DATA MINING
Study of Restaurants in US

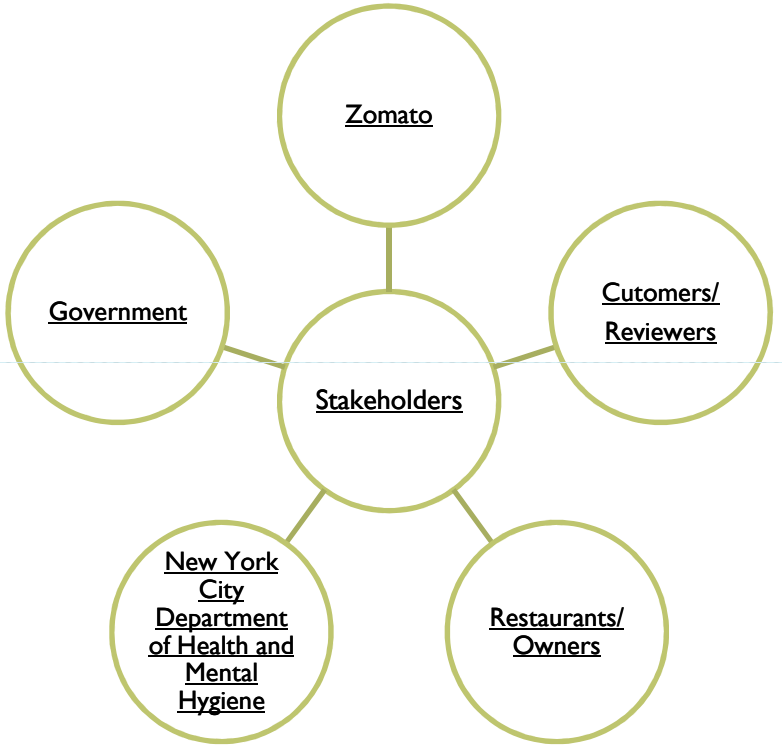
-Gunjan Batra

-Amol Kumtakar

-Krithika Raghavan

Objective - Data Analytics and Text Mining

- Using Tableau, Analyze the restaurant data set of 5 cities of US based on following parameters
  - Cost of restaurant
  - Cuisines
  - User review rating
  - Locality
  - Health Inspection Grade
  - Text in the Reviews
- Using WEKA, Use different classification algorithms to build a model for Text Data of User reviews to be classified as Positive/ Negative



Data description

Source	www.zomato.com http://www.nyc.gov/html/doh/html/environmental/food-service-inspection.shtml
Type	Numeric, String, Nominal
Size	495 Rows of Hotel Records
Range of values	99 Records each from New York, Chicago, Houston, San Francisco, Detroit
Prediction Class label	
Final Attributes	13+ 15 Reviews of each New York Restaurant

Restaurant_ID	Restaurant_Name	Restaurant_URL	Restaurant_Address	Restaurant_Locality	Restaurant_City	Restaurant_Zipcode	Restaurant_Cuisines	Restaurant_Avg_cost_for_two	Restaurant_UserRating	Restaurant_UserRatin	Restaurant_UserRatingVotes	Health Inspection
16781904	Momofuku	https://ww	171 1st Aver	East Village	New York Ci	10003	Asian, Rame	60	4.1	Excellent	1530	A
16767139	Gramercy Ta	https://ww	42 E 20th Str	Union Squa	New York Ci	10003	American	160	3.7	Very Good	1754	A
16760100	Balthazar	https://ww	80 Spring St	Soho	New York Ci	10012	French, Cafe	140	3.7	Very Good	4005	A
16775039	Peter Luger	https://ww	178 Broadw	Williamsbu	New York Ci	11211	Steakhouse,	150	3.8	Very Good	2091	A
16783153	Shake Shack	https://ww	366 Columb	Upper West	New York Ci	10024	American, B	30	4.2	Excellent	1524	A
16783998	The Halal Gu	https://ww	6th Avenue	Theater Dis	New York Ci	10019	Middle Easte	25	4.4	Excellent	493	A
16761344	Buddakan	https://ww	75 9th Aven	Meatpackin	New York Ci	10011	Chinese, Fus	150	3.9	Very Good	1483	A
16761402	Burger Joint	https://ww	Le Parker M	Theater Dis	New York Ci	10019	American, B	25	4.1	Excellent	1381	A
16785398	Shake Shack	https://ww	691 8th Ave	Hell's Kitch	New York Ci	10036	American, B	30	4.4	Excellent	856	A

Data Preprocessing

Web Script were written for Zomato Developer API

20 data records were obtained for each city at a time in the form of json files which was converted to csv

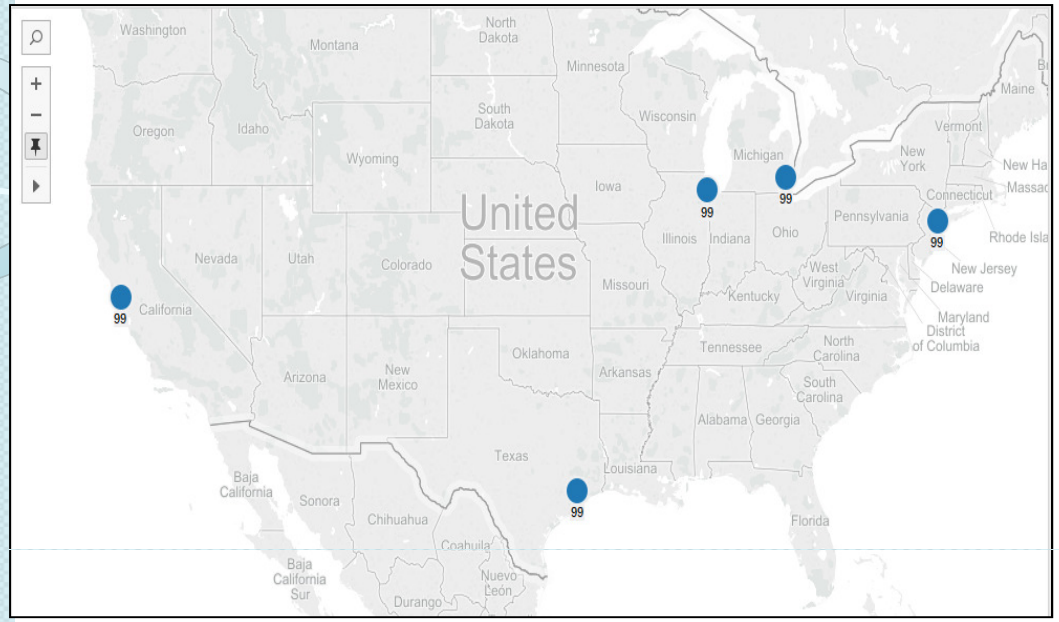
Combined the data set to obtain a collective data of 495 restaurants for exploration in Tableau and Weka

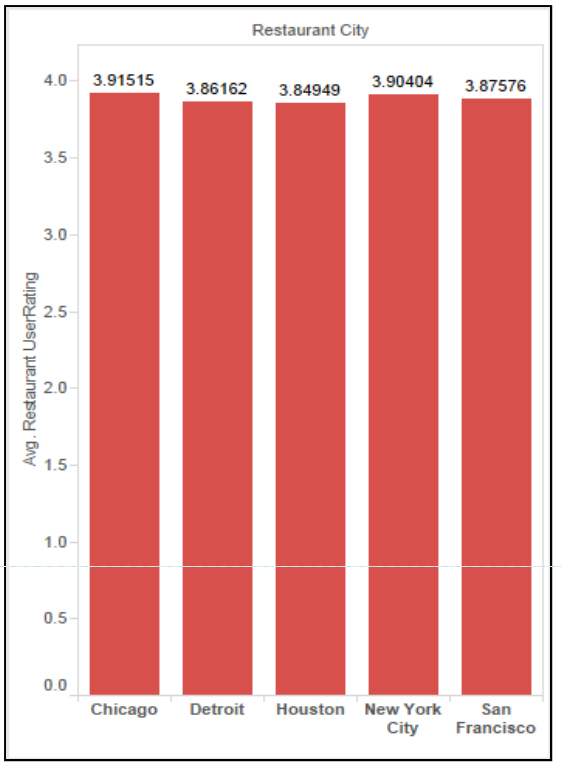
After exploration in Weka and Tableau the attributes were reduces from 31 to 13

For New York City hotels, manually collected the New York City Department of Health and Mental Hygiene Inspection Results Rating and User reviews of the Hotels(1545 reviews)

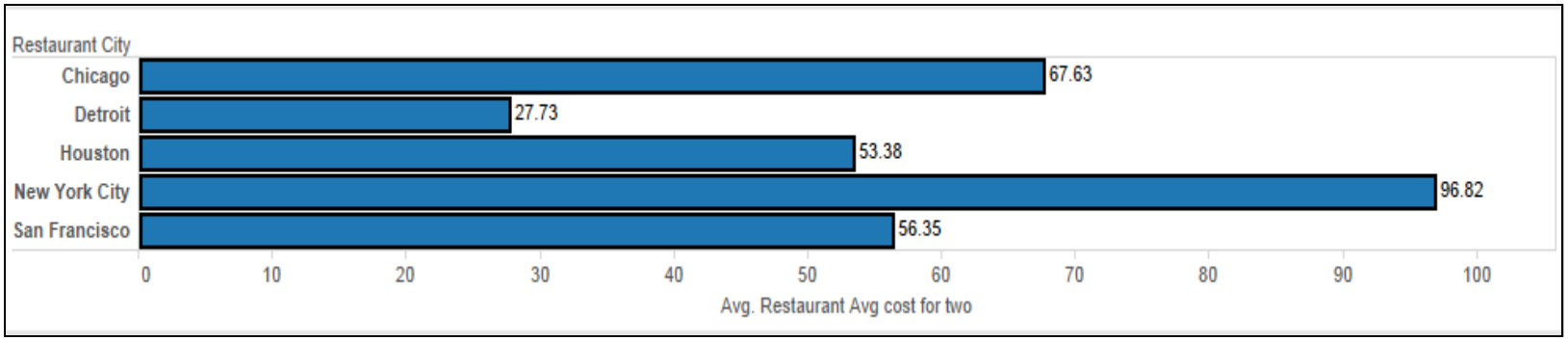
Reviews collected were converted into manually written arff files under two folders – Positive and Negative
Each folder contained **67 files and 70 files** each respectively corresponding to every restaurant

Data Set Description



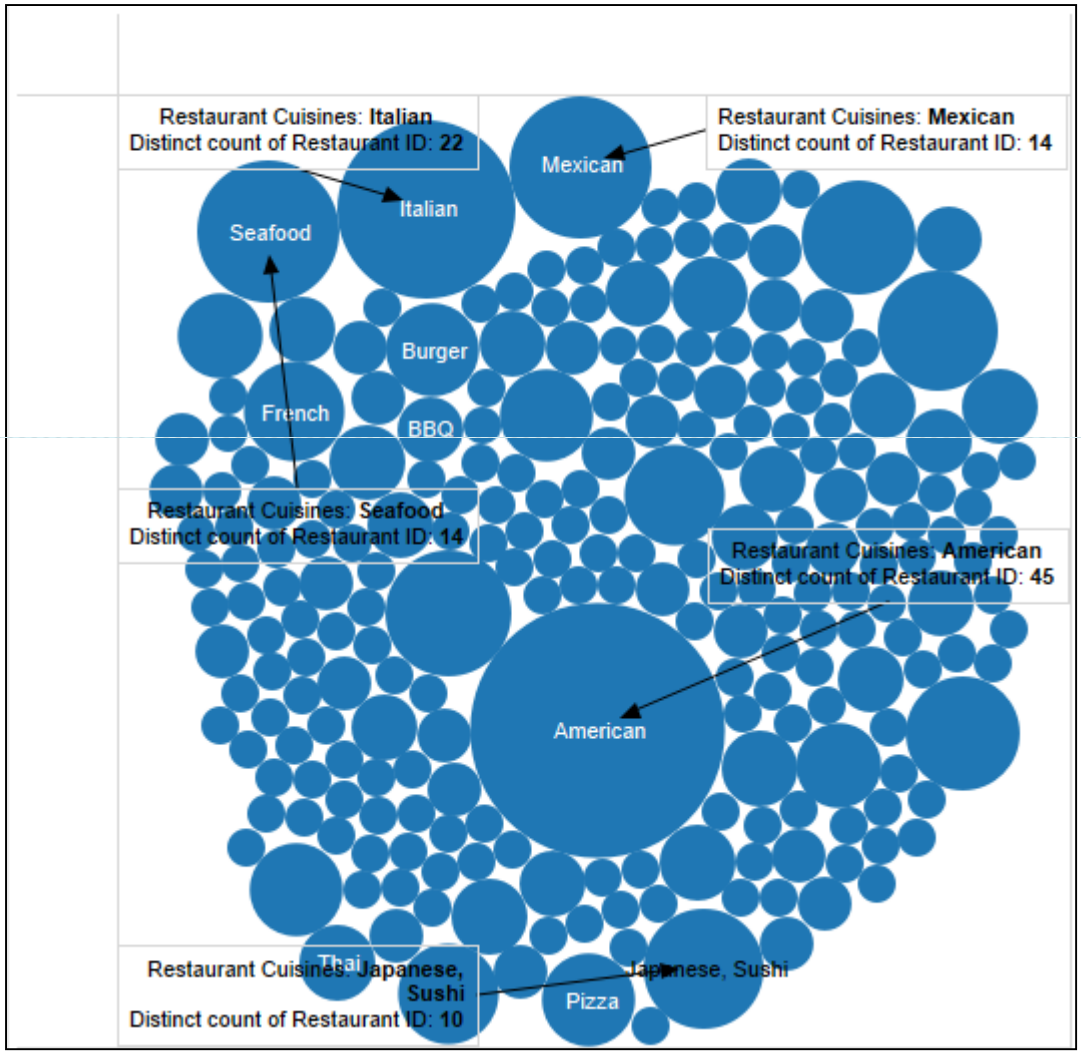


New York City has most expensive restaurants.
Chicago restaurants have the best user ratings



