

# **PREDICTIVE MODELING TEAM 3**

## **TARGETED APPROACH TO ATTAIN HAPPINESS**

### **Using Somerville Happiness Survey**

**Abhishek Gorla  
Agastya Kommanamanchi  
Gowtham Kommineni  
Elmedina Loga  
Jitender Tewari**

## Contents

EXECUTIVE SUMMARY .....	3
PROBLEM STATEMENT .....	3
METHODOLOGY .....	4
CONCLUSIONS AND RECOMMENDATIONS .....	144
RECOMMENDATIONS .....	145
REFERENCES .....	177
APPENDIX.....	188

\*\*\* NOTE- Please refer to the Appendix for all the images referenced in the text.

# EXECUTIVE SUMMARY

We have sourced the data compiled by the local government of Somerville conducted across three years as the sample to develop insights about happiness. We effectively cleaned the survey data and through pre-visualization, identified the parameters that are most likely to play a role in a person's happiness.

First, through linear regression, we identified parameters with high statistical significance that determine happiness of a person. Then, we developed various models to classify the respondents into two groupings of "Happy" people and "Not Happy" people.

Further, we sub-divided these "Not Happy" people into various focus groups through clustering method which will allow governments to better target their programs and resources.

# PROBLEM STATEMENT

Happiness is increasingly becoming an acceptable measure of success for governments and companies alike, across the world. The primary challenge we are dealing with, through our project, is to make unhappy people happy. To achieve this, we have identified two objectives:

- To identify the determinants that make people Happy.
- To accurately identify the people who are Not Happy.

Solving the above problems helps us make insightful targeted recommendations to improve the happiness determinants for the unhappy people.

# METHODOLOGY

We have taken the SEMMA approach and observed the patterns in the general happiness of people.

## **SAMPLE:**

The survey data we have selected was already a sample data that represented the population of Somerville. The sample was designed to cover various races and gender, income and age groups. The survey was conducted across three years (2011, 2013, 2015) where questions were sent out to the entire population in 2011 whereas random samples were selected for 2013 and 2015.

We observed that majority of our data was from 2011 as the data was compiled from the responses given from the general populace of Somerville. Coming to the data for years 2013 and 2015 the number of observations were significantly less even as the samples were created to have statistical significance.

So, our analysis was carried out separately for all 3 years to get insights across the years as well. But our major focus was around the data of year 2011 for predictive purposes as a larger dataset allows for better training to get accurate results.

## EXPLORE AND MODIFY:

### ANOMALIES and ACTIONS:

Exploration	Modification
<b>Same question was framed in different words across various years:</b> “The overall quality of public schools in your community_2011” and “The overall quality of public schools_2013_2015”.	Both the Columns were merged taking the scale into consideration.
<b>Different questions having similar meaning</b> e.g. “The effectiveness of the local police_2011_2013” and “Your trust in the local police_2015”	Both the columns were merged because both talk about the satisfaction of the same metric.
<b>Inconsistency in options provided across different years</b> e.g. There were multiple combinations of Races which were provided in different years.  In 2011 and 2013, Combinations of White, Black, American, Hispanic, Non-Hispanic, Asian, Native American and Native African, Pacific Islanders were allowed. However, in 2015, All the possible granular options were given, and respondents were asked to mark all those related to him/her.	Options have been normalized across years by projecting the data to the non-granular options.
<b>For few Categorical features binning was different across years.</b> e.g. In case of Annual Household Income, the binning for the income groups have been made on a different interval for different years.  For Example: In 2011, The options given were binned with an interval of \$10000. E.g. 10000- \$19999, \$20000- \$29999 etc. But in 2013, 2015, The options given were binned with an interval of \$25000. E.g. \$25000- \$49999, \$50000- \$74999, \$75000- \$99999 etc	Subsets have been merged to create common bins and the variables have been regrouped accordingly.
<b>Few of the categorical features consisted of options which were differently spelled in different years.</b>  For example, all options in Gender were in small letters for the year 2011 like male. But for years 2013 and 2015 they were camel cased like Male and Female.	All the variables have been recoded to have same values.

### The scale of the data for some continuous features was different in different years

For example, The Rating of public schools was done on a scale of 5 for the year 2011 and for 2013 and 2015 it was done on a scale of 10.

Columns which needed combining have been normalized to have same scale and the rest were left for JMP to take care of.

### OUTLIER ANALYSIS:

The data we received had very few outliers. There were 15 outliers in total and all them can be classified as data entry errors as the observations in these columns had categorical values from 1 to 10. Similarity with other people column had two outliers of 75 and 110, likelihood of seeking advice column had five outliers with values of 108

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers (Count)
Satisfaction at Somerville	5	10	-10	25	0
Similarity with other people	4	9	-11	24	2 75 110
Likelihood of Seeking advice	3	10	-18	31	5 108(5)
Satisfaction with Neighborhood	5	10	-10	25	0
Pride in Somerville_2015	5	10	-10	25	0
Satisfaction with Availability of Information About City Services_2015	3	5	-3	11	0
Availability of Affordable Housing	2	4	-4	10	0
Satisfaction with the Cost of housing	1	4	-8	13	0
Public Schools	0	4	-12	16	1 62
Physical Setting of Somerville	2	4	-4	10	7 33(7)
Trust in Police	2	5	-7	14	0
Availability of Social Community Events	3	5	-3	11	0
Walking at Night	4.2	10	-13.2	27.4	0
Physical Setting of Neighborhood	4	10	-14	28	0
Satisfaction With appearance of parks	0	5	-15	20	0
Ward	2	7	-13	22	0
Precinct	1	3	-5	9	0

### HANDLING OF OUTLIERS:

Since the percentage of outliers is minimal, we have amputated them from our analysis.

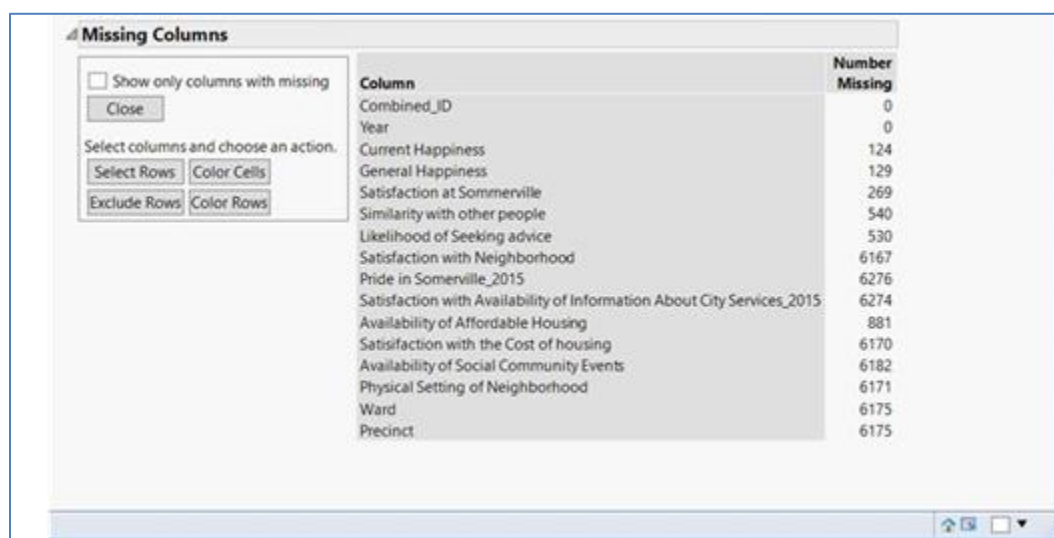
### HANDLING MISSING VALUES:

Majority of the missing values present in the data is a direct result of the different questions that were asked across the years. For example,

“In.general..how.similar.are.you.to.other.people.you.know.\_2011” had missing values for observations pertaining to years 2013 and 2015.

Since this is a survey data with an extensive range of questions, some of which are personal in nature, it was no surprise that we have observed plenty of ‘NA’ (No Response) and ‘R’ (Refused to Answer).

Which neighbourhood do you live in? We observed a lot of missing values. This can be interpreted as a concern for privacy as people were not willing to disclose that information.



Column	Number Missing
Combined_ID	0
Year	0
Current Happiness	124
General Happiness	129
Satisfaction at Somerville	269
Similarity with other people	540
Likelihood of Seeking advice	530
Satisfaction with Neighborhood	6167
Pride in Somerville_2015	6276
Satisfaction with Availability of Information About City Services_2015	6274
Availability of Affordable Housing	881
Satisfaction with the Cost of housing	6170
Availability of Social Community Events	6182
Physical Setting of Neighborhood	6171
Ward	6175
Precinct	6175

### MISSING VALUE ANALYSIS:

The *features* that have a very high percentage of missing values have been excluded from our analysis.

Example: Satisfaction in Neighbourhood (6167 missing Values) and Pride in Somerville (6276 missing values.) have been amputated.

The continuous columns with considerably minimal missing values have been imputed using the “Multivariate Normal Imputation Method”.

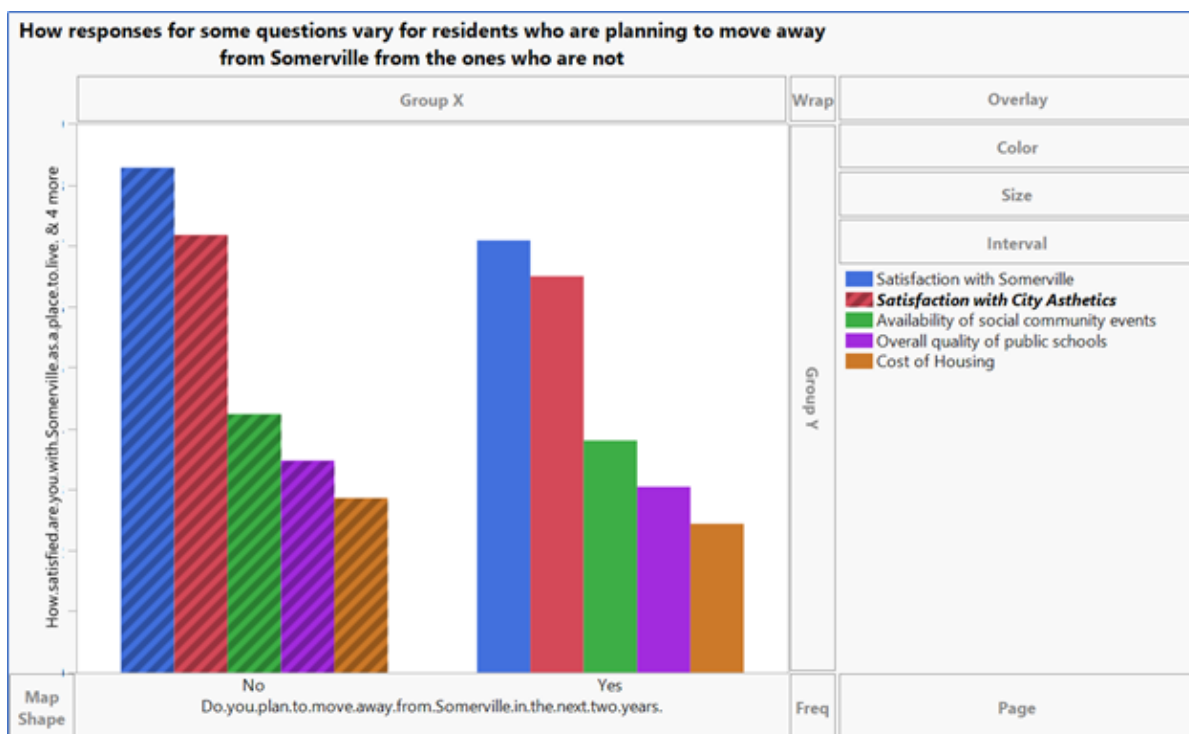
The Categorical columns where a blank was left as a response have been recoded into 'R' (Refused to answer). Since the survey was handed out with all the questions to the responders the only possibility for a blank value is when the respondent has ignored to answer which can be considered as “Refused to answer”.

