MGMT 635: Data Mining & Analysis for Managers

Project Report

SPRING 2020

Submitted To: Dr. Stephan P. Kudyba

TEAM

Abhishek Reddy - ak2626

Chandni Mandaviya - csm4

Namita Mahindra - nm648

Sirisha Bojjireddy - sb2423

Part-1 Input-data analysis

1.1 Objective:

You will receive an excel file with data that includes descriptions of sporting goods retail branches. You are required to analyze the data and adjust it (e.g. formats, content) to focus your research. In other words, you will be required to "troubleshoot" the data file and tell your data warehousing person what is wrong with it.

For example, are there errors; what variables are unnecessary, in order to perform a data mining analysis. You are to return the adjusted file along with any other requirements that are stipulated on that file.

1.2 Data items description

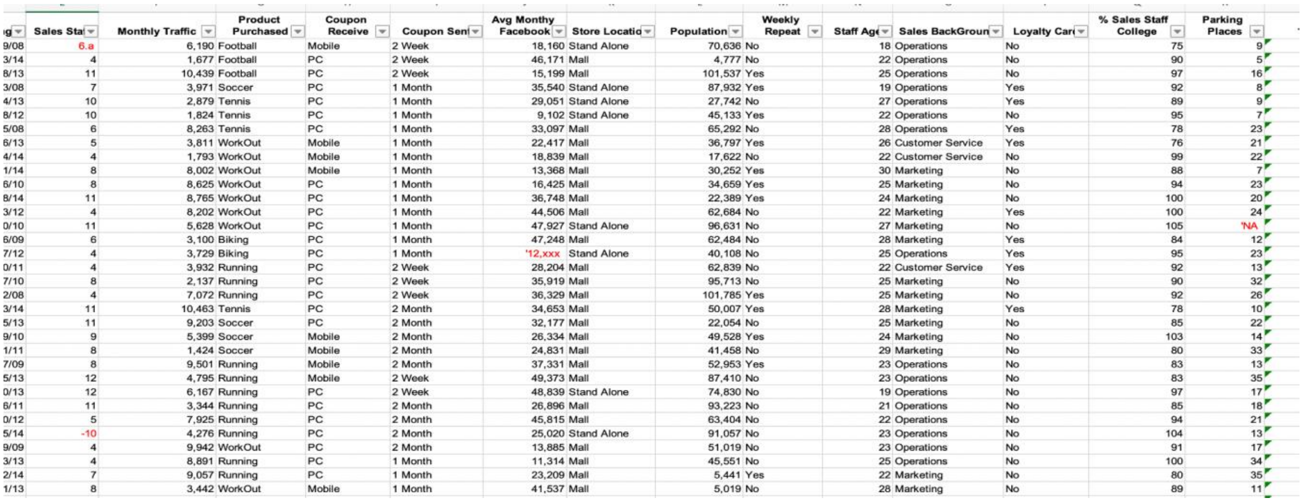
Column	**Description**	
Region	Region of where the store resides	Driver Variable
Store ID	The ID of a particular store	Driver Variable
Ship to Store	Whether the store receives internet orders (products purchased online) that can be picked up by customers	Driver Variable
Store Opening	The day the store opened its doors for business	Driver Variable
Sales Staff	The average amount of store workers that are present on a daily basis	Driver Variable
Monthly Traffic	The average amount of shoppers that visit the store on a monthly basis (this is highly correlated with Sales)	Driver Variable

Product Purchased	The product type that was purchsed by a customer	Driver Variable
Coupon Receive	How a particular customer retrieved a coupon that was emailed to them (either on a PC/laptop, or on a mobile device)	Driver Variable
Coupon Sent	The time deadline that was stipulated on the coupon sent to customers by the store (e.g. coupon must be redeemed in 2 weeks, 1 month, 2 months)	Driver Variable
Avg Monthly Facebook	The average monthly traffic that the store's facebook page receives	Driver Variable
Store Location	Whether a store is a stand alone building or is in a mall.	Driver Variable
Population	The total population within 20 miles of where the store resides	Driver Variable
Weekly Repeat	Whether the email sent by the store, containing the coupon was sent multiple times in a given week to customers	Driver Variable
Staff Age	The age of a staff worker in the retail store	Driver Variable
Sales Background	The day the store opened its doors for business	Driver Variable
Loyalty Card	Whether a customer used a loyalty card when purchasing a product	Driver Variable
% Sales Staff College	The percentage of the salesforce that has a bachelor’s degree from college	Driver Variable
Total Sales	Monthly sales generated by the particular branch (it has been determined that sales is a function of store traffic)	Target Variable

1.3 Data Cleansing Steps

Step1	Since it is asked to analyze on retailers in the North-East Region, we have filtered the Region only to North East Neglecting Midwest, South West, South East and West Coast. We deleted the regions except the North East.
Step 2	Sales Staff is the average amount of store workers that are present daily. It contains negative number and invalid number format. We have removed the insignificant values from the Sales Staff Column using a Data filter.
Step 3	In ‘Avg Month Facebook’ Column, there is one value which is not significant. And thus, removed that respective value using Data Filter.
Step 4	In the ‘Weekly Repeat’ column, there is a value which is wrong to the context of the column, we have filtered out that value from the column.
Step 5	In ‘Parking Places’, there is NA in one of the rows, we have filtered that respective row.
Step 6	Once the dataset is filtered and cleaned, we have now 77 rows of data ready to do Mining