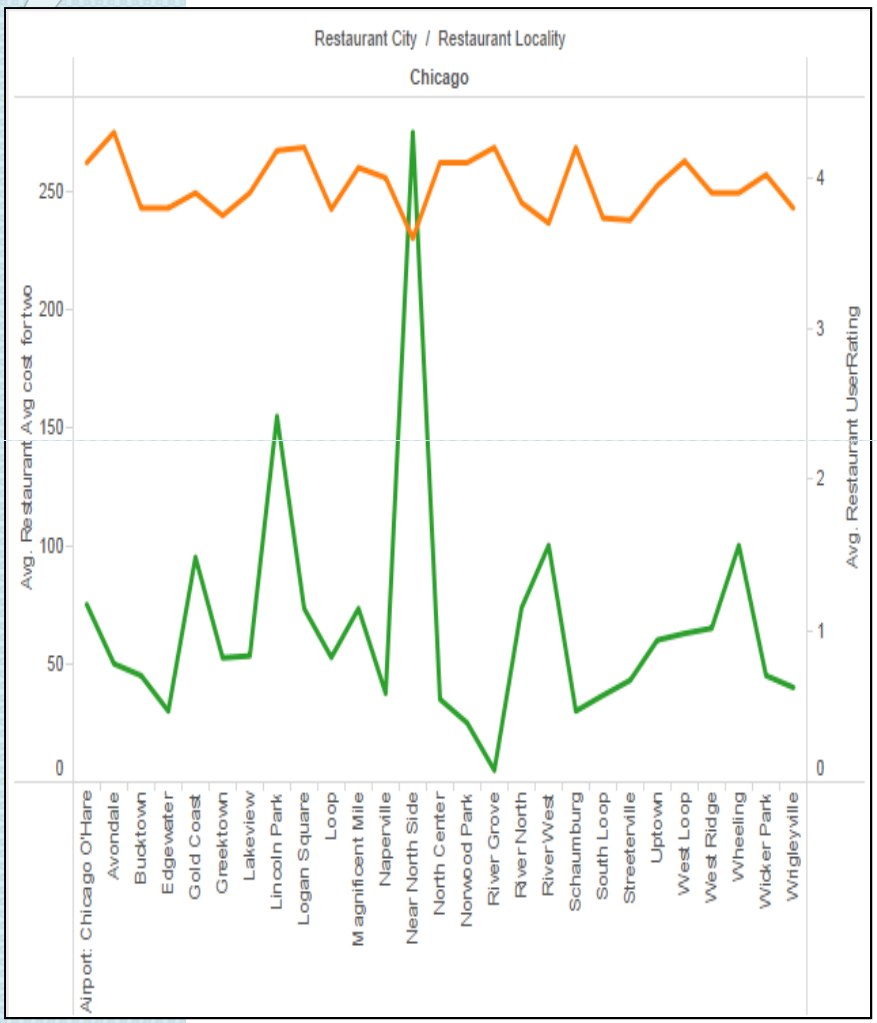
Cuisines Analysis

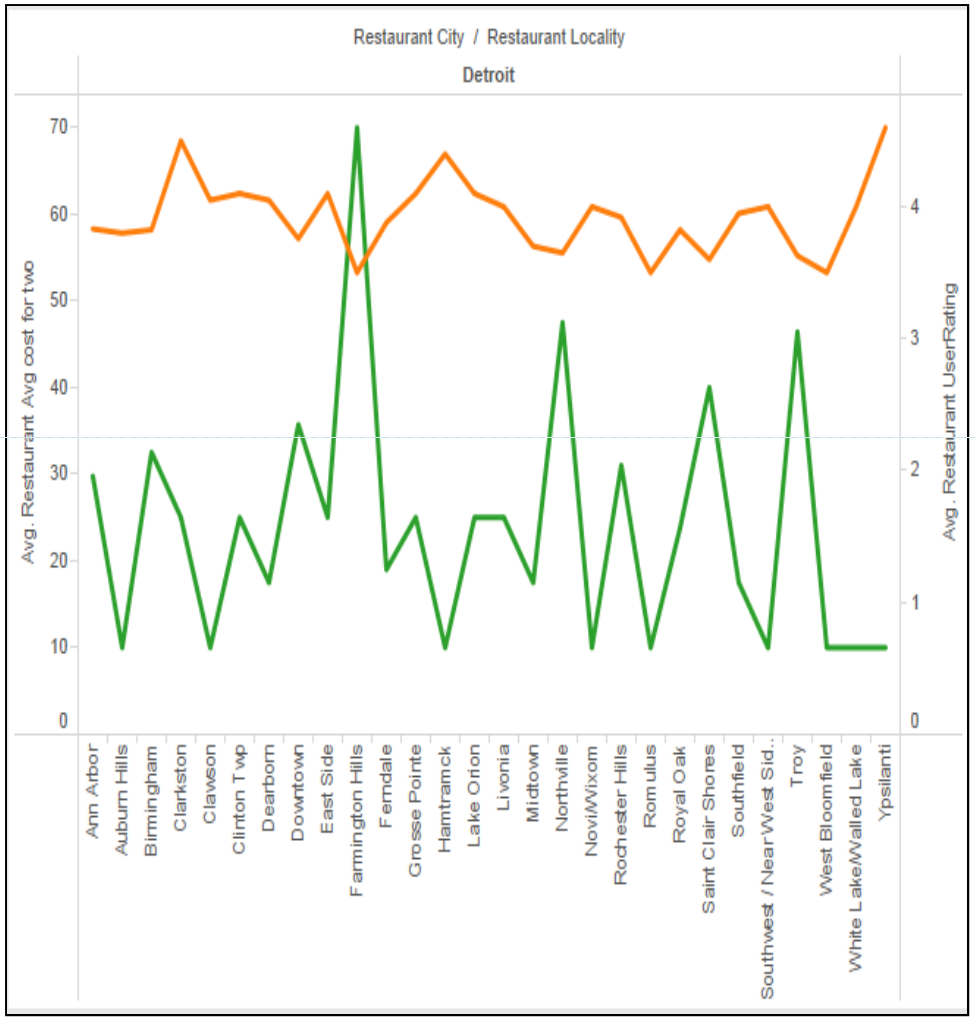
Count of Restaurants for various cuisine types



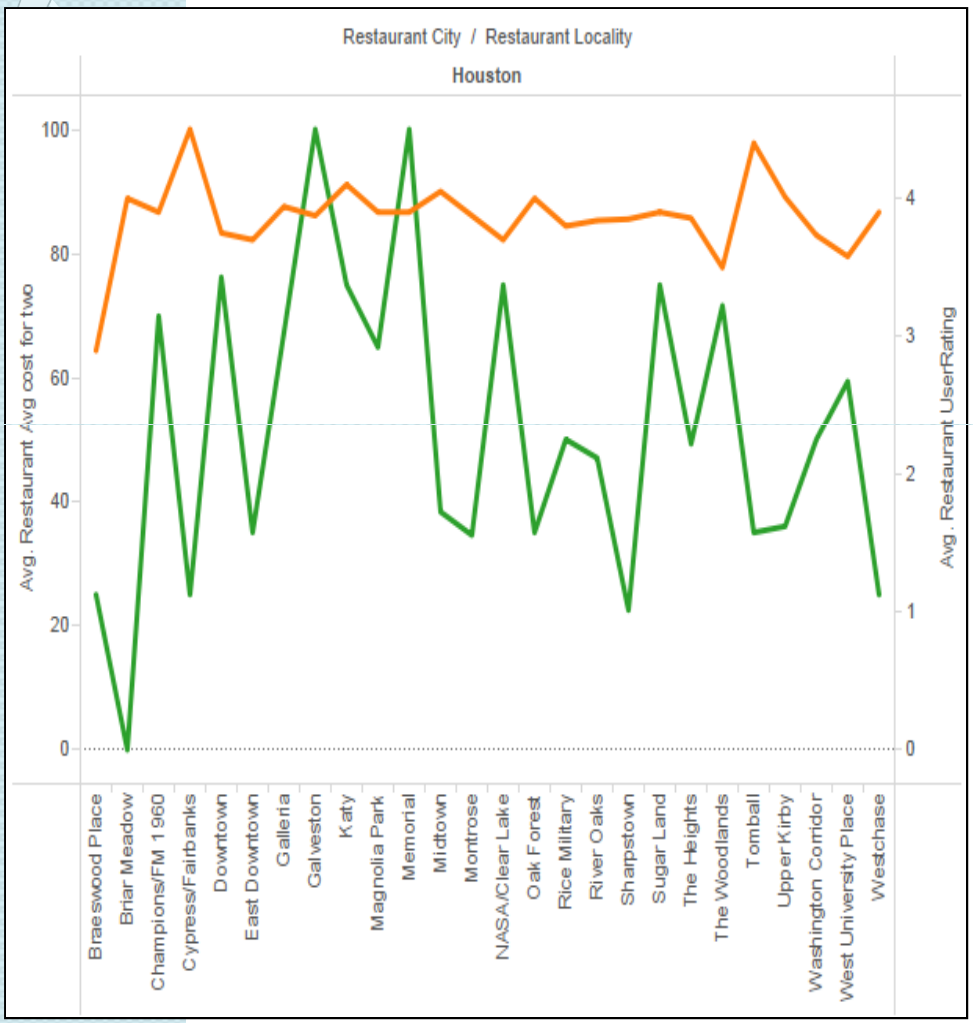
The most popular cuisine in our data set is American which is in 45 restaurants, followed by Italian, Mexican, Seafood and Japanese

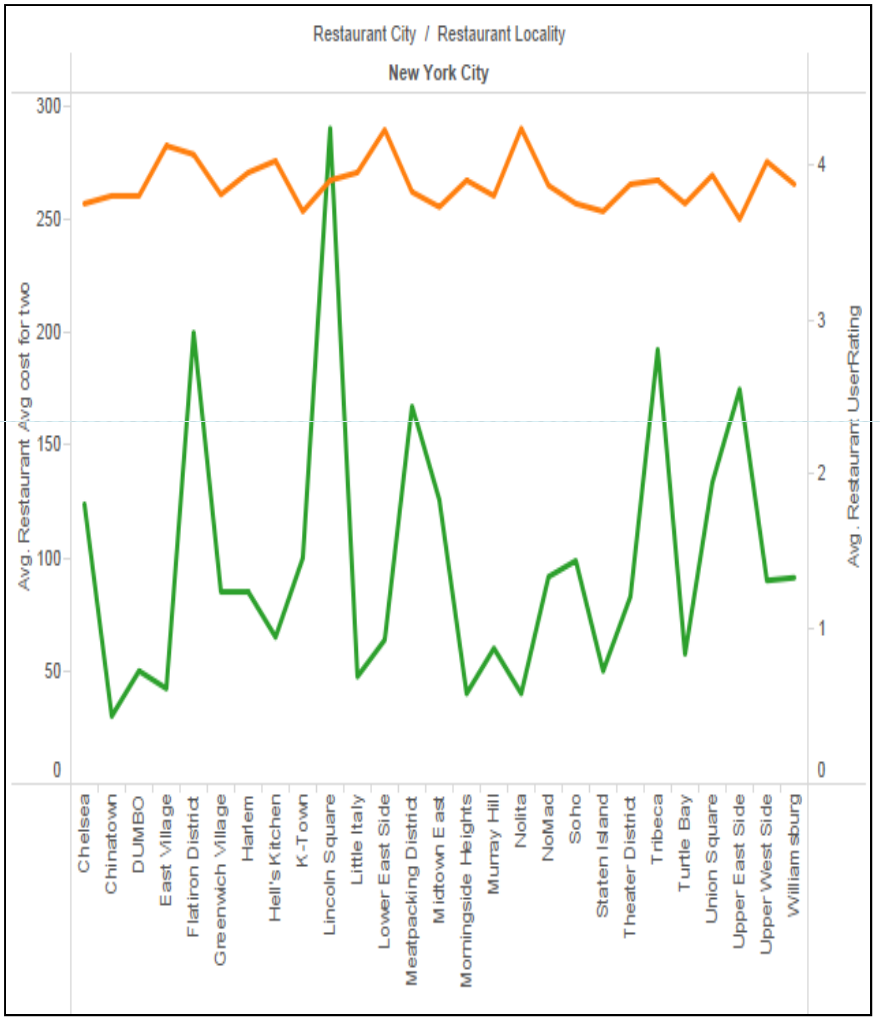
Study of relationship between Average Cost for two and Average User Rating of a Restaurant



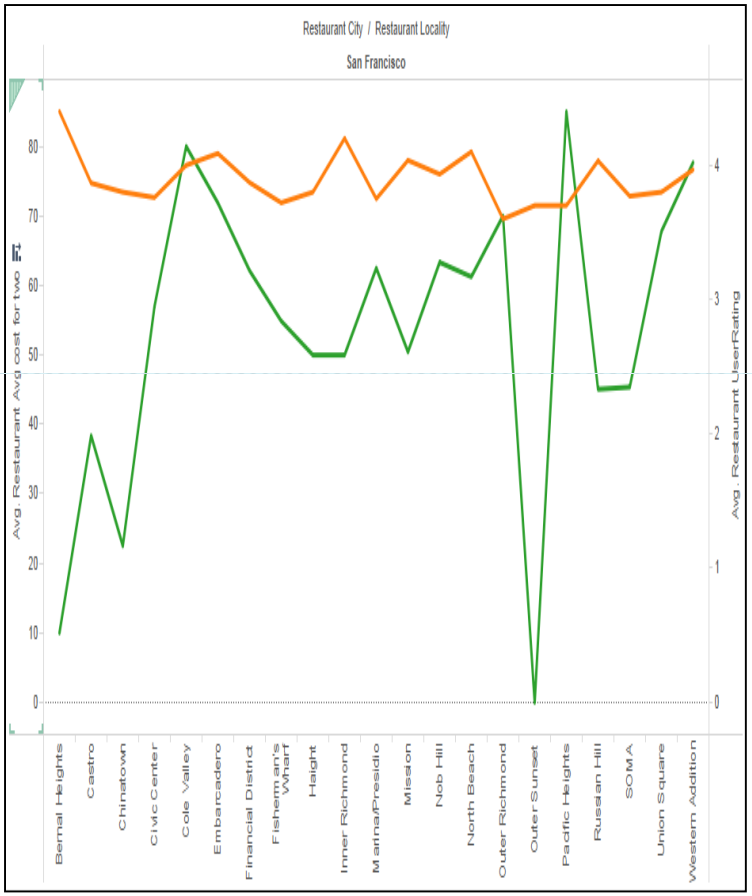


Study of relationship between Average Cost for two and Average User Rating of a Restaurant





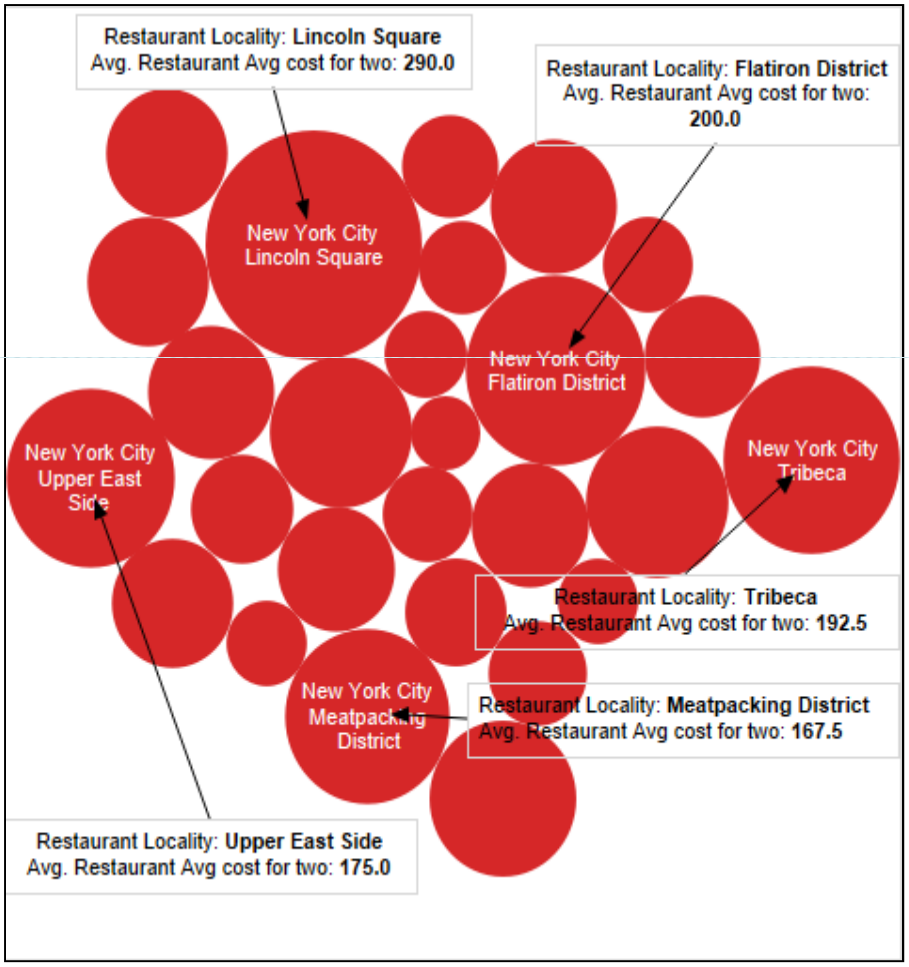
Study of relationship between Average Cost for two and Average User Rating of a Restaurant



Pearson Correlation Coefficient between the entire data set of Average Cost for two and User review rating is -0.0946664771605779 which mean that they are probably not correlated.

New York City – An Exploration

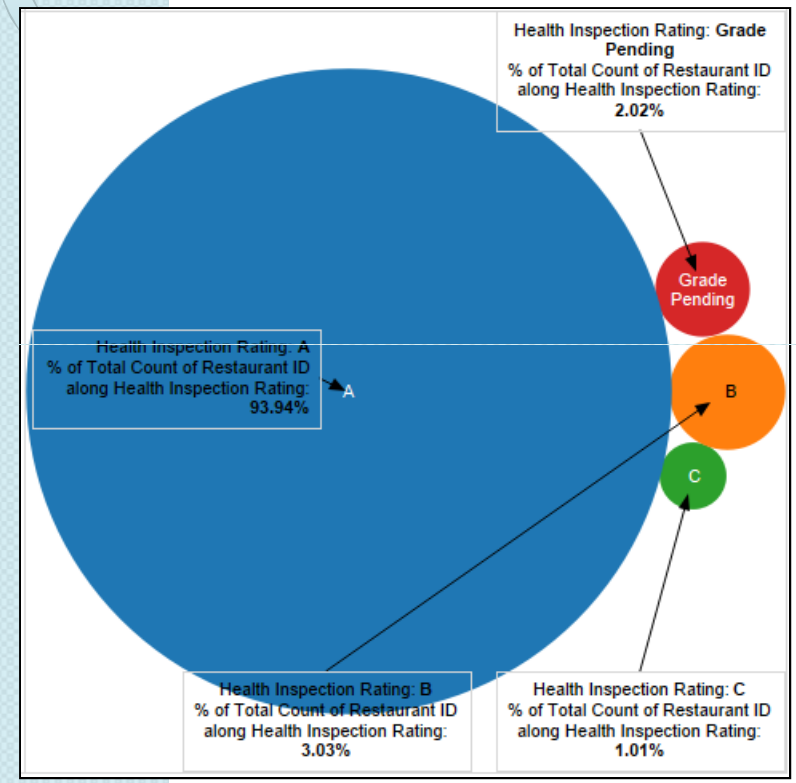
Locality wise – Average cost of two people



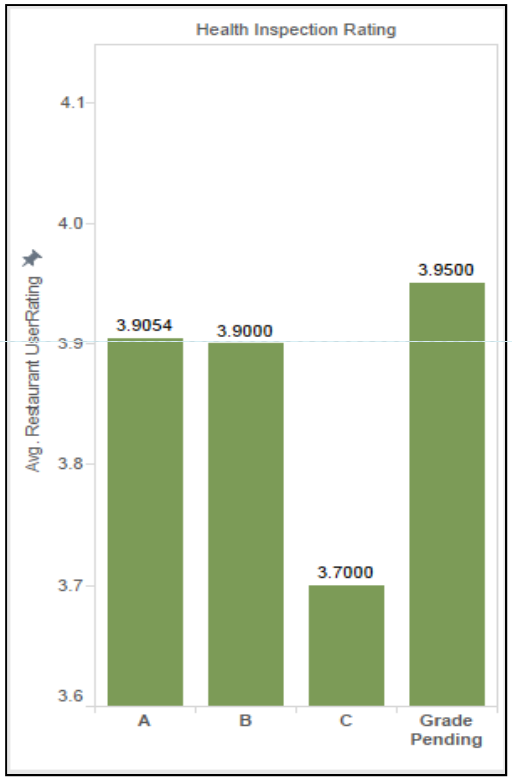
The most expensive restaurants in New York City are in Lincoln Square, followed by Flatiron District, Upper East Side, Tribeca and Meat Packing District

New York City – An Exploration
Results of Health Inspection Grades

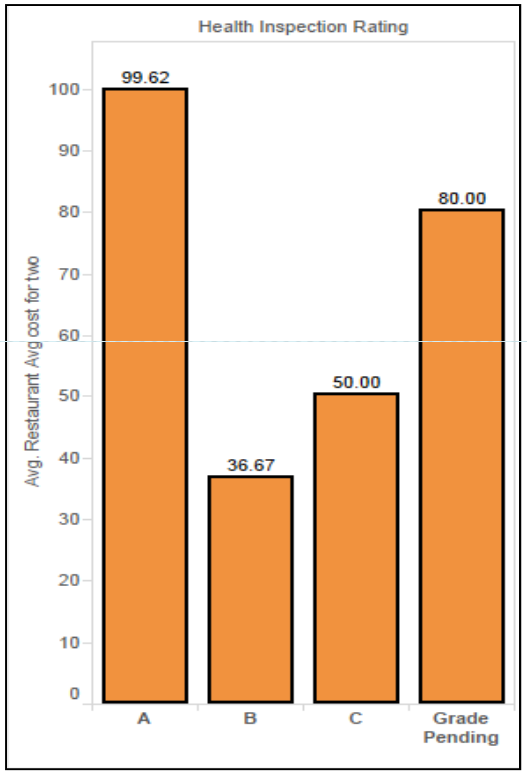
Percentage of Restaurants in each grade



Average User rating of restaurants in each grade



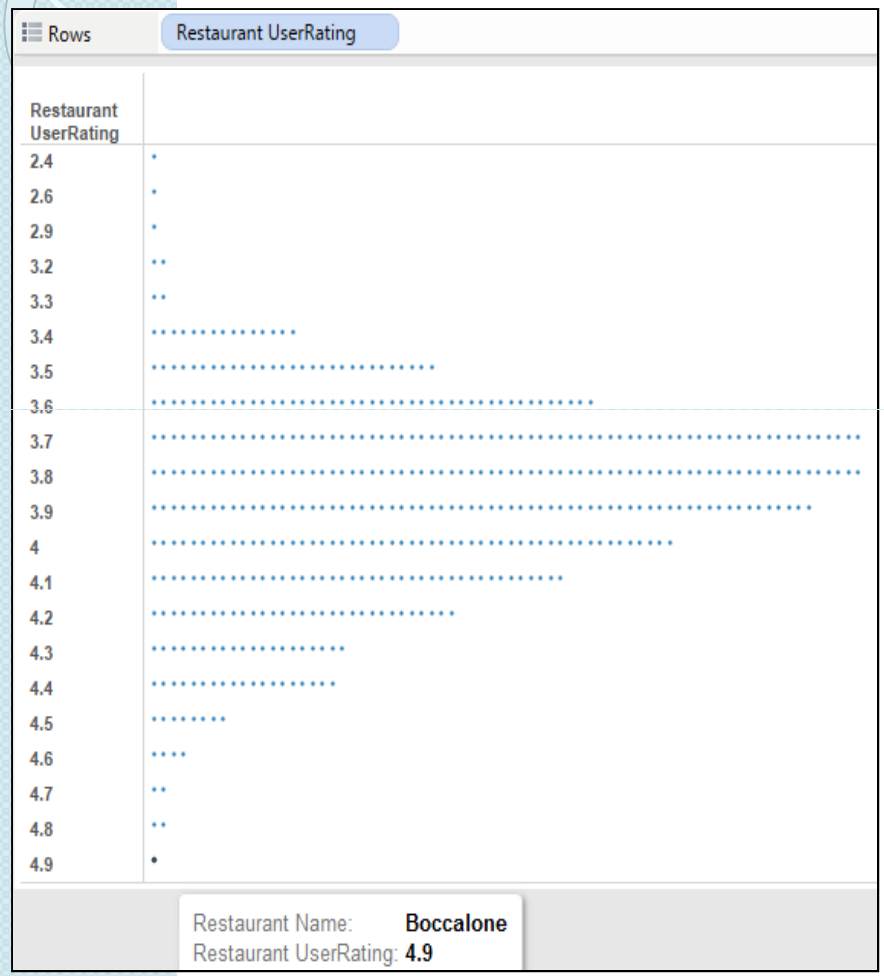
Average Health Inspection rating of restaurants in each grade