Initially we started of with 52 columns and have eventually reduced it to around 18 columns which are consistent, reliable and can be used for modelling purposes.

We tried to retain the diversity of the features while at the same time identified features that are most relevant for further modelling purpose by exploring the data for functional significance as follows.

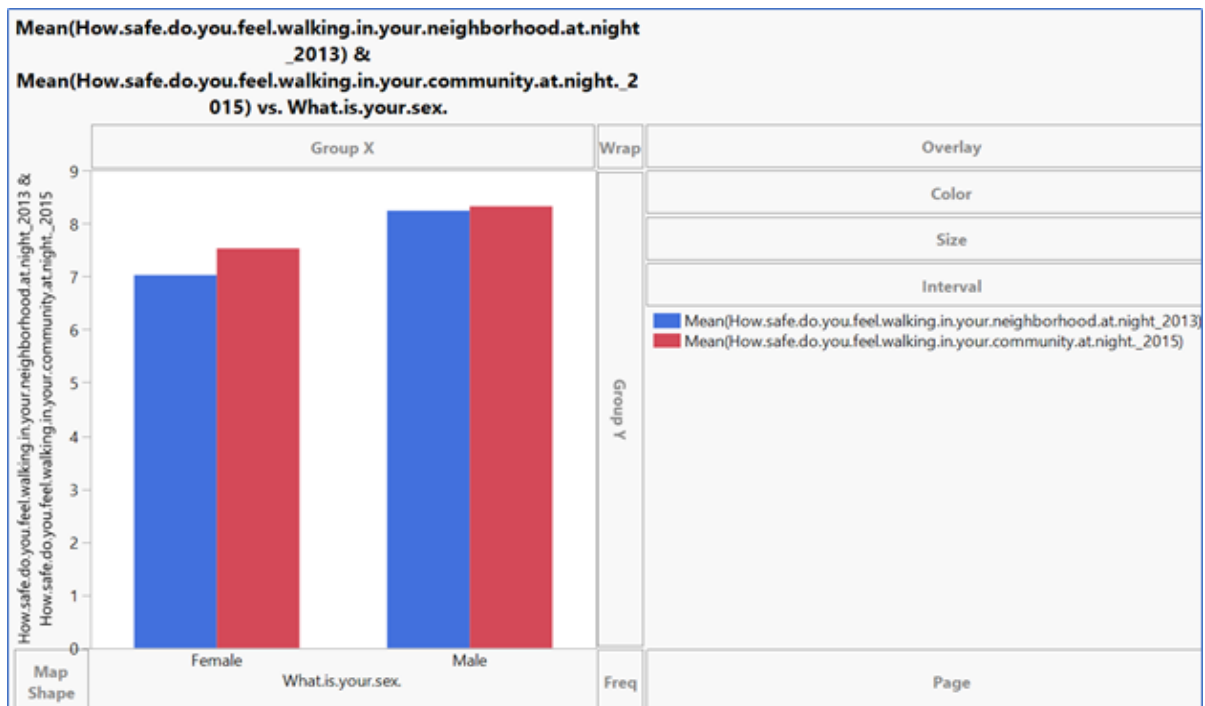
### FUNCTIONAL EXPLORATION:

The below graph helps us understand how some of the parameters may play a vital role in deciding to plan a future in Somerville –





Below graph shows Gender wise distribution for question “How safe do you feel walking in your neighborhood at night?”



Below table shows average scores for some of the questions answered by the respondents -

Questions concerning	2011	2013	2015
General Happiness	7.63	7.75	8.06
Satisfaction at Somerville	7.59	7.67	8.23
Availability of Affordable Housing	2.96	2.98	3.05
Public Schools	3.17	3.56	3.28
Trust in Police	3.68	3.89	3.75
Similarity with Other People	6.69	6.76	6.95
Likelihood of Seeking Advice	6.53	6.62	6.67

Note: As we can already see that most of the parameters including general happiness have kept increasing significantly. This shows how effective the city policies have been adopted to make people happy.

## MODEL BUILDING:

Before starting to model, it is important to identify the predicted variable and check for any kind of issues. With the problem at hand being a conversion problem, it is important to have a categorical variable as the predicted. So, we created a new variable “Happy or Not” based on a cut-off which was subjective but has been selected from well-known researches.

The variable “Happy or Not” was made based on a different scale and comparing it with the existing General happiness values.

Scale OF Happiness	Description
Equal to 0	Not Happy
Equal to 5	Neither Happy nor Not Happy
Equal to 7.5	Fairly Happy
Equal to 10	Completely Happy (Can’t get better)

*“A General Happiness value of 7.5 was selected as a cutoff for creating a categorical feature “Happy or Not”. Greater than or Equal to 7.5 is recoded to Happy (1) and Less than 7.5 to Not Happy (0).”*

The continuous variable of General Happiness was used only for the purpose of linear regression, so that important determinants for happiness can be found by using the p values obtained.

**\*\*\*Please Note that all the numerical values to describe each model and for comparison of models pertain to the TEST data partition alone.**

## LINEAR REGRESSION:

Since Linear Regression is sensitive to multicollinearity, observing at the scatterplot matrix helps us find such patterns. We built a linear regression model with 13 variables and found that 8 of those variables are statistically significant. Out of them Similarity with others and Satisfaction with Somerville with high log worth really stood out. Scatterplot matrix can be observed in IMG-1.1

Since the scatterplot matrix doesn't show any collinearity, we haven't excluded any features from the analysis. We have converted each categorical column into n-1 dummy variables before we moved into modelling.

On Executing the model linear regression model with all the variables has given out the following results of p-values showing the significance of each feature in IMG-1.2. The parameter estimates and the summary details can be seen in the IMG-1.3 and IMG-1.4

Since Linear Regression is only used to find the predominant features for the happiness, we have observed 10 features which are significant- "Similarity with other people, satisfaction at somerville, age, gender, annual household income, likelihood of seeking advice, trust in police, physical setting of somerville. availability of affordable housing, public schools."

These selected features will be used for further modelling. And Since our predicted variable is categorical variable, we will only use algorithms which deal with them.

## LOGISTIC REGRESSION:

Using the above obtained features for logistic regression modelling, we have observed the ROC curve which has an AUROC value of 0.725 in the IMG-2.1.

We can also see the misclassification rate of 28.2% and a percentage of prediction of Zeros for the test data as 45% from IMG-2.2

## **DECISION TREES:**

As we know that decision trees are one of the easiest algorithms to predict the categorical variables, On modelling using the same features extracted from linear regression we can see the output of the Decision trees from the images mentioned in the appendix.

The area under the ROC curve comes out to be 0.7150 from IMG-3.1 and the total accuracy of the model comes out to be 69.2% from IMG-3.2 and a total accuracy in the prediction of a zero is 46%.

## **BOOSTED TREE:**

As we know that a boosted tree is decision tree which was built to minimize the error of previous step in the current step using the gradient descent framework making it a convex optimization problem. Using a boosted tree decreases the variance of the model resulting in overfitting. Knowing this we have not given it a much priority since the survey data is prone to over fitting. But the results from the Boosted tree can be seen in IMG-4.1, IMG-4.2.

The total accuracy of the model comes out to be 69.8% and an area under the ROC curve of 0.744. The accuracy of Zero prediction comes out to be 45%

## **RANDOM FOREST/ BOOTSTRAP FOREST:**

Since the boosted tree has been not reliable due to overfitting problem, the random forest method helps us by randomly making a number of trees with different features and rows of data making it more statistically significant. But the problem with random forest method is the huge inconsistency of the model output, which cannot be reliable.

However, on building this model the output, area under the ROC curve comes out to be 0.767 with a total accuracy of 71.8%, but the accuracy of Zeros comes out to be 44%. These values can be calculated from IMG-5.1 and IMG-5.2

## NEURAL NETWORKS:

Neural Networks are one of the most non intuitive models which can be built. So, the choice of features as an input should be done with care. However, we have used only the significant features which were a result of the linear regression model. Here we are trying to predict the “Happy or Not” variable using the above-mentioned features.

We have tried various number of nodes and type of activation functions to get the best fit which can predict the zeros of the data more accurately. The best model parameters of the model have been observed as only one layer with 6 neurons- 3 neurons with NTanH activation function and other 3 have a Gaussian activation function with a learning rate of 0.085.

The area under the ROC curve is found to be 0.737 in IMG-6.1 and a total accuracy of 70.6% has been calculated from IMG-6.2. But the accuracy of the zeros has been calculated as 57%.

## MODEL SELECTION: IMG-7

We observed that all the models that we have tested showcased similar metrics in terms of misclassification rate, model accuracy and R Square values. We have decided to choose our model that helped us best solve the problem we are dealing with. Since we are dealing with a classification problem, we have zeroed down the models to Logistic Regression, Decision Trees and Neural Network as they are most relevant to our problem statement.

Among them, we have chosen “*Neural Network*” method as the best model for our project. Since, we needed to identify the unhappy people among the respondents, we selected the model that provided us with the highest accuracy of 0’s. Neural Network model provided us with a Prediction of 0’s accuracy of over 57%, highest among all the models tested after varying multiple parameters. This improved prediction of unhappy people allowed us to perform clustering among them thereby allowing for better targeting of government programs.

We have also observed that the accuracy of models is acutely low when these models were performed on 20113 and 2015 data. Therefore, we excluded data from these two years from our models.

We have chosen Random Forest method as the optimum model for our dataset. Not only has this given us good accuracy, the classification tree method is best suited for our business case as we intend to classify as either happy or not happy. This method gives us the flexibility. This model also avoids all the sampling biases since we have worked on a survey data.

## CONCLUSIONS

Our model suggests that a sense of community, not a fancy house with all the material things in it, is the most important determinant of happiness. What mattered most was feeling of security and a neighborhood that promotes social interaction. “Satisfaction with Somerville” came out to be a crucial predictor in all our models.

Another important determinant for happiness is “Similarity with Other people”. The responders cared less about what they actually had and rated their happiness based on what they had in comparison to others in their community.

Interestingly, the top two determinants of happiness have been not-material attributes such as “Satisfaction with neighborhood” and “Similarity with other people” trumping many other material factors.

So, does money even buy happiness? In short, the answer is yes, but only to an extent.

