Exercise 1

Yahya Jabary (11912007)

About

Dataset: https://www.kaggle.com/datasets/new-york-city/nyc-inspections

Tableau Public: https://public.tableau.com/app/discover

Comments must argue for:

- **Correctness**: argue why you are correct, answered the question and visually conveyed the information.

- **Effectivity**: argue why you answered question well, why this is the best solution you could come up with and what else you've tried → you can use as many slides as you want to show your failed attempts.

Question 1

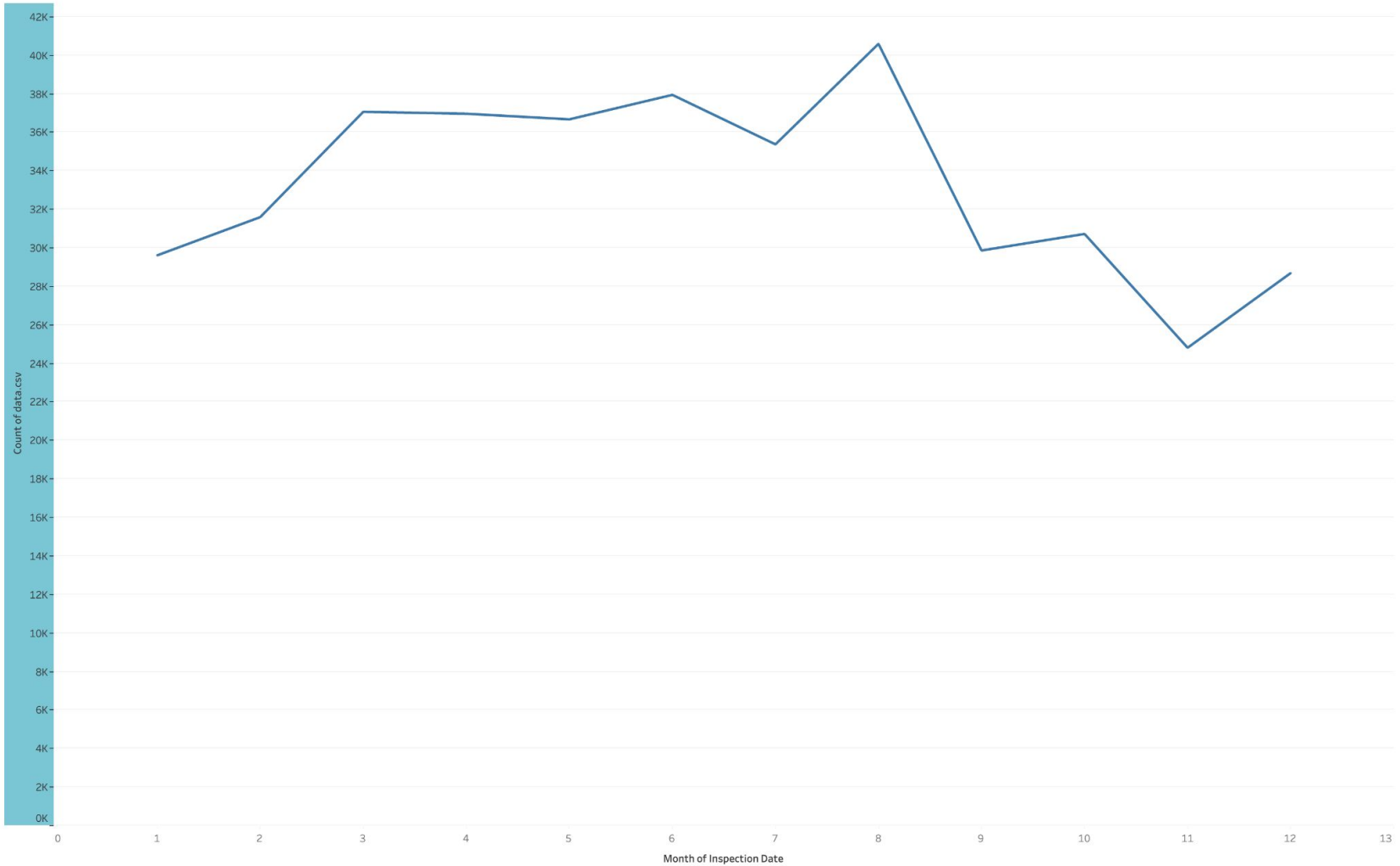
How does the number of inspections change over time (use month as the level of temporal granularity)?

