[WEEK3] Exercise on Graph Design

Hyunsoo Luke Kim

2019-02-02

```
## function ()
## .Internal(getwd())
## <bytecode: 0x55b772151c28>
## <environment: namespace:base>
```

Question 1.

- How does the number of inspections change over time (use month as the level of temporal granularity)?
- Does the number of inspections increase or decreate over time?
- Are there any peak times? Is there any seasonal effect (like inspections being more common during cretain seasons or months)?

Load the data

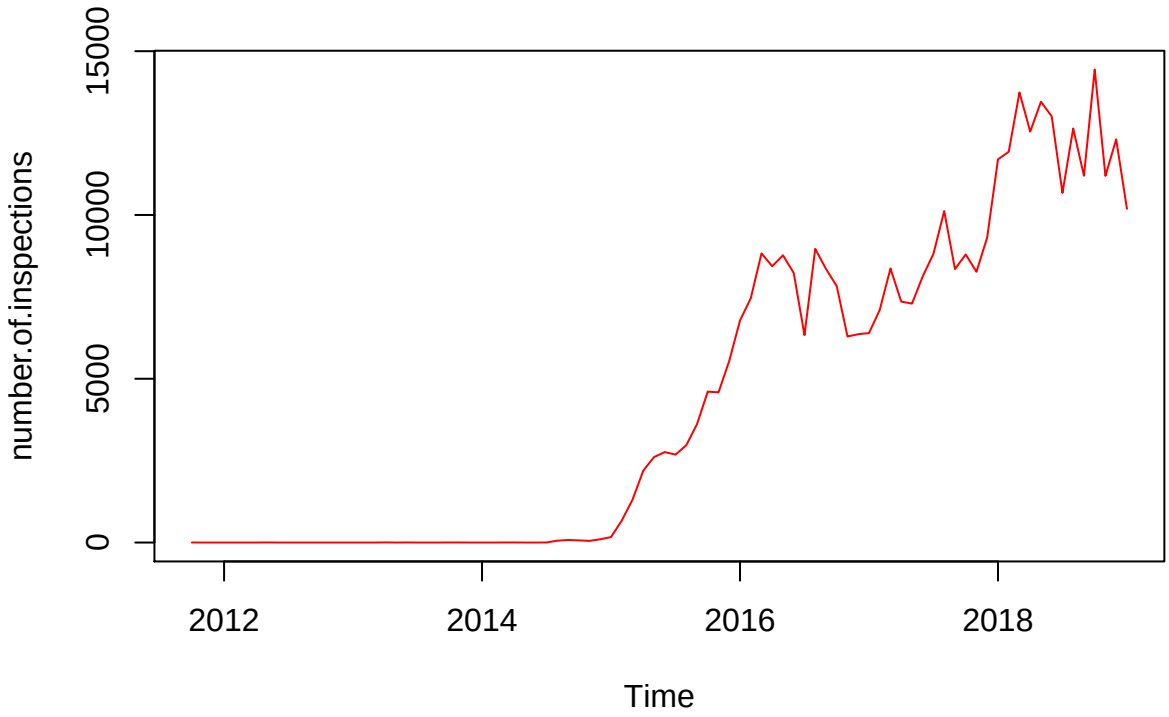
```
## 'data.frame':    383996 obs. of  1 variable:
##  $ date: Factor w/ 1391 levels "01/02/2015","01/02/2016",..: 984 962 3 1249 1228 408 747 1385 194 11
```

Summarize the data by month

```
##   month YEAR max_Frequency
## 1    10 2011             1
## 2    11 2011             0
## 3    12 2011             0
## 4     1 2012             0
## 5     2 2012             0
## 6     3 2012             0
```

```
##    month YEAR max_Frequency
## 83     8 2018         12641
## 84     9 2018         11196
## 85    10 2018         14436
## 86    11 2018         11193
## 87    12 2018         12308
## 88     1 2019         10188
```

