Finding the safe restaurants in the LA county

Han

May 22, 2020

1. Introduction

1.1 Background and Problem

As a senior preparing for graduate school, I am desired to go to California in the future. I have been to Los Angeles 3 times and I love the sunshine, the beach and various cuisines in restaurants. However, in recent years, food safety has become an increasing concern. Food products may be exposed to various levels of risks in different processes, including processing, packaging, and distribution.

1.2 Interest

Both visitors and locals would be very interested in this project. By analyzing data of around 40,000 restaurants, I will figure out where people can find safe restaurants in LA county.

2. Data Acquisition and Cleaning

2.1 Data sources

A dataset about Restaurant Inspections and violations of the LA County could be found on Kaggle. It contains over 40,000 restaurants in 85 out of 88 cities in the LA county.

The dataset is originally from Los Angeles County Environmental Health and it was last updated two years ago.

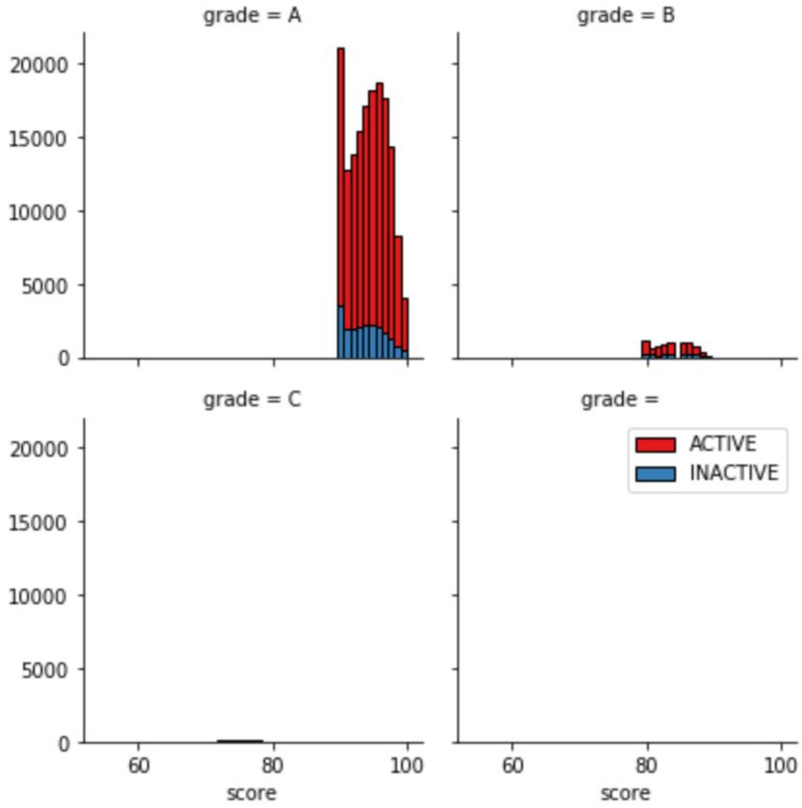
2.2 Data cleaning

Data downloaded from Kanggle and were combined into one table. As there were a lot of description data, I deleted some columns that would not be used. There were also many missing values in scores, grades, facility_zip, which were essential data would be used in the analysis, so I deleted data that has missing values in such columns.