Does the number of inspections increase or decrease over time?

Are there any peak times?

Is there any seasonal effect (like inspections being more common during certain seasons or months)?

Sheet 1


42K
40K
38K
36K
34K
32K
30K
28K
26K
24K
22K
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
Count of data.csv
0
1
2
3
4
5
6
7
8
9
10
11
12
13
Month of Inspection Date

Question 1: Comments

I believe that this is the most straightforward way to communicate the answer to this question. Lines that connect discrete function values give a sense of continuity. By looking at the differential we can tell whether there is an increase/decrease in inspections.

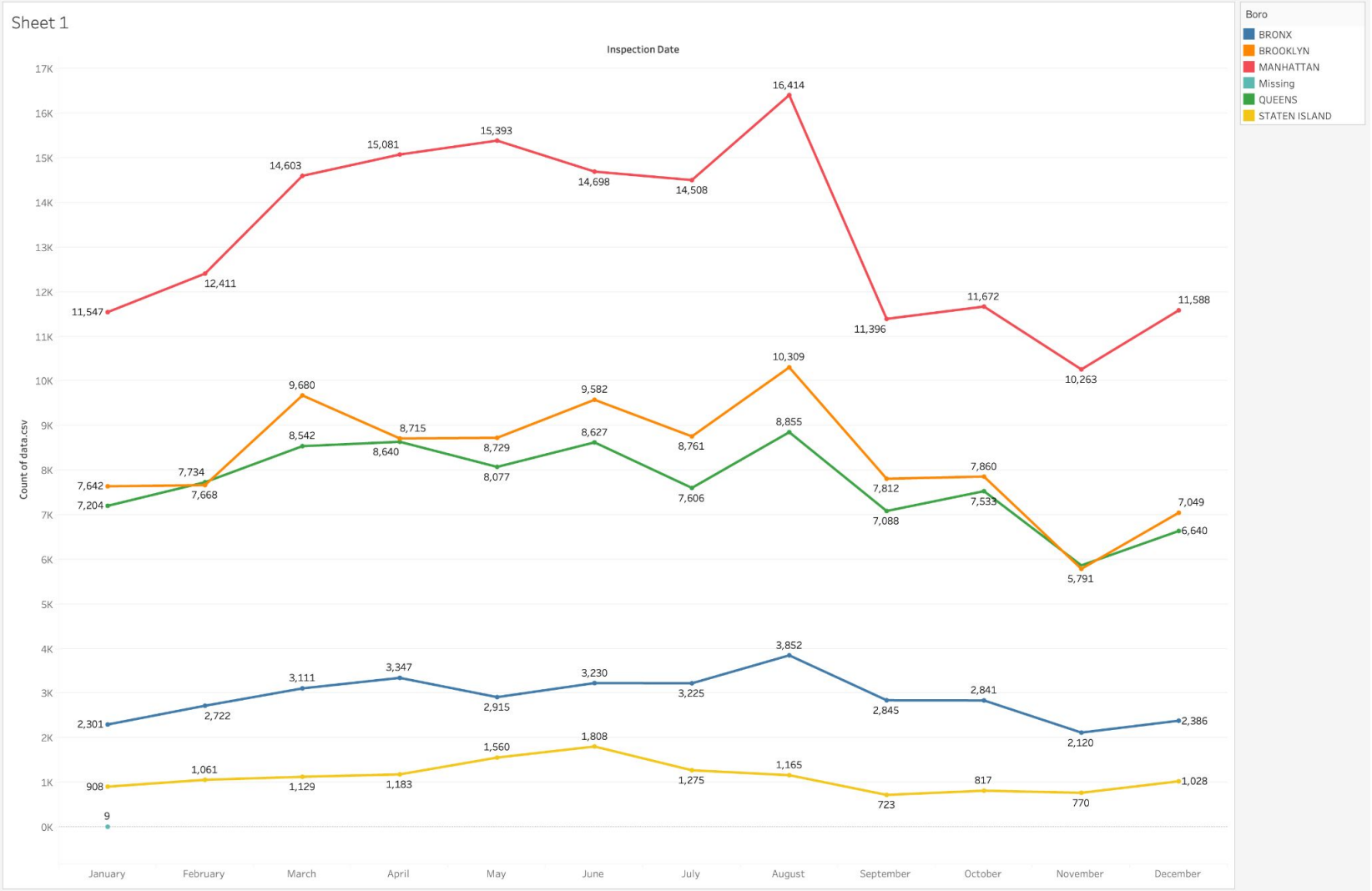
We could also add more metrics (like numeric values, delta, …) but I believe that they'd be redundant for this example.

Additional info we can read:

- The inspections have increased overall: `f(x_last) > f(x_first) = 28691 > 29611`
- Peak times: local/global maxima: "August" seems to be particularly popular.

Question 2

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


Sheet 1
Inspection Date
Count of data.csv
Boro
BRONX
BROOKLYN
MANHATTAN
Missing
QUEENS
STATEN ISLAND
17K
16K
15K
14K
13K
12K
11K
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
January
February
March
April
May
June
July
August
September
October
November
December
11,547
12,411
14,603
15,081
15,393
14,698
14,508
16,414
11,396
11,672
10,263
11,588
7,642
7,734
9,680
8,715
8,729
9,582
8,761
10,309
7,812
7,860
5,791
7,049
7,204
7,668
8,542
8,640
8,077
8,627
7,606
8,855
7,088
7,533
6,640
2,301
2,722
3,111
3,347
2,915
3,230
3,225
3,852
2,845
2,841
2,120
2,386
908
1,061
1,129
1,183
1,560
1,808
1,275
1,165
723
817
770
1,028
9

Question 2: Comments

I’d argue the same way as in the previous example.

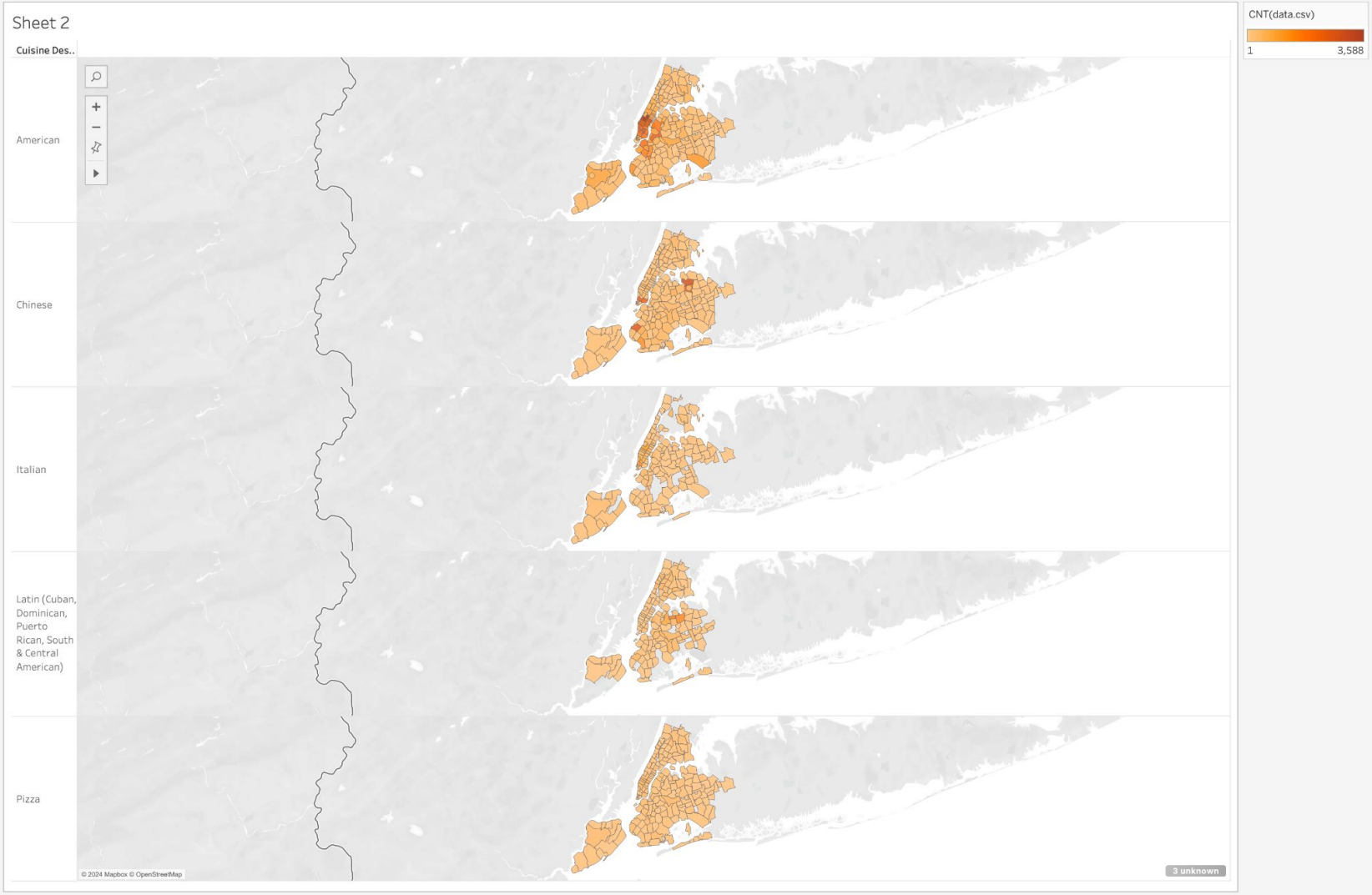
This time I additionally added mark labels and a legend to have some complementary quantitative data as a point of reference.

