An Analysis of Restaurant Cleanliness in New York City

Anh Vu Nguyen, Anh Nguyen, Cody Couture

December 17, 2015

I. Introduction

Over the past several decades, Americans have drastically increased the amount of time and money that they spend at restaurants. In fact, today Americans spend more money eating in restaurants than they do buying food in grocery stores. On average, Americans eat at a restaurant nearly 6 times a week. However, despite spending so much time eating out, many Americans tend to spend relatively little time considering the impact that this food may have on their health. As the recent outbreaks at Chipotle restaurants across the country have shown, the food that we eat in restaurants is not always safe. In fact, according to the CDC, nearly 50 million Americans will become ill this year as a result of foodborne sickness. Of those who are affected, over 100,000 will be hospitalized and approximately 3,000 will die. Given the substantial risk posed by unsanitary food preparation, it is essential that consumers have the best information possible when choosing to eat at a restaurant. As a result, we as a society have created a system of restaurant inspections and accompanying safety ratings that assist restaurant goers in choosing a restaurant. Although the safety rating system is far from perfect, it is a useful tool in protecting the health of American consumers.

In our study, we examine a data set of New York City food safety inspections. Due to its large population, large variety of restaurants, and a culture of eating out, New York City is an ideal candidate for observation. We are particularly interested in examining how food safety ratings vary across different types of cuisine. Certain types of cuisine, such as Chinese food, often have a reputation for being of a lower quality, and therefore, more dangerous to eat. If such a reputation is accurate, then consumers should adjust their behavior to account for this risk. However, if the reputation is inaccurate, then there is no reason for consumers to discriminate against these restaurants. Moreover, New York City has famously struggled with pest problems, particularly rodents. We want to determine if such a problem currently exists within New York City restaurants and, if so, determine the severity of the problem.

Since its introduction, the NYC restaurant inspection system has been scrutinized by a multitude of studies. Some of these studies attempt to model the grades over a specific period of time while others examine the accuracy of the inspection ratings. The latter of these two avenues of research has yielded a variety of conflicting opinions about the trustworthiness of this grading system. One study by iQuantNY seems to suggest that the grading system is flawed while another article suggests that, while the ratings are not perfect, they do what they are intended to do. We expand on the previous literature of modelling restaurant inspections by highlighting differences in food safety rating across restaurant type and location. We hope this model will help to reduce the number of illnesses caused by foodborne sickness.

II. Data Collection and Cleanup

The entire dataset was loaded from NYC Open Data Portal on November 14, 2015. The raw dataset contained of 484,743 rows and was loaded into RStudio for data cleanup process.
There were several considerations made while cleaning up the dataset to be use for this study:

1. **We only consider the most recent restaurant inspection**: This is because the most recent inspection is also the most relevant inspection. A person eating at a restaurant will likely care less about a bad inspection from three years ago than a bad inspection from three weeks ago.

2. **We take out those inspections without a grade**: This is due to the system of grading used by The New York City Department of Health and Mental Hygiene (DOHMH), where a restaurant is given a second chance if they did poorly on their first inspection.

3. **We look at whether rodents and/or roaches were sighted**: This is significant since pests have been deemed a problem in NYC by people living there, both historically and more recently.

4. **Linking Income with Zipcode**: Since the dataset includes full addresses, we were able to merge an external dataset that includes median income and demographic data and link it using each restaurant's specific zipcode.

5. **Focus on specific cuisine**: NYC is a melting pot of cultures and thus many types of cuisine are available. However, we choose to only consider the most popular cuisines. Those in consideration are:

   - American
   - Asian [1]
   - Chinese
   - Italian
   - Latin/Spanish [2]
   - Pizza
   - Sandwiches

After this process, our dataset was shrunk down to 11,525 rows of observation. Below are the important variables that will be used in this study and a short description for each.

Variable name	Description
camis	Unique Key for each restaurant, set by DOHMH.
dba	Name of restaurant as registered with DOHMH.
boro	Borough that the restaurant was located in.
zipcode	Zipcode for the address of the restaurant.
score	Score from the last restaurant inspection.
grade	Grade from the last restaurant inspection.
rodent_list	Whether rodents were sighted in the last inspection (True/False)
roach_list	Whether roaches were sighted in the last inspection (True/False)
median_income.list	Median income associated with the Zipcode of the restaurants
cuisine_group_factor	Cuisine for the restaurant

The reason we chose to separate Chinese and Asian into two separate groups is that the number of observations for Chinese restaurants is quite high on its own (2,457 observations). Also, since Chinese food is more commonly associated with low quality than Thai, Korean, or Japanese food, it seemed important to keep Chinese food into its own group.

It is also important to point out that most of the restaurants falling under the Asian category are cuisines from Southeast Asian countries. We did not include other Asian cuisines from South Asia or Middle East since the historical and cultural background of these areas are distinct from Southeast Asian cuisines.

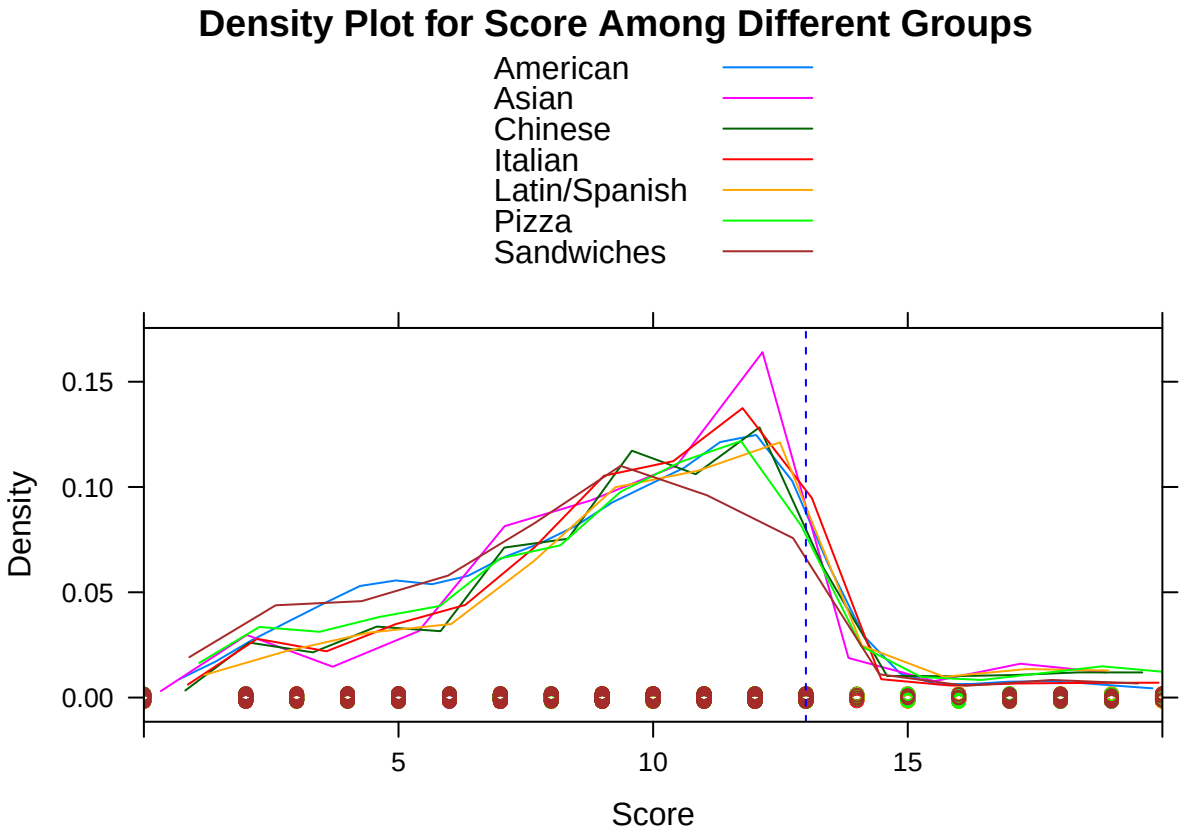
After much consideration, we have chosen to assess the restaurant's cleanliness based on the score they get from the inspection instead of the grade. This is because the grade distribution is highly skewed to A with relatively few B and C grades. This is not a surprise based on several factors. First, as the graph below shows, there is a strong possibility that an inspector is biased towards giving a restaurant the minimum grade required to get an A. The drastic drop in density after score 13 shows that a restaurant rarely gets a score

[1] Combination of Asian, Thai, Japanese, Korean, Vietnamese/Cambodian/Malaysia
[2] Combination of Latin (Cuban, Dominican, Puerto Rican, South & Central American), Tapas, Spanish, Mexican, Caribbean

of 14 despite the fact the scores are numerically so similar. Also, as stated above, the inspections can be done twice for those restaurants that fail to get an A in the first inspections, which should result in a higher average grade. Finally, the jump between a B and A does not tell us as much as the jump from a 29 points to 0 points.

```
##  min Q1 median Q3 max     mean       sd     n missing
##    0  8     10 12  80 10.99393 6.393777 11525       0
```

The score ranges from 0 to 80, with an average score of 11. Since the third quantile gets a value of 12, at least 75% of our restaurants score an 'A' on the inspection.



III. Are Chinese Restaurants the Worst?

1. Are Chinese Restaurants Worse Than Other Restaurants?

We tackled this question by first running a t-test to find the 95% confident interval for the mean score of Chinese restaurants.

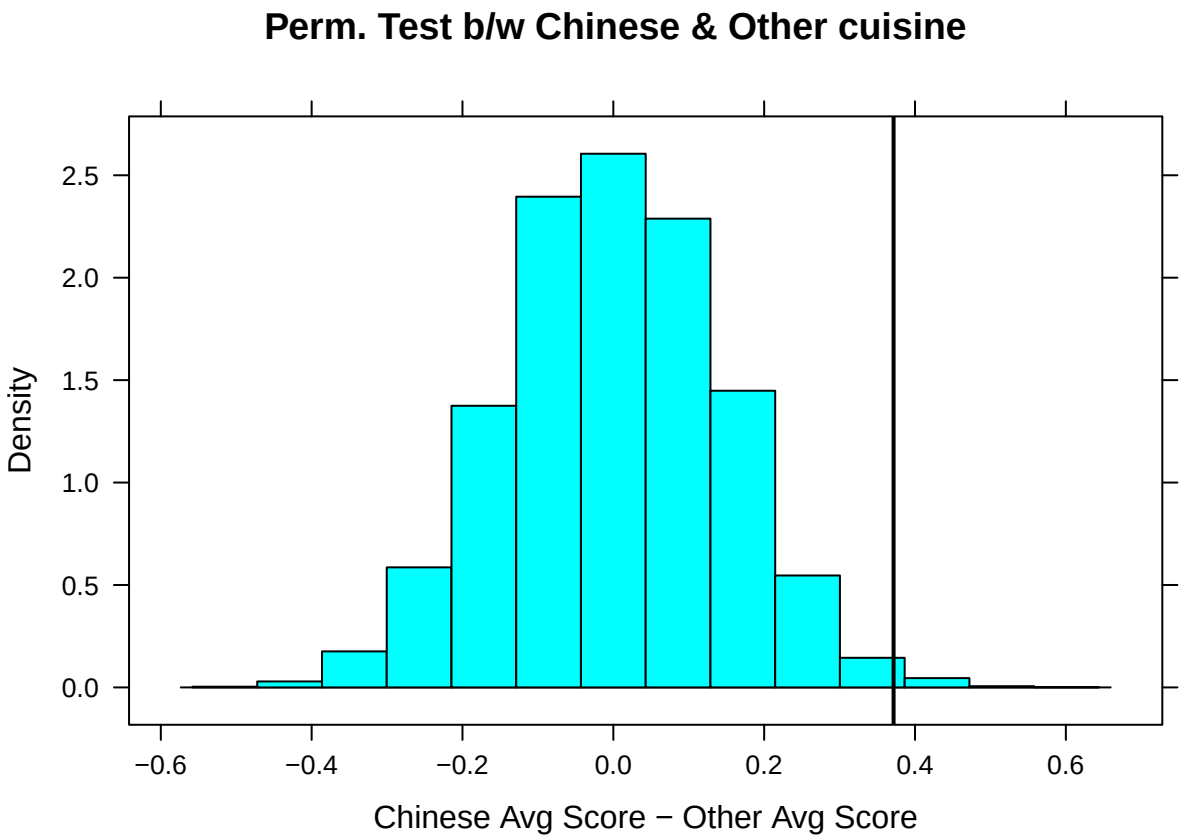
```
## [1] 11.03644 11.53627
## attr(,"conf.level")
## [1] 0.95
```

From the statistics above, we are 95% confident that the mean score of Chinese restaurants will be between 11 and 11.5, equivalent to scoring an 'A' in the inspection.

Next, in order to get a sense of how clean Chinese restaurants are in comparison with other restaurants, we did a permutation test, with mean score as the test statistic. However, since we also want a robust statistics, which are not sensitive to outliers, we also did a permutation test with trimmed mean. The trimmed mean omits a certain fraction of the low and high values, and calculates the mean of remaining values. Our assumption, or alternative hypothesis, for this test was Chinese restaurants would have a higher average score, meaning they are not as clean as other restaurants.

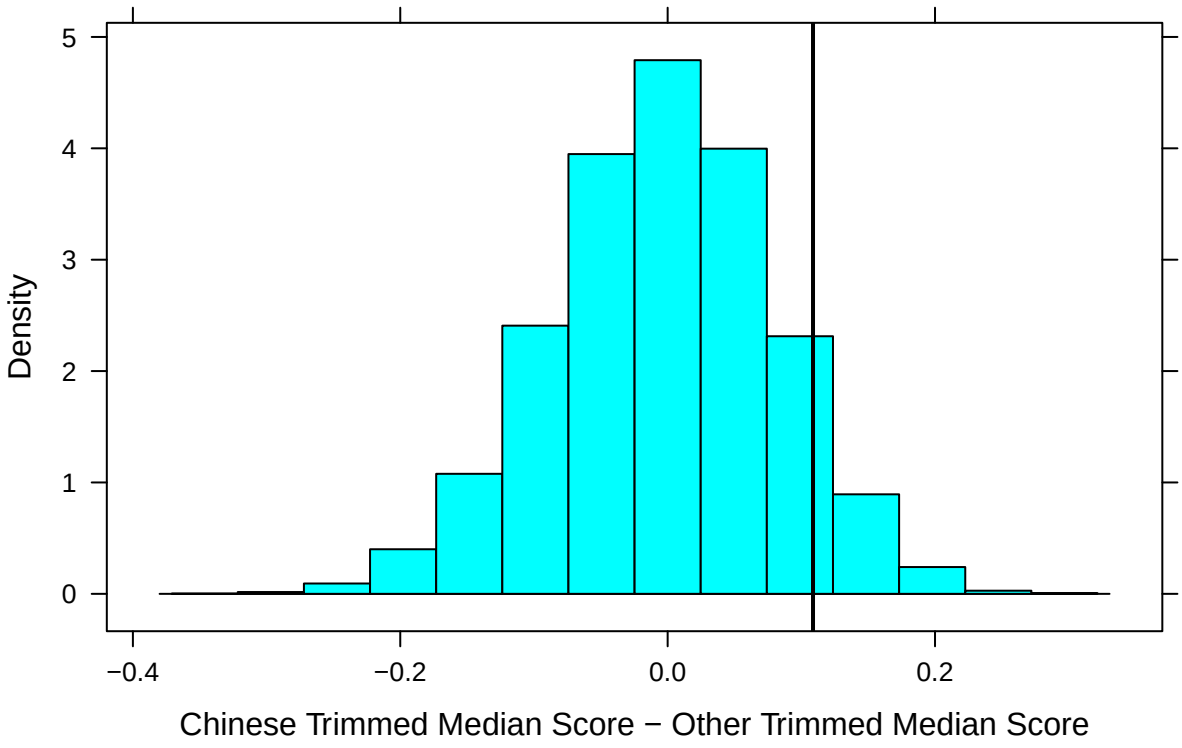
$$H_o : \mu_{Chinese} = \mu_{Asian}$$

$$H_a : \mu_{Chinese} > \mu_{Asian}$$




```
## [1] 0.0059
```

We got a p-value of 0.006, which is smaller than 0.05, suggesting overwhelming evidence to reject the null hypothesis. We concluded that Chinese restaurants on average have higher score than that of other restaurants.




```
## [1] 0.0828
```

For the trimmed mean permutation test, we get a p-value of 0.09, which is higher than 0.05. We concluded that when we removed the Chinese restaurants with 25% lowest and 25% highest scores, the difference in mean score between Chinese and other restaurants is not as extreme anymore. We came to the conclusion that, Chinese restaurants might not be as clean as others, yet, this might happen due to the fact that there exists some Chinese restaurants with extremely high scores.

2. Is There a Cuisine Worse than Chinese?

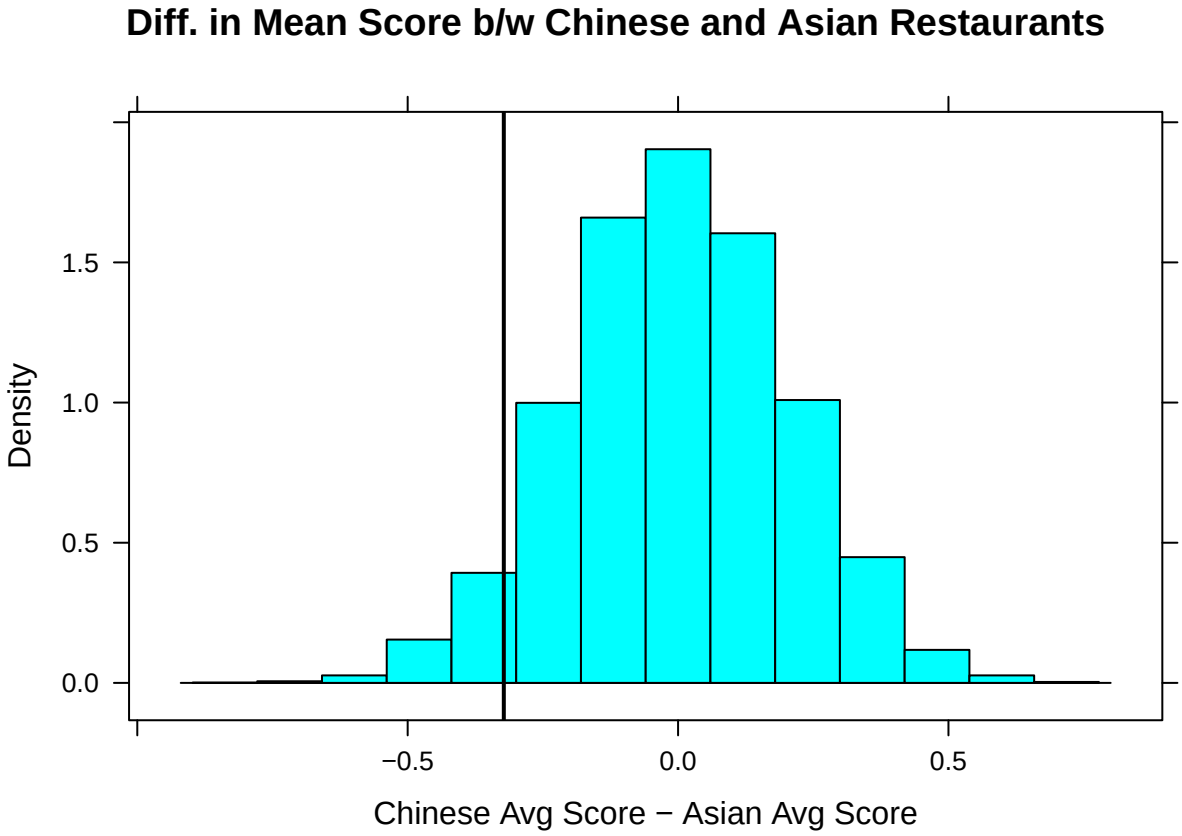
After establishing that Chinese food might not be as clean as other cuisines, we now attempt to answer the question of whether it deserves the stereotype of being the "dirtiest." We consider Asian food, another cuisine with high average score (11.61) and test to see whether its score is better or worse than that of Chinese food. We started by calculating the confidence interval at 95% confidence for the difference in mean score between Chinese and Asian restaurants.

```
## [1] -0.72664195  0.08210737
## attr(,"conf.level")
## [1] 0.95
```

The interval above passes through zero, giving us no strong evidence of difference in mean score right now. Therefore, we proceed to a permutation test for the difference in mean score between Chinese and Asian restaurants, followed by the same test with the trimmed mean score.
Let $\mu_{Chinese}$ and μ_{Asian} be the mean score of Chinese and Asian restaurants respectively. We have the following hypothesis:

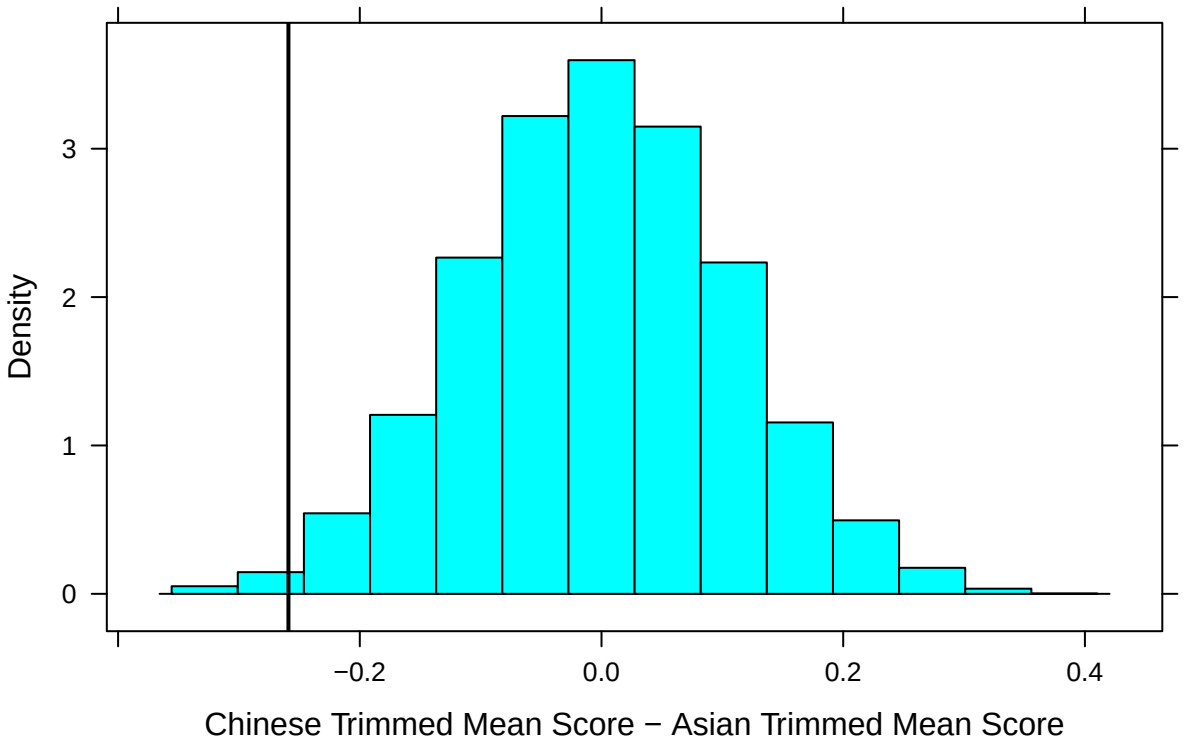
$$H_o : \mu_{Chinese} = \mu_{Asian}$$

$$H_a : \mu_{Chinese} > \mu_{Asian}$$



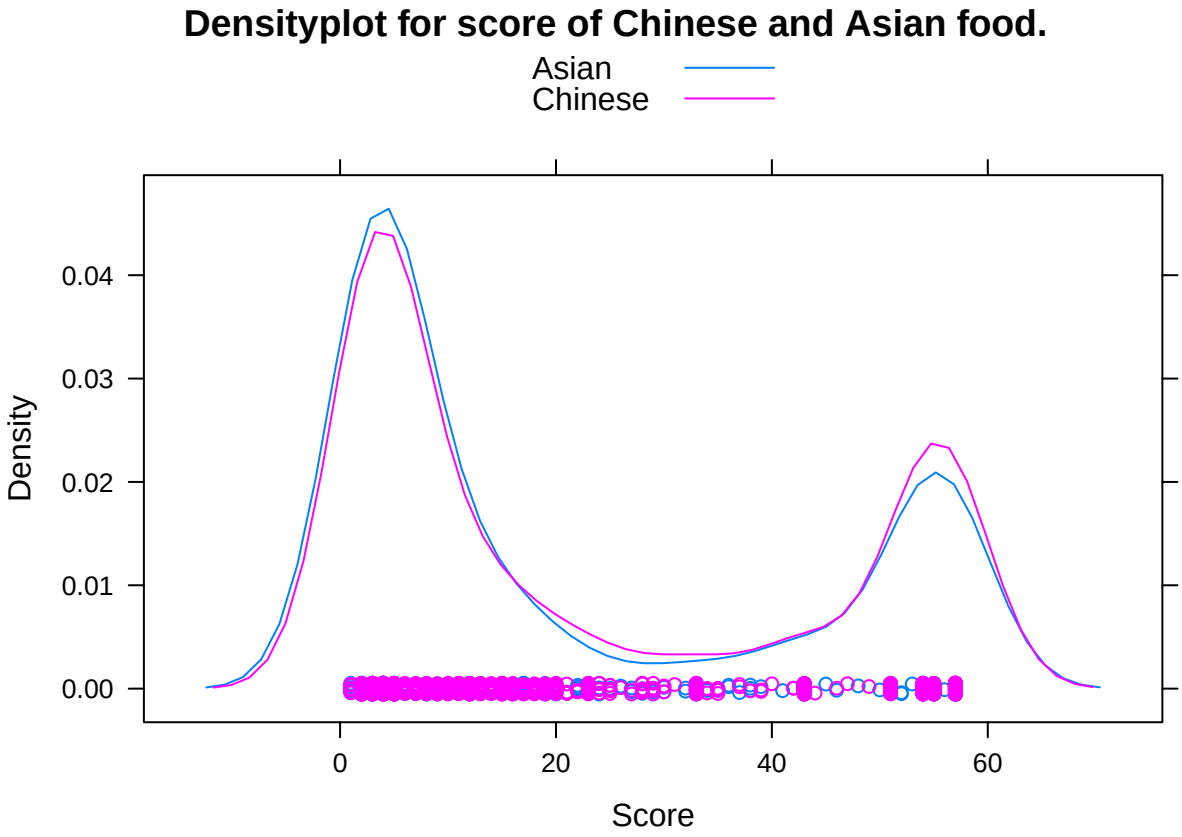

```
## [1] 0.9431
```

We see that p-value is 0.94, which is higher than the α value of 0.05. Therefore, we fail to reject the alternative hypothesis that Chinese cuisine scores on average higher than that of Asian cuisine.




```
## [1] 0.9933
```

Again, we see a high p-value of 0.99. Therefore, we fail to reject the alternative hypothesis that Chinese cuisine scores on average higher than that of Asian cuisine.
Therefore, in both cases, we have failed to reject the null hypothesis, suggesting that Chinese cuisine might not be the dirtiest but might actually be the same as other Asian countries cuisines.
It is also interesting to note the difference between the mean and trimmed mean p-value. The p-value for trimmed mean is higher than that for mean. This means that the difference in observed statistics are more outside the range of normal chance variation for trimmed mean than for mean. Therefore, it might suggest that Chinese restaurants' score might be driven by some really high scoring restaurants while Asian restaurants' score is just in general higher than Chinese's, with not as many outliers that being as extreme as Chinese's.
We confirm this with the density plot below:



As we can see, Asian cuisine has a lower density of outliers with score of above 40. This means that it does not have as many outliers on that end. However, for the peak close to 10, we can see that there is higher concentration of Asian cuisine. The peak of Chinese food is slightly to the left of that for Asian food. This could be an explanation for the differences in mean and trimmed mean's p-values.

3. Regression Modeling

Inspired by the previous findings, we expanded our test, looking at the income variable and trying to see if there exists any relationship between restaurants score and median income.

a. Simple Linear Model

We started by running the linear regression test to build the model linking score and income.

```
##
## Call:
## lm(formula = sqrt(score) ~ median_income.list, data = restaurant.clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2232 -0.3898 -0.0151  0.2905  5.7363
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.240e+00  2.073e-02 156.307   <2e-16 ***
## median_income.list -6.566e-07  3.065e-07  -2.142   0.0322 *
```

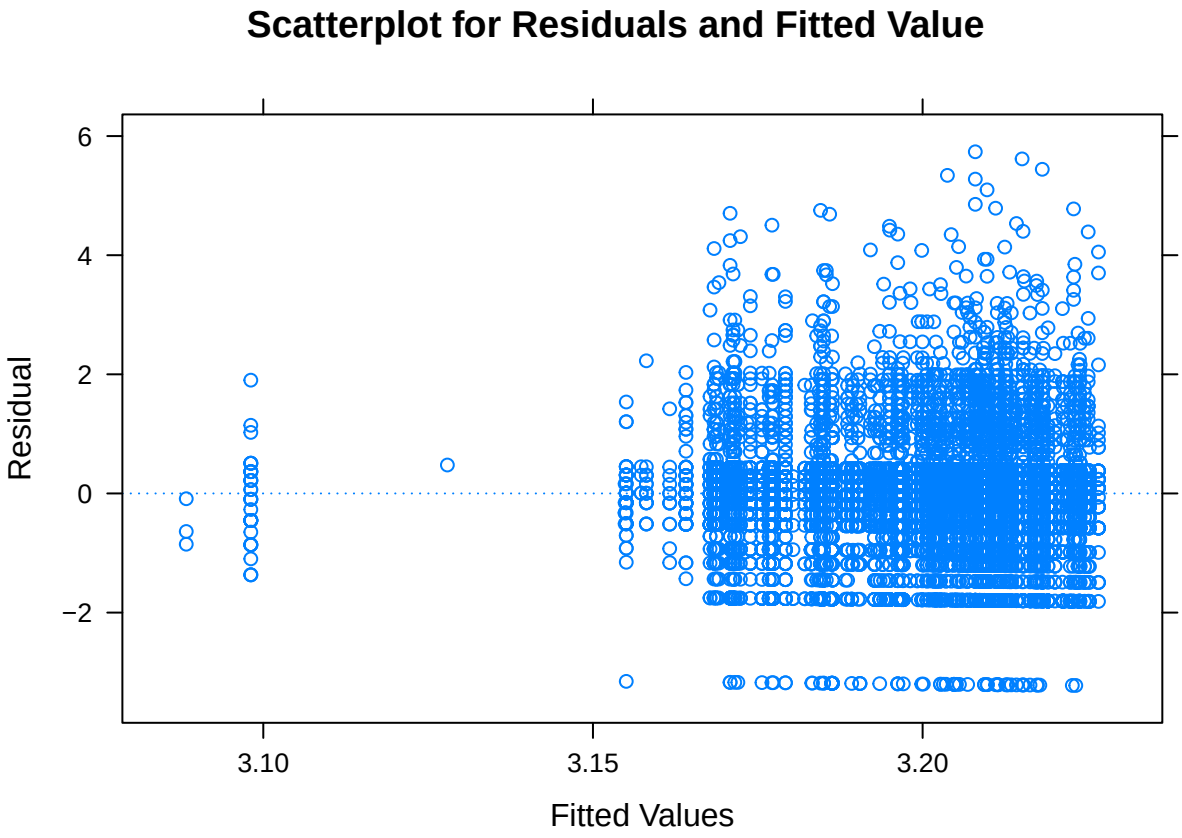
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8853 on 11373 degrees of freedom
##   (150 observations deleted due to missingness)
## Multiple R-squared:  0.0004034,  Adjusted R-squared:  0.0003155
## F-statistic:  4.59 on 1 and 11373 DF,  p-value: 0.03219
```

From the summary above, the least-squares equation line fitting the relationship between restaurants score and median income will be,

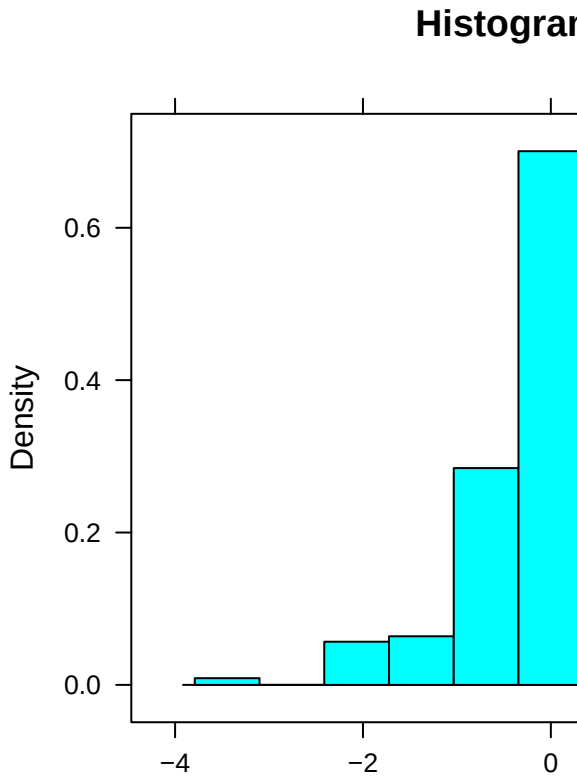
$$\widehat{\sqrt{Score}} = 3.6 - 6.57 \times 10^{-7} \times MedianIncome$$

In other words, for each increase of \$10,000 in median income, $\sqrt{Score}$ decreases by 0.07 point. Also with a p-value of 0.03, which is less than 0.05, Median Income is a significant variable. Nevertheless, the R-squared value is approximately 0, indicating that this model might not be the best at explaining the relationship between Score and Median Income.

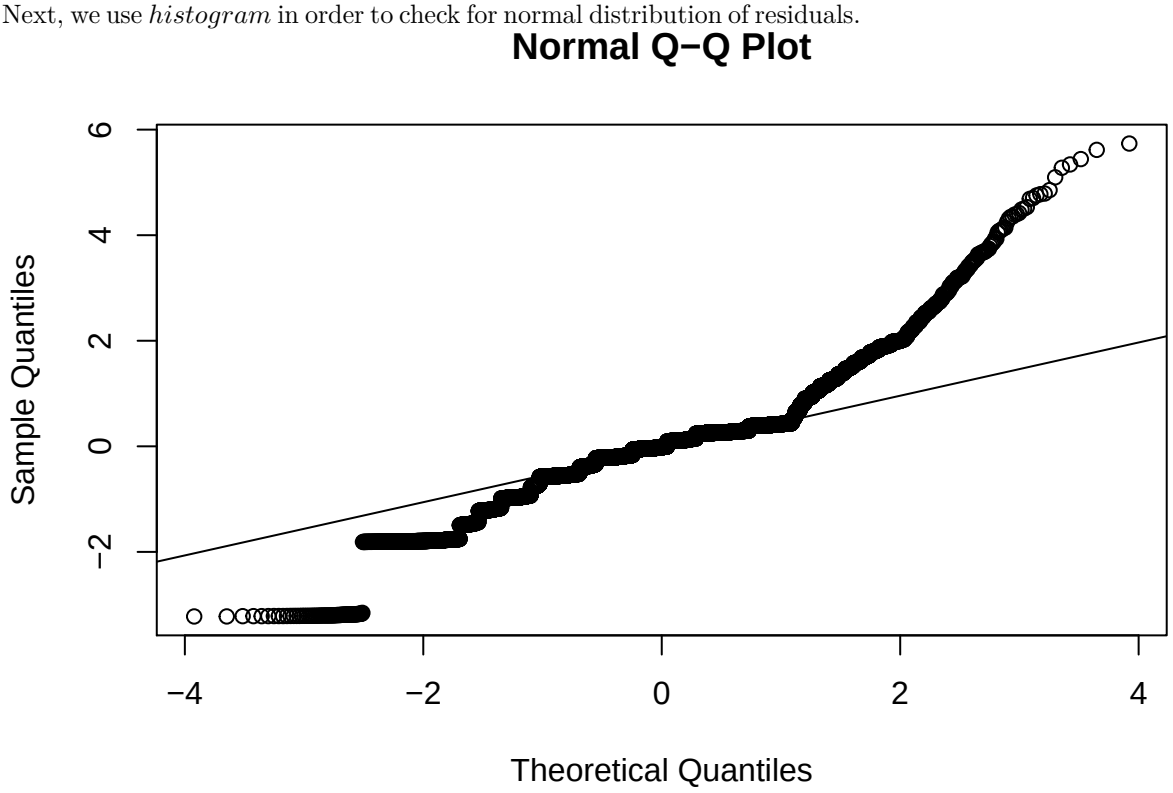
Assumptions Our first assumption was that there exists a weak linear negative relationship between Score and Median Income.



We can see that the condition of constant variability does not appear to be met fully as the values scatter unevenly above and below the horizontal line. The plot also shows a gap between 3.1 and 3.15, which might be interesting to study further.



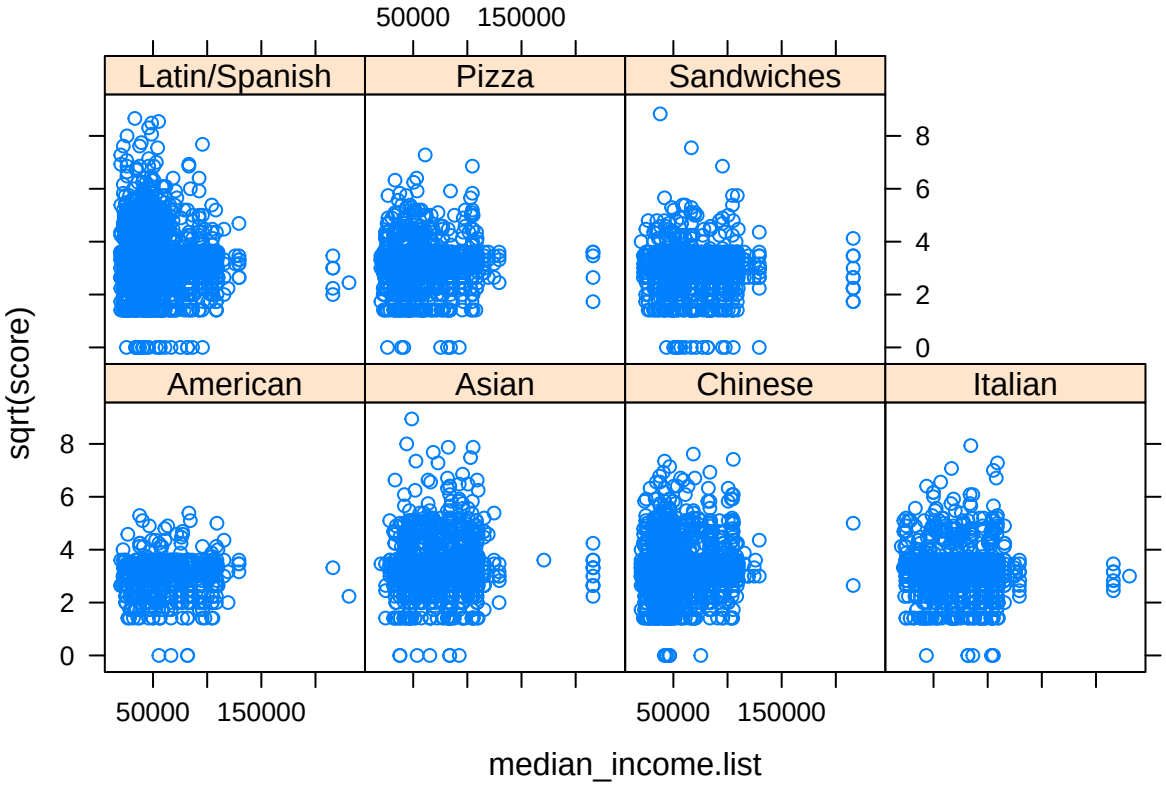
Next, we use *histogram* in order to check for normal distribution of residuals.



From the graphs above, we can see that the distribution of residuals is strongly skewed-right. Hence, our linear model might not be good enough to explain fully the relationship between median income and score.

b. Multivariable Linear Model

Since our previous finding suggested that scores seem to vary across different cuisines, we decided to include Cuisine into our linear regression model.




```
##
## Call:
## lm(formula = sqrt(score) ~ cuisine_group_factor + median_income.list,
##     data = restaurant.clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3012 -0.4203  0.0363  0.3258  5.9049
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.000e+00  4.262e-02  70.388  < 2e-16
## cuisine_group_factorAsian         3.069e-01  4.292e-02   7.151 9.13e-13
## cuisine_group_factorChinese       2.550e-01  4.137e-02   6.163 7.37e-10
## cuisine_group_factorItalian       1.927e-01  4.375e-02   4.405 1.07e-05
## cuisine_group_factorLatin/Spanish 2.863e-01  4.072e-02   7.030 2.19e-12
## cuisine_group_factorPizza         1.269e-01  4.542e-02   2.794  0.00522
## cuisine_group_factorSandwiches   -6.739e-02  4.630e-02  -1.456  0.14556
## median_income.list               -1.421e-07  3.273e-07  -0.434  0.66422
##
## (Intercept)                       ***
## cuisine_group_factorAsian         ***
## cuisine_group_factorChinese       ***
## cuisine_group_factorItalian       ***
```

```
## cuisine_group_factorLatin/Spanish ***
## cuisine_group_factorPizza         **
## cuisine_group_factorSandwiches
## median_income.list
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8782 on 11367 degrees of freedom
##   (150 observations deleted due to missingness)
## Multiple R-squared:  0.01702,    Adjusted R-squared:  0.01641
## F-statistic: 28.12 on 7 and 11367 DF,  p-value: < 2.2e-16
```

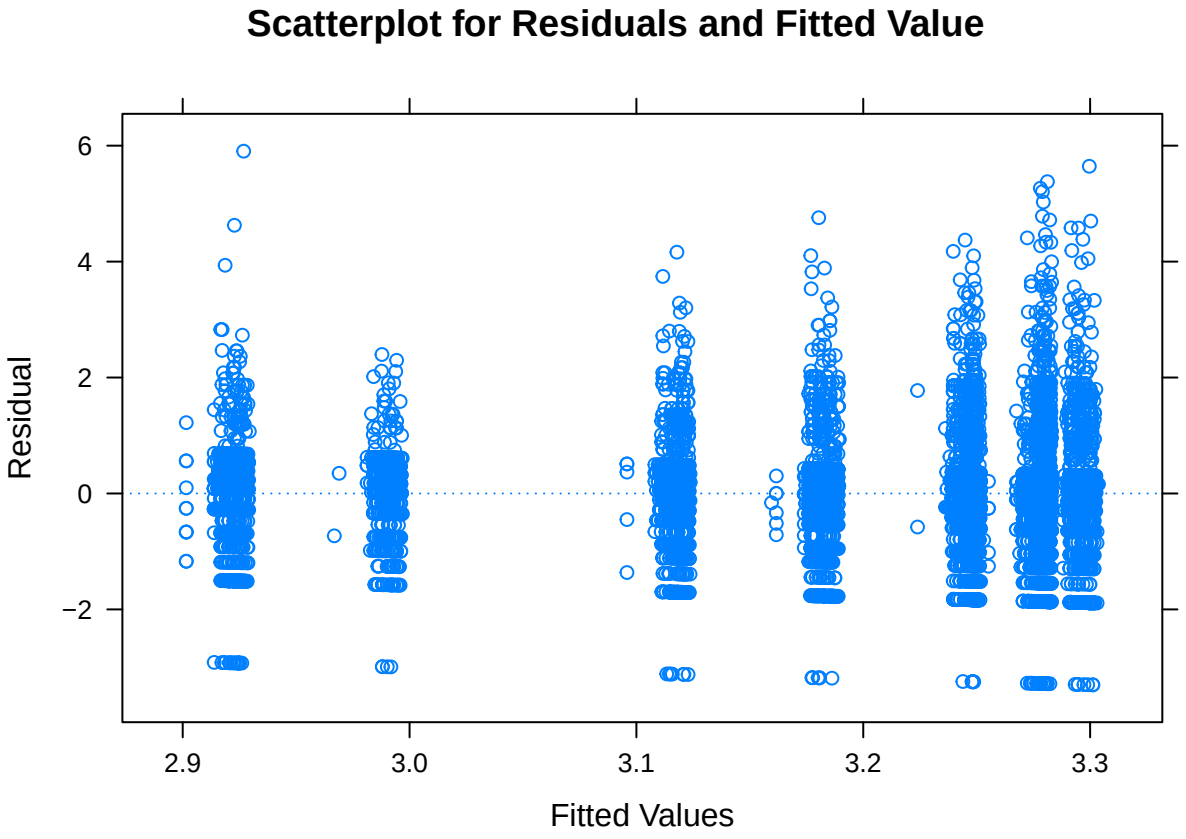
The least-squares equation line fitting the relationship between Score, Cuisine, and Median Income will be,

$$\widehat{\sqrt{Score}} = 3 + 3 \times 10^{-1} \times Asian + 2.6 \times 10^{-1} \times Chinese + 1.9 \times 10^{-1} \times Italian + 2.9 \times 10^{-1} \times Latin/Spanish +$$

$$1.3 \times 10{-}1 \times Pizza - 6.7 \times 10{-}2 \times Sandwiches - 1.4 \times 10{-}7 \times MedianIncome$$

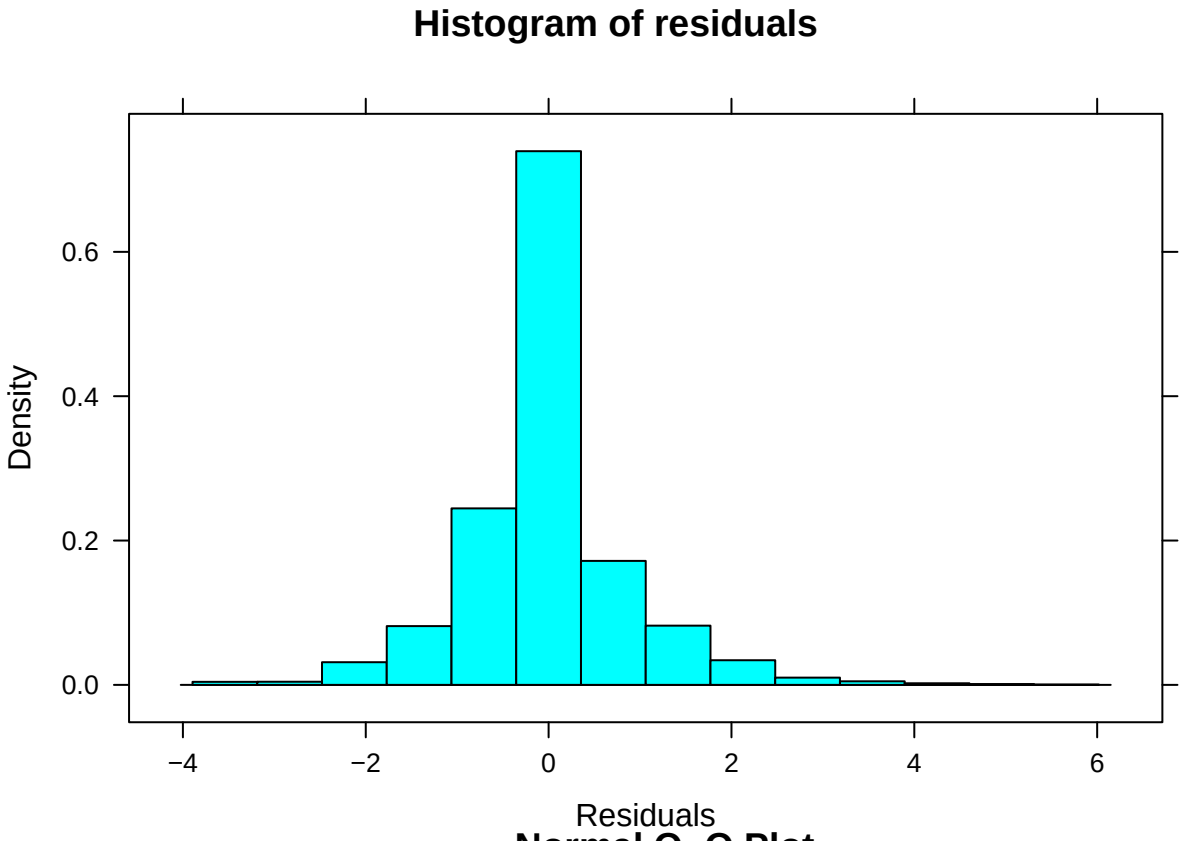
From the model above, we see that American cuisine is chosen as the reference group. Since most of the coefficients of other types of cuisine are positive, with the exception of Sandwiches, it seems that being American cuisine will give us a lower score. It is also important to point out that the p-value of Median Income now is 0.66, indicating that it is no more a significant variable, with Cuisine in the model. It seems that the significance of income on score can be explained by the fact that cleaner cuisines, such as American restaurants, are more likely to be opened in higher income zip codes. Essentially, it looks like income is positively related to score, but what seems to be the case is that income is related to cuisine and cuisine is related to score. American restaurants tend to be opened in wealthy areas and Asian restaurants are more likely to be open in lower income areas. There is not a significant difference in an Asian restaurant in a high income area compared to a low income area just a difference in Asian and American.

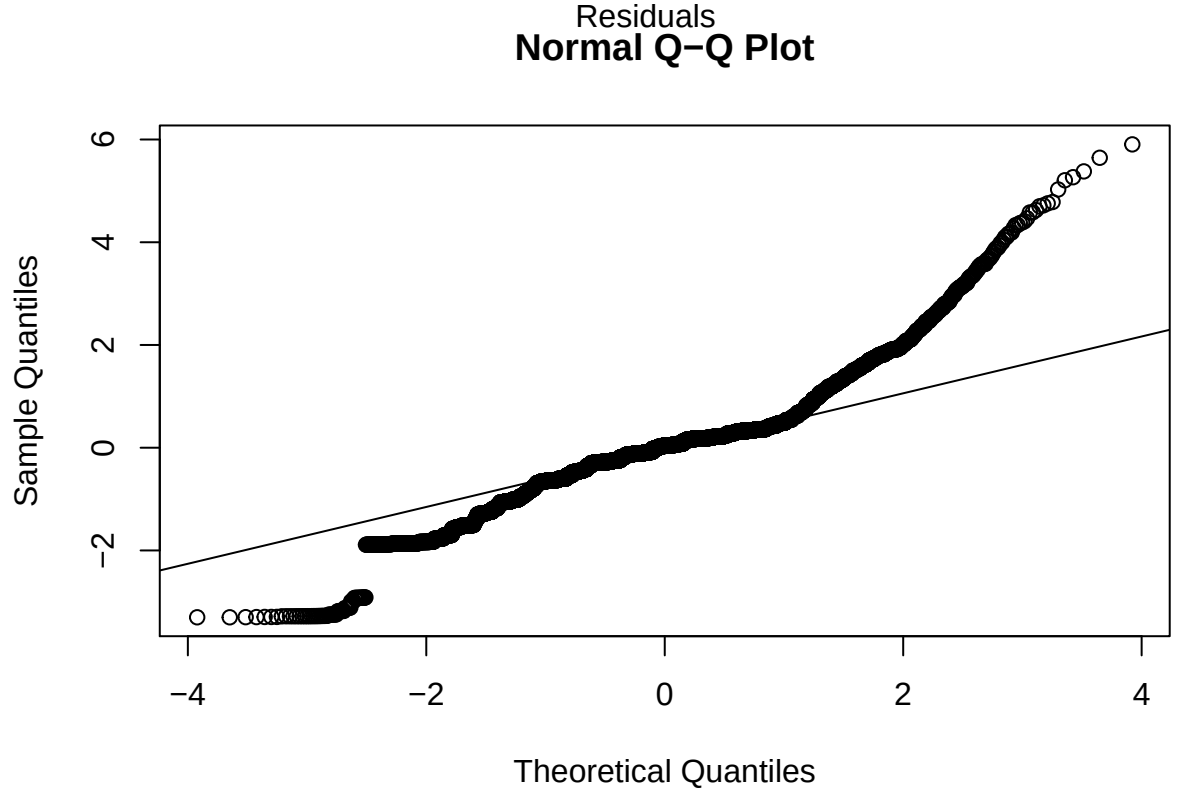
Assumptions We check for linearity with the following plot:



The plot above shows that the points are scattered more evenly around line 0 than the previous model, indicating that the condition is better met when we include cuisine into our model. However, we should note a potential fanning out pattern as fitted values get bigger.

We also check for normal distribution of residuals:



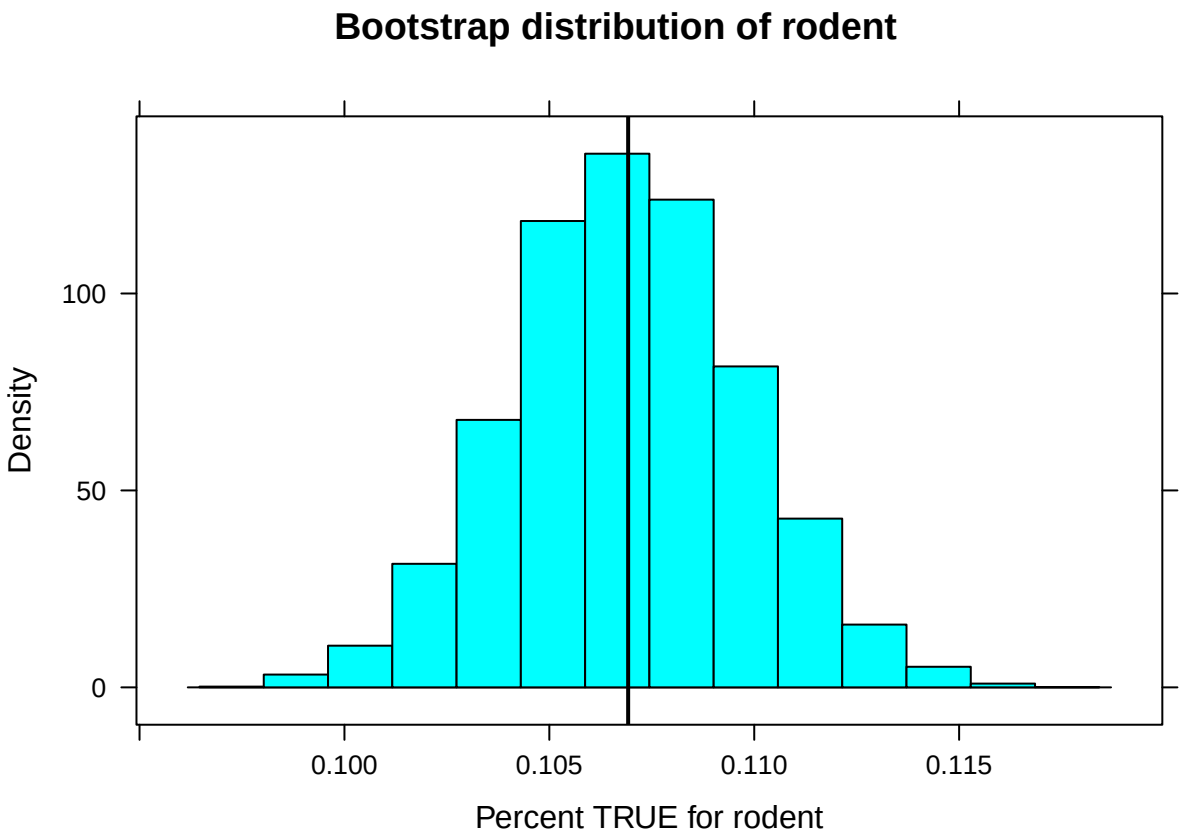


The distribution of residuals is quite similar to that of the previous model, so we have the same conclusion as before.

IV. The Pest Problem in NYC

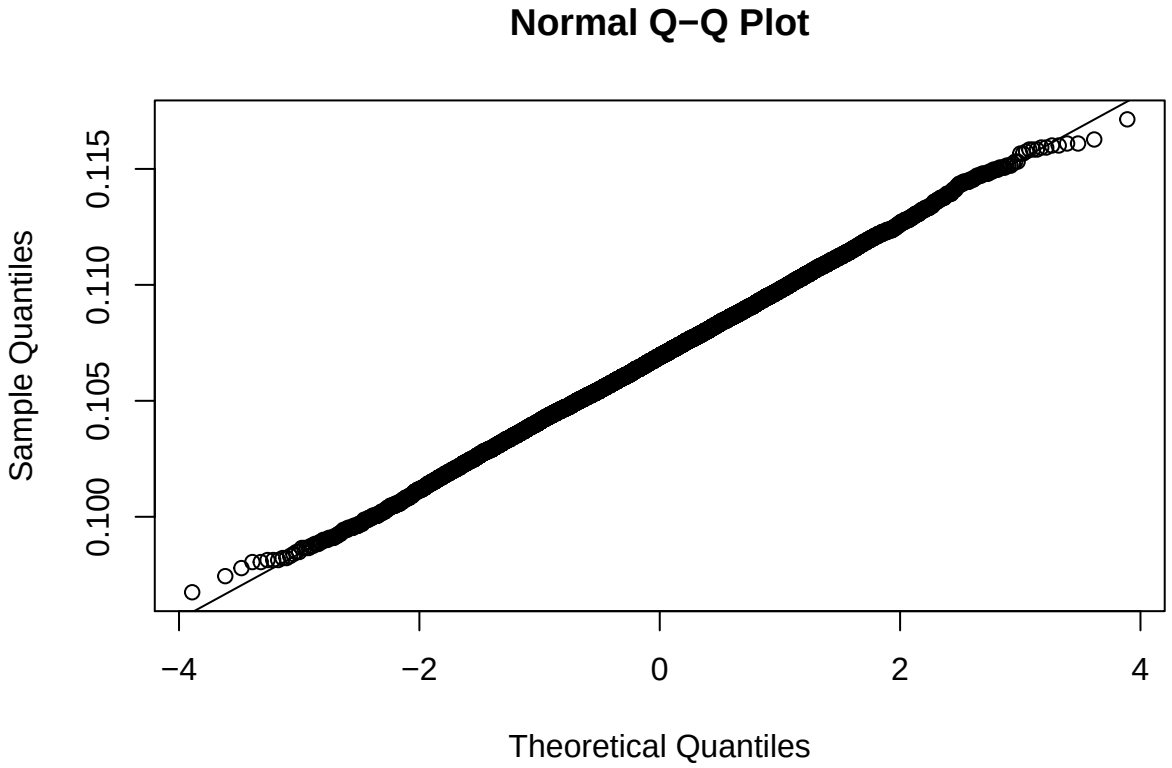
Historically, it has not been uncommon for restaurants to have pests such as rodents and roaches. We run bootstrap tests in an attempt to determine the true percentage of restaurants with rodents or roaches during their most recent inspection.

How common are rodents in NYC



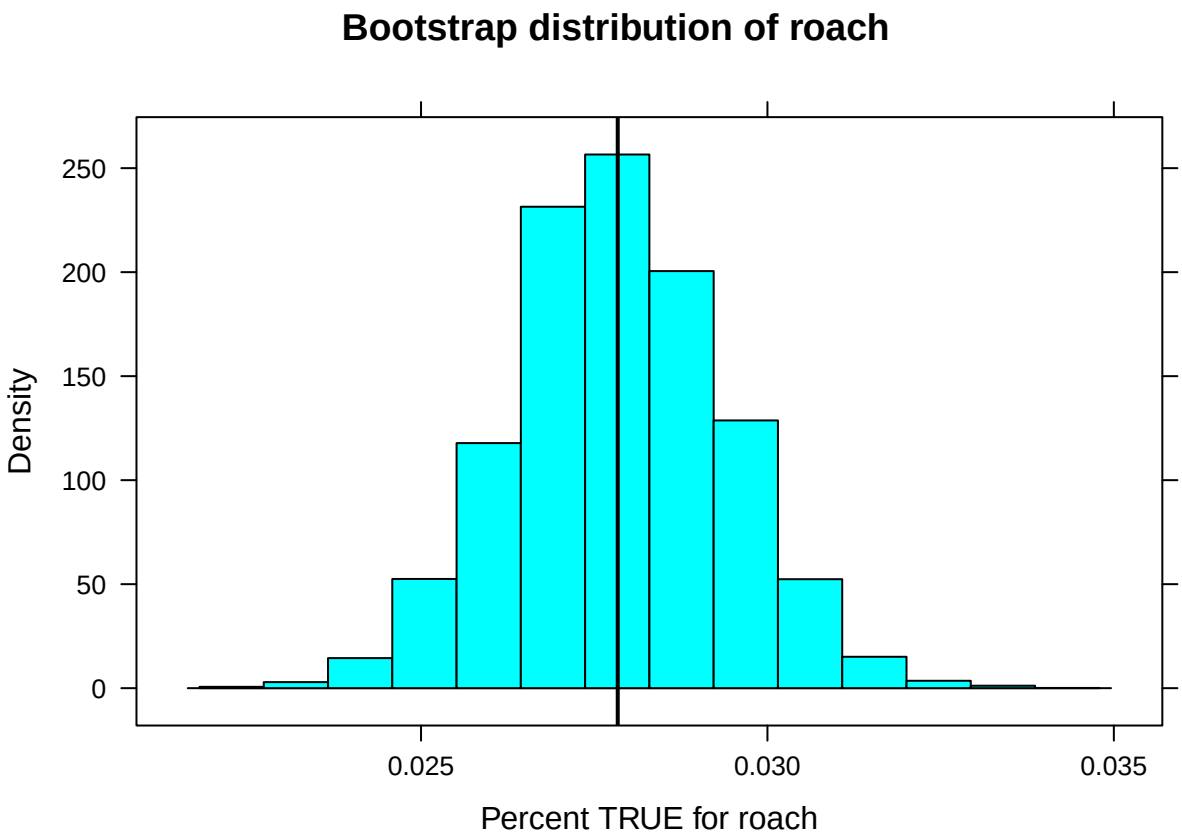

```
## [1] 0.1069237
```

```
## [1] 0.002859941
```



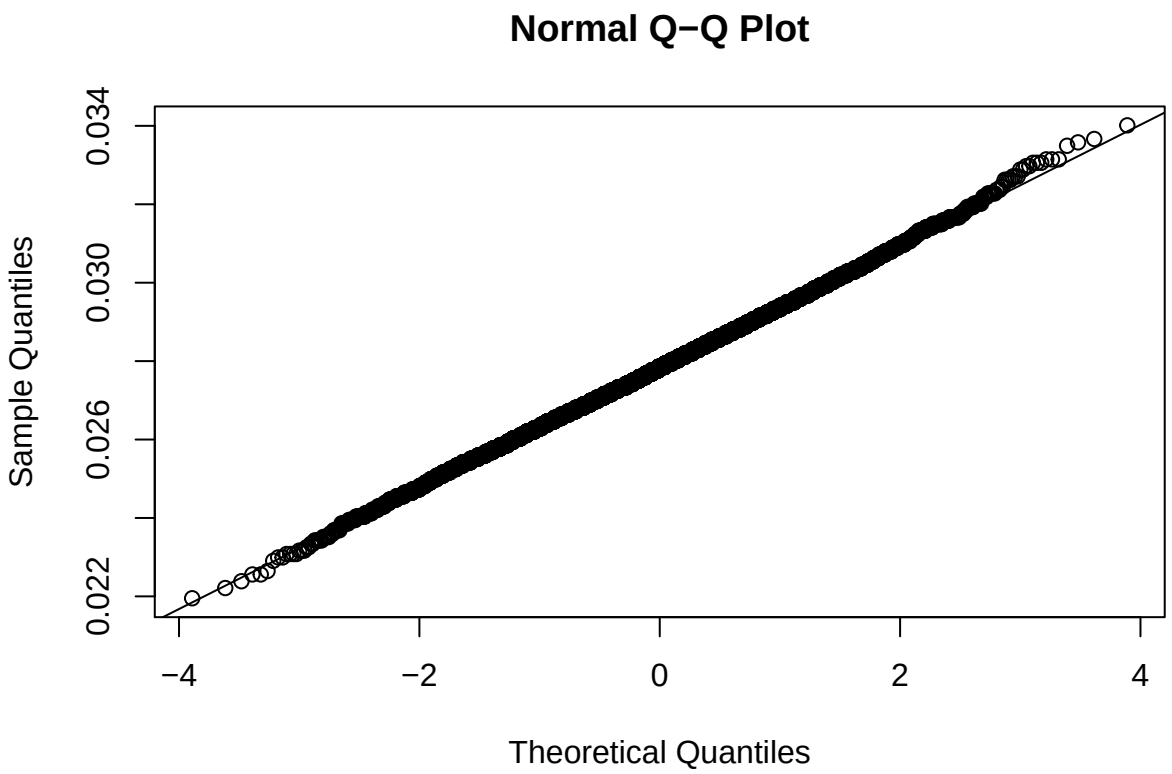
The bootstrap suggests that the true percentage of restaurants in New York City in which inspections found rodents is between 10% and 11.5%, with the most likely guess being 10.7%.

How common are roaches in NYC



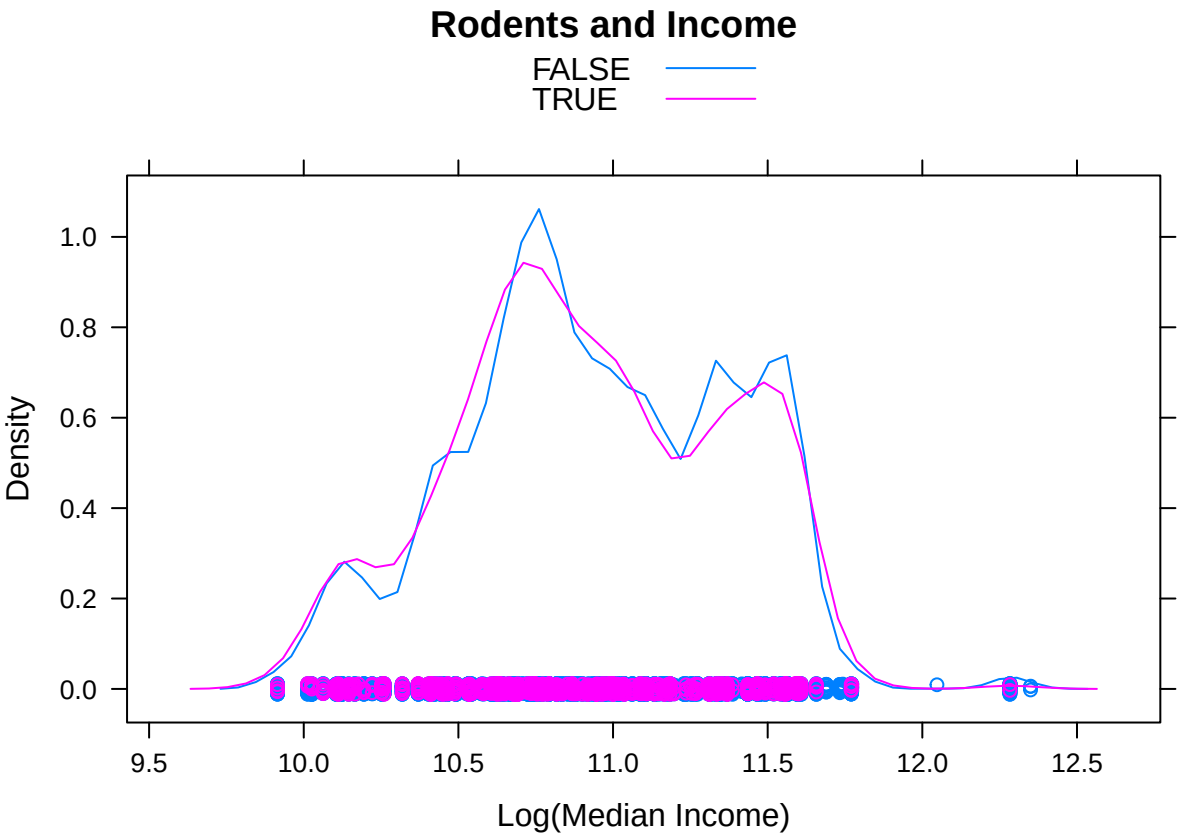

```
## [1] 0.02783931
```

```
## [1] 0.001537997
```



The bootstrap suggests that the true percentage of restaurants in New York City in which inspections found roaches is between 2.4% and 3.2%, with the most likely guess being 2.8%.

Rodents and Income



The income of the zip code that the restaurants are in does not seem to vary based upon whether rodents were found during the last inspection. This suggests that the presence of rodents in a restaurant is not dependent upon income. In other words, restaurants in wealthier areas are just as likely to have rodents as restaurants in poorer areas of the city.

Roaches and Income

The income of the zip code that the restaurants are in appears to vary based upon whether roaches were found during the last inspection. This suggests that the presence of roaches in a restaurant is dependent upon the income of that zip code. In other words, restaurants in wealthier areas are less likely to have roaches than restaurants in poorer areas of the city.

V. Areas for Improvement

While the results of this study are interesting, there are several notes that should be made. First, the data on restaurants comes entirely from New York City. While many of these results may apply to other cities, New York City is unique and as such, it is difficult to generalize. Secondly, our results are contingent to a certain extent on the accuracy of the restaurant inspections. If these inspections are biased against particular cuisines or geographic locations, obviously the results of this paper may also be inaccurate. Thirdly, we only considered a handful of cuisines and the categorization was somewhat arbitrary. It is possible that other, less popular cuisines are actually significantly worse than those included in this paper. Similarly, a different categorization may yield different results. A more thorough analysis of the subject could shed light on such details.

VI. Potentials for Future Projects

There are several potential continuations to this project. Firstly, it is interesting to see roaches being correlated to income. This brings up the question of whether income disparity could be linked with presence of roaches. Currently, there have been studies between housing disrepair condition and pest infestation. From the public health standpoint, cockroach allergens have been studied for association with prevalence of childhood asthma and thus knowing the above relationship could help answer more public health questions. Also, while we consider a handful of cuisines and the most recent inspection, future projects could lift these restrictions and obtain a more complete picture.

VII. Conclusion

In this paper we used restaurant inspection scores in New York City to answer two questions: Are Chinese restaurants less sanitary than other types restaurants and how bad is the pest problem in New York City? We found that while Chinese restaurants are worse than the average restaurant in New York City, Chinese restaurants are not the worst. In fact, Asian restaurants appeared to perform worse than Chinese restaurants. Interestingly, income does not seem to have a significant affect on the score of a restaurant after we account for type of cuisine. Furthermore, we found that rodents and roaches are a significant problem in New York City Restaurants, being found in 10% and 3% of restaurants respectively. Curiously, while roaches are more likely to be found in lower income zip codes, the presence of roaches does not appear to vary based upon a region's income.