Digging In on Data: Identifying At-risk Restaurants Using NYC DOHMH Data

By: SB, Ted Stetzel, AS, CT
STAT9660:Final Project, Baruch College

I. Executive Summary

The restaurant business represents a large portion of the New York City economy as both the city's 8.5 million residents and the 47 million tourists who visit the city each year frequent the city's +22,000 restaurants. The objective of this study is to help the DOHMH take a more proactive approach to restraint inspections by creating a predictive model that utilizes available data to identify at-risk restaurants that the agency can work with to prevent future complaints and closures. The potential benefits of this model include reduction of incident reporting to the DOHMH, increased tax revenue due to closure prevention, greater diversification of restaurants in New York City and a reduction in the DOHMH's operating budget.

Targeting each restaurant's previous score, various models including a Decision Tree, Regression Tree, PCA and Liner Regression models were created. All these models are used to analyze the dataset, help to interpret each factor. In conclusion it was determined that a regression based built using stepwise selection using a combination of Principle Component Variables to describe variability in the numerous violation types and other non-PCA variables was the best prediction model.

II. Introduction

Dining out is a significant part of most people's lives whether it is a special occasion or it is a daily occurrence. Depending on the restaurant, you may or may not be able to see how the food is prepared and have to go on good faith that the restaurant is preparing your food in a safe and sanitary way. Despite a large amount of trust being put into the people who prepare food and other preventative measures that have been put in place, problems still arise. Going out to eat and

finding something unexpected in your food or getting sick afterwards can be an uncomfortable experience. Unsanitary conditions of restaurants can also cost city governments in terms of lost revenue from sales taxes, payroll taxes, and other tourism dollars.

This is especially important in New York City where dining out is big business. New Yorkers have the highest average spend for diner[i] and over 47 million tourist visit the city every year[ii] so it should come as no surprise the city has a very rigorous restraint inspection system to protect its residents and visitors. Since 2010 the DOHMH has been assigning grades to the approximately 24,000 restaurants across all five boroughs. An A is top rating; this would require a restaurant to have 13 or less violation points. If a restaurant was cited with 14-27 violation points they would be rated a B, and if a restaurant has more than 28 violation points they would earn a C. If a severe instances it may be necessary for the DOHMH to shut down a restaurant if the violations are severe enough assuming they can't be quickly remedied in time for reasonable reopening. If you've ever walked around NYC, you'll likely notice restaurant will have their posting near the front door showing their latest score, which restaurants are required to post their score within feet of their main entrance as per the department of health guidelines. By design, the posting guideline keeps all restaurant owners honest and accountable of their inspection since some dimers would likely be hesitant to dine anywhere did not receive an A.

III. Methodology

Given that the New York City Department of Health and Mental Hygiene (DOHMH) has a small portion of its staff of 6,000 dedicated to restaurant inspections, the agency has to take a reactive stance and rely heavily on complaints from residents to be alerted to violations of the sanitary code.[iii] Unfortunately by the time a complaint is filed, a resident may already have become ill or injured as a result of their experience.

While this grading system encourage restaurants to meet the DOHMH's requirements they can have a negative impact such lost tax revenue for the city or having property sit idle because a restaurant has been shut down a business for reasons that could have been prevented. Our study seeks to create a model that will help the DOHMH identify at-risk restaurants in order to provide these restaurants with extra assistance in order to prevent them from failing or scoring low in an inspection. The questions our study seeks to answer are:

- Are there a few benchmark violations that cause most restaurants to fail?
- Is location or cuisine type indicative of a passing or failing score?
- Does the average income of your patronage affect your likelihood to pass?

IV. Data Preparation

The majority of the data was collected from NYC Open Data website. The New York Department of Health and Mental Hygiene (DOHMH) has made available a regularly updated list of New York City restaurant inspection results since the restaurant grading program started in July 2010. The data set is organized by events (such as inspections, re-openings and violations) so instead of being organized by restaurant, each of these +500,000 events is represented a row in the dataset. Since our study seeks to predict grades on a per restaurant basis, the data needed to be reorganized so that each row represents a restaurant and violations are tracked as separate columns associated with the restaurant in that row. To achieve this configuration we imported the event based data set in the form of a CSV file from NYC OpenData into Microsoft excel and created a pivot table.

Once the data was in a pivot table, it was first organized so that each restaurant unique ID (CAMIS) was a row and new columns were created for each of the violation types (a full list of

all violations are provided in Appendix 1). Our group also wanted to investigate the effect that population and average income might have on restaurant grades. To do so we imported data made available by the University of Michigan's Population Studies Center that listed the median household income between the years 2006-2010 for every U.S. zip code

V. Data Exploration

Before building any models to find out which variables contribute the most to restaurant scores, an examination of the data was performed using SAS Enterprise Miner. First, the location of each restaurant was examined on a per-borough basis. Not surprisingly we found that the majority of points in our dataset are located in Manhattan followed by Queens and Brooklyn. This is expected since being the commercial hub of New York City, Manhattan has the highest concentration of restaurants. As for the cuisine description majority of the restaurants were classified themselves as American Restaurant followed by Chinese and Pizza. It would be interesting to explore the idea of how a cuisine type determines the food grade since factors such as ingredients and preparation procedures differ by cuisine.

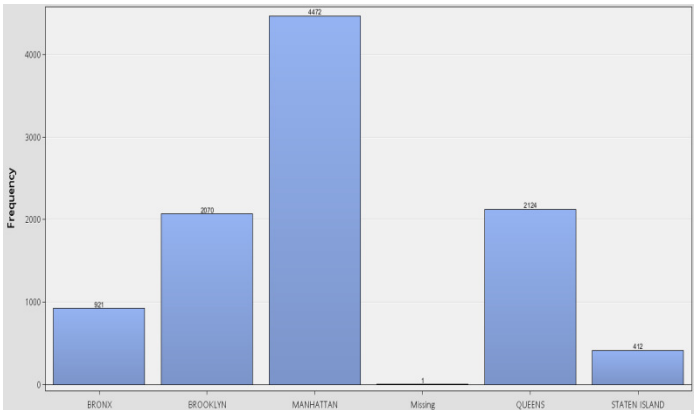


Figure 1: Distribution of Restaurants in NYC by Borough

An inspection of previously assigned grade shows that the majority of the restaurants in the received a grade of A. This imbalance means that we may have to sample from the population of restaurants in order to have a more equal balance with our target, at-risk restaurants.

Delving into the variables which have the greatest impact on predicting our target variable we see that the four most important variables in order of importance are as follows:

- Last Inspection Date,
- Variable 08_A (Facility not vermin proof)
- Variable 02_G (Cold item held above 41° F)
- Number of Inspections, Variable 02_B
- (hot item not being held above 140 ° F)

Figure 2 shows the relation of the lastscore with variable _08A (Facility not vermin proof) which according to our initial exploration has the highest explanatory power.

