PLAN 372 HW 2

Lucas Mayo

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(lubridate)
```

Load the data from CSV.

```r
install.packages("ggplot2")
```

```
Warning: package 'ggplot2' is in use and will not be installed
```

(the above and below is a result from where I made a mistake trying to follow a guide to expand my graphs, and uninstalled a crucial package, and thus caused issues with most of my code).

```r
restaurant_inspections <- read.csv("C:/Users/lucas/Documents/plan372/plan372-hw2/restaurant_
```

```r
data <- read_csv("C:/Users/lucas/Documents/plan372/plan372-hw2/restaurant_inspections.csv")
```

```
Rows: 3875 Columns: 12
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (8): HSISID, DESCRIPTION, TYPE, INSPECTOR, NAME, RESTAURANTOPENDATE, CI...
dbl  (3): OBJECTID, SCORE, PERMITID
dttm (1): DATE_

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

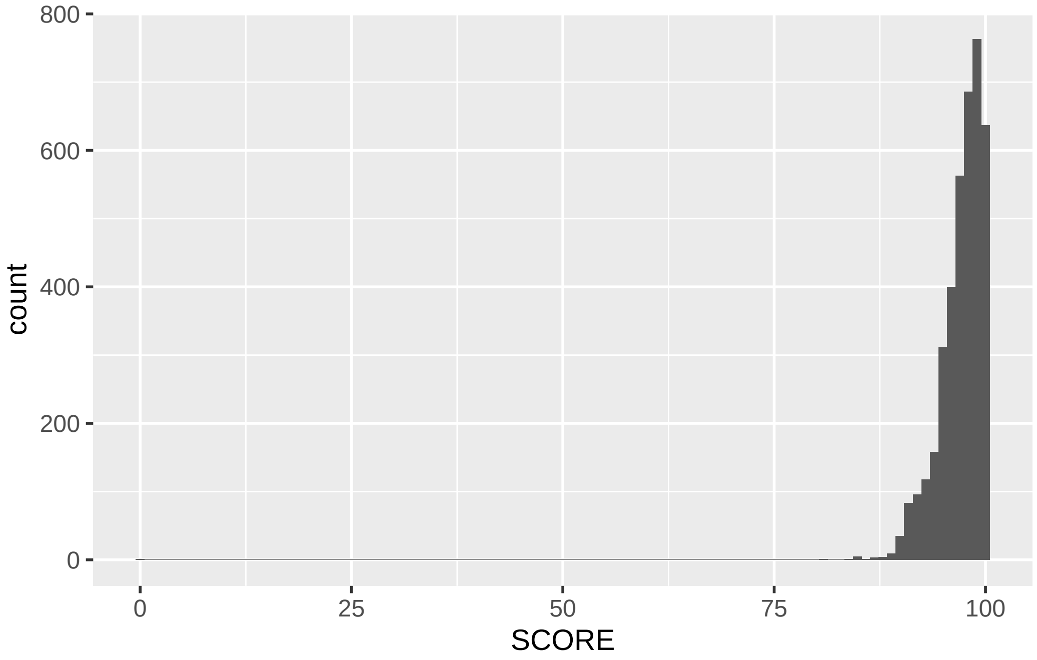
```
head(data)
```

```
# A tibble: 6 x 12
  OBJECTID HSISID SCORE DATE_               DESCRIPTION TYPE  INSPECTOR PERMITID
     <dbl> <chr>  <dbl> <dttm>              <chr>       <chr> <chr>        <dbl>
1 25137654 04092~  97   2017-10-22 04:00:00  <NA>       Insp~ Karla Cr~    13405
2 25115128 04092~  96   2019-02-27 05:00:00 "*Notice* ~ Insp~ Meghan S~    13939
3 25123164 04092~  98.5 2019-03-04 05:00:00 "*NOTICE* ~ Insp~ Kaitlyn ~    20554
4 25128895 04092~  90.5 2019-03-23 04:00:00 "Opening c~ Insp~ Angela M~    15506
5 25124786 04092~  97.5 2019-04-24 04:00:00 "*NOTICE* ~ Insp~ Patricia~    14839
6 25108274 04092~  98   2019-05-14 04:00:00 "*NOTICE* ~ Insp~ Maria Po~     8851
# i 4 more variables: NAME <chr>, RESTAURANTOPENDATE <chr>, CITY <chr>,
#   FACILITYTYPE <chr>
```

Used the head(data) like we commonly do in class in order to ensure that it was loaded correctly.

Question 1. Visualizing distribution of inspection scores

```
ggplot(restaurant_inspections, aes(x= SCORE)) +
  geom_histogram(bins = 100)
```



Additionally, at this point 296 rows containing missing values or values outside the scale range were removed. This could potentially impact the average scores and how they are presented in this histogram, as some scores that were lost could have been lower than the 75 mark and add additional points.

Here I attempted to utilize what we did in class with line plots and apply it to a histogram. The data mentioned that I should use a better value, as it was originally using 'bins = 30' so I changed it to 100.