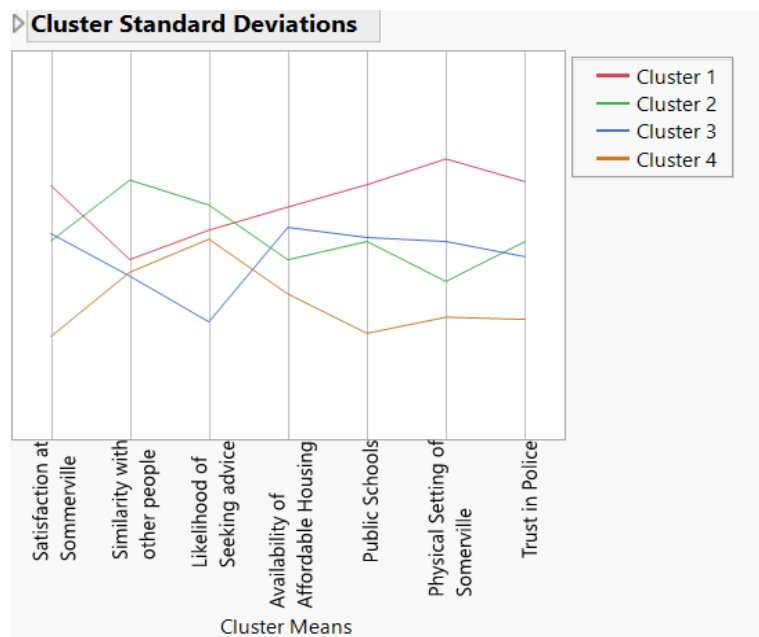
Increase in annual income has little effect on the general happiness. This is no mystery. This is called Easterlin Paradox. People compare their incomes with other people, and if others become richer, they feel less happy at any given income. The relationship between wealth and happiness is not strictly linear in the sense that it’s not the more money you have, the happier you are. Once you get to a certain point and you have your basic needs met, it takes more and more money to increase your happiness.

What was also interesting was the factors that did not matter when it comes to determination of happiness.

Race and Gender seem to have a minimal impact on whether a person is happy or not.

# RECOMMENDATIONS

- **Development of Neighbourhoods:** Expanding public spaces like parks give people space to disconnect from everyday life and connect with community and their inner self.
- **Reduction of inequality:** Constantly judging our worth and social standing by comparing ourselves with others, happiness is generally relative. Therefore, focus should not just be on increasing the overall size of the pie but on who gets how much of the pie.
- **Targeted Approach:** Instead of a blanket implementation of the above recommendations, we suggest a more targeted approach based on the weights of various determinants across different clusters.



For focus group 1, for example, the government can expend its resources on developing the physical setting of neighbourhoods while for focus group 2, people are more likely to be happier when inequalities are reduced. Such targeted spending will save valuable time and resources for the governments while maximizing the outcomes of their programs.

- Achieving happiness requires the same approach as losing weight. There is no magic bullet. There is no one thing that we can do that will make everyone happy for the rest of their lives. Our

experience working on the project only reinstated this fact. We would also be wrong to make generalized observations based on a survey confined to a single city. We would therefore recommend governments and Companies, above all, to undertake more such efforts on understanding what makes people happy because what gets measured gets managed. And it is important to make sure people are happy because happy people are productive people.



# REFERENCES

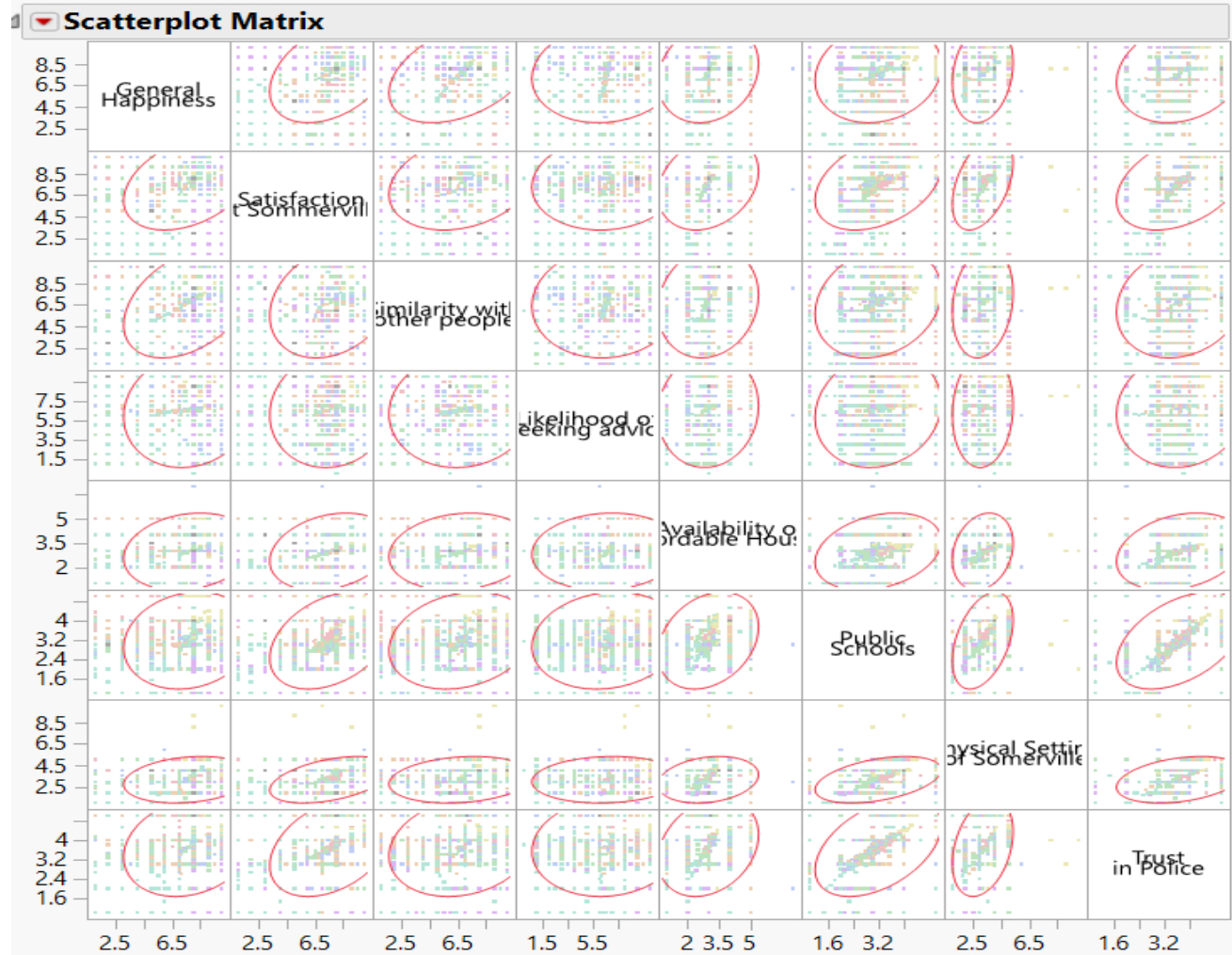
Source of Dataset - <https://catalog.data.gov/dataset/somerville-happiness-survey-responses-2011-2013-2015>

Information about the awards Somerville has received - <https://www.somervillema.gov/about>

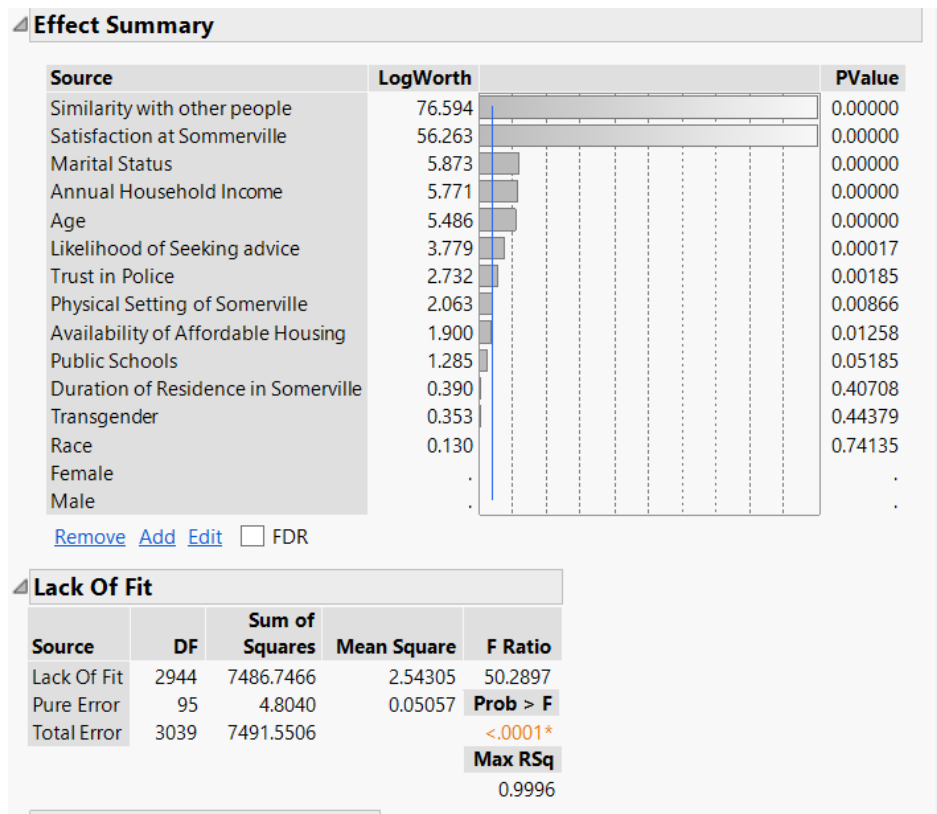
# APPENDIX

## Linear Regression Model Output:

The correlations are estimated by Row-wise method.



