Restaurant Inspections in NYC

I. Description of the data

For this case study, I am using a New York Restaurants Inspection dataset downloaded off kaggle.com. It is originally provided by the Department of Health and Mental Hygiene (DOHMH). Every restaurant in New York City is scheduled for an unannounced inspection at least once a year. During the inspection, an inspector checks for compliance with city and state food safety regulations and marks points for any condition that violates these rules. The dataset contains 399.918 observations and 18 columns. It covers the years from 2011 until 2017.

CAMIS	This is a unique identifier for the entity (restaurant)
BDA	This field represents the name (doing business as) of the entity (restaurant)
BORO	Borough in which the entity (restaurant) is located
BUILDING	Building number for establishment (restaurant) location
STREET	Street name for establishment (restaurant) location
ZIPCODE	Zip code of establishment (restaurant) location
PHONE	Phone Number
CUISINE.DESCRIPTION	This field describes the entity (restaurant) cuisine
INSPECTION.DATE	This field represents the date of inspection
ACTION	This field represents the actions that is associated with each restaurant inspection.
VIOLATION.CODE	Violation code associated with an establishment (restaurant) inspection
VIOLATION.DESCRIPTION	Violation description associated with an establishment (restaurant) inspection
CRITICAL.FLAG	Indicator of critical violation
SCORE	Total score for a particular inspection
GRADE	Grade associated with the inspection

GRADE.DATE	The date when the current grade was issued to the entity (restaurant)
RECORD.DATE	The date when the extract was run to produce this data set
INSPECTION.TYPE	A combination of the inspection program and the type of inspection performed

Fig. 1 Description of the data

This data is enough to answer my questions, so no other external dataset was used in the study.

II. Cleaning the data

Before I started doing any work on the dataset, I read its documentation, inspected it and performed any necessary data transformations and data manipulations in order to get it into a usable format that wouldn't skew my insights. There were a lot of observations that had NA or NULL values, so I removed those. I transformed the Grade and Critical Flag variables into ordered factors. Since the Inspection Date was in the format of mm/dd/yyyy, in order to be able to use it, I divided it up into three columns: day, month and year.
On closer inspection, I noticed some values from the Grades column were not matching up with the respective scores they were derived from, so I fixed that as well.
The Grades are derived in the following way:

- restaurant scores between 0 and 13 received an "A" grade
- restaurant scores between 14 and 27 received a "B" grade
- restaurant scores higher than 28 received a "C" grade

Lastly, I removed any irrelevant variables that I was not going to use in the analysis.

This left me with the following reduced dataset of 376.588 observations and 13 variables:

Variable Name	Data Type	Dimension or Measure
CAMIS	Nominal	Dimension
DBA	Nominal	Dimension
BORO	Nominal	Dimension
ZIPCODE	Nominal	Dimension
CUISINE.DESCRIPTION	Nominal	Dimension
MONTH	Ordinal	Dimension
DAY	Quantitative	Dimension
YEAR	Quantitative	Dimension

ACTION	Nominal	Dimension
VIOLATION.CODE	Nominal	Measure
VIOLATION.DESCRIPTION	Nominal	Measure
SCORE	Quantitative	Measure
GRADE	Ordinal/Nominal	Measure