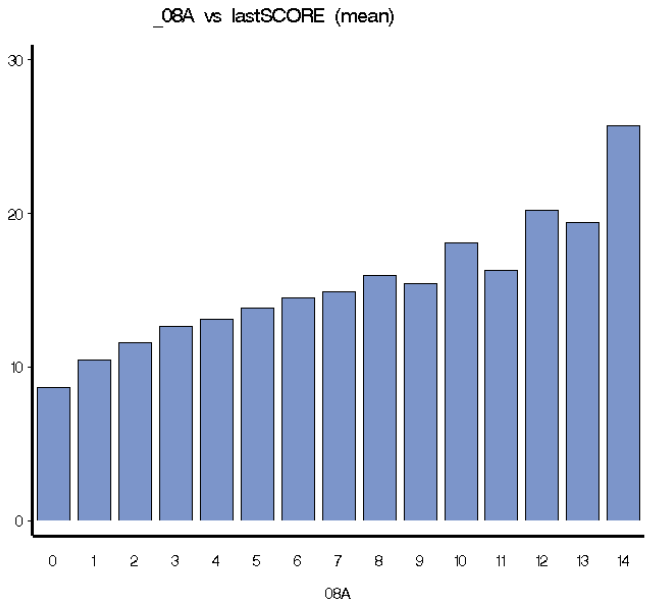


Figure 2: Occurrence of Vermin Violations

It is interesting to note that in addition to the three violations which are expected to impact the score which eventually translates into the restaurant grade there are two more explanatory variables. These include the Last Inspection Date and Number of Inspections. At this stage, the potential importance of these two variables can be explained in the sense that once a restaurant has had a few inspections it becomes familiar with the evaluation criteria and better able to handle the inspection. Moreover when the last inspection date is recent a restaurant can gauge easily the latest patterns the inspectors are looking for. These initial patterns combined with our modelling results will better enable us to make predictions about the way the grading system works in New York.

One interesting aspect of our data is that the income measures of the neighborhood do not play a very significant role in explaining the last scores that are received by the different restaurants in consideration. Normally we would expect that restaurants located in a higher mean income area would be of higher quality. We believe this can be explained from the fact that the grading system itself is keeping a check on the different restaurants in New York City. All restaurant owners are putting in substantial effort to get a decent grade regardless of the fact whether they are located in a rich neighborhood or not. Moreover the grading criteria are based on basic measures which can be easily employed by any food facility without intensive investment.

VI. Data Analysis

Decision/Regression Trees

In order distinctly bucket restaurants into an at-risk category we first tried decisions trees as a prediction model. Since our target variable last score is a continuous variable, a new binary variable named AtRisk was added to the model. All restaurants with a grade of 26 or higher were marked as AtRisk since they were a minor violation of being marked with a C grade. Additionally the full list of 22,000+ was sampled so that an equal proportion of A, B and C graded restaurant were evaluated into the model (654 per letter). Using a 60/40 training/validation a decision tree was built using the new AtRisk variable as the target. It should be noted that average prior score was removed from the model since it dominated the majority of the nodes.

The decision tree found that the most important variable was the presence of a single 05D (Hand washing facility not provided in or near food preparation area and toilet room) violation as the most important variable for determining if a restraint is at-risk. If a restaurant had at least one

prior 05D and a 09C (Food contact surface not properly maintained), it was much more likely to be at-risk than those restaurants that haven't had those two previous violations.

Taking a non-binary approach to the target variable provided very different results. Using a regression tree and targeting the last score variable showed that a restaurant with 2 or more 08A violation (facility not vermin proof) averaged a score of 17 compared to 24 to those that had one or fewer vermin violations. If a restaurant did not have a vermin violation, then the most important violation from the decision tree, 05D (Hand washing facility not provided) which dropped the average score by 5 points when present. Given that the overall the regression model performed better in terms of over fitting and the binary model did not have a vermin related violations highly ranked tree model, the linear regression model seems to be a better fit.

Though a decision tree model provided a lot of predictive information, one of the most important factors, average_score, dominated the model and could not be interpreted well in decision tree. Since the relationship between average_score and the last_score (target) is linear in nature, a linear regression model is more appropriate as a predicting model.

Regression Analysis Introduction/Dimension Reduction

Given that our liner approach to decision tree proved promising we decided to analye the data using linear regression. Before a regression model could be created, some form of dimension reduction was required since the original data set contained over 100 prediction variables; so principle component analysis was utilized. Following PCA a stepwise variable selection was utilized to build a regression model.

Only the violation variables were used to create the PCA variables and variables such as cuisine type and borough were not included and lastScore was set as the target. We first set 0.8 of cumulative variance explained as the cut-off value for number of PCA variables to include however this required the inclusion of 68 principle components so that threshold was scrapped.

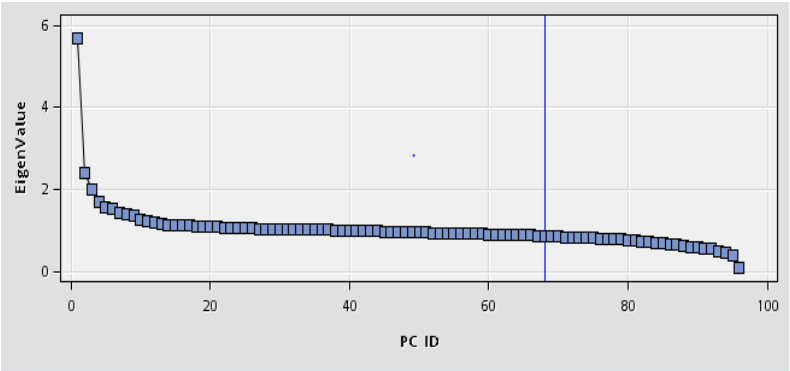


Figure 3: Cumulative variance explained for PCA eigenvalues

We then examined the eigenvalue plot (Figure 3) to visually inspect for the point when adding additional variables became meaningless. As expected, first few PCs had the greatest cumulative gains, and as others were added additional gains became very small. We concluded to include PC1, PC2, PC3, PC4, and PC5 to the new dataset. An examination of these variables showed the the following variables contributed highly to each of the chose PCAs:

PC	ViolationCode	Description
PC1	08A	Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.
PC2	15K	Operator failed to make good faith effort to inform smokers of the Smoke-free Act prohibition of smoking.
PC3	10K	Immersion basket not provided, used or of incorrect size.
PC4	06G	HACCP plan not approved or approved HACCP plan not maintained on premises.
PC5	15K	Operator failed to make good faith effort to inform smokers of the Smoke-free Act prohibition of smoking.

We extracted the newly created PCA variables and then used Microsoft Excel's "vlookup" function to combine these with non-PCA variables such as borough and cuisine type. In the end, we created a dataset containing PC1 to PC5, and all the other variables as seen in the chart to the right.

Name	Role	Level
CUISINEDESCRI	Classificatior	Nominal
DBA	Classificatior	Nominal
PC4	Input	Interval
PC5	Input	Interval
PC3	Input	Interval
ZipPop	Input	Interval
ZipMedIncome	Input	Interval
ZipMeanIncome	Input	Interval
numOfInspectior	Input	Interval
average_score	Input	Interval
BORO	Input	Nominal
PC1	Input	Interval
PC2	Input	Interval

Figure 4: List of Variables entered into stepwise regression

Stepwise Regression

To build the final model we used Stepwise regression (with a 60/40 training/validation sample). Using the results of the stepwise selection process to we could identify and exclude any variables those make little contribution to predict the target variable automatically. We set the last_score as the target, used the variables below as input or classifications, and rejected all the others.

We did not include type of cuisine as an input since there are too many types to provide an productive lift, Therefore, we only include one categorical factor-BORO (borough), which was one of the most import factors based on the results of data exploration and decision tree.

According to the output, average_score was entered first entered into the model, which matched showed that restaurants that have gotten A's in the past will continue to get A's. Number of inspections was the second variable to be included into the model and then PC1, PC2, PC5, BORO, and zip population entered the model consecutively. Instead of using all 11 varibles, the stepwise profess helped us filter cut the number down to 7 variables: BORO PC1 PC2 PC5 ZipPop average_score numOfInspections.