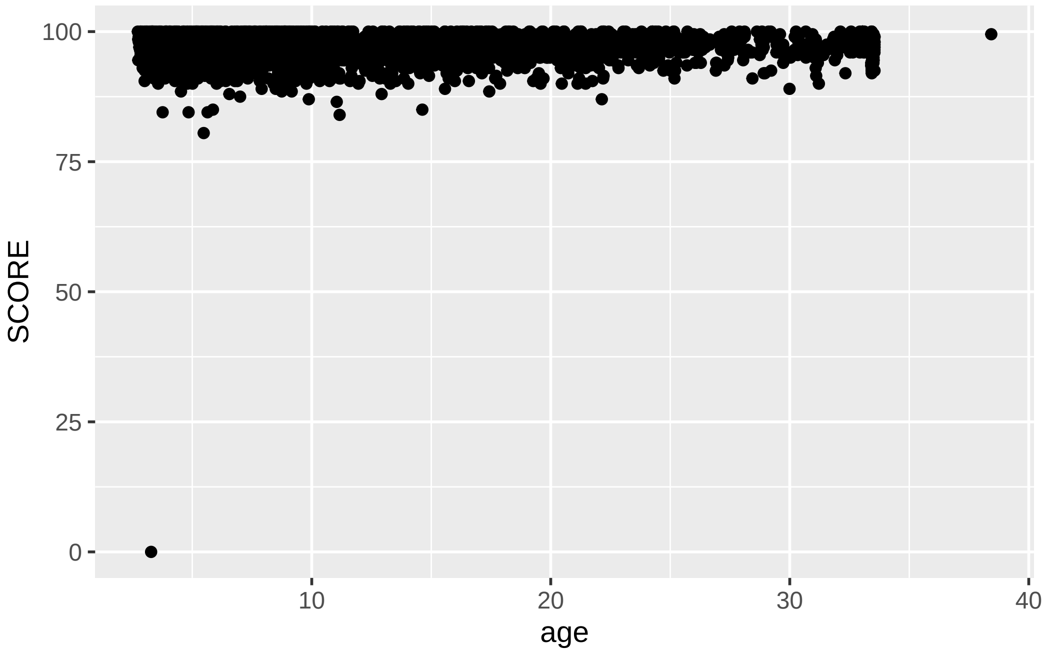
Question 2. Older and Newer Restaurant Inspection Scores

```
data$RESTAURANTOPENDATE <- ymd_hms(data$RESTAURANTOPENDATE)
data$age <- as.numeric(difftime(Sys.Date(), data$RESTAURANTOPENDATE, units = "days")) /365
```

Here I first attempted to convert the date, but ran into issues using only ymd. I hate to look at the results produced earlier in the head(data) command, to realize that it was most likely the lack of inclusion of hour and minutes that were giving me error messages.

```
ggplot(data, aes(x = age, y = SCORE)) +
  geom_point()
```

```
Warning: Removed 296 rows containing missing values or values outside the scale range
(`geom_point()`).
```



Here I used geom_point to create a scatter plot, as while scrolling the geom_ commands it seemed to be the best reflection. There does seem to be a very slight trend, because the newer restaurants have a few scores that dip closer to the 80 range, while the older restaurants do not, but it does seem very slight at best.

Question 3. Varying Inspection Score

```
different_cities <- unique(data$CITY)
print(different_cities)
```

```
 [1] "CARY"                   "RALEIGH"                "KNIGHTDALE"
 [4] "CLAYTON"                "FUQUAY VARINA"          NA
 [7] "GARNER"                 "MORRISVILLE"            "RESEARCH TRIANGLE PARK"
[10] "RTP"                    "WENDELL"                "Cary"
[13] "APEX"                   "Apex"                   "WILLOW SPRING"
[16] "HOLLY SPRINGS"          "ROLESVILLE"             "ZEBULON"
[19] "Raleigh"                "WAKE FOREST"            "NEW HILL"
[22] "FUQUAY-VARINA"          "Zebulon"                "Morrisville"
[25] "Wake Forest"            "Holly Springs"          "ANGIER"
[28] "Fuquay Varina"          "NORTH CAROLINA"         "MORRISVILE"
[31] "Fuquay-Varina"          "HOLLY SPRING"           "Garner"
```

I attempted to move on to question 3, before realizing that I did not know all the different city names and variations used in the data.