Fig. 2 Cleaned up dataset

III. Answering the questions

1. Which is the safest area to eat in NYC?

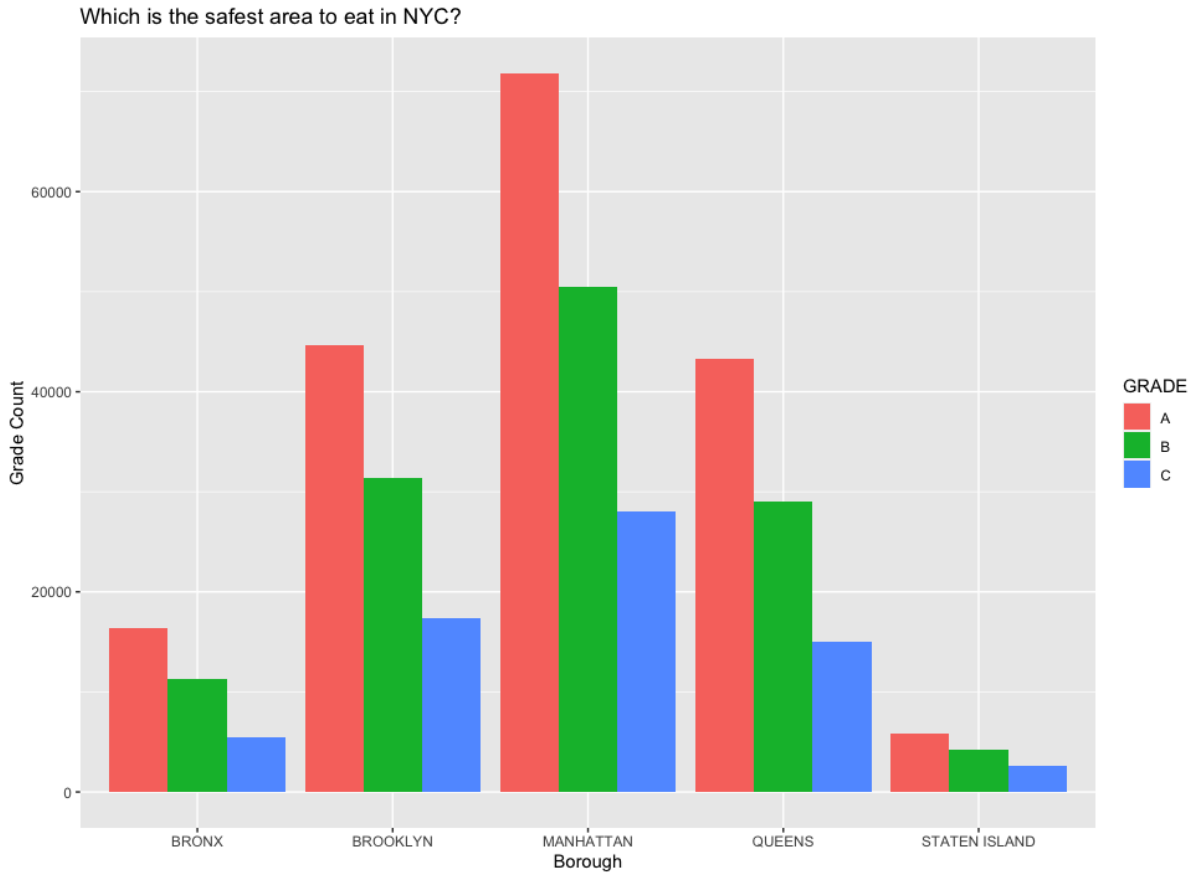

Fig.3 Safest area to eat in NYC

It seems like Manhattan might be the safest option to eat in NYC if you want to account for cleanness. Might this be because Manhattan just has the most number of restaurants?
In order to answer this question, I thought about which variables to use. I could have picked *Score* instead of *Grades*, but I also wanted to see the difference between all three grades. I had already done all the important data cleaning on the dataset, so the only thing that was left to do was to group the grades by borough and create a new variable to plot on the Y-axis:

Grade Count(Q). Since I am looking to pick the borough with the highest number of A's, it is important to use the visual encodings that allow me to perceive the differences between each category. Following Bertin's Semiology of Graphics regarding visual encoding, I deemed size/length the most suitable to visualize the variable *Grade* on the X-axis, which in this context becomes nominal. It makes it very easy to perceive differences. Regarding the differences between the number of A, B and C's, using colour hue as a third variable to encode them is very fitting in order to differentiate between them.

Effectiveness ranking

- estimating magnitude by using length is very accurate, following "Automating the design of graphical presentations of relational information", by Mackinlay. This makes the visualisation readily perceived, aligns with human perception and allows us to process the data faster and more accurately.
- using colour hue to differentiate between the grades, want them to be perceived as unordered because they're just categories.

