

DATA MINING

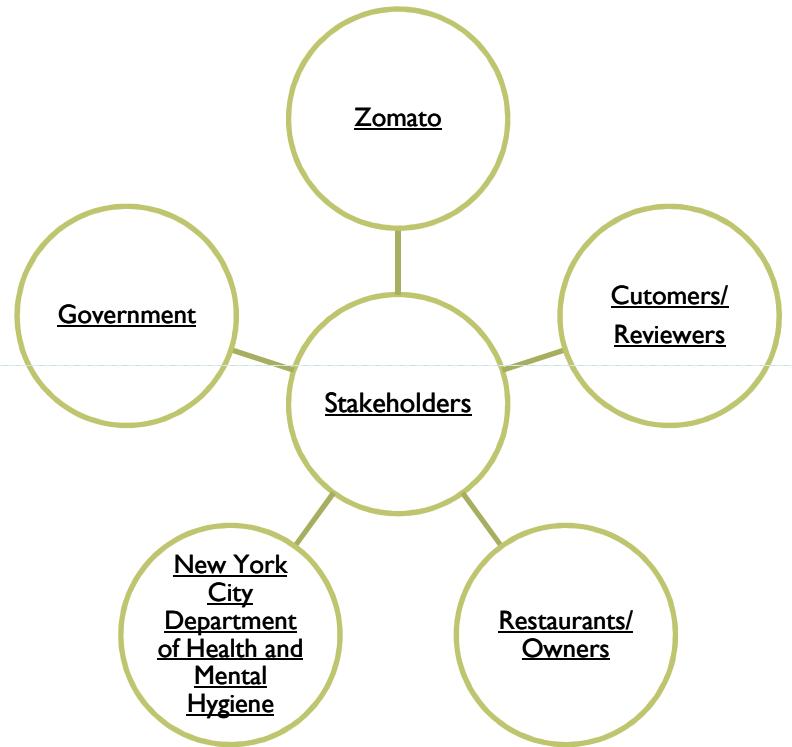
Study of Restaurants in US



- 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_UserId	Restaurant_RatingVotes	Health Inspection
16781904	Momofuku	https://www.zomato.com/momofuku	171 1st Ave	East Village	New York City	10003	Asian, Ramen	60	4.1	Excellent	1530	A
16767139	Gramercy Tavern	https://www.zomato.com/gramercy-tavern	42 E 20th St	Union Square	New York City	10003	American	160	3.7	Very Good	1754	A
16760100	Balthazar	https://www.zomato.com/balthazar	80 Spring St	Soho	New York City	10012	French, Cafe	140	3.7	Very Good	4005	A
16775039	Peter Luger	https://www.zomato.com/peter-luger	178 Broadway	Williamsburg	New York City	11211	Steakhouse, BBQ	150	3.8	Very Good	2091	A
16783153	Shake Shack	https://www.zomato.com/shake-shack	366 Columbus	Upper West Side	New York City	10024	American, Burgers	30	4.2	Excellent	1524	A
16783998	The Halal Guys	https://www.zomato.com/the-halal-guys	6th Avenue	Theater District	New York City	10019	Middle Eastern	25	4.4	Excellent	493	A
16761344	Buddakan	https://www.zomato.com/buddakan	75 9th Avenue	Meatpacking	New York City	10011	Chinese, Fusion	150	3.9	Very Good	1483	A
16761402	Burger Joint	https://www.zomato.com/burger-joint	Le Parker Meridien	Theater District	New York City	10019	American, Burgers	25	4.1	Excellent	1381	A
16785398	Shake Shack	https://www.zomato.com/shake-shack	691 8th Avenue	Hell's Kitchen	New York City	10036	American, Burgers	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

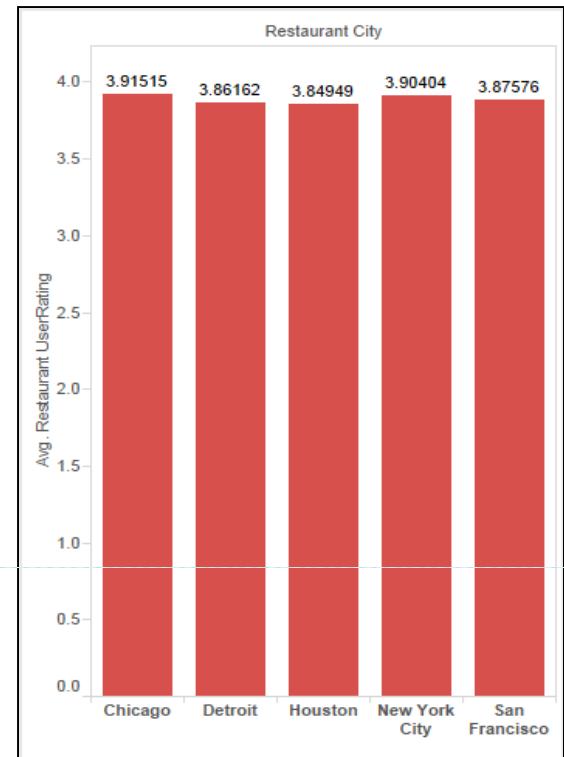
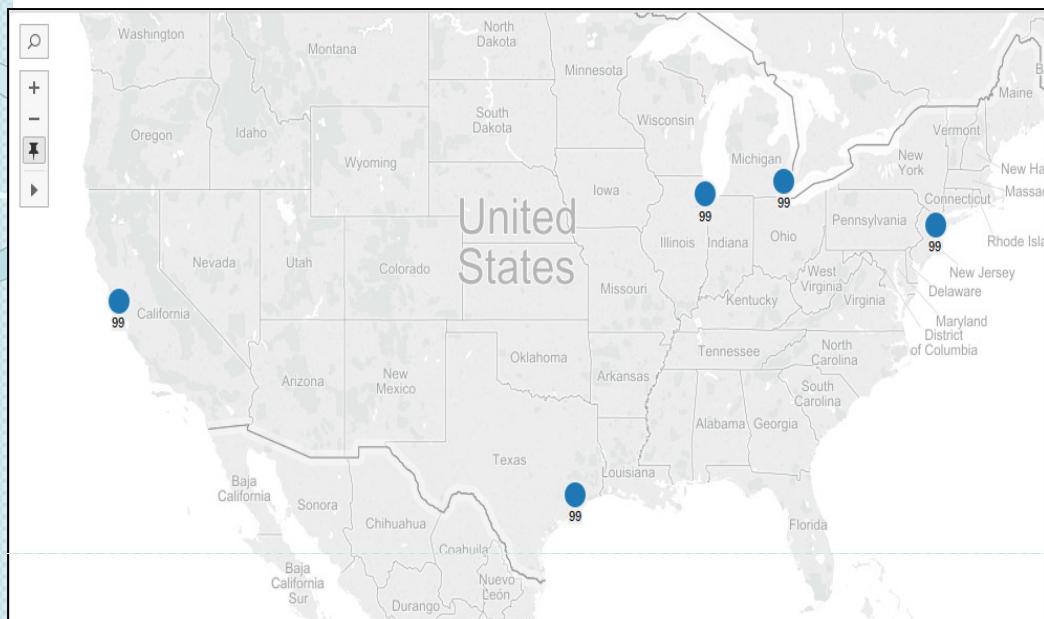
Combined the data set to obtain a collective data of 495 restaurants for exploration in Tableau and Weka

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

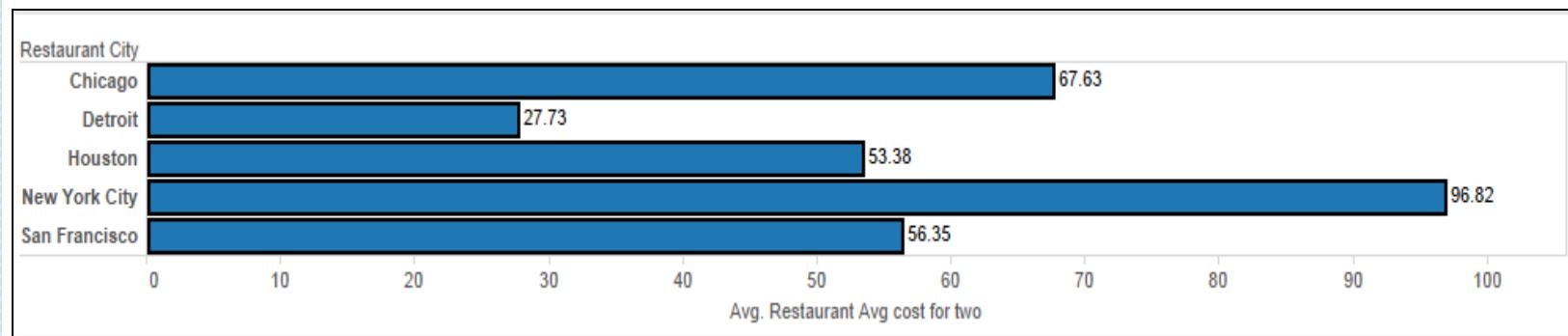
For New York City hotels, manually collected the New York City Department of Health and Mental Hygiene Inspection Results Rating and User reviews of the Hotels(1545 reviews)

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

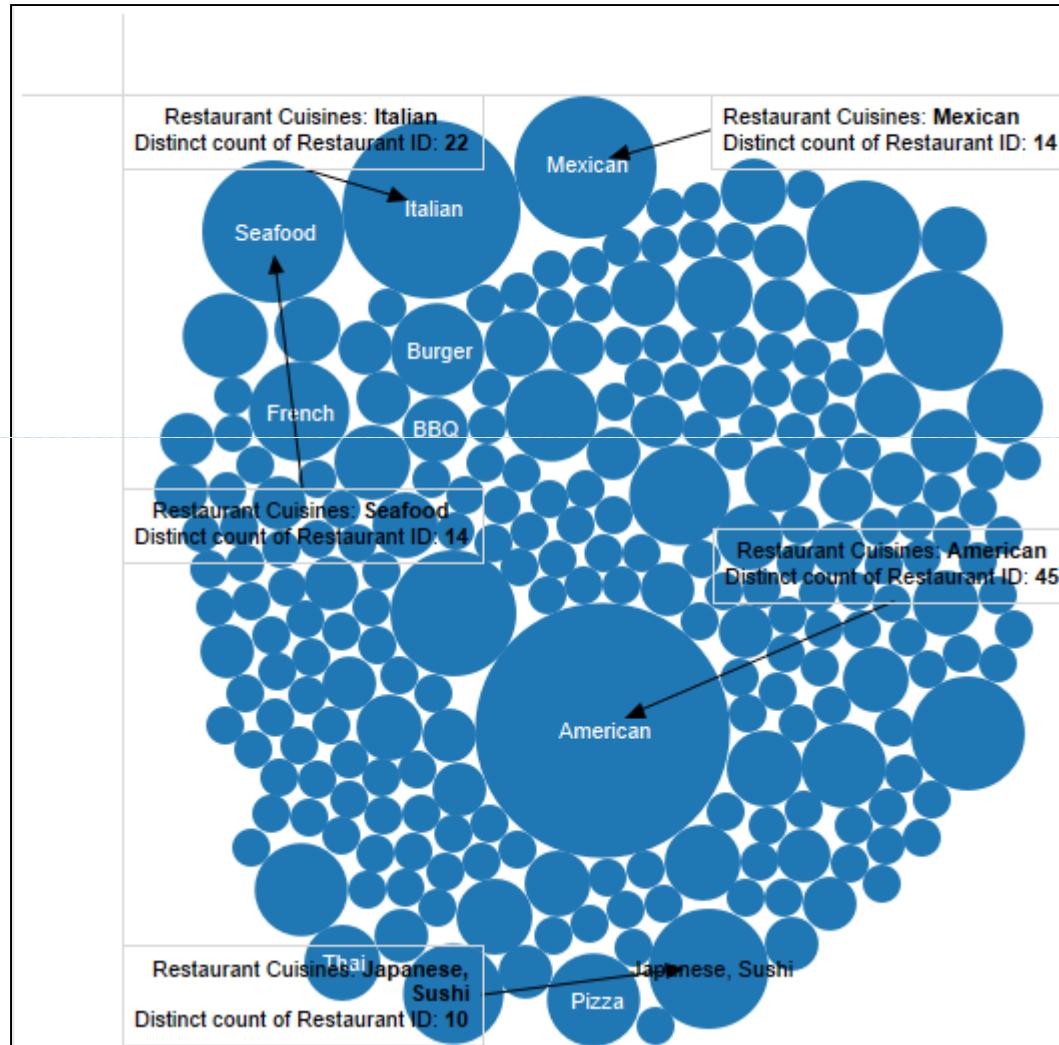


New York City has most expensive restaurants.
Chicago restaurants have the best user ratings