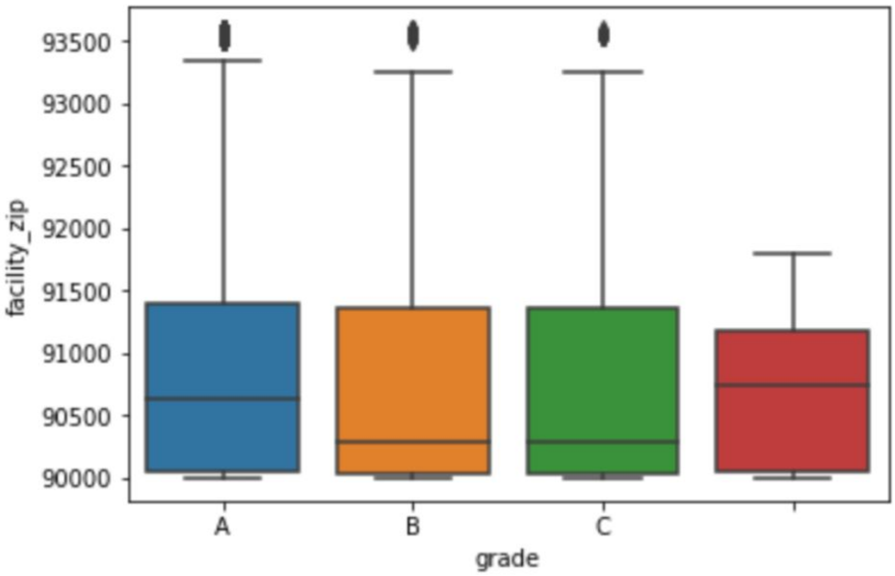
There are few problems with the datasets. One of the most important problems was facility_zip since the zip codes were in different types. Some of the zip codes were too specific, showing like '90005-1234'. In order to keep all the data in the same format, I deleted the second part and only contains the general zip codes. After cleaning, there were 19,1343 samples and 15 features in the data.

3. Exploratory Data Analysis

First of all, I explore the relationship among grade, score and program status. According to the diagram below, we can see restaurants with higher scores and grades tend to maintain active status.

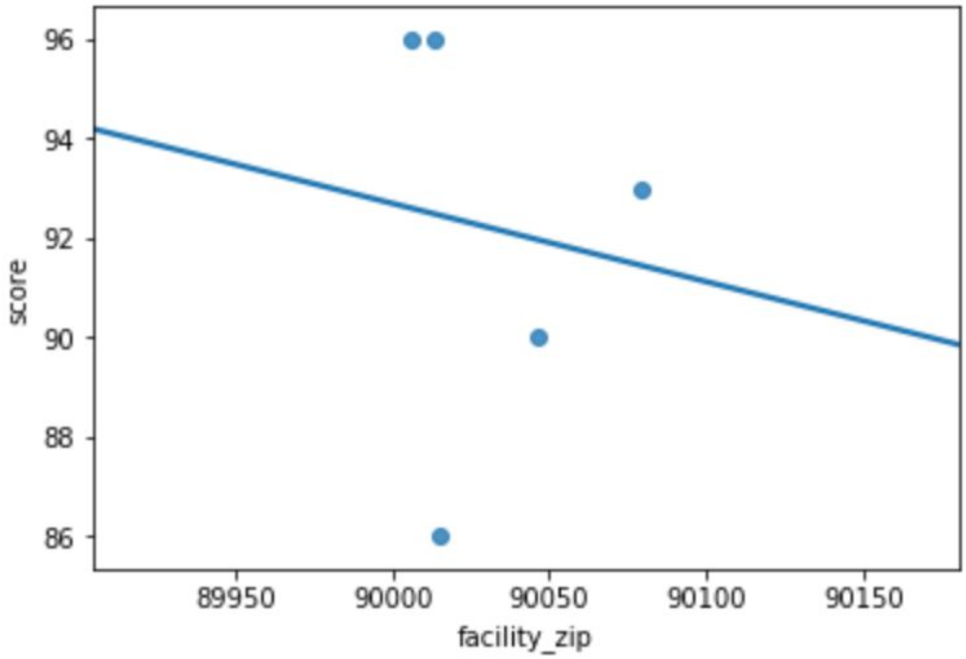


Then I used df.groupby(['grade'])['facility_city'].value_counts(normalize=True) to figure out city with highest grade. It shows in the dataset, Los Angeles and Carson has more than half of high-graded cities. I also made a boxplot between the location of restaurants and the their grades.



As there are 443 cities in total, the data were very decentralized. After comparing the grades, I find Los Angeles has most high-graded restaurants and decide to focus on data in the city of Los Angeles. By making regression of restaurants in Los Angeles, we can see the score of restaurants and the location of restaurants are negatively related. In general, food restaurants located in

89950 is safer.


score
facility_zip