Expressiveness ranking

- the visualisation doesn't communicate anything more than what I set it out to

Perceptual processing

Here, length is a pre-attentive feature that allows us to quickly draw answers from the plot. One Gestalt Grouping principle I followed was Proximity, which I used to group the grade categories for each borough.

Overall, this visualization is very efficient in answering the question at hand. One is able to figure out which area has the most number of A's almost instantaneously.

2. What restaurants to avoid?

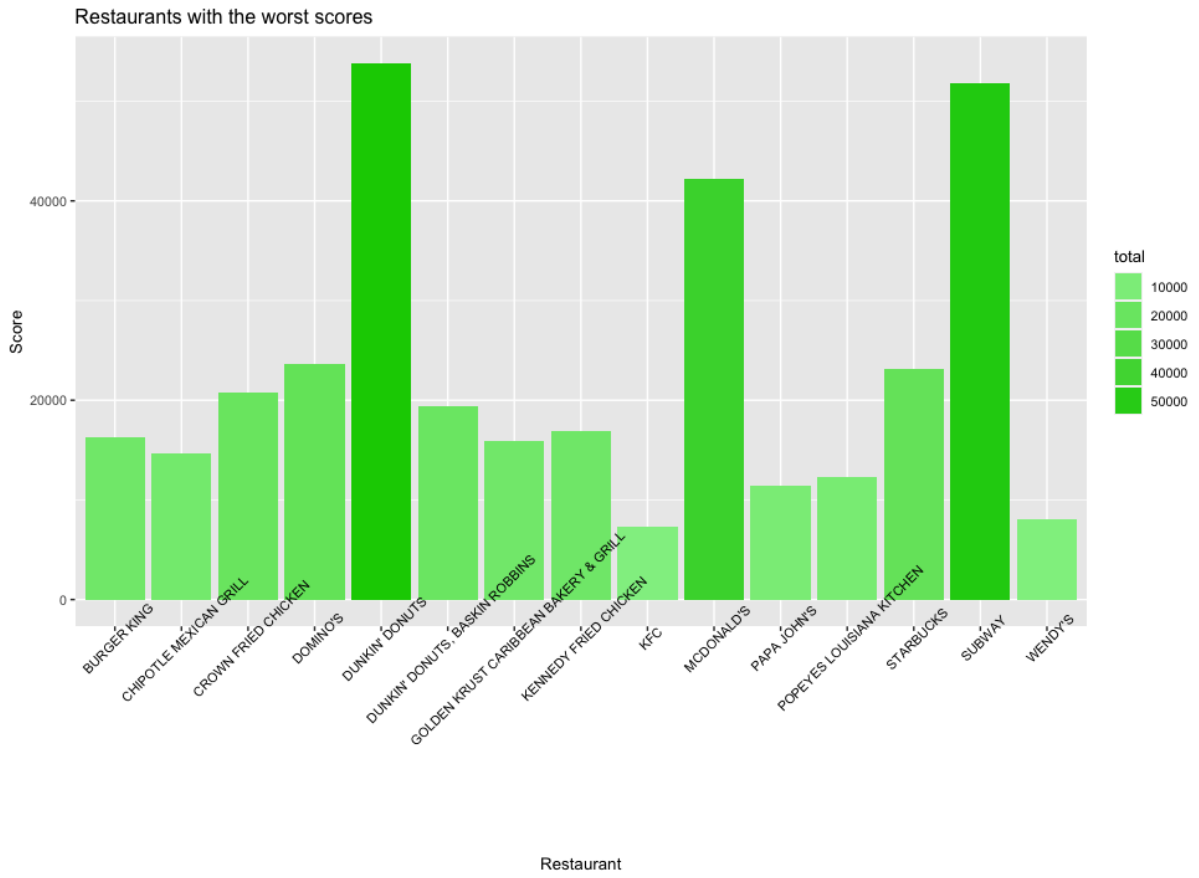

Fig. 4 Restaurants with the worst scores

I am starting off with a general question in order to get a better understanding of what New York City has to offer in terms of food quality. It seems like one would be better off avoiding all the big restaurants and fast-food chains.
For this visualization, I aggregated all the scores over the years by restaurant, ordered them in descending order and finally taking only the first 10 observations to visualise.