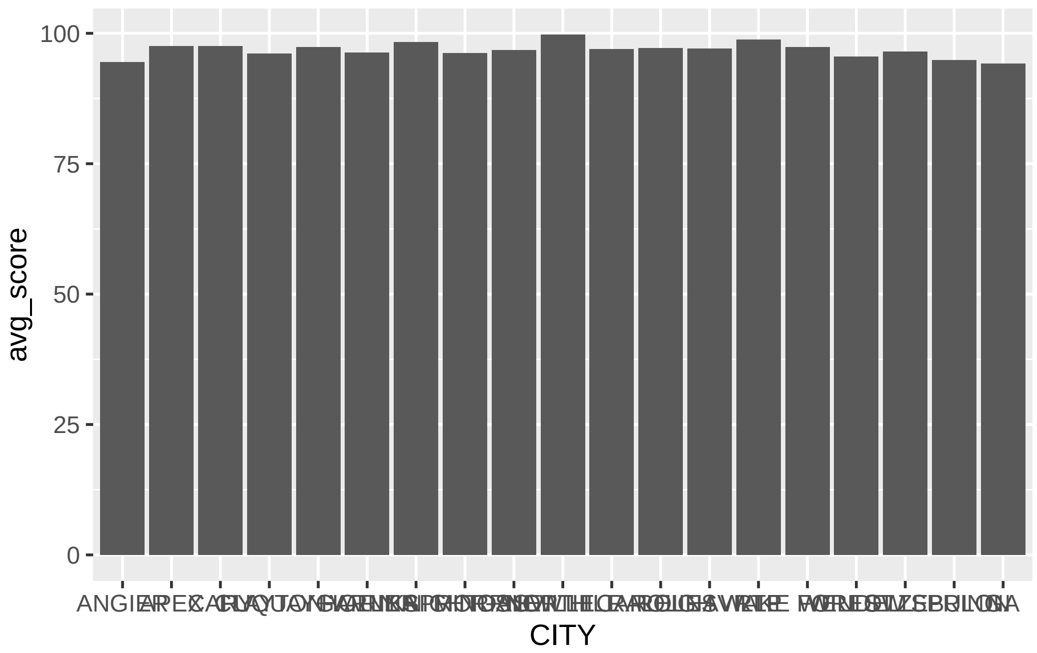
```
data$CITY <- str_to_upper(data$CITY)
data$CITY <- recode(data$CITY, "FUQUAY VARINA" = "FUQUAY-VARINA",
                    "WAKE FOREST" = "WAKE FOREST",
                    "MORRISVILE" = "MORRISVILLE",
                    "HOLLY SPRING" = "HOLLY SPRINGS",
                    "RESEARCH TRIANGLE PARK" = "RTP",
                    "CARY" = "CARY",
                    "RALEIGH" = "RALEIGH",
                    "KNIGHTDALE" = "KNIGHTDALE",
                    "CLAYTON" = "CLAYTON",
                    "GARNER" = "GARNER",
                    "MORRISVILLE" = "MORRISVILLE",
                    "WENDELL" = "WENDELL",
                    "APEX" = "APEX",
                    "WILLOW SPRING" = "WILLOW SPRING",
                    "ROLESVILLE" = "ROLESVILLE",
                    "ZEBULON" = "ZEBULON",
                    "NEW HILL" = "NEW HILL",
                    "ANGIER" = "ANGIER")
```

I made an effort to recode like we did the SFpark exercise here.

```
city_scores <- data |>
  group_by(CITY) |>
  summarize(avg_score = mean(SCORE, na.rm = TRUE)) |>
  ungroup()
```

Using the above I attempted to ensure that NA values would be ignored.

```
ggplot(city_scores, aes(x = CITY, y= avg_score)) +
  geom_col()
```



This was very hard to read the city names properly, so I looked up how to change that and spread them out in order to properly see how scores vary by city.

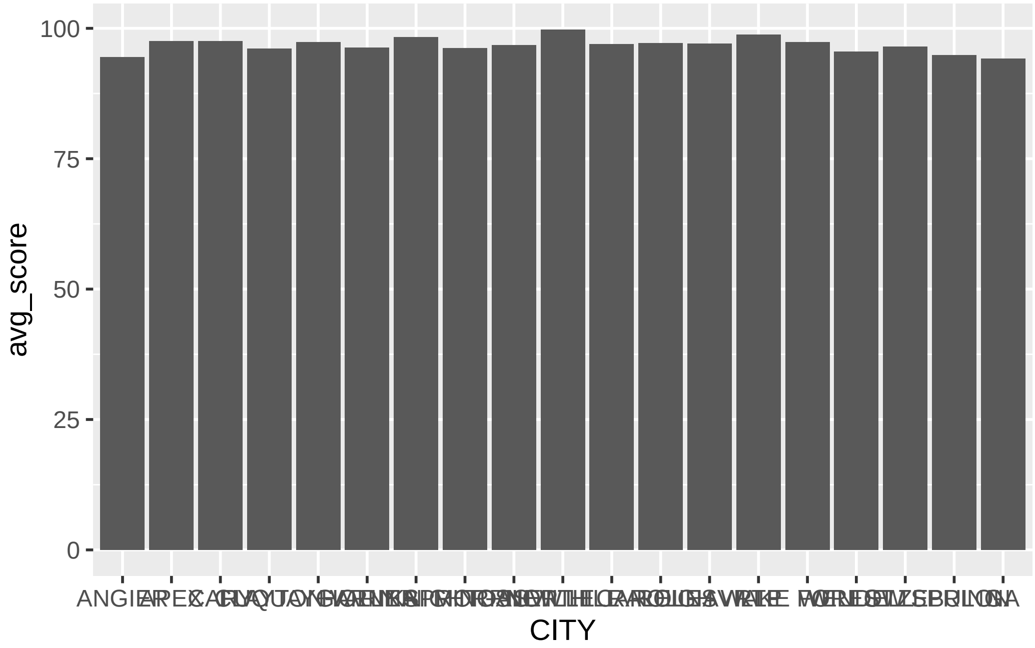
```r
install.packages("ggplot2")
```

```
Warning: package 'ggplot2' is in use and will not be installed
```