However, the estimate beta of ZipPop is almost 0 so we deleted this variable from our final model (Appendix 4 displays some important statistics and estimates). Based on result, we have the predictive equation below:

Predicted _Score	=	if BORO=BRONX	4.21+0.41+0.57*PC1+0.67*PC2+ 0.56*PC5+0.56*average_score-0.38*numberOfInspections
		if BORO=BROOKLYN	4.21-4.33+0.57*PC1+0.67*PC2+ 0.56*PC5+0.56*average_score-0.38*numberOfInspections
		if BORO=MANHATON	4.21-0.14+0.57*PC1+0.67*PC2+ 0.56*PC5+0.56*average_score-0.38*numberOfInspections
		if BORO=QUEENS	4.21-0.277+0.57*PC1+0.67*PC2+ 0.56*PC5+0.56*average_score-0.38*numberOfInspections
		if BORO=STATEN ISLAND	4.21+0+0.57*PC1+0.67*PC2+ 0.56*PC5+0.56*average_score-0.38*numberOfInspections
		if BORO is missing	4.21-0.32+0.57*PC1+0.67*PC2+ 0.56*PC5+0.56*average_score-0.38*numberOfInspections

The R-square of the model is 0.231, and the validate average square error is 39.54 which is a little above the train average square error. Also, the plot in the Appendix 4 shows that there is no over-fitting within the model.

VII. Conclusion

We considered all the analysis results from data exploration, decision tree, and principle components analysis, combing all the important factors to a new dataset (Grades3). Eventually, we ran the stepwise regression, reaching out to a predicting equation of the Score. R-square and MSE are decent of our model, which is a proof of a good predictive model.

The variables that contributed most to the model are ones that could be controlled by the restaurant. Factors such as medium income per zipcode provided no predictive insight. Instead prior violations played a critical part.

If this study were being performed with additional resources and proper funding it would be interesting to include additional data that would enable a more effective model. A few potential factors that would be interesting to delve into would be:

- How do being a Franchise, small chain or mom & pop compare to one another?

- Explore the relationship of other factors not available to during this study such as food price, outdoor seating, presence of a reservation system, and hours of operation.
- Is a good yelp rating indicative of a restaurant grade?
- Since vermin control is a powerful prediction factor, could we use information about upcoming construction in close proximity to a restaurant to predict grade?

Appendix 1: List of Violation Codes

Code	Violation Description
02A	Food not cooked to required minimum temperature.
02B	Hot food item not held at or above 140º F.
02C	Hot food item that has been cooked and refrigerated is being held for service without first being reheated to 1 65º F or above within 2 hours.
02D	Precooked potentially hazardous food from commercial food processing establishment that is supposed to be heated, but is not heated to 140º F within 2 hours.
02E	Whole frozen poultry or poultry breasts, other than a single portion, is being cooked frozen or partially thawed.
02F	Meat, fish or molluscan shellfish served raw or undercooked without prior notification to customer.
02G	Cold food item held above 41º F (smoked fish and reduced oxygen packaged foods above 38 ºF) except during necessary preparation.
02H	Food not cooled by an approved method whereby the internal product temperature is reduced from 140º F to 70º F or less within 2 hours, and from 70º F to 41º F or less within 4 additional hours.
02I	Food prepared from ingredients at ambient temperature not cooled to 41º F or below within 4 hours.
02J	Reduced oxygen packaged (ROP) foods not cooled by an approved method whereby the internal food temperature is reduced to 38º F within two hours of cooking and if necessary further cooled to a temperature of 34º F within six hours of reaching 38º F.
03A	Food from unapproved or unknown source or home canned. Reduced oxygen packaged (ROP) fish not frozen before processing; or ROP foods prepared on premises transported to another site.
03B	Shellfish not from approved source, improperly tagged/labeled; tags not retained for 90 days.
03C	Eggs found dirty/cracked; liquid, frozen or powdered eggs not pasteurized.
03D	Canned food product observed swollen, leaking or rusted, and not segregated from other consumable food items .
03E	Potable water supply inadequate. Water or ice not potable or from unapproved source. Cross connection in potable water supply system observed.
03F	Unpasteurized milk or milk product present.
03G	Raw food not properly washed prior to serving.
04A	Food Protection Certificate not held by supervisor of food operations.
04B	Food worker prepares food or handles utensil when ill with a disease transmissible by food, or have exposed infected cut or burn on hand.
04C	Food worker does not use proper utensil to eliminate bare hand contact with food that will not receive adequate additional heat treatment.
04D	Food worker does not wash hands thoroughly after using the toilet, coughing, sneezing, smoking, eating, preparing raw foods or otherwise contaminating hands.
04E	Toxic chemical improperly labeled, stored or used such that food contamination may occur.
04F	Food, food preparation area, food storage area, area used by employees or patrons, contaminated by sewage or liquid waste.
04G	Unprotected potentially hazardous food re-served.
04H	Raw, cooked or prepared food is adulterated, contaminated, cross-contaminated, or not discarded in accordance with HACCP plan.
04I	Food item spoiled, adulterated, contaminated or cross-contaminated. Unprotected food re-served.
04J	Appropriately scaled metal stem-type thermometer or thermocouple not provided or used to evaluate temperatures of potentially hazardous foods during cooking, cooling, reheating and holding.

04K	Evidence of rats or live rats present in facility's food and/or non-food areas.
04L	Evidence of mice or live mice present in facility's food and/or non-food areas.
04M	Live roaches present in facility's food and/or non-food areas.
04N	Filth flies or food/refuse/sewage-associated (FRSA) flies present in facilitys food and/or non-food areas. Filth flies include house flies, little house flies, blow flies, bottle flies and flesh flies. Food/refuse/sewage-associated flies include fruit flies, drain flies and Phorid flies.
04O	Live animals other than fish in tank or service animal present in facility's food and/or non-food areas.
05A	Sewage disposal system improper or unapproved.
05B	Harmful, noxious gas or vapor detected. CO ~1 3 ppm.
05C	Food contact surface improperly constructed or located. Unacceptable material used.
05D	Hand washing facility not provided in or near food preparation area and toilet room. Hot and cold running water at adequate pressure to enable cleanliness of employees not provided at facility. Soap and an acceptable hand-drying device not provided.
05E	Toilet facility not provided for employees or for patrons when required.
05F	Insufficient or no refrigerated or hot holding equipment to keep potentially hazardous foods at required temperatures.
05H	No facilities available to wash, rinse and sanitize utensils and/or equipment.
05I	Refrigeration used to implement HACCP plan not equipped with an electronic system that continuously monitors time and temperature.
06A	Personal cleanliness inadequate. Outer garment soiled with possible contaminant. Effective hair restraint not worn in an area where food is prepared.
06B	Tobacco use, eating, or drinking from open container in food preparation, food storage or dishwashing area observed.
06C	Food not protected from potential source of contamination during storage, preparation, transportation, display or service.
06D	Food contact surface not properly washed, rinsed and sanitized after each use and following any activity when contamination may have occurred.
06E	Sanitized equipment or utensil, including in-use food dispensing utensil, improperly used or stored.
06F	Wiping cloths soiled or not stored in sanitizing solution.
06G	HACCP plan not approved or approved HACCP plan not maintained on premises.
06H	Records and logs not maintained to demonstrate that HACCP plan has been properly implemented.
06I	Food not labeled in accordance with HACCP plan.
07A	Duties of an officer of the Department interfered with or obstructed.
08A	Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.
08B	Covered garbage receptacle not provided or inadequate, except that garbage receptacle may be uncovered during active use. Garbage storage area not properly constructed or maintained; grinder or compactor dirty.
08C	Pesticide use not in accordance with label or applicable laws. Prohibited chemical used/stored. Open bait station used.
09A	Canned food product observed dented and not segregated from other consumable food items.
09B	Milk or milk product undated, improperly dated or expired or Thawing procedures improper.
09C	Food contact surface not properly maintained.
10A	Toilet facility not maintained and provided with toilet paper, waste receptacle and self-closing door.

10B	Plumbing not properly installed or maintained; anti-siphonage or backflow prevention device not provided where required; equipment or floor not properly drained; sewage disposal system in disrepair or not functioning properly.
10C	Lighting inadequate; permanent lighting not provided in food preparation areas, ware washing areas, and storage rooms.
10D	Mechanical or natural ventilation system not provided, improperly installed, in disrepair and/or fails to prevent excessive build-up of grease, heat, steam condensation vapors, odors, smoke, and fumes.
10E	Accurate thermometer not provided in refrigerated or hot holding equipment.
10F	Non-food contact surface improperly constructed. Unacceptable material used. Non-food contact surface or equipment improperly maintained and/or not properly sealed, raised, spaced or movable to allow accessibility for cleaning on all sides, above and underneath the unit.
10G	Food service operation occurring in room used as living or sleeping quarters.
10H	Proper sanitization not provided for utensil ware washing operation.
10I	Single service item reused, improperly stored, dispensed; not used when required.
10J	""Wash handssign not posted at hand wash facility.
10K	Immersion basket not provided, used or of incorrect size. Incorrect manual technique. Test kit and thermometer not provided or used. Improper drying practices.
15E	Out-of package sale of tobacco products observed.
15H	Sign prohibiting sale of tobacco products to minors not conspicuously posted.
15I	No Smoking and/or 'Smoking Permitted sign not conspicuously posted. Health warning not present on 'Smoking Permitted
15J	Ashtray present in smoke-free area.
15K	Operator failed to make good faith effort to inform smokers of the Smoke-free Act prohibition of smoking.
15L	Smoke free workplace smoking policy inadequate, not posted, not provided to employees.
15S	Flavored tobacco products sold or offered for sale.
15T	Original label for tobacco products sold or offered for sale.
16A	A food containing artificial trans fat, with 0.5 grams or more of trans fat per serving, is being stored, distributed, held for service, used in preparation of a menu item, or served.
16B	The original nutritional fact labels and/or ingredient label for a cooking oil, shortening or margarine or food item sold in bulk, or acceptable manufacturers documentation not maintained on site.
16C	Caloric content not posted on menus, menu boards or food tags, in a food service establishment that is 1 of 15 or more outlets operating the same type of business nationally under common ownership or control, or as a franchise or doing business under the same name, for each menu item that is served in portions, the size and content of which are standardized.
16E	Caloric content range (minimum to maximum) not posted on menus and or menu boards for each flavor, variety and size of each menu item that is offered for sale in different flavors, varieties and sizes.
16F	Specific caloric content or range thereof not posted on menus, menu boards or food tags for each menu item offered as a combination meal with multiple options that are listed as single items.
18B	Document issued by the Board of Health, Commissioner or Department unlawfully reproduced or altered.
18C	Notice of the Department of Board of Health mutilated, obstructed, or removed.
18D	Failure to comply with an Order of the Board of Health, Commissioner, or Department.
18F	Permit not conspicuously displayed.
18G	Manufacture of frozen dessert not authorized on Food Service Establishment permit.
18I	Choking first aid poster not posted. Alcohol and Pregnancy warning sign, inspection report sign; not posted. CPR sign not posted, equipment (resuscitation masks, adult & pediatric, latex gloves) not provided.

20A	Food allergy information poster not conspicuously posted where food is being prepared or processed by food workers.
20B	Food allergy information poster not posted in language understood by all food workers.
20D	Choking first aid poster not posted. Alcohol and pregnancy warning sign not posted. Resuscitation equipment: exhaled air resuscitation masks (adult & pediatric), latex gloves, sign not posted. Inspection report sign not posted.
20E	Letter Grade or Grade Pending card not conspicuously posted and visible to passersby.
20F	Current letter grade card not posted.
22A	Nuisance created or allowed to exist. Facility not free from unsafe, hazardous, offensive or annoying conditions.
22B	Toilet facility used by women does not have at least one covered garbage receptacle.
22C	Bulb not shielded or shatterproof, in areas where there is extreme heat, temperature changes, or where accidental contact may occur.
22E	ROP processing equipment not approved by DOHMH.
99B	Other general violation.

We utilized the count function to record the number of times that each violation occurred at each restaurant. Additionally, we also imported several other fields from this set such as zip, borough, and cuisine type. Also, the table and remove duplicate values functions in excel were utilized to sort and reduce the data by most recent date to get each restaurant's most recent inspection date and our target variable, most recent score. It was also at this point that restaurants that did not have an inspection date more recent than January 2013 were removed from the data set. Finally phone number, street and address were removed from the dataset at this point because they possess no predictive information.

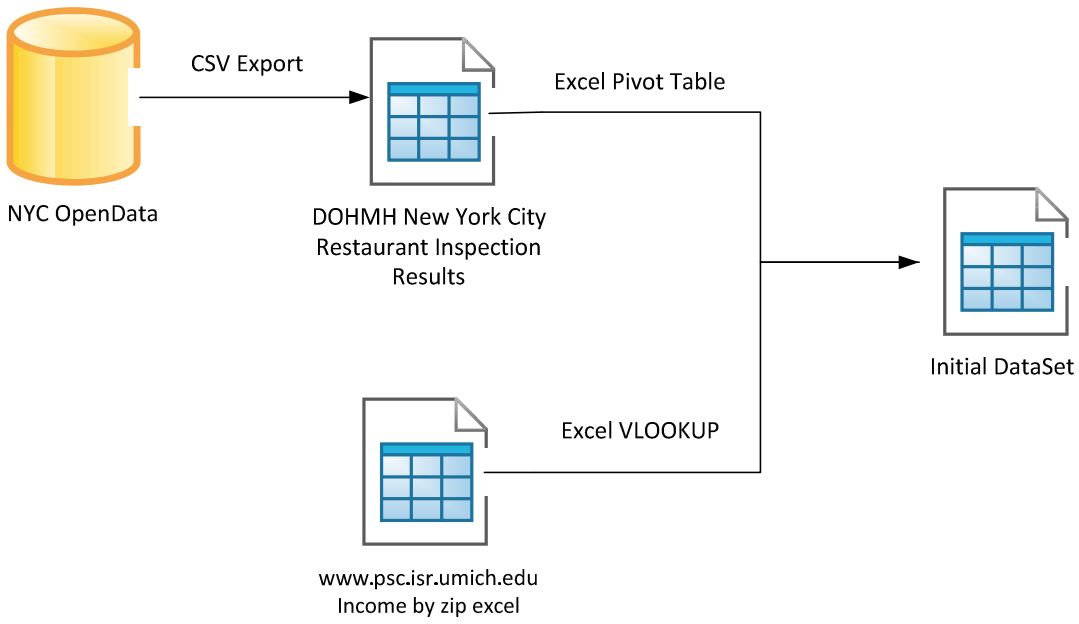
Appendix 2: Misc. Data Collection

1. xlsx to sas7bdat SAS code:

In order to bring our data into SAS Enterprise Miner we used the import procedure in core SAS to transform the excel data set into a permanent SAS dataset. We then copied the corresponding sas7bdat created by SAS core into a library for import into SAS Enterprise Miner.

```
PROC IMPORT OUT= NYC.grades DATAFILE= "C:\sas\grades.xlsx"
            DBMS=xlsx REPLACE;
     SHEET="grades";
     GETNAMES=YES;
RUN;
```

2. Graphic representation of data cleanup/combination process:



3. List of all variables:

Our initial data library is provided below.