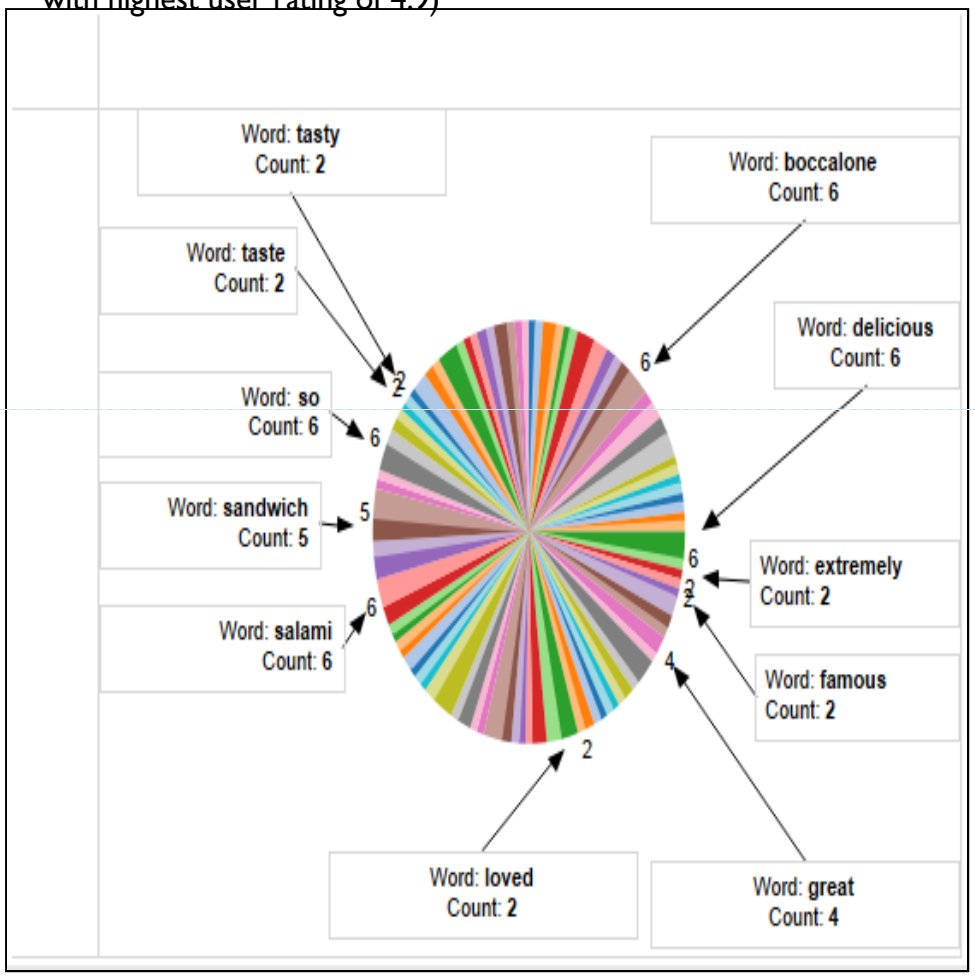
- The average user review rating for A grade restaurants is much higher than C grade restaurants
- The average cost for two in A grade restaurants is much higher than in B and C grade restaurants

Data Exploration – Review Text

The dots represent the restaurants

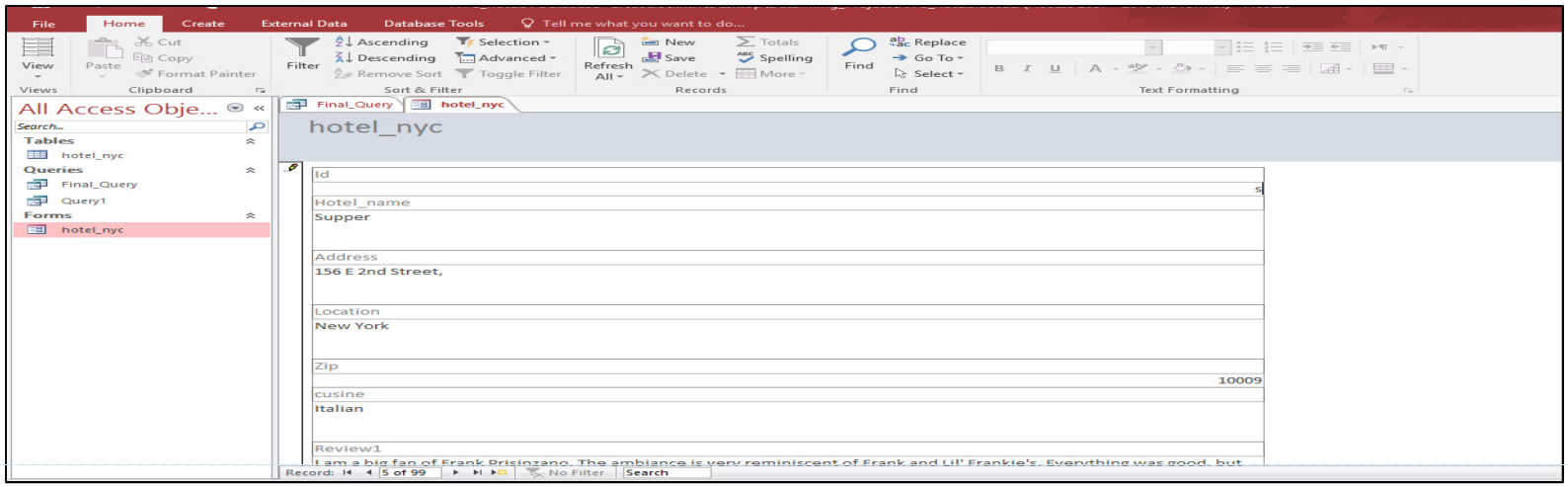


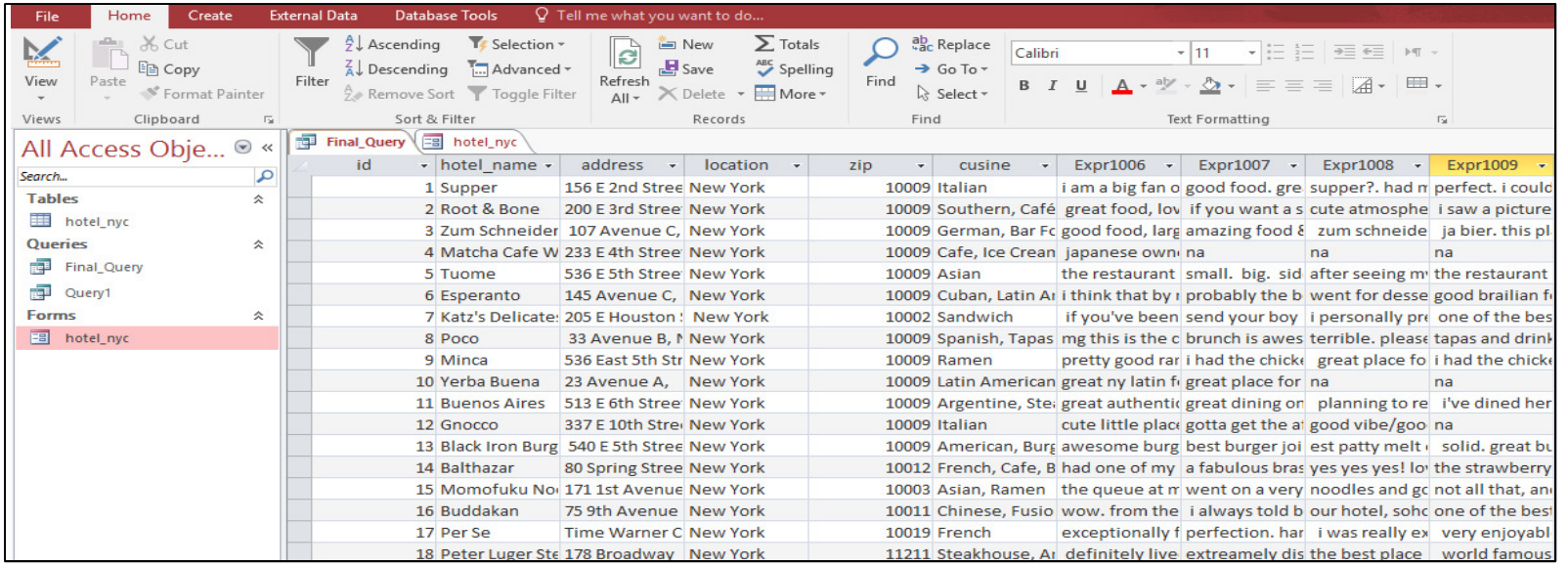
Study of Word count of 15 user reviews of Buccalone, (the restaurant with highest user rating of 4.9)



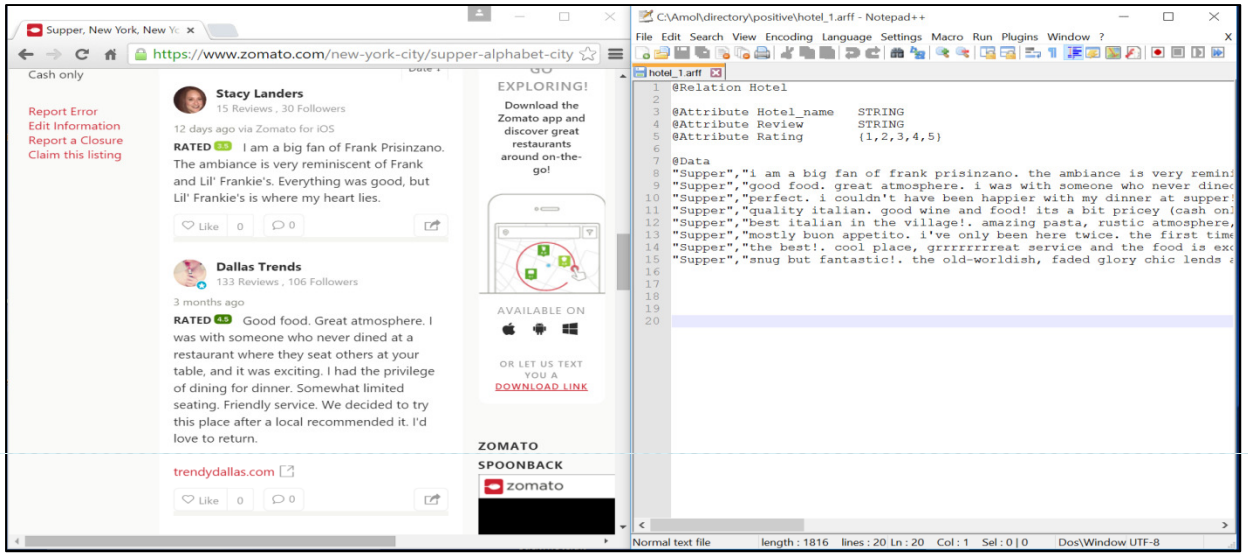
Words adding to positive review of Buccalone : delicious, tasty, taste, famous, great, love

Text Mining - WEKA

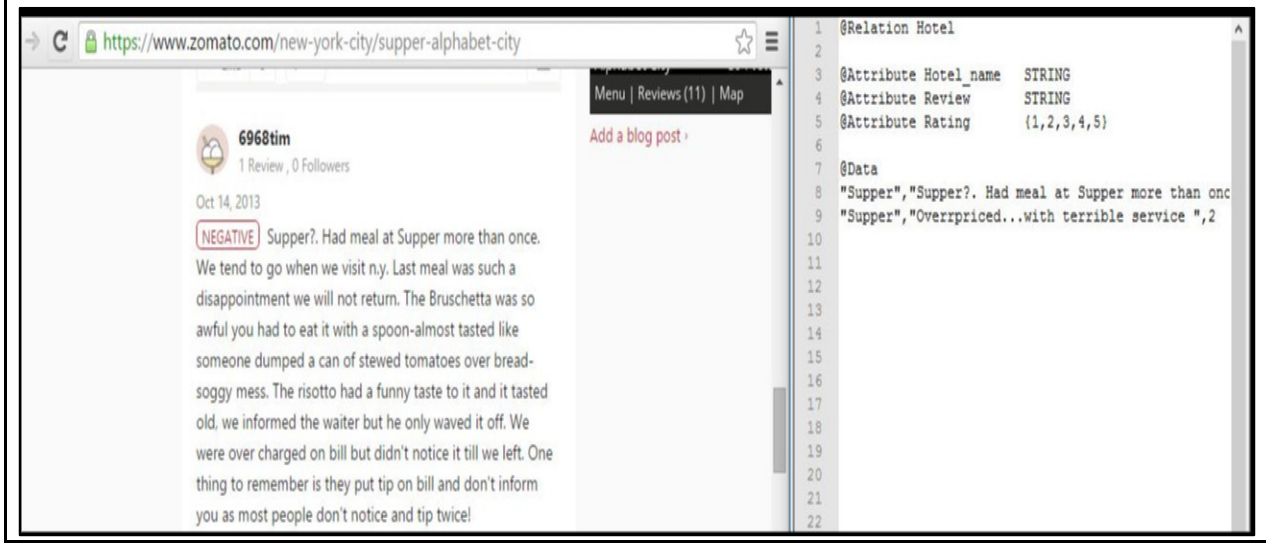




Text Mining - Weka



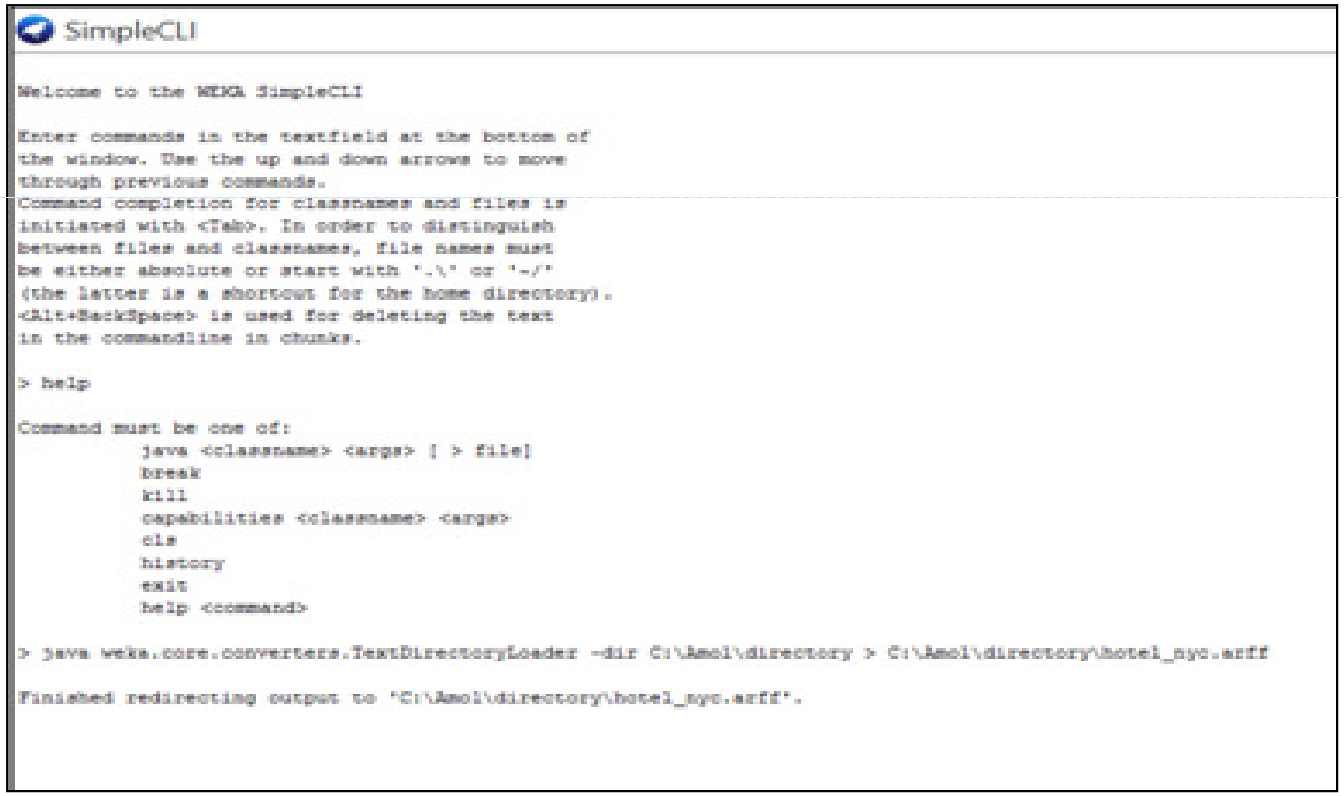
Positive Review arff



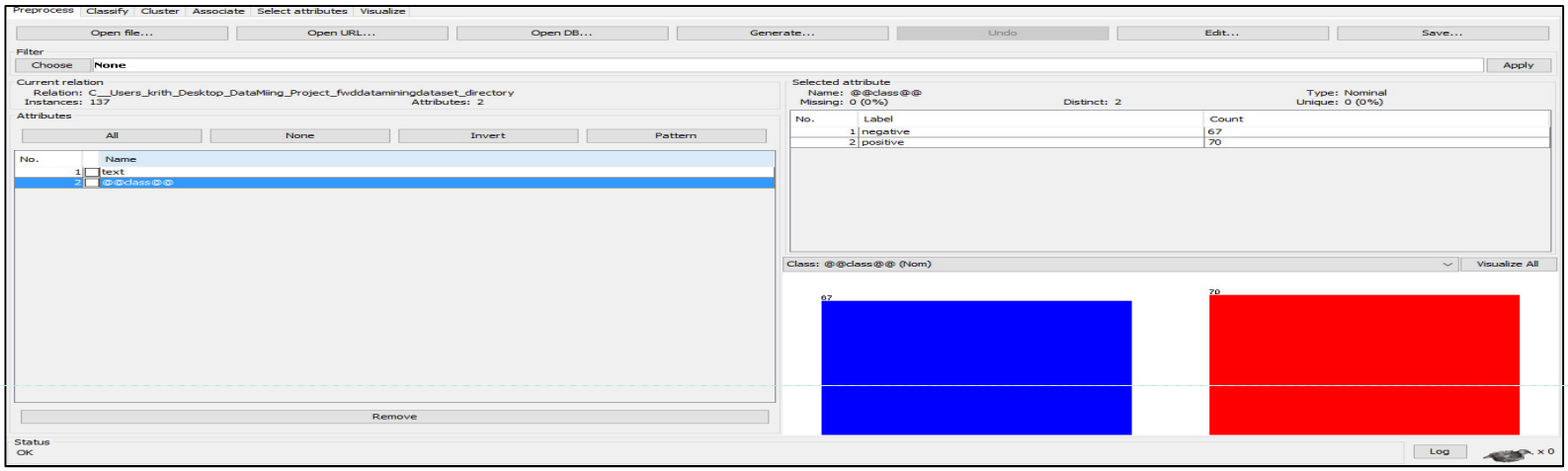
Negative Review arff

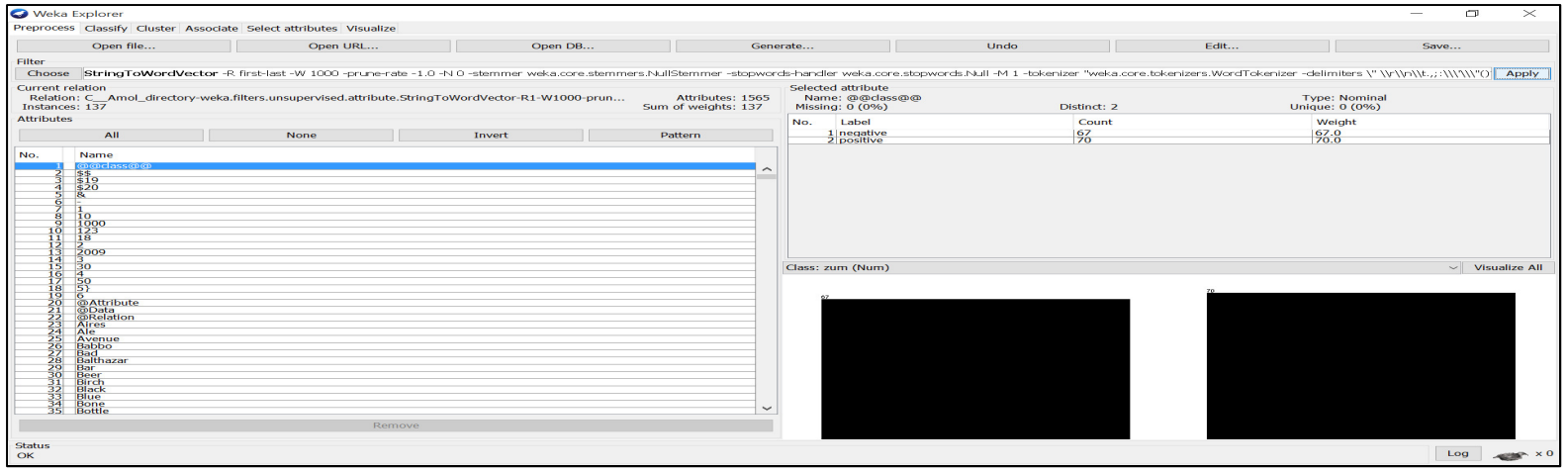
Text Mining - Weka

- CLI (command line interface) script to convert all the arff files into single data