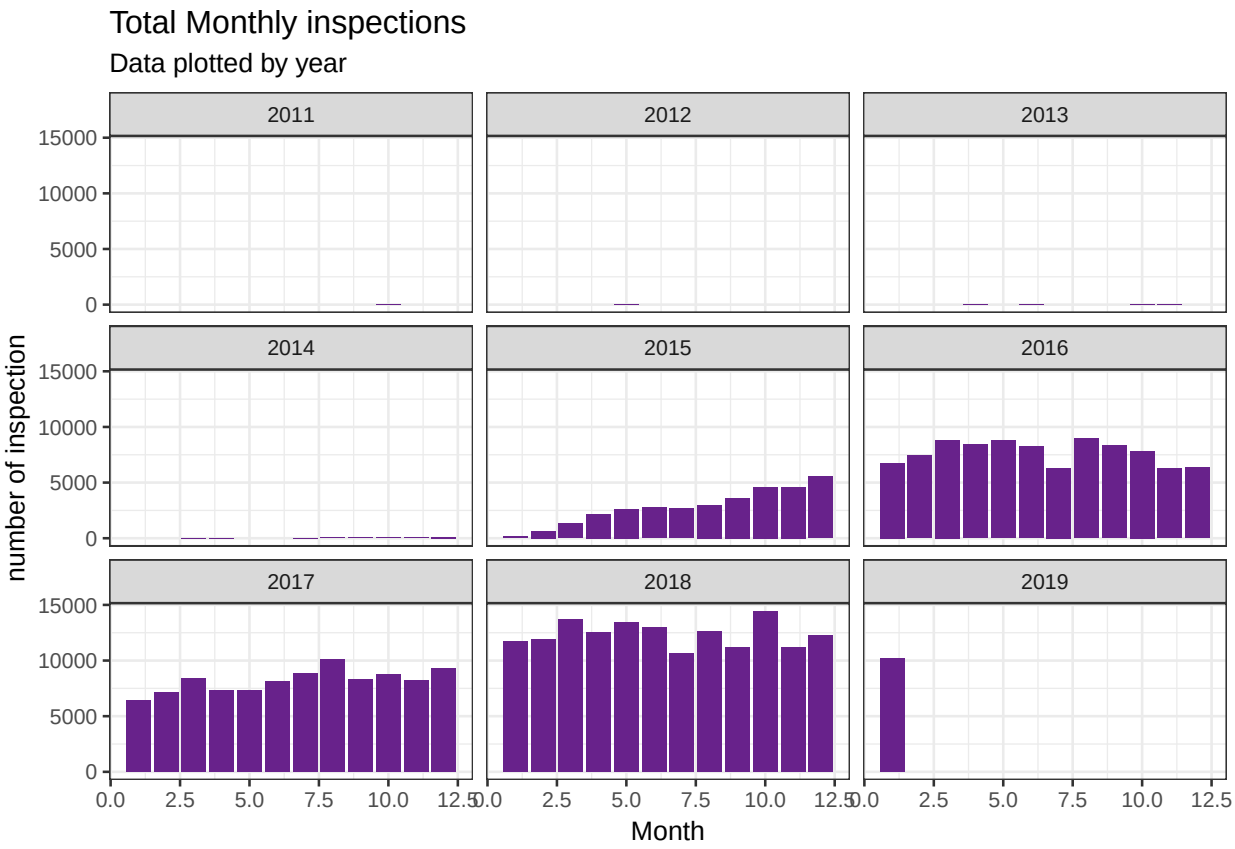
[Graph1] Visualization the data by month



Analysis the graph 1

- I used **Line chart** because it visualize how a quantity changes in relation to time.
- As you can see, the number of inspections has been increasing since July 2014.
- Before July 2014, there were just few inspections.
- The peak time was **October 2018** and number of inspections were **14436**.
- To see seasonal effect, I added an additional attribute which is month.

[Graph2] Visualization data using two adttributes(year, month) by faceting



Analysis the graph 2

- I used the **faceting technique** to splited data into a number of plots. So, I could see monthly data with year.
- Also, I employed **bar chart** because it is useful in case reading every single value accurately.

- Unfortunately, I could not see seasonal effect exactly through this graph.
- Neverthelass, I wanted to check if there were some seasonal patterns in 2016, 2017, 2018.