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.


Sheet 2
Cuisine Des..
American
Chinese
Italian
Latin (Cuban, Dominican, Puerto Rican, South & Central American)
Pizza
CNT(data.csv)
1
3,588
© 2024 Mapbox © OpenStreetMap
3 unknown

Question 3: Comments

I believe using a map is the most effective way to communicate geospatial data.

I used a geographical heatmap for each of the top 5 cuisines.

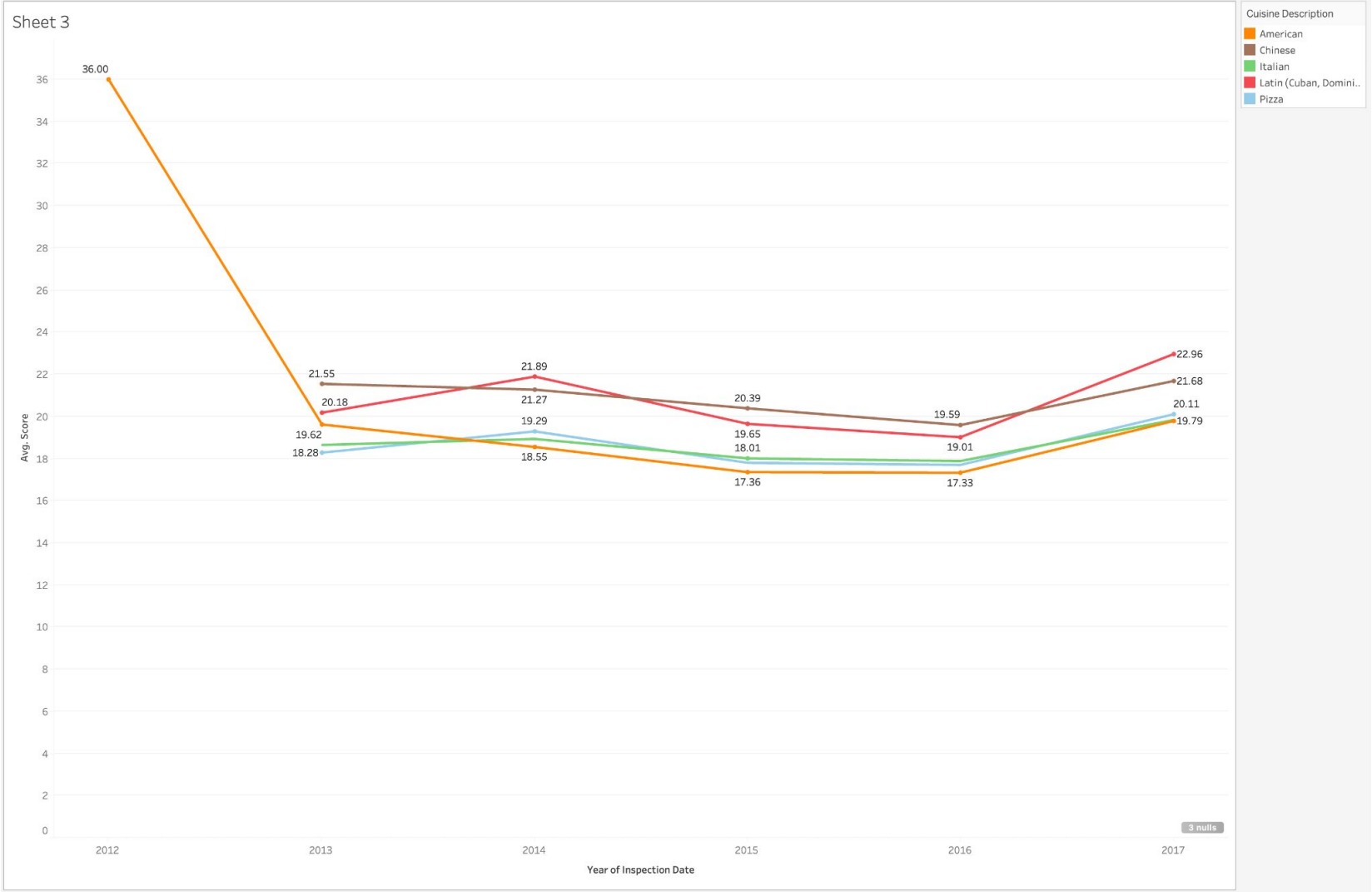
Furthermore, I customized the color mapping to communicate the feeling of “heat” and intensity better.

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.