Plot elements

- **X - axis:** Nominal variable *Restaurant*
- **Y - axis:** Quantitative variable *Score*

Visual encoding

I'm using length to express the quantities for scores, which makes it easy to compare between them for each restaurant.

Effectiveness ranking

- it's easy to compare two data points thanks to using length as the encoding variable

3. What cuisine has always had the worst grades?

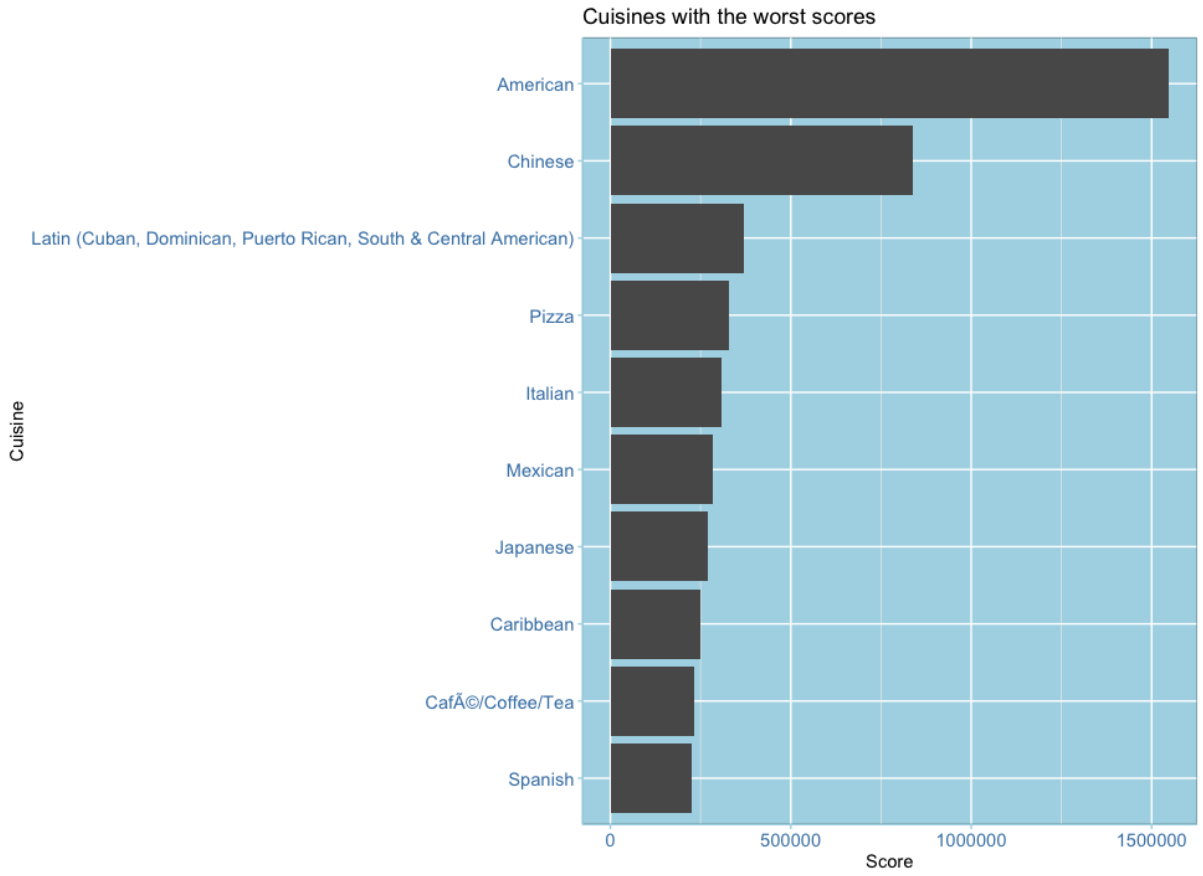

Fig. 5 Worst grades by cuisine

It is readily apparent that American cuisine is the worst in terms of food safety in NYC.
In order to make important comparisons easy to see, I arranged the data along the Y scale in a descending way starting from the highest score.