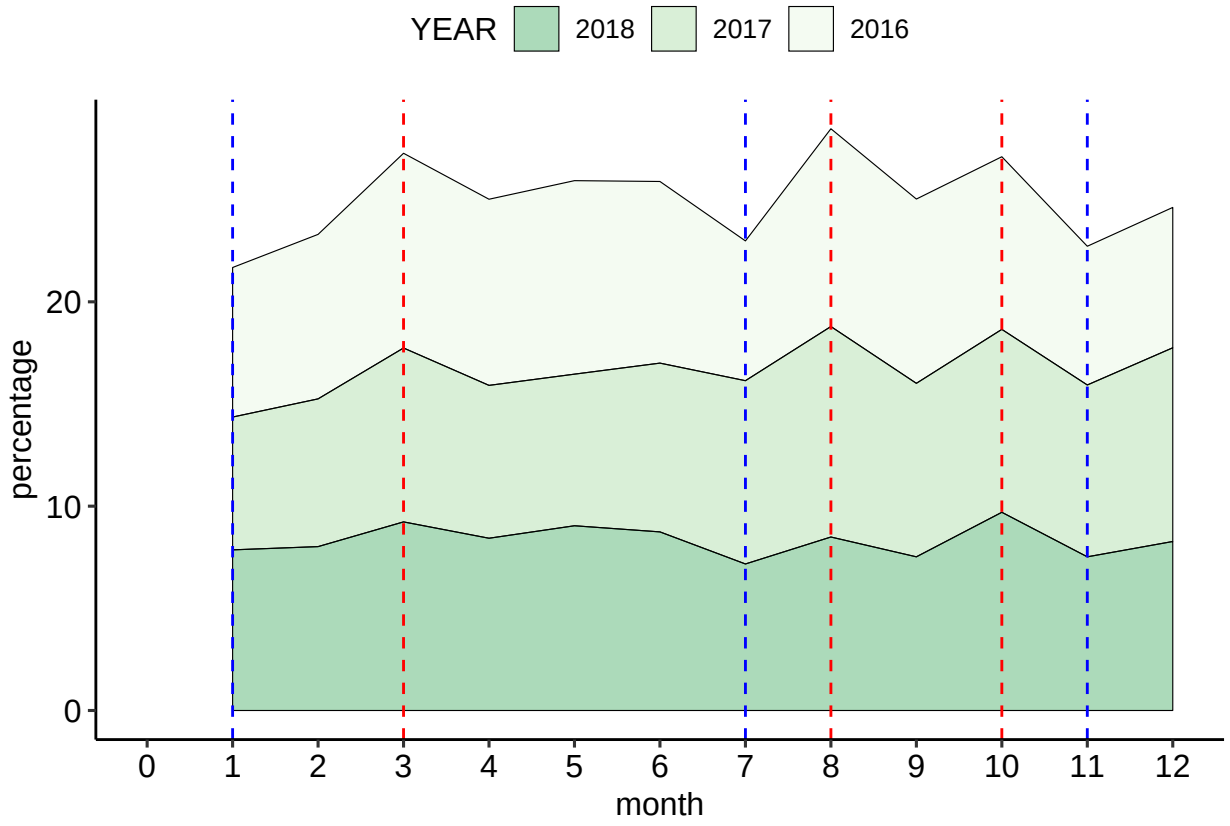
- To find monthly effects, I used **stacked technique** and compared data by representing as the **percentage**.

Representing data as the percentage in 2016, 2017, 2018

```
##   month YEAR max_Frequency percentage
## 1     1 2016          6777       7.32
## 2     2 2016          7456       8.05
## 3     3 2016          8826       9.53
## 4     4 2016          8435       9.11
## 5     5 2016          8768       9.47
## 6     6 2016          8229       8.89
```

```
##    month YEAR max_Frequency percentage
## 31     7 2018         10673       7.17
## 32     8 2018         12641       8.49
## 33     9 2018         11196       7.52
## 34    10 2018         14436       9.70
## 35    11 2018         11193       7.52
## 36    12 2018         12308       8.27
```

[Graph3] Stacked area graph



Analysis the graph 3

- I used **stacked area chart** to find which month has a highest or lowest proportion values.
- I can see easily how are the proportion of values different each month through this stacked area chart.
- As you can see, there are **red dash lines which show rising patterns** in this graph. The proportion of inspections increased **every March, August, October** in 2016, 2017, 2018.
- Also, the **blue dash lines which display decreasing patterns** indicate that the proportion of inspections downsized **every January, July, November**.

Question 2.

- Is there any difference in how the number of inspections changes over time in the 5 different boroughs of New York City?

Load the data

```
## 'data.frame':    383869 obs. of  2 variables:
##  $ BORO           : Factor w/ 5 levels "BRONX","BROOKLYN",..: 2 2 3 3 4 1 5 4 2 2 ...
##  $ INSPECTION.DATE: Factor w/ 1391 levels "01/02/2015","01/02/2016",..: 984 962 3 1249 1228 408 747 1
```