This is the mistake that I referenced earlier, and now after various restarts and attempted fixes are hoping to ignore.

```r
ggplot(city_scores, aes(x= CITY, y=avg_score)) +
  geom_col() +
  theme(axis.txt.x = element_text(margin=margin(t=10)))
```

```
Warning in plot_theme(plot): The `axis.txt.x` theme element is not defined in
the element hierarchy.
```



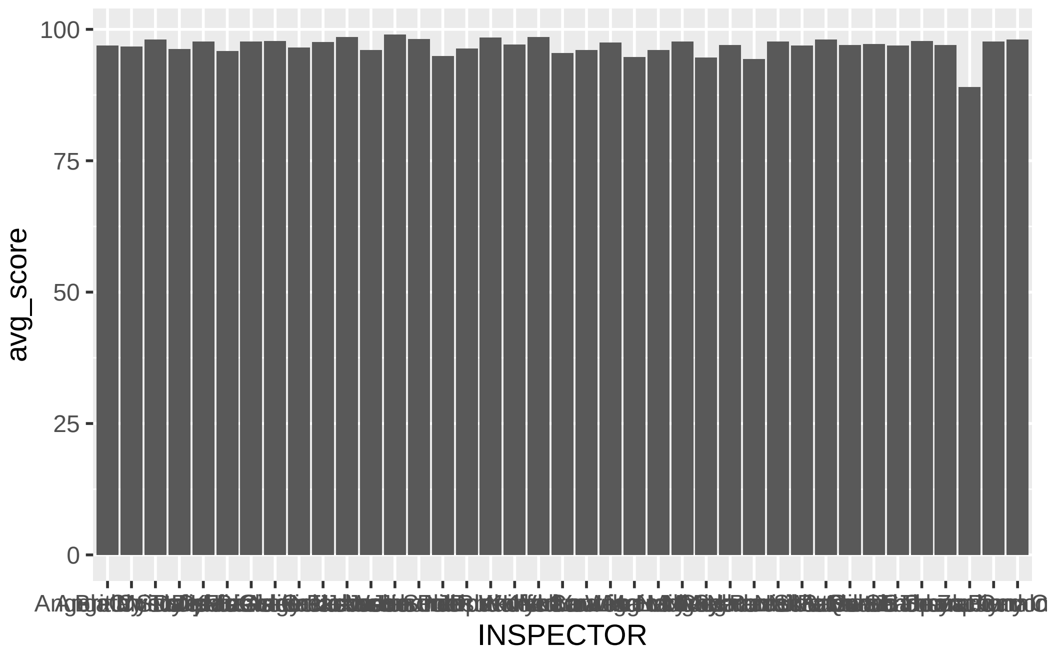
There seems to be fairly little variation of inspection scores by city.

Question 4. Varying Inspection by Inspectors

```r
inspector_scores <- data |>
  group_by(INSPECTOR) |>
  summarize(avg_score =mean(SCORE, na.rm = TRUE)) |>
  ungroup()
```

Here I attempted to group by the scores by inspectors in order to create the graph representing the differences in scores.

```r
ggplot(inspector_scores, aes(x = INSPECTOR, y = avg_score)) +
  geom_col()
```



There does not seem to be a large amount of variation in the scores by the inspectors for the most part, but there is one inspector that seems to give out scores on average that are lower than any of the other inspectors.

Question 5. Sample Size

```r
sample_sizes <- data |>
  group_by(CITY) |>
  summarize(count = n()) |>
  ungroup()
```

Here I attempted to assess the sample sizes by creating a summary of the groups, and then afterwards displaying it and assessing the inspections per city.

```r
print(sample_sizes)
```