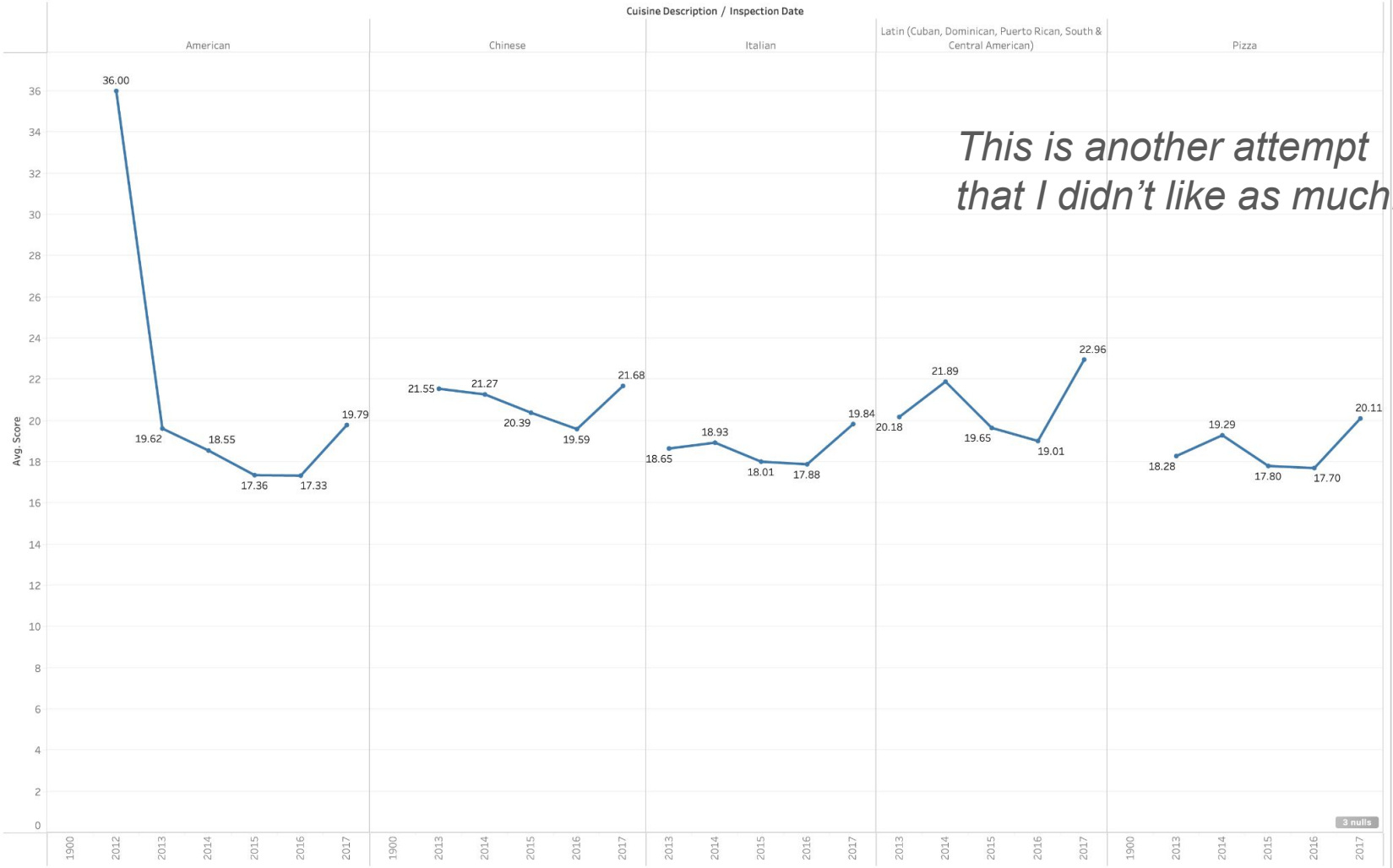

Sheet 3
Cuisine Description
American
Chinese
Italian
Latin (Cuban, Domini..
Pizza
Avg. Score
Year of Inspection Date
2012
2013
2014
2015
2016
2017
36.00
21.55
20.18
19.62
18.28
21.89
21.27
19.29
18.55
20.39
19.65
18.01
17.36
19.59
19.01
17.33
22.96
21.68
20.11
19.79
3 nulls

Question 4: Comments

Not quite sure - But I understood “Are there cuisines that tend to have consistently lower/higher average scores compared to the others?” as: “Get the average score of the top 5 most popular cuisines and then see how they changed **throughout time** - find out if they stayed consistently higher / lower than the others”.

I believe this graph communicates the change of average popularity of multiple cuisines throughout time very well.