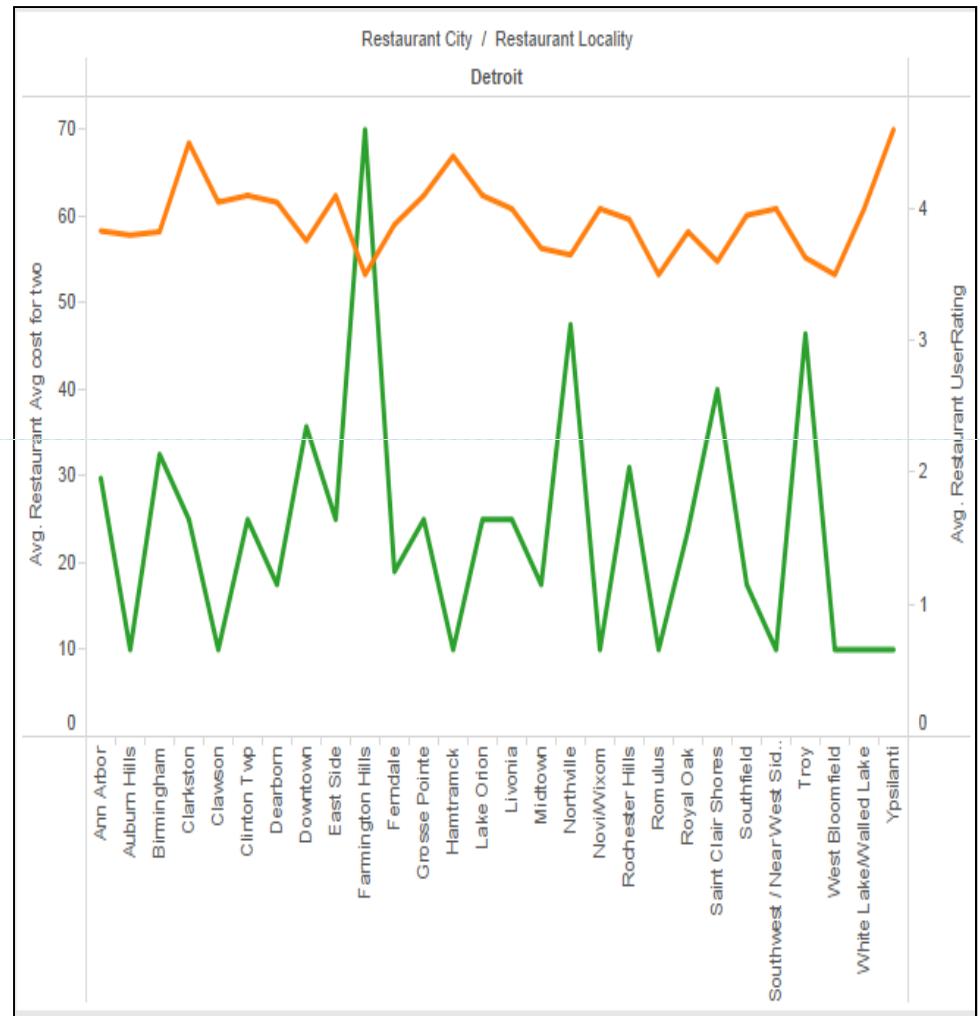
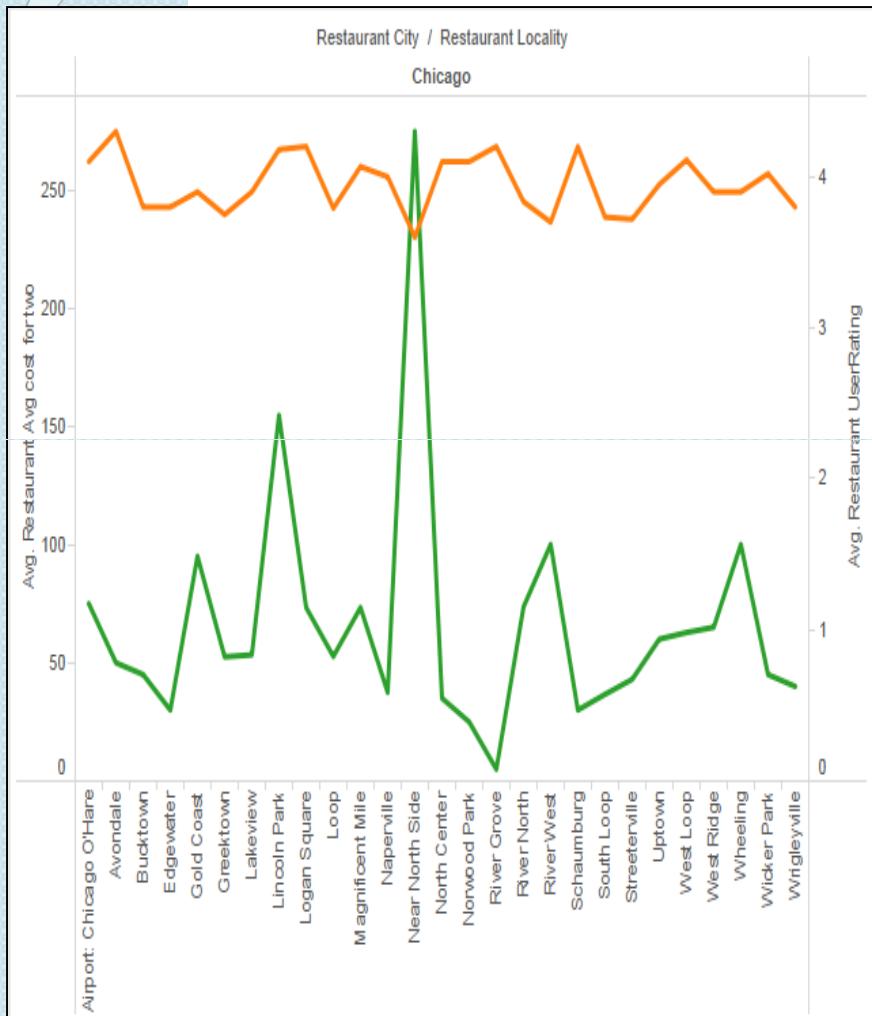
Cuisines Analysis

Count of Restaurants for various cuisine types

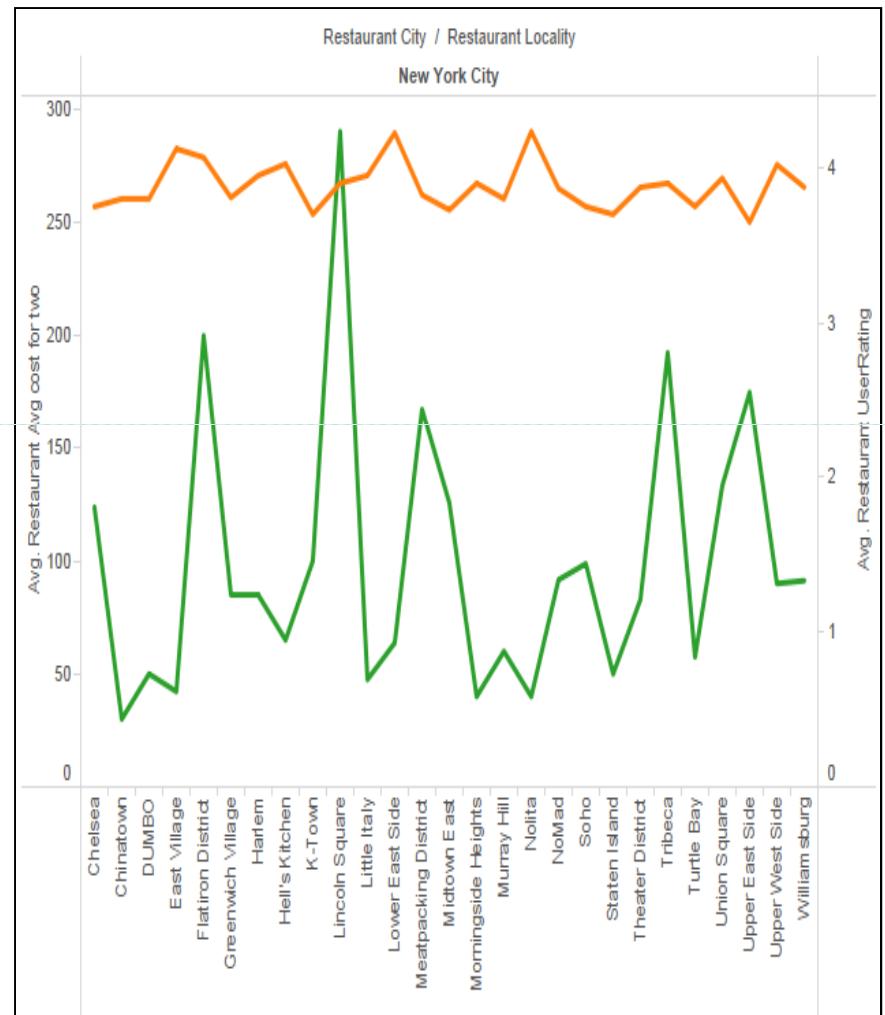
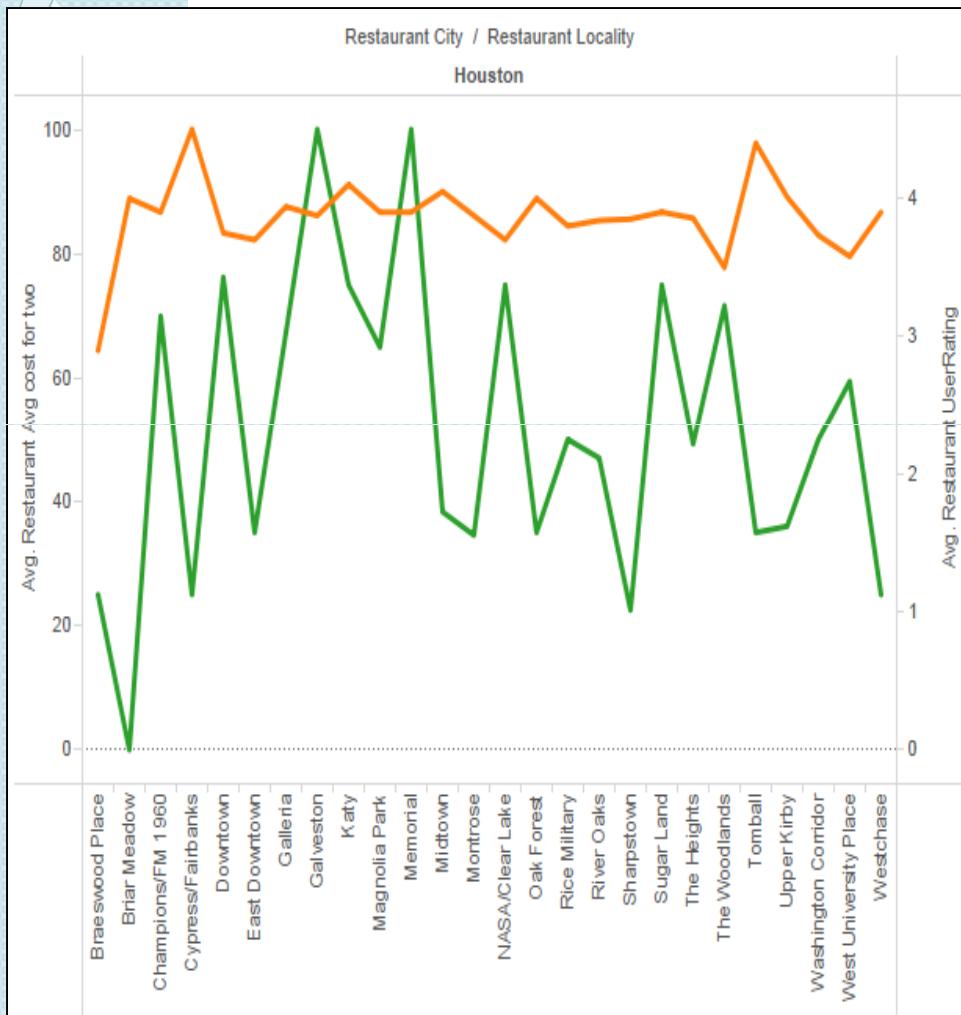


The most popular cuisine in our data set is American which is in 45 restaurants, followed by Italian, Mexican, Seafood and Japanese

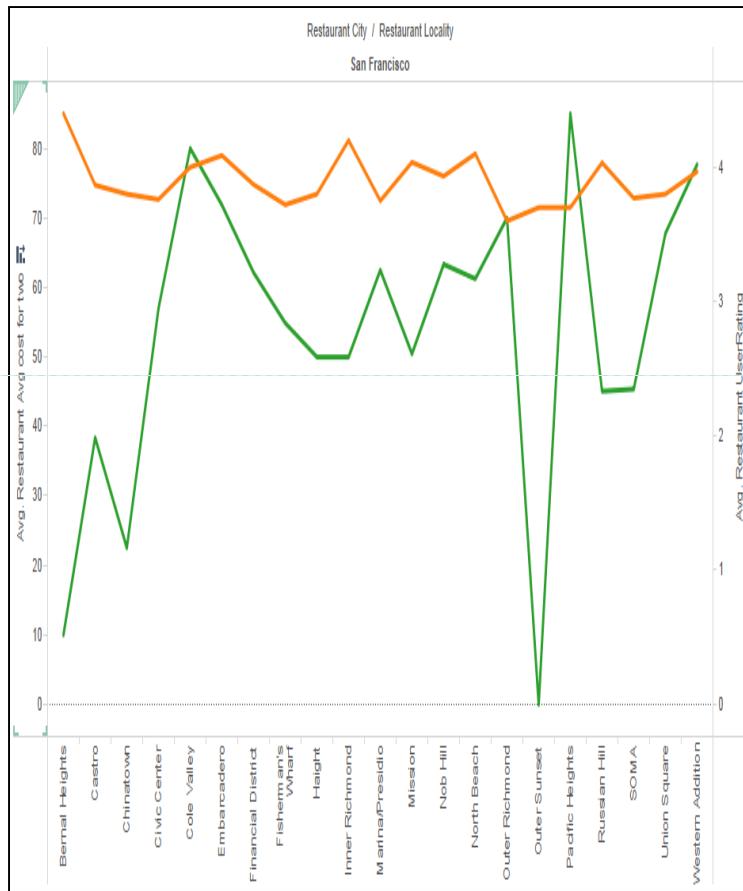
Study of relationship between Average Cost for two and Average User Rating of a Restaurant



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



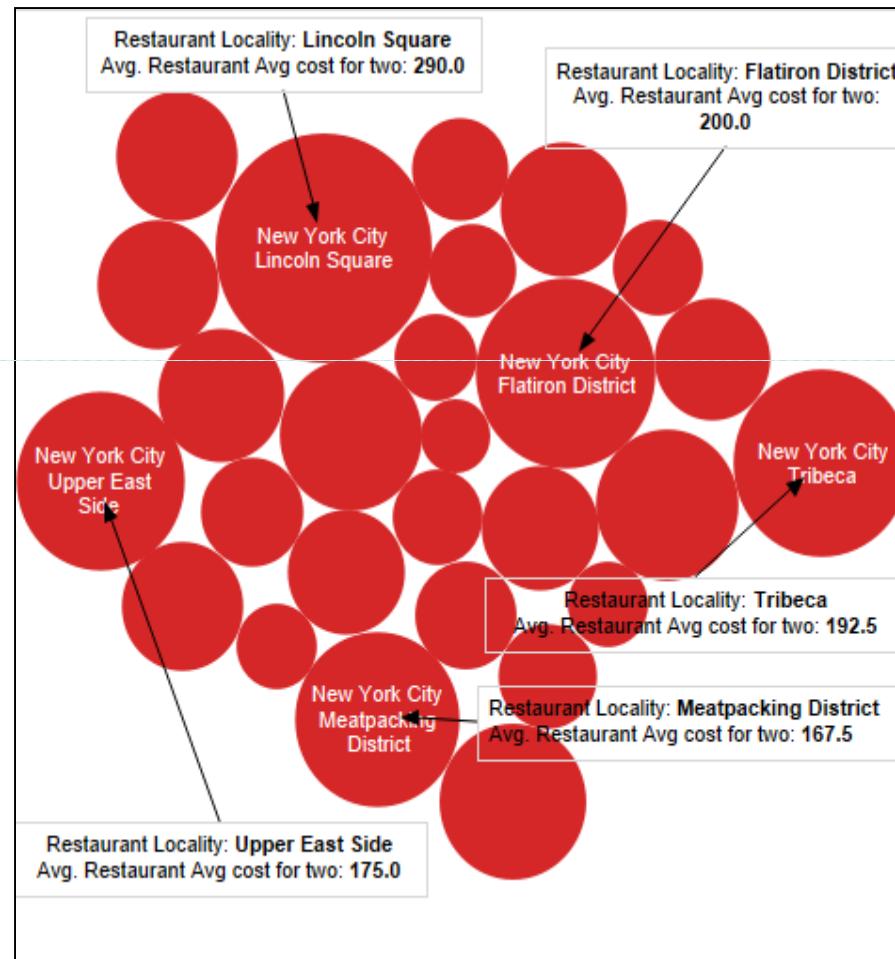
Study of relationship between Average Cost for two and Average User Rating of a Restaurant



Pearson Correlation Coefficient between the entire data set of Average Cost for two and User review rating is -0.09466647716057779 which mean that they are probably not correlated.

New York City – An Exploration

Locality wise – Average cost of two people

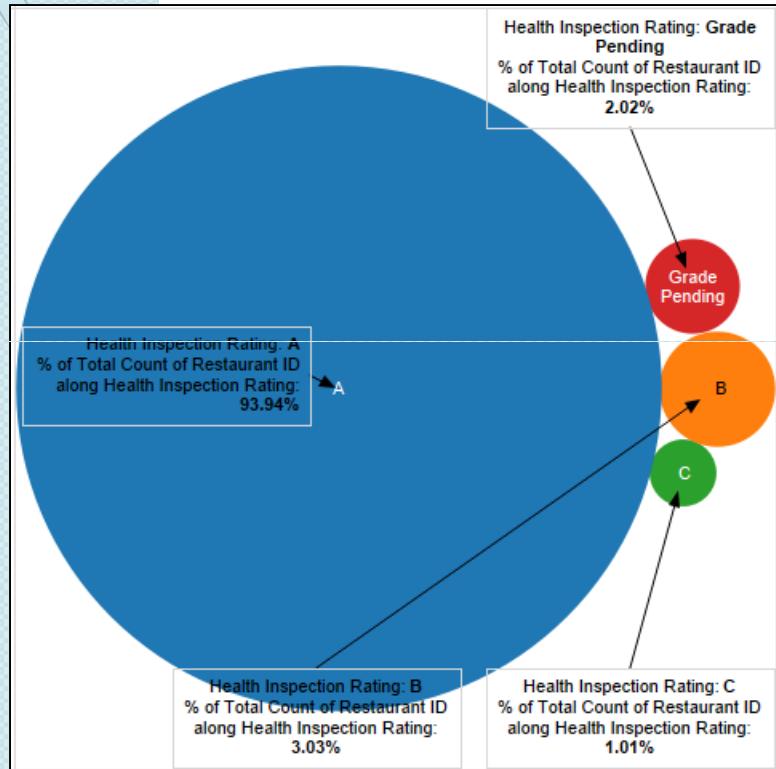


 The most expensive restaurants in New York City are in Lincoln Square, followed by Flatiron District, Upper East Side, Tribeca and Meat Packing District

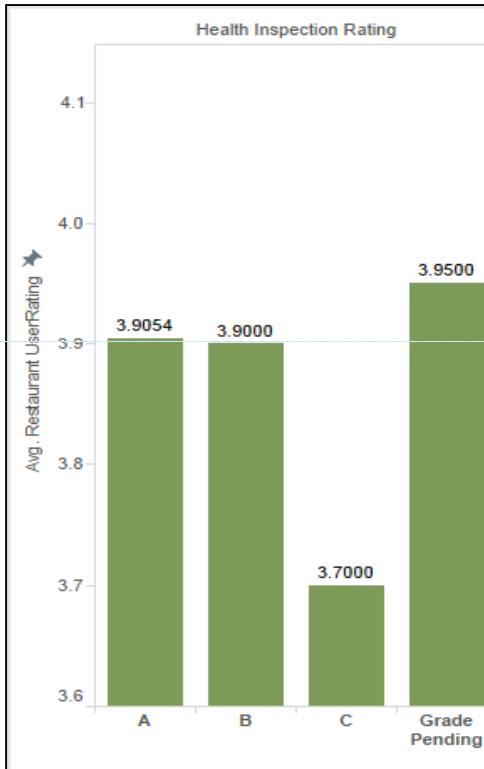
New York City – An Exploration

Results of Health Inspection Grades

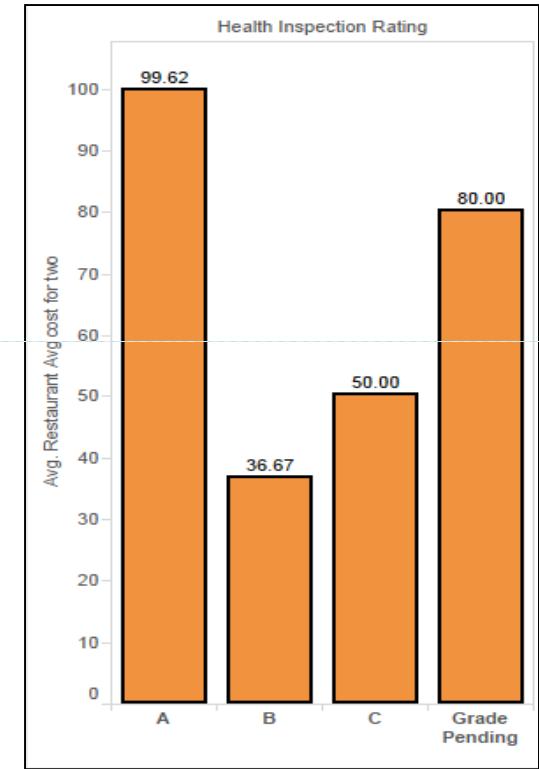
Percentage of Restaurants in each grade



Average User rating of restaurants in each grade



Average Health Inspection rating of restaurants in each grade

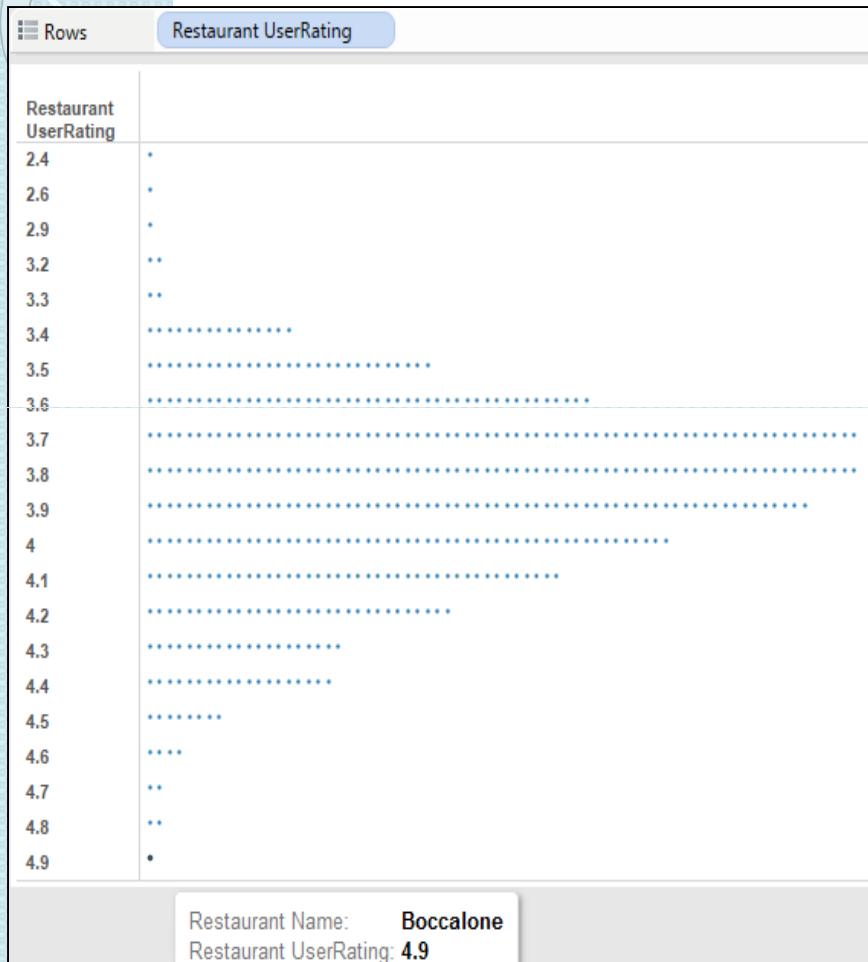


The average user review rating for A grade restaurants is much higher than C grade restaurants

