Data Visualization from Scratch

NYC Data Science Academy Student: Zach Stone

Introduction

- NYC's Department of Health and Mental Hygiene (DOHMH) conducts unannounced inspections of restaurants at least once a year to check food handling, food temperature, personal hygiene, and vermin control. Since 2010, NYC restaurants have to prominently post their Grade (e.g. A/B/C) which empowers diners with decision-making information and incentivizes establishments to improve their hygiene.
- Download the csv file from here and rename it to `data.csv`. Move the file to the same directory of your markdown file and use relative path to read it.

The dataset was originally from NYC Open Data.

- After you finish the lab, please push your rmarkdown file (**w/o data**) to the corresponding folder under the homework repository.

1. Data Preprocessing

1.1 Load libraries

```
library(tidyverse)
```

1.2 Load dataset

```
raw_df = read.csv(url('https://lukepublicbucket.s3.us-east-2.amazonaws.com/nyc_dohmh_20210724.csv'))
write.csv(raw_df, 'data.csv')
raw.df <- readr::read_csv("data.csv")
```

1.3 Clean your data

1. Convert all the column names to lower cases and rename the cuisine description column to cuisine , inspection date`to`inspection.date‘.
2. Convert the `inspection.date` column from character to date format.
3. If we want to perform analysis for each month, then the date column would be pretty annoying because you have different date for each month. Mutate a new column called `month` that extract the month from `inspection.date` and convert it to numeric. For example, 03/24/2016 -> 03
4. We have scores for some of the observations but their grades are missing. Impute the missing values in grade column with the following rules:
   - 0 <= score < 14: A
   - 14 <= score < 28: B
   - score >= 28: C
   - You can ignore the other grades
5. Rename the description from the action column to something short so it won't blow up your graph.
   - "Violations were cited in the following area(s)." => "violations"
   - "Establishment Closed by DOHMH. Violations were cited in the following area(s) and those requiring immediate action were addressed." => "closed"

- "Establishment re-closed by DOHMH" => "reclosed"
- "No violations were recorded at the time of this inspection." => "no violations"
- "Establishment re-opened by DOHMH" => "reopened"
- Hint: `gsub()` function might be helpful. The function takes regular expression as the `pattern` parameter, which means `()` is treated as a special character. You might want to set `fixed=TRUE` in the `gsub()` function to leave the pattern as it is.
- **reclosed** means they failed the second time during the same inspection cycle and **reopened** means they passed.

6. We want to filter out missing values to make our life easier for further analysis.
   - Filter out missing values (if any) from the `boro` column.
   - Filter out missing values and negative values (if any) from the `score` column.
   - Filter out any `inspection date` that doesn't make any sense (if any).
7. Select the following columns from `raw.df`: `camis`, `boro`, `cuisine`, `inspection.date`, `action`, `score`, `grade`, `month`
8. Return only the unqiue inspections from the previous step and save it as a new data frame called `inspections`. The reason is one inspection might have multiple observations with different violation code but their actions are the same, so we count them as one.

```
clean_df = raw.df
```

```
# 1. Convert all the column names to lower cases and rename the columns that have empty space.
names(clean_df)  = names(clean_df) %>%
  str_to_lower() %>%
  str_replace_all(.,'cuisine.description','cuisine')

# 2. Convert the inspection.date column from character to date format.
clean_df$inspection.date  = as.Date(clean_df$inspection.date,format = '%m/%d/%y')

#3. If we want to perform analysis for each month, then the date column would be pretty annoying becaus

clean_df = clean_df %>%
  mutate(month = as.numeric(format(inspection.date,'%m')))


# 4. We have scores for some of the observations but their grades are missing. Impute the missing value
#     + 0 <= score < 14: A
#     + 14 <= score < 28: B
#     + score >= 28: C
#     + You can ignore the other grades


clean_df$grade = if_else(condition = clean_df$grade == "",
        true = cut(clean_df$score, breaks = c(0,14,28, Inf), labels = c('A', 'B', 'C')),
        false = factor(clean_df$grade))
```

```
## Warning in `[<-.factor`(`*tmp*`, i, value = structure(c(2L, 2L, 1L, 1L, :
## invalid factor level, NA generated
```

```
# 5. Rename the description from the action column to something short so it won't blow up your graph.
clean_df$action = clean_df$action %>%
  gsub(pattern = 'Violations were cited in the following area(s).', replacement = 'violations', fixed =
  gsub(pattern = 'Establishment Closed by DOHMH. Violations were cited in the following area(s) and thos
  gsub(pattern = 'Establishment re-opened by DOHMH.', replacement = 'reopened', fixed = TRUE) %>%
  gsub(pattern = 'No violations were recorded at the time of this inspection.', replacement = 'no viola
  gsub(pattern = 'Establishment re-closed by DOHMH.', replacement = 'reclosed', fixed = TRUE)
```

```r
# 6. We want to filter out missing values to make our lives easier for further analysis.
clean_df = clean_df %>%
  filter(boro != 0) %>%
  filter(score >= 0)

# 7. Select the following columns from raw.df: camis, boro, cuisine, inspection.date, action, score, gr
# 8. Unique inspections

inspections = clean_df %>%
  select(camis, boro, cuisine, inspection.date, action, score, grade, month) %>%
  unique()
```