Plot elements

- **X - axis:** Quantitative variable *Score*
- **Y - axis:** Nominal variable *Cuisine*

Visual encoding

I am using length as an encoding variable, which is excellent for the question at hand.

Effectiveness ranking

- it is easy to compare two data points since the length is interpreted as a quantitative value
- the visualisation facilitates our brain to more accurately perceive the information presented

Expressiveness ranking

- the visualisation communicates the truth and doesn't allow for other connotations that are not inherent in the dataset.

4. When was the safest time to eat in NYC?

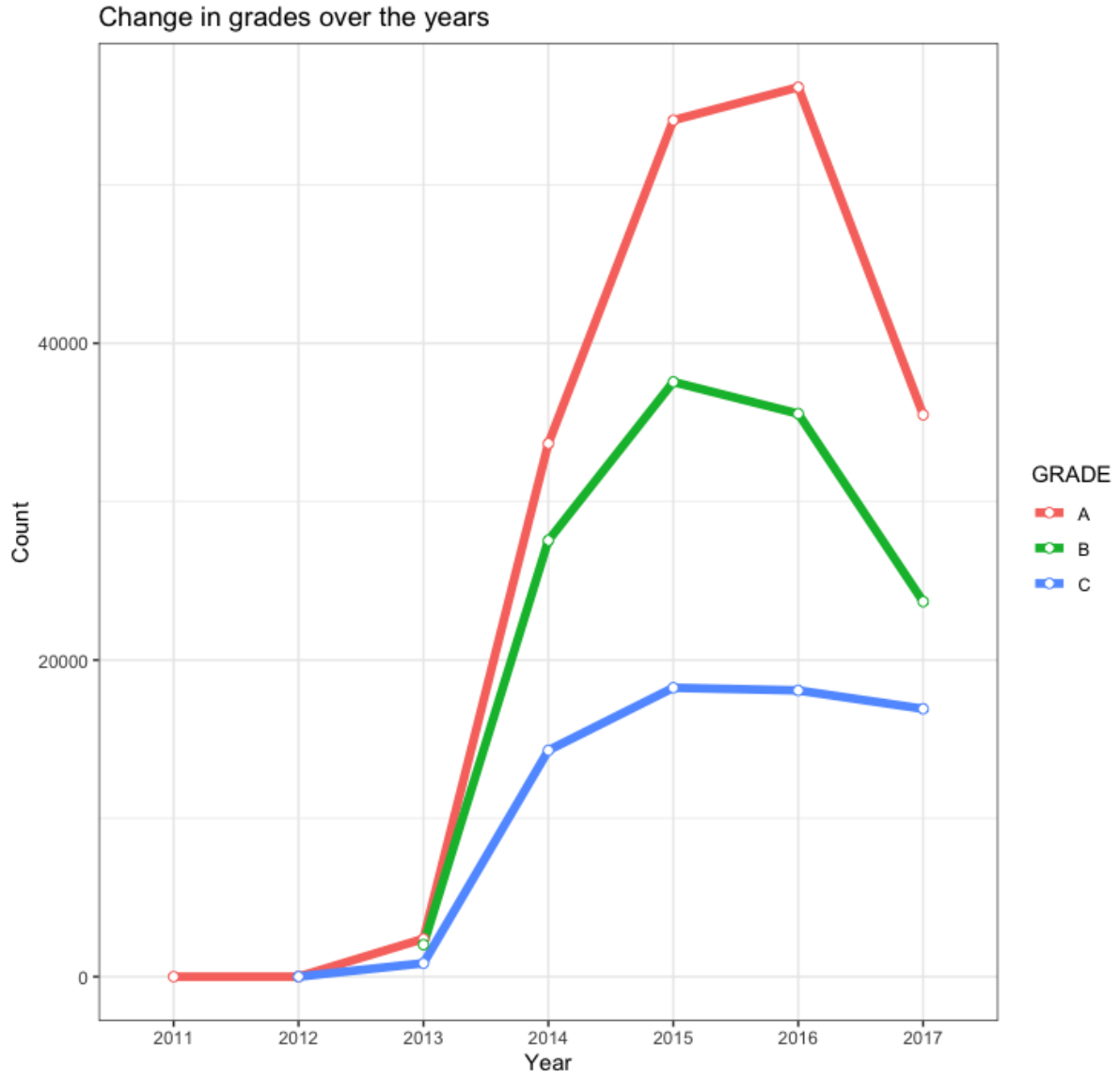

Fig. 6 Changes in grades over the years

It seems like NYC was at its peak in 2016 in terms of food safety. We can see that that year has the highest number of A grades.

The reason for the low grades from 2011 to 2013 was that only a few inspections were recorded during that period.

Visual encoding

- using a line graph to depict quantitative value change during a continuous period of time