IMG-1.1



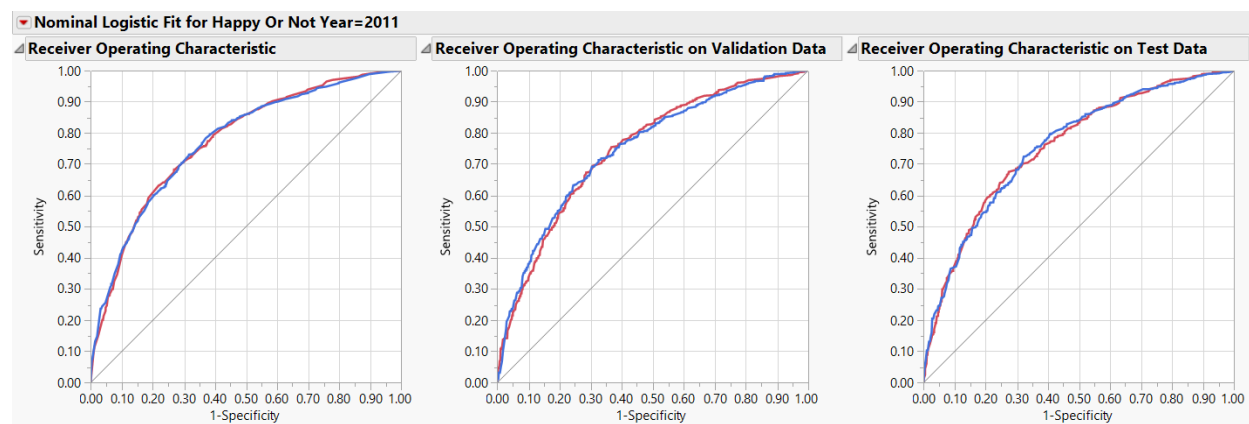
IMG-1.2

Summary of Fit		Crossvalidation			
RSquare	0.310026	Source	RSquare	RASE	Freq
RSquare Adj	0.301853	Training Set	0.3100	1.5606	3076
Root Mean Square Error	1.570075	Validation Set	0.2300	1.6552	1538
Mean of Response	7.636974	Test Set	0.2791	1.5714	1537
Observations (or Sum Wgts)	3076				

IMG-1.3

Parameter Estimates					
Term		Estimate	Std Error	t Ratio	Prob> t
Intercept		-0.0223304	0.153333	-0.15	0.8821
Satisfaction at Somerville		0.3085067	0.018988	16.25	<.0001*
Similarity with other people		0.2787605	0.014551	19.16	<.0001*
Likelihood of Seeking advice		0.0464401	0.012317	3.77	0.0002*
Availability of Affordable Housing		0.0765654	0.030663	2.50	0.0126*
Public Schools		-0.086734	0.04459	-1.95	0.0518
Physical Setting of Somerville		0.1025127	0.039022	2.63	0.0087*
Trust in Police		0.127687	0.040982	3.12	0.0019*
Transgender[0]		0.1843027	0.240631	0.77	0.4438
Male[0]	Biased	0.0115244	0.030168	0.38	0.7025
Female[0]	Zeroed	0	0	.	.
Age[18-21]		0.5469671	0.273571	2.00	0.0457*
Age[22-25]		0.3887036	0.112203	3.46	0.0005*
Age[26-30]		0.1770578	0.085306	2.08	0.0380*
Age[31-40]		-0.096112	0.078684	-1.22	0.2220
Age[41-50]		-0.262255	0.088494	-2.96	0.0031*
Age[51-60]		-0.298625	0.099164	-3.01	0.0026*
Age[61+]		-0.06042	0.098408	-0.61	0.5393
Marital Status[Divorced]		0.0294376	0.094024	0.31	0.7542
Marital Status[Married]		0.138249	0.070292	1.97	0.0493*
Marital Status[Never Married]		-0.250982	0.072141	-3.48	0.0005*
Marital Status[R]		0.2233041	0.163228	1.37	0.1714
Race[African-American]	Biased	-0.257365	0.245291	-1.05	0.2942
Race[Asian]	Biased	-0.624206	2.746893	-0.23	0.8203
Race[Asian/Pacific Islander]	Biased	-0.025024	0.207032	-0.12	0.9038
Race[Hispanic / Latino]	Biased	-0.357159	0.307443	-1.16	0.2454
Race[Hispanic, Asian/Pacific Islander]	Biased	-0.565069	1.13882	-0.50	0.6198
Race[Native American]	Biased	-0.144871	0.352311	-0.41	0.6810
Race[Other]	Biased	1.4045053	1.127204	1.25	0.2129
Race[R]	Biased	-0.018796	0.209487	-0.09	0.9285
Race[White]	Biased	0.7299278	1.586407	0.46	0.6455
Race[White, Hispanic / Latino]	Zeroed	0	0	.	.
Duration of Residence in Somerville[0-11 Years]		0.1141496	0.067441	1.69	0.0906
Duration of Residence in Somerville[11-17 Years]		0.0025281	0.093671	0.03	0.9785
Duration of Residence in Somerville[18+]		0.0016694	0.074683	0.02	0.9822
Annual Household Income[\$10,000 to \$49,999]		-0.208451	0.059452	-3.51	0.0005*
Annual Household Income[\$100,000 and up]		0.2864415	0.067669	4.23	<.0001*
Annual Household Income[\$50,000 to \$99,999]		0.0627005	0.057207	1.10	0.2731
Annual Household Income[Less than \$10,000]		-0.200754	0.119605	-1.68	0.0934

## Logistic Regression Model Output:



IMG-2.1

## Confusion Matrix

Training			Validation			Test		
Actual	Predicted		Actual	Predicted		Actual	Predicted	
Happy Or Not	Count		Happy Or Not	Count		Happy Or Not	Count	
	0	1		0	1		0	1
0	541	590	0	276	325	0	264	310
1	255	1690	1	139	798	1	128	835

## Effect Summary

Source	LogWorth	PValue
Similarity with other people	46.716	0.00000
Satisfaction at Somerville	29.414	0.00000
Annual Household Income	5.542	0.00000
Age	3.385	0.00041
Marital Status	3.314	0.00049
Physical Setting of Somerville	3.036	0.00092
Likelihood of Seeking advice	2.926	0.00119
Availability of Affordable Housing	0.979	0.10505
Public Schools	0.825	0.14963
Trust in Police	0.697	0.20084

## Fit Details

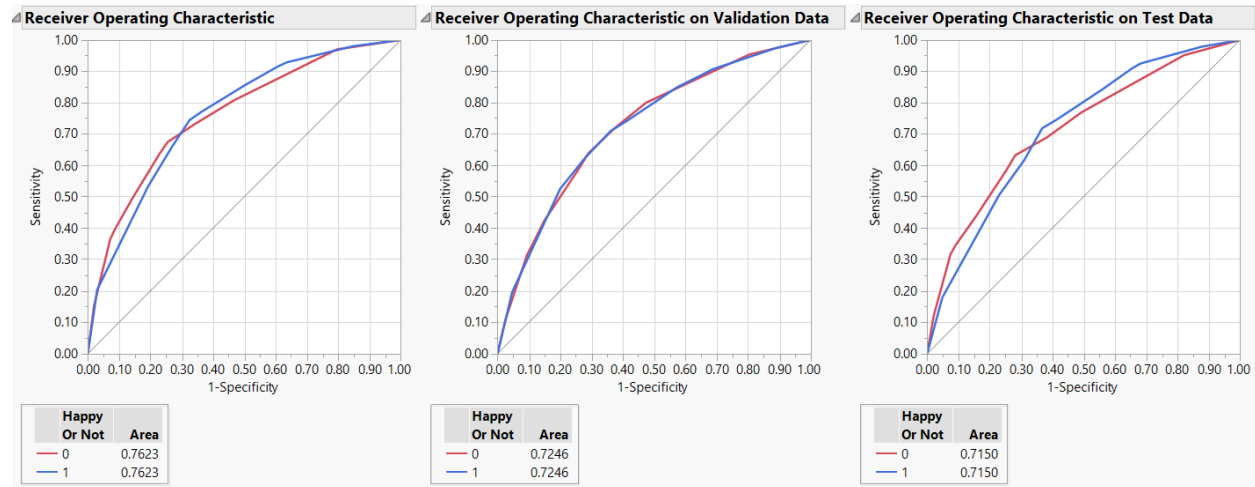
Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1708	0.1374	0.1491	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.2751	0.2277	0.2440	$(1 - (L(0)/L(\text{model}))^{(2/n)}) / (1 - L(0)^{(2/n)})$
Mean -Log p	0.5454	0.5771	0.5622	$\sum -\text{Log}(p[j]) / n$
RMSE	0.4283	0.4437	0.4359	$\sqrt{\sum (y[j] - p[j])^2 / n}$
Mean Abs Dev	0.3667	0.3812	0.3753	$\sum  y[j] - p[j]  / n$
Misclassification Rate	0.2747	0.3017	0.2850	$\sum (p[j] \neq pMax) / n$
N	3076	1538	1537	n

## Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	5.01937429	0.328162	233.95	<.0001*
Satisfaction at Somerville	-0.3216397	0.0292857	120.62	<.0001*
Similarity with other people	-0.3072115	0.0220757	193.66	<.0001*
Likelihood of Seeking advice	-0.0598066	0.0184436	10.51	0.0012*
Availability of Affordable Housing	-0.074021	0.0456653	2.63	0.1050
Public Schools	0.09799954	0.0682043	2.06	0.1508
Physical Setting of Somerville	-0.1930667	0.0583582	10.94	0.0009*
Trust in Police	-0.0782156	0.0611527	1.64	0.2009
Age[18-21]	-0.3210702	0.4135306	0.60	0.4375
Age[22-25]	-0.4514107	0.159093	8.05	0.0045*
Age[26-30]	-0.2566482	0.1169141	4.82	0.0282*
Age[31-40]	0.04066344	0.1107758	0.13	0.7136
Age[41-50]	0.24620651	0.1274792	3.73	0.0534
Age[51-60]	0.39023092	0.1393271	7.84	0.0051*
Age[61+]	0.00105778	0.1349005	0.00	0.9937
Marital Status[Divorced]	-0.085895	0.1406561	0.37	0.5414
Marital Status[Married]	-0.1350404	0.1054392	1.64	0.2003
Marital Status[Never Married]	0.32315119	0.107473	9.04	0.0026*
Marital Status[R]	-0.1009042	0.2439591	0.17	0.6792
Annual Household Income[\$10,000 to \$49,999]	0.28770866	0.0882344	10.63	0.0011*
Annual Household Income[\$100,000 and up]	-0.4395772	0.1015017	18.76	<.0001*
Annual Household Income[\$50,000 to \$99,999]	-0.1004892	0.0846273	1.41	0.2351
Annual Household Income[Less than \$10,000]	0.17534693	0.180152	0.95	0.3304

IMG-2.2

## Decision Tree Model Output:



IMG-3.1

	RSquare	N	Number of Splits
Training	0.170	3076	9
Validation	0.105	1538	
Test	0.110	1537	