2. Data Visualization

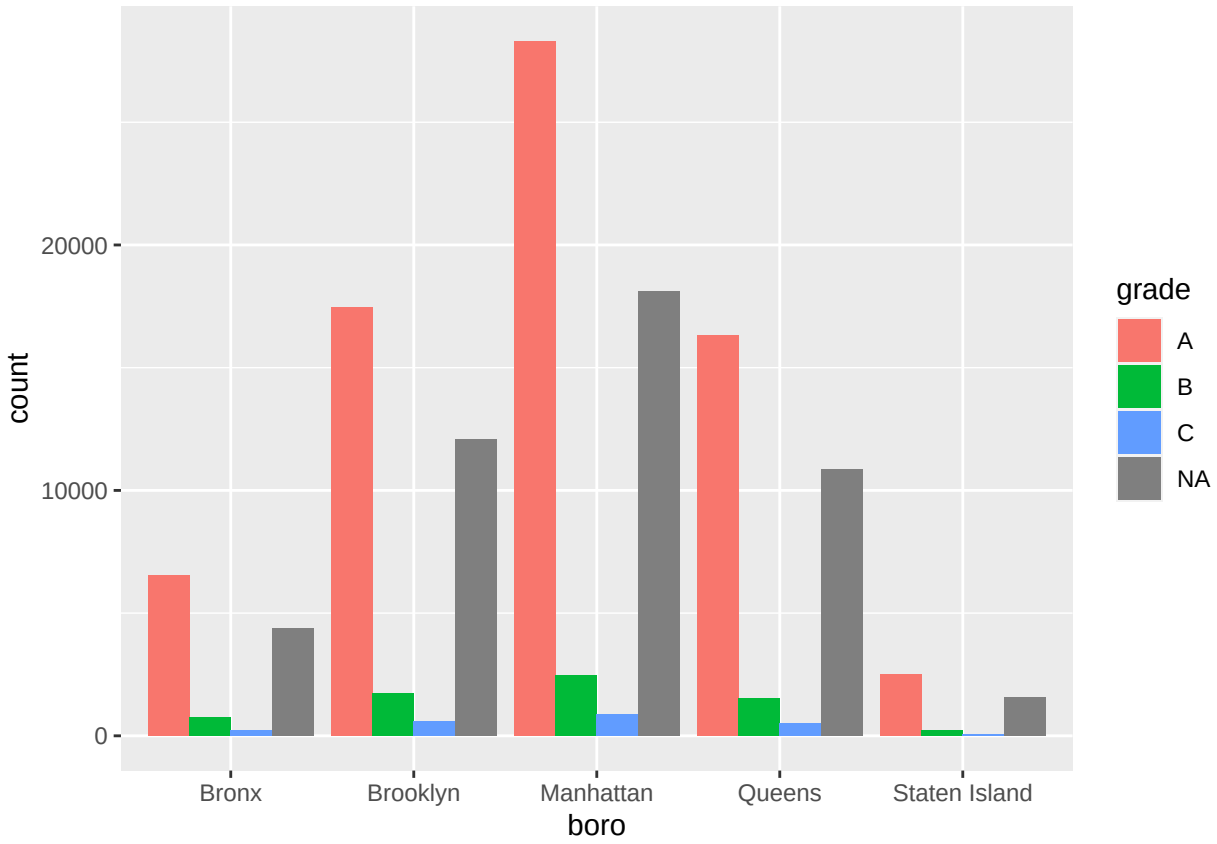
Example questions we want to answer from this dataset

- How do a restaurant's location and type of cuisine affect its inspection results?
- Do restaurants in Manhattan have better scores than those of restaurants in Queens or the Bronx?
- Are Manhattan restaurants cleaner than those in Queens or the Bronx?
- Do restaurants of your favorite cuisine perform better or worse in health inspections than the rest of the restaurants?

2.1 Inspection Grade vs Borough

- What is your conclusion?

```r
#Bar plot: Restaurants by borough and grade
#Unclear what this means - the same restaurant can have many different grades in the dataset. Tallying
inspections %>%
  ggplot(data =., aes(x = boro)) +
    geom_bar(position = 'dodge', aes(fill = grade))
```



Manhattan has the most restaurant inspections, and it looks like the most A's. However, when I looked at a 'fill' bar graph, the proportions looked similar across boroughs.