Field	Description
NYC Open Data	
SCORE (Target)	Date of the most recent inspection - only restaurants w/ grade more recent than 1/1/2013
CAMIS	A unique ID used by NYC Health Department
DBA	The name of the restaurant
BORO	The borough of New York that the restaurant is located in
ZIPCODE	The zip code of the restaurant
CUISINE DESCRIPTION	The type of cuisine recorded by the health department. (e.g. Italian, American, Mexican, café)
INSPECTION DATE	Date of the most recent inspection.
_ViolationCode	98 columns containing the count of previous violations for that specific code. Full list is provided in Appendix 1
Derived from NYC Open Data	
NumPrevInspections	Number of total previous inspections
AvgPrevScore	Average score from all previous exams
www.psc.isr.umich.edu	
ZipPopulation	Population for restaurants by zip code
ZipMedianIncome	Median income for restaurants by zip code

Appendix 3: Audited Data Points

Audit Process

Since the cleanup and transformation process required manual manipulation, we decided to perform an audit to ensure the integrity of the dataset was still intact before beginning the data exploration stage. In order to create a random sample for audit, all values where assigned a random number using the RAND() function in excel and then sorted from smallest to largest. The top 22 values (just over 0.1% of the sampling frame) where kept and audited. Each of these audit points was manually checked using the NYC Open Data website and original copy of the University of Michigan's Population Studies Center Median income by zip excel sheet.

#	audit number	CAMIS	DBA	Result
1	0.00000	50002480	CITI CAFE/ROHATYN ROOM	OK
2	0.00001	41016740	SATGURU SWEETS & CATERING	OK
3	0.00002	40974766	CHIKALICIOUS DESSERT BAR	OK
4	0.00009	50006377	LOS TAQUITOS DEL TIO	OK
5	0.00019	41685731	GOLDEN KITCHEN	OK
6	0.00024	41603386	NEW HAPPY JOY	OK
7	0.00024	41508258	BELLA MAMA ROSE	OK
8	0.00029	40806791	GINGER'S BAR	OK
9	0.00030	41467170	HUNAN GLATT KOSHER	OK
10	0.00034	41695257	NO. 7 SUB	OK
11	0.00042	41468049	AU BON PAIN	OK
12	0.00064	41407168	HACHI	OK
13	0.00068	40968428	MR. DENNEHY'S	OK
14	0.00070	41573878	SURFISH BISTRO	OK
15	0.00072	40728353	UNIQUE LOUNGE	OK
16	0.00079	41433965	NO 1 CHINESE RESTAURANT	OK
17	0.00084	40400096	CIRCLE'S GRILL	OK
18	0.00088	41313395	FIVE NAPKIN BURGER	OK
19	0.00093	40392063	WALTER'S BAR	OK
20	0.00094	41634185	EMPELLON COCINA	OK
21	0.00102	41340507	SUBWAY	OK

Appendix 3-Part of Grades 3

lastSCORE	CAMIS	DBA	BORO	ZIPCODE	ZipMedIncome	ZipMeanIncome	ZipPop	CUISINEDESCRIPTION	average score	numOfIns	LastGRA	PC1	PC2	PC3	PC4	PC5
6	30075445	MORRIS PARK BAKE SH	BRONX	10462	46085.2485	51310.3756	71130	Bakery	13.50	9	A	4.29948	-0.766	-0.04847	-0.5987	0.041
8	30112340	WENDY'S	BROOKLYN	11225	39782.409	52484.8359	61858	Hamburgers	14.00	6	A	3.26189	-0.28	0.02462	0.6286	0.075
2	30191841	DJ REYNOLDS PUB ANI	MANHATTAN	10019	84786.4665	133175.3716	39048	Irish	11.60	5	A	3.0706	-0.259	-0.03654	-0.3177	-0.08
5	40356018	RIVIERA CATERER	BROOKLYN	11224	34958.4279	45362.8356	49491	American	12.20	5	A	2.54378	0.1304	-0.01404	-0.4401	-0.25
20	40356068	TOV KOSHER KITCHEN	QUEENS	11374	49707.9331	65129.5937	39224	Jewish/Kosher	16.43	7	B	5.1419	-1.26	-0.02588	-0.9516	0.52
38	40356151	BRUNOS ON THE BOUL	QUEENS	11369	54732.4966	65142.6645	37429	American	22.83	7	Z	3.96802	1.7169	-0.07451	0.4442	-1.6
9	40356442	KOSHER ISLAND	STATEN ISLAI	10314	78413.7436	90427.5476	85260	Jewish/Kosher	11.20	5	A	2.49112	-0.533	0.02131	0.6868	0.714
10	40356483	WILKEN'S FINE FOOD	BROOKLYN	11234	67952.9473	77076.9976	81033	Delicatessen	15.20	11	A	7.17868	-0.542	-0.00046	-0.0923	-0.51
12	40356649	REGINA CATERERS	BROOKLYN	11219	34969.9857	50736.5471	85825	American	12.14	7	A	4.82212	-0.592	-0.07781	-1.9649	-0.45
12	40356731	TASTE THE TROPICS IC	BROOKLYN	11226	38763.3706	49672.5306	106461	Ice Cream, Gelato, Yog	9.60	5	A	2.55444	-0.626	-0.00513	-0.1667	0.381
11	40357217	WILD ASIA	BRONX	10460	23644.951	33567.5699	53752	American	6.00	3	A	1.35745	-0.311	0.0004	0.2045	0.091
5	40357437	C & C CATERING SERVI	BROOKLYN	11214	35963.5161	48902.3504	78896	American	11.40	5	A	1.67391	-0.093	-0.02816	-0.4006	-0.2
22	40358429	MAY MAY KITCHEN	BROOKLYN	11208	33875.818	44514.588	84692	Chinese	29.55	13	Z	12.805	-2.807	-0.06091	-2.2822	0.927
3	40359480	1 EAST 66TH STREET KI	MANHATTAN	10065	108154.9505	191673.5228	31500	American	6.20	5	A	1.45649	0.1496	0.03988	0.2225	-0.22
25	40359705	NATHAN'S FAMOUS	BROOKLYN	11224	34958.4279	45362.8356	49491	Hotdogs	16.25	9	0	6.12061	0.7985	0.05919	0.4826	-1.19
40	40360045	SEUDA FOODS	BROOKLYN	11223	39420.9802	53995.6663	74659	Jewish/Kosher	18.25	8	0	6.51057	-0.329	-0.14704	-2.8817	-0.95
9	40360076	CARVEL ICE CREAM	BROOKLYN	11218	48010.6089	61311.5804	75418	Ice Cream, Gelato, Yog	8.00	3	A	1.36907	-0.307	-0.00742	-0.2608	0.114
25	40361322	CARVEL ICE CREAM	QUEENS	11004	75303.738	84521.7605	13390	Ice Cream, Gelato, Yog	15.17	6	Z	3.31979	0.4391	0.0513	-0.6992	-0.76
10	40361521	GLORIOUS FOOD	MANHATTAN	10021	113800.3846	198530.1081	44173	American	21.00	12	A	9.5587	-1.006	-0.05926	-1.1696	-0.26
11	40361606	THE MOVABLE FEAST	BROOKLYN	11215	81331.0535	109383.9243	59307	American	11.20	5	A	3.33987	-0.135	-0.08435	-1.6308	-0.66
12	40361618	SAL'S DELI	QUEENS	11356	67807.8323	80098.4197	20443	Delicatessen	9.50	4	A	1.98052	-0.237	0.00525	0.2036	0.152
10	40361708	BULLY'S DELI	MANHATTAN	10003	89998.5339	139330.9968	53609	Delicatessen	10.50	6	A	2.83122	0.3353	0.07837	0.3407	-0.5
2	40361998	STEVE CHU'S DELI & GF	QUEENS	11106	44030.5142	56437.4173	40963	Delicatessen	6.25	4	A	1.46989	-0.422	-0.00543	0.1228	0.397
10	40362098	HARRIET'S KITCHEN	MANHATTAN	10024	110998.9883	191408.2026	61485	Chicken	13.67	12	A	9.20176	-1.685	-0.15798	-4.2321	-0.69
13	40362264	P & S DELI GROCERY	MANHATTAN	10025	65001.2415	112009.46	96117	American	22.14	8	A	5.9943	-0.511	0.01986	-0.0963	-0.35
13	40362274	ANGELIKA FILM CENTE	MANHATTAN	10012	81316.7066	135998.6472	26464	American	10.50	4	A	1.69198	0.1533	0.04967	0.0551	-0.25
12	40362432	HO MEI RESTAURANT	QUEENS	11368	43962.6552	54259.4037	95662	Chinese	16.27	11	A	7.12827	-1.788	-0.06483	-1.5185	0.773
10	40362715	THE COUNTRY CAFE	MANHATTAN	10005	115133.2855	163762.6601	1517	Turkish	18.75	8	A	6.37658	-1.441	-0.03919	-0.8502	0.401
2	40362869	SHASHEMENE INT'L RE	BROOKLYN	11203	47788.2901	60025.2785	78886	Caribbean	10.00	5	A	2.7899	0.1818	0.02954	-0.507	-1.01
10	40363093	CARVEL ICE CREAM	BRONX	10466	48469.7389	57368.0989	68662	Ice Cream, Gelato, Yog	13.75	8	A	5.24457	-0.235	-0.04643	-1.1723	-0.64
8	40363098	DUNKIN' DONUTS	BROOKLYN	11201	92174.8101	140857.9804	48101	Donuts	9.50	4	A	1.8366	-0.325	0.00246	-0.0025	0.126
10	40363117	MEJLANDER & MULGA	BROOKLYN	11209	56906.5301	78504.7606	69646	American	9.17	7	A	2.41658	-0.525	-0.02311	-0.2964	0.08
12	40363151	OLIVE'S	MANHATTAN	10012	81316.7066	135998.6472	26464	Sandwiches/Salads/Mi	13.50	8	A	4.16598	-1.079	-0.02386	-0.5563	0.268
9	40363289	HAPPY GARDEN	BRONX	10474	24244.2862	31476.2226	11365	Chinese	12.00	6	A	3.61342	-0.842	-0.01655	-0.742	0.102
12	40363298	CAFE METRO	MANHATTAN	10018	84799.0772	112292.3727	4255	American	13.86	8	A	5.28227	-0.98	-0.03701	-0.2256	0.032
11	40363333	TONY'S DELI	QUEENS	11385	47469.8714	57794.7135	95262	Delicatessen	14.44	9	A	5.96773	-1.455	0.0263	-1.357	0.305
12	40363426	LEXLER DELI	MANHATTAN	10174	104023	146411	1	Sandwiches/Salads/Mi	7.40	5	A	1.8947	-0.377	0.00344	0.2358	0.042
3	40363427	BAGELS N BUNS	STATEN ISLAI	10314	78413.7436	90427.5476	85260	Delicatessen	15.00	7	A	4.47829	-0.699	0.02647	0.5202	0.091
7	40363565	HOT BAGELS	QUEENS	11379	61824.3232	76167.0077	33886	Bagels/Pretzels	19.42	12	A	8.63383	-0.543	0.0361	-0.6227	-0.43
3	40363590	SNACK TIME GRILL	QUEENS	11418	57063.3834	66711.6469	36587	American	7.29	7	A	2.40478	-0.713	0.00862	0.1706	0.546
9	40363630	LORENZO & MARIA'S	MANHATTAN	10028	105456.4514	192257.802	40914	Continental	20.50	8	A	5.28085	-0.778	0.06335	0.1575	-0.22
3	40363644	DOMINO'S PIZZA	MANHATTAN	10016	96760.4462	144872.3901	49904	Pizza	11.60	10	A	5.23068	-0.6	-0.06245	-1.477	-0.58
9	40363685	BERKELY	MANHATTAN	10022	93106.6559	158965.178	26460	American	11.57	7	A	4.32583	-0.934	0.03843	-0.4722	0.158
0	40363744	SONNY'S HEROS	BROOKLYN	11236	59567.5509	69801.3086	94098	American	1.00	3	A	0.22386	-0.006	0.00317	0.0332	-0.05
12	40363834	CARVEL ICE CREAM	STATEN ISLAI	10305	58937.658	75468.7376	37014	Ice Cream, Gelato, Yog	6.33	3	A	1.50639	-0.26	-0.01968	-0.4349	-0.08
17	40363920	NEW GOLDEN BILLION	BROOKLYN	11212	27911.2921	37845.6579	85400	Chinese	10.75	5	0	2.2337	-0.341	-0.01846	-0.0668	-0.05
13	40363945	DOMINO'S PIZZA	MANHATTAN	10023	108285.5918	194365.9259	59002	Pizza	17.17	7	A	4.07673	0.4104	0.18445	0.1013	-1.26
5	40364179	SPOON BREAD CATERI	MANHATTAN	10025	65001.2415	112009.46	96117	American	18.13	8	A	6.09096	-0.413	-0.05179	-1.1002	-0.68
10	40364220	KOSHER BAGEL HOLE	BROOKLYN	11230	47496.1618	64348.0001	88449	Jewish/Kosher	9.25	4	A	1.38604	-0.37	0.08591	0.3562	0.12
10	40364262	KOSHER BAGEL HOLE	BROOKLYN	11230	47496.1618	64348.0001	88449	Jewish/Kosher	9.33	4	A	1.48921	-0.462	0.01081	0.2613	0.199
13	40364286	PLAZA BAGELS & DELI	STATEN ISLAI	10306	75212.0632	91445.7035	54812	Delicatessen	14.17	7	A	3.7195	-1.25	0.01041	0.1202	0.904
12	40364296	HAPPY GARDEN	BRONX	10458	25642.1721	34873.0857	77840	Chinese	16.88	8	A	6.15123	-1.601	-0.04327	-1.4571	0.533
4	40364299	B & M HOT BAGEL & GI	STATEN ISLAI	10308	75667.7216	85725.9159	23972	Delicatessen	19.44	11	A	6.93342	-0.075	-0.06375	-1.0311	-1.02
12	40364304	TEXAS ROTISSERIE	MANHATTAN	10038	55937.3774	89760.0452	15435	Chicken	15.22	11	A	5.18342	-1.166	0.04149	0.0014	0.369
9	40364305	PHILADELHIA GRILLE E	BROOKLYN	11209	56906.5301	78504.7606	69646	Italian	11.83	6	A	3.27248	-1.087	0.00265	-0.1744	0.75
11	40364335	PETER LUGER STEAKHO	BROOKLYN	11211	37632.4577	52060.9141	84434	Steak	27.25	8	A	7.61833	-1.171	-0.06231	-0.8889	0.051
9	40364347	METROPOLITAN CLUB	MANHATTAN	10022	93106.6559	158965.178	26460	American	13.91	11	A	7.21124	-1.572	0.00215	-1.6518	0.233
5	40364355	PALM RESTAURANT	MANHATTAN	10017	102523.5025	149723.7834	16231	American	16.80	10	A	7.69422	-1.286	-0.09533	-1.7792	-0.03
11	40364362	21 CLUB	MANHATTAN	10019	84786.4665	133175.3716	39048	American	16.43	7	A	5.9879	-0.448	0.01652	0.5176	-0.11
5	40364363	MANHEM CLUB	BRONX	10465	61446.3421	73646.2435	41282	American	10.25	4	A	2.34582	0.705	-0.02618	-0.3716	-1.09
13	40364373	ISLE OF CAPRI RESTUR	MANHATTAN	10065	108154.9505	191673.5228	31500	Italian	14.78	9	A	6.50866	-0.099	-0.12128	-1.9641	-1.09
10	40364389	OLD TOWN BAR & RES	MANHATTAN	10003	89998.5339	139330.9968	53609	American	17.00	5	A	3.23682	-0.04	-0.03703	-0.355	-0.27
2	40364404	POLISH NATIONAL HO	BROOKLYN	11222	54366.9033	69808.6353	40003	Polish	15.40	5	A	2.9014	0.8523	-0.0217	0.0744	-1.06