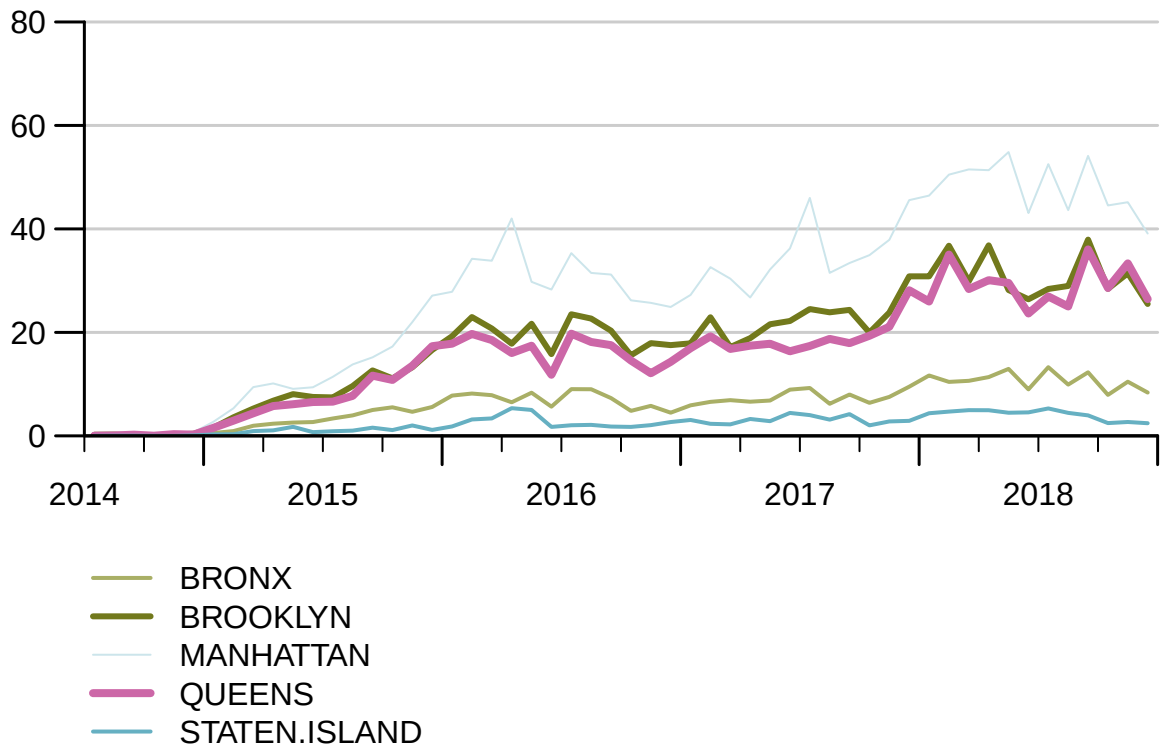
Summarize the data by month

- I selected the data which is from Aug 2014 because there were few inspections before July 2014.

```
##   month YEAR          BORO max_Frequency
## 1     8 2014      BROOKLYN            21
## 2     8 2014     MANHATTAN            32
## 3     8 2014        QUEENS             3
## 4     8 2014         BRONX             0
## 5     8 2014 STATEN ISLAND             0
## 6     9 2014         BRONX             5
```

```
##     month YEAR          BORO max_Frequency
## 265    12 2018 STATEN ISLAND           268
## 266     1 2019         BRONX           837
## 267     1 2019      BROOKLYN          2549
## 268     1 2019     MANHATTAN          3911
## 269     1 2019        QUEENS          2643
## 270     1 2019 STATEN ISLAND           244
```

[Graph4] Visualization the data by month



Analysis the graph 4

- I used **separate lines** because it is easy to compare each borough values.
- I got a 5 differnt lines by visualiszation and I **divided the number of inspectors by 100**. The y-axis represents the number of observations.*
- As you can see, the number of inspections in **MANHATTAN is always higher** then others' one throughout every year.
- Next, the number of inspections of **BROOKLYN and Queens are very similar** over the years.
- Finally, the number of inspections of **BRONX and STATEN ISLAND have very slightly increased** compared to others' one. Especially, **STATEN ISLAND has the lowest values of them** throughout the years.

Question 3.

- How are cuisines types distributed across the New York area? Are there geographical areas where certain cuisines tend to concentrate (that is are there any areas where certain cuisines are more prevalent than others)? NOTE: focus only on the top 5 most frequent "Cuisine Description" categories.

Load the data and data selection

- I selected the data which is from Aug 2014 because there were few inspections before July 2014.

```
## 'data.frame':    383857 obs. of  3 variables:
##  $ BORO               : Factor w/ 6 levels "BRONX","BROOKLYN",..: 2 2 3 3 5 1 6 5 2 2 ...
##  $ DATE               : Date, format: "2018-09-19" "2017-09-14" ...
##  $ CUISINE.DESCRIPTION: Factor w/ 85 levels "Afghan","African",..: 14 3 3 47 8 20 47 17 81 17 ...
```