Fit Details

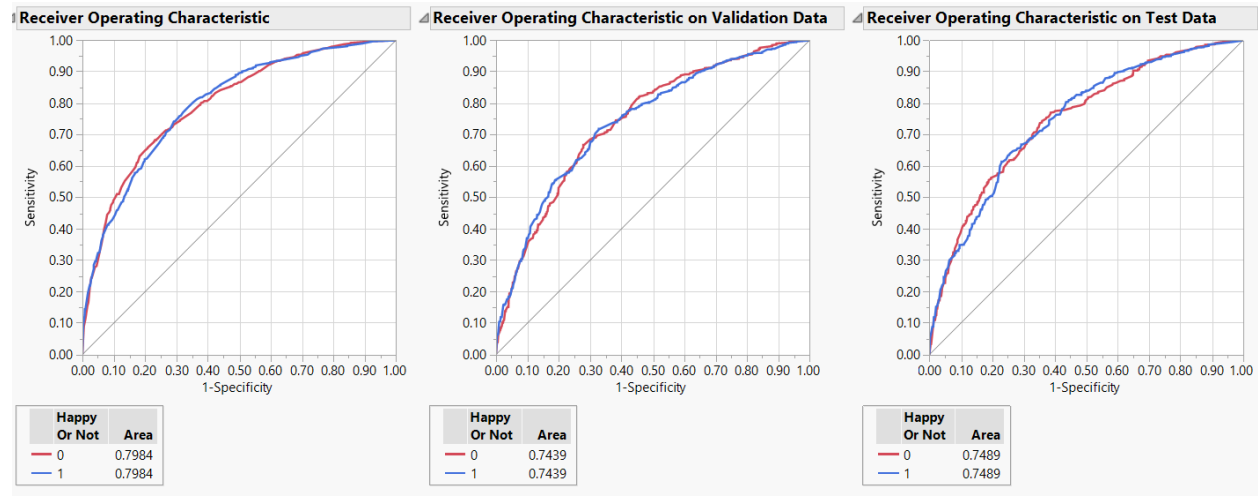
Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1700	0.1050	0.1105	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.2739	0.1776	0.1852	$(1 - (L(0)/L(\text{model}))^{2/n}) / (1 - L(0)^{2/n})$
Mean -Log p	0.5459	0.5989	0.5878	$\sum -\text{Log}(p[j])/n$
RMSE	0.4278	0.4529	0.4468	$\sqrt{\sum (y[j] - p[j])^2/n}$
Mean Abs Dev	0.3663	0.3904	0.3854	$\sum  y[j] - p[j] /n$
Misclassification Rate	0.2757	0.3160	0.3077	$\sum (p[j] \neq p\text{Max})/n$
N	3076	1538	1537	n

Confusion Matrix

Training			Validation			Test		
Actual	Predicted Count		Actual	Predicted Count		Actual	Predicted Count	
Happy Or Not	0	1	Happy Or Not	0	1	Happy Or Not	0	1
0	563	568	0	255	346	0	252	322
1	280	1665	1	140	797	1	151	812

IMG-3.2

## Boosted Tree Model Output:



IMG-4.1

### Specifications

Target Column:	Happy Or Not	Number of training rows:	3076
Validation Column:	Validation	Number of validation rows:	1538
Number of Layers:	50	Number of test rows:	1537
Splits per Tree:	3		
Learning Rate:	0.1		
Overfit Penalty:	0.0001		

### Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2046	0.1332	0.1444	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.3225	0.2214	0.2370	$(1 - (L(0)/L(\text{model}))^{2/n}) / (1 - L(0)^{2/n})$
Mean -Log p	0.5232	0.5799	0.5653	$\sum -\text{Log}(p[j]) / n$
RMSE	0.4172	0.4453	0.4380	$\sqrt{\sum (y[j] - p[j])^2 / n}$
Mean Abs Dev	0.3655	0.3914	0.3854	$\sum  y[j] - p[j]  / n$
Misclassification Rate	0.2507	0.3140	0.2856	$\sum (p[j] \neq p\text{Max}) / n$
N	3076	1538	1537	n

### Confusion Matrix

Training			Validation			Test		
Actual	Predicted		Actual	Predicted		Actual	Predicted	
Happy Or Not	0	1	Happy Or Not	0	1	Happy Or Not	0	1
0	574	557	0	263	338	0	251	323
1	214	1731	1	145	792	1	116	847

IMG-4.2

## Random Forest Model Output:

### Specifications

Target Column:	Happy Or Not	Training Rows:	3076
Validation Column:	Validation	Validation Rows:	1538
		Test Rows:	1537
Number of Trees in the Forest:	100	Number of Terms:	10
Number of Terms Sampled per Split:	2	Bootstrap Samples:	3076
		Minimum Splits per Tree:	10
		Minimum Size Split:	6

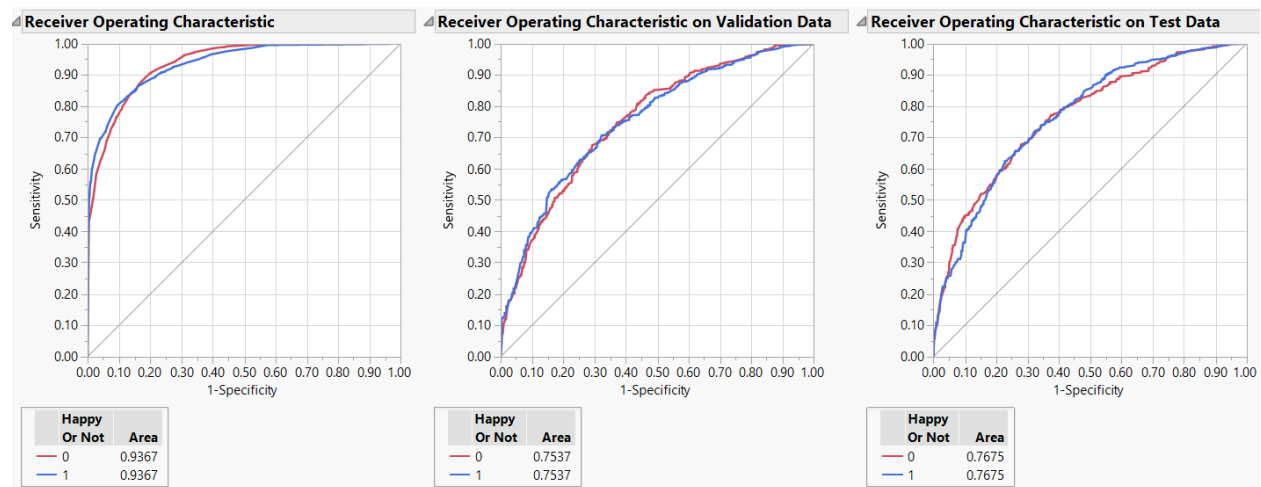
### Confusion Matrix

Training			Validation			Test		
Actual	Predicted		Actual	Predicted		Actual	Predicted	
Happy Or Not	0	1	Happy Or Not	0	1	Happy Or Not	0	1
0	722	409	0	261	340	0	260	314
1	88	1857	1	124	813	1	103	860

### Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3671	0.1474	0.1663	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.5234	0.2426	0.2690	$(1 - (L(0)/L(\text{model}))^{2/n}) / (1 - L(0)^{2/n})$
Mean -Log p	0.4163	0.5705	0.5509	$\sum -\text{Log}(p[j]) / n$
RMSE	0.3573	0.4414	0.4311	$\sqrt{\sum (y[j] - p[j])^2 / n}$
Mean Abs Dev	0.3168	0.3910	0.3819	$\sum  y[j] - p[j]  / n$
Misclassification Rate	0.1616	0.3017	0.2713	$\sum (p[j] \neq p\text{Max}) / n$
N	3076	1538	1537	n

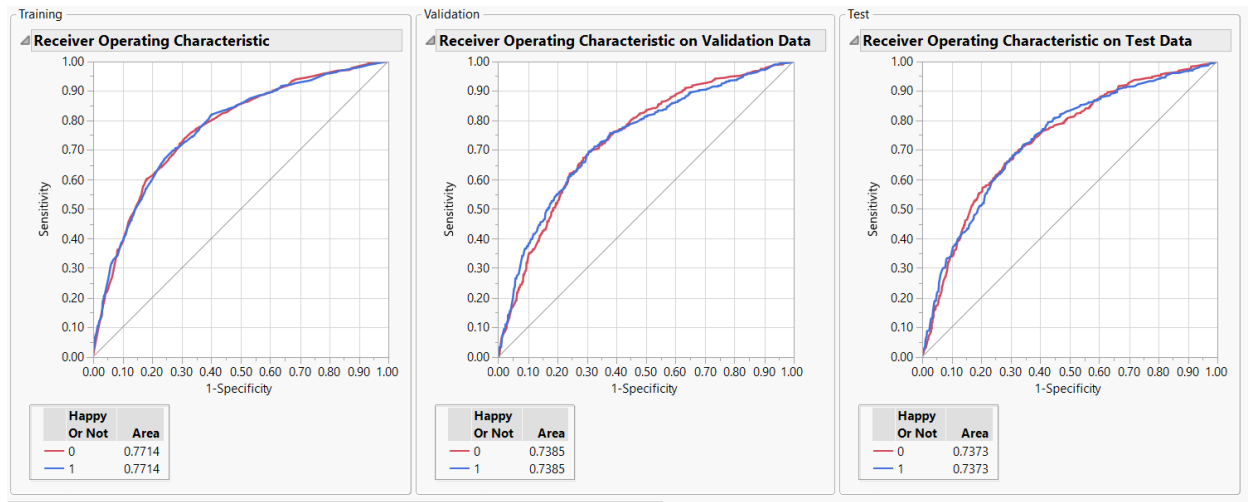
IMG-5.2



IMG-5.1



## Neural Network Model Output:



IMG-6.1

Model NTanH(3)NGaussian(3)					
Training		Validation		Test	
Happy Or Not		Happy Or Not		Happy Or Not	
Measures	Value	Measures	Value	Measures	Value
Generalized RSquare	0.2716793	Generalized RSquare	0.211729	Generalized RSquare	0.2069952
Entropy RSquare	0.1684695	Entropy RSquare	0.1269076	Entropy RSquare	0.1245674
RMSE	0.4282369	RMSE	0.4465408	RMSE	0.4435309
Mean Abs Dev	0.369944	Mean Abs Dev	0.3882092	Mean Abs Dev	0.3858821
Misclassification Rate	0.2652796	Misclassification Rate	0.3029909	Misclassification Rate	0.2934288
-LogLikelihood	1682.2884	-LogLikelihood	898.46489	-LogLikelihood	889.09083
Sum Freq	3076	Sum Freq	1538	Sum Freq	1537
Confusion Matrix		Confusion Matrix		Confusion Matrix	
Actual	Predicted	Actual	Predicted	Actual	Predicted
Happy	Count	Happy	Count	Happy	Count
Or Not	0 1	Or Not	0 1	Or Not	0 1
0	688 443	0	355 246	0	330 244
1	373 1572	1	220 717	1	207 756
Confusion Rates		Confusion Rates		Confusion Rates	
Actual	Predicted	Actual	Predicted	Actual	Predicted
Happy	Rate	Happy	Rate	Happy	Rate
Or Not	0 1	Or Not	0 1	Or Not	0 1
0	0.608 0.392	0	0.591 0.409	0	0.575 0.425
1	0.192 0.808	1	0.235 0.765	1	0.215 0.785

