Data Wrangling
Predicting Critical Health and Safety Violations at Food Establishments

The first step was collecting the data from two different sources. The restaurant inspections results were obtained from the Nevada Health District Restaurant Inspections webpage, using their API. The Yelp academic dataset files were downloaded directly from Yelp. All the JSON files were then normalized to manipulate them using Pandas and some initial processing was done before merging them.

For the restaurant inspections results dataset, which had over 160,000 records and 24 attributes, a subset was taken to include:
Address, category_name, city, current_demeritis[1], current_grade, inspection_date, inspection_time, inspection_demerits, inspection_grade, inspection_result, violations, inspection_type, location_coordiantes, location_name, permit_status, restaurant_name and zip code.

The main data cleaning steps included:

- removing trailing and leading spaces from column names and values in the category name and restaurant name attributes.
- removing special characters from the restaurant name values, as this is one of the keys used to merge with the Yelp data.
- splitting zip codes to leave only the first five digits (e.g. 89103-4004 became 89103)
- splitting the date to create a year and a month column to use as keys during merging
- creating a binary target variable called 'is_compliant', with 1, if the inspections results were 'compliant', 'A grade', 'approved', or 'no further action' and 0 otherwise.

For the Yelp data, only two files were used, the business data and reviews data. For the business file, which had over 180,000 records, the main data cleaning steps included:

- filtering the data to include only records for the state of Nevada
- selecting business id, name, location, review count, categories, attributes, and zip code.
- performing the same first four steps done to the inspection results dataset.

For the reviews file, which had over 5 million records, the main cleaning steps included:

- selecting only the date, rating and business id attributes. The review text was not included for this analysis.
- grouping the data by date and business id and get the daily rating average for each business before merging it with the business data. This was done to mitigate

[1] As noted in previous report ('Predicting Critical Health and Safety Violations at Food Establishments – Proposal'), each violation is assigned a value, demerit, that results in a grade. 'A's are given to establishments that are compliant (0 – 10 demerits). 'B's (11 – 20) and 'C's (21-40) are downgrades indicating critical or major violations.

'unnecessary' data duplication when merging the Yelp file with the inspections file, since the text was not considered.

After the pre-processing, the datasets were merged to include only values that were present in both data sets, using name, year, month, and zip code as keys. Using the year and month, ensured the ratings and business attributes were applicable.

Once the merge was done, some additional data processing was done:
Feature engineering:

- since the 'category_names' and the 'categories' attributes were categorical with multiple levels, similar levels were combined to create separate binary attributes, such as serves_alcohol, or to identify the cuisine type, japanese, mexican, thai, etc.
- A column with the number of days between inspection and rating was also created.

Missing values:

- Since the merged dataset has over 35,000 values, four attributes with over 20,000 missing values were removed.
- For the three attributes that have between 2,000 and 3,500 missing values, a predictive model will be used to fill in the missing values using existing data.

As the project progresses, additional processing will be done.