Find Top 5 cousines by frequency

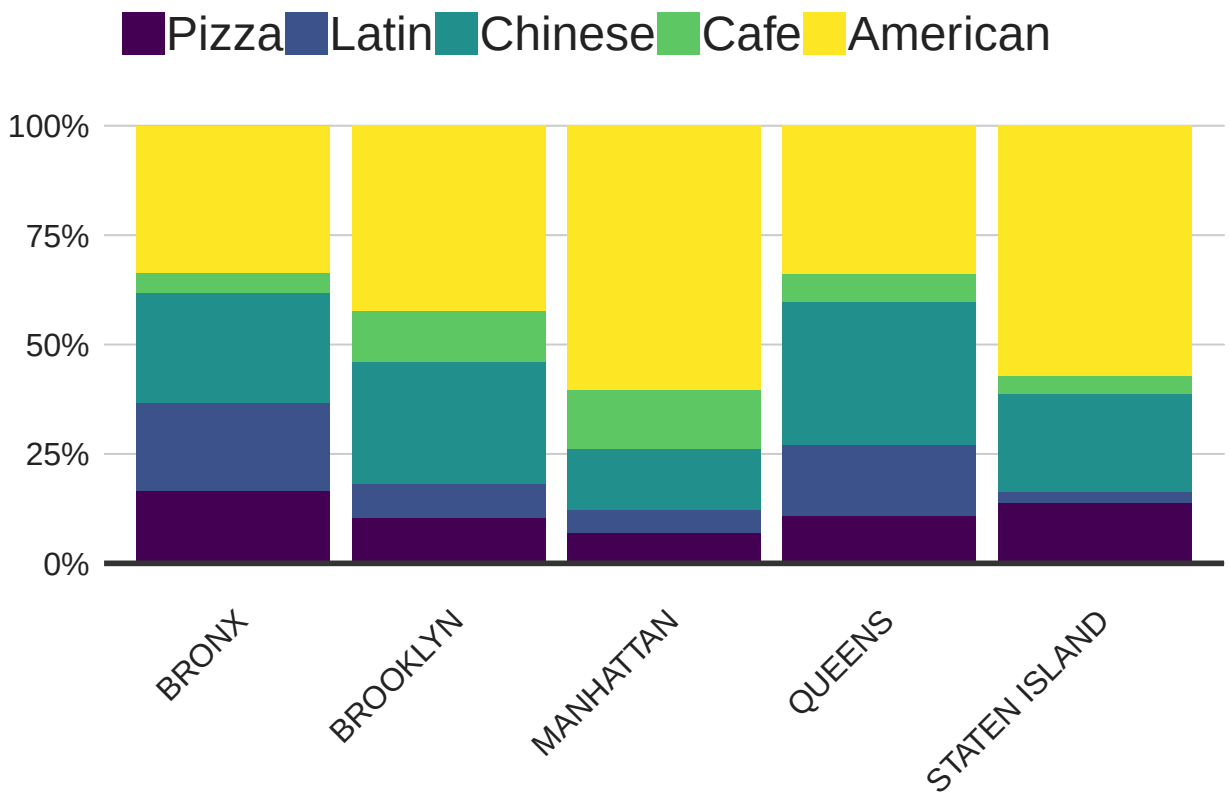
```
## # A tibble: 10 x 2
##    CUISINE.DESCRIPTION                                                Frequency
##    <fct>                                                                  <int>
##  1 American                                                               83358
##  2 Chinese                                                                40013
##  3 "CafA\xa8\xcf/Coffee/Tea"                                              18064
##  4 Pizza                                                                  17224
##  5 Latin (Cuban, Dominican, Puerto Rican, South & Central Americ~         16749
##  6 Italian                                                                16068
##  7 Mexican                                                                15488
##  8 Japanese                                                               13532
##  9 Caribbean                                                              13455
## 10 Bakery                                                                 11702
```

- As you can see, Top5 most frequent cuisines are American, Chinese, Cafe, Pizza, Latin
- So, I used the data which only include these cuisines to compare the distribution of them each borough.

Reconstruct the data with top5 cuisines

```
##             BORO               DATE        CUISINE.DESCRIPTION
##  BRONX        :16642   2018-10-31:   422   American:83358
##  BROOKLYN     :43426   2018-01-30:   411   Cafe    :18064
##  MANHATTAN    :71689   2018-10-23:   401   Chinese :40013
##  QUEENS       :38187   2019-01-10:   399   Latin   :16749
##  STATEN ISLAND: 5464   2018-06-05:   397   Pizza   :17224
##                        2018-03-28:   396
##                        (Other)   :172982
```

[Graph5] Visualization the data with the stacked bar graph



Analysis the graph 5

- Before visualizaion, I reexpressed the values in terms of percentages. This is because percentages make comparison between values easier. And, I used **staked bar graph.**
- There are ratios of five cuisines in five bar which indicates each borough.
- From this, I could see that every borough has **the highest ratio of American cuisine, especailly at Manhattan and Staten Island.**
- Next, **Chinese cuisine has the second high ratios** all borough and particularly **Queens** has very similar ratios between American cuisine and Chinese cuisine.

Question 4.

- How does the average score compare across different cuisine types? Are there cuisines that tend to have consistently lower/higher average scores compared to the others? NOTE: focus only on the top 5 most frequent "Cuisine Description" categories.

Load the data and data selection

- I selected the data which is from Aug 2014 because there were few inspections before July 2014.

- And I only focused on top 5 most frequnet cuisine types which are American, Chinese, Cafe, Latin, Pizza.