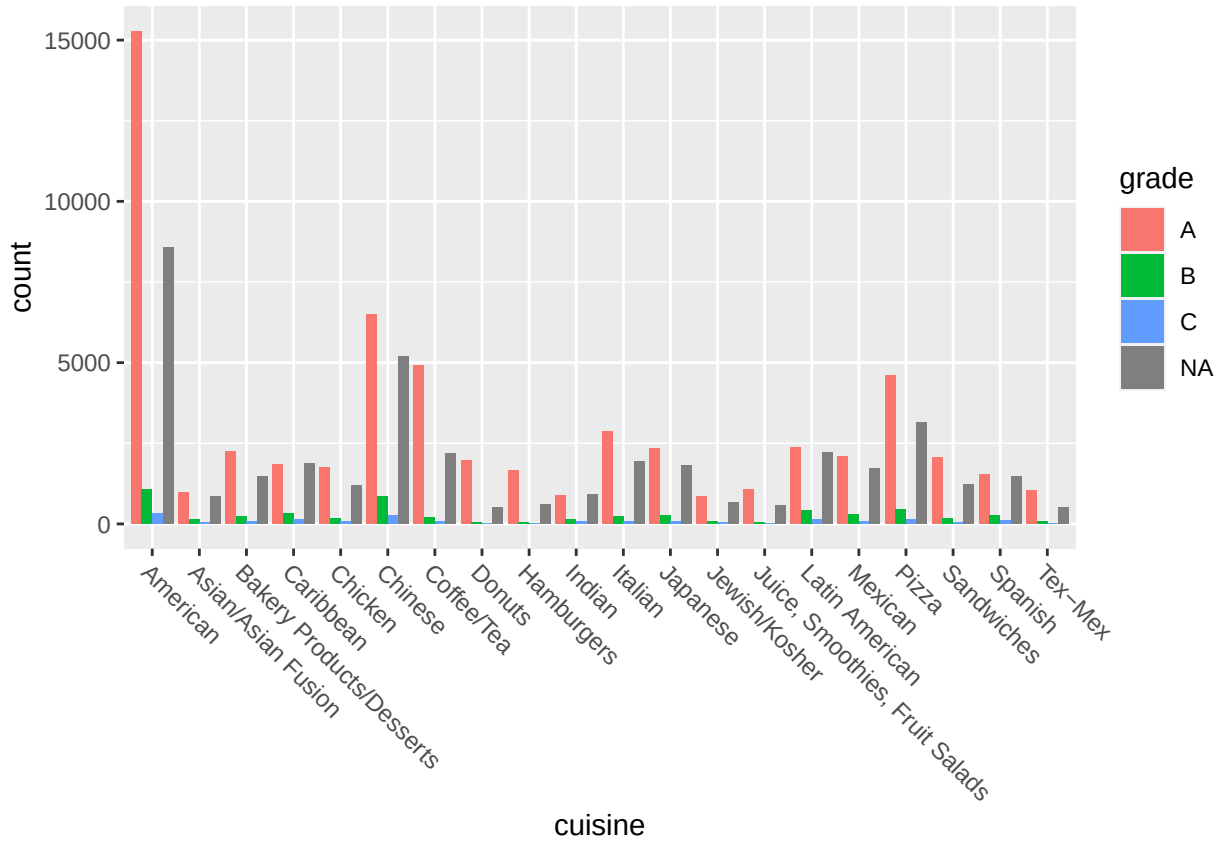
2.2 Inspection Grade vs Cuisine

- There are too many cuisine in the dataset and it will make the plot hard to read. Let's just focus on the top 20 cuisines.
- What is your conclusion?

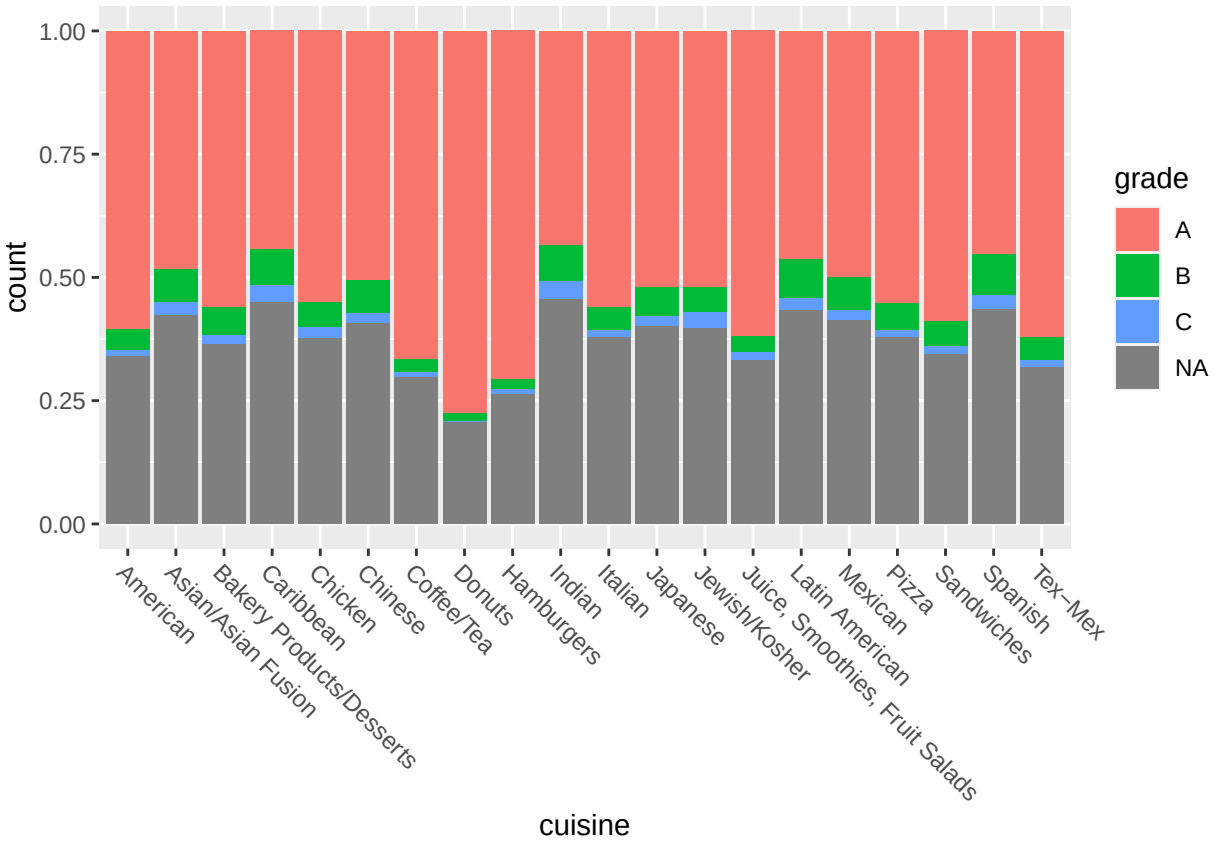
```
#Bar plot: Restaurants by cuisine and gradec

cuisine20 = summarize(group_by(inspections,cuisine),count = n()) %>%
  slice_max(count, n= 20)

semi_join(x = inspections, y = cuisine20, "cuisine") %>%
  ggplot(data =., aes(x=cuisine)) +
    geom_bar(aes(fill = grade), position = 'dodge') +
  theme(axis.text.x=element_text(angle=-45, hjust = 0))
```




```
semi_join(x = inspections, y = cuisine20, "cuisine") %>%
  ggplot(data =., aes(x=cuisine)) +
    geom_bar(aes(fill = grade), position = 'fill') +
  theme(axis.text.x=element_text(angle=-45, hjust = 0))
```



We can see that inspections of small-item/order restaurants, like donuts, coffee/tea, hamburgers, and juice bars have a smaller ratio of non-A restaurants.

2.3 Scores vs. Borough

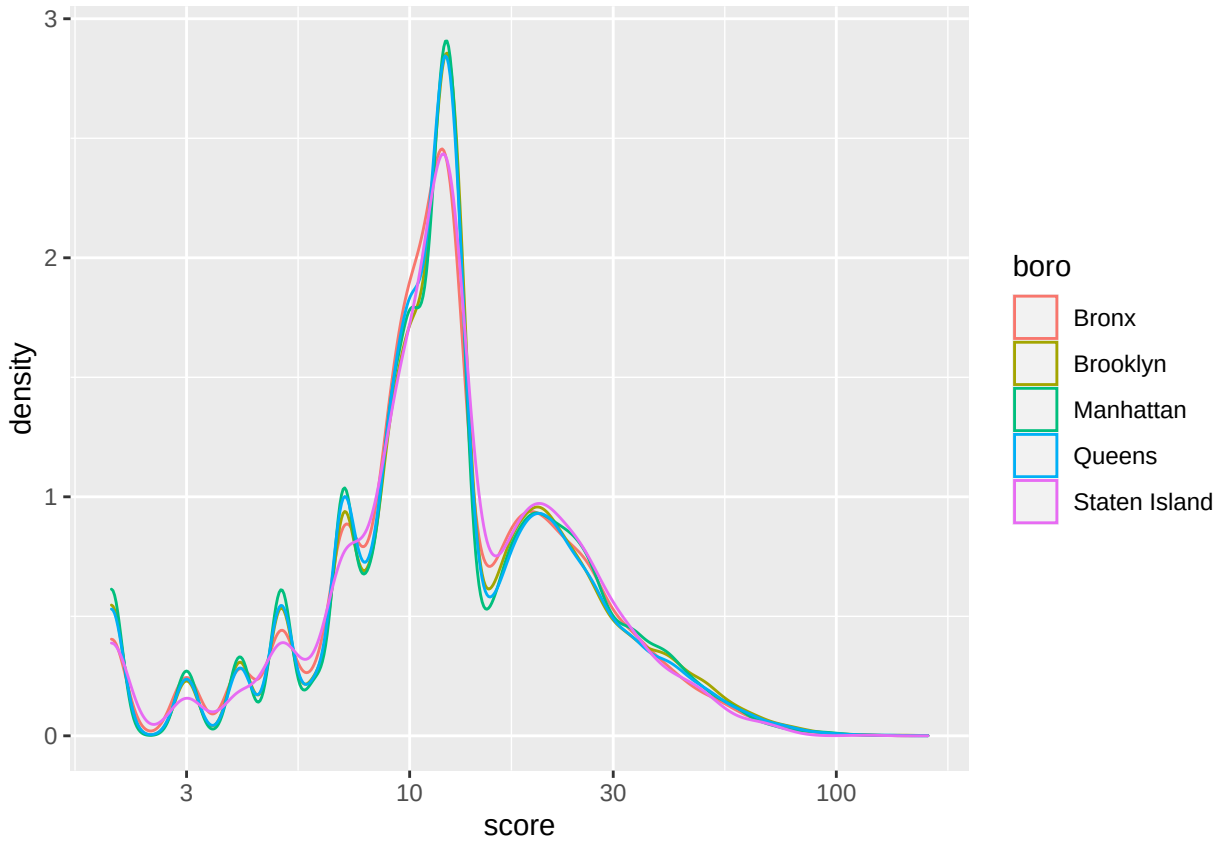
- Since grades couldn't differentiate boroughs, let's plot restaurants by scores instead and use a density plot to account for the disparity in number of restaurants by borough.
- What is your conclusion?

```
#Density plot: Restaurants by score and borough
#again, not sure what this means, since the data contain multiple inspections for each restaurant with
#I will plot the inspections by score instead of restaurants
g = ggplot(inspections, aes(x = score, color = boro))

g + geom_density() + scale_x_log10()
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 1574 rows containing non-finite values (stat_density).
```



Similar distribution among boroughs, slightly more spread out in Staten Island.

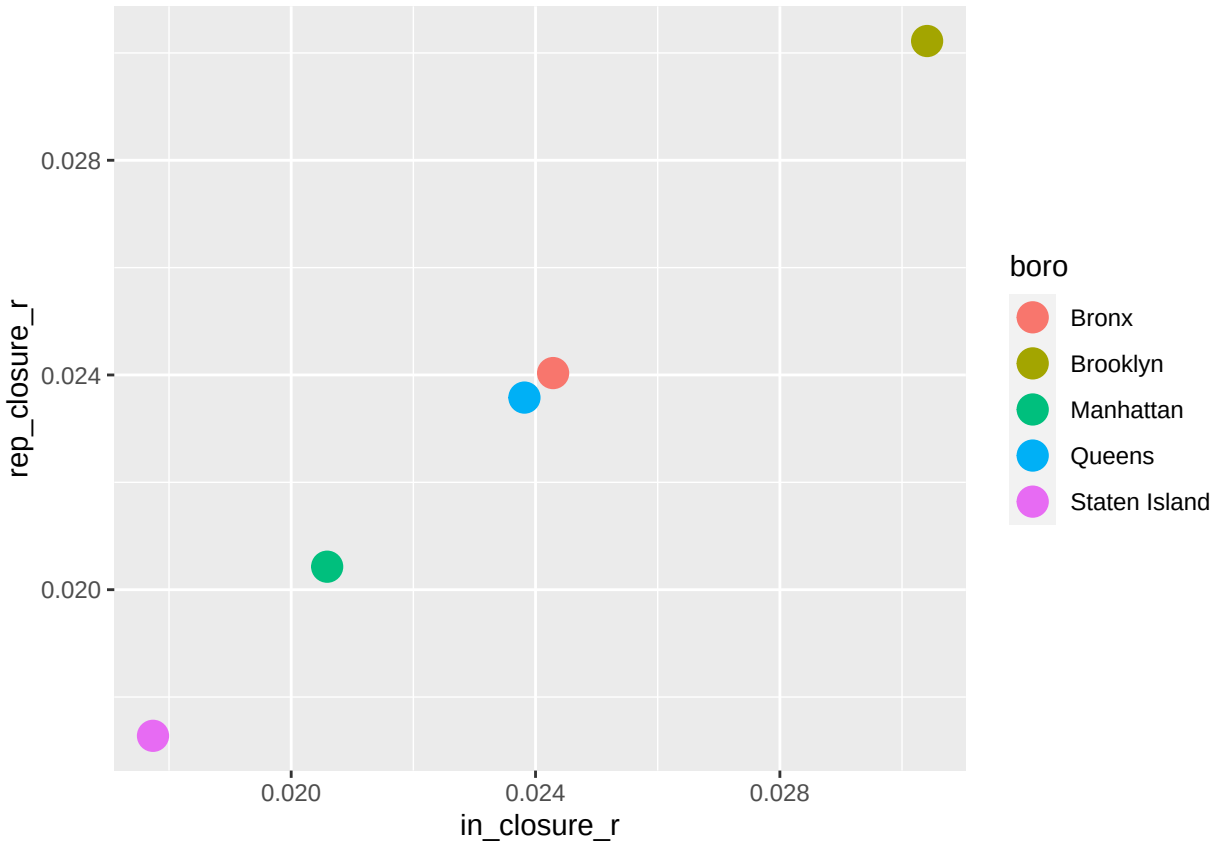
What about inspection closures?

- Scores don't tie directly to restaurant closures (e.g. public health hazard)

2.4 Closures vs. Borough

- Reclosed means they failed the second time during the same inspection cycle.
- Defined the following ratios:
  - Inspection closure ratio for each borough: % of inspections that lead to the restaurant being closed (including closed and reclosed)
  - Repeat closure ratio for each borough: % of restaurants that were closed more than once for different inspection cycles (just count the number of closed)
- What is your conclusion?

```
# It could be either a barplot with two different ratios for each borough or a scatterplot with two rat
inspections %>%
  group_by(boro) %>%
  summarize(in_closure_r = sum(action == 'closed' | action == 'reclosed')/n(),
            rep_closure_r = sum(action == 'closed')/n()) %>%
  ggplot(data = ., aes(x = in_closure_r, y = rep_closure_r, color = boro)) +
    geom_point(size = 5)
```



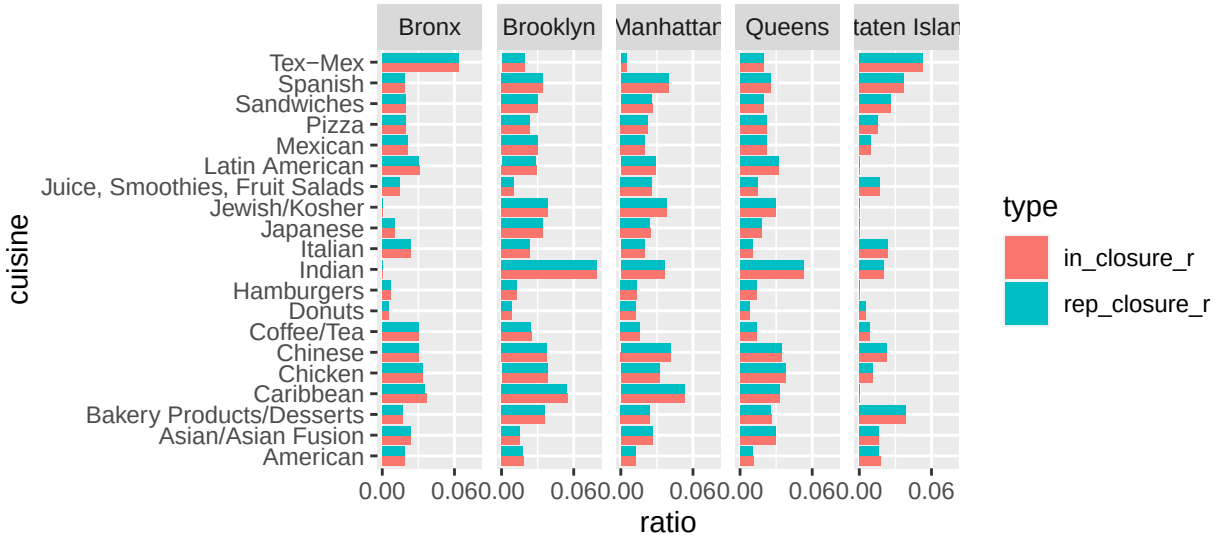
We can see a strict ordering by percentage of closures which holds along both ratios, with Staten Island having the smallest proportion, followed by Manhattan, Queens, Bronx, and Brooklyn with the most.

2.5 Closures vs. Cuisine and Borough

- Finally, what if we combined both dimensions of location and cuisine? Intuitively, certain cuisines could fare better or worse in health inspections depending on the neighborhood. Used faceted bar plots of inspection closure ratios by borough with the top 20 cuisine types.
- What is your conclusion?

```
semi_join(x = inspections, y = cuisine20, "cuisine") %>%
 group_by(boro, cuisine) %>%
  summarize(in_closure_r = sum(action == 'closed' | action == 'reclosed')/n(),
            rep_closure_r = sum(action == 'closed')/n()) %>%
  ungroup() %>%
  pivot_longer(values_to = 'ratio', cols = c(in_closure_r, rep_closure_r), names_to  = c('type')) %>%
  ggplot(data =., aes(y=cuisine, x = ratio, fill = type)) +
    geom_col(position = 'dodge') +
  scale_x_continuous(breaks = c(0,0.06)) +
    facet_grid(. ~ boro) +
    theme(aspect.ratio = 4)
```

```
## `summarise()` has grouped output by 'boro'. You can override using the
## `.groups` argument.
```



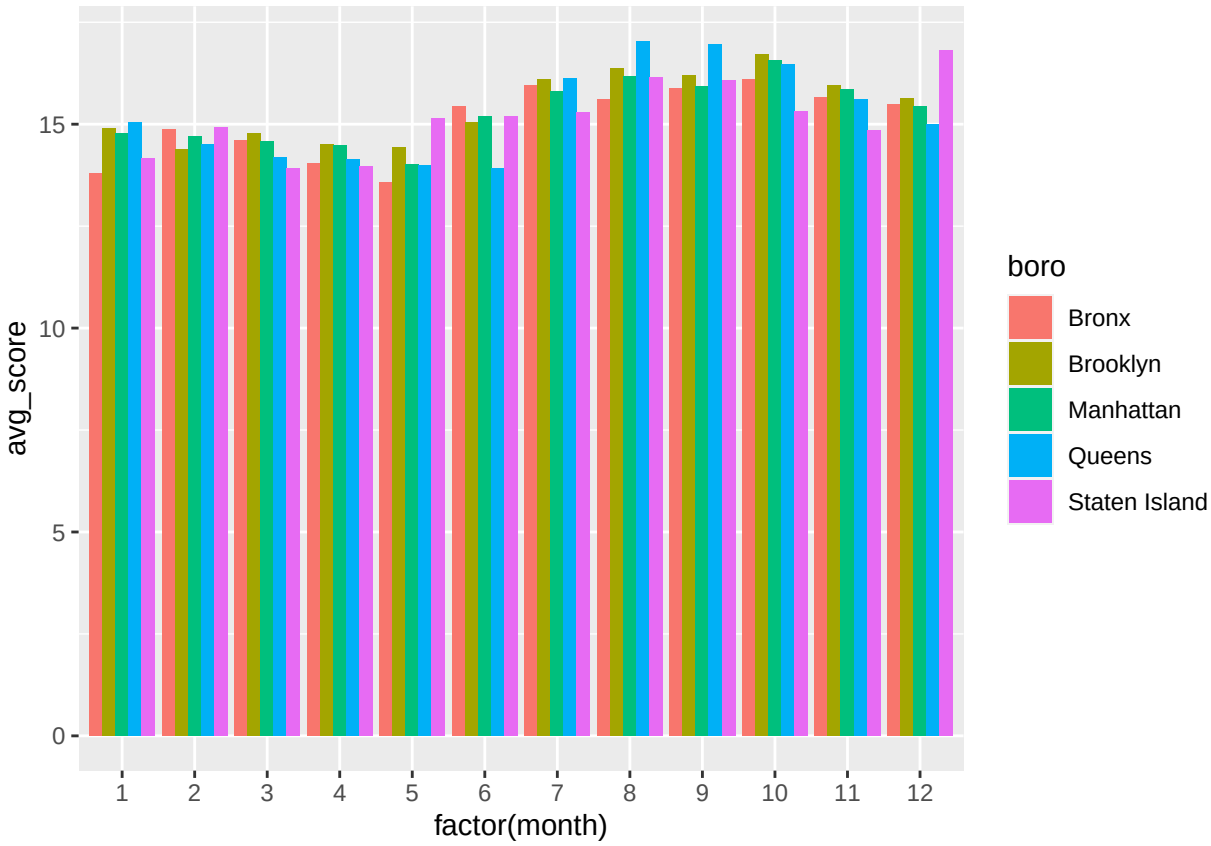
Indian restaurants have the highest closure ratio in Brooklyn and queens, while Tex-max has the highest closure ratio in Bronx and Staten Island. However, these trends do not hold in the the other boroughs.

2.6 Trend of score by month

- What is your conclusion?

```
# Find the trend of average scores by month and borough. Graph your result
inspections %>%
  group_by(boro, month) %>%
  summarize(avg_score = mean(score)) %>%
  ungroup() %>%
  ggplot(data = ., aes(x = factor(month), y = avg_score, fill = boro)) +
  geom_col(position = 'dodge')
```

```
## `summarise()` has grouped output by 'boro'. You can override using the
## `.groups` argument.
```



Scores seem to be slightly higher July - Nov.

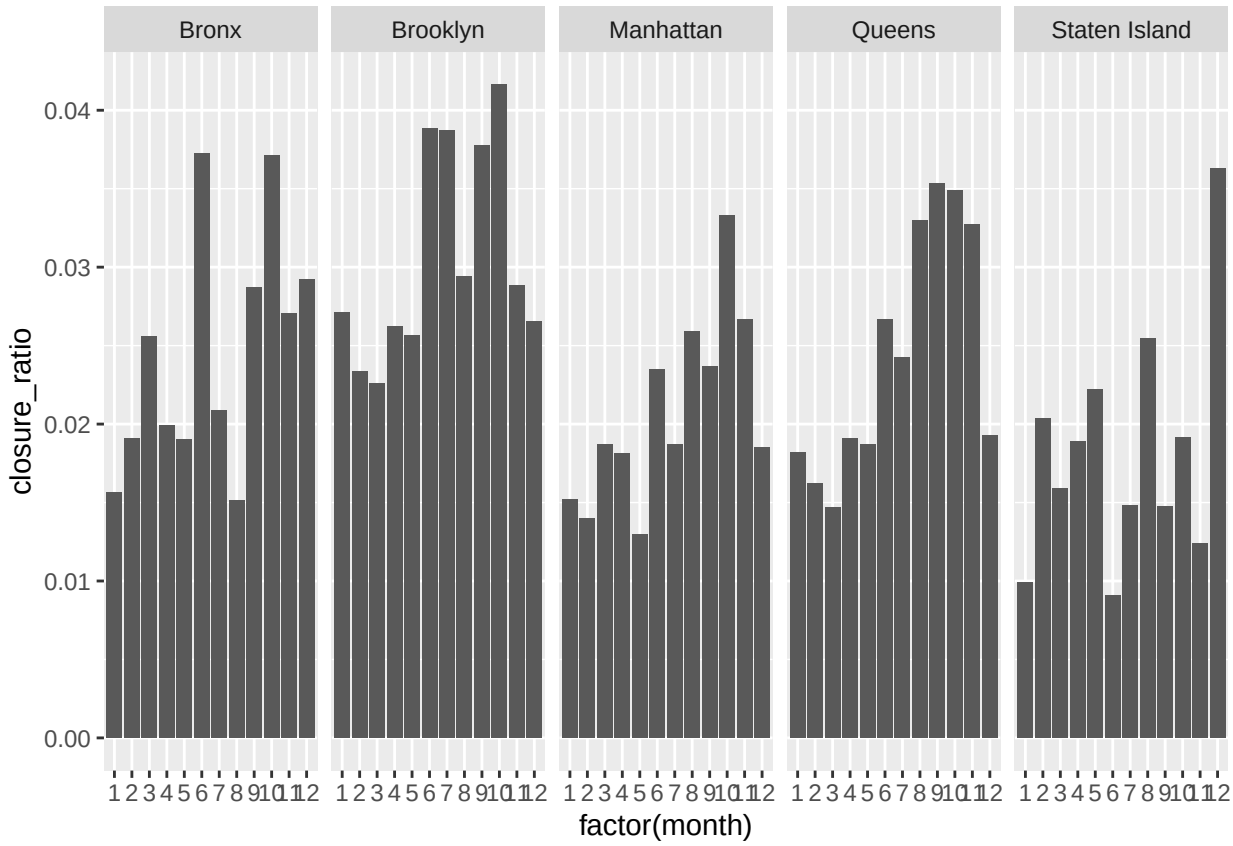
2.7 Trend of inspection closure ratio by month

- What is your conclusion?

