# Template Data Quality Report

## Overview

This report will outline the initial data quality findings on amenities data obtained from the [overpass turbo](https://overpass-turbo.eu/) API for the New York City area. This report will include an overview of the dataset, and a review of the continuous and categorical features, including histograms and bar charts. On initial review, this dataset contains a lot of missing data for most features. The data that is present appears to be reasonable and logical, however a number of columns will need to be dropped.

## Summary

This dataset consists of information on different amenities in New York City, including bars, restaurants, and cafés. There are significant amounts of missing values and most features will need to be dropped. There are no duplicate rows. Distribution of the data is consistent with expectations. While the missing data is not ideal, this data is still suitable for use as place data.

## Review Logical Integrity

Test 1: No date in x is before Feb 2022 or after the upload date (22nd May).

* 0 instances.

## Review Continuous Features

### Descriptive Statistics

There are 3 continuous features in this dataset:

* ID
  + This is the Identification number of each place in OSM. Each row has a unique value. There are no missing values.
* Latitude
  + This is the latitude of each place in OSM. Each row has a unique value. There are no missing values. Latitude is centred around 40.72, roughly the centre of NYC.
* Longitude
  + This is the longitude of each place in OSM. Each row has a unique value. There are no missing values. Longitude is centred around -73.95, roughly the center of NYC.

### Histograms

All Histograms can be found in the appendix as a summary sheet. All features show a plausible distribution.

## Review Categorical Features

There are 43 categorical features in the dataset. For brevity, only columns of interest will be listed. See the appendix for full descriptive statistics on all features.

* City
  + This has 71 unique values. The most common is New York. 66% of values are missing.
* House Number
  + This has 4269 unique values. The most common is 2. 34% of values are missing.
* State
  + This has 2 unique values. The most common is NY. The second most common (which needs to be removed) is NJ. 69% of values are missing, however these can all be imputed as NY.
* Amenity
  + The type of amenity. This has 24 unique values. The most common is restaurant. There are no missing values.
* Cuisine
  + This has 1303 unique values. The most common is pizza. 33% of values are missing.
* Name
  + This has 11435 unique values. 3% of values are missing.
* Opening Hours
  + This has 4405 unique values. The most common is 24/7. 53% of values are missing.
* Phone
  + This has 7598 unique values. There are some duplicate values, likely for large chains.
* Website
  + This has 6647 unique values. 54% of values are missing.
* Wheelchair
  + This has 6 unique values. The most common is yes. 87% of values are missing.
* Vegan
  + This has 4 unique values. The most common is yes. 96% of values are missing.
* Vegetarian
  + This has 4 unique values. The most common is yes. 96% of values are missing.
* Opening Hours
  + This has 4405 unique values. The most common is 24/7. 53% of values are missing.

## Actions to take

x actions will be taken:

* Missing name
  + Drop rows with missing name.
* Vegan/Vegetarian
  + Change “only” and “limited” values to “yes”.
* State
  + Drop rows with state as “NJ”
* Wheelchair
  + Change “designated” and “dedicated” to “yes”.

# Appendix

## Continuous Features

### Descriptive Statistics

Table with statistics on continuous features.

### Histograms

A graph of a graph

Description automatically generated with medium confidence

## Categorical Features

### Descriptive Statistics

Table with statistics on categorical features.





### Box Plots

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A group of blue and black graphs

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated