HW01: Coding Portion SOLUTIONS

PSTAT 100: Spring 2024 (Instructor: Ethan P. Marzban)

MEMBER 1 (NetID 1) MEMBER 2 (NetID 2)
MEMBER 3 (NetID 3)

Coding Portion: Health Inspections

> **Tip**
>
> Don't try to answer the sub-questions within any one question in list format - write your answers narratively, referencing code output wherever necessary. Additionally, think of the Coding Portions of the Homework Assignments as Mini-Mini-Projects. Specifically, some of the questions that are asked of you may be open-ended, which is by design! Feel free to stop by office hours (either the Instructor's or the TAs') to discuss!

The County of Los Angeles Department of Public Health routinely publishes results of environmental health inspections for several types of businesses (e.g. restaurants, apartments, etc.) at this link. In this part of our homework, we will investigate some of the results of these health inspections; the specific dataset we will be using can be found in the `data` subdirectory, with the file name `safety_ratings.csv`, and includes the following variables:

- `Facility`: the name of the facility being reviewed
- `Last Routine Inspection`: date of the last routine inspection (as of early March 2024)
- `Score`: score of the last routine inspection
- `Address`: address of the facility being reviewed
- `City`: city of the facility being reviewed (for those unaware, the county of Los Angeles is comprised of several smaller cities; e.g. Burbank, Santa Monica, etc.)

Also included in the `data` subdirectory is a file called `city_info.csv`, which contains selected information about the various cities included in the County of Los Angeles (data accessed and modified from this source). Specifically, the `city_info.csv` dataset contains the following variables:

- `City_Name`: the name of the city
- `Supervisorial_District`: the Supervisorial District of the city
- `Class`: the class of the city
- `Population_2010`: the population of the city in 2010
- `Inc_Yr`: the year of Incorporation of the city
- `Inc_Month`: the month of Incorporation[1] of the city
- `Inc_Day`: the day of month of Incorporation of the city

[1] **Incorporation**, in an urban geography context, refers to the act of officially forming a city.

Part 1: Exploring the Cities

Let's start off by exploring the cities included in the County of Los Angeles (i.e. by exploring the `city_info.csv` file)

> **! Question 1**
>
> - According to the `city_info.csv` file, how many cities are located in the county of Los Angeles?
> - What was the total (aggregated) population of cities in Los Angeles in 2010?
> - What was the most recent city to be Incorporated in the County of Los Angeles?
> - What was the oldest city to be Incorporated in the County of Los Angeles?

ANSWERS TO QUESTION 1:

Replace this line with your answers

We start off by reading in the data:

```
library(tidyverse)
city_info <- read.csv("data/city_info.csv")
```

Skimming through the first few lines of the dataframe, we see that each city is listed as an individual row in our table. Hence, to find the number of cities, we simply need to use the `nrow()` function:

```
nrow(city_info)
```

```
[1] 88
```

Therefore, we see that there are 88 cities included in the dataset. To find the total population, we simply sum up the values in the `population` column:

```
sum(city_info$Population_2010)
```

```
[1] 9345804
```

There are a couple of ways we can find the most recent and oldest cities to be incorporated. I'll demonstrate how we can use some of the `tidyverse` functions to accomplish this. First, let's mutate the incorporation year and month variables to be numerical and ordinal, respectively; then, we'll sort the rows of our table chronologically (I'm also displaying only a few columns, just so everything fits on the page):

```
city_info <- city_info %>%
  mutate(Inc_Yr = as.numeric(Inc_Yr)) %>%
  mutate(Inc_Month = factor(Inc_Month,
```

```
                              ordered = T,
                              levels = c("Jan.", "Feb.", "March", "April",
                                         "May", "June", "July", "Aug.",
                                         "Sept.", "Oct.", "Nov.", "Dec."))) %>%
    group_by(Inc_Yr, Inc_Month, Inc_Day)

  city_info %>%
    arrange("Inc_Yr", "Inc_Month", "Inc_Day", .by_group = T) %>%
    select(
      City_Name,
      Inc_Yr,
      Inc_Month,
      Inc_Day
    )
```

```
# A tibble: 88 x 4
# Groups:   Inc_Yr, Inc_Month, Inc_Day [88]
   City_Name      Inc_Yr Inc_Month Inc_Day
   <chr>           <dbl> <ord>       <int>
 1 Los Angeles      1850 April           4
 2 Pasadena         1886 June           19
 3 Santa Monica     1886 Dec.            9
 4 Monrovia         1887 Dec.           15
 5 Pomona           1888 Jan.            6
 6 South Pasadena   1888 Feb.           29
 7 Compton          1888 May            11
 8 Redondo Beach    1892 April          29
 9 Long Beach       1897 Dec.           13
10 Whittier         1898 Feb.           28
# i 78 more rows
```

To extract out the oldest and newest cities (based on incorporation date), we can simply select the first and last rows of this reordered dataframe:

```
  (city_info %>%
    arrange("Inc_Yr", "Inc_Month", "Inc_Day", .by_group = T) )[c(1, 88),]
```

```
# A tibble: 2 x 7
# Groups:   Inc_Yr, Inc_Month, Inc_Day [2]
  City_Name   Supervisorial_District Class      Population_2010 Inc_Yr Inc_Month
  <chr>       <chr>                  <chr>                <int>  <dbl> <ord>
1 Los Angeles 2,4                    Charter            4094764   1850 April
2 Calabasas   3                      General L~           23788   1991 April
# i 1 more variable: Inc_Day <int>
```

So, the oldest city to be incorporated was Los Angeles and the most recent was Calabasas.

As a Data Scientist, it is important that we understand as many of the variables in our dataset as possible (which sometimes involves drawing on **domain knowledge.**) Google is a great resource for this! For example, it's not entirely obvious (from our dataset alone) what the "`Class`" of a city refers to.

! Question 2

- What are the different classes of cities?
- Use Google to look up what differences in these classes of cities, and write down a few.

ANSWERS TO QUESTION 2:

Replace this line with your answers

There are two main types of cities: Charter cities, and General Law cities. From this source:

> 'There are two types of cities in California: "charter cities," which operate under the city's local charter, and "general law cities," which operate under the general laws of the state. Charters can also contain (self-imposed) limitations on city activities.'

Similarly, not all of us may know what the different Supervisorial Districts of Los Angeles are.

! Question 3

- Use Google to look up how many Supervisorial Districts there are in the County of Los Angeles, and write down their names.

ANSWERS TO QUESTION 3:

Replace this line with your answers

From this source, there are give Supervisorial Districts within the county of LA, named "First District", "Second District", "Third District", "Fourth District", and "Fifth District".

Alright, let's flex our statistical knowledge a bit.

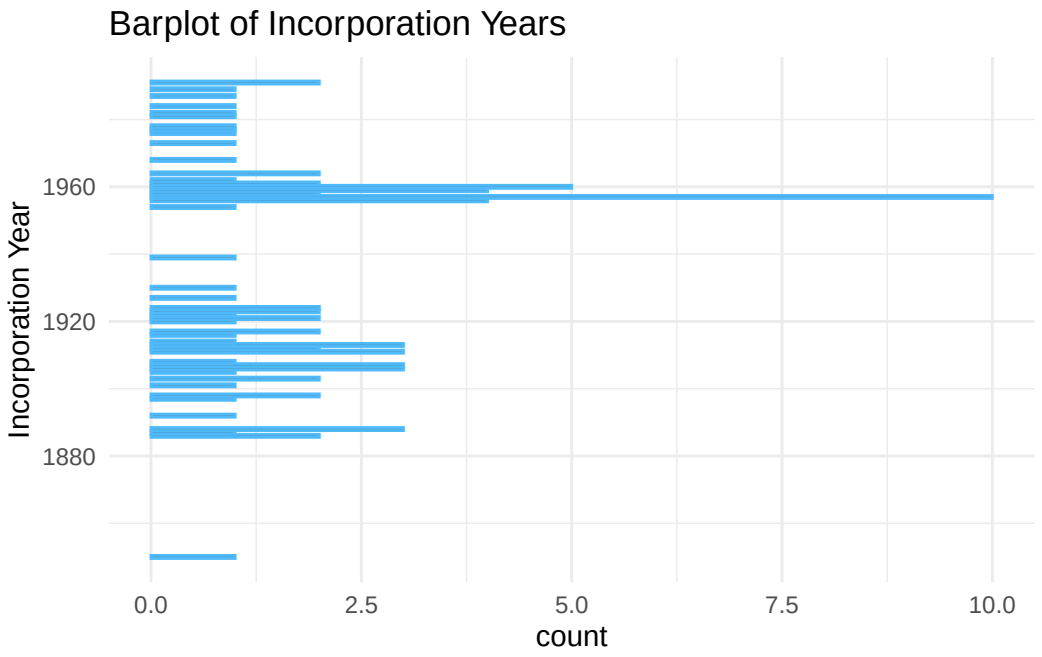
! Question 4

- Generate a barplot of Incorporation Year, and identify which year/s saw the greatest number of cities incorporated.
- Does there appear to be a month in which Incorporations typically occur? Answer this question using a graph.

ANSWERS TO QUESTION 4:

Replace this line with your answers

```
city_info %>%
  ggplot(aes(y = Inc_Yr)) +
  geom_bar(col = "#4ebafc") +
  theme_minimal() +
  ylab("Incorporation Year") +
  ggtitle("Barplot of Incorporation Years")
```



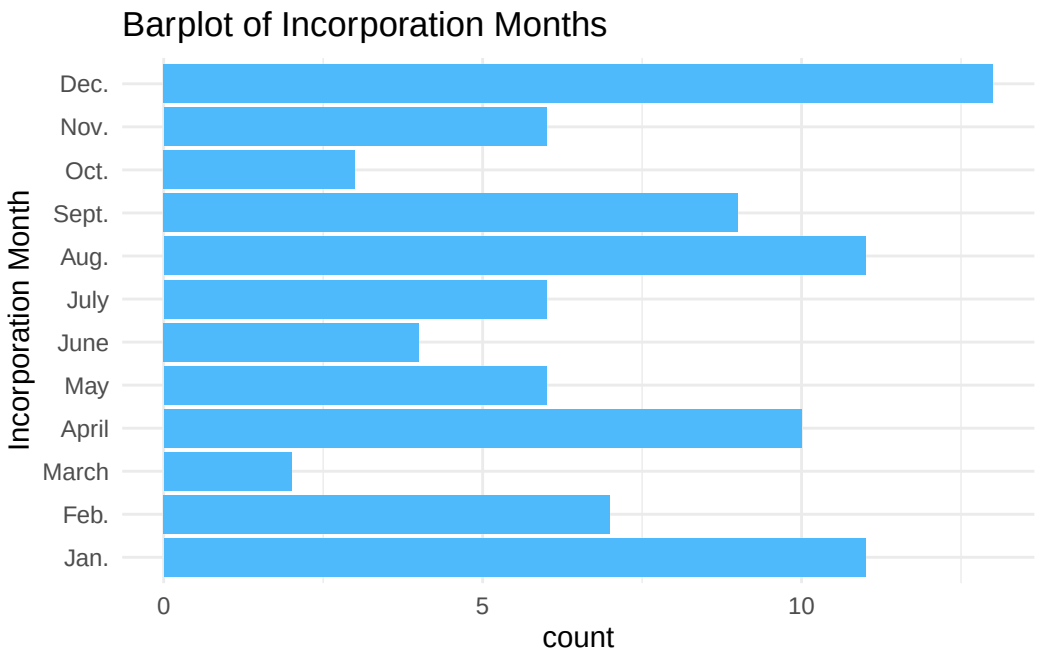
From this barplot, it appears that the greatest number of incorporations occurred sometime right before 1960. If we wanted the exact year, we can use the `table()` function:

```
city_info$Inc_Yr %>% table() %>% which.max()
```

```
1957
  30
```

So, the year with the most incorporations (of which there were 30) was 1957. Similarly, to explore incorporations by month:

```
city_info %>%
  ggplot(aes(y = Inc_Month)) +
  geom_bar(fill = "#4ebafc") +
  theme_minimal() +
  ylab("Incorporation Month") +
  ggtitle("Barplot of Incorporation Months")
```



It appears as though December contains the most amount of Incorporations, however August and January are fairly close behind. Conversely, October and March seem to contain the fewest number of incorporations.

We can also flex our `tidyverse` skills.

! Question 5

- Use the `group_by()` function to group the `city_info` dataset by Supervisorial Districts, and compute the total (aggregate) population within each Supervisorial District.

ANSWERS TO QUESTION 5:

Replace this line with your answers

```r
city_info %>%
  group_by(Supervisorial_District) %>%
  summarise(
    tot_pop = sum(Population_2010, na.rm = T)
  )
```

```
# A tibble: 6 x 2
  Supervisorial_District tot_pop
  <chr>                   <int>
1 1                      1280200
2 2                       616599
3 2,4                    4094764
4 3                       263935
5 4                      1611949
6 5                      1478357
```

Part 2: Exploring the Restaurants and Ratings

Alright, let's turn our attention to the restaurants that were reviewed.

! Question 6

- How many restaurants were included in the dataset?

ANSWERS TO QUESTION 6:

Replace this line with your answers

Let's again start by reading in the dataset:

```r
safety_ratings <- read.csv("data/safety_ratings.csv")
```

Let's also take a look at the first few rows of our dataframe:

```r
safety_ratings %>% head()
```

```
                  Facility Last.Routine.Inspection Score              Address
1 ARIEL COURT APTS SPA POOL              2020-01-31    NA       535 GAYLEY AVE
2            EAGLE CATERING              2020-08-06    90 7782 SAN FERNANDO RD
3                 WORLD OIL              2022-06-21    98      478 W ARROW HWY
4              LOWE'S #1852              2023-09-06   100       13500 PAXTON ST
5         LA VERNE CAR WASH              2023-01-23    95  914 W FOOTHILL BLVD
```

```
6                  THE LOOP              2021-08-25     99    1100 W COVINA BLVD
          City
1 LOS ANGELES
2  SUN VALLEY
3       COVINA
4     PACOIMA
5    LA VERNE
6   SAN DIMAS
```

It seems as though each restaurant appears on its own line, meaning we can compute the total number of restaurants by simply counting the number of rows:

```
safety_ratings %>% nrow()
```

```
[1] 129205
```

If you skim through the dataframe, you might notice several restaurants located at 380 World Way.

! Question 7

- What major building is located at 380 World Way? (Use Google!) Why does it make sense that there might be many restaurants listed as having this location?
- How many restaurants are located at this address?

ANSWERS TO QUESTION 7:

Replace this line with your answers

Okay, that's enough preliminary exploration (for now). Let's turn our attention to the heart of this dataset: the safety ratings!

! Question 8

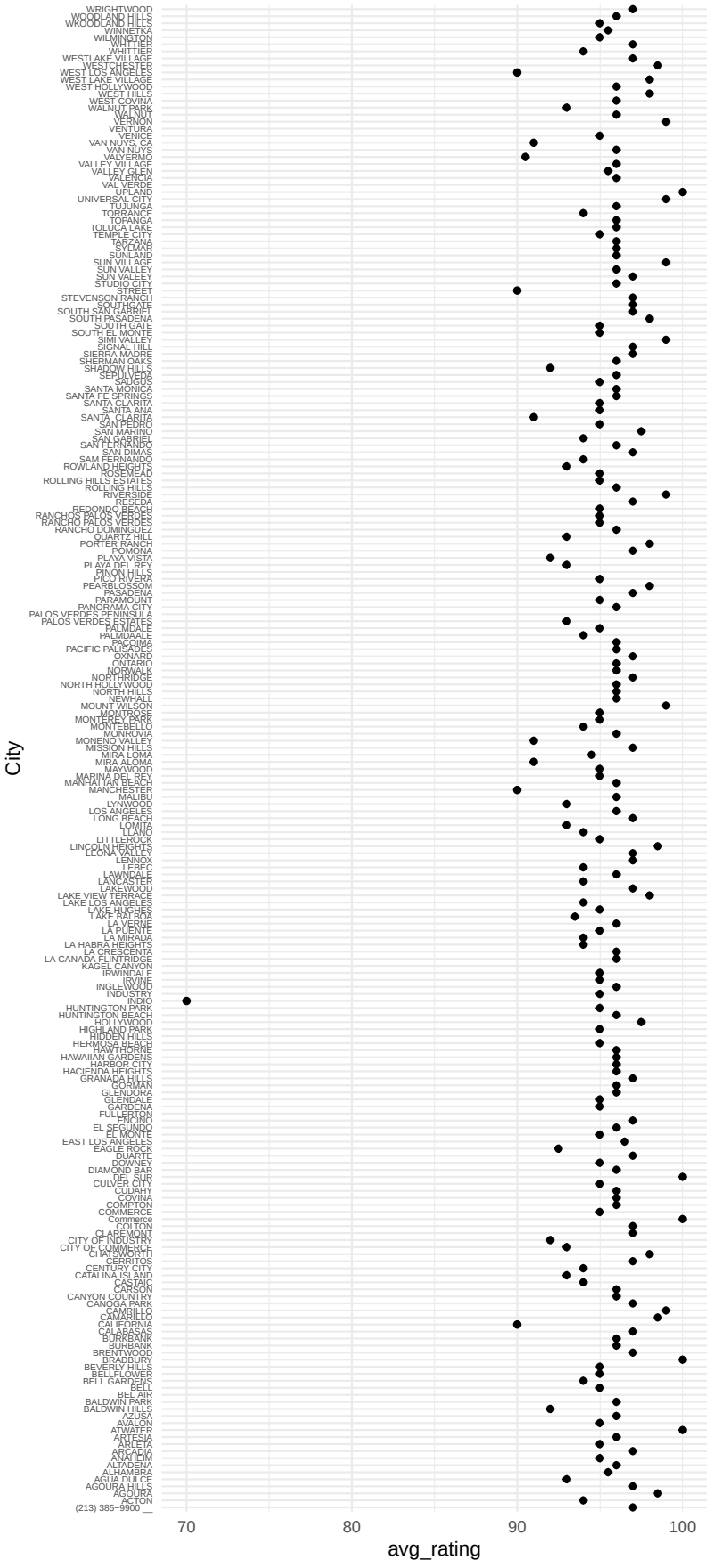
- Group the `safety_ratings` dataframe by city, and compute the median safety rating within each city.
- Use this to produce a graph with city name on the y-axis and average (**median**) score on the x-axis. Play around with axis text size and figure margins to make the figure as long as possible.

ANSWERS TO QUESTION 8:

Replace this line with your answers

```
safety_ratings %>%
  group_by(City) %>%
  summarise(
    avg_rating = median(Score, na.rm = T)
  ) %>%
  ggplot(aes(x = avg_rating,
             y = City)) +
  geom_point() +
  theme_minimal() +
  theme(
    axis.text.y = element_text(size = 5)
  )
```

```
Warning: Removed 8 rows containing missing values or values outside the scale range
(`geom_point()`).
```


City
avg_rating
70
80
90
100

Now, the graphic we produced in the question above is a bit misleading, because we know that not all cities have the same number of restaurants! As such, number of restaurants surveyed might be a confounding variable that artificially inflates (or deflates) a city's average safety rating.

! Question 9

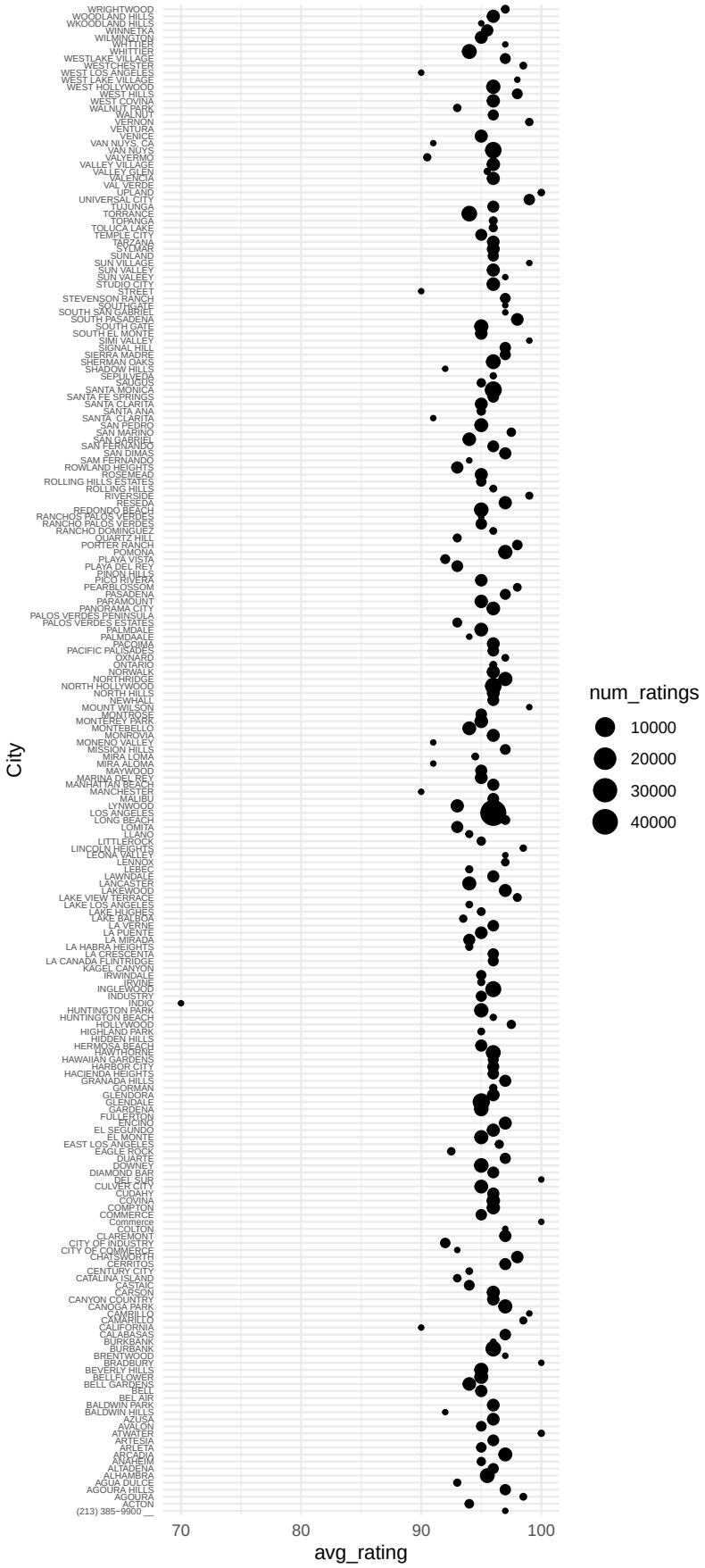
- Re-do your plot from the previous question, this time scaling each point according to the number of restaurants that were included in the city. As a hint, the first portion of your plot should look like this:

ANSWERS TO QUESTION 9:

Replace this line with your answers

```
safety_ratings %>%
  group_by(City) %>%
  summarise(
    avg_rating = median(Score, na.rm = T),
    num_ratings = n()
  ) %>%
  ggplot(aes(x = avg_rating,
             y = City)) +
  geom_point(aes(size = num_ratings)) +
  theme_minimal() +
  theme(
    axis.text.y = element_text(size = 5)
  ) +
  scale_size_continuous(transform = "sqrt")
```

```
Warning: Removed 8 rows containing missing values or values outside the scale range
(`geom_point()`).
```


City
avg_rating
70
80
90
100
num_ratings
10000
20000
30000
40000

Now, this plot is actually revealing something else about our dataset. Note, for example, that our plot includes a city called "(213) 385-9900 ___". This is clearly a mis-input.

! Question 10

- What was the name of the restaurant whose `City` was listed as `(213) 385-9900 __`?
- Use Google to look up this restaurant, and find which city it is really located in. Then, replace its `City` value (in the safety ratings dataframe) with the correct city.

ANSWERS TO QUESTION 10:

Replace this line with your answers

```
safety_ratings %>%
  filter(City == "(213) 385-9900 __")
```

```
         Facility Last.Routine.Inspection Score                    Address
1 TONKATSUZIP INC              2023-10-26    97 928 S WESTERN AVE STE #127
               City
1 (213) 385-9900 __
```

Looks like this restaurant is called "Tonkatsuzip Inc." Googling the address of this restaurant, we find this yelp listing, leading us to believe that the true city of this location is "Los Angeles". Let's fix this:

```
safety_ratings$City[
  which(safety_ratings$City == "(213) 385-9900 __")
] <- "LOS ANGELES"
```

Let's check that our replacement was successful:

```
safety_ratings %>%
  filter(Facility == "TONKATSUZIP INC")
```

```
         Facility Last.Routine.Inspection Score                    Address
1 TONKATSUZIP INC              2023-10-26    97 928 S WESTERN AVE STE #127
         City
1 LOS ANGELES
```

Additionally, note that our plot contains both a city called "Woodland Hills" and a city called "Wkoodland Hills". This is *also* clearly a mis-input!

! Question 11

- List out the unique values of the `City` variable as they appear in the safety ratings dataframe. Identify which values you believe to be typos (e.g. "Wkoodland Hills"); write down a list of these misspelled cities.
- Replace the misspelled city values with their correct spelling (e.g. replace all instances of "Wkoodland Hills" with "Woodland Hills", etc.)

ANSWERS TO QUESTION 11:

Replace this line with your answers

Let's list out the (current) unique values of the `City` variable:

```
safety_ratings$City %>% unique() %>% sort()
```

```
  [1] "ACTON"                "AGOURA"               "AGOURA HILLS"
  [4] "AGUA DULCE"           "ALHAMBRA"             "ALTADENA"
  [7] "ANAHEIM"              "ARCADIA"              "ARLETA"
 [10] "ARTESIA"              "ATWATER"              "AVALON"
 [13] "AZUSA"                "BALDWIN HILLS"        "BALDWIN PARK"
 [16] "BEL AIR"              "BELL"                 "BELL GARDENS"
 [19] "BELLFLOWER"           "BEVERLY HILLS"        "BRADBURY"
 [22] "BRENTWOOD"            "BURBANK"              "BURKBANK"
 [25] "CALABASAS"            "CALIFORNIA"           "CAMARILLO"
 [28] "CAMRILLO"             "CANOGA PARK"          "CANYON COUNTRY"
 [31] "CARSON"               "CASTAIC"              "CATALINA ISLAND"
 [34] "CENTURY CITY"         "CERRITOS"             "CHATSWORTH"
 [37] "CITY OF COMMERCE"     "CITY OF INDUSTRY"     "CLAREMONT"
 [40] "COLTON"               "Commerce"             "COMMERCE"
 [43] "COMPTON"              "COVINA"               "CUDAHY"
 [46] "CULVER CITY"          "DEL SUR"              "DIAMOND BAR"
 [49] "DOWNEY"               "DUARTE"               "EAGLE ROCK"
 [52] "EAST LOS ANGELES"     "EL MONTE"             "EL SEGUNDO"
 [55] "ENCINO"               "FULLERTON"            "GARDENA"
 [58] "GLENDALE"             "GLENDORA"             "GORMAN"
 [61] "GRANADA HILLS"        "HACIENDA HEIGHTS"     "HARBOR CITY"
 [64] "HAWAIIAN GARDENS"     "HAWTHORNE"            "HERMOSA BEACH"
 [67] "HIDDEN HILLS"         "HIGHLAND PARK"        "HOLLYWOOD"
 [70] "HUNTINGTON BEACH"     "HUNTINGTON PARK"      "INDIO"
 [73] "INDUSTRY"             "INGLEWOOD"            "IRVINE"
 [76] "IRWINDALE"            "KAGEL CANYON"         "LA CANADA FLINTRIDGE"
 [79] "LA CRESCENTA"         "LA HABRA HEIGHTS"     "LA MIRADA"
 [82] "LA PUENTE"            "LA VERNE"             "LAKE BALBOA"
 [85] "LAKE HUGHES"          "LAKE LOS ANGELES"     "LAKE VIEW TERRACE"
```

```
 [88] "LAKEWOOD"                "LANCASTER"               "LAWNDALE"
 [91] "LEBEC"                   "LENNOX"                  "LEONA VALLEY"
 [94] "LINCOLN HEIGHTS"         "LITTLEROCK"              "LLANO"
 [97] "LOMITA"                  "LONG BEACH"              "LOS ANGELES"
[100] "LYNWOOD"                 "MALIBU"                  "MANCHESTER"
[103] "MANHATTAN BEACH"         "MARINA DEL REY"          "MAYWOOD"
[106] "MIRA ALOMA"              "MIRA LOMA"               "MISSION HILLS"
[109] "MONENO VALLEY"           "MONROVIA"                "MONTEBELLO"
[112] "MONTEREY PARK"           "MONTROSE"                "MOUNT WILSON"
[115] "NEWHALL"                 "NORTH HILLS"             "NORTH HOLLYWOOD"
[118] "NORTHRIDGE"              "NORWALK"                 "ONTARIO"
[121] "OXNARD"                  "PACIFIC PALISADES"       "PACOIMA"
[124] "PALMDAALE"               "PALMDALE"                "PALOS VERDES ESTATES"
[127] "PALOS VERDES PENINSULA" "PANORAMA CITY"           "PARAMOUNT"
[130] "PASADENA"                "PEARBLOSSOM"             "PICO RIVERA"
[133] "PINON HILLS"             "PLAYA DEL REY"           "PLAYA VISTA"
[136] "POMONA"                  "PORTER RANCH"            "QUARTZ HILL"
[139] "RANCHO DOMINGUEZ"        "RANCHO PALOS VERDES"     "RANCHOS PALOS VERDES"
[142] "REDONDO BEACH"           "RESEDA"                  "RIVERSIDE"
[145] "ROLLING HILLS"           "ROLLING HILLS ESTATES"   "ROSEMEAD"
[148] "ROWLAND HEIGHTS"         "SAM FERNANDO"            "SAN DIMAS"
[151] "SAN FERNANDO"            "SAN GABRIEL"             "SAN MARINO"
[154] "SAN PEDRO"               "SANTA  CLARITA"          "SANTA ANA"
[157] "SANTA CLARITA"           "SANTA FE SPRINGS"        "SANTA MONICA"
[160] "SAUGUS"                  "SEPULVEDA"               "SHADOW HILLS"
[163] "SHERMAN OAKS"            "SIERRA MADRE"            "SIGNAL HILL"
[166] "SIMI VALLEY"             "SOUTH EL MONTE"          "SOUTH GATE"
[169] "SOUTH PASADENA"          "SOUTH SAN GABRIEL"       "SOUTHGATE"
[172] "STEVENSON RANCH"         "STREET"                  "STUDIO CITY"
[175] "SUN VALEEY"              "SUN VALLEY"              "SUN VILLAGE"
[178] "SUNLAND"                 "SYLMAR"                  "TARZANA"
[181] "TEMPLE CITY"             "TOLUCA LAKE"             "TOPANGA"
[184] "TORRANCE"                "TUJUNGA"                 "UNIVERSAL CITY"
[187] "UPLAND"                  "VAL VERDE"               "VALENCIA"
[190] "VALLEY GLEN"             "VALLEY VILLAGE"          "VALYERMO"
[193] "VAN NUYS"                "VAN NUYS, CA"            "VENICE"
[196] "VENTURA"                 "VERNON"                  "WALNUT"
[199] "WALNUT PARK"             "WEST COVINA"             "WEST HILLS"
[202] "WEST HOLLYWOOD"          "WEST LAKE VILLAGE"       "WEST LOS ANGELES"
[205] "WESTCHESTER"             "WESTLAKE VILLAGE"        "WHITTIER"
[208] "WHTTIER"                 "WILMINGTON"              "WINNETKA"
[211] "WKOODLAND HILLS"         "WOODLAND HILLS"          "WRIGHTWOOD"
```

- `PALMDAALE` (should be `PALMDALE`)
- `WKOODLAND HILLS` (should be `WOODLAND HILLS`)
- `SAM FERNANDO` (should be `SAN FERNANDO`)
- `WEST LAKE VILLAGE` (should be `WESTLAKE VILLAGE`)
- `SOUTH GATE` (should be `SOUTHGATE`)

Let's now perform our replacements:

```
safety_ratings$City[
  which(safety_ratings$City == "PALMDAALE")
] <- "PALMDALE"

safety_ratings$City[
  which(safety_ratings$City == "WKOODLAND HILLS")
] <- "WOODLAND HILLS"

safety_ratings$City[
  which(safety_ratings$City == "SAM FERNANDO")
] <- "SAN FERNANDO"

safety_ratings$City[
  which(safety_ratings$City == "WEST LAKE VILLAGE")
] <- "WESTLAKE VILLAGE"

safety_ratings$City[
  which(safety_ratings$City == "SOUTH GATE")
] <- "SOUTHGATE"
```

Finally, note that there is a city called "California" in our dataset, that has a suspiciously small point on our plot (indicating that there is a suspiciously small amount of restaurants included in this city).

! Question 12

- How many restaurants have a `City` value of `"California"`?
- Use Google to look up each of these restaurants; replace their `City` value with their correct city locations (as identified by Google).

ANSWERS TO QUESTION 12:

Replace this line with your answers

```
safety_ratings %>%
  filter(City == "CALIFORNIA")
```

```
     Facility Last.Routine.Inspection Score             Address       City
1 LALIS PIZZA              2023-04-13    90 7902 CALIFORNIA AVE CALIFORNIA
```

Looks like Lali's Pizza is the only restaurant with a city listed as "California." A quick Google search reveals that the correct city for this location should be Huntington Beach:

```
safety_ratings$City[
  which(safety_ratings$City == "CALIFORNIA")
] <- "HUNTINGTON BEACH"
```

Part 3: Further Exploration of Ratings

Do more populous cities seem to have different average safety rantings than less populous cities? This is the main question we're going to try and answer in this part, by using plots.

! Question 13

- Merge the safety ratings and cities dataframes. As a hint: you may need to use the `toupper()` function somewhere in this step. Display the first few rows of the merged dataframe.

ANSWERS TO QUESTION 13:

Replace this line with your answers

```
safety_ratings_merged <- left_join(
  safety_ratings,
  city_info %>% mutate(City_Name = toupper(City_Name)),
  by = join_by(City == City_Name)
)

safety_ratings_merged %>% head()
```

```
                  Facility Last.Routine.Inspection Score              Address
1 ARIEL COURT APTS SPA POOL              2020-01-31    NA       535 GAYLEY AVE
2            EAGLE CATERING              2020-08-06    90 7782 SAN FERNANDO RD
3                 WORLD OIL              2022-06-21    98      478 W ARROW HWY
4              LOWE'S #1852              2023-09-06   100       13500 PAXTON ST
5         LA VERNE CAR WASH              2023-01-23    95  914 W FOOTHILL BLVD
6                 THE LOOP              2021-08-25    99   1100 W COVINA BLVD
         City Supervisorial_District       Class Population_2010 Inc_Yr
1 LOS ANGELES                   2,4     Charter         4094764   1850
2  SUN VALLEY                  <NA>        <NA>              NA     NA
3      COVINA                     5 General Law           49622   1901
4     PACOIMA                  <NA>        <NA>              NA     NA
5    LA VERNE                     5 General Law           34051   1906
6   SAN DIMAS                     5 General Law           36946   1960
  Inc_Month Inc_Day
```

```
1      April       4
2       <NA>      NA
3       Aug.      14
4       <NA>      NA
5      Sept.      11
6       Aug.       4
```

The reason we needed to use the `toupper()` function is that city names in the `safety_ratings` dataframe were listed in all-caps whereas city names in the `city_info` dataframe were listed in mixed case.

Now that we have both the safety rating information as well as the populations in a single dataframe, it's time to begin formatting our dataframe into a format that `ggplot()` will recognize.

First, notice that not all cities included in the safety ratings dataframe appear in the cities dataframe. (This is largely because the safety ratings dataframe includes *neighborhoods* and a few neighboring cities of LA, whereas the cities dataframe includes only cities that were formally incorporated into the county of LA). To simplify our considerations, let's here on out focus only on cities that were formally incorporated into the county of LA.

! Question 14

- Make a dataframe that includes only the following variables: `City`, `Supervisorial_District`, `Score`. Group this dataframe by `City` and `Supervisorial_District`, and compute the average rating of each city/supervisorial district combination along with the population of the underlying city. Remove all cities with a missing `Supervisorial_District` value. The first few rows of your final table should look something like this:

City	Supervisorial_District	med_score	pop
AGOURA HILLS	3	97	23387
ALHAMBRA	5	95.5	89501
ARCADIA	5	97	56719
ARTESIA	4	96	17608

Hint: When displaying the population values, think about what summarizing metric you might be able to use to extract out the desired population value. (You could also consider simply appending the population column from the original `cities` dataframe.)

ANSWERS TO QUESTION 14:

Replace this line with your answers

```r
safety_ratings_merged %>%
  group_by(
    City
  ) %>%
  summarise(
    Supervisorial_District = first(Supervisorial_District),
    med_score = median(Score, na.rm = T),
    pop = first(Population_2010)
  ) %>%
  filter(!is.na(pop))
```

```
# A tibble: 86 x 4
   City         Supervisorial_District med_score   pop
   <chr>        <chr>                      <dbl> <int>
 1 AGOURA HILLS 3                           97   23387
 2 ALHAMBRA     5                           95.5 89501
 3 ARCADIA      5                           97   56719
 4 ARTESIA      4                           96   17608
 5 AVALON       4                           95    3559
 6 AZUSA        1                           96   49207
 7 BALDWIN PARK 1                           96   81604
 8 BELL         1                           95   38867
 9 BELL GARDENS 1                           94   77312
10 BELLFLOWER   4                           95   47002
# i 76 more rows
```

Okay, this is looking pretty good! Let's start making some plots.

! Question 15

- Use your dataframe from the above question to create a scatterplot of median safety ratings (on the y-axis) and population (on the x-axis). Color your plot based on supervisorial district.

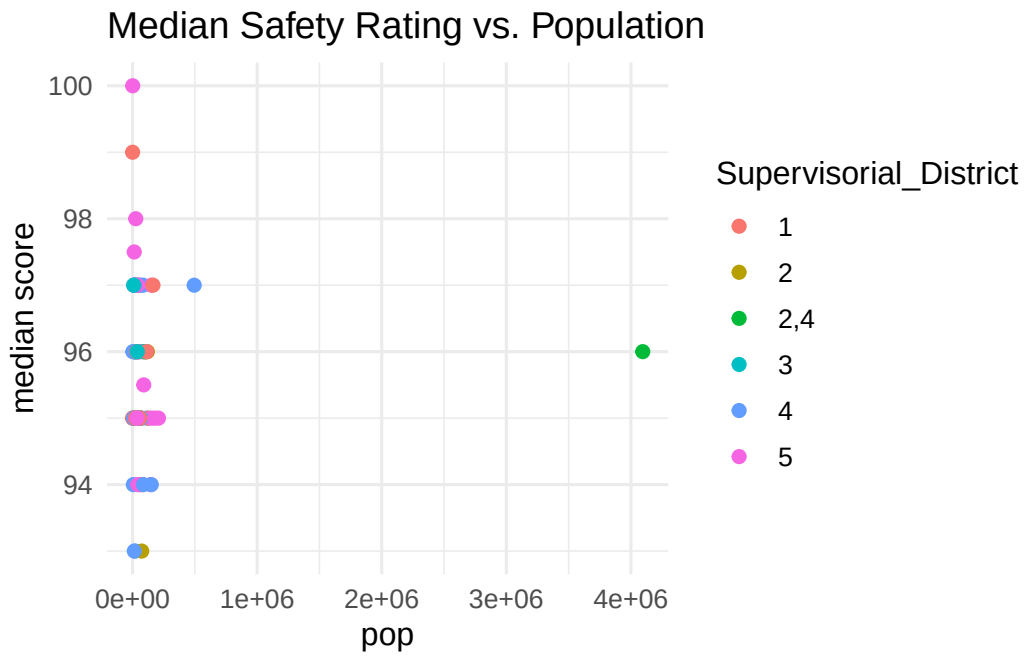
ANSWERS TO QUESTION 15:

Replace this line with your answers

```r
safety_ratings_merged %>%
  group_by(
    City
  ) %>%
  summarise(
```

```
  Supervisorial_District = first(Supervisorial_District),
  med_score = median(Score, na.rm = T),
  pop = first(Population_2010)
) %>%
filter(!is.na(pop)) %>%
ggplot(aes(x = pop,
           y = med_score,
           group = Supervisorial_District)) +
geom_point(aes(col = Supervisorial_District),
           size = 2) +
theme_minimal(base_size = 12) +
ylab("median score") +
ggtitle("Median Safety Rating vs. Population")
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



The different supervisorial districts are getting a bit muddled - coloring might not have been the best choice. When it comes to displaying variations across categories, another option available to us is **facetting**.

! Question 16

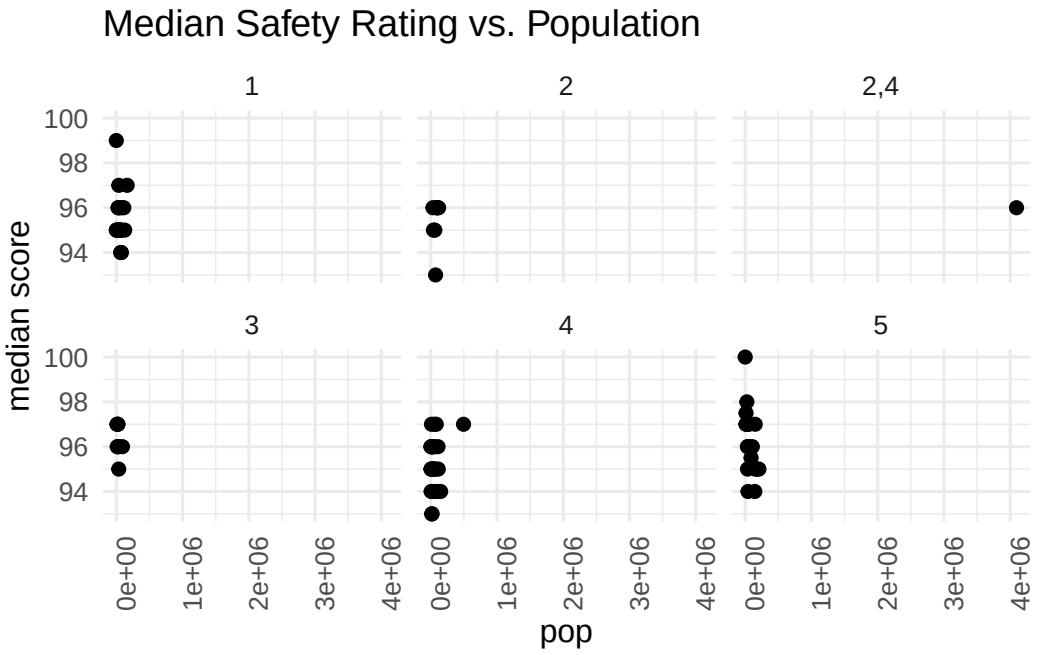
- Make another scatterplot of median safety ratings (on the y-axis) and population (on the x-axis); this time, use the `facet_wrap()` function to facet based on supervisorial district.

ANSWERS TO QUESTION 16:

Replace this line with your answers

```
safety_ratings_merged %>%
  group_by(
    City
  ) %>%
  summarise(
    Supervisorial_District = first(Supervisorial_District),
    med_score = median(Score, na.rm = T),
    pop = first(Population_2010)
  ) %>%
  filter(!is.na(pop)) %>%
  ggplot(aes(x = pop,
             y = med_score,
             group = Supervisorial_District)) +
  geom_point(size = 2) +
  theme_minimal(base_size = 12) +
  facet_wrap(~Supervisorial_District) +
  ylab("median score") +
  ggtitle("Median Safety Rating vs. Population") +
  theme(axis.text.x = element_text(angle = 90))
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```



Finally, as mentioned many times throughout this course, interpreting our plots is a key part of being a good datascientist.

! Question 17

- Does there appear to be a relationship between median safety ratings and population? Does the nature of the relationship appear to change across supervisorial districts?

ANSWERS TO QUESTION 17:

Replace this line with your answers

It doesn't appear as though there is a relationship between average safety rating and population size; furthermore, this lack of relationship doesn't appear to change across supervisorial districts.