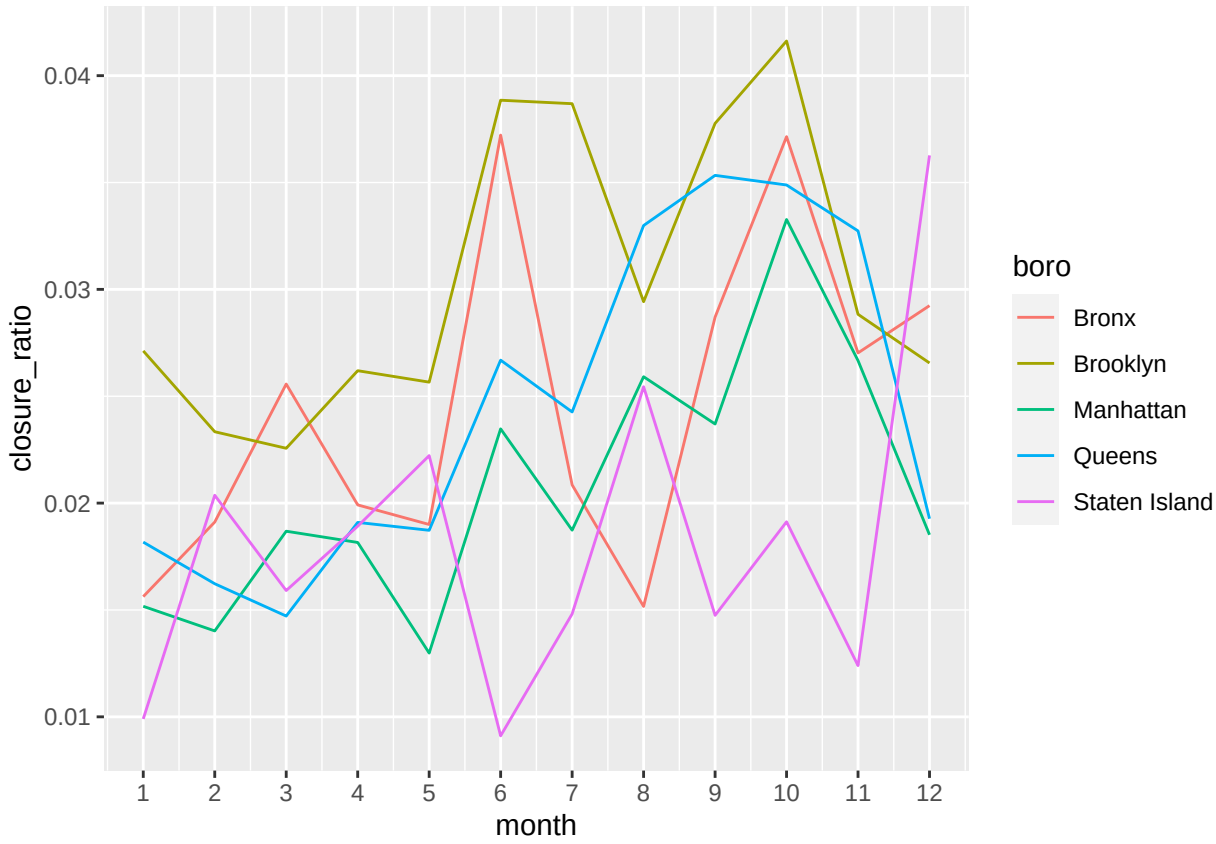
```
# Find the trend of the inspection closure ratio by month and borough. Graph your result.
cl_mon_boro = inspections %>%
  group_by(boro, month) %>%
  summarize(closure_ratio = sum(action == 'closed'| action == 'reclosed')/n())
```

```
## `summarise()` has grouped output by 'boro'. You can override using the
## `.groups` argument.
```

```
ggplot(data = cl_mon_boro, aes(x = factor(month), y = closure_ratio)) +
  geom_col() +
  facet_grid(. ~ boro)
```