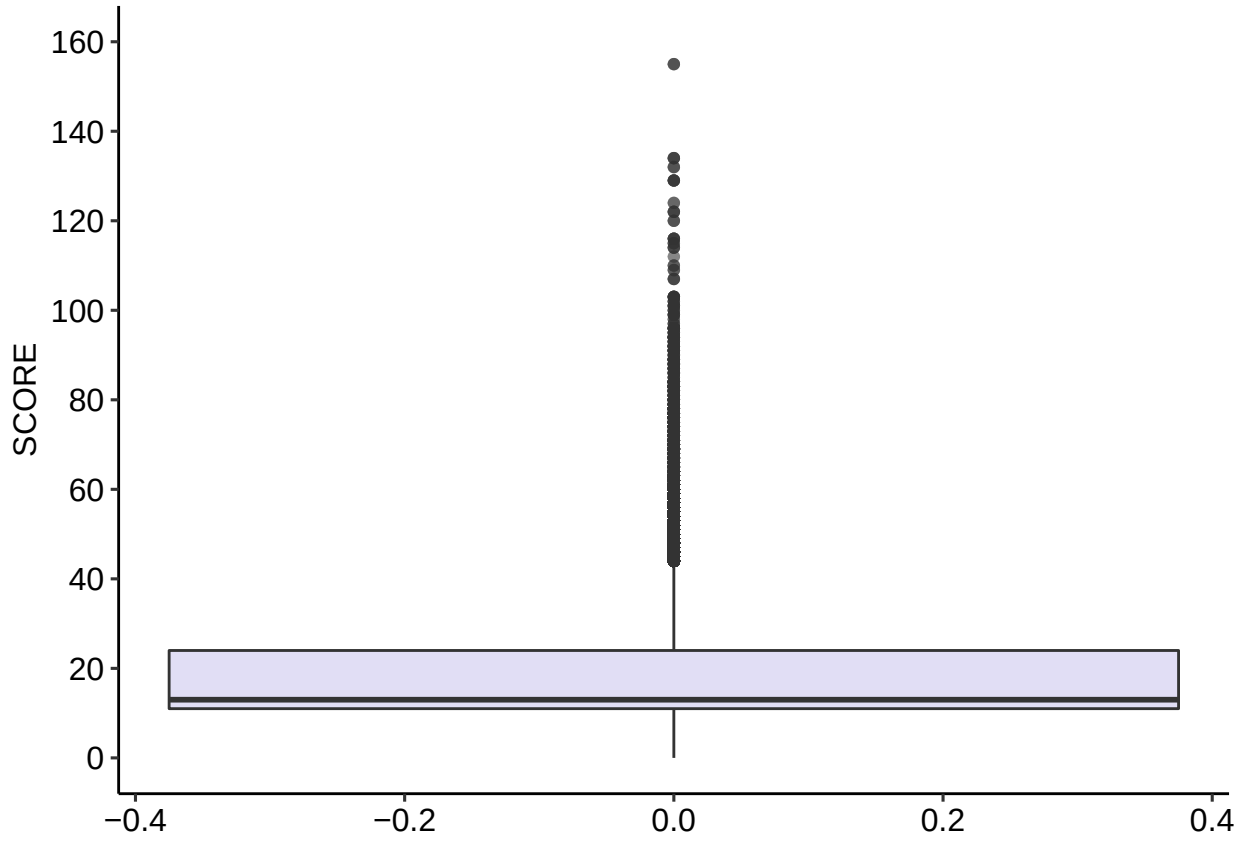
```
##      SCORE                DATE        CUISINE.DESCRIPTION
##  Min.   : -1.00   2018-10-31:   422   American:83427
##  1st Qu.: 11.00   2018-01-30:   411   Cafe    :18064
##  Median : 13.00   2018-10-23:   401   Chinese :40013
##  Mean   : 19.12   2019-01-10:   399   Latin   :16749
##  3rd Qu.: 24.00   2018-06-05:   397   Pizza   :17224
##  Max.   :155.00   2018-03-28:   396
##  NA's   :8169     (Other)   :173051
```

Delete raws which contain missing values(SCORE)

```
## 'data.frame':    167308 obs. of  3 variables:
##  $ SCORE              : int  12 61 70 12 6 13 28 19 8 12 ...
##  $ DATE               : Factor w/ 1328 levels "2014-08-01","2014-08-07",..: 1221 902 992 1087 383 97
##  $ CUISINE.DESCRIPTION: Factor w/ 5 levels "American","Cafe",..: 2 1 1 3 4 3 4 2 1 1 ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:8169] 5 15 20 50 77 114 120 125 205 295 ...
##   .. ..- attr(*, "names")= chr [1:8169] "5" "15" "20" "50" ...
```

Distribution of all scores with the box plot

- The min value is '-1' and the max value is '155'. But, interestingly, the most scores are near '20'.