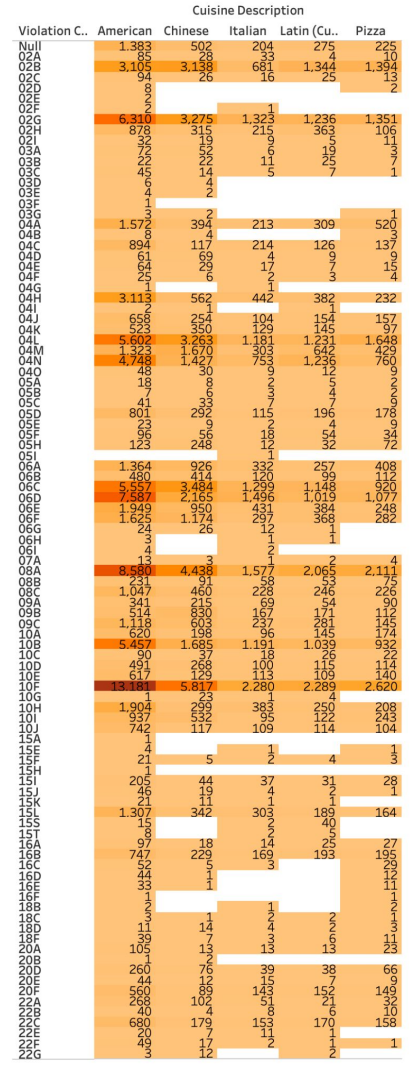
Sheet 3


Cuisine Description / Inspection Date
American
Chinese
Italian
Latin (Cuban, Dominican, Puerto Rican, South & Central American)
Pizza
Avg. Score
36.00
19.62
18.55
17.36
17.33
19.79
21.55
21.27
20.39
19.59
21.68
18.65
18.93
18.01
17.88
19.84
20.18
21.89
19.65
19.01
22.96
18.28
19.29
17.80
17.70
20.11
3 nulls
This is another attempt that I didn't like as much.

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?


Cuisine Description
Violation C..
American
Chinese
Italian
Latin (Cu..
Pizza
Null 1,383 502 204 275 225
02A 85 28 33 4 10
02B 3,105 3,138 681 1,344 1,394
02C 94 26 16 25 13
02D 8 2
02E 2
02F 2 1
02G 6,310 3,275 1,323 1,236 1,351
02H 878 315 215 363 106
02I 32 19 9 5 11
03A 72 52 6 19 3
03B 22 22 11 25 7
03C 45 14 5 7 1
03D 6 4
03E 4 2
03F 1
03G 3 2 1
04A 1,572 394 213 309 520
04B 8 4 3
04C 894 117 214 126 137
04D 61 69 4 9 9
04E 64 29 17 7 15
04F 25 6 2 3 4
04G 1 1
04H 3,113 562 442 382 232
04I 2 1 1
04J 658 254 104 154 157
04K 523 350 129 145 97
04L 5,602 3,263 1,181 1,231 1,648
04M 1,323 1,670 303 642 429
04N 4,748 1,427 753 1,236 760
04O 48 30 9 12 9
05A 18 8 2 5 2
05B 7 6 3 4 2
05C 41 33 7 7 9
05D 801 292 115 196 178
05E 23 9 2 4 9
05F 96 56 18 54 34
05H 123 248 12 32 72
05I 1
06A 1,364 926 332 257 408
06B 480 414 120 99 112
06C 5,557 3,484 1,299 1,148 920