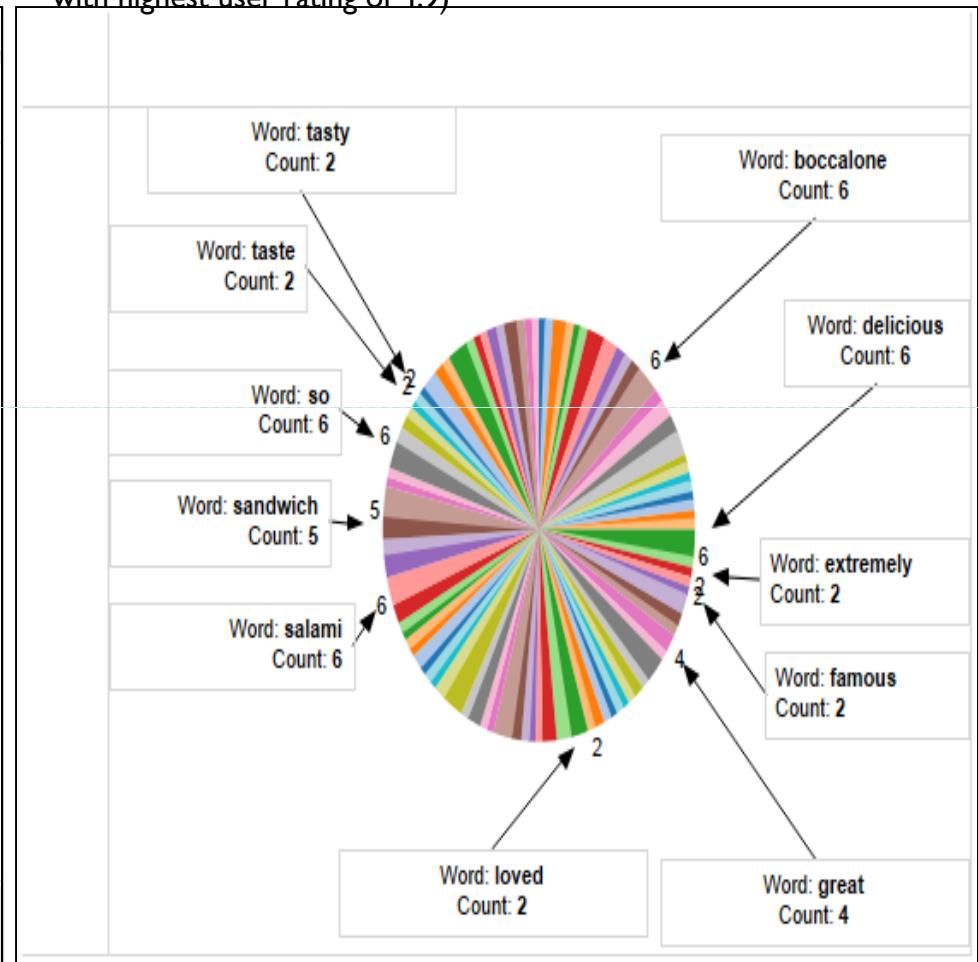
The average cost for two in A grade restaurants is much higher than in B and C grade restaurants

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

This screenshot shows a Microsoft Access form titled "hotel_nyc". The form contains fields for "Id", "Hotel_name" (Supper), "Address" (156 E 2nd Street), "Location" (New York), "Zip" (10009), and "cuisine" (Italian). Below the form is a text area labeled "Review1" containing a single-line review: "I am a big fan of Frank Piscinazzo. The ambience is very reminiscent of Frank and Lili's. Frankie's. Everything was good, but". At the bottom, it shows "Record: 14 5 of 99" and "Search" buttons.

This screenshot shows a Microsoft Access query results grid titled "Final_Query" for the "hotel_nyc" table. The grid has columns for "id", "hotel_name", "address", "location", "zip", "cuisine", and several Expr columns (Expr1006 through Expr1009). The data includes rows for various restaurants like Supper, Root & Bone, Zum Schneider, etc., with their addresses and cuisines. The "Expr" columns contain the raw text reviews from the "Review1" field of each record.

	id	hotel_name	address	location	zip	cuisine	Expr1006	Expr1007	Expr1008	Expr1009
1	Supper	156 E 2nd Street	New York	10009	Italian	i am a big fan of good food. gre	supper?. had n	perfect. i could		
2	Root & Bone	200 E 3rd Street	New York	10009	Southern, Café	great food, lov	if you want a s	cute atmosphe	i saw a picture	
3	Zum Schneider	107 Avenue C	New York	10009	German, Bar Fc	good food, larg	amazing food &	zum schneide	ja bier. this pl	
4	Matcha Cafe W	233 E 4th Street	New York	10009	Cafe, Ice Cream	japanese own	na	na	na	
5	Tuome	536 E 5th Street	New York	10009	Asian	the restaurant	small. big. sid	after seeing m	the restaurant	
6	Esperanto	145 Avenue C	New York	10009	Cuban, Latin Am	i think that by	i probably the b	went for desse	good braillan	
7	Katz's Delicatessen	205 E Houston St	New York	10002	Sandwich	if you've been	send your boy	i personally pre	one of the bes	
8	Poco	33 Avenue B, N	New York	10009	Spanish, Tapas	mg this is the c	brunch is awes	terrible. please	tapas and drink	
9	Minca	536 East 5th Str	New York	10009	Ramen	pretty good rai	i had the chick	great place fo	i had the chick	
10	Yerba Buena	23 Avenue A,	New York	10009	Latin American	great ny latin fi	great place for	na	na	
11	Buenos Aires	513 E 6th Street	New York	10009	Argentine, Ste	great authentic	great dining on	planning to re	i've dined her	
12	Gnocco	337 E 10th Street	New York	10009	Italian	cute little plac	gotta get the a	good vibe/goo	na	
13	Black Iron Burg	540 E 5th Street	New York	10009	American, Burg	awesome burg	best burger joi	est patty melt	solid. great bl	
14	Balthazar	80 Spring Street	New York	10012	French, Cafe, B	had one of my	a fabulous bras	yes yes! lo	the strawberry	
15	Momofuku Noodle Bar	171 1st Avenue	New York	10003	Asian, Ramen	the queue at m	went on a very	noodles and go	not all that, an	
16	Buddakan	75 9th Avenue	New York	10011	Chinese, Fusion	wow. from the	i always told b	our hotel, sohc	one of the best	
17	Per Se	Time Warner Center	New York	10019	French	exceptionally f	perfection. har	i was really ex	very enjoyabl	
18	Peter Luger Steakhouse	178 Broadway	New York	12111	Steakhouse, Am	definitely live	extremely dis	the best place	world famous	

Text Mining - Weka

Supper, New York, New York

<https://www.zomato.com/new-york-city/supper-alphabet-city>

Cash only

Stacy Landers 15 Reviews , 30 Followers
12 days ago via Zomato for iOS

RATED 4.5 I am a big fan of Frank Prisinzano. The ambiance is very reminiscent of Frank and Lil' Frankie's. Everything was good, but Lil' Frankie's is where my heart lies.

Like 0 Comment 0

Dallas Trends 133 Reviews , 106 Followers
3 months ago

RATED 4.5 Good food. Great atmosphere. I was with someone who never dined at a restaurant where they seat others at your table, and it was exciting. I had the privilege of dining for dinner. Somewhat limited seating. Friendly service. We decided to try this place after a local recommended it. I'd love to return.

trendydallas.com

Like 0 Comment 0

EXPLORING!
Download the Zomato app and discover great restaurants around on-the-go!

AVAILABLE ON 

OR LET US TEXT YOU A [DOWNLOAD LINK](#)

ZOMATO SPOONBACK 

C:\Amo\directory\positive\hotel_1.arff - Notepad++

File Edit Search View Encoding Language Settings Macro Run Plugins Window ?

hotel_1.arff

```
1 @Relation Hotel
2
3 @Attribute Hotel_name STRING
4 @Attribute Review STRING
5 @Attribute Rating {1,2,3,4,5}
6
7 @Data
8 "Supper", "i am a big fan of frank prisinzano. the ambiance is very remin
9 "Supper", "good food, nice atmosphere. i was with someone who never din
10 "Supper", "terrible, i couldn't have been happier with my dinner at supper
11 "Supper", "quality italian, good wine and food! its a bit pricey (cash on
12 "Supper", "best italian in the village!. amazing pasta, rustic atmosphere
13 "Supper", "mostly buon appetito. i've only been here twice, the first time
14 "Supper", "the best!. cool place, grrrrrrrreat service and the food is ex
15 "Supper", "snug but fantastic!. the old-worldish, faded glory chic lends
16
17
18
19
20
```

Positive Review arff

→ C <https://www.zomato.com/new-york-city/supper-alphabet-city>

6968tim
1 Review, 0 Followers

Oct 14, 2013

NEGATIVE Supper?. Had meal at Supper more than once. We tend to go when we visit n.y. Last meal was such a disappointment we will not return. The Bruschetta was so awful you had to eat it with a spoon-almost tasted like someone dumped a can of stewed tomatoes over bread-soggy mess. The risotto had a funny taste to it and it tasted old. we informed the waiter but he only waved it off. We were over charged on bill but didn't notice it till we left. One thing to remember is they put tip on bill and don't inform you as most people don't notice and tip twice!

Menu | Reviews(11) | Map

Add a blog post >

```
1 @Relation Hotel
2
3 @Attribute Hotel_name STRING
4 @Attribute Review STRING
5 @Attribute Rating {1,2,3,4,5}
6
7 @Data
8 "Supper","Supper?. Had meal at Supper more than onc
9 "Supper","Overpriced...with terrible service ",2
10
11
12
13
14
15
16
17
18
19
20
21
22
```

Negative Review arff

Text Mining - Weka

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



The screenshot shows the WEKA SimpleCLI interface. It features a title bar with the text "SimpleCLI" and a small icon. Below the title bar is a welcome message: "Welcome to the WEKA SimpleCLI". The main area contains several lines of text providing instructions for using the command-line interface, including details about command completion and history. At the bottom of the interface, there is a text input field where commands can be typed. A command has been entered into this field: "java weka.core.converters.TextDirectoryLoader -dir C:\Amol\directory > C:\Amol\directory\hotel_nyc.arff". The output of this command is displayed below the input field, stating "Finished redirecting output to 'C:\Amol\directory\hotel_nyc.arff'".

```
SimpleCLI

Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.

Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with ".." or "../"
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

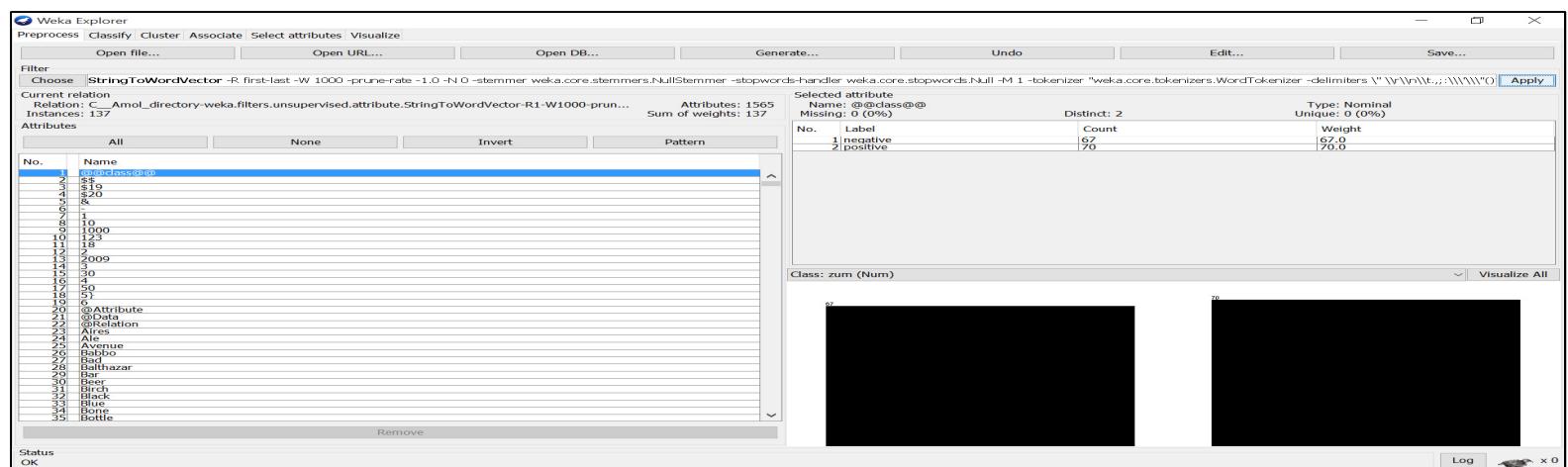
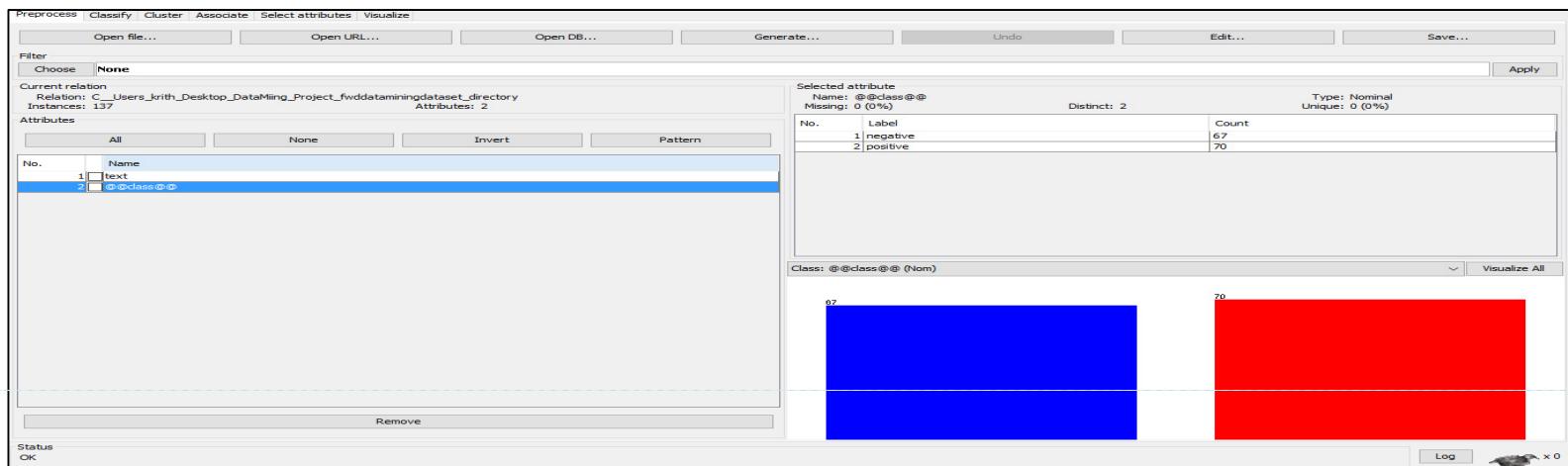
> help

Command must be one of:
    java <classname> <args> | > file
    break
    kill
    capabilities <classname> <args>
    cls
    history
    exit
    help <command>

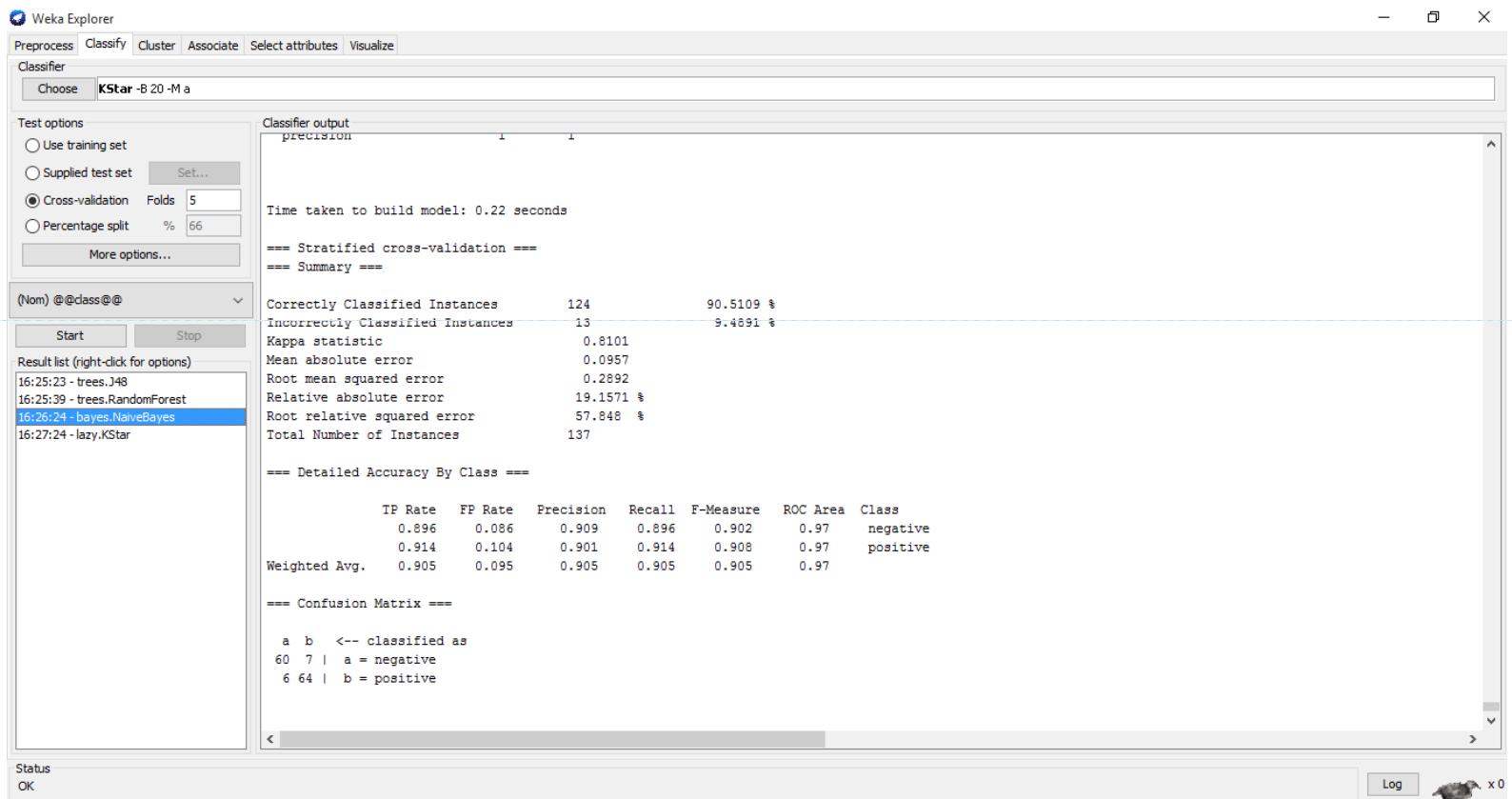
> java weka.core.converters.TextDirectoryLoader -dir C:\Amol\directory > C:\Amol\directory\hotel_nyc.arff

Finished redirecting output to 'C:\Amol\directory\hotel_nyc.arff'.
```