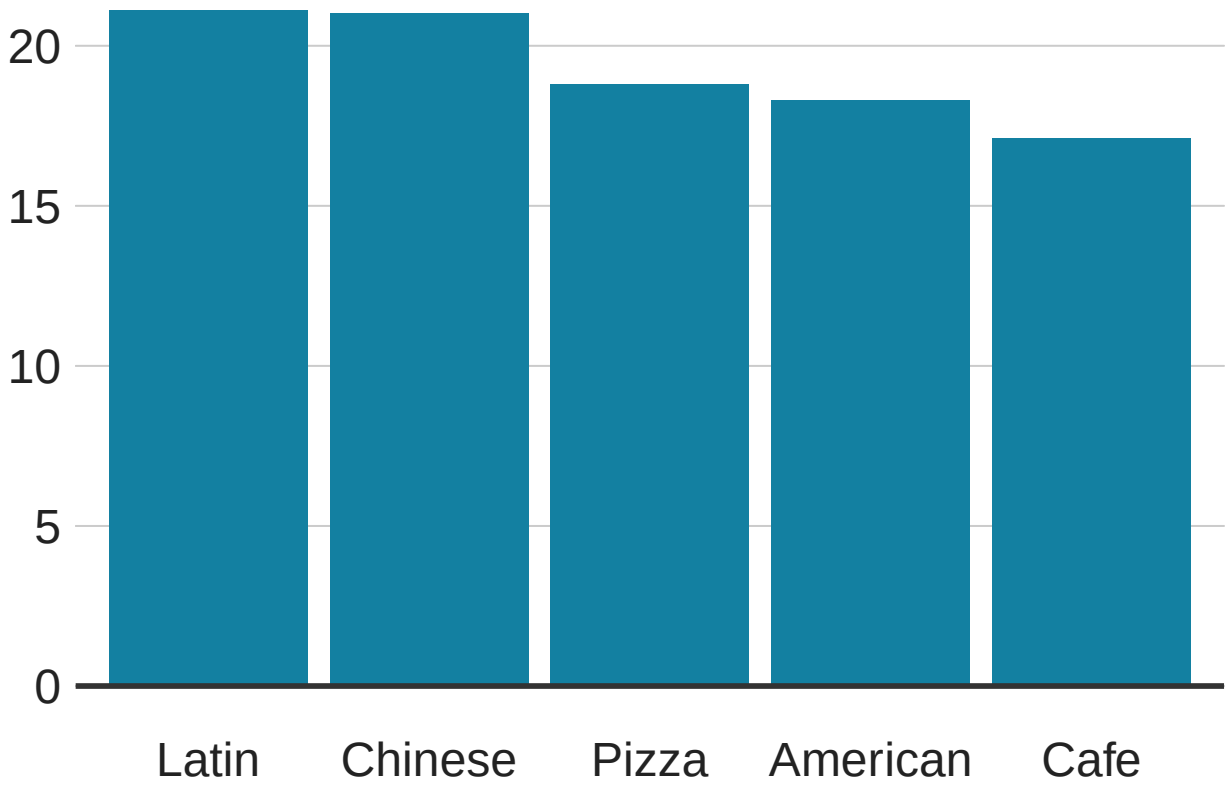
The average score compare across different cuisine types.

- For solving question 4, I aggregated all the categories and calculated the average across all scores that belong to each of the cuisines.

```
## # A tibble: 5 x 2
##   CUISINE.DESCRIPTION Mean.Score
##   <fct>                    <dbl>
## 1 Latin                     21.1
## 2 Chinese                   21.0
## 3 Pizza                     18.8
## 4 American                  18.3
## 5 Cafe                      17.1
```

```
## [1] 19.26207
```

[Graph6] Visualization the data with the bar graph



Analysis the graph 6

- I used the bar graph to see how the average amount of scores distributes across cuisine types.
- As you can see, there is no significant difference with the average score across cuisine types. - **The average score is 19.26** and the average score of each cuisine is almost close to it though **The Latin cuisine has the highest average score(21.09)** .

Question 5: Is there a relationship between cuisine type and violation? For instance, do some cuisine types tend to have more of some type of violations that other cuisine types?

Load the data

- I selected the data which is from Aug 2014 because there were few inspections before July 2014.

```
## 'data.frame':    379249 obs. of  3 variables:
##  $ VIOLATION.CODE     : Factor w/ 99 levels "02A","02B","02C",..: 61 30 36 60 61 29 55 42 68 57 ...
##  $ DATE               : Factor w/ 1381 levels "2014-08-01","2014-08-05",..: 1273 946 1037 1006 683 1
##  $ CUISINE.DESCRIPTION: Factor w/ 85 levels "Afghan","African",..: 14 3 3 47 8 20 47 17 81 17 ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:4731] 75 137 142 580 643 648 781 931 950 1008 ...
##   .. ..- attr(*, "names")= chr [1:4731] "75" "137" "142" "580" ...
```

Data selection

- I wanted to use the **Matrix chart** to solve this question because it represents how a quantity distributes across two categories.
- But, a problem was that each attribute have too many levels(VIOLATION.CODE = 100, CUISINE.DESCRIPTION = 85). So, it was hard to visualize them at once.
- Thus, I selected only the top 5 most frequent levels of cuisine types and the top 15 most frequent levels of violation code types.

```
## # A tibble: 15 x 2
##    VIOLATION.CODE Frequency
##    <fct>              <int>
##  1 10F                53962
##  2 08A                41531
##  3 04L                28113
##  4 06D                25992
##  5 06C                25337
##  6 10B                22623
##  7 02G                22589
##  8 04N                20840
##  9 02B                19787
## 10 04H                 8205
## 11 04M                 8112
## 12 06E                 7666
## 13 06F                 7212
## 14 04A                 7097
## 15 06A                 6467
```

```
##  VIOLATION.CODE          DATE        CUISINE.DESCRIPTION
##  10F    :25510   2018-10-31:   346   American:66453
##  08A    :18651   2018-10-23:   331   Cafe    :13941
##  06D    :12710   2018-02-01:   326   Chinese :32494
##  04L    :12358   2018-06-05:   321   Latin   :13252
##  06C    :11610   2018-01-30:   310   Pizza   :13848
##  02G    :10565   2018-05-09:   308
##  (Other):48584   (Other)   :138046
```