Appendix 4 Stepwise regression output

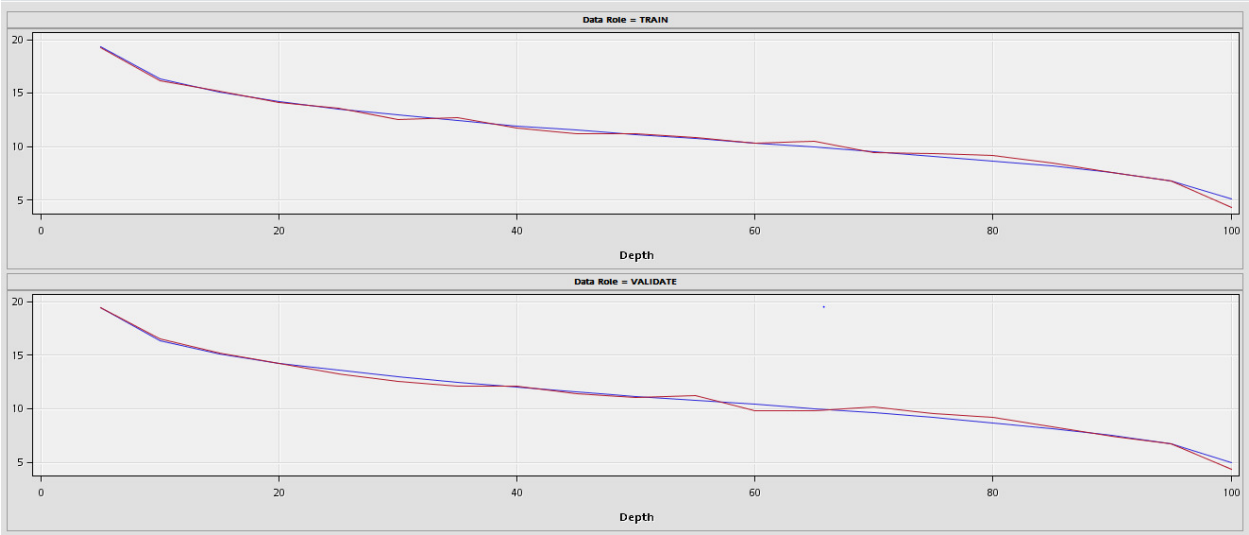
Model Fit Statistics

R-Square	0.2310	Adj R-Sq	0.2303
AIC	45930.8238	BIC	45932.8241
SBC	46020.1336	C(p)	23.8488

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	4.2091	0.3744	11.24	<.0001
BORO	BRONX	1	0.4090	0.3366	1.22	0.2243
BORO	BROOKLYN	1	-0.4330	0.3087	-1.40	0.1608
BORO	MANHATTAN	1	-0.1439	0.2962	-0.49	0.6270
BORO	Missing	1	-0.3196	1.4459	-0.22	0.8251
BORO	QUEENS	1	-0.2771	0.3075	-0.90	0.3676
BORO	STATEN ISLAND	0	0	.	.	.
PC1		1	0.5748	0.0650	8.84	<.0001
PC2		1	0.6708	0.1205	5.56	<.0001
PC5		1	0.5613	0.1325	4.24	<.0001
ZipPop		1	-4.97E-6	2.429E-6	-2.05	0.0408
average_score		1	0.5606	0.0169	33.19	<.0001
numOfInspections		1	-0.3864	0.0408	-9.47	<.0001

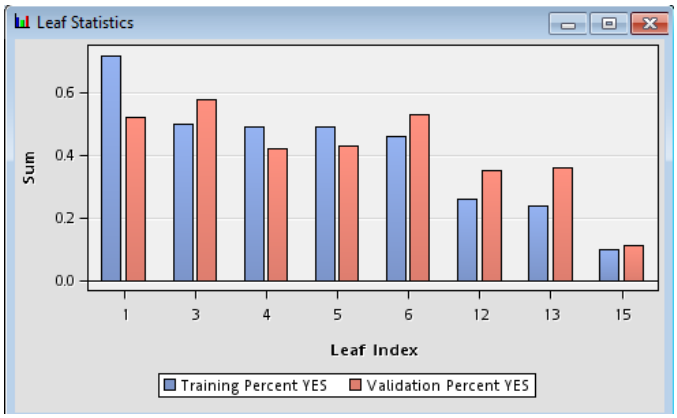
Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	45933.93	.
ASE	Average Squared Error	38.07665	39.53918
AVERR	Average Error Function	38.07665	39.53918
DFE	Degrees of Freedom for Error	12602	.
DFM	Model Degrees of Freedom	12	.
DFT	Total Degrees of Freedom	12614	.
DIV	Divisor for ASE	12614	8409
ERR	Error Function	480298.9	332484.9
FPE	Final Prediction Error	38.14917	.
MAX	Maximum Absolute Error	51.4472	76.84238
MSE	Mean Square Error	38.11291	39.53918
NOBS	Sum of Frequencies	12614	8409
_NW_	Number of Estimate Weights	12	.
RASE	Root Average Sum of Squares	6.170628	6.288018
RFPE	Root Final Prediction Error	6.176501	.
RMSE	Root Mean Squared Error	6.173566	6.288018
SBC	Schwarz's Bayesian Criterion	46023.24	.
SSE	Sum of Squared Errors	480298.9	332484.9
SUMW	Sum of Case Weights Times Freq	12614	8409

