big-data-analytics/197)

## **Who funded the creation of the dataset?** If there is an associated grant, provide the grant number.

The New York City Health Department

# Dataset Composition

## **What are the instances?** (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

The instances are the check for compliance in food handling, food temperature, personal hygiene and vermin control. Each violation of a regulation gets a certain number of points. At the end of the inspection, the inspector totals the points, and this number is the restaurant's inspection score—the lower the score, the better the grade.

## **How many instances of each type are there?**

There are totally 379350 rows of restaurants in the dataset, with each row containing information about the business name, borough it located, building number, street address and so on.

On each column, the datatype is consistent

## **What data does each instance consist of?** “Raw” data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

CAMIS ID number(integer), Restaurant name(string), borough(string), building number(integer), street(string), zip code(integer), business phone(integer), cuisine description(string), inspection date(time series), action(string), violation code(string), violation description(string), critical flag(string), inspection score(integer), restaurant grade(empty or string), grade and record date(time series), inspection type(string).

## **Is everything included or does the data rely on external resources?** (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with *any* of the data?

All included

# Data Collection Process

## **How was the data collected?** (e.g., hardware apparatus/sensor, manual human curation, software pro- gram, software interface/API; how were these con- structs/measures/methods validated?)

The data was collected by inspectors through unannounced inspection (to ensure fairness). The inspection includes cycle inspection/re-inspection, cycle inspection/initial inspection, pre-permit(non-operational)/initial inspection, pre-permit(operational)/initial inspection, administrative miscellaneous/ initial inspection. After the inspection process, there would be an inspection score given by the inspector and finally a restaurant grade was drawn according to that score.

## **Who was involved in the data collection process?** (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

The New York City Health Department, the inspectors from the department, restaurants (including the café and bakery store) that are in all boroughs through New York city.

## **Over what time-frame was the data collected?**

6/13/2017--10/10/2018

## **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

Each instance contains the basic information about each restaurant like the restaurant’s name, address, cuisine type and its violation description in detail. But when It comes to violation code, it’s hard to figure out the meaning of violation code because it’s just a simple version of the violation description. I may want to delete that column. Also, the “inspection type” doesn’t contribute much to my research since it is related with the type inspectors inspecting the restaurants.

## **Does the dataset contain all possible instances?** Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

The dataset is only an example of New York city’s restaurants and my conclusion or the recommendation advice from this dataset can only be applied to people who consider going to a restaurant in New York city.

## **If the dataset is a sample, then what is the population?**

## What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of in- stances)? How does this affect possible uses?

The population would be the current existing restaurants in New York City, it can bot be used to represent the larger set (national restaurant rating maybe) since nationally the restaurants varieties and cuisine types would be larger. And due to geographic element, people in different regions have different food preferences, which would influence the number of restaurant in each kind of cuisine.

## **Is there information missing from the dataset and why?**

## (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

There are some missing values in the restaurant grade and grade date column. However, the grade date is not so important since it’s a yearly updated dataset. But the missing values regarding the restaurant grade will affect my analysis for the best or worst restaurant in certain area. So, I decide to filter these restaurants with grade score to continue my project.

# Data Preprocessing

## **What preprocessing/cleaning was done?** (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)

I would firstly delete these restaurants with no inspection score since these data is not helpful for me to find the best or worst one based on its inspection score. For the missing values in grade date, it’s not a big influent element to my research since this data is yearly updated. Then I would drop columns such as the building number, street address which is the basic geographic information.

## **Was the “raw” data saved in addition to the preprocessed/cleaned data?** (e.g., to support unanticipated future uses)

Yes

# Dataset Distribution

## **How is the dataset distributed?** (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

API, website. It is also been cited and utilized in some other dataset or research.

## **When will the dataset be released/first distributed?** (Is there a canonical paper/reference for this dataset?)

The dataset is distributed on enigma see [link](https://public.enigma.com/datasets/new-york-city-new-york-restaurant-inspections/fa7ab996-fb43-4e86-80e7-f8e82ccba15f) on 10/10/2018

## **What license (if any) is it distributed under?** Are there any copyrights on the data?

The copyrights belong to The New York City Health Department

# Dataset Maintenance

## **Will the dataset be updated?** How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum?

The dataset is updated annually, the new dataset would be updated after another inspection process been conducted.

## **Is there a repository to link to any/all papers/systems that use this dataset?**

KEY USE CASES:

1. Contrast restaurant grades across New York City boroughs
2. Compare how restaurant grades have changed over time
3. Investigate restaurant grades across different cuisines

## **If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?** If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

It is available to be downloaded as csv file and can be analyzed.

# Legal and Ethical Considerations

**If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?** (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

According to the dataset, we can only imply that the data are collected by inspectors from the New York City Health Department to help grade the restaurant based on the violation these restaurants may have. It’s obvious that the restaurant owner will definitely not be informed since this is an unannounced inspection.

**If it relates to other ethically protected subjects, have appropriate obligations been met?** (e.g., medical data might include information collected from animals)

I don’t think so, since it’s only related with the violation that restaurant may have.