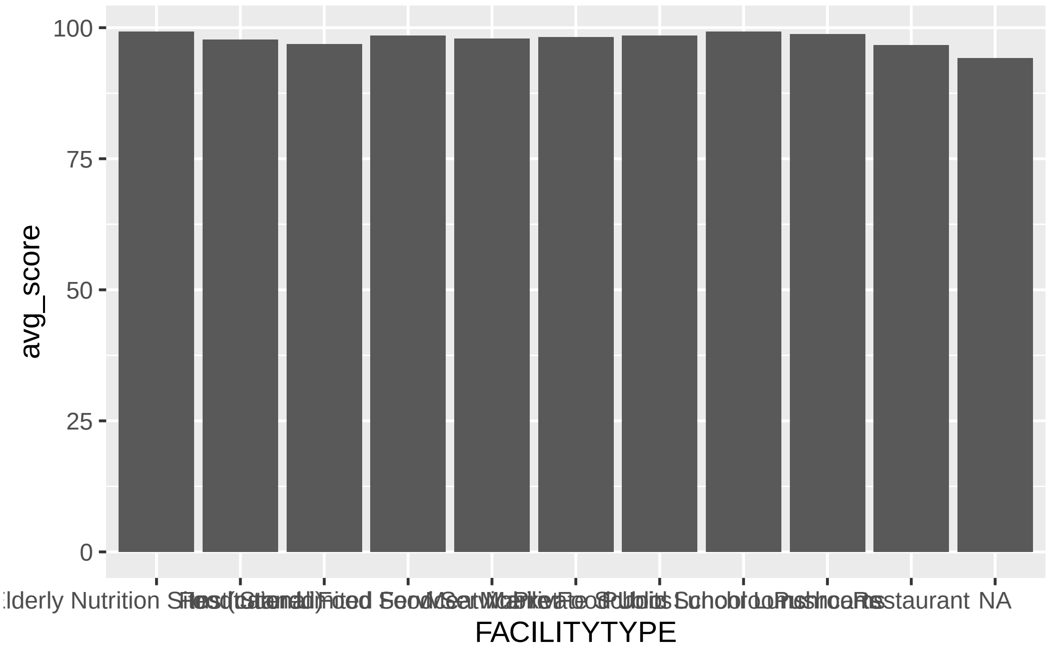
```
# A tibble: 19 x 2
   CITY           count
   <chr>          <int>
 1 ANGIER             1
 2 APEX             185
 3 CARY             573
 4 CLAYTON            4
 5 FUQUAY-VARINA    114
 6 GARNER           133
 7 HOLLY SPRINGS    107
 8 KNIGHTDALE        81
 9 MORRISVILLE      174
10 NEW HILL           2
11 NORTH CAROLINA     1
12 RALEIGH         1895
13 ROLESVILLE        24
14 RTP                2
15 WAKE FOREST      196
16 WENDELL           35
17 WILLOW SPRING      2
18 ZEBULON           50
19 <NA>             296
```

From viewing this, it can be seen that there could be a possible explanation for the results I came to above. There are various cities in which there is not a statistically significant amount of results for the sample sizes. For example, the cities of New Hill, Clayton, and Angier have results that could potentially not be representative.

Question 6. Food-Service Facility Records

```r
facility_scores <- data |>
  group_by(FACILITYTYPE) |>
  summarize(avg_score = mean(SCORE, na.rm = TRUE)) |>
  ungroup()

ggplot(facility_scores, aes(x = FACILITYTYPE, y = avg_score)) +
  geom_col()
```



The scores for restaurants do not seem to be higher than other types of facilities, they all seem to hover around the same scores. If anything, it appears that restaurant is scoring lower than the other food facilities.

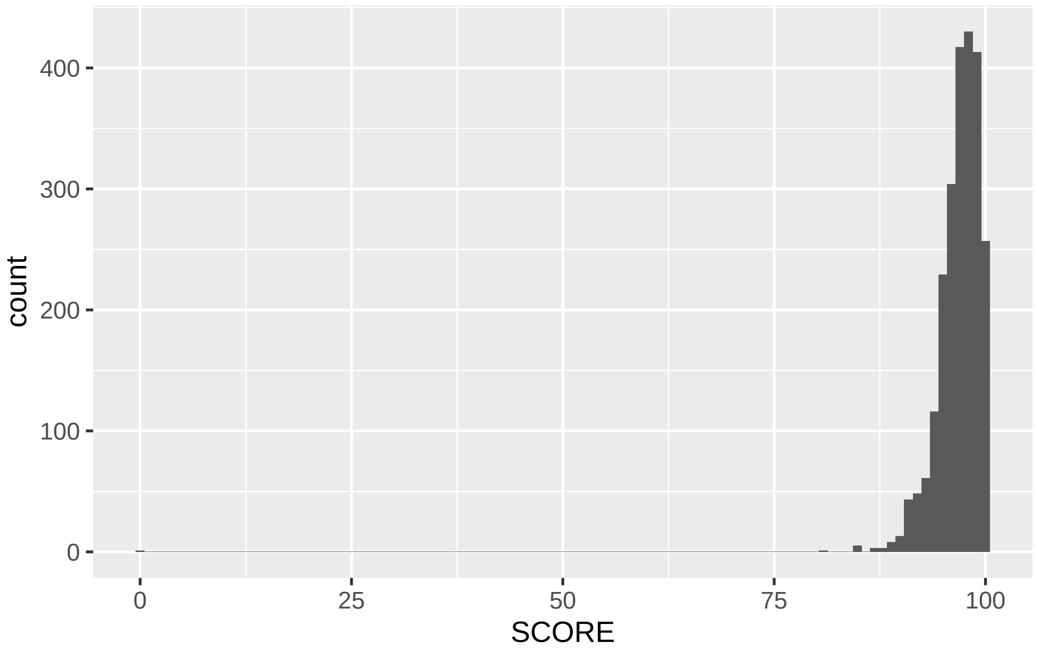
Question 7. Sanitation in Restaurants.

```r
restaurant_data <- data |>
  filter(FACILITYTYPE == "Restaurant")
```

Filtering only for restaurants.