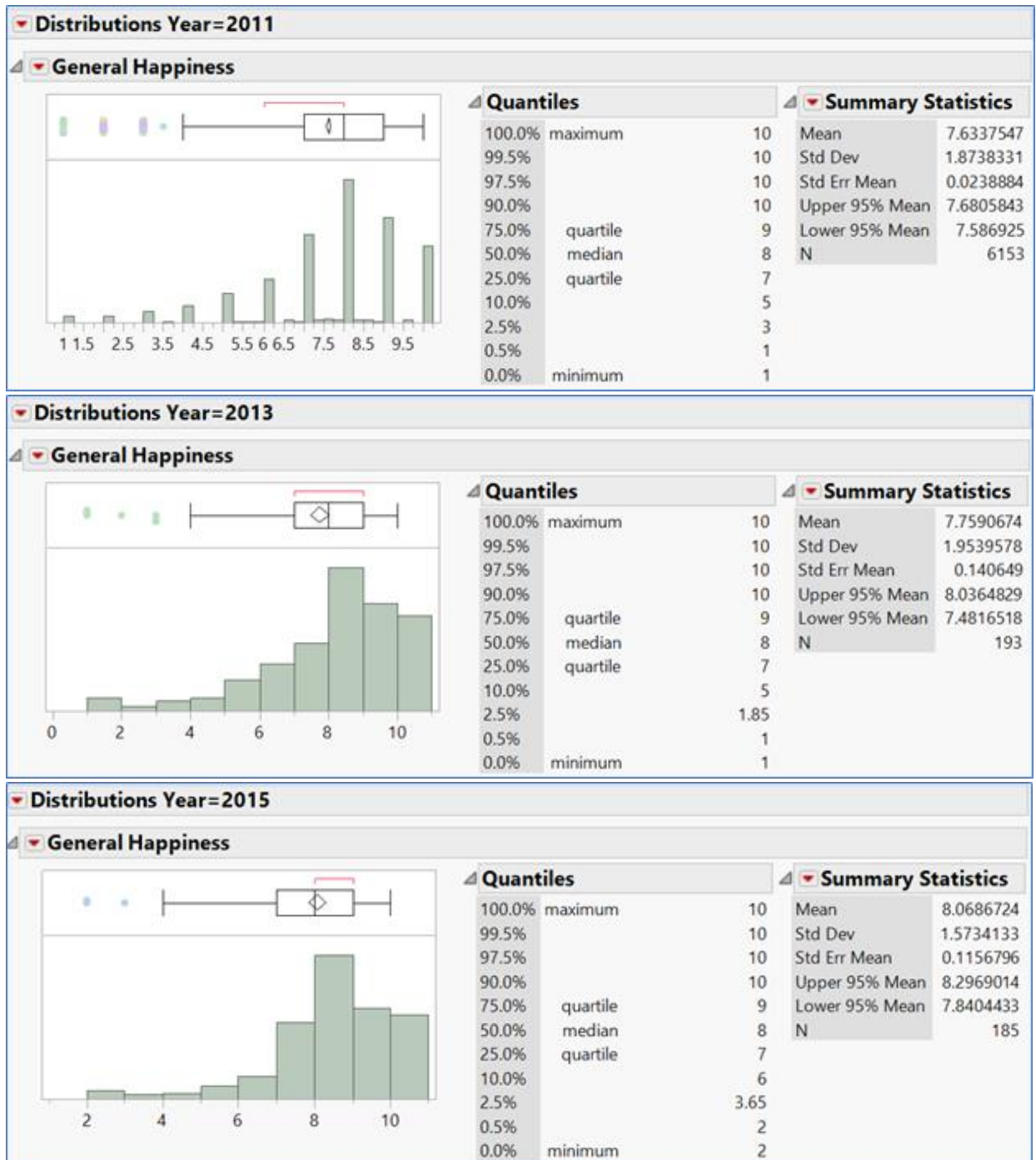
IMG-6.2

**Model Comparison:**

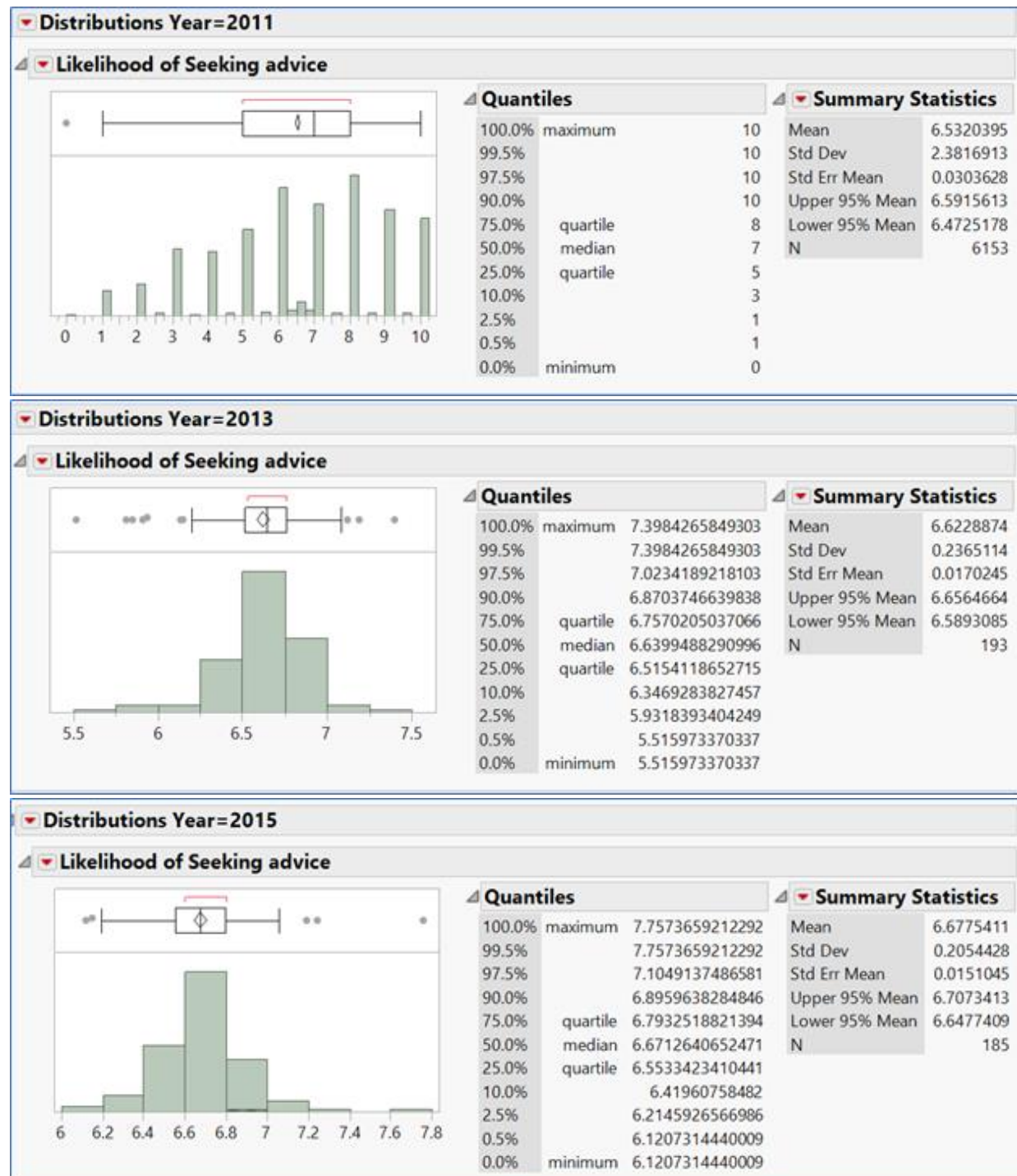
	Neural Nets	Random Forest	Logistic Regression
<b>Total Accuracy On Test Data</b>	70.6%	71.8%	71.5%
<b>Area Under the ROC curve</b>	0.737	0.767	0.757
<b>Prediction Of Zero's</b>	57%	44%	45%

**IMG-7**

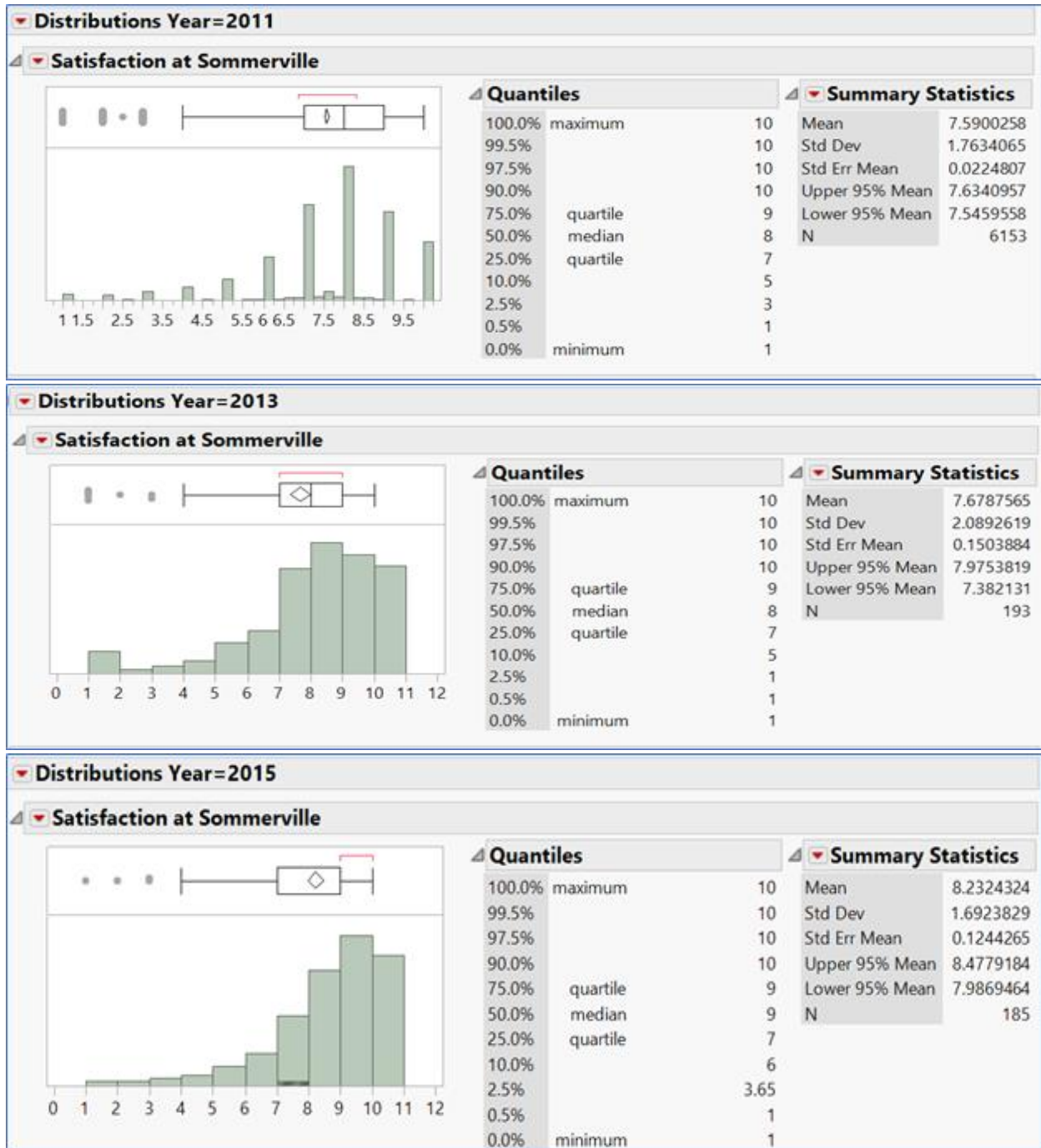
## Distribution of General Happiness Year Wise –