Expressiveness ranking

- expresses only the information that is inherent in the dataset

Overall, it appears that there has been a positive impact of the inspections on food safety, however, it is unknown why there is a downgrade in all grade categories between 2016 and 2017.

IV. Follow up questions

5. Which borough has the most number of restaurants?

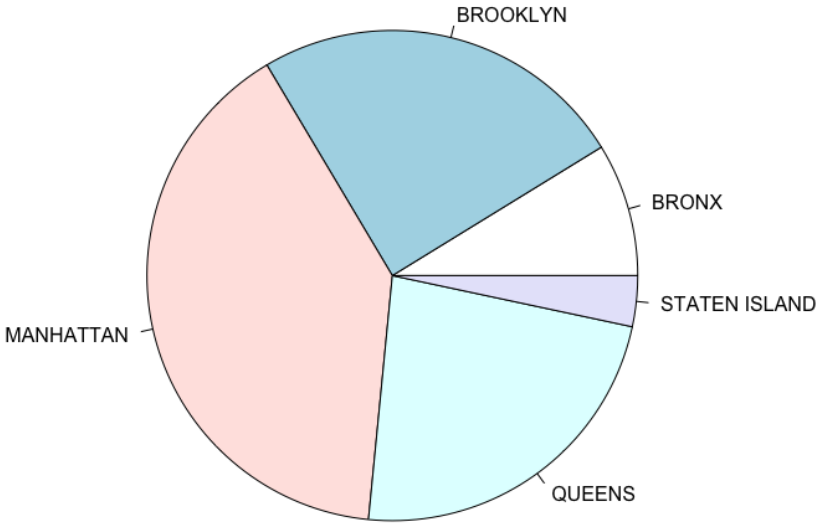

Fig. 7 Which borough has the most restaurants

This is a follow-up question from question 1. It seems I was right about Manhattan having the most restaurants out of all of the boroughs. The choice of a pie chart here is fitting because there are not a lot of categories to visualize and thus it is easy to see a part-of-whole comparison.

Visual encoding

- using size to express magnitude

Effectiveness ranking

- it is very quick and error-free to find the biggest part of the pie chart thanks to the small number of categories.

Expressiveness ranking

- the graph expresses all of the data necessary and nothing more.