Analysis of the distribution of inspection scores:

```r
ggplot(restaurant_data, aes(x = SCORE)) +
  geom_histogram(bins = 100)
```



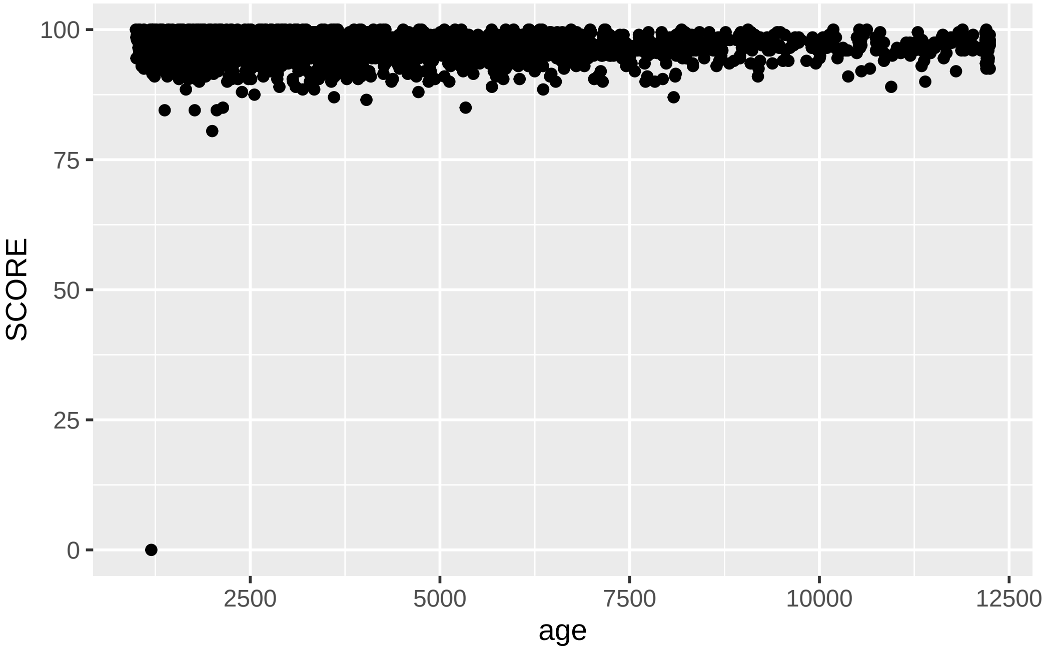
A somewhat similar trend to the before, with a bit of dip for the 100 scores.

Analysis of the trend of scores by age of restaurant

```
restaurant_data$RESTAURANTOPENDATE <- ymd_hms(restaurant_data$RESTAURANTOPENDATE)
restaurant_data$age <- as.numeric(difftime(Sys.Date(), restaurant_data$RESTAURANTOPENDATE))
```

Here is converting the restaurants open dates to the correct date format of year, month, date, and calculating age.

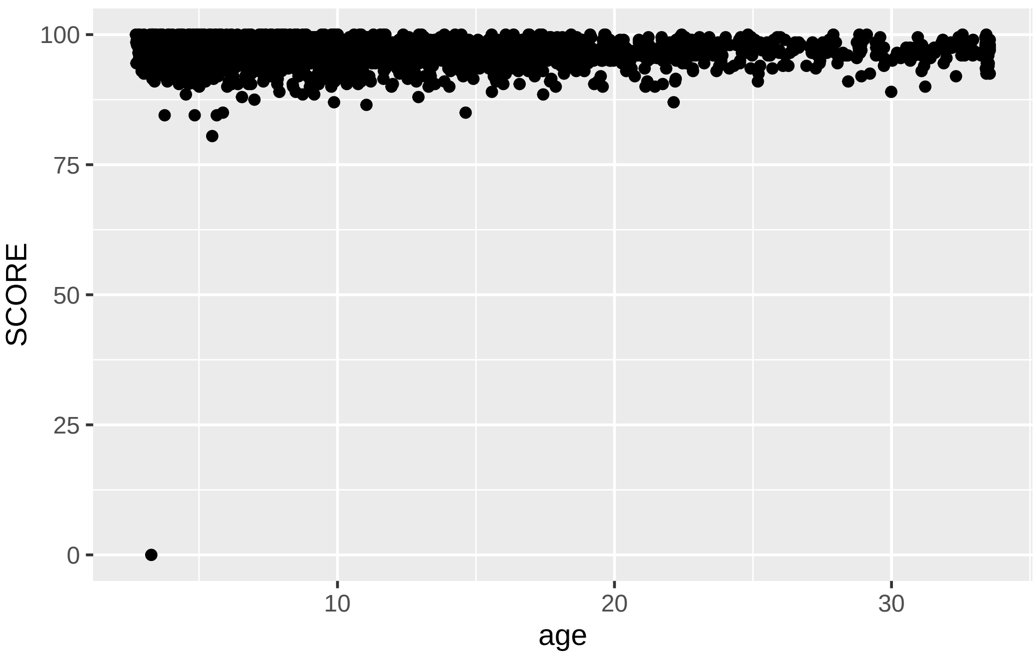
```
ggplot(restaurant_data, aes(x= age, y = SCORE)) +
  geom_point()
```