06D 7,587 2,165 1,496 1,019 1,077
06E 1,949 950 431 384 248
06F 1,625 1,174 297 368 282
06G 24 26 12 1
06H 3 1 1
06I 4 2
07A 13 3 1 2 4
08A 8,580 4,438 1,577 2,065 2,111
08B 231 91 58 53 75
08C 1,047 460 228 246 226
09A 341 215 69 54 90
09B 514 830 167 171 112
09C 1,118 603 237 281 145
10A 620 198 96 145 174
10B 5,457 1,685 1,191 1,039 932
10C 90 37 18 26 22
10D 491 268 100 115 114
10E 617 129 113 109 140
10F 13,181 5,817 2,280 2,289 2,620
10G 1 23 1 4
10H 1,904 299 383 250 208
10I 937 532 95 122 243
10J 742 117 109 114 104
15A 1
15E 4 1 1
15F 21 5 2 4 3
15H 1
15I 205 44 37 31 28
15J 46 19 4 2 1
15K 21 11 1 1
15L 1,307 342 303 189 164
15S 15 2 40
15T 8 2 5
16A 97 18 14 25 27
16B 747 229 169 193 195
16C 52 5 3 29
16D 44 1 12
16E 33 1 11
16F 1 1
18B 2 1 2
18C 3 1 2 2 1
18D 11 14 4 2 3
18F 39 7 3 6 11
20A 105 13 13 13 23
20B 1 2
20D 260 76 39 38 66
20E 44 12 15 7 9
20F 560 89 143 152 149
22A 268 102 51 21 32
22B 40 4 8 6 10
22C 680 179 153 170 158
22E 20 7 11 1
22F 49 17 2 1 1
22G 3 12 2

Question 5: Comments

Since we don’t want to aggregate but compare all possible combinations between the row and column item-types a grid-heatmap is the ideal visualization.

Furthermore it is the most effective way to immediately recognize “spots” / anomalies on a spectrum which helps in the detection of violations.