TO EAT OR NOT TO EAT, THAT IS THE QUESTION?

A data derived analysis on New York City's restaurants.

Artemas Wang

Data science has recently become a trendy buzzword in many a social circle. Yet what exactly is data science and what does it really do? According Hadley Wickham and Garrett Grolemund, data science is a field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data (Wickham, 2017). But how do these methods help people make decisions and why are these data driven insights important? This project aspired to answer these questions by examining a New York City governmental dataset on restaurants in the five boroughs and their various cuisines, scores, and grades. By conducting an in-depth analysis on the dataset using R and applying a variety of visualizations with several R packages, the results of the analysis yielded insightful answers to the confounding questions around data science and its impact.

Description of Data and Methodology

The overarching ethos of this project was to help consumers make better informed decisions about dining in New York City. The dataset used for this project can be found on the New York State government website under health and restaurant grades. It is updated on a monthly basis by the Department of Health and Mental Hygiene (DOHMH) and the most recent update of the version was used for this analysis, dated December 8th, 2019. The original dataset had a total of 25 variables and 397,024 observations. However, for the purposes of this analysis, only 7 variables were used.

Variables	Zip Code
Restaurant Name	Cuisine
NYC (Boro) Borough	Score
Street	Grade

Moreover, because the project is concerned with helping consumers make better dining decisions, only relevant variables such as the restaurant's name, street, cuisine type, scores, grades, and zip code within New York City's boroughs were selected. The aforementioned variables were selected to guide the project and sought to answer a variety of questions that included:

- The borough in NYC that had the highest and lowest average restaurant score.
- The letter grade given the most often in each borough.
- The borough with the most restaurants.
- Whether there exists a correlation between a restaurant's score and the cuisine type?
- Whether there exists a correlation between a restaurant's score and zip code?

A variety of R packages were also used for this project and included libraries that could manipulate, analyze, model, and visualize the dataset to provide the consumer a greater understanding of New York City restaurants and the various nuances within their scores. The packages used in this analysis were:

R Package	**Usage**
tidyverse	Data Manipulation
dplyr	Data Manipulation
tidyr	Data Manipulation
data.table	Data Manipulation
MASS	Data Analysis/Modeling
ggplot2	Data Visualization

Packages like tidyverse, dplyr, tidyr, and data.table helped manipulate and sort the data and its many observations. The MASS package allowed for a chi-squared test to be conducted

for independence and to check for p-value and significance. Finally, the ggplot2 package provided a visualization detailing the relationship between scores, grades and zip code.

Data Tidying and Analysis

The process and methodology of this analysis will now be discussed. Once the data set was downloaded and read into RStudio, it was then tidied to achieve coherence. Beginning with an overview of the variables (column names) and the amount of observations (restaurants), a general understanding of the data set was conducted. Once the necessary variables were selected by relevance, feature engineering was performed to prepare the data set for analysis. Of note during this process was the amount of NA and missing values within the data set. To address this, a nested ifelse() loop was applied to the data set so that every restaurant that had a score also had a grade. Restaurants without a score were filtered with the complete.cases() function. Furthermore, only unique() restaurants with complete scores were used in the analysis and duplicate restaurants were also removed. Finally, according to the New York Data governmental website, the grading scale used to critique restaurants was: ***A restaurant inspection score of 0-13 resulted in an 'A', 14-27 points a 'B;, and 28-40 a 'C', 40 or more is resulted in temporary closure of the restaurant.***

To address the questions asked of the data set, the tidied data was then indexed by New York City's five boroughs. Once this was completed, mathematical analysis was conducted to ascertain the mean(), max(), and min() scores of all the restaurants in each borough. The results yielded the following:

Borough	**Average Score**
Bronx	Average: **15.68**
Brooklyn	Average: **16.04**
Manhattan	Average: **15.86**

Queens	Average: **15.78**
Staten Island	Average: **15.65**

The results of the mathematical analysis indicated that the average grade for all the boroughs was a B, and the borough with the highest average score was Brooklyn with Staten Island the possessing lowest average score. The range between mean scores within boroughs could be because of the amount of restaurants in each borough, as the higher the volume of restaurants the higher the variance of scores, which could affect the mean. A quick look at the amount of restaurants in each borough reflects this assumption (total restaurant grades in appendix).

Borough	**Total Restaurants**
Bronx	Total Restaurants: **10538**
Brooklyn	Total Restaurants: **28836**
Manhattan	Total Restaurants: **44757**
Queens	Total Restaurants: **26193**
Staten Island	Total Restaurants: **3900**

Although the number of restaurants per borough could be alarming, it must be noted that New York State law stipulates that any establishment that serves food or beverage needed to be examined, therefore a neighborhood coffee shop is graded for cleanliness and hygiene the same as a Michelin starred restaurant.

To provide greater insight, a chi-squared test using the MASS package was conducted on the restaurants within each borough and its relation to three distinct variables, the restaurant's score, cuisine type, and zip code. The test was predicated on two hypotheses with relation to a restaurant's score. The hypotheses were:

H_0 - No relationship between cuisine type and restaurant score.

H_0 - No relationship between zip code and restaurant score.

The findings of chi-squared test however proved insightful as there showed some correlation between the aforementioned variables. In the table below, the test indicated that there was indeed some correlation between the cuisine type and restaurant score.

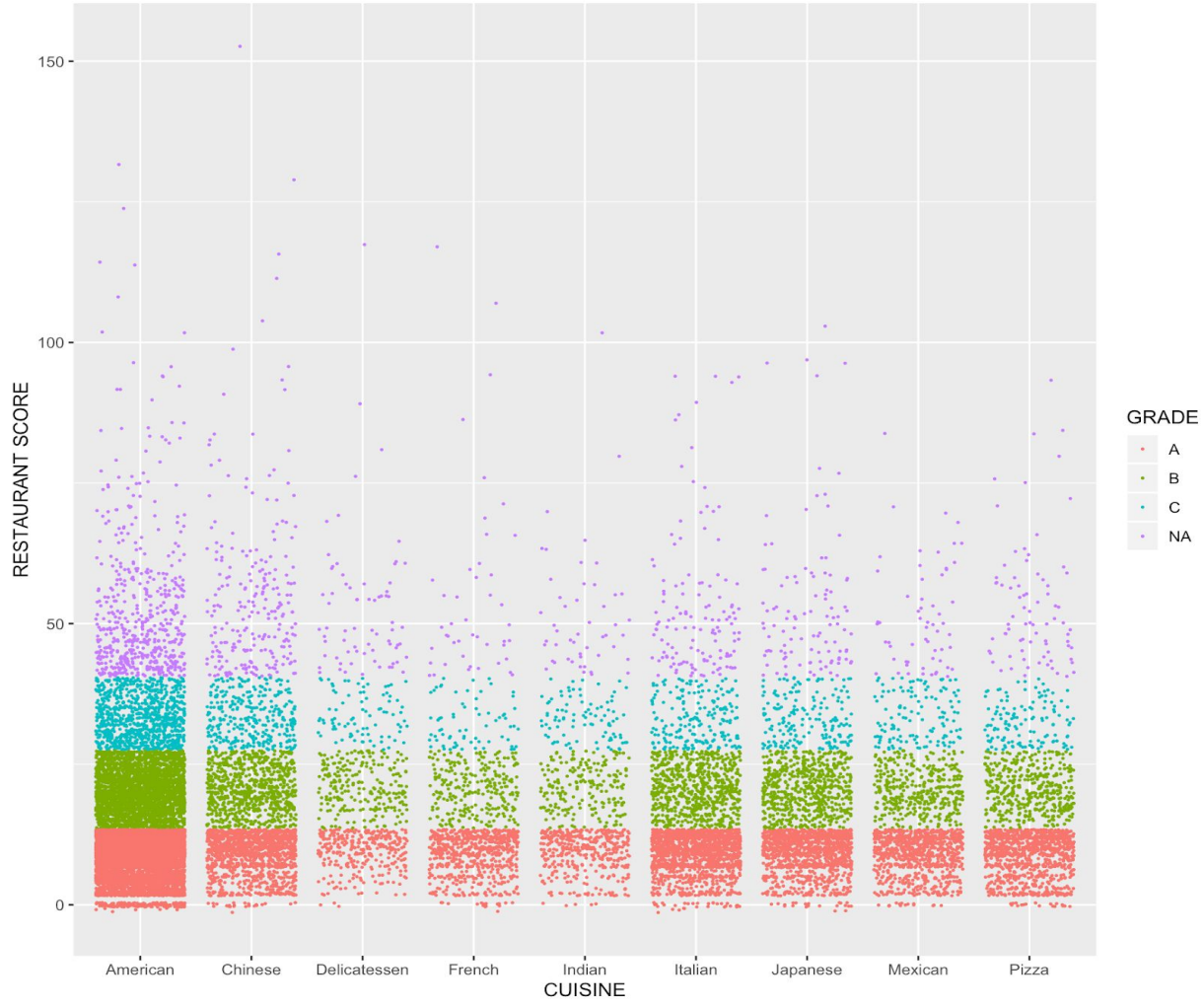
Borough	**p-value**
Bronx	Cuisine / Score: **0.3879** Zip Code / Score: **0.9982**
Brooklyn	Cuisine / Score: **2.2e-16** Zip Code / Score: **0.9217**
Manhattan	Cuisine / Score: **2.2e-16** Zip Code / Score: **1**
Queens	Cuisine / Score: **2.2e-16** Zip Code / Score: **0.8582**
Staten Island	Cuisine / Score: **0.9988** Zip Code / Score: **0.7745**

Therefore, at the .05 significance level, the chi-squared test indicated that the null hypothesis, cuisine type is not related to restaurant score, can be rejected in three boroughs, Brooklyn, Manhattan, and Queens. It must be noted however, that the p-values for cuisine types, of which there are 84 unique cuisine varieties, could possibly depend also on the total amount of restaurants within the three largest boroughs. To account for this possible discrepancy, additional feature engineering was needed to fit the chi-squared test before visualization.

Data Visualization

As such, the data was tidied again but only for the borough of Manhattan, with only the top ten variety of cuisines represented based on volume. Consequently, the extra tidying

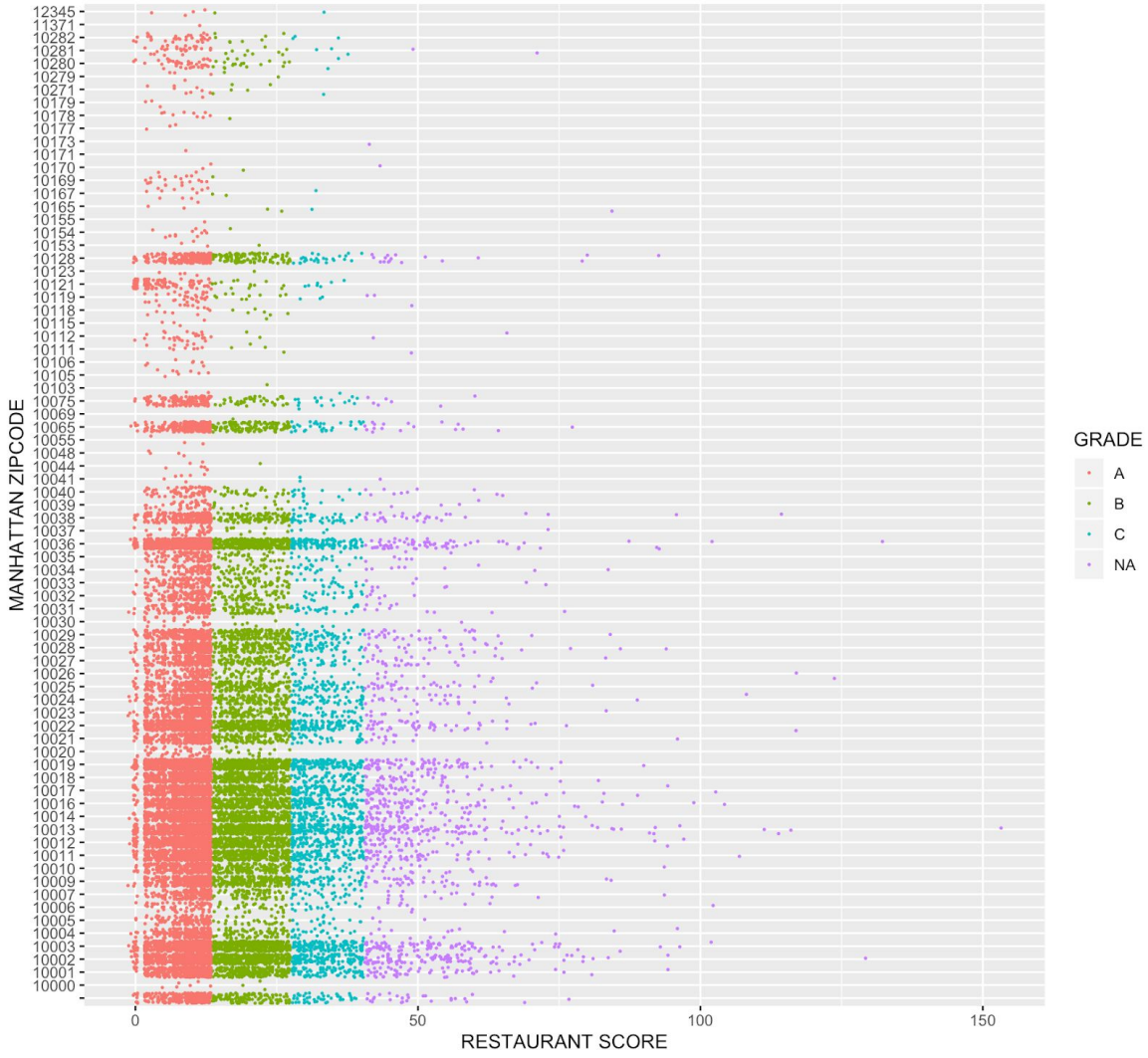
provided a more accurate analysis of the variables questioned. The two ggplot2 visualizations below now accurately displays the relationships between restaurant scores, cuisine type, and zip code. In the first visualization, the restaurant score and the cuisine type were placed on the y-axis and x-axis respectively. In the second visualization, the restaurant's zip code in Manhattan and restaurant score were placed on the y-axis and x-axis accordingly. Both plots had a legend that displayed their grade colored as a factor A, B, C, or NA.



As the visualization displays, there exists a higher correlation between a restaurant's score and it's cuisine type for American and Chinese restaurants, and less so for Delicatessens and French Cuisine. Despite a few outliers in purple that failed the cleanliness test, the other varieties of cuisine seem on par with the mean grade for Manhattan. Whether this distinction

exists because of the nature of the cuisine type, how it is prepared, or the types of ingredients involved, the choice is up to the diner and their preferences. Presumably, a qualitative analysis on food preparation and employee training might shed more light on this particular relationship.

With regards to a possible relationship between a restaurant's location (zip code) and it's score, the visualization suggests that there is such a relationship.



The visualization above indicates that there exists a relationship between restaurant location and score. The more densely populated the zip code, the lower the restaurant score, and the less densely populated the zip code, the better the restaurant score is on average. The zip

codes towards the bottom of the visualization are the locations around midtown Manhattan, while the zip codes towards the top of the visualization are more residential areas like the Upper West and Upper East sides of Manhattan where residential buildings outnumber commercial buildings. With restaurant scores and zip code, in plain speech, the more the messier.

Future Considerations

While the project itself yielded some interesting conclusions, there is still so much more information in this data set yet to be answered. In retrospect, the sheer amount of restaurants analyzed made for any concrete resolutions to be taken with a grain of salt. Moreover, additional feature engineering would have possibly yielded better results and visualizations in general. Likewise, a more nuanced approach to the handling of missing restaurants, values, and scores could have offered more meticulous answers. However, given the nature of the data, how it was compiled, and the Department of Health's interesting lack of transparency in explaining why values went missing added further complexity to the project and subsequent analysis. Despite this, and as it pertains to Manhattan, the analysis does provide a Tufte approved visualization that there does exist a correlation between a restaurant's score and its cuisine type and zip code. But what does this all mean for diners? Well, the proof is in the pudding.

REFERENCES

Matloff, N. (2011). *The Art of R Programming: A Tour of Statistical Software Design*. San Francisco, CA: No Starch Press.

Tufte, E. (2001). *The Visual Display of Quantitative Data (2nd ed.)*. Cheshire, CT: Graphics Press.

Wickam, H. & Grolemund, G (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol, CA: O'Reilly Media, Inc.

R PACKAGES USED AND CONSIDERED

R Package	Usage
tidyverse	Data Manipulation
dplyr	Data Manipulation
tidyr	Data Manipulation
data.table	Data Manipulation
MASS	Data Analysis/Modeling
randomForest	Data Analysis/Modeling
rplot	Data Analysis/Modeling
rpart.plot	Data Analysis/Modeling
ggplot2	Data Visualization

APPENDIX

```
```

```
library(tidyverse)
library(data.table)
library(dplyr)
library(tidyr)
library(ggplot2)

# Uploading dataset from https://data.cityofnewyork.us/browse?q=food+grades
restaurants_original <- fread(paste0("NYC Restaurant Inspections.csv"), header = T,
stringsAsFactors = F, data.table = T)
save(restaurants_original, file = "restaurants_original.Rdata")
restaurants <- restaurants_original

### UNDERSTANDING DATA SET ###
# colnames(restaurants)
# [1] "CAMIS"                 "DBA"                   "BORO"                  "BUILDING"
# [5] "STREET"                "ZIPCODE"               "PHONE"                 "CUISINE
DESCRIPTION"
# [9] "INSPECTION DATE"       "ACTION"                "VIOLATION CODE"        "VIOLATION
DESCRIPTION"
# [13] "CRITICAL FLAG"         "SCORE"                 "GRADE"                 "GRADE DATE"
# [17] "RECORD DATE"           "INSPECTION TYPE"       "Latitude"              "Longitude"
# [21] "Community Board"       "Council District"      "Census Tract"          "BIN"
# [25] "BBL"                   "NTA"

# Changing column names, easier to understand
colnames(restaurants)[2] = "RESTAURANT NAME"
colnames(restaurants)[3] = "NYC BORO"
colnames(restaurants)[8] = "CUISINE"

### DATA TIDYING ###
restaurants <- dplyr::select(restaurants, "RESTAURANT NAME","NYC
BORO","STREET","ZIPCODE","CUISINE","SCORE","GRADE")
str(restaurants)
colnames(restaurants)
class(restaurants)
nrow(restaurants)  # 379,819
ncol(restaurants) # 7

# Only accept restaurants with NYC gov documented restaurant scores, complete.cases
which(is.na(restaurants$SCORE))  # 17,205 restaurants without a score
restaurant_scores <- restaurants[ , c("SCORE")]
restaurants_complete <- restaurants[complete.cases(restaurant_scores), ]

# Feature Engineering
# Adding Scores and Grades based on NYC Grading Scale to fill in missing values

# An inspection score of 0-13 is an A, 14-27 points is a B, and 28-40 is a C, 40 or more is
Temporary Closure.
restaurants_complete$GRADE <- ifelse(restaurants_complete$SCORE <= 13, 'A',
                                     ifelse(restaurants_complete$SCORE <= 27 &
restaurants_complete$SCORE > 13, 'B',
```

```
                                          ifelse(restaurants_complete$SCORE <= 40 &
restaurants_complete$SCORE > 27, 'C', 'NA')))

# Duplicate rows removed
restaurants_complete <- restaurants_complete[order(restaurants_complete$`NYC BORO`),]
complete <- distinct(restaurants_complete)
str(complete)

# Indexing Restaurants into Borough
unique(complete$`NYC BORO`)  # "0" Bronx" "Brooklyn" "Manhattan" Queens" "Staten Island"
which(complete$`NYC BORO` == 'Bronx')
bronx <- complete[25:10562,]

which(complete$`NYC BORO` == 'Brooklyn')
brooklyn <- complete[10563:39398,]

which(complete$`NYC BORO` == 'Manhattan')
manhattan <- complete[39399:84155,]

which(complete$`NYC BORO` == 'Queens')
queens <- complete[84156:110348,]

which(complete$`NYC BORO` == 'Staten Island')
staten_island <- complete[110349:114248,]

allboros <- c(bronx, brooklyn, manhattan, queens, staten_island)

# Amount of Restaurants, Min, Max, and Mean restaurant scores in each Borough
# BRONX
length(bronx$`NYC BORO`) # 10538 Restaurants
max(bronx$SCORE)  # 164
min(bronx$SCORE)  # -1
mean(bronx$SCORE)  # 15.68
# BROOKLYN
length(brooklyn$`NYC BORO`) # 28836 Restaurants
max(brooklyn$SCORE) # 151
min(brooklyn$SCORE) # -1
mean(brooklyn$SCORE)  # 16.04
# MANHATTAN
length(manhattan$`NYC BORO`) # 44757 Restaurants
max(manhattan$SCORE) # 153
min(manhattan$SCORE) # -1
mean(manhattan$SCORE)  # 15.86
# QUEENS
length(queens$`NYC BORO`) # 26193 Restaurants
max(queens$SCORE)  # 164
min(queens$SCORE)  # -1
mean(queens$SCORE)  # 15.78
# STATEN ISLAND
length(staten_island$`NYC BORO`) # 3900
max(staten_island$SCORE)  # 117
min(staten_island$SCORE) # -1
mean(staten_island$SCORE) # 15.65

# Which letter grade is given out most often per Borough
length(grep("A", bronx$GRADE)) # 7007
length(grep("B", bronx$GRADE)) # 2649
```

```
length(grep("C", bronx$GRADE)) # 882
length(grep("NA", bronx$GRADE)) # 488

length(grep("A", brooklyn$GRADE)) # 19459
length(grep("B", brooklyn$GRADE))  # 7084
length(grep("C", brooklyn$GRADE))  # 2293
length(grep("NA", brooklyn$GRADE))  # 1608

length(grep("A", manhattan$GRADE))  # 30099
length(grep("B", manhattan$GRADE))  # 10992
length(grep("C", manhattan$GRADE))  # 3666
length(grep("NA", manhattan$GRADE))  # 2245

length(grep("A", queens$GRADE))  # 17792
length(grep("B", queens$GRADE))  # 6323
length(grep("C", queens$GRADE))  # 2078
length(grep("NA", queens$GRADE))  # 1331

length(grep("A", staten_island$GRADE))  # 2548
length(grep("B", staten_island$GRADE))  # 1028
length(grep("C", staten_island$GRADE))  # 324
length(grep("NA", staten_island$GRADE))  # 159

# Visualization for Cuisine Type and Restaurant Scores
length(unique(complete$CUISINE)) # 84 unique cuisine types.

ggplot(complete, aes(complete$SCORE, complete$CUISINE, color = complete$CUISINE)) +
  geom_point(show.legend = FALSE) +
  xlab('RESTAURANT SCORE') +
  ylab('CUISINE TYPE')

### DATA ANALYSIS ###
library(MASS)

# Chi-squared test for independence
# P <= .05 reject null, P >= .05 cannot reject null
# Null Hypothesis - No relationship between Cuisine and Score
# Null Hypothesis - No relationship between Zipcode and Score

bronx_cuisine_score <- chisq.test(table(bronx$CUISINE, bronx$SCORE))
# data:  table(bronx$CUISINE, bronx$SCORE)
# X-squared = 4811.2, df = 4784, p-value = 0.3879
bronx_zipcode_score <- chisq.test(table(bronx$ZIPCODE, bronx$SCORE))
# data:  table(bronx$ZIPCODE, bronx$SCORE)
# X-squared = 2107.8, df = 2300, p-value = 0.9982
ggplot(bronx, aes(bronx$SCORE, bronx$CUISINE, color = bronx$CUISINE)) +
  geom_point(show.legend = FALSE)

brooklyn_cuisine_score <- chisq.test(table(brooklyn$CUISINE, brooklyn$SCORE))
# data:  table(brooklyn$CUISINE, brooklyn$SCORE)
# X-squared = 10123, df = 8970, p-value < 2.2e-16
brooklyn_zipcode_score <- chisq.test(table(brooklyn$ZIPCODE, brooklyn$SCORE))
# data:  table(brooklyn$ZIPCODE, brooklyn$SCORE)
# X-squared = 4464.8, df = 4600, p-value = 0.9217
ggplot(brooklyn, aes(brooklyn$SCORE, brooklyn$CUISINE, color = brooklyn$CUISINE)) +
  geom_point(show.legend = FALSE)
```

```
manhattan_cuisine_score <- chisq.test(table(manhattan$CUISINE, manhattan$SCORE))
# data:  table(manhattan$CUISINE, manhattan$SCORE)
# X-squared = 10997, df = 9322, p-value < 2.2e-16
manhattan_zipcode_score <- chisq.test(table(manhattan$ZIPCODE, manhattan$SCORE))
# data:  table(manhattan$ZIPCODE, manhattan$SCORE)
# X-squared = 8075.9, df = 9794, p-value = 1
ggplot(manhattan, aes(manhattan$SCORE, manhattan$ZIPCODE, color = manhattan$CUISINE)) +
  geom_point(show.legend = FALSE)

queens_cuisine_score <- chisq.test(table(queens$CUISINE, queens$SCORE))
# data:  table(queens$CUISINE, queens$SCORE)
# X-squared = 10263, df = 8778, p-value < 2.2e-16
queens_zipcode_score <- chisq.test(table(queens$ZIPCODE, queens$SCORE))
# data:  table(queens$ZIPCODE, queens$SCORE)
# X-squared = 7166.6, df = 7296, p-value = 0.8582
ggplot(queens, aes(queens$SCORE, queens$ZIPCODE, color = queens$ZIPCODE)) +
  geom_point(show.legend = FALSE)

staten_island_cuisine_score <- chisq.test(table(staten_island$CUISINE, staten_island$SCORE))
# data:  table(staten_island$CUISINE, staten_island$SCORE)
# X-squared = 3710.2, df = 3975, p-value = 0.9988
staten_island_zipcode_score <- chisq.test(table(staten_island$ZIPCODE, staten_island$SCORE))
# data:  table(staten_island$ZIPCODE, staten_island$SCORE)
# X-squared = 867.75, df = 900, p-value = 0.7745
ggplot(staten_island, aes(staten_island$SCORE, staten_island$ZIPCODE, color =
staten_island$ZIPCODE)) +
  geom_point(show.legend = FALSE)

# FEATURE ENGINEERING FOR MANHATTAN TO IMPROVE VISUALIZATION
as.factor(manhattan$ZIPCODE)
str(manhattan)
length(unique(manhattan$CUISINE)) # 80 unique cuisines

# Top 10 restaurants by volume
as.data.frame(table(manhattan$CUISINE))
# American 12,661
# Chinese 2715
# Italian 2639
# Japanese 2139
# Pizza 1667
# Mexican 1490
# French 1053
# Indian 761
# Delicatessen 716
# Mediterranean 632

patterns <- c('American','Chinese','Italian','Japanese','Pizza',
              'Mexican','French','Indian','Delicatessen','Meditteranean')

manhattan_top10 <- manhattan[manhattan$CUISINE %in% patterns, ]

# HYPOTHESIS TESTING FOR MANHATTAN RESTAURANTS
manhattan_top10_cuisine_score <- chisq.test(table(manhattan_top10$CUISINE,
manhattan_top10$SCORE))
# data:  table(manhattan_top10$CUISINE, manhattan_top10$SCORE)
# X-squared = 1348.5, df = 872, p-value < 2.2e-16
```

```
manhattan_top10_zipcode_score <- chisq.test(table(manhattan_top10$ZIPCODE,
manhattan_top10$SCORE))
# data:  table(manhattan_top10$ZIPCODE, manhattan_top10$SCORE)
# X-squared = 7183.5, df = 8284, p-value = 1

### DATA VISUALIZATIONS ###
# Cuisine and Score - Manhattan
ggplot(manhattan_top10, aes(manhattan_top10$CUISINE, manhattan_top10$SCORE, color = GRADE)) +
  geom_jitter(show.legend = TRUE, size = 0.25) +
  xlab('CUISINE') +
  ylab('RESTAURANT SCORE')
# Zip Code and Score - Manhattan
ggplot(manhattan_top10, aes(manhattan_top10$SCORE, manhattan_top10$ZIPCODE, color = GRADE)) +
  geom_jitter(show.legend = TRUE, size = 0.25) +
  xlab('RESTAURANT SCORE') +
  ylab('MANHATTAN ZIPCODE')

# OTHER TESTS CONSIDERED BUT NOT USED IN REPORT

library(rpart.plot)
library(rpart)

art <- rpart(SCORE ~ CUISINE + ZIPCODE, data = manhattan_top10)
rpart.plot(art)
summary(art)
printcp(art)

#CP            nsplit rel error    xerror       xstd
#1 0.01254235      0 1.0000000 1.0001092 0.01923212
#2 0.01000000      1 0.9874577 0.9876399 0.01894708

# Multiple Regression Test
tester1 <- lm(SCORE ~ CUISINE, data = manhattan_top10)
summary(artey)

# ANOVA Test
tester2 <- aov(SCORE ~ CUISINE, data = manhattan_top10)
summary(arte)

# t.test
t.test(manhattan_top10$SCORE)

```