Here is creating a scatter plot of the points of data, and seem to be very similar to the previous scatter plot as well. However, I forgot to convert the amount of days to years so I proceeded to do it in the next line.

```
restaurant_data$RESTAURANTOPENDATE <- ymd_hms(restaurant_data$RESTAURANTOPENDATE)
restaurant_data$age <- as.numeric(difftime(Sys.Date(), restaurant_data$RESTAURANTOPENDATE, un
```

```
ggplot(restaurant_data, aes(x= age, y = SCORE)) +
  geom_point()
```



The outlier for age in the previous scatter plot including non-restaurants seems to have disappeared however, as there are no 40+ years of age restaurants here.

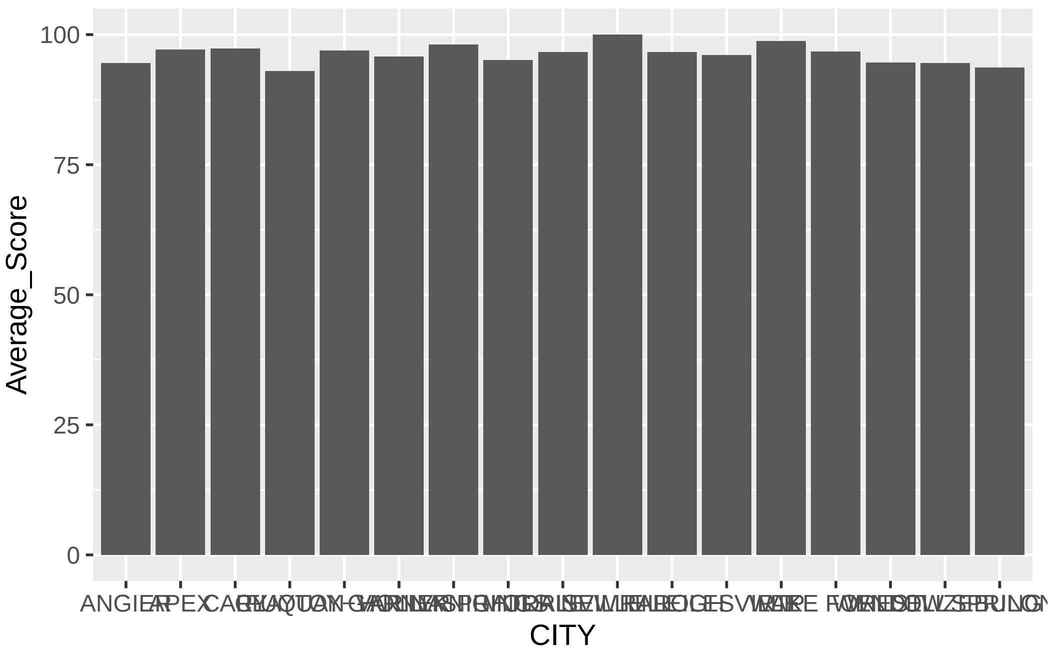
Analysis of the inspection scores by city:

```
restaurant_data$CITY <- str_to_upper(restaurant_data$CITY)
restaurant_data$CITY <- recode (restaurant_data$CITY, "FUQUAY VARINA" = "FUQUAY-VARINA", "WA
```

Here I attempt to clean the names of the cities.

```
city_scores <- restaurant_data |>
  group_by(CITY) |>
  summarise(Average_Score = mean(SCORE, na.rm = TRUE)) |>
  ungroup()

ggplot(city_scores, aes(x= CITY, y = Average_Score)) +
  geom_col()
```