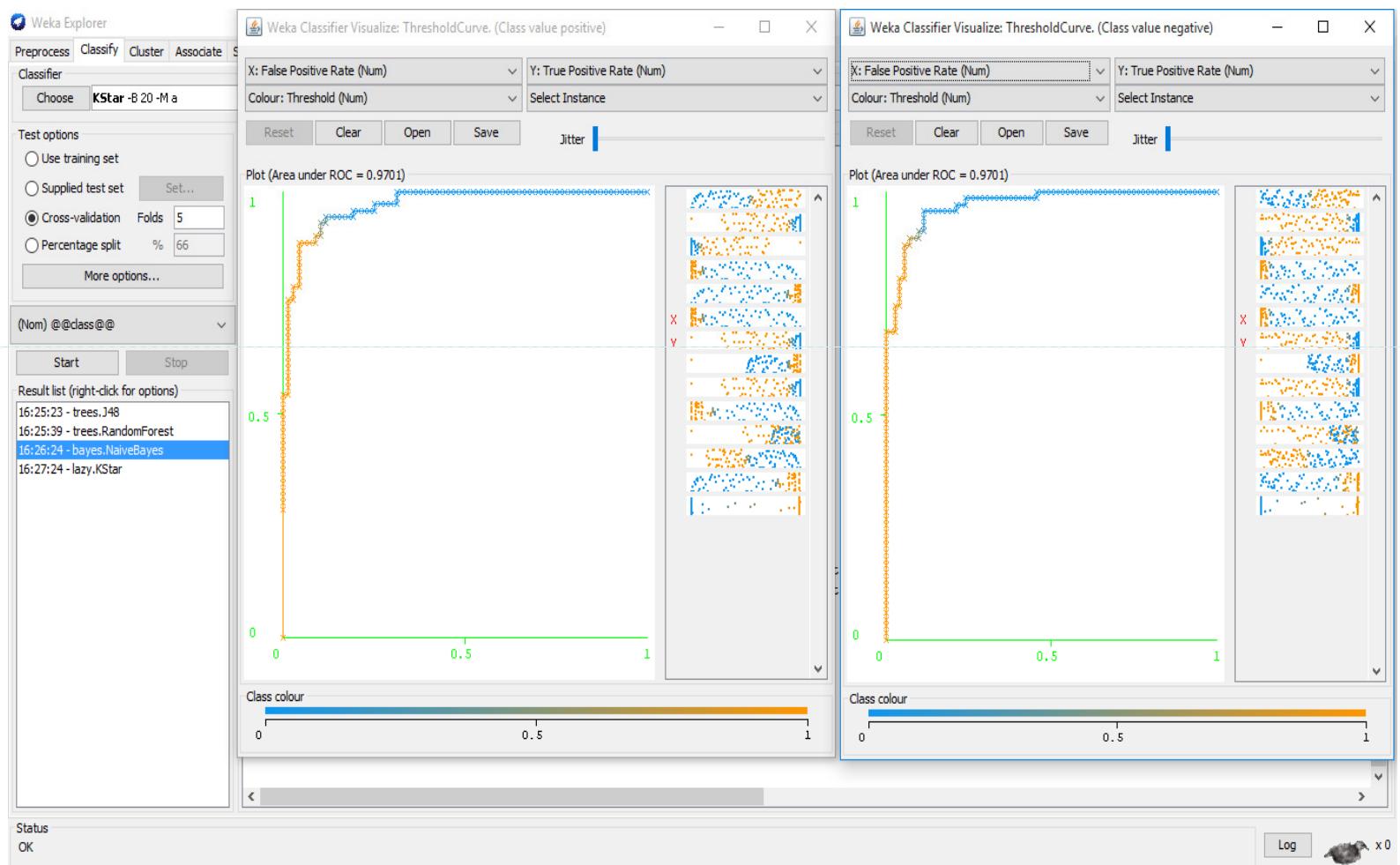
Text Mining - Weka



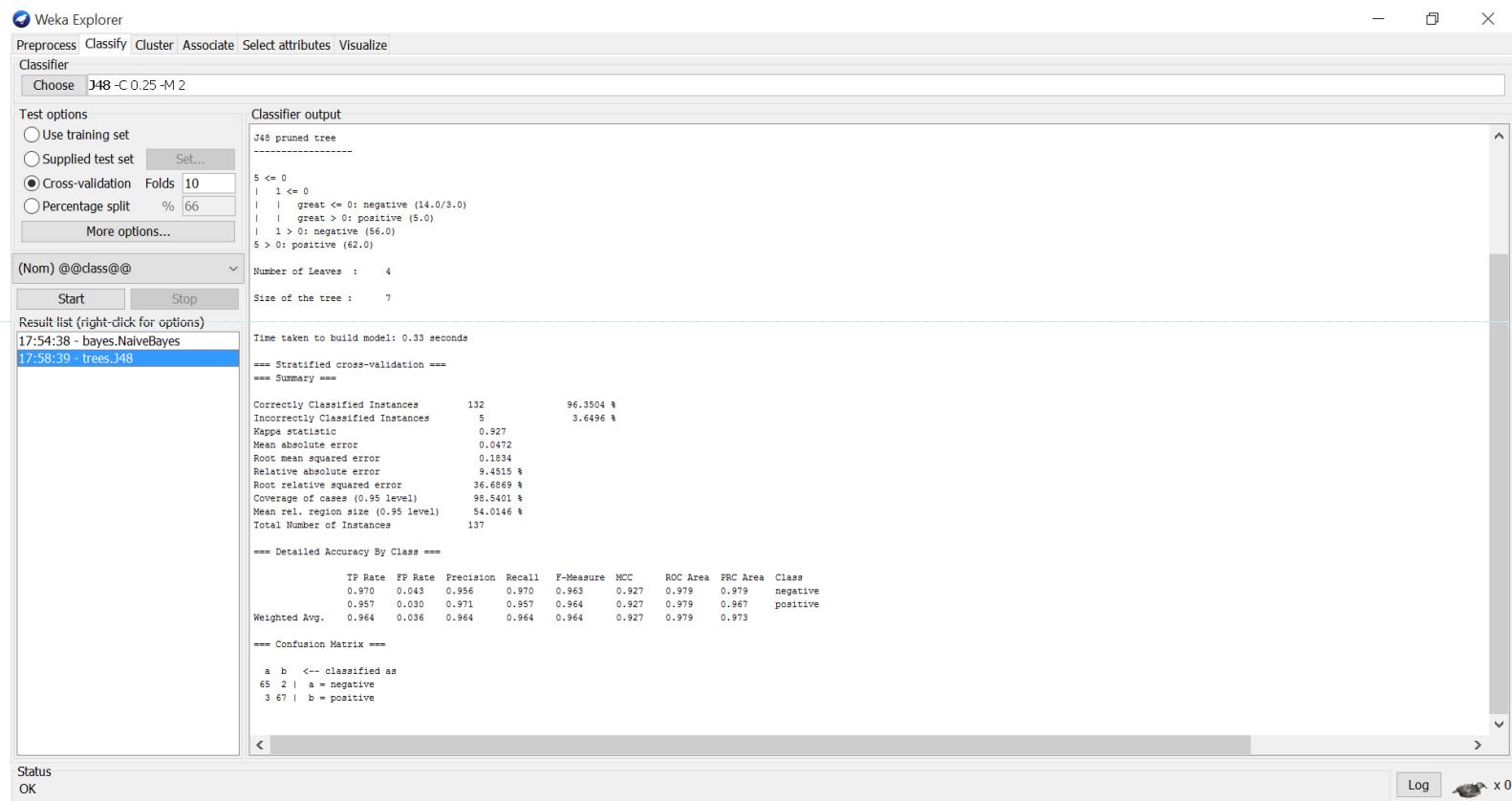
Naïve Bayes Classification - Confusion Matrix



