Restaurant Data Analysis

AUTHOR
WGE

Restaurant Cleanliness Result Analysis

Initial setup

Here I install libraries I will need.

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(ggplot2)
library(stringr)
```

Here I put the data into a proper object to work with

```
restaurant_data <-read_csv("restaurant_inspections.csv")
```

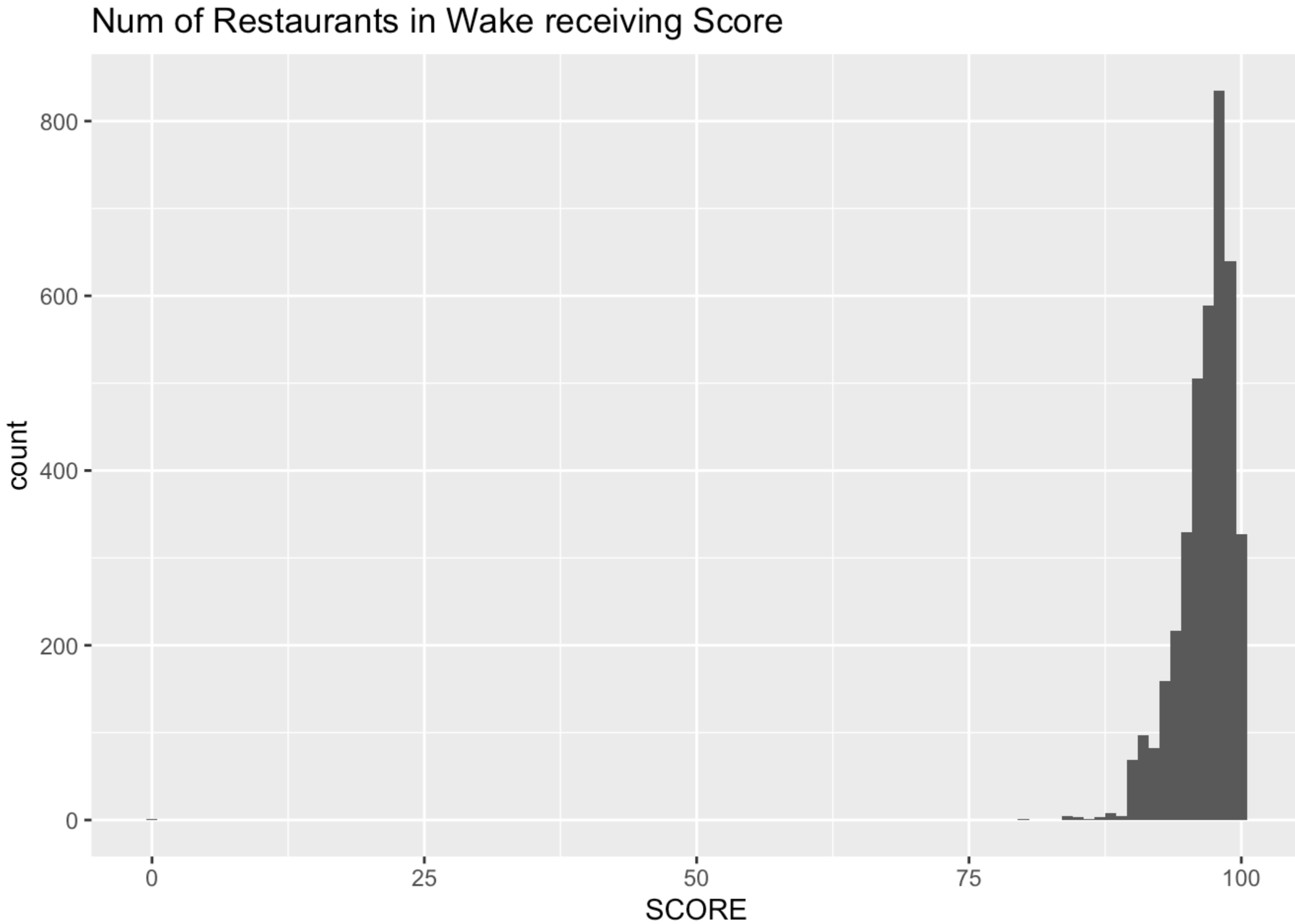
```
Rows: 3875 Columns: 12
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (8): HSISID, DESCRIPTION, TYPE, INSPECTOR, NAME, RESTAURANTOPENDATE, CI...
dbl  (3): OBJECTID, SCORE, PERMITID
dttm (1): DATE_

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.
```

Starting analysis

Now, I will make a histogram of restaurant inspection scores' distribution

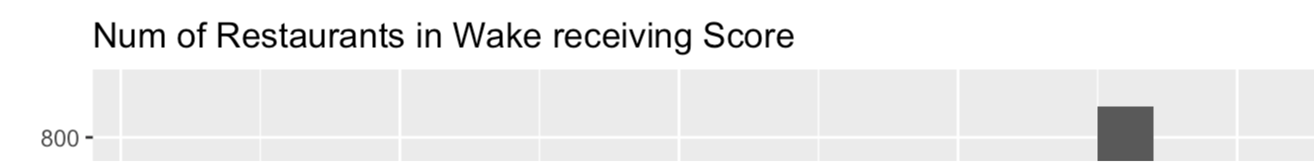
Now, I will make a histogram of restaurant inspection scores' distribution

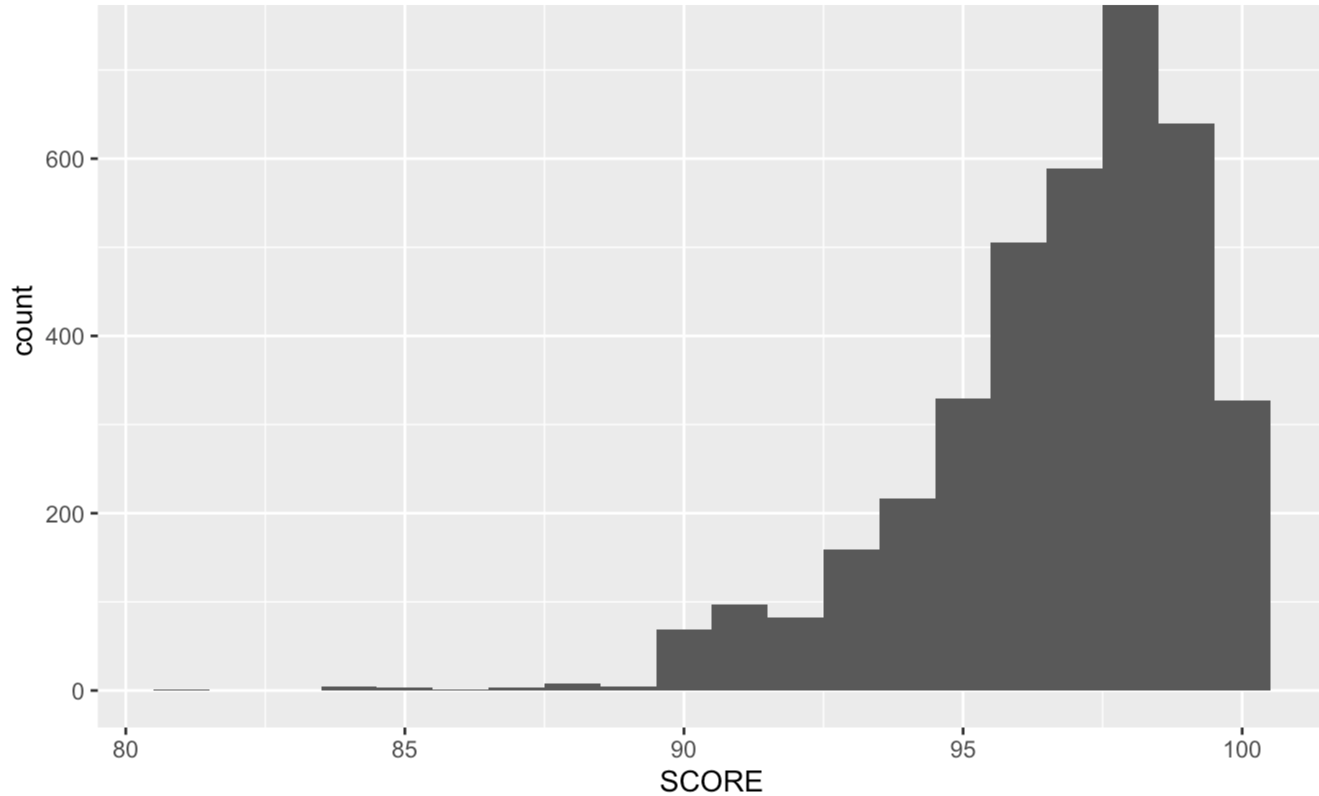
```
ggplot(data = restaurant_data,
       mapping = aes(x = SCORE)) +
  geom_histogram(binwidth = 1)+
  ggtitle("Num of Restaurants in Wake receiving Score")
```



As you can see, there's a strong tendency of the score to be in the low 80s at most–and yet it extends all the way to 0. Looking into the dataset, it can be found that there is a single 0 score in the dataset and only one–which can be discounted for graph purposes to look at the main substance of the binning better, at least.

```
nozeroset <- subset(restaurant_data, SCORE>0)

ggplot(data = nozeroset,
       mapping = aes(x = SCORE)) +
  geom_histogram(binwidth = 1)+
  ggtitle("Num of Restaurants in Wake receiving Score")
```





Much better for analysis!

Age of Restaurants

Some restaurants have been around a lot longer–is there a correlation between restaurant age and their general scoring tendency?

I have once again removed the 0 score value, as it makes the graph similarly annoying to read.

```
class(nozeroset$RESTAURANTOPENDATE)
```

```
[1] "character"
```

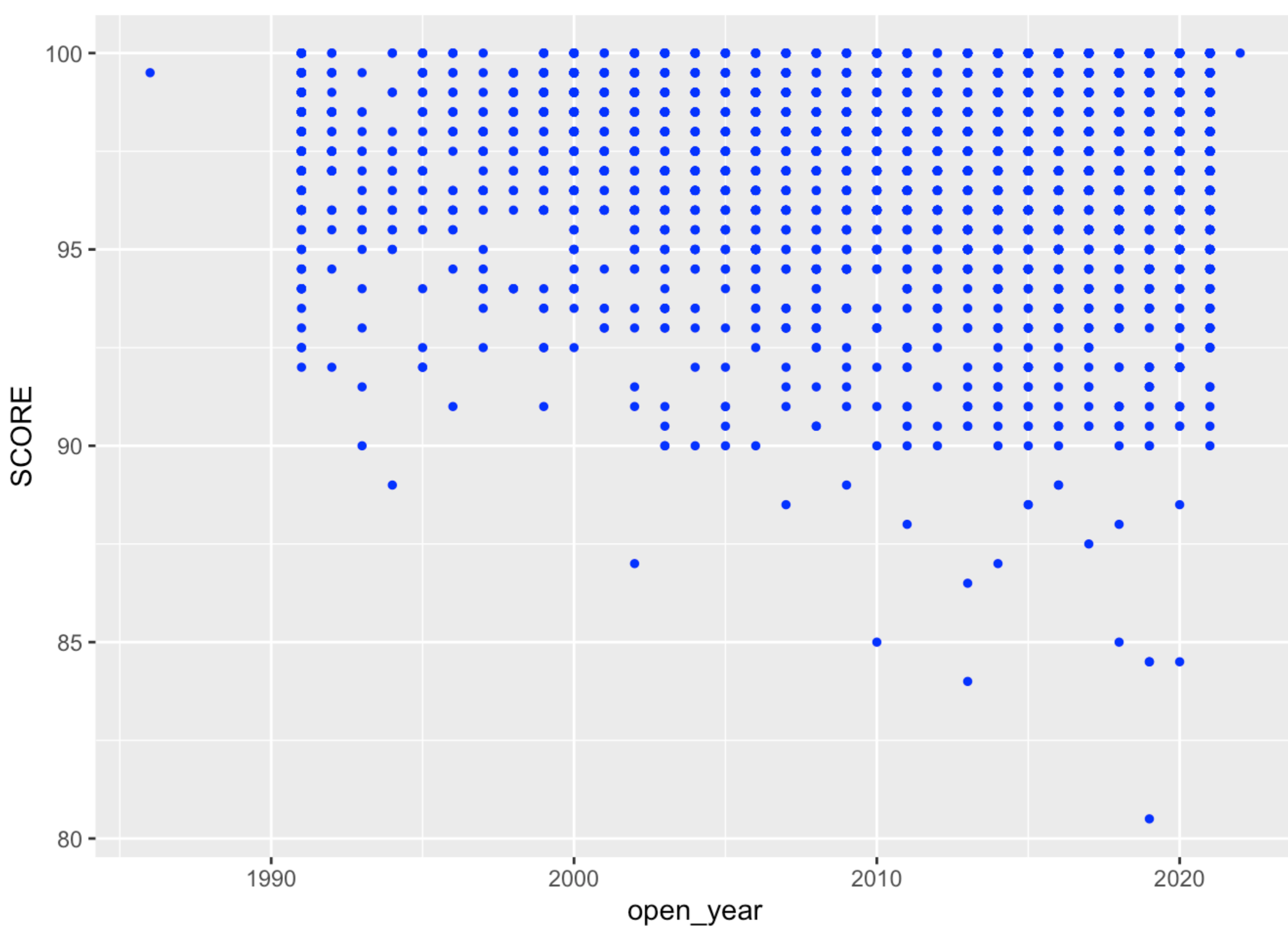
```
head(nozeroset)
```

```
# A tibble: 6 × 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>
```

```
#   FACILITYTYPE <chr>
```

```r
#2017-10-22 04:00:00

restaurant_data <- restaurant_data %>%
  mutate(date1 = str_sub(RESTAURANTOPENDATE, end = -13)) %>%
  mutate(open_date = as.Date(date1,format = "%Y/%m/%d"),
         open_year = year(open_date))

#had to cut off the timestamp at the end, then convert to a date object

#redefining nozeroset
nozeroset <- subset(restaurant_data, SCORE>0)


ggplot(data = nozeroset,mapping = aes(x = open_year, y = SCORE))+
  geom_point(col = "blue", size = 1)#+
```

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



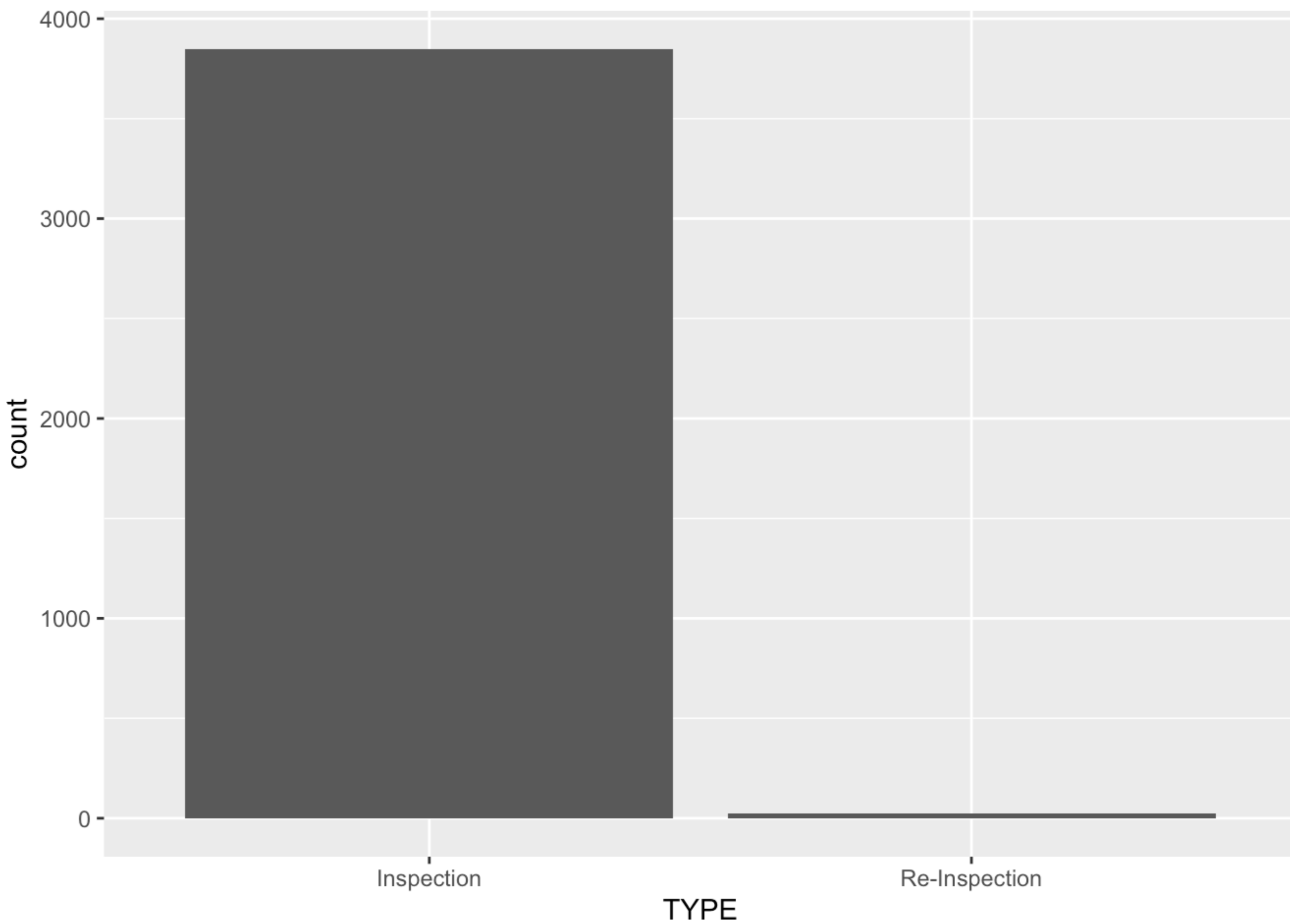

```r
  geom_smooth(col = "red")
```

```
geom_smooth: na.rm = FALSE, orientation = NA, se = TRUE
stat_smooth: na.rm = FALSE, orientation = NA, se = TRUE
position_identity
```

It does not appear to have that much correlation–there's a small trend towards score decreasing as the open year gets closer to the modern day, but it doesn't appear major.

My first hypothesis would be thinking that re-inspections could simply bump up likelihood of older restaurants having a better final score–but Re-Inspections appear to be quite rare in the dataset, as you can see.

```
ggplot(restaurant_data,mapping = aes(x = TYPE))+
  geom_bar()
```



Inspection scores by city

Do inspection scores vary by Wake County city? First, we must standardize the spellings of all the city names, and correct for errors

```
unique(restaurant_data$CITY)
```

```
 [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"
```

```
test <-restaurant_data %>%
  mutate(CITY = str_to_upper(CITY)) %>%
  mutate(CITY = str_replace(CITY,"RTP","RESEARCH TRIANGLE PARK")) %>%
  mutate(CITY = str_replace(CITY,"FUQUAY VARINA","FUQUAY-VARINA")) %>%
  mutate(CITY = str_replace(CITY,"HOLLY SPRING","HOLLY SPRINGS")) %>%
  mutate(CITY = str_replace(CITY,"HOLLY SPRINGSS","HOLLY SPRINGS")) %>%
  mutate(CITY = str_replace(CITY,"MORRISVILE","MORRISVILLE"))

unique(test$CITY)
```

```
 [1] "CARY"                 "RALEIGH"              "KNIGHTDALE"
 [4] "CLAYTON"              "FUQUAY-VARINA"        NA
 [7] "GARNER"               "MORRISVILLE"          "RESEARCH TRIANGLE PARK"
[10] "WENDELL"              "APEX"                 "WILLOW SPRING"
[13] "HOLLY SPRINGS"        "ROLESVILLE"           "ZEBULON"
[16] "WAKE FOREST"          "NEW HILL"             "ANGIER"
[19] "NORTH CAROLINA"
```

```
restaurant_data<-test

unique(restaurant_data$CITY)
```

```
 [1] "CARY"                 "RALEIGH"              "KNIGHTDALE"
 [4] "CLAYTON"              "FUQUAY-VARINA"        NA
 [7] "GARNER"               "MORRISVILLE"          "RESEARCH TRIANGLE PARK"
[10] "WENDELL"              "APEX"                 "WILLOW SPRING"
[13] "HOLLY SPRINGS"        "ROLESVILLE"           "ZEBULON"
[16] "WAKE FOREST"          "NEW HILL"             "ANGIER"
[19] "NORTH CAROLINA"
```

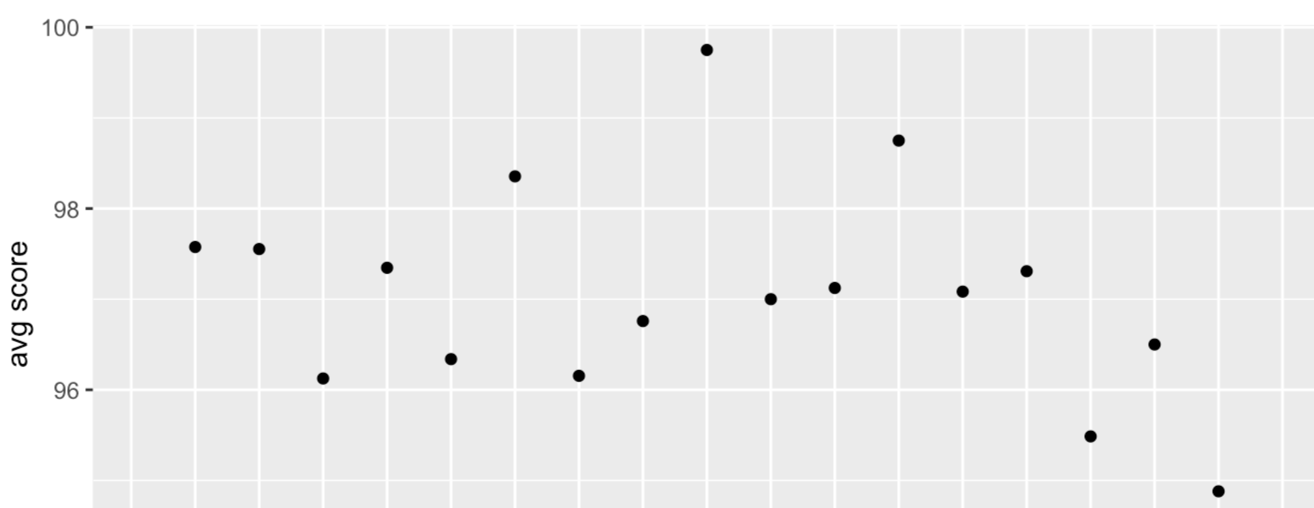
```
#RTP -> research triangle park
#FQ-V -> FQ v
#holly spring -> holly springs
#morris vile -> ville
```

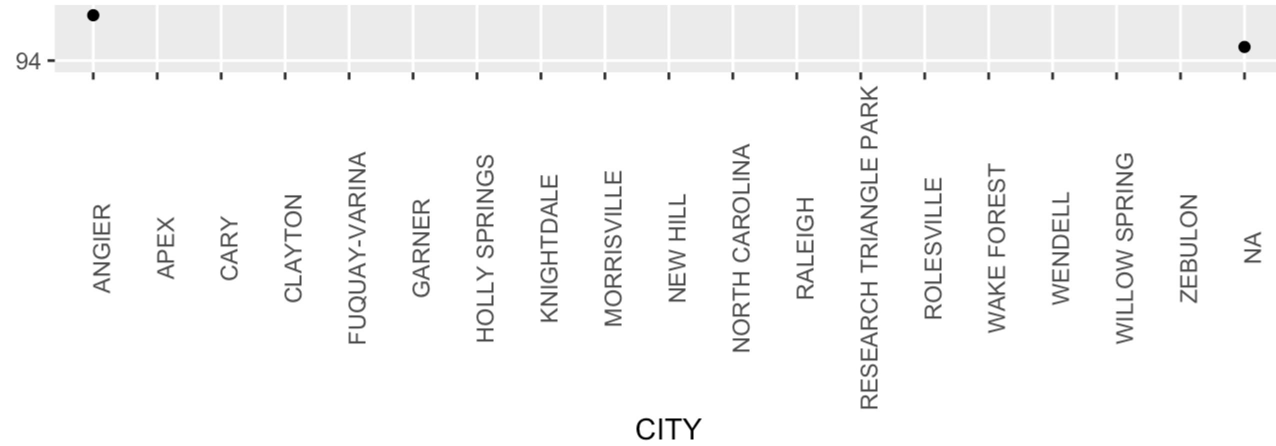
Now, with 19 unique locations, 18 discounting N/A...

```
test2 <- restaurant_data %>%
  group_by(CITY) %>%
  summarize(city_average_inspection = mean(SCORE))

test2
```

```
# A tibble: 19 × 2
   CITY                      city_average_inspection
   <chr>                                       <dbl>
 1 ANGIER                                       94.5
 2 APEX                                         97.6
 3 CARY                                         97.6
 4 CLAYTON                                      96.1
 5 FUQUAY-VARINA                                97.3
 6 GARNER                                       96.3
 7 HOLLY SPRINGS                                98.4
 8 KNIGHTDALE                                   96.2
 9 MORRISVILLE                                  96.8
10 NEW HILL                                     99.8
11 NORTH CAROLINA                               97
12 RALEIGH                                      97.1
13 RESEARCH TRIANGLE PARK                       98.8
14 ROLESVILLE                                   97.1
15 WAKE FOREST                                  97.3
16 WENDELL                                      95.5
17 WILLOW SPRING                                96.5
18 ZEBULON                                      94.9
19 <NA>                                         94.2
```

```
ggplot(test2, mapping = aes(x=CITY,y=city_average_inspection))+
  geom_point()+
  ylab("avg score")+
  theme(axis.text.x = element_text(angle=90))
```





It does appear to vary somewhat! incorporating other factors such as city average wealth, or city population, or competition level for restaurants, or a number of other elements could potentially be interesting to look into.

Inspector variance

Wake county has a team of inspectors, who have likely changed somewhat over the years to boot. Do inspection results vary by inspector?

We can use much the same technique as the previous section, hopefully with less cleaning-up beforehand.

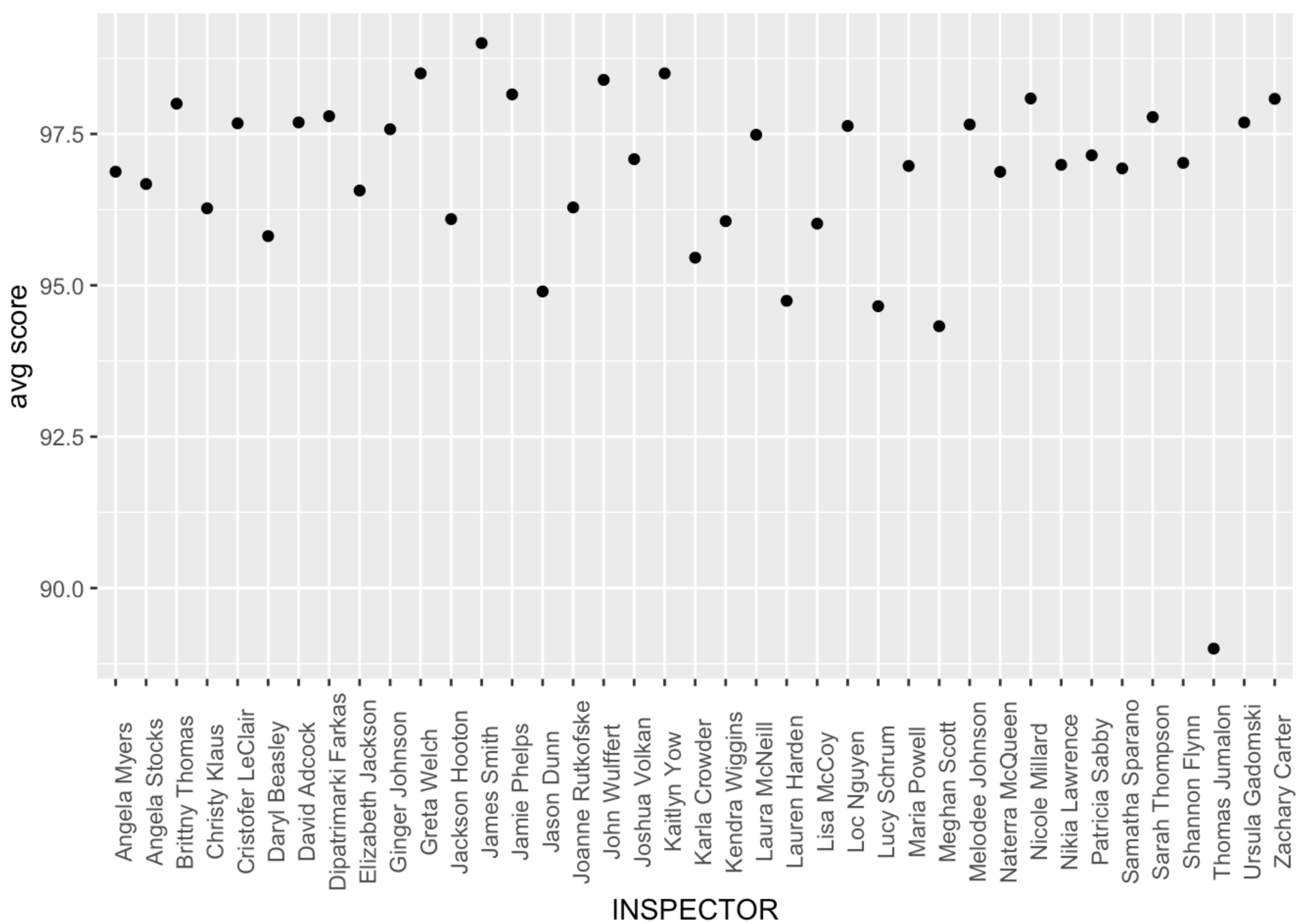
```
unique(restaurant_data$INSPECTOR)
```

```
 [1] "Karla Crowder"       "Meghan Scott"        "Kaitlyn Yow"
 [4] "Angela Myers"        "Patricia Sabby"      "Maria Powell"
 [7] "David Adcock"        "Jason Dunn"          "Laura McNeill"
[10] "Joanne Rutkofske"    "Nicole Millard"      "Loc Nguyen"
[13] "Brittny Thomas"      "Christy Klaus"       "Zachary Carter"
[16] "Greta Welch"         "Lucy Schrum"         "Ginger Johnson"
[19] "Jamie Phelps"        "John Wulffert"       "Naterra McQueen"
[22] "James Smith"         "Joshua Volkan"       "Lisa McCoy"
[25] "Ursula Gadomski"     "Cristofer LeClair"   "Shannon Flynn"
[28] "Jackson Hooton"      "Lauren Harden"       "Elizabeth Jackson"
[31] "Daryl Beasley"       "Dipatrimarki Farkas" "Samatha Sparano"
[34] "Melodee Johnson"     "Sarah Thompson"      "Thomas Jumalon"
[37] "Nikia Lawrence"      "Kendra Wiggins"      "Angela Stocks"
```

```
test3 <- restaurant_data %>%
  group_by(INSPECTOR) %>%
  summarize(average_by_inspector = mean(SCORE)) %>%
  ungroup()

test3
```

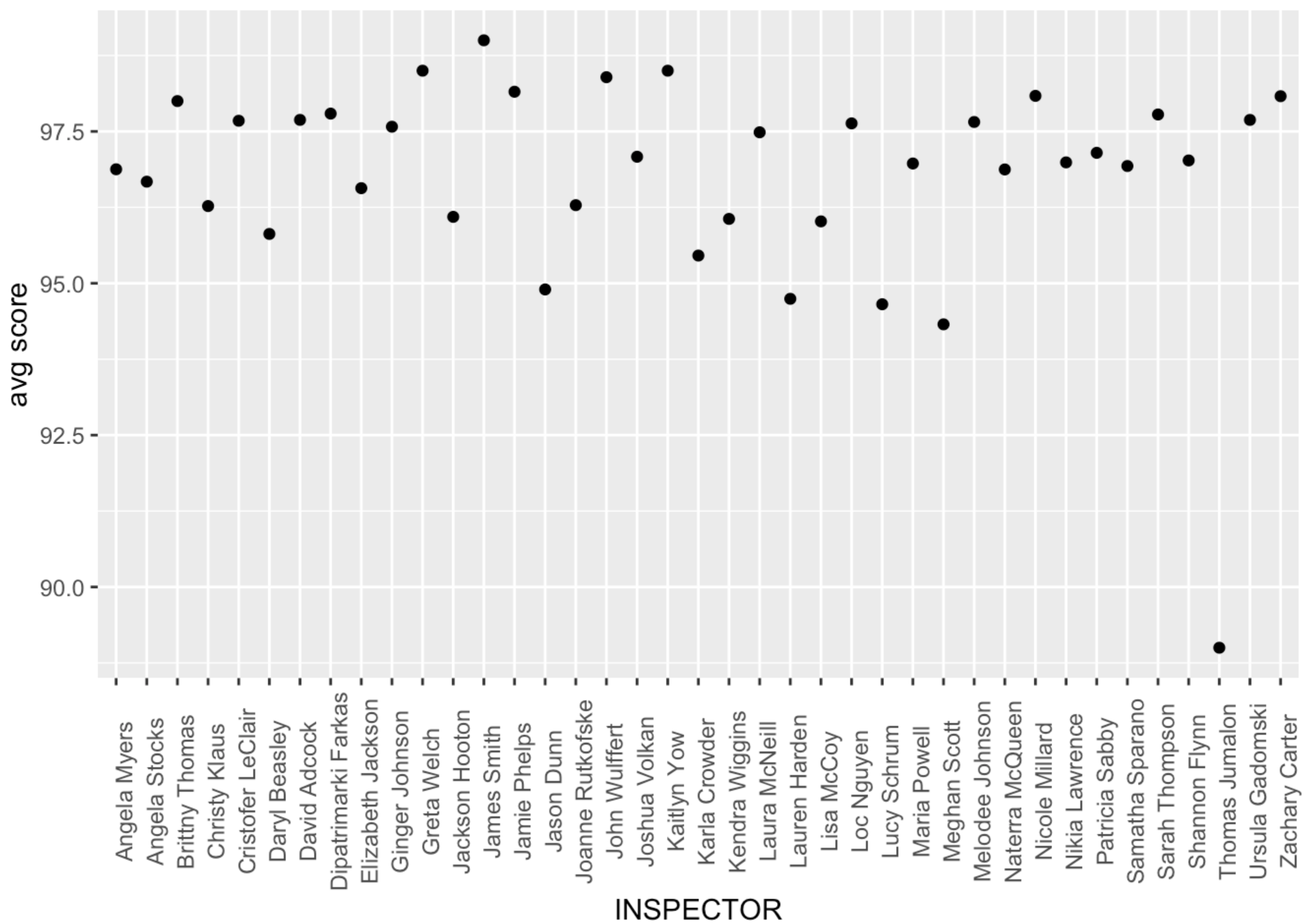
```
# A tibble: 39 × 2
   INSPECTOR            average_by_inspector
   <chr>                               <dbl>
 1 Angela Myers                         96.9
 2 Angela Stocks                        96.7
 3 Brittny Thomas                       98  
 4 Christy Klaus                        96.3
 5 Cristofer LeClair                    97.7
 6 Daryl Beasley                        95.8
 7 David Adcock                         97.7
 8 Dipatrimarki Farkas                  97.8
 9 Elizabeth Jackson                    96.6
10 Ginger Johnson                       97.6
# ℹ 29 more rows
```

```r
ggplot(test3, mapping = aes(x=INSPECTOR,y=average_by_inspector))+
  geom_point()+
  ylab("avg score")+
  theme(axis.text.x = element_text(angle=90))
```



Strange! there appears to be one particular wild outlier from the largely-homogenous averages...But I wonder if this might be our culprit from earlier graphs.

```
#renewing it
nozeroset <- subset(restaurant_data, SCORE>0)

test4 <- restaurant_data %>%
  group_by(INSPECTOR) %>%
  summarize(average_by_inspector = mean(SCORE)) %>%
  ungroup()

ggplot(test4, mapping = aes(x=INSPECTOR,y=average_by_inspector))+
  geom_point()+
  ylab("avg score")+
  theme(axis.text.x = element_text(angle=90))
```



Fascinatingly, it isn't! examining the dataset indicates that it's Inspector Meghan Scott who assigned the lone 0 score–not this outlier Inspector Thomas Jumalon. What could be causing this?

##Sample Sizes

Perhaps it's the sample size to blame? How many inspections has each inspector carried out?

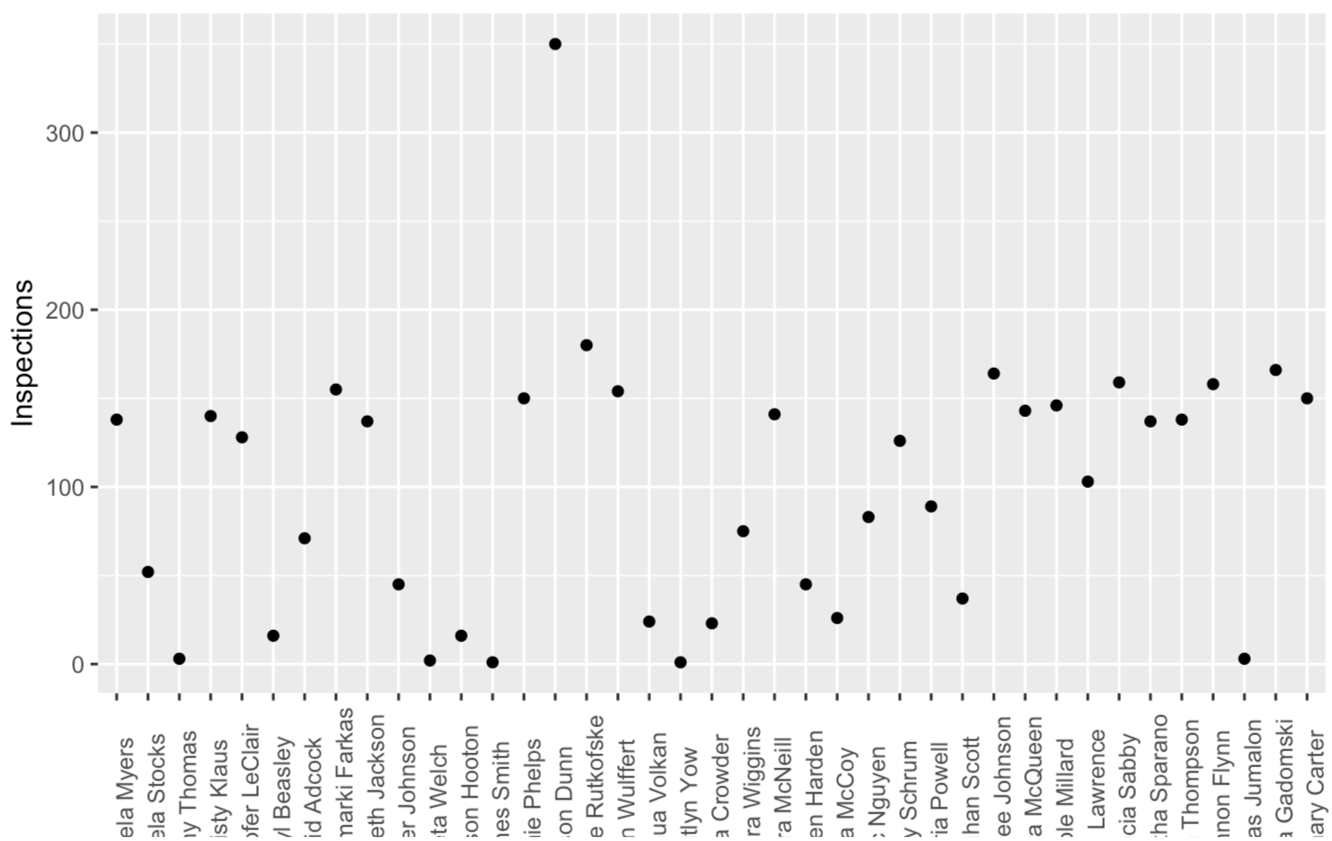
```
samplesize_inspector <-restaurant_data %>%
  mutate(inspected = 1) %>%
  group_by(INSPECTOR) %>%
```

```
  group_by(INSPECTOR) %>%
  summarize(inspections = sum(inspected)) %>%
  ungroup()

samplesize_inspector
```

```
# A tibble: 39 × 2
   INSPECTOR           inspections
   <chr>                     <dbl>
 1 Angela Myers                138
 2 Angela Stocks                52
 3 Brittny Thomas                3
 4 Christy Klaus               140
 5 Cristofer LeClair           128
 6 Daryl Beasley                16
 7 David Adcock                 71
 8 Dipatrimarki Farkas         155
 9 Elizabeth Jackson           137
10 Ginger Johnson               45
# ℹ 29 more rows
```

Illustrated on a graph, this becomes

```
ggplot(samplesize_inspector, mapping = aes(x=INSPECTOR,y=inspections))+
  geom_point()+
  ylab("Inspections")+
  theme(axis.text.x = element_text(angle=90))
```



And here the true culprit of the outlier is revealed–the fact that Thomas Jumalon only performed 3 inspections, along with a number of other inspectors who performed few looks. Thomas Jumalon likely happened to make a few lower grades in those mere 3 inspections, without the evident average of 95+ – creating a stark outlier.

In fact, we can check what his were.

```
jumalon <- subset(restaurant_data, INSPECTOR=="Thomas Jumalon")

jumalon
```

```
# A tibble: 3 × 15
  OBJECTID HSISID SCORE DATE_               DESCRIPTION TYPE  INSPECTOR PERMITID
     <dbl> <chr>  <dbl> <dttm>              <chr>       <chr> <chr>        <dbl>
1 25096315 04092…    91 2021-09-07 04:00:00 "Follow-Up… Insp… Thomas J…     1887
2 25131283 04092…    91 2022-01-27 05:00:00 "Inspectio… Insp… Thomas J…    21266
3 25126866 04092…    85 2022-01-28 05:00:00 "The facil… Insp… Thomas J…     9680
# ℹ 7 more variables: NAME <chr>, RESTAURANTOPENDATE <chr>, CITY <chr>,
#   FACILITYTYPE <chr>, date1 <chr>, open_date <date>, open_year <dbl>
```

You can see that of his 3 inspections, two were low and one was significantly low, relatively speaking, leading to his notably outlying average score.

Restaurant relative to others

Are restaurants more cleanly than other types of food facilities that are inspected in this dataset?

```
unique(restaurant_data$FACILITYTYPE)
```

```
 [1] "Food Stand"                    "Restaurant"
 [3] "Mobile Food Units"             "Pushcarts"
 [5] NA                              "Elderly Nutrition Sites (catered)"
 [7] "Private School Lunchrooms"     "Meat Market"
 [9] "Institutional Food Service"    "Public School Lunchrooms"
[11] "Limited Food Service"
```

```
comparison_r <- restaurant_data %>%
  group_by(FACILITYTYPE) %>%
  summarize(avg_score = mean(SCORE)) %>%
  ungroup()

comparison_r
```

```
# A tibble: 11 × 2
   FACILITYTYPE                      avg_score
   <chr>                                 <dbl>
 1 Elderly Nutrition Sites (catered)      99.2
 2 Food Stand                             97.7
 3 Institutional Food Service             96.9
 4 Limited Food Service                   98.5
 5 Meat Market                            98.0
 6 Mobile Food Units                      98.1
 7 Private School Lunchrooms              98.5
 8 Public School Lunchrooms               99.2
 9 Pushcarts                              98.8
10 Restaurant                             96.7
11 <NA>                                   94.2
```

No, in fact, it appears as though restaurants are on the whole less cleanly than other facility types, bar none but "N/A".

However, I have a suspicion that...

```
samplesize_facil <-restaurant_data %>%
  mutate(pip = 1) %>%
  group_by(FACILITYTYPE) %>%
  summarize(locations = sum(pip)) %>%
  ungroup()

samplesize_facil
```

```
# A tibble: 11 × 2
   FACILITYTYPE                      locations
   <chr>                                 <dbl>
 1 Elderly Nutrition Sites (catered)         8
 2 Food Stand                              661
 3 Institutional Food Service               46
 4 Limited Food Service                      1
 5 Meat Market                              93
 6 Mobile Food Units                       181
 7 Private School Lunchrooms                13
 8 Public School Lunchrooms                185
 9 Pushcarts                                39
10 Restaurant                             2352
11 <NA>                                    296
```

...the sample size means that this kind of result is not surprising, although it may still be illustrative of something. It is likely that something is simply that there are far more restaurants than any other category by a factor of magnitude, and thus naturally they will vary far more, dragging the average down somewhat.

ANALYSIS FOR RESTAURANTS

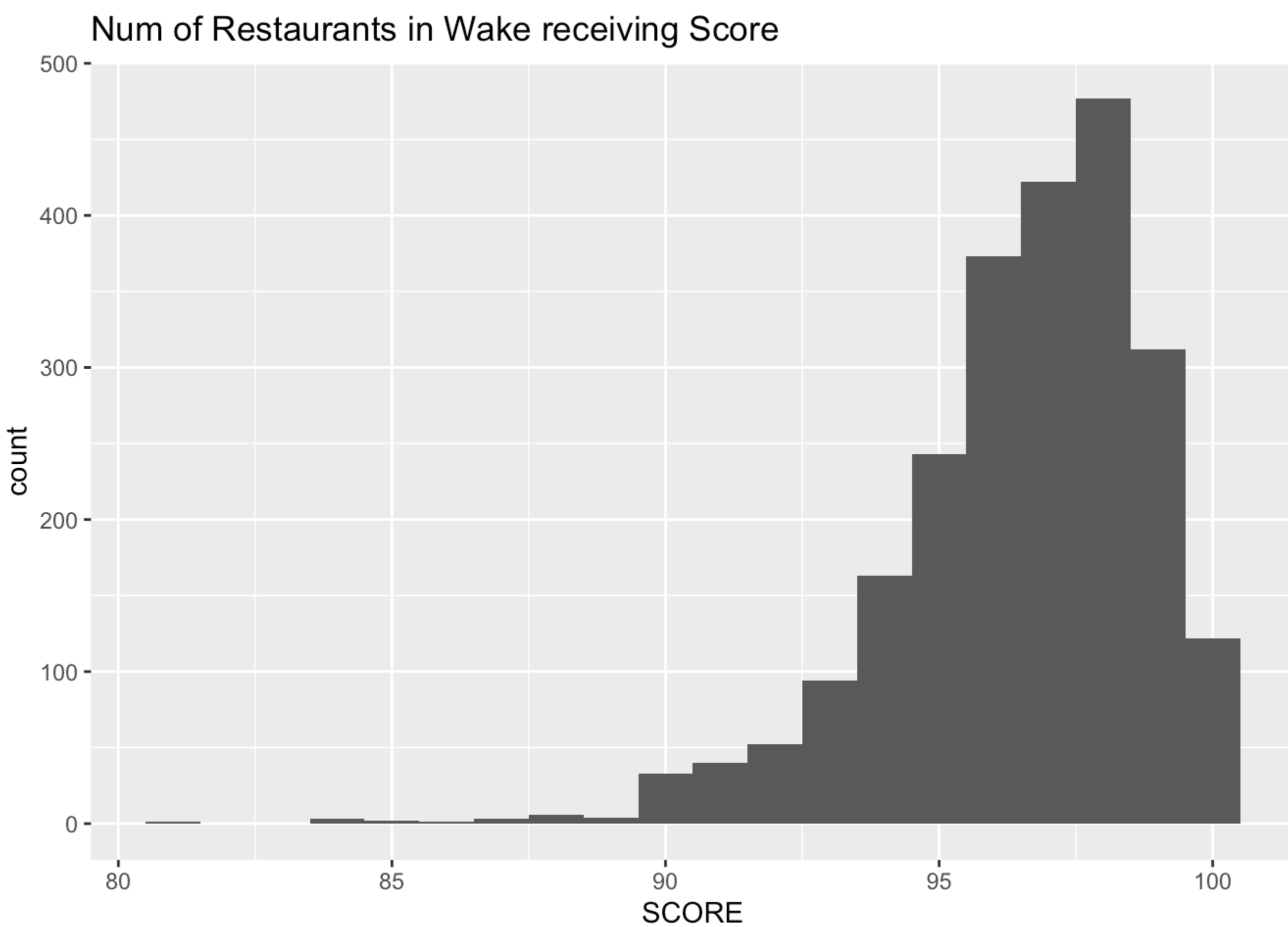
Since restaurants are where the general public is most likely to interact with the food-service system, Wake County Public Health is particularly interested in sanitation in restaurants.

Thus, here are the above analyses restricted to restaurant type facilities.

```
only_restaurants <- subset(restaurant_data, FACILITYTYPE=="Restaurant")
```

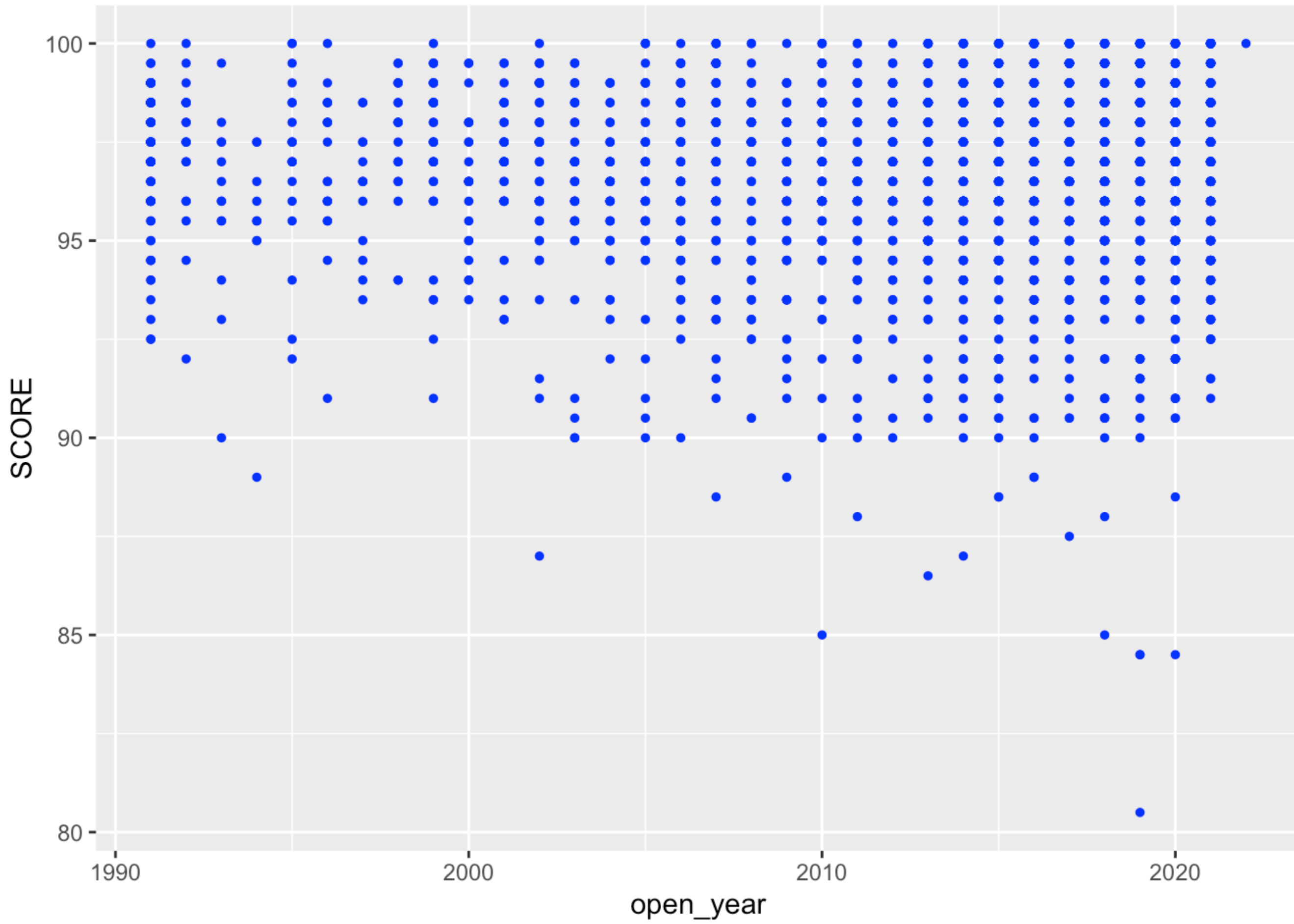
Histogram of overall scores of restaurants,

```
r_nozeroset <- subset(only_restaurants, SCORE>0)

ggplot(data = r_nozeroset,
       mapping = aes(x = SCORE)) +
  geom_histogram(binwidth = 1)+
  ggtitle("Num of Restaurants in Wake receiving Score")
```



Newer versus older Restaurants

```
ggplot(data = r_nozeroset,mapping = aes(x = open_year, y = SCORE))+
  geom_point(col = "blue", size = 1)#+
```




```
geom_smooth(col = "red")
```

```
geom_smooth: na.rm = FALSE, orientation = NA, se = TRUE
stat_smooth: na.rm = FALSE, orientation = NA, se = TRUE
position_identity
```

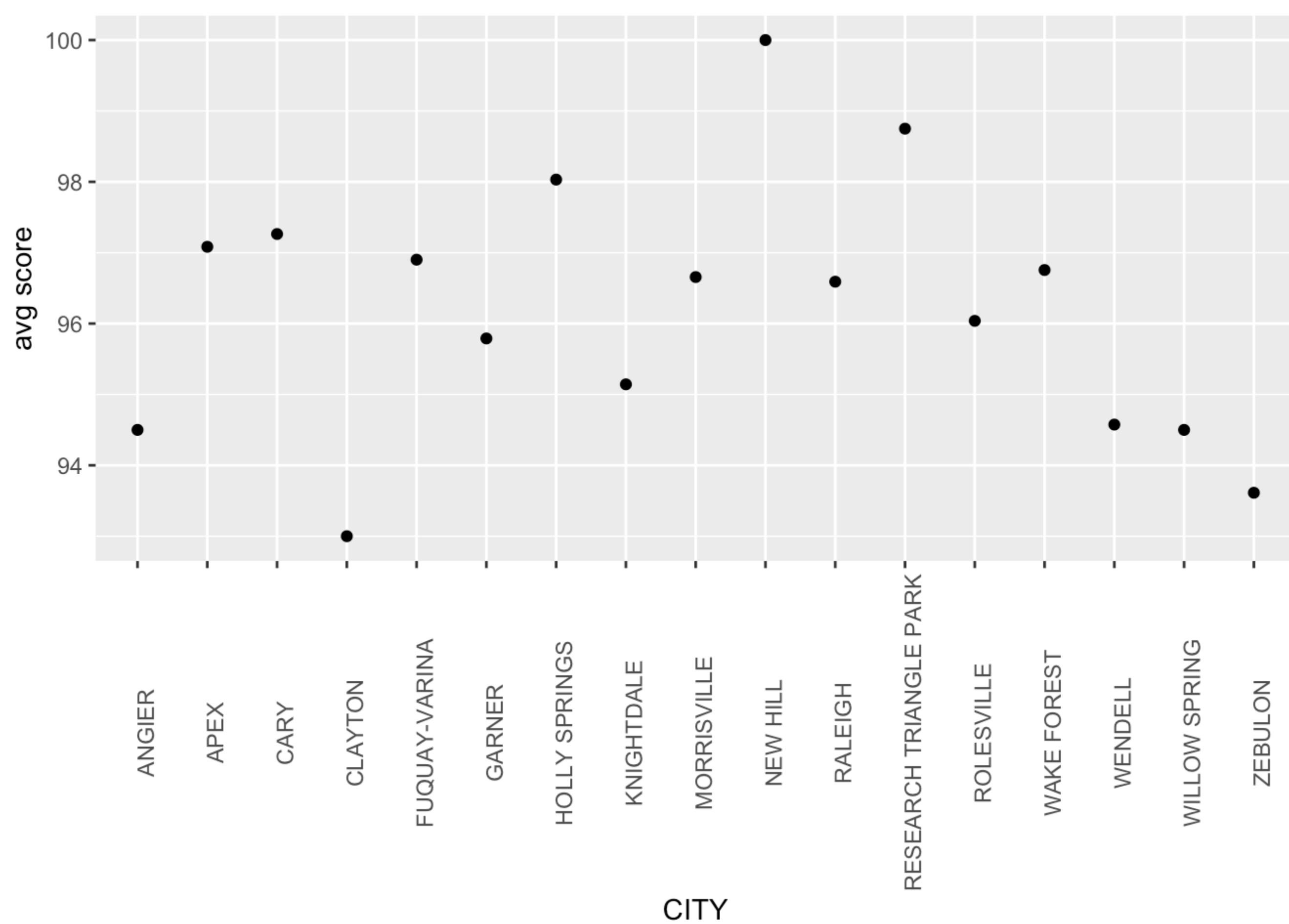
Variation by city–as I've already cleaned the data, I needn't do it again, luckily.

```
test5 <- only_restaurants %>%
  group_by(CITY) %>%
  summarize(city_average_inspection = mean(SCORE))

test5
```

```
# A tibble: 17 × 2
   CITY                    city_average_inspection
   <chr>                                     <dbl>
 1 ANGIER                                     94.5
 2 APEX                                       97.1
 3 CARY                                       97.3
 4 CLAYTON                                    93
```

```
 4 CLAYTON                                 93
 5 FUQUAY-VARINA                           96.9
 6 GARNER                                  95.8
 7 HOLLY SPRINGS                           98.0
 8 KNIGHTDALE                              95.1
 9 MORRISVILLE                             96.7
10 NEW HILL                               100
11 RALEIGH                                 96.6
12 RESEARCH TRIANGLE PARK                  98.8
13 ROLESVILLE                              96.0
14 WAKE FOREST                             96.8
15 WENDELL                                 94.6
16 WILLOW SPRING                           94.5
17 ZEBULON                                 93.6
```

```r
ggplot(test5, mapping = aes(x=CITY,y=city_average_inspection))+
  geom_point()+
  ylab("avg score")+
  theme(axis.text.x = element_text(angle=90))
```



It varies by city much like the previous data–and with somewhat different outliers, which is interesting. Clayton, for instance, dropped significantly without the evidently balancing influence of other food facilities–it would be interesting to analyze more deeply here, especially checking what kinds of food facilities are most common in what cities based on this notable shift.

kinds of food facilities are most common in what cities based on this notable shift...

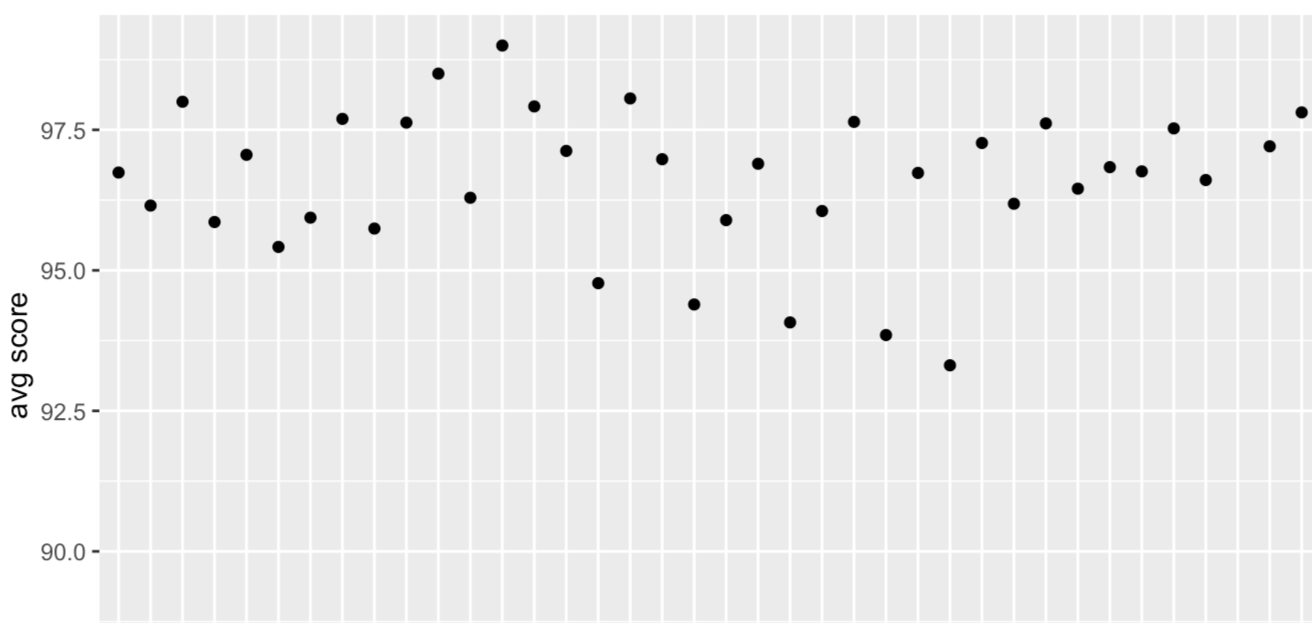
Similarly, inspector variance

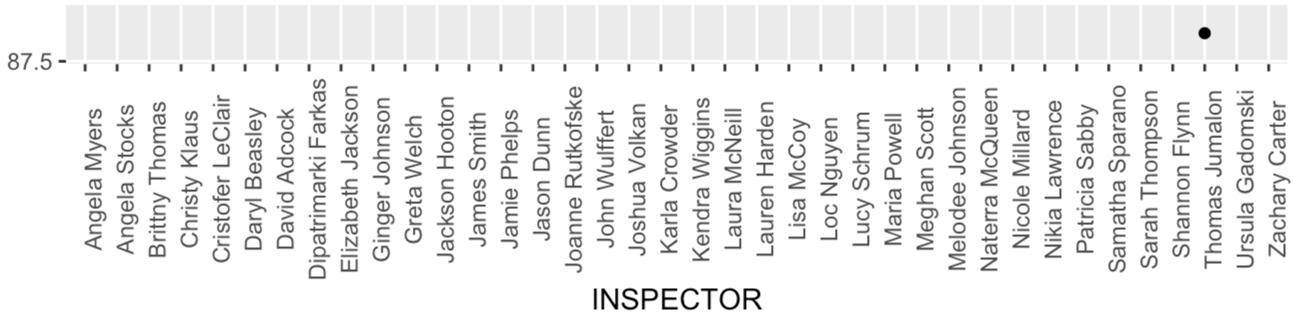
```
unique(only_restaurants$INSPECTOR)
```

```
 [1] "Meghan Scott"        "Maria Powell"        "Laura McNeill"
 [4] "Nicole Millard"      "Joanne Rutkofske"    "Loc Nguyen"
 [7] "Brittny Thomas"      "Patricia Sabby"      "Zachary Carter"
[10] "Greta Welch"         "Lucy Schrum"         "Ginger Johnson"
[13] "Jamie Phelps"        "John Wulffert"       "Naterra McQueen"
[16] "James Smith"         "Lisa McCoy"          "Ursula Gadomski"
[19] "Cristofer LeClair"   "Lauren Harden"       "Jackson Hooton"
[22] "Shannon Flynn"       "David Adcock"        "Elizabeth Jackson"
[25] "Daryl Beasley"       "Dipatrimarki Farkas" "Joshua Volkan"
[28] "Samatha Sparano"     "Melodee Johnson"     "Angela Myers"
[31] "Christy Klaus"       "Sarah Thompson"      "Karla Crowder"
[34] "Nikia Lawrence"      "Jason Dunn"          "Kendra Wiggins"
[37] "Angela Stocks"       "Thomas Jumalon"
```

of some interest is the fact that it was 39 previously–there are some inspectors who evidently only inspect non-restaurant facilities.

```
test6 <- only_restaurants %>%
  group_by(INSPECTOR) %>%
  summarize(average_by_inspector = mean(SCORE)) %>%
  ungroup()

ggplot(test6, mapping = aes(x=INSPECTOR,y=average_by_inspector))+
  geom_point()+
  ylab("avg score")+
  theme(axis.text.x = element_text(angle=90))
```




87.5
Angela Myers
Angela Stocks
Brittny Thomas
Christy Klaus
Cristofer LeClair
Daryl Beasley
David Adcock
Dipatrimarki Farkas
Elizabeth Jackson
Ginger Johnson
Greta Welch
Jackson Hooton
James Smith
Jamie Phelps
Jason Dunn
Joanne Rutkofske
John Wulffert
Joshua Volkan
Karla Crowder
Kendra Wiggins
Laura McNeill
Lauren Harden
Lisa McCoy
Loc Nguyen
Lucy Schrum
Maria Powell
Meghan Scott
Melodee Johnson
Naterra McQueen
Nicole Millard
Nikia Lawrence
Patricia Sabby
Samatha Sparano
Sarah Thompson
Shannon Flynn
Thomas Jumalon
Ursula Gadomski
Zachary Carter
INSPECTOR