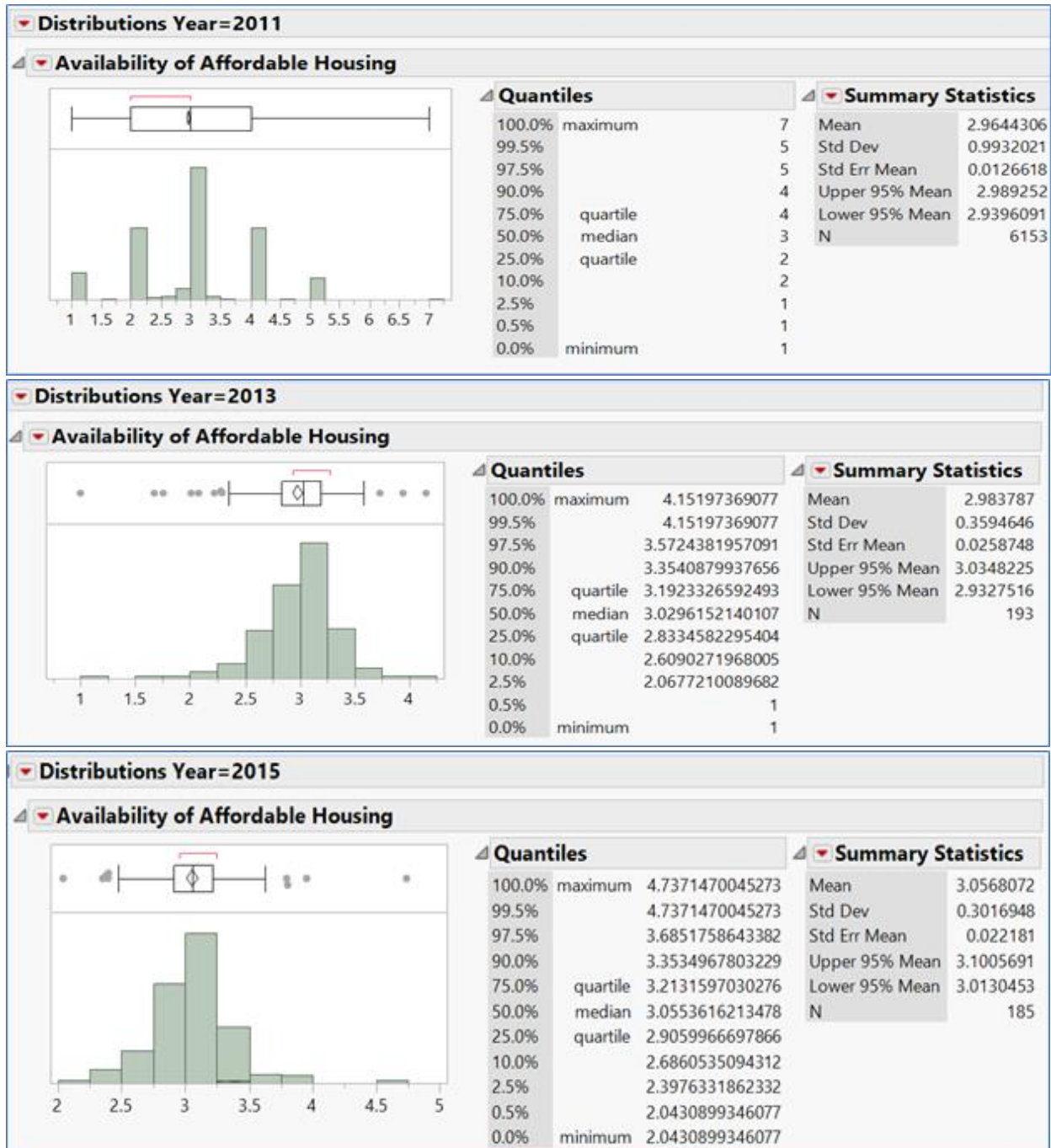
## Distribution of Likelihood of Seeking Advice Year Wise –