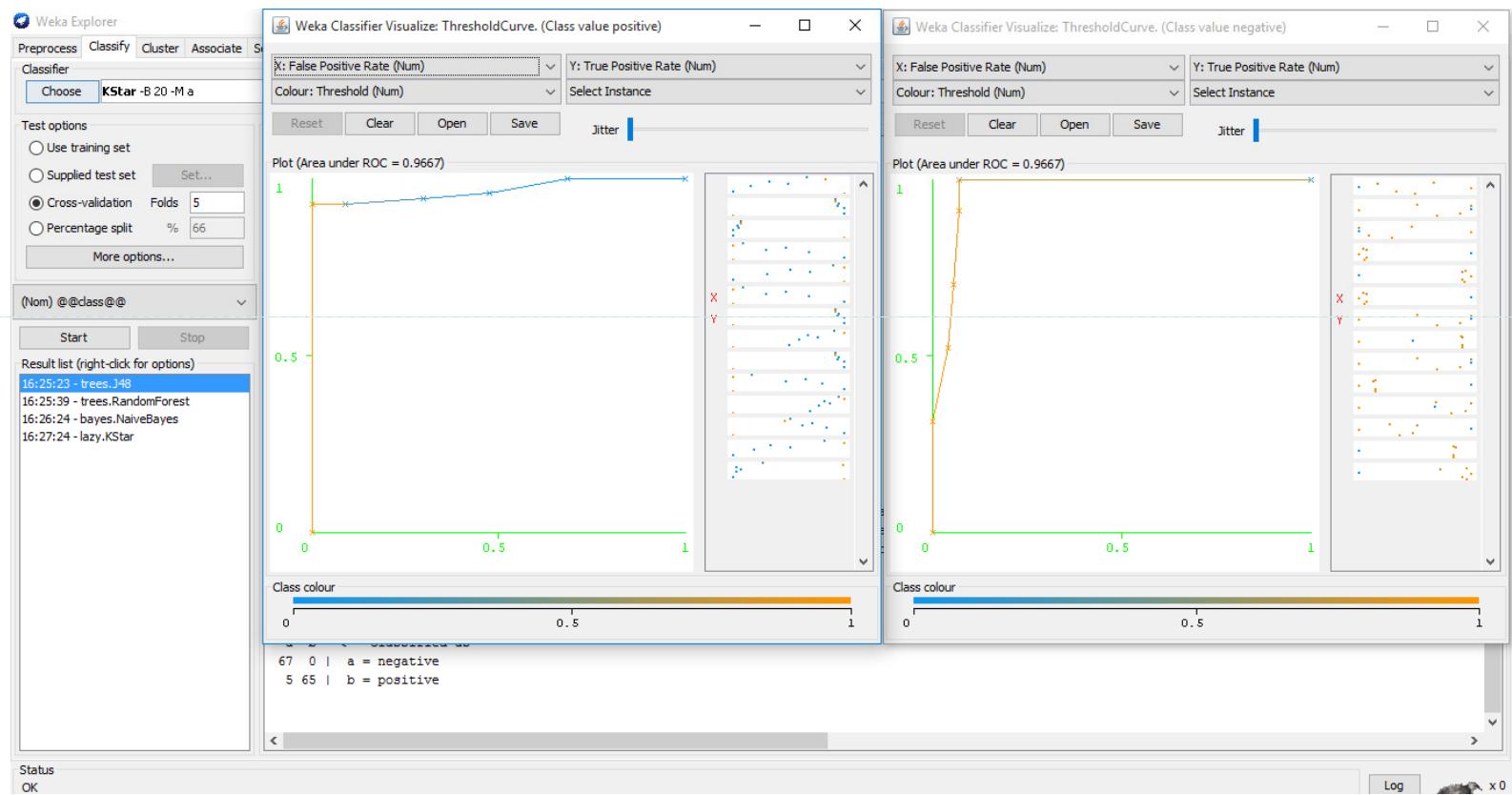
Naïve Bayes Classification - ROC Curve



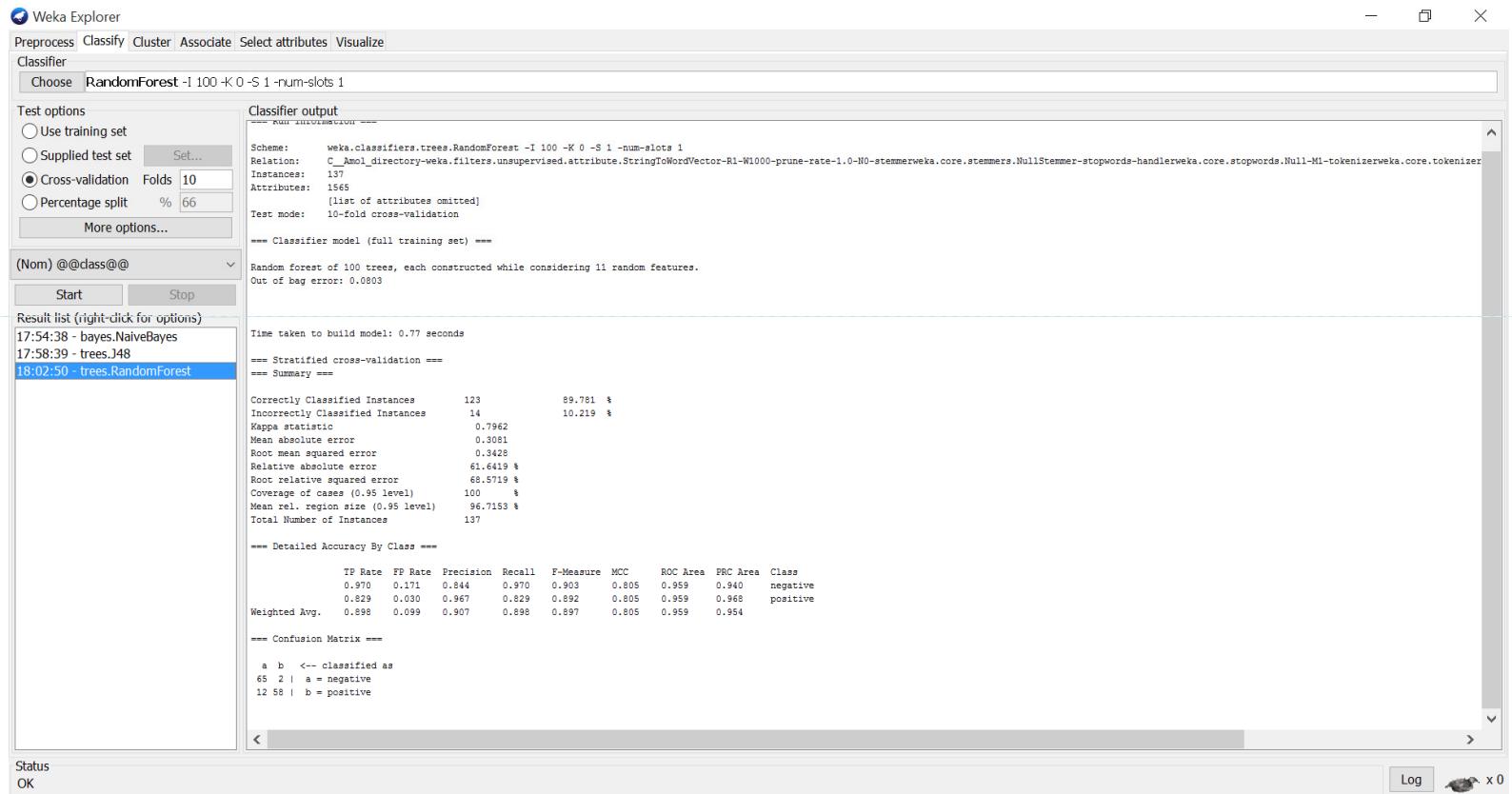
J 48 Classification - Confusion matrix



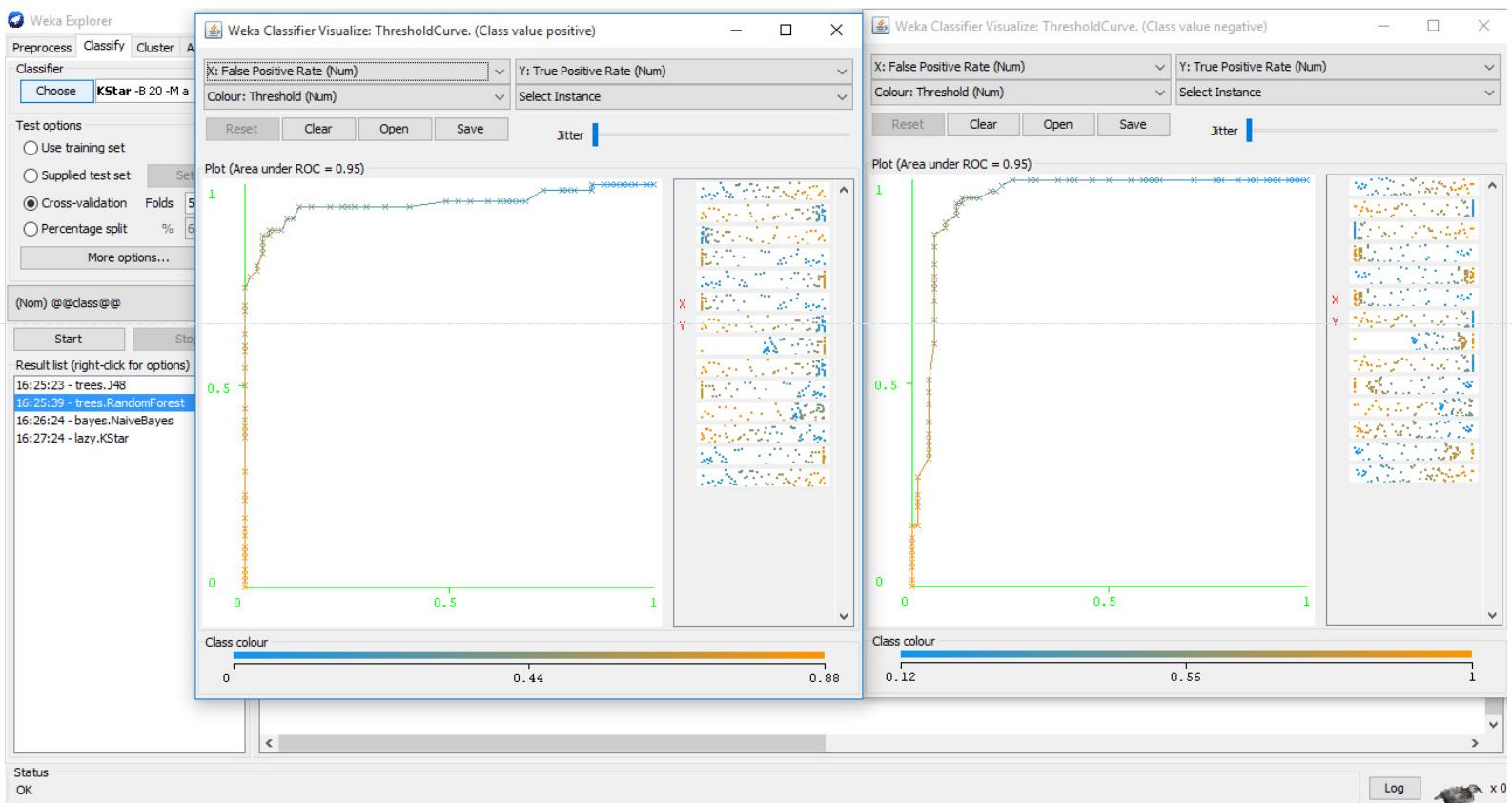
J 48 Classification - ROC Curve



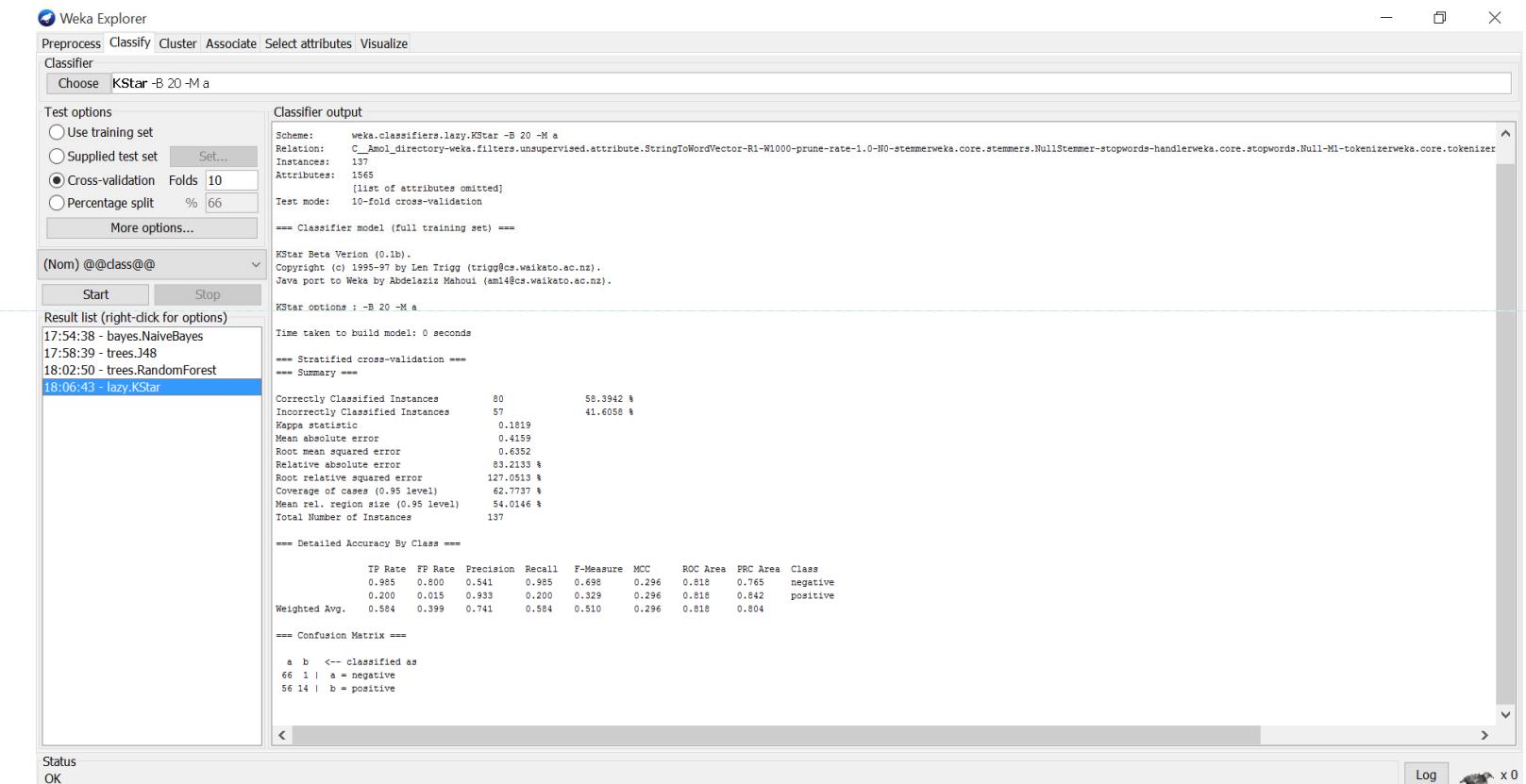
Random Forest Classification - Confusion matrix



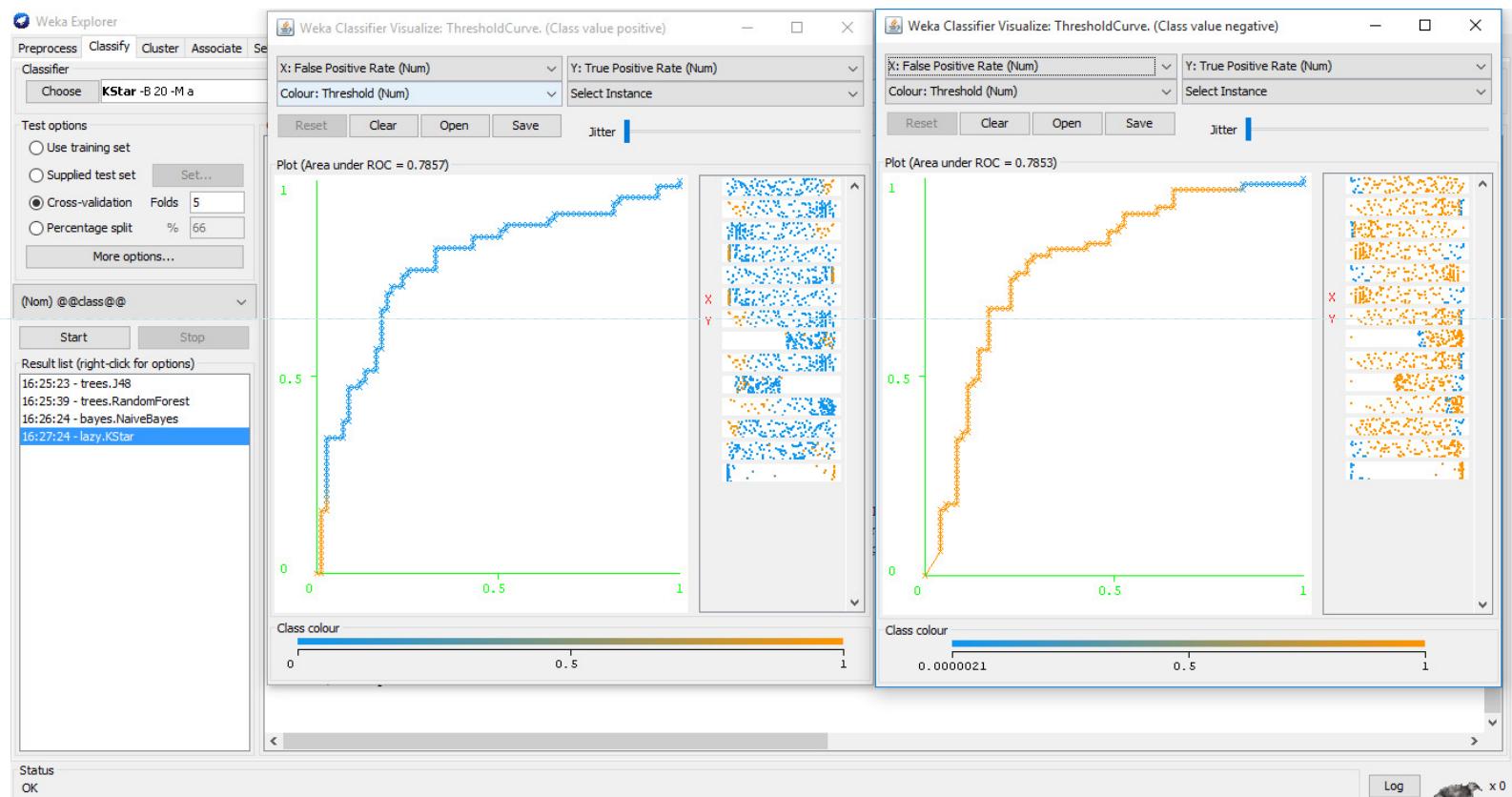
Random Forest Classification - ROC Curve