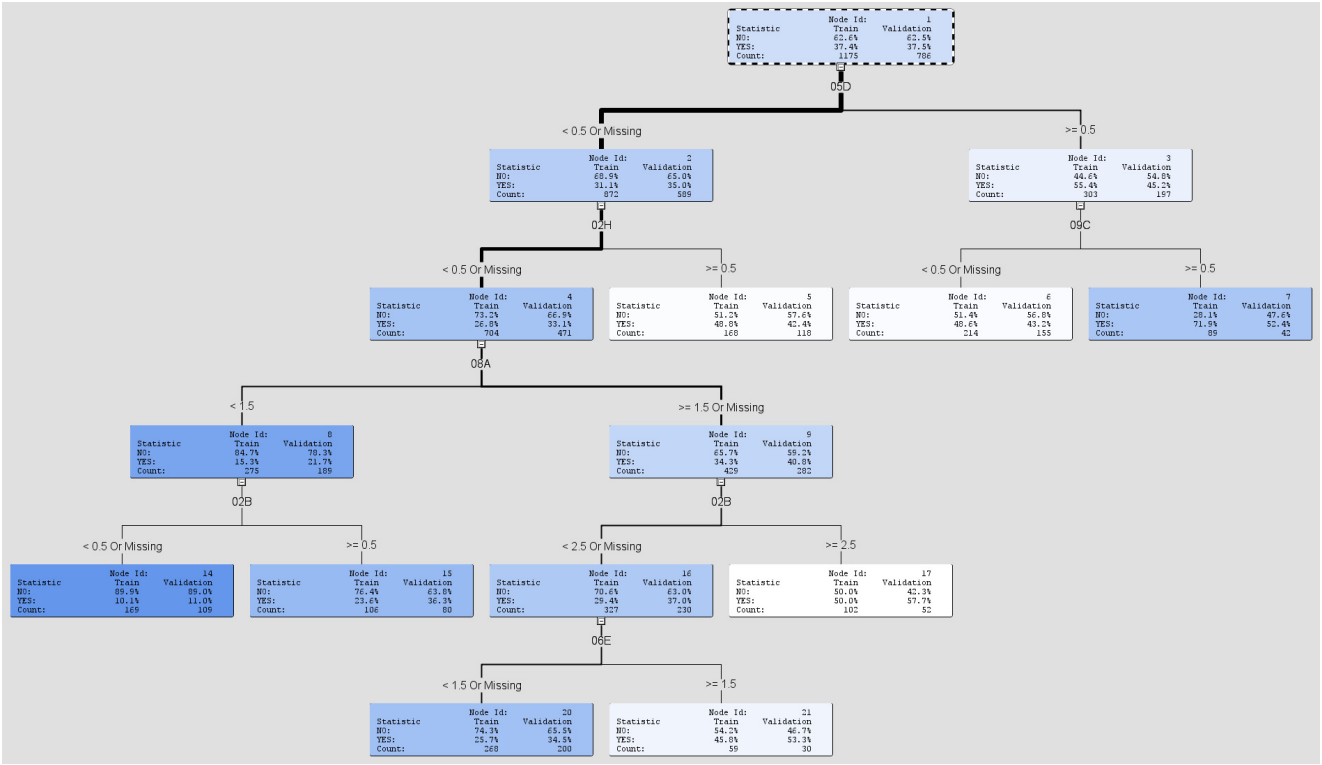


Appendix 5 decision tree/regression tree output

Binary Decision tree:

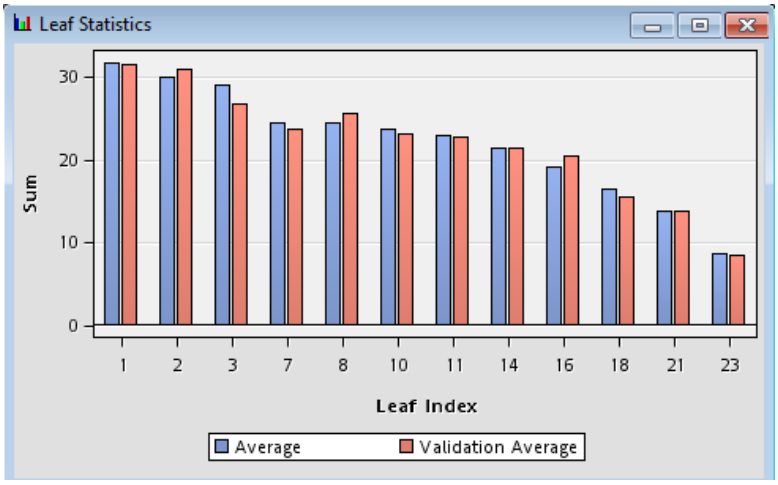
Statistics Label	Train	Validation
Sum of Frequencies	1175.00	786.00
Misclassification Rate	0.34	0.37
Maximum Absolute Error	0.90	0.90
Sum of Squared Errors	478.37	355.38
Average Squared Error	0.20	0.23
Root Average Squared Error	0.45	0.48
Divisor for ASE	2350.00	1572.00
Total Degrees of Freedom	1175.00	NaN

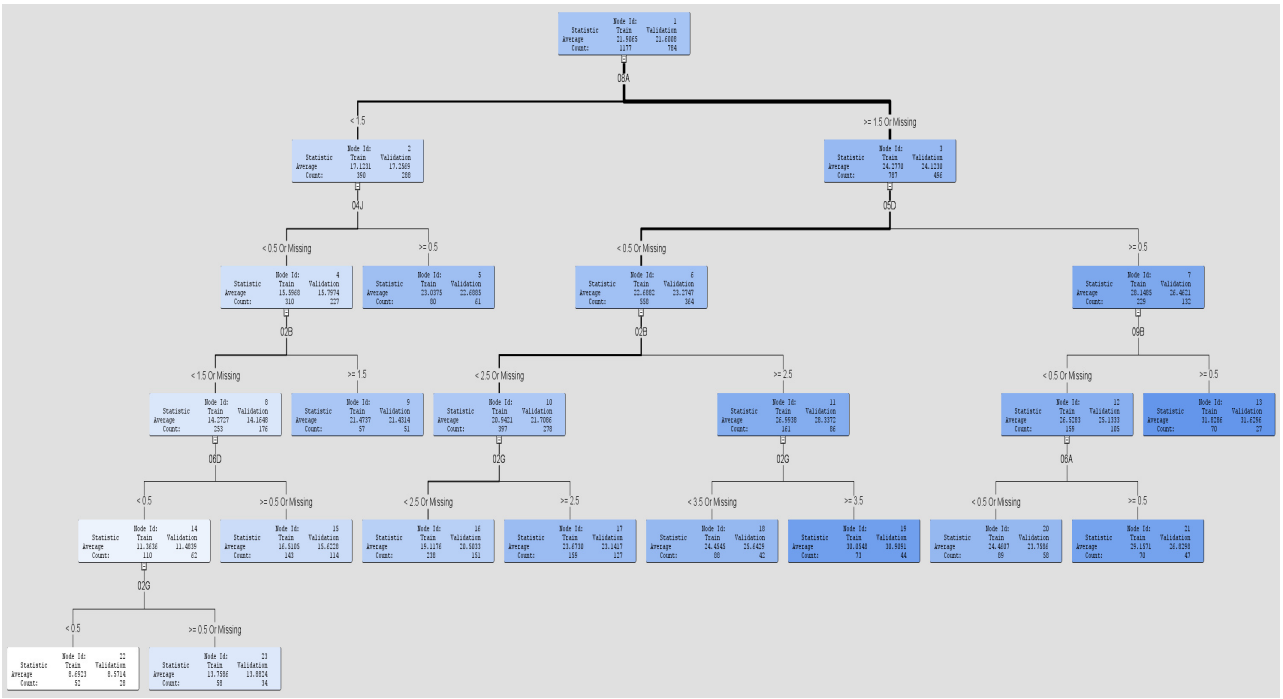




Regression tree:

Statistics Label	Train	Validation
Sum of Frequencies	1177.00	784.00
Maximum Absolute Error	75.96	43.95
Sum of Squared Errors	148125.16	109028.49
Average Squared Error	125.85	139.07
Root Average Squared Error	11.22	11.79
Divisor for ASE	1177.00	784.00
Total Degrees of Freedom	1177.00	NaN





[i] https://www.zagat.com/b/dining-trends-survey-tipping-pet-peeves-and-more#2

[ii] http://www.forbes.com/2010/04/28/tourism-new-york-lifestyle-travel-las-vegas-cities_slide_10.html

Refrences

[iii] http://www.nyc.gov/html/doh/downloads/pdf/report/nycdohmh-triennial09-11.pdf