6. Where should I go if I want to eat the best Chinese?

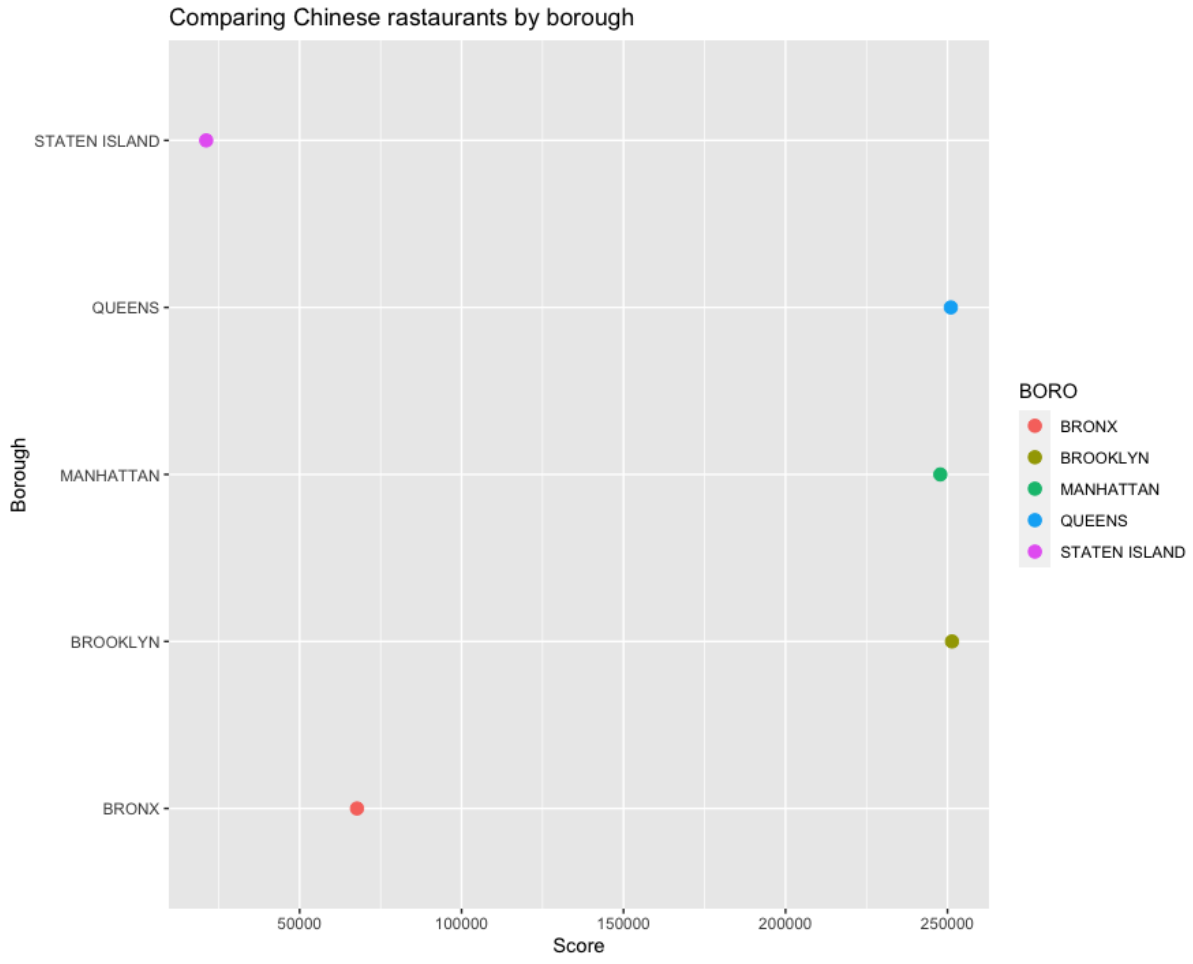

Fig. 8 Comparing Chinese restaurants by borough

It seems like Staten Island is the best place to go if I want to eat a good Chinese.
Since I am encoding a quantitative variable in order to express magnitude, I could have also used length, resulting in a bar plot. However, I have already used a barplot already and I wanted something different. Position is still a good choice according to Bertin's Levels of Organisation.

Plot elements

- **X - axis:** Quantitative variable *Score*
- **Y - axis:** Nominal variable *Borough*

Visual encoding

- Using position to encode score which makes it easy and fast to interpret quantitative variables.
- Also using colour to encode the boroughs. Although it is not necessary and we could do without it by only reading along the Y scale.

Effectiveness ranking

- there are only 5 categories to compare, thus they don't overlap, which would have made this visualization less effective.

Expressiveness ranking

- expresses only the information that is inherent in the dataset

V. Discussion

Overall, I consider the visualizations to have answered my questions really well. Although, if I wanted to dive even deeper into the dataset I can use interaction in the future.