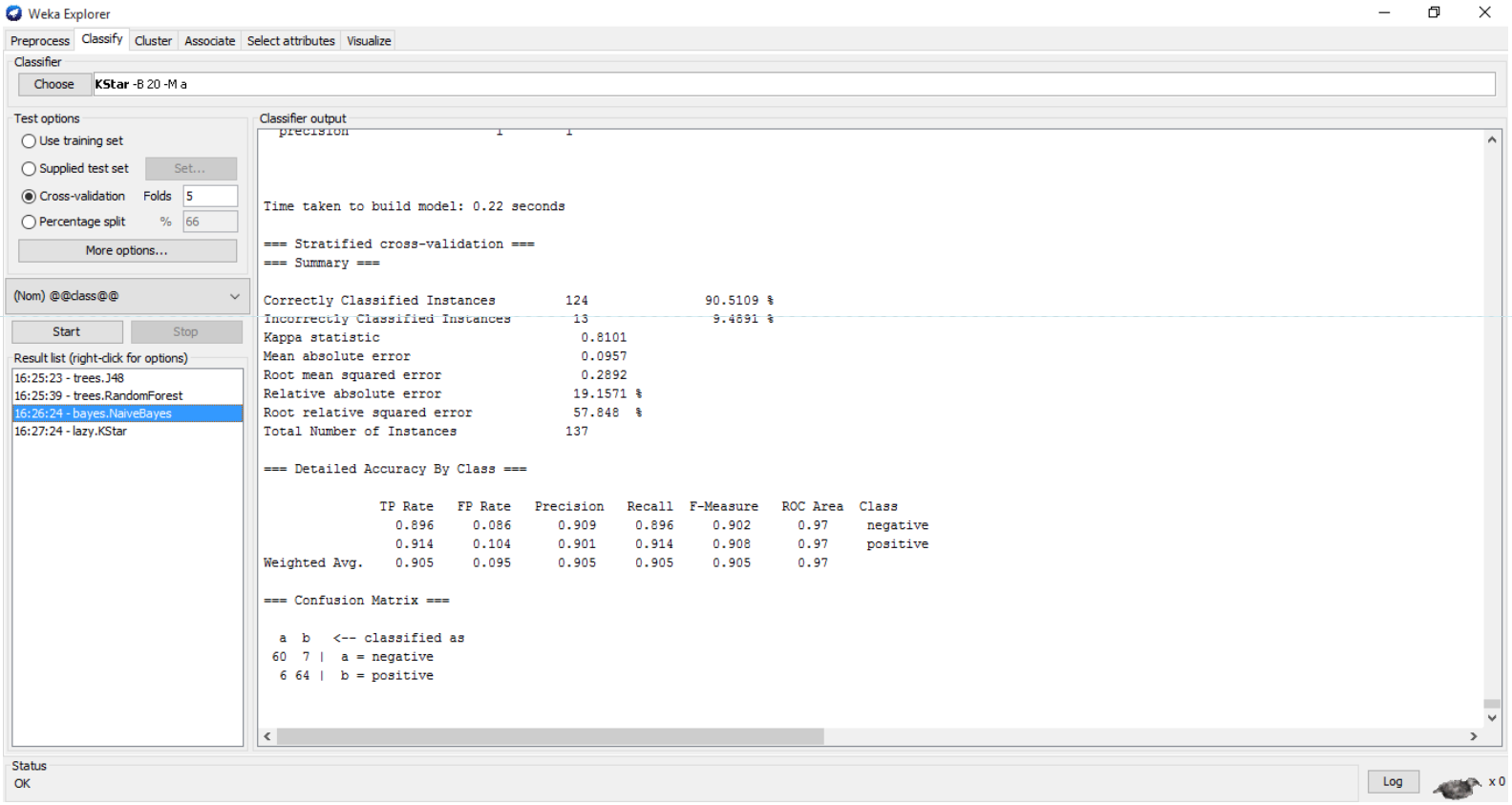


Text Mining - Weka

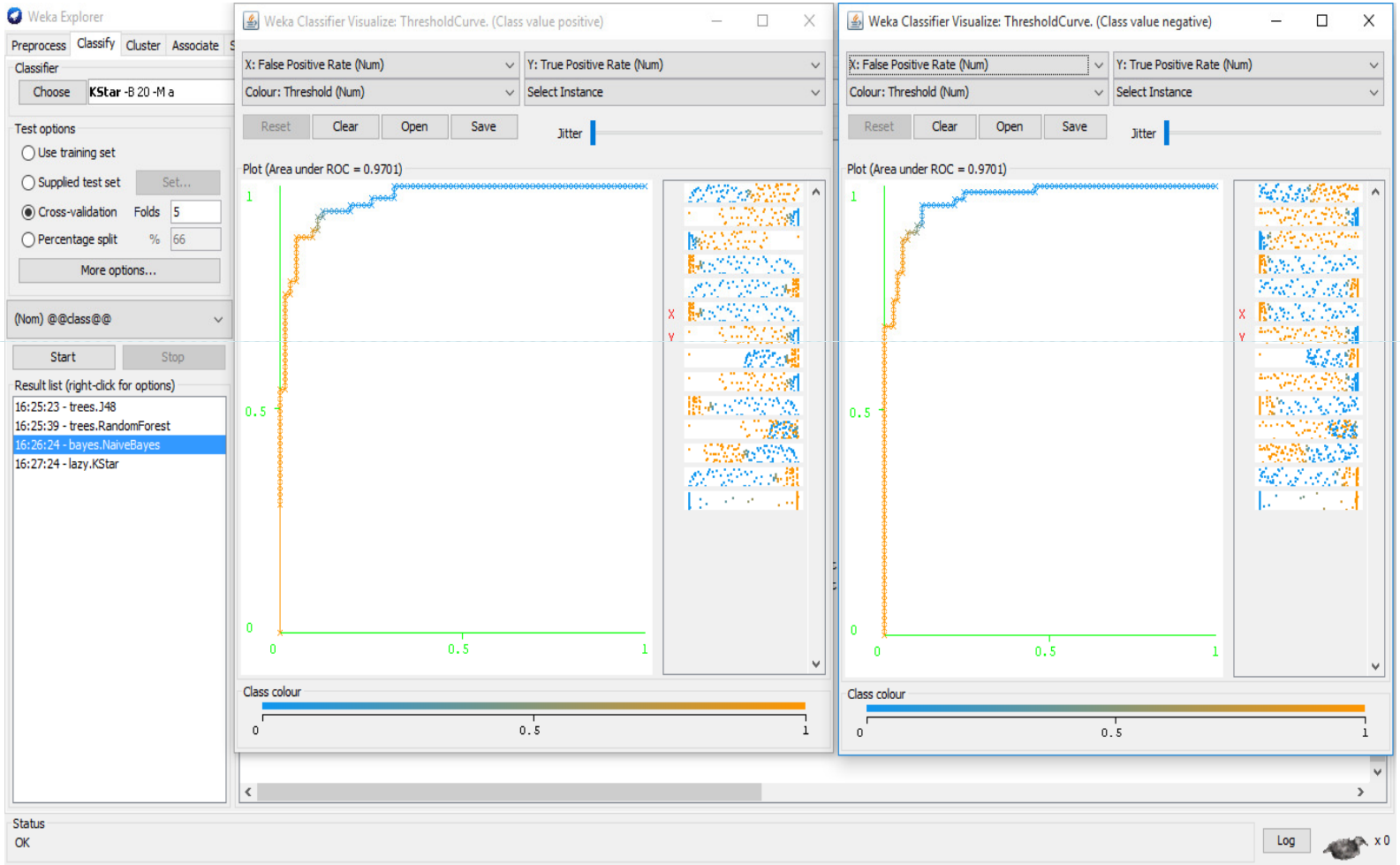




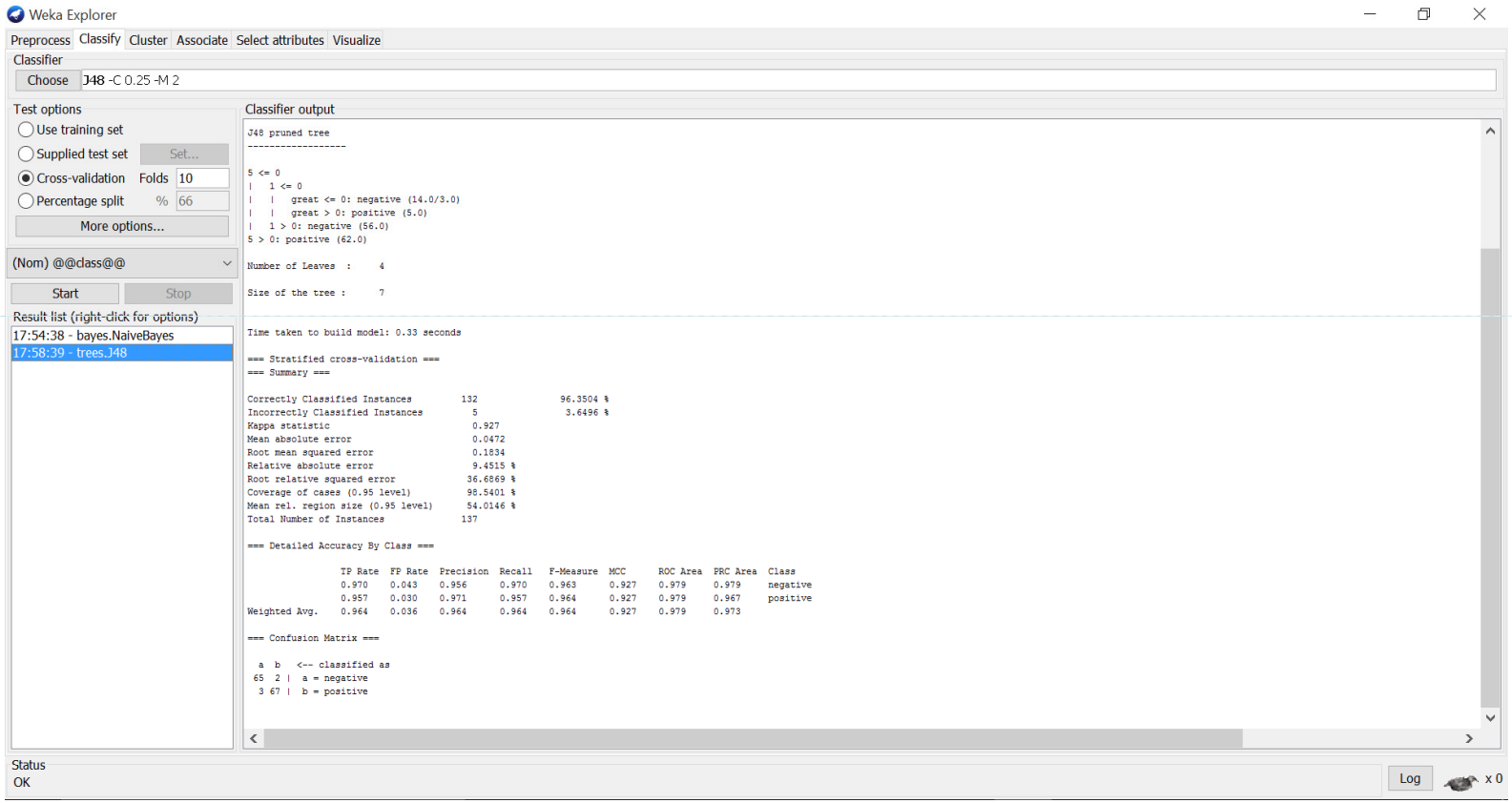
Naïve Bayes Classification - Confusion Matrix



