Stat 27850/30850: Group mini-project # 3

Data The data set for the third project is the San Francisco restaurant inspection data set, from
`https://data.sfgov.org/Health-and-Social-Services/Restaurant-Scores-LIVES-Standard/pyih-qa8i`
This data set records every food safety inspection of restaurants in San Francisco, over a time period spanning 2016–2019. For each inspection, the available data includes the restaurant ID / name / location, the date of the inspection, and the outcome (e.g., 'Low Risk', what violations were found, etc).

Run the provided script `get_restaurant_data.R` to download the data. This script also creates a clean subset of the data, the data frame `data1`, with the following variables:

- `month`, `day`, `year` record the date of the inspection
- `business_id` is the unique ID identifying the restaurant
- `risk_category` is one of 'None', 'Low Risk', 'Moderate Risk', 'High Risk'
- `inspection_score` is an integer score (100 is the best score)
- `zipcode` is the zip code of the restaurant, and `latitude` and `longitude` give the coordinates of the location, in degrees (note: at San Francisco's latitude, one degree of latitude corresponds to approximately 69 miles, and one degree of longitude corresponds to approximately 55 miles).

A few notes about the data set:

- This subset (`data1`) includes only random/"routine" inspections—we have removed other types of inspections, e.g., inspections of a new facility / followup after a violation was found / etc. We also did not include variables like the description of the violation, etc. You are not restricted to using only `data1`, but you may prefer to use this simpler/cleaner data in the interest of time.
- Due to missing values etc, `data1` contains only about 17% of the original data set—this may introduce extremely high bias into the data set, but for the purpose of this project we will ignore this issue.
- If multiple violations were found at a single inspection, then this inspection appears as multiple entries—in other words, you will see two or more rows in `data1` that have the same ID and the same date. The `inspection_score` is the same across all of these rows. The `risk_category` may be the same (e.g., two low risk violations were found at a single inspection), which will lead to identical rows in the data set—these are not errors/duplicates.
- The row names in `data1` correspond to the row names in the full data frame `data` (so that you can check the full entry for additional variables, if you like)
- In total, there are 3235 rows in `data1`, corresponding to 1977 unique inspections. The data set is sorted first by ID, then by date (so that all inspections for a single restaurant are grouped together).

Here is the beginning of `data1`:

	month	day	year	business_id	risk_category	inspection_score	zipcode	latitude	longitude
12623	2	26	2018	10030	None	100	94103	37.76686	-122.4190
9527	9	5	2017	10083	Moderate Risk	81	94111	37.79402	-122.4013
17937	9	5	2017	10083	Moderate Risk	81	94111	37.79402	-122.4013
10661	4	26	2019	10083	Moderate Risk	86	94111	37.79402	-122.4013
15284	9	19	2017	10280	Moderate Risk	92	94102	37.78203	-122.4198
17767	4	5	2018	10280	Low Risk	88	94102	37.78203	-122.4198
10621	10	18	2018	10280	High Risk	81	94102	37.78203	-122.4198
12919	10	18	2018	10280	Low Risk	81	94102	37.78203	-122.4198
15055	10	18	2018	10280	Low Risk	81	94102	37.78203	-122.4198

```
16433    10  18 2018       10280      Low Risk              81   94102 37.78203 -122.4198
13452     3   1 2019       10280      Low Risk              88   94102 37.78203 -122.4198
18323     3   1 2019       10280      Low Risk              88   94102 37.78203 -122.4198
15118     5  23 2017       10282     High Risk              91   94118 37.78731 -122.4468
15532     6  12 2018       10282      Low Risk              92   94118 37.78731 -122.4468
......
```

Guidelines for group & report Please see guidelines for mini-project 1.
The grading rubric will be the same as mini-project 1.

Assignment Please choose *either* question 1 or question 2 for your project.

- **Question 1:** Geographical clusters.
  - (a) Since the inspections are carried out by people who need to travel to the restaurant in person, it would be plausible that nearby restaurants are likely to be inspected on the same day for efficiency. Do we see any evidence of this in the data set? (The zip code is only a coarse measure of location, so we should use latitude/longitude for a finer measure.) Be sure to consider possible sources of confounding, e.g., higher rates of inspections in certain locations, seasonal effect like more frequent inspections during certain weeks/months, day-of-week effect, overall increase in frequency of inspections, etc. You can consider permutation based approach or something more parametric/model-based.
  - (b) If a restaurant was recently inspected and/or cited for a violation, does that influence the hygiene at nearby restaurants—do they improve their sanitation in fear of an inspection, which would result in higher inspection scores/lower risk categories at inspection time?
- **Question 2:** Time between inspections.
  - (a) First, we will construct prediction intervals for the time of the next inspection. Suppose that a restaurant i is inspected on date j. At this time, we would like to say, with 95% confidence, that the next random inspection will be at least XXX days from now. This prediction needs to be calculated as a function of information available on date j (that is, we will not update it later if we see that, e.g., many inspections are occurring nearby to restaurant i a few months in the future).

    Using the data from 2016–2018, construct a method for giving this type of (one-sided) prediction interval—the model you use is up to you, and you should be sure to consider relevant features/confounders as described in question 1. You should use a holdout method or some other method for calibrating your coverage, i.e., ensuring that your predictive interval covers the actual inspection date 95% of the time as claimed.

    Next, assess the accuracy of your method on the 2019 data (that is, for every restaurant that was in the data set in 2016–2018, calculate its prediction interval based on the last date it had a random inspection (i.e., "The next inspection will be at least XXX days from now"), and then verify if the date of the next inspection indeed came at least XXX many days later (note—if no inspection occurred in 2019 then we can count this as a success).
  - (b) Next, we will test a hypothesis related to the predictive inference question above. Does the outcome of the most recent inspection (i.e., inspection score and/or risk categories) affect how soon the restaurant will have its next routine inspection?