## Distribution of Satisfaction at Somerville Year Wise –

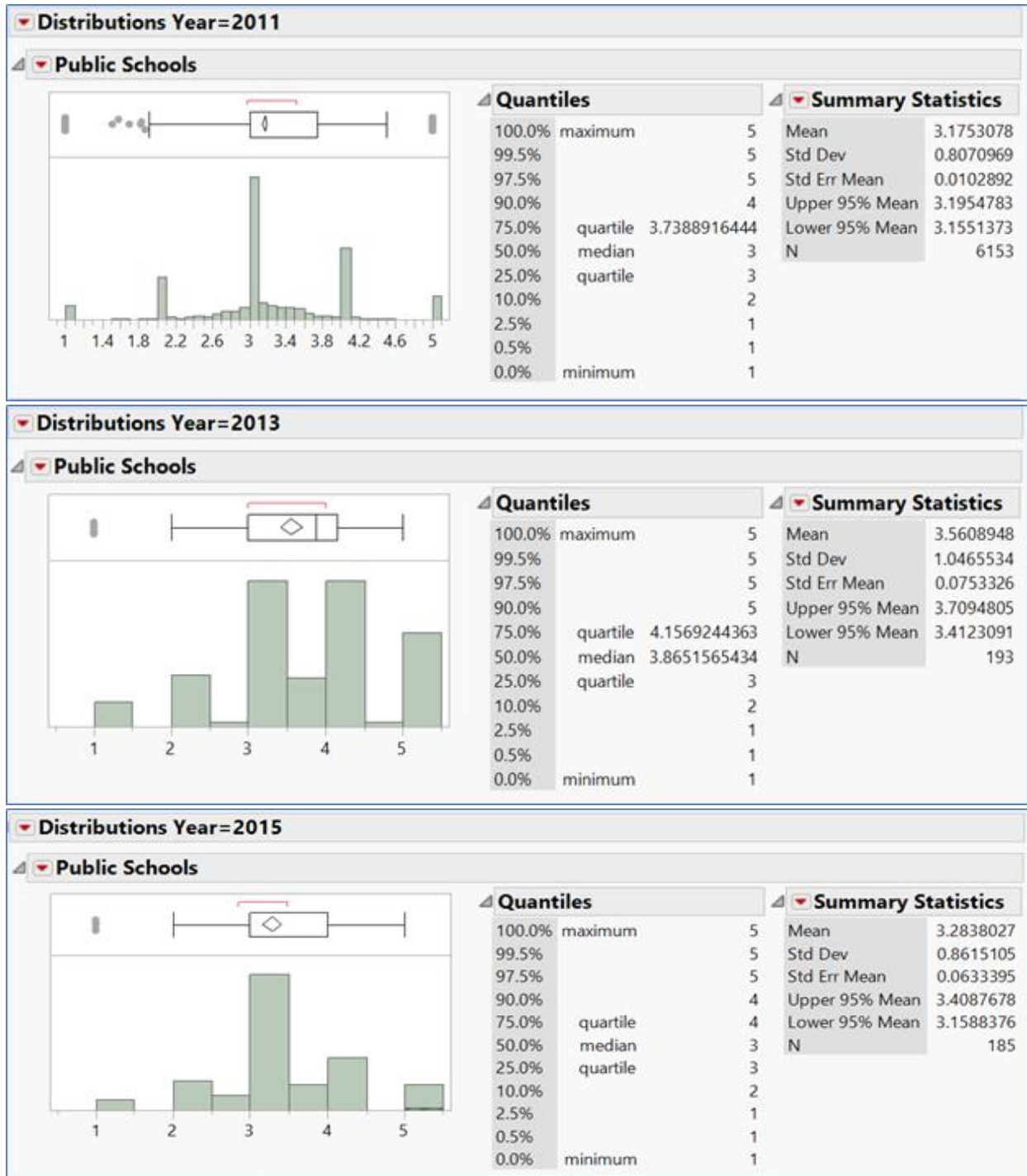


## Distribution of Availability of Affordable Housing Year Wise –

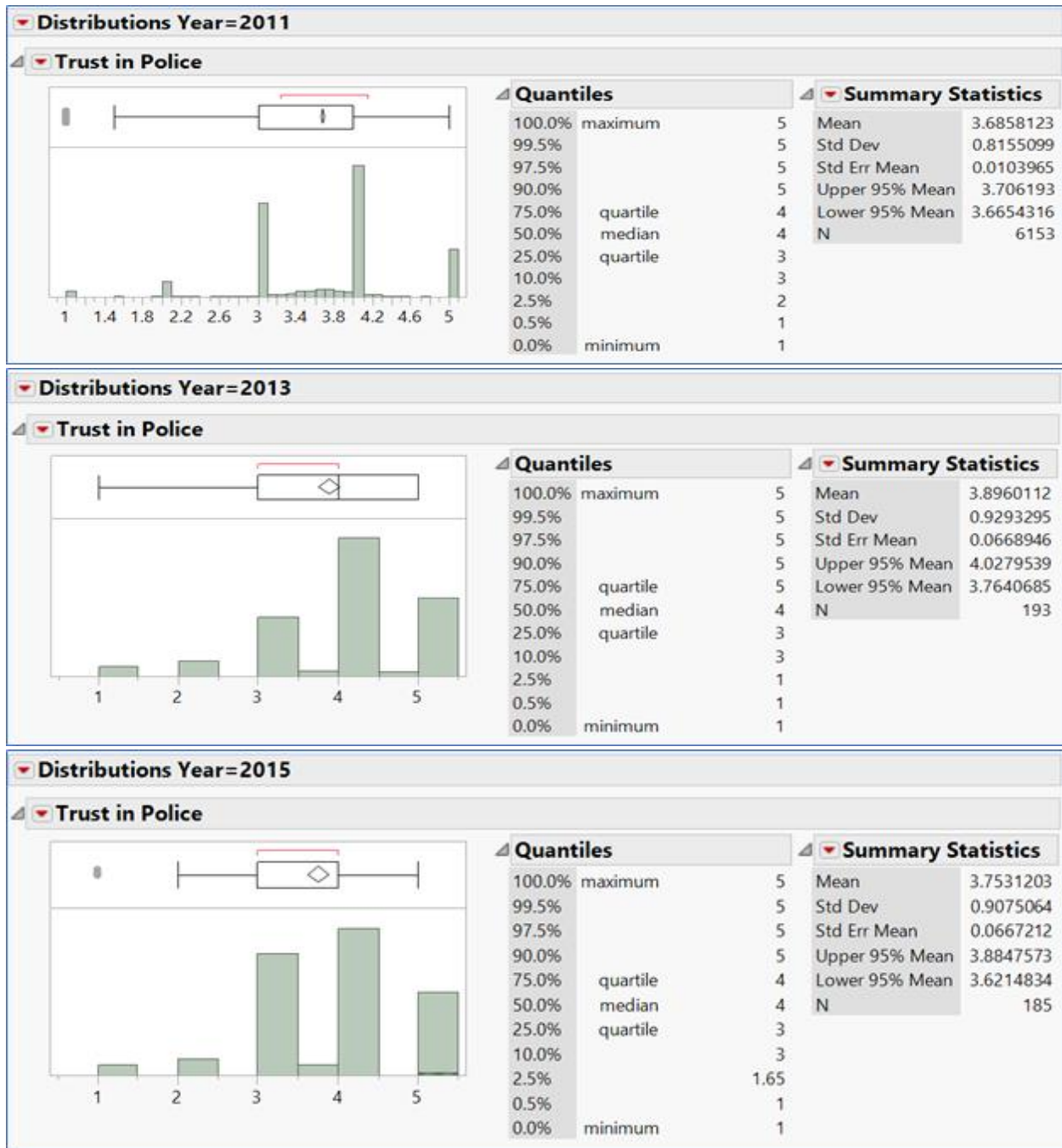




## Distribution of the Quality of Public Schools Year Wise –

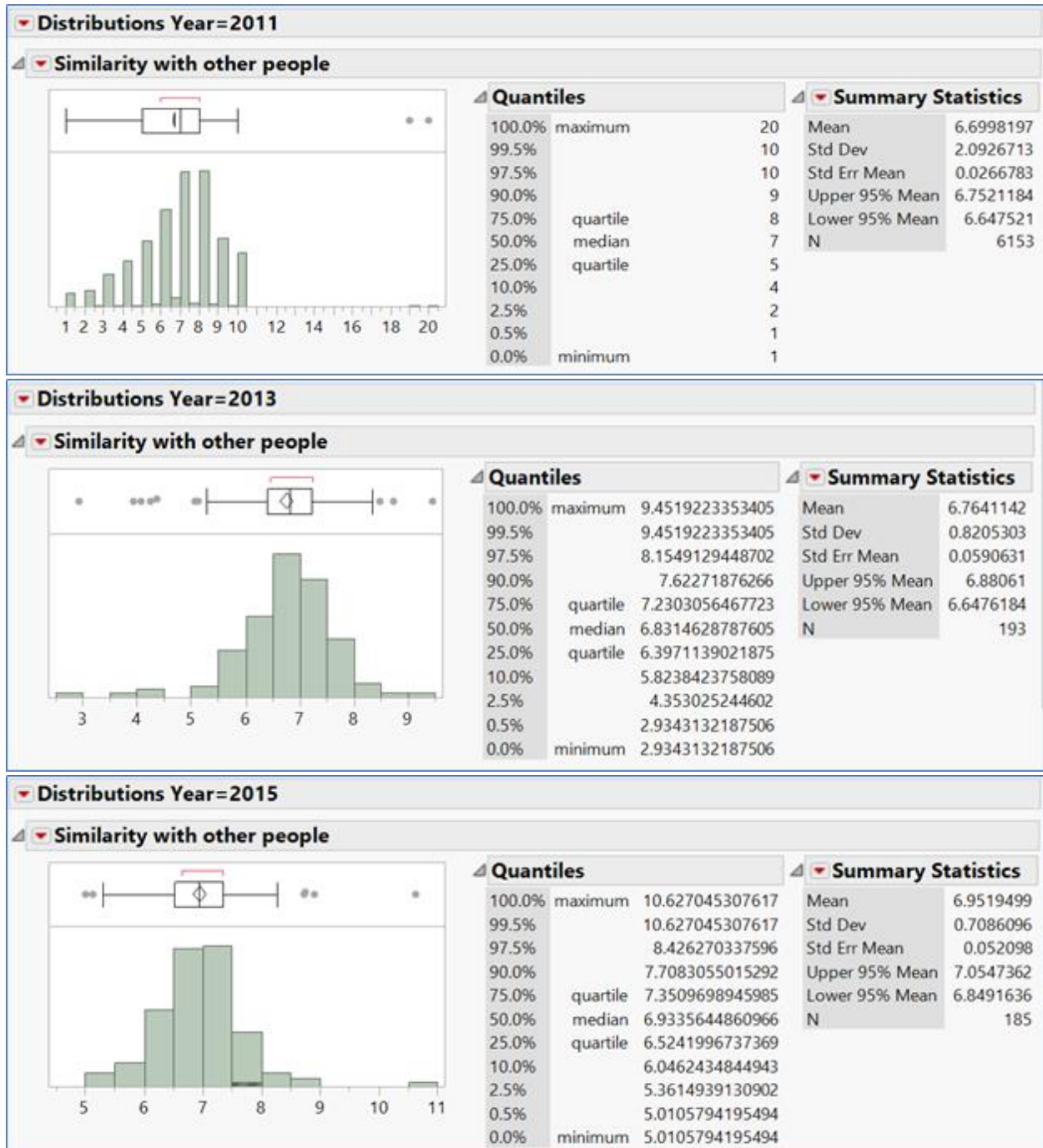


## Distribution of Trust in Police Year Wise –

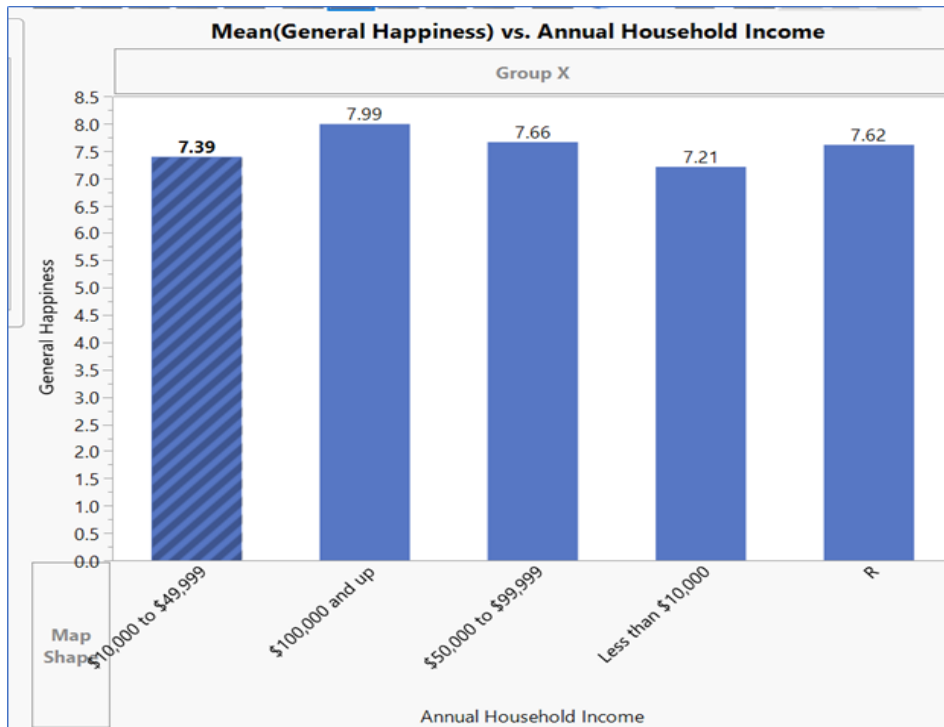




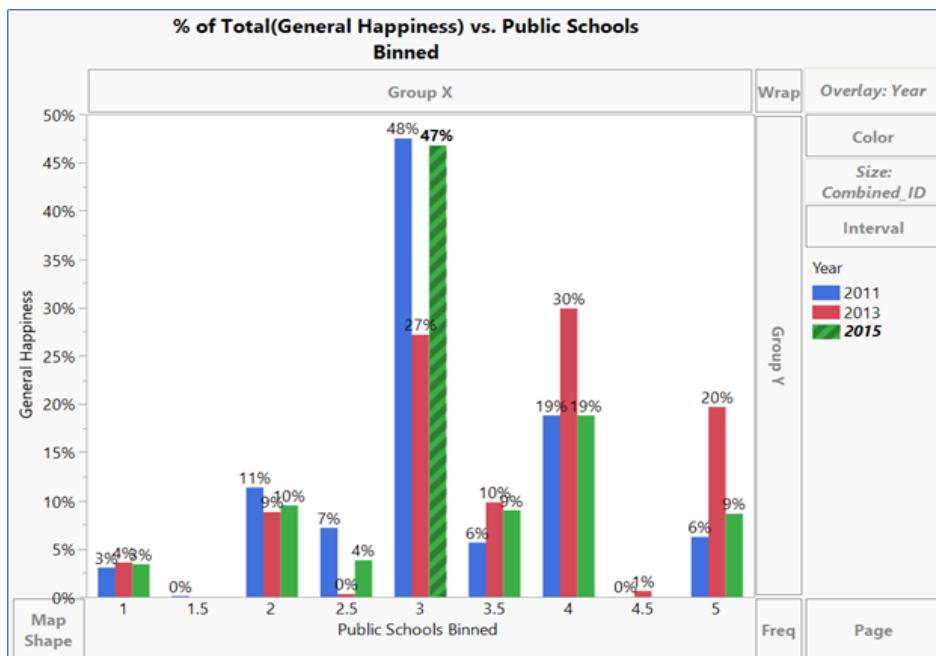
## Distribution of Similarity with Other people Year Wise –



### Distribution of Annual House Income with General Happiness -

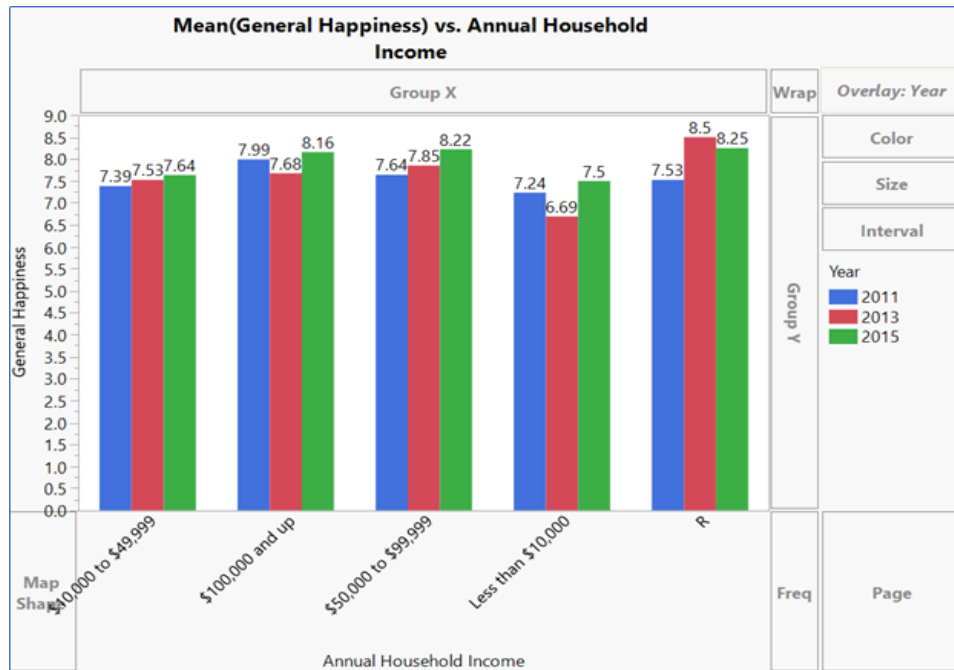


The below graph shows the percentage of total respondents who have been contributed toward a particular rating –



The below graph shows how different income groups have rated their general happiness by year wise

-



The below graph shows the yearwise distribution of general happiness level across the different ratings for “trust in police” -