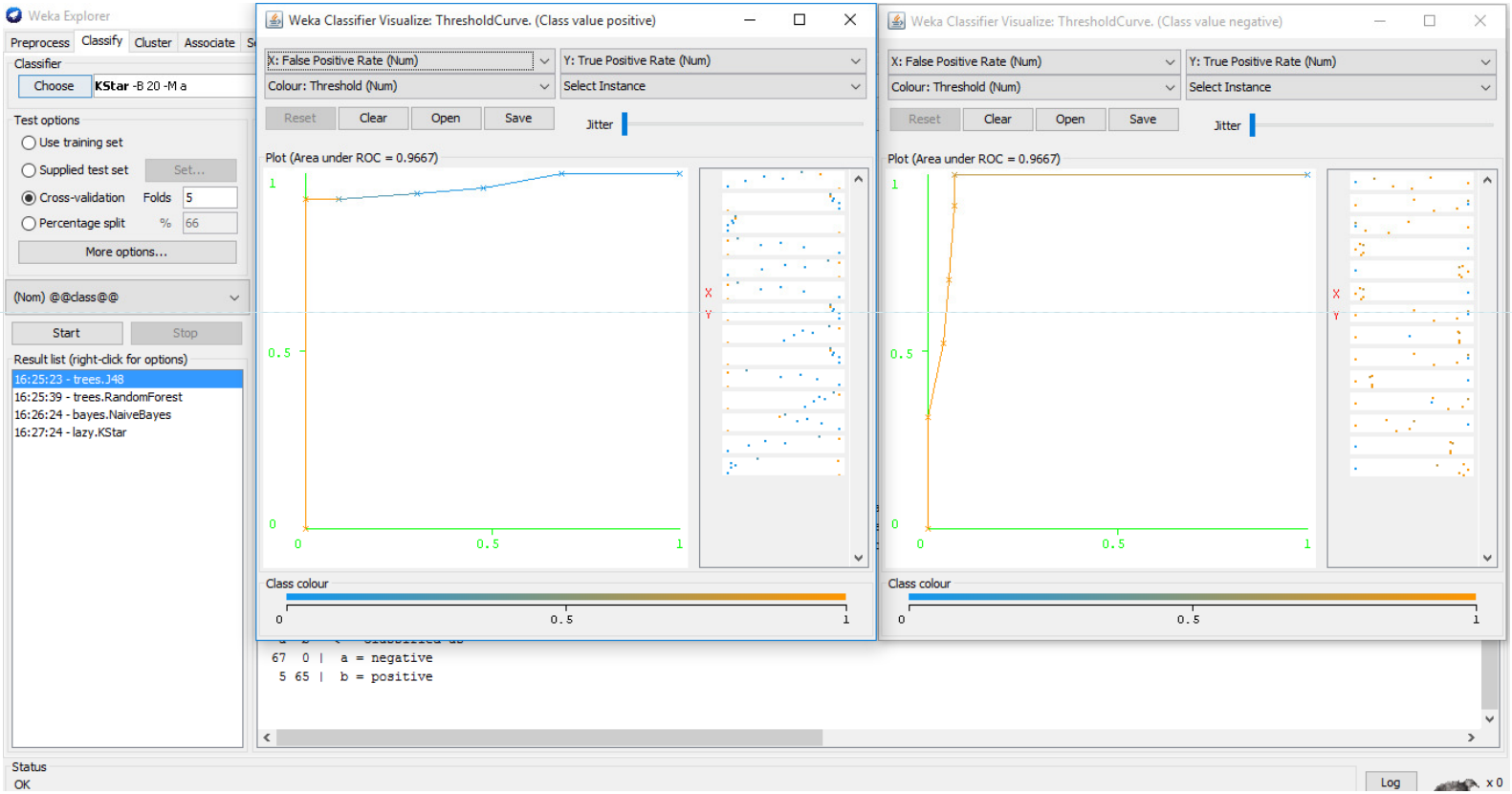
Naïve Bayes Classification - ROC Curve



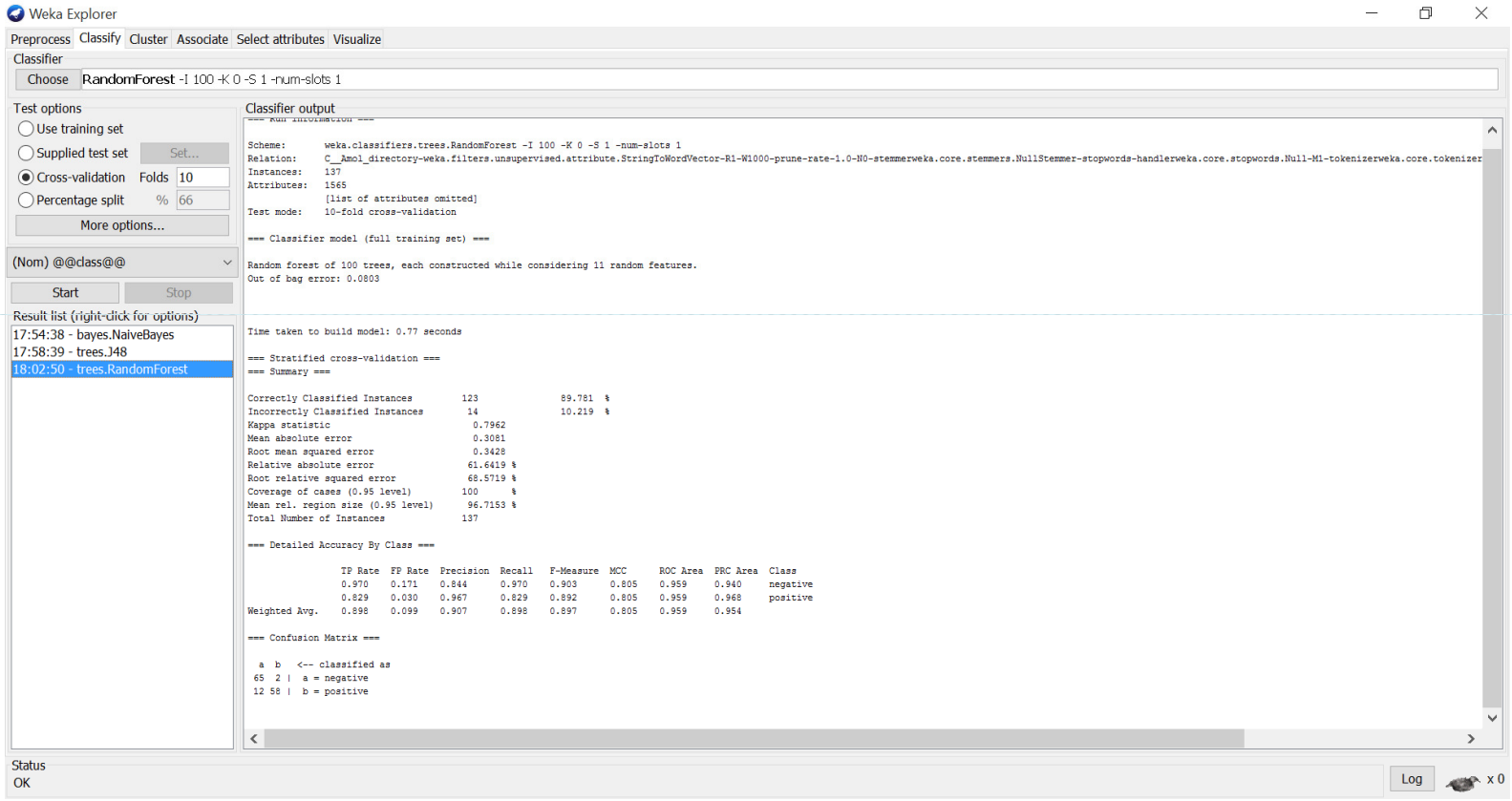
J 48 Classification - Confusion matrix



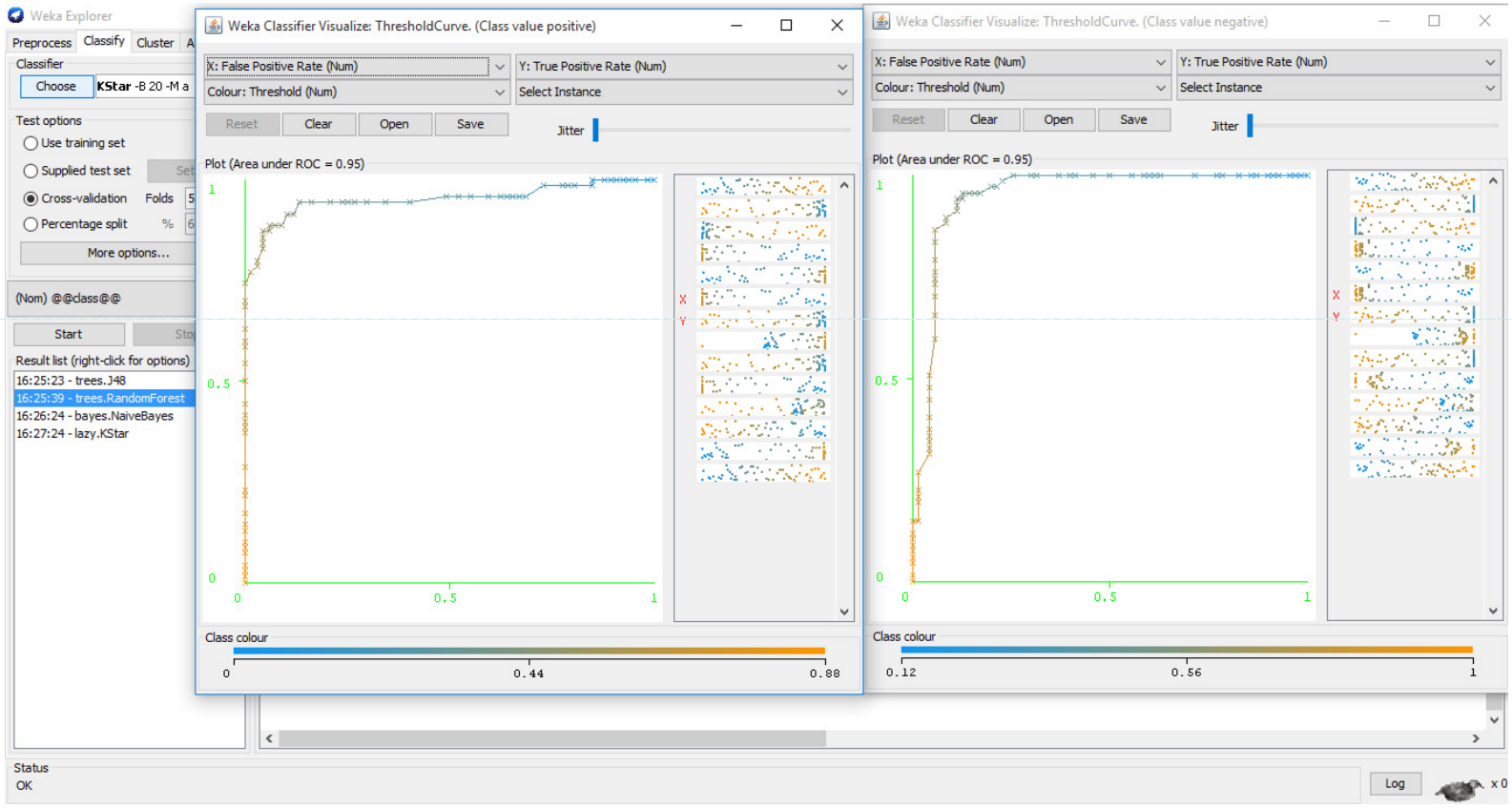
J 48 Classification - ROC Curve



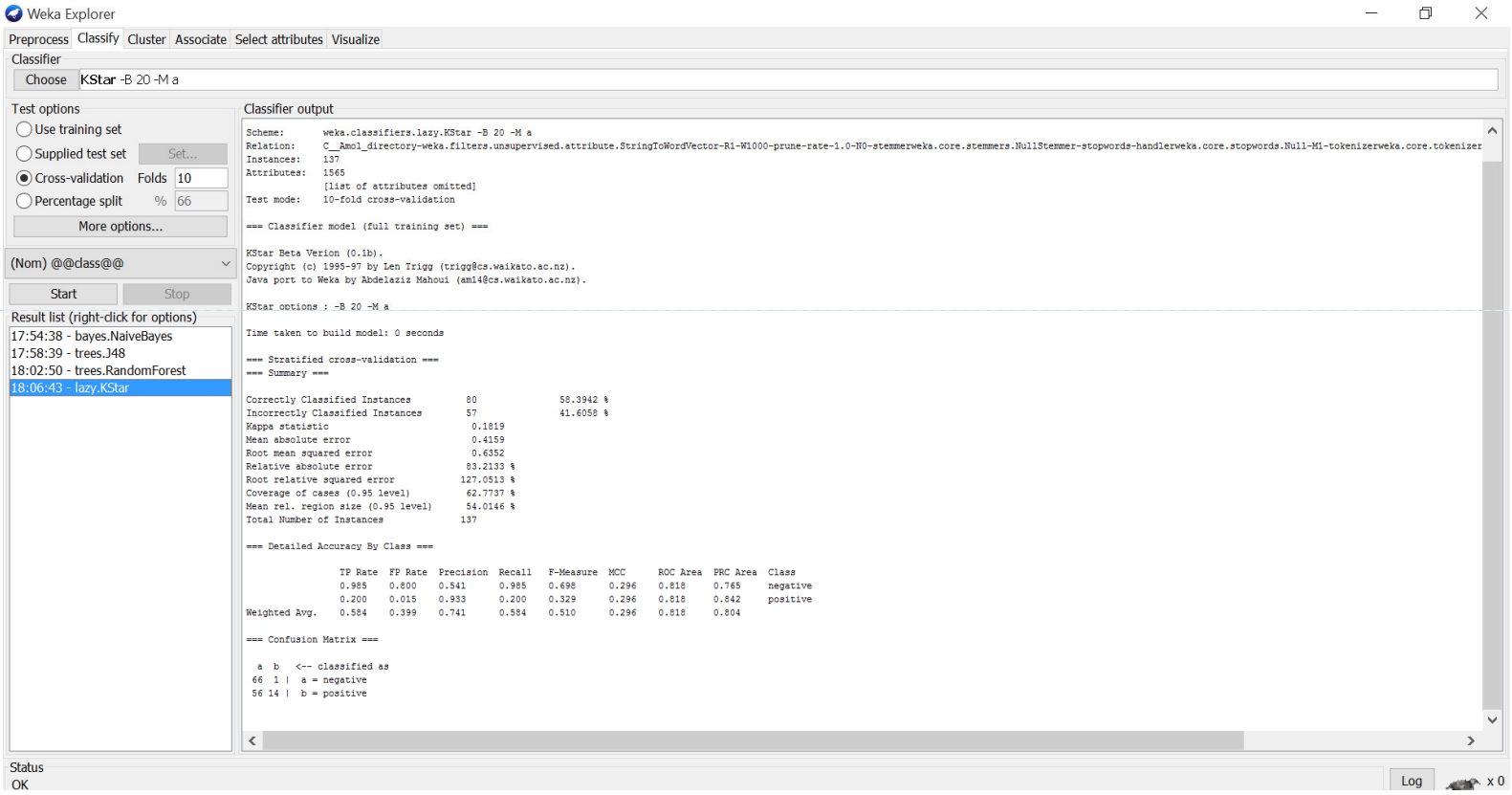
Random Forest Classification - Confusion matrix



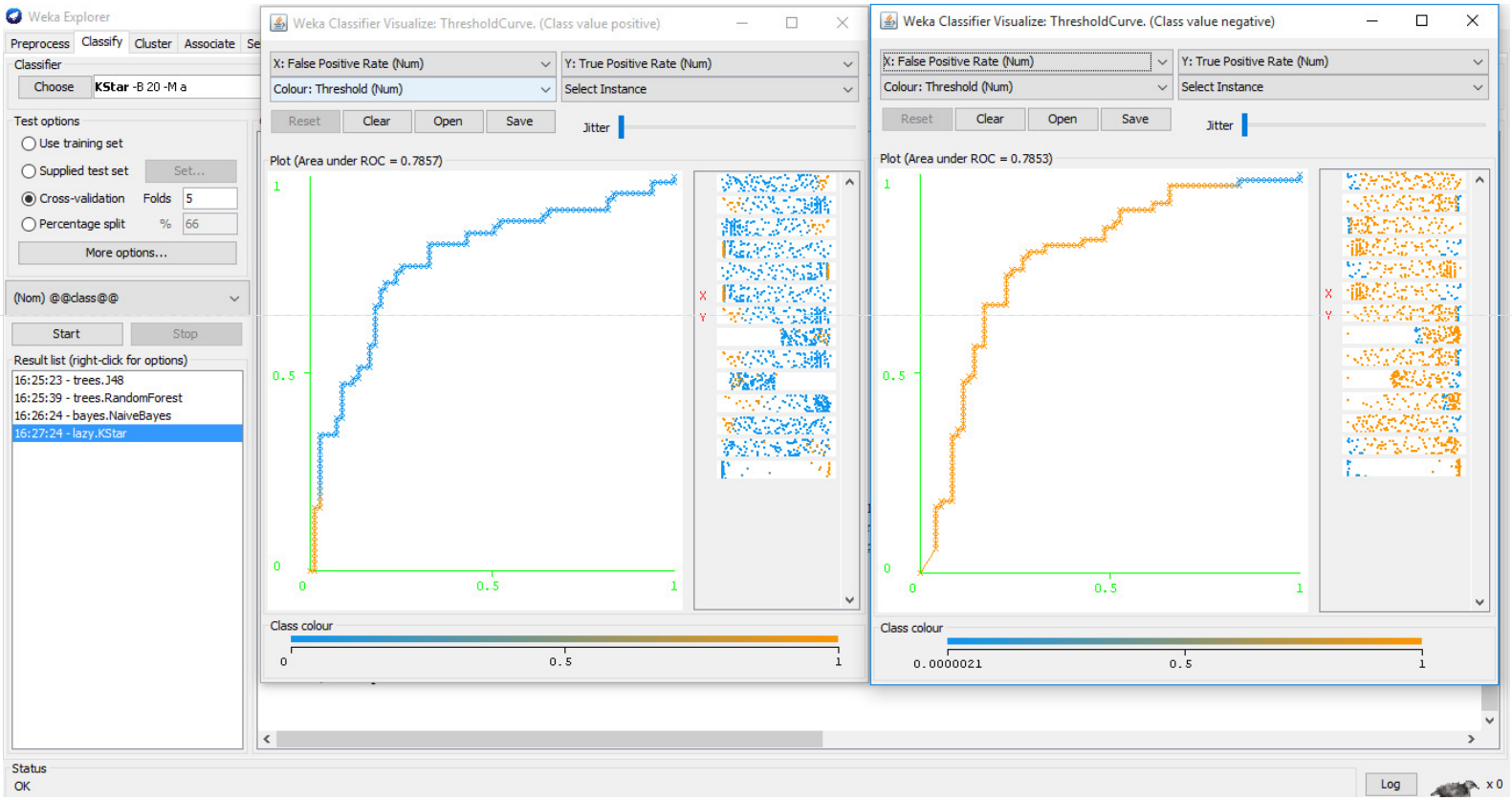
Random Forest Classification - ROC Curve