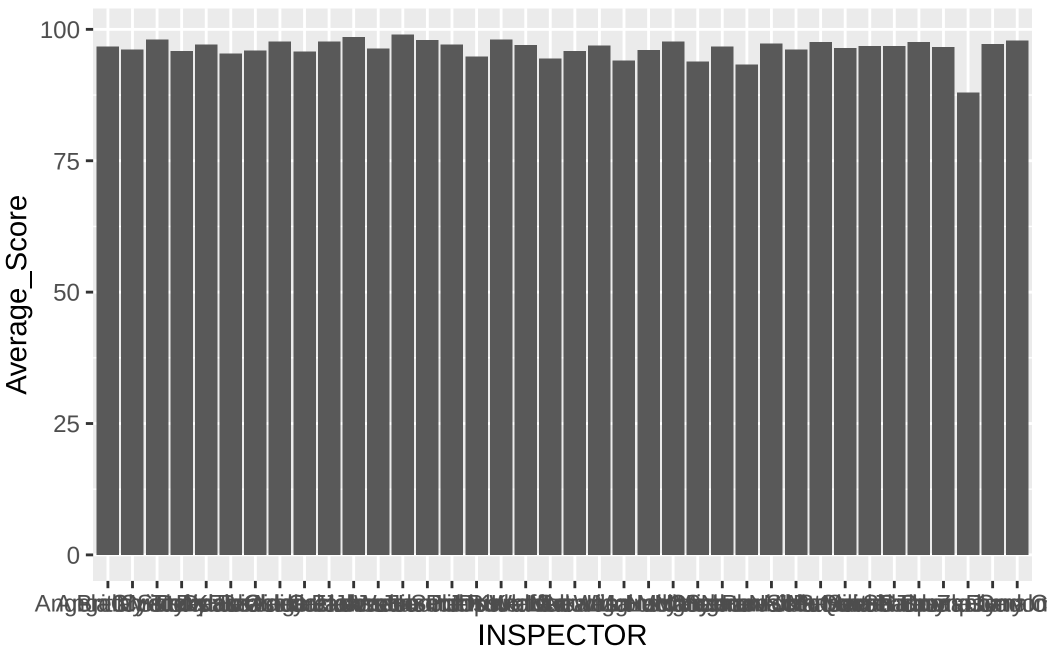


Above is where I created a column bar plot of the average scores by citites, and analyzed that there is very little variance between the average scores of the cities, similar to the lack of variance between all food facilities.

Analysis 4 Inspector Score Variance

```r
inspector_scores <- restaurant_data |>
  group_by(INSPECTOR) |>
  summarize(Average_Score = mean(SCORE, na.rm = TRUE)) |>
  ungroup()

ggplot(inspector_scores, aes(x = INSPECTOR, y= Average_Score)) +
  geom_col()
```

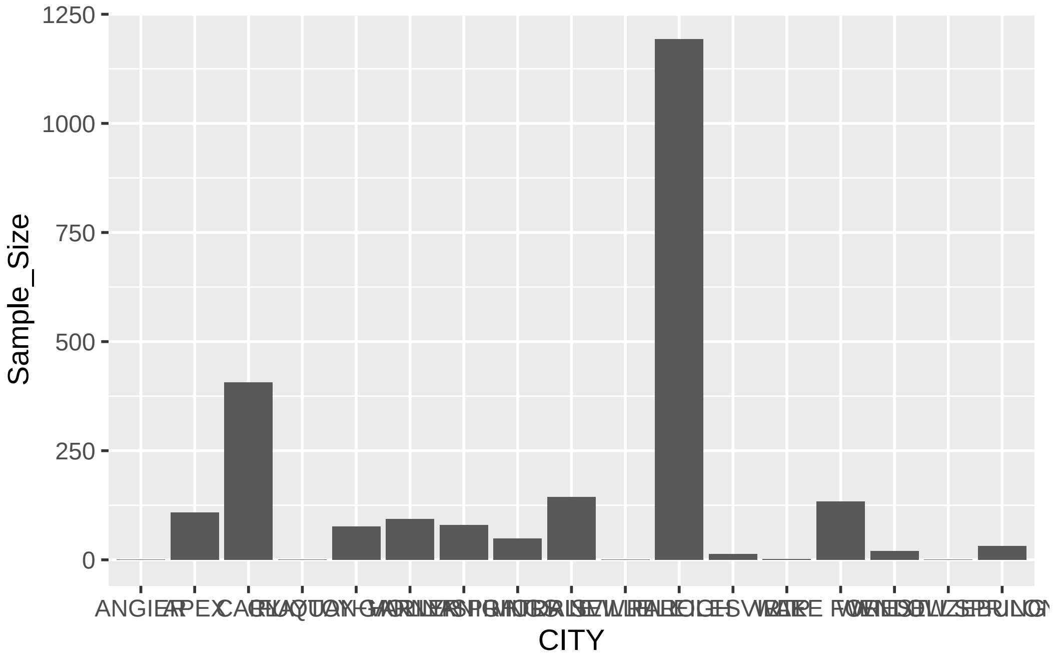


This graph also seems to follow the trend of the previous one, which had average scores by the inspectors around the same mark, with one individual with a significantly lower average score as well.

Analysis of Sample Size

```r
city_sample_sizes <- restaurant_data |>
  group_by(CITY) |>
  summarise(Sample_Size = n()) |>
  ungroup()

ggplot(city_sample_sizes, aes (x=CITY, y= Sample_Size)) +
  geom_col()
```



Here, rather than using a list like I did previously, decided to create a bar graph to showcase the differing degrees of samples sizes. This looks to be even more extreme than the previous one that analyzed all food facilities, as none are even close to matching the highest sample amount.

https://github.com/LucasWMayo/plan372-hw2

Above is the requested link to the Github repository.