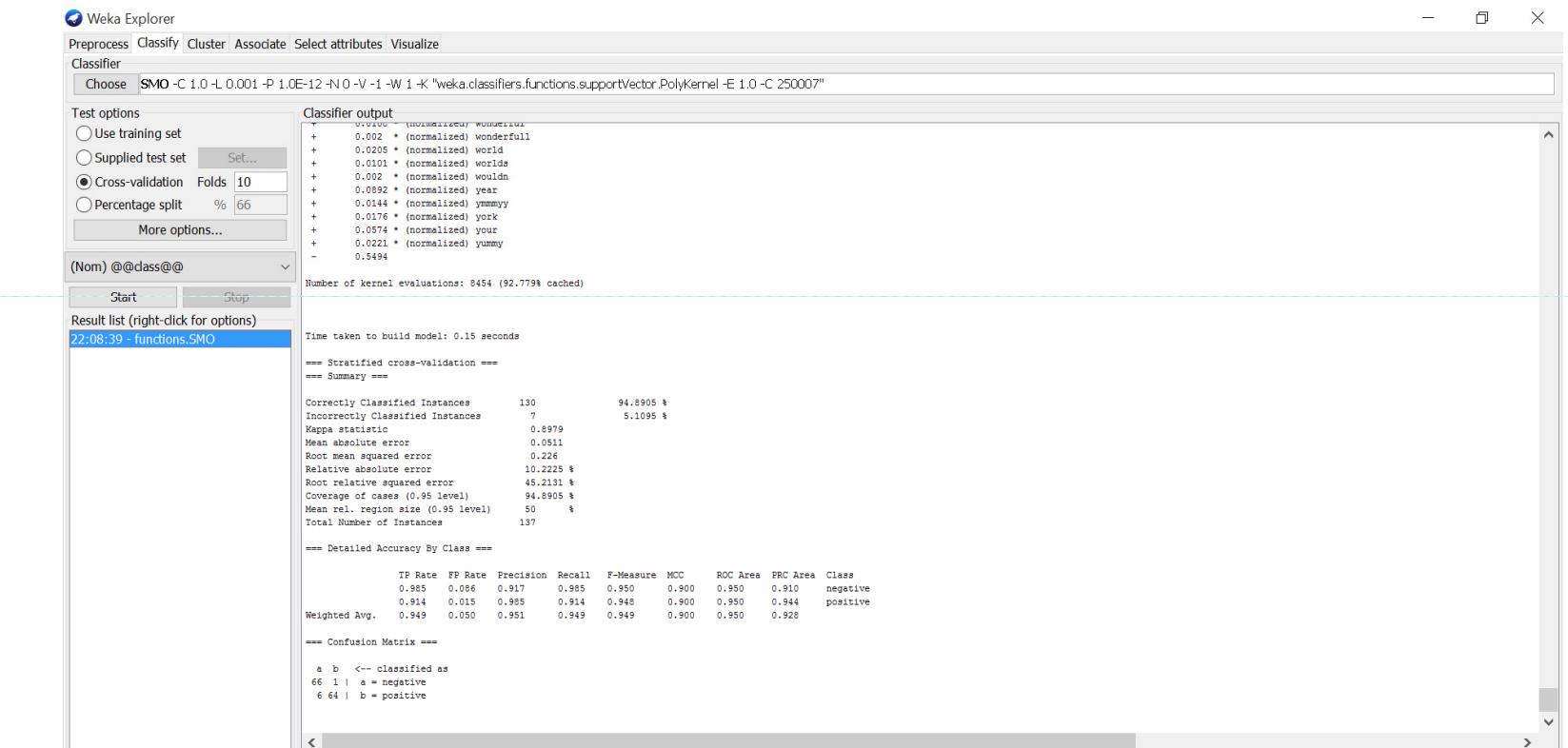
KStar Classification - Confusion matrix



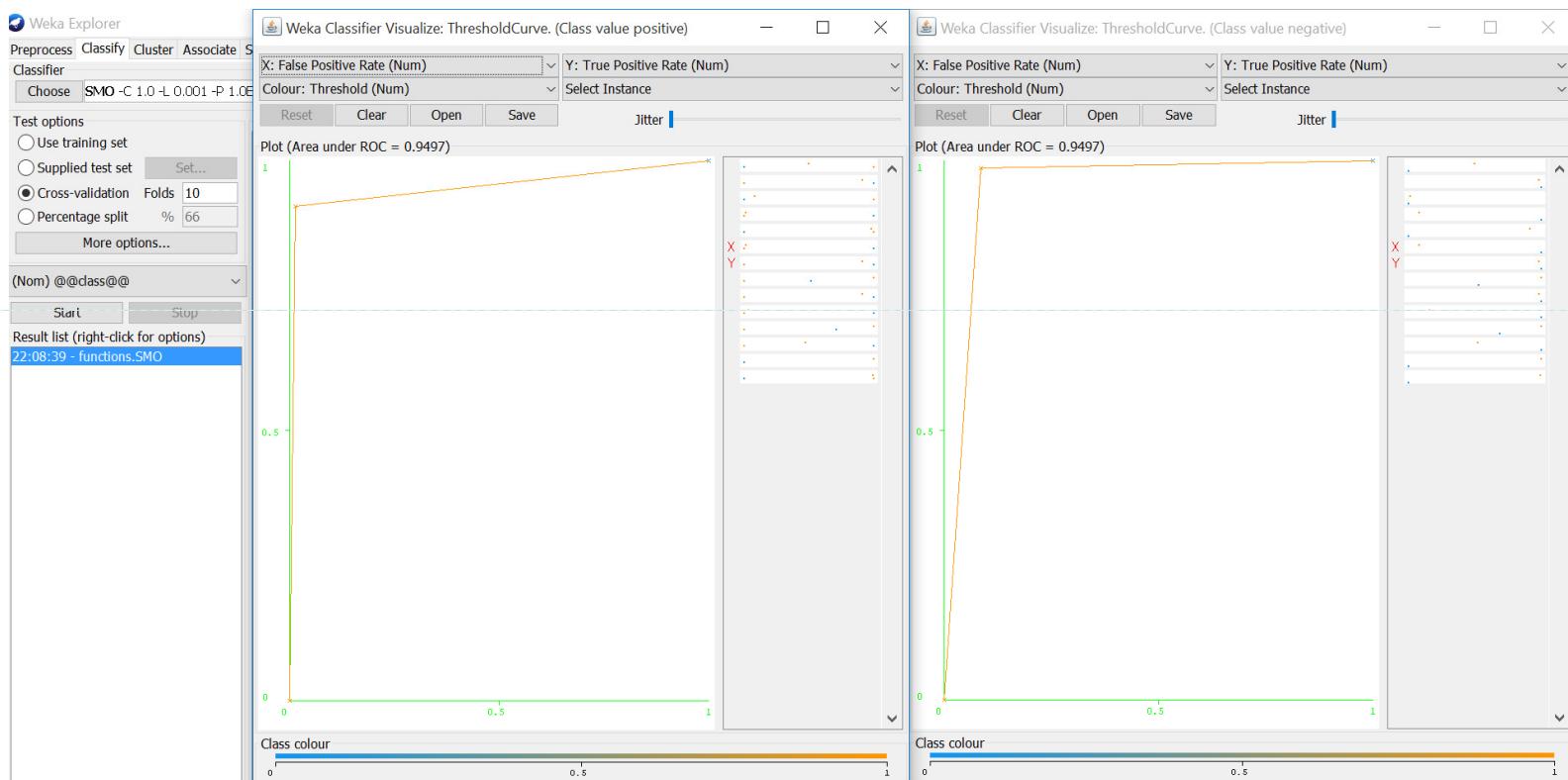
KStar Classification - ROC Curve



Support Vector Machine Classification - Confusion matrix



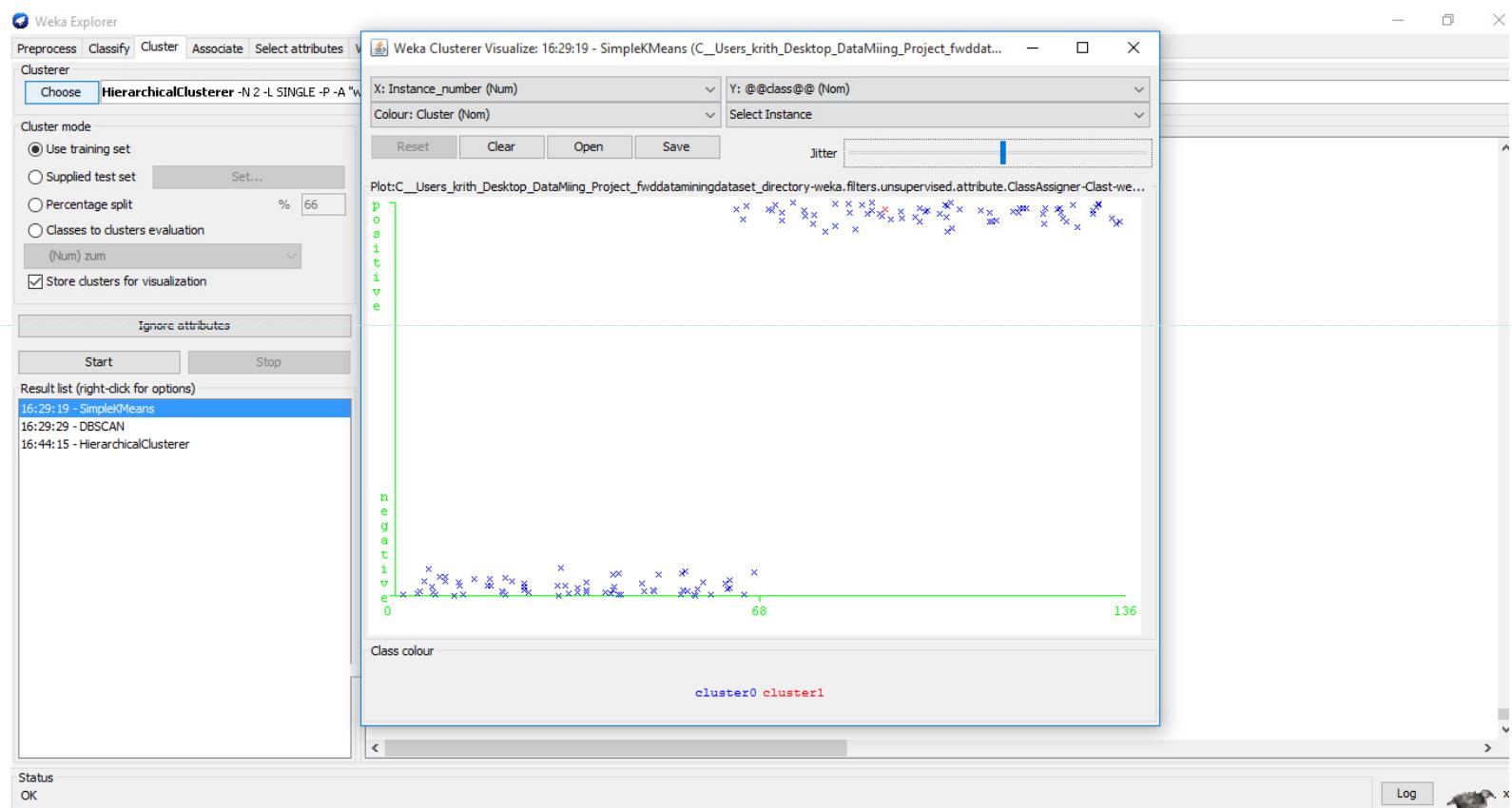
Support Vector Machine Classification - ROC Curve



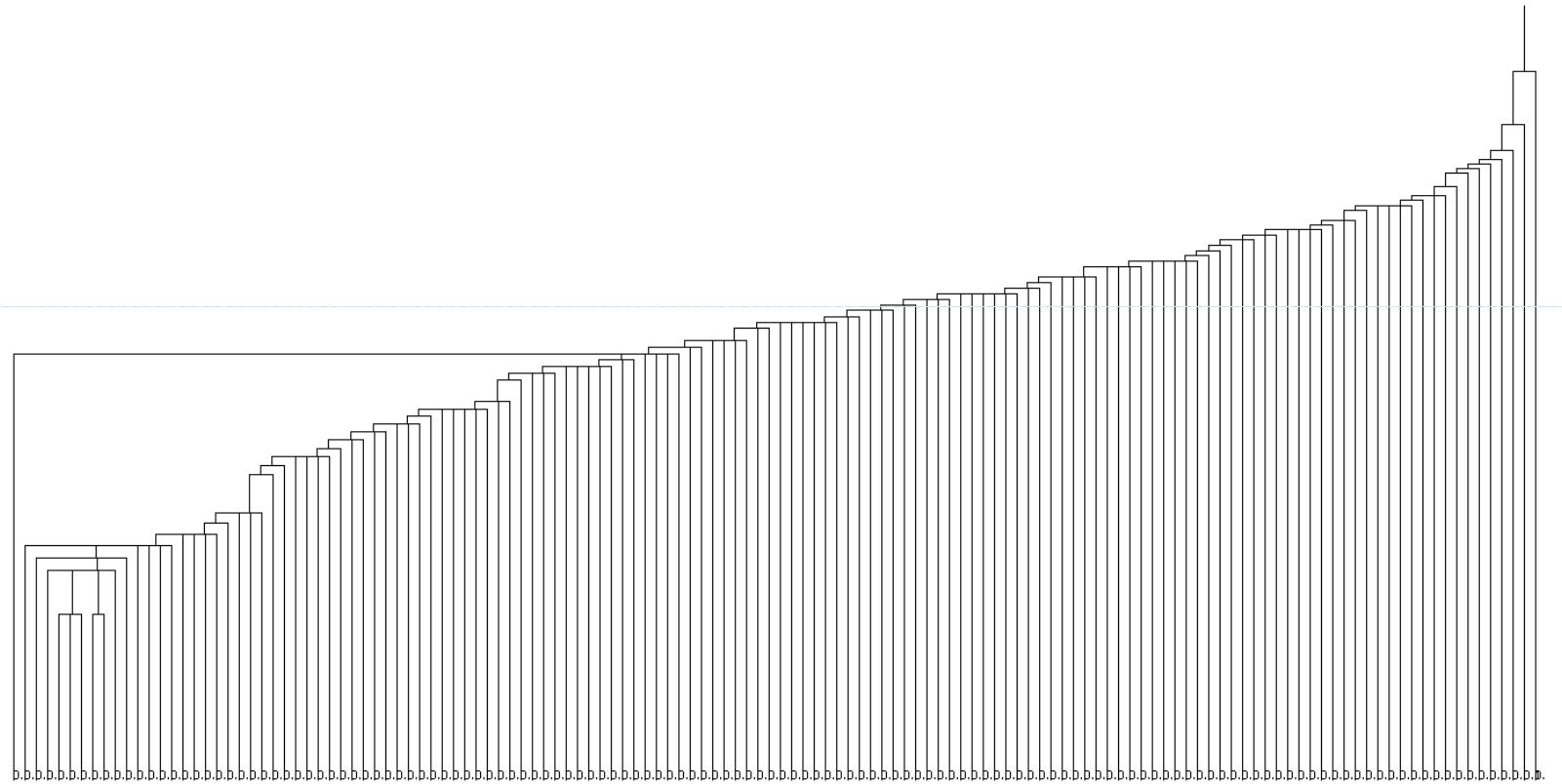
Comparison of Classification Models

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??