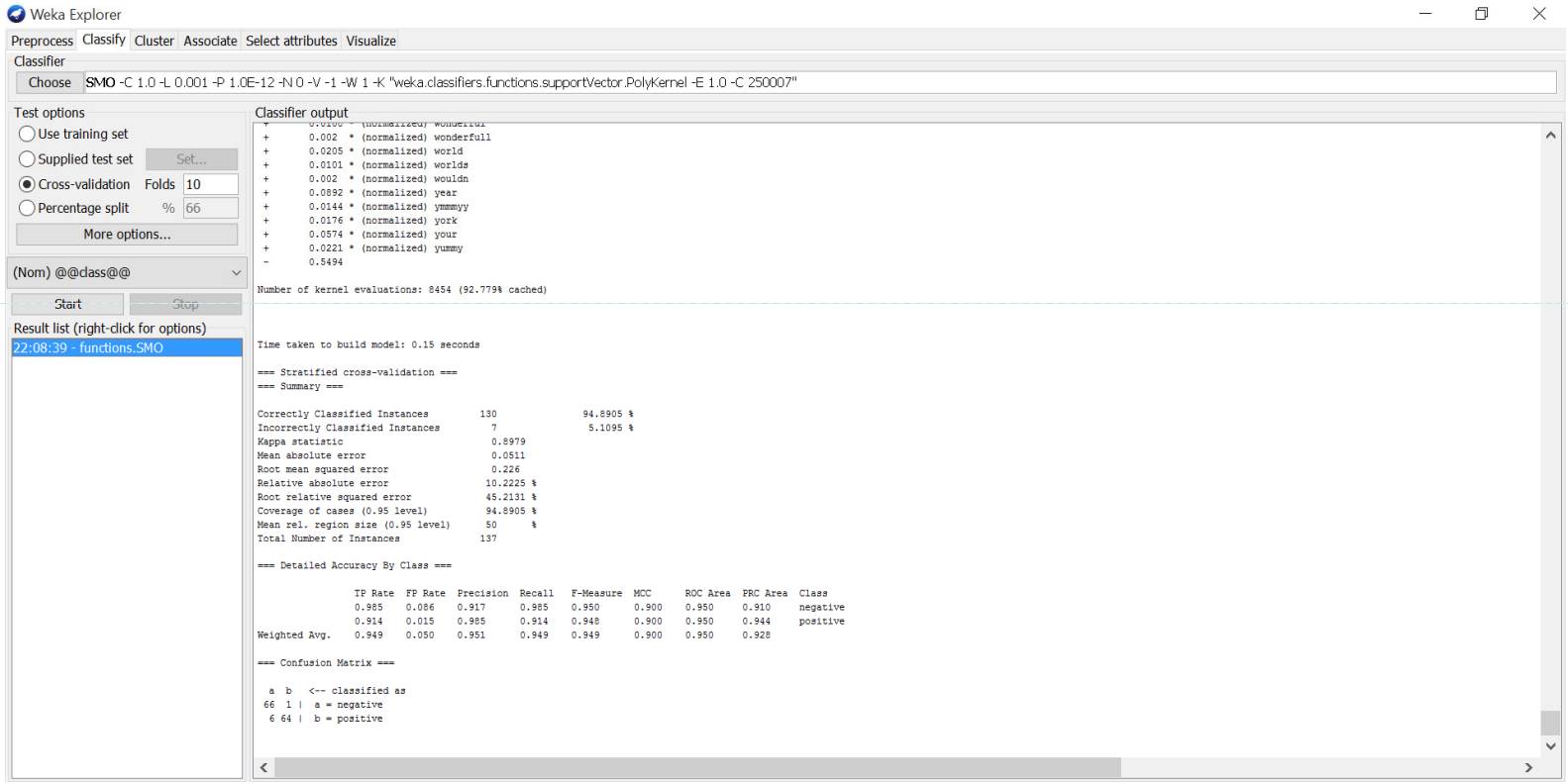
KStar Classification - Confusion matrix



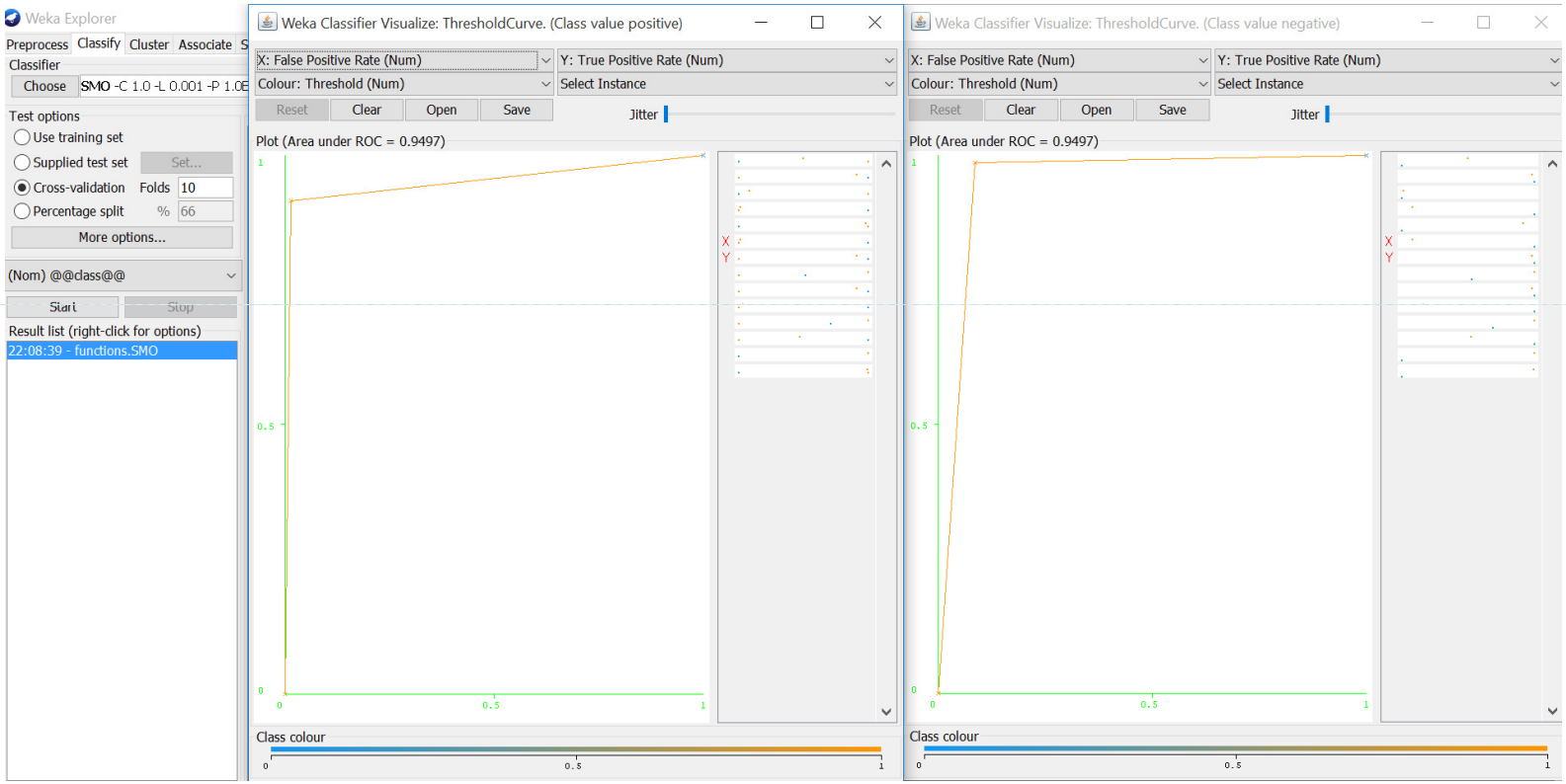
KStar Classification - ROC Curve



Support Vector Machine
Classification - Confusion matrix



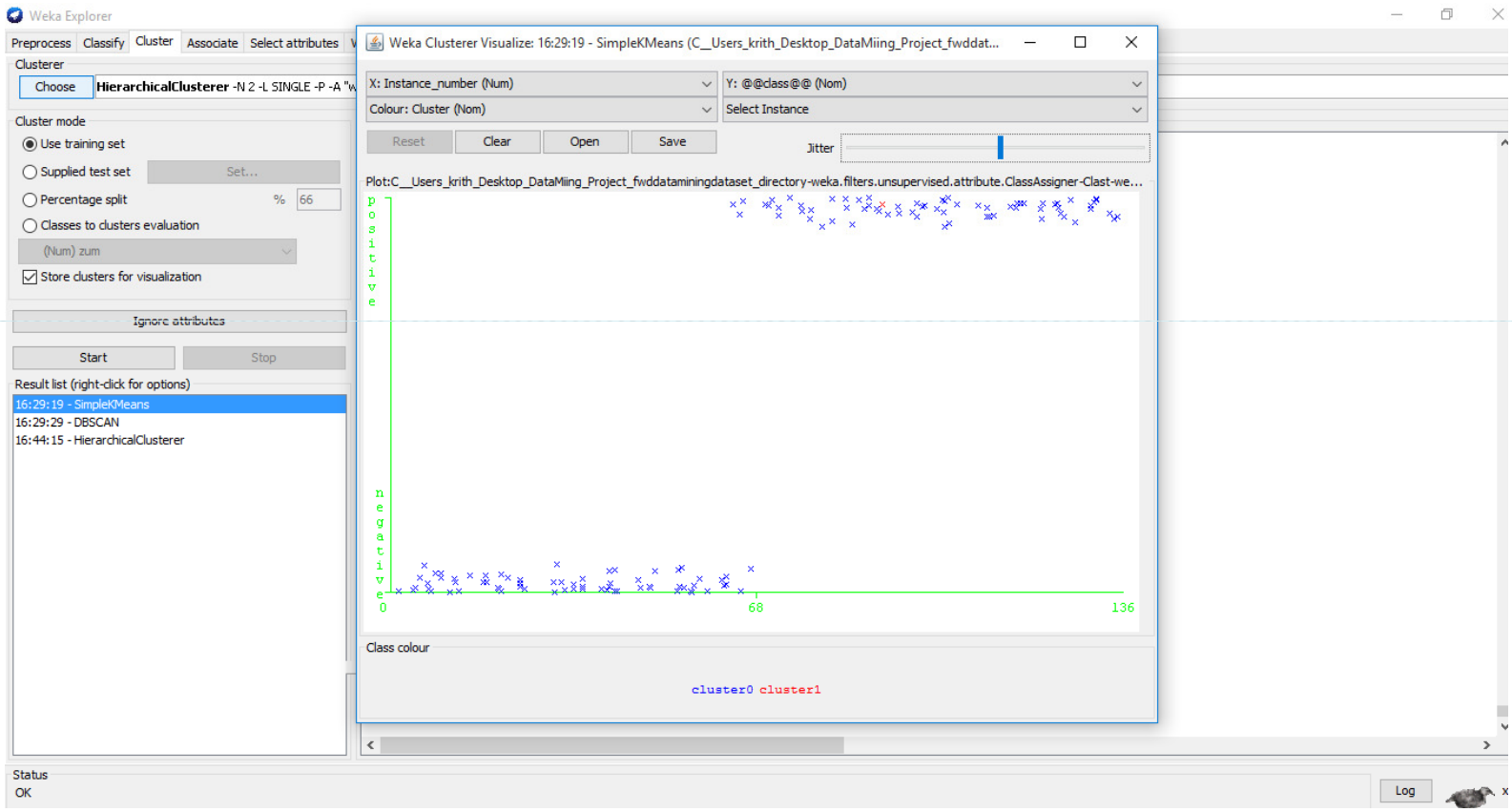
Support Vector Machine Classification - ROC Curve



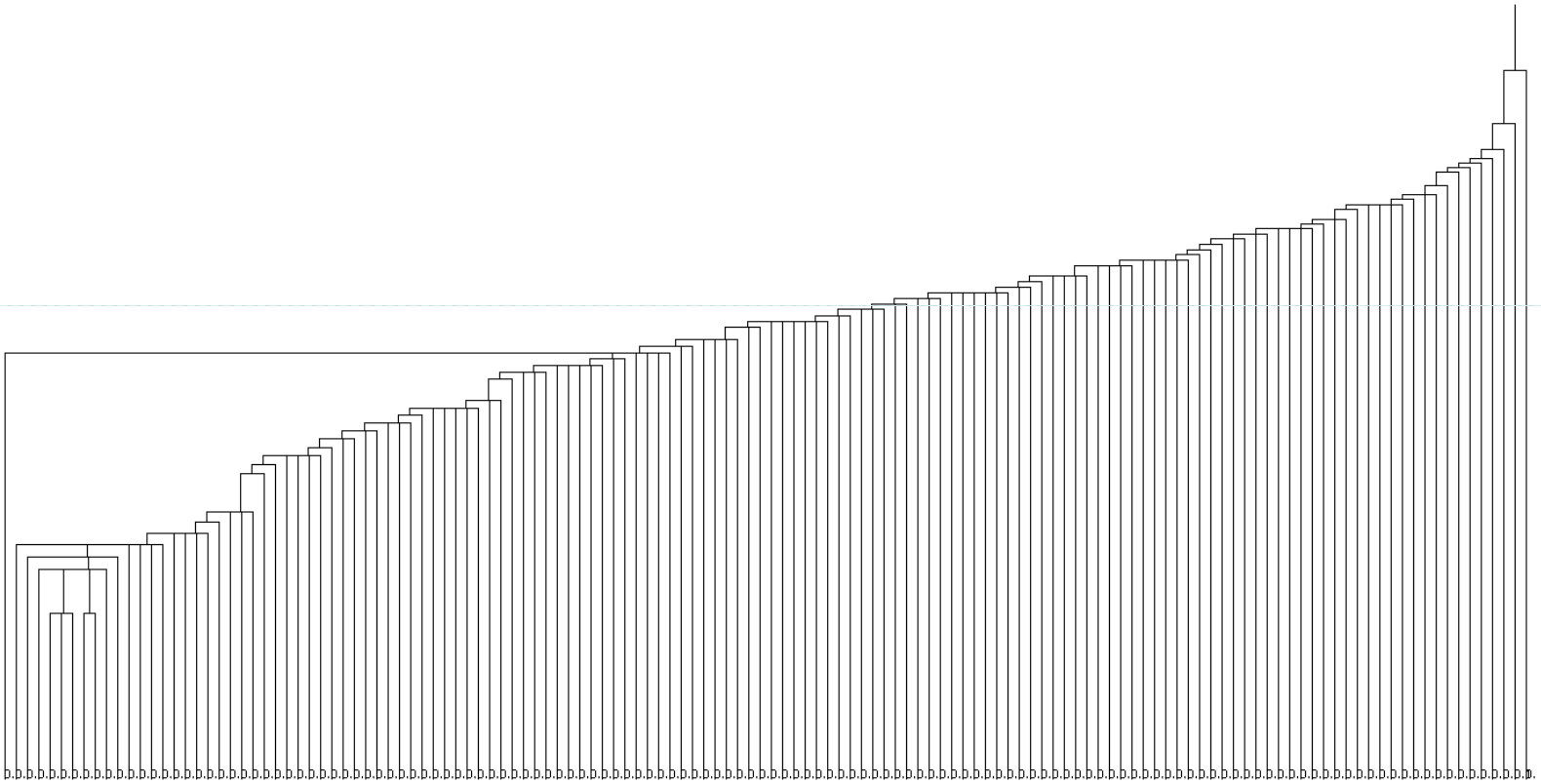
Comparison of Classification Models

S. No.	Naïve Bayes	J 48	Random Forest	K - Star	Support Vector Machine
Area under ROC	0.9701	0.966	0.95	0.78	0.94
Correctly classified Instance (%)	90.5	96.35	89.7	58.3	94.8
Time taken to build model (secs)	0.22	0.33	0.77	0	0.15

Simple K Means Clustering



Hierarchical Clustering



Conclusions

Tableau

- New York City has most expensive restaurants.
- Chicago restaurants have the best user ratings
- The most popular cuisine in our data set is American which is in 45 restaurants, followed by Italian, Mexican, Seafood and Japanese
- The most expensive restaurants in New York City are in Lincoln Square, followed by Flatiron District, Upper East Side, Tribeca and Meat Packing District
- The average user review rating for A grade restaurants is much higher than C grade restaurants
- The average cost for two in A grade restaurants is much higher than in B and C grade restaurants

Weka

- J48 is the best classification model with less false positive rate.
- SVM is the fastest classification model with less misclassification error.

THANK YOU

QUESTIONS??