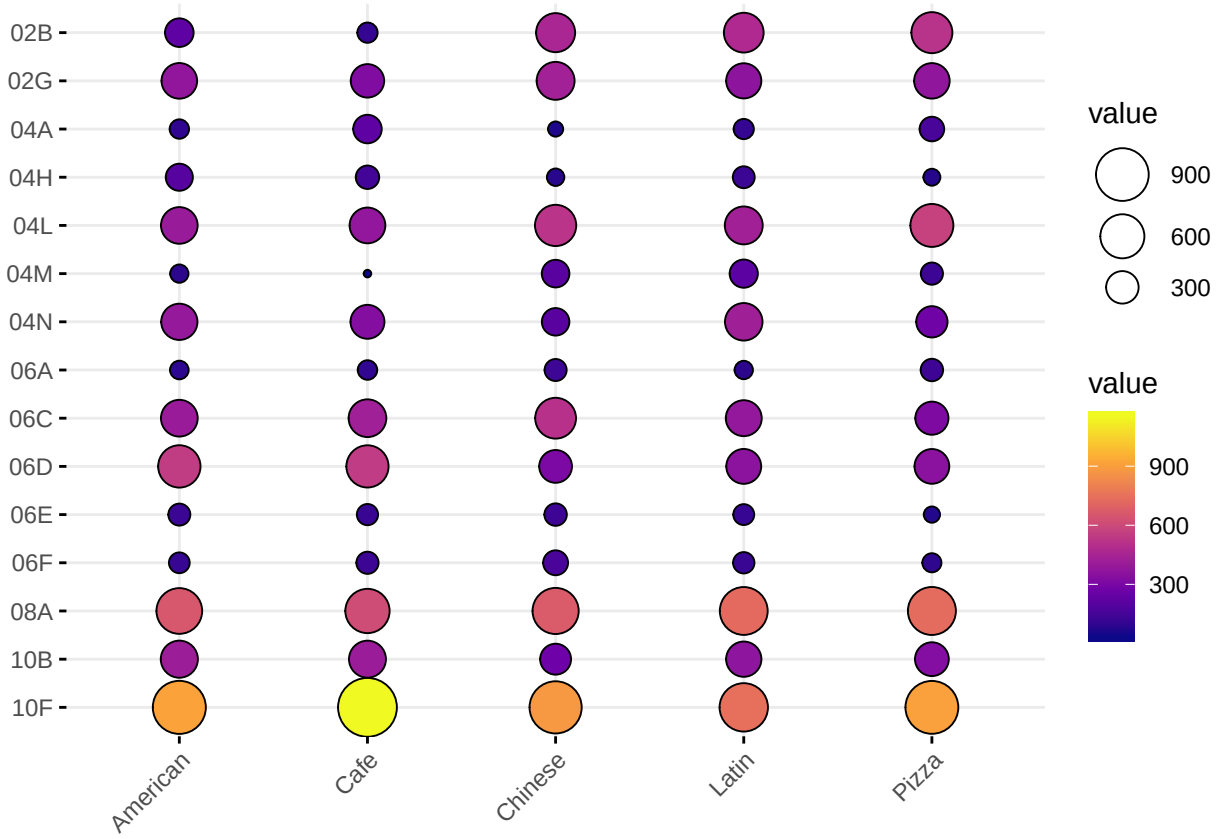
Transform the data as matrix and Normalize the data

- To use the Matrix chart, I fisrt converted the data form to the matrix.
- This martix data has 15 rows and 5 culomns. Each value means the frequency of violation in terms of each cuisine type.

- But, the normalization is needed for comparison of each value with different cuisine types.
- So, I used this formaiton: normalization = 5000 * value / The aggregation of violation in terms of each cuisine type.

```
##     American     Cafe  Chinese    Latin    Pizza
## 02B 224.1434 105.4444 456.6997 475.7772 510.9041
```

[Graph7] Visualization the data with the matrix graph



Analysis the graph 7

- In this matrix chart, every value has its own balloon. Its size indicates the amount of violation.
- As you can see, **every balloon size is larger than others at 10F**, especially Cafe of cuisine types is the biggest one.
- **The next is 08A**, every balloon size is large.
- This visualization shows us **some of the violations which are '10F' and '08A' happened more than others regardless of cuisine types**.