# STAT231 Group C Revised Project Proposal

*Kenny Chen, Jessica Yu*

*November 12, 2019*

## Contents

### Group Letter: C

### Group Members:

Kenny Chen and Jessica Yu

### Title:

Examining healthcode violation trends and similarities between Las Vegas restaurants

### Purpose:

We want to use spatial data to map restaurants in Las Vegas to examine if certain types of healthcode violations occur in a specific area/district more frequently than others. We want to combine that information with sentiment analysis of Yelp reviews to also show users "similar" restaurants.

### Data:

We will be using the dataset provided by Yelp that contains information about restaurants in Las Vegas along with reviews as well as Las Vegas Health Department data on healthcode violations. Sample of the yelp data and healthcode violation is in the revised folder and a sample of the code is shown below.

Out of 10 restaurants we chose from the Yelp Dataset, only 3 were able to be matched with inspection data when joining the two tables together.

```r
## reading in files
lvinspection <- read.csv("Restaurant_Inspections_Open_Data.csv")
lvrestaurant <-readRDS("lasVegasExcerpt.Rds")
lvviolation <-read.csv("Restaurant_Inspection_Violation_Codes.csv")
```

```r
## cleaning inspection data
    # making latitude + longitude columns w/rounding
lvinspection <- lvinspection %>%
  separate(Location_1, c("latitude","longitude"),sep=",")
```

```r
lvinspection$latitude <-
  as.numeric(gsub("\\(", "", lvinspection$latitude))

lvinspection$longitude <-
  as.numeric(gsub("\\)", "", lvinspection$longitude))

lvinspection <-lvinspection %>%
  mutate_at(vars(latitude,longitude), funs(round(.,3)))

 # make restaurant names lowercase and getting first 2 letters
lvinspection$Restaurant_Name<-
  tolower(lvinspection$Restaurant_Name)
lvinspection$Restaurant_Name <-
  (str_extract(lvinspection$Restaurant_Name , "^[a-z]{2}"))


## cleaning restaurant data
    #rounding lat + long
lvrestaurant <- lvrestaurant %>%
  mutate_at(vars(latitude,longitude), funs(round(.,3)))%>%
  select(name,address,city,state,postal_code,latitude,longitude)

    # making restaurant names lowercase and getting first 2 letters
lvrestaurant$name<-tolower(lvrestaurant$name)
lvrestaurant$name <- (str_extract(lvrestaurant$name, "^[a-z]{2}"))


table<-left_join(lvrestaurant,lvinspection,
                 by = c("latitude"="latitude","city"="City",
                                                   "postal_code"="Zip",
                        "name"="Restaurant_Name"))
```

## Variables:

We will be looking at longitude,lattitude,postal code, type of cuisine, price range, number and type of healthcode violations, and conducting sentiment analysis on Yelp reviews.

## End Product:

Ultimately, we would like to have some sort of shiny app that can visualize the restaurant data across a map of Las Vegas. Our inspiration for our project is: https://www.reddit.com/r/dataisbeautiful/comments/8hys9k/the_city_is_alive_the_population_of_manhattan/ We would allow users to view where certain types of violations occur most frequently, as well as choosing criteria for what kinds of similar restaurants they would like to see. (For example, viewing similar expensive Chinese restaurants)…