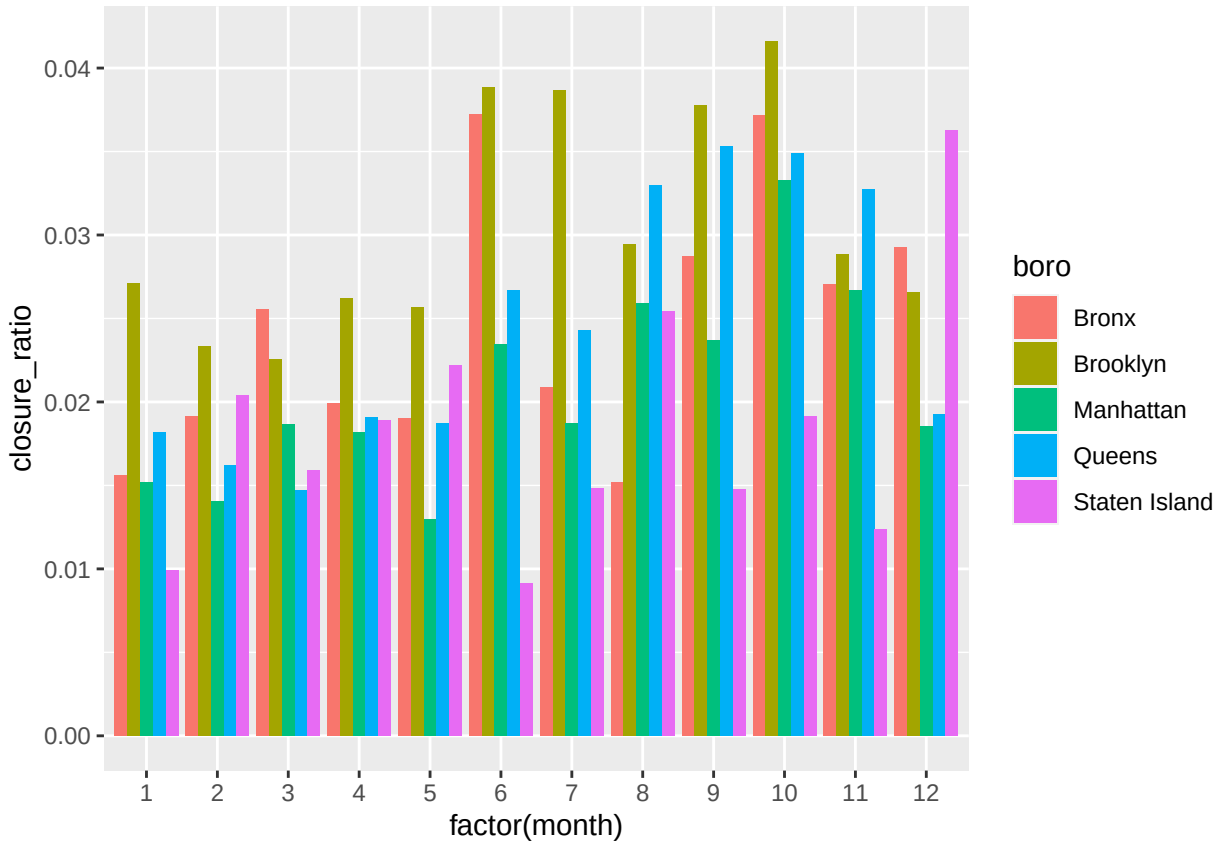



```
ggplot(data = cl_mon_boro, aes(x = month, y = closure_ratio, color = boro)) +
  scale_x_continuous(breaks = 1:12) +
  geom_line()
```

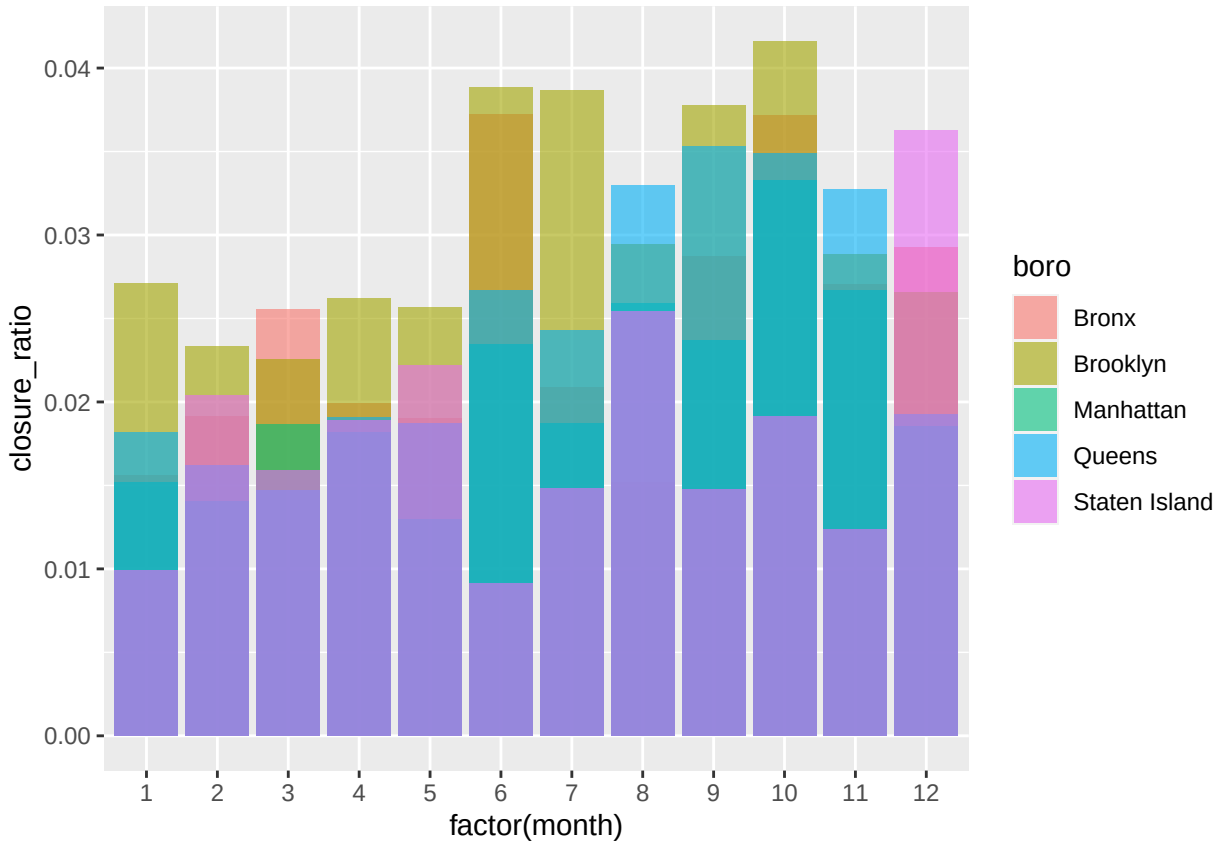



```
g = ggplot(data = cl_mon_boro, aes(x = factor(month), y = closure_ratio, fill = boro))

g + geom_col(position = 'dodge')
```




```
g + geom_col(alpha = 0.6, position = 'nudge')
```



The closure ratio is highest in October for the Bronx and Manhattan. It is also relatively high for Brooklyn and Queens at this time. June/July also has higher ratios of closures in Brooklyn and Staten Island.