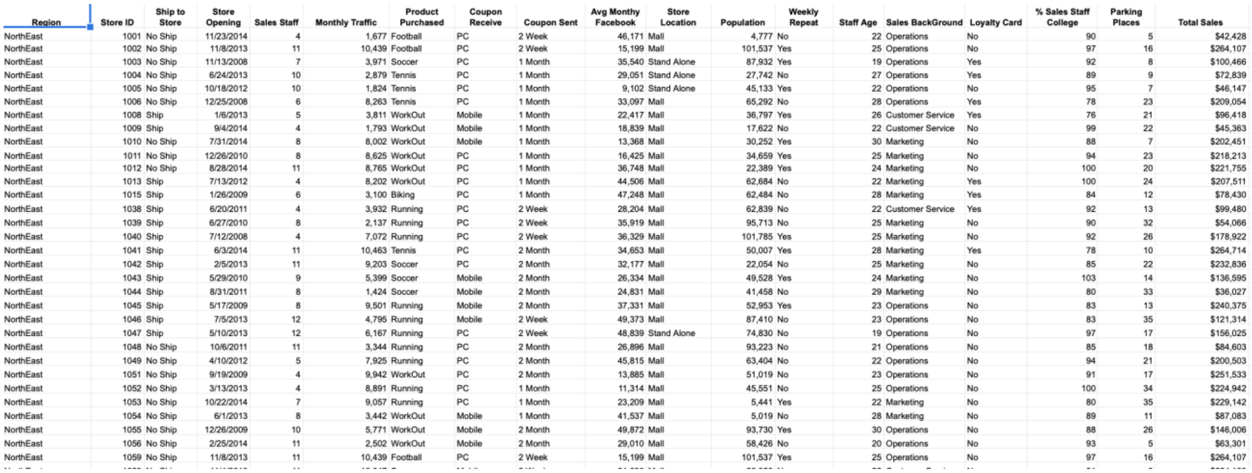
1.4 a) Screenshot of Original Dataset:



g	Sales Sta	Monthly Traffic	Product Purchased	Coupon Receive	Coupon Sen	Avg Monthy Facebook	Store Locatio	Population	Weekly Repeat	Staff Ag	Sales BackGroun	Loyalty Car	% Sales Staff College	Parking Places
9/08	6.a	6,190	Football	Mobile	2 Week	18,160	Stand Alone	70,636	No	18	Operations	No	75	9
3/14	4	1,677	Football	PC	2 Week	46,171	Mall	4,777	No	22	Operations	No	90	5
8/13	11	10,439	Football	PC	2 Week	15,199	Mall	101,537	Yes	25	Operations	No	97	16
3/08	7	3,971	Soccer	PC	1 Month	35,540	Stand Alone	87,932	Yes	19	Operations	Yes	92	8
4/13	10	2,879	Tennis	PC	1 Month	29,051	Stand Alone	27,742	No	27	Operations	Yes	89	9
8/12	10	1,824	Tennis	PC	1 Month	9,102	Stand Alone	45,133	Yes	22	Operations	No	95	7
5/08	6	8,263	Tennis	PC	1 Month	33,097	Mall	65,292	No	28	Operations	Yes	78	23
6/13	5	3,811	WorkOut	Mobile	1 Month	22,417	Mall	36,797	Yes	26	Customer Service	Yes	76	21
4/14	4	1,793	WorkOut	Mobile	1 Month	18,839	Mall	17,622	No	22	Customer Service	No	99	22
1/14	8	8,002	WorkOut	Mobile	1 Month	13,368	Mall	30,252	Yes	30	Marketing	No	88	7
6/10	8	8,625	WorkOut	PC	1 Month	16,425	Mall	34,659	Yes	25	Marketing	No	94	23
8/14	11	8,765	WorkOut	PC	1 Month	36,748	Mall	22,389	Yes	24	Marketing	No	100	20
3/12	4	8,202	WorkOut	PC	1 Month	44,506	Mall	62,684	No	22	Marketing	Yes	100	24
0/10	11	5,628	WorkOut	PC	1 Month	47,927	Stand Alone	96,631	No	27	Marketing	No	105	'NA
6/09	6	3,100	Biking	PC	1 Month	47,248	Mall	62,484	No	28	Marketing	Yes	84	12
7/12	4	3,729	Biking	PC	1 Month	'12,xxx	Stand Alone	40,108	No	25	Operations	Yes	95	23
0/11	4	3,932	Running	PC	2 Week	28,204	Mall	62,839	No	22	Customer Service	Yes	92	13
7/10	8	2,137	Running	PC	2 Week	35,919	Mall	95,713	No	25	Marketing	No	90	32
2/08	4	7,072	Running	PC	2 Week	36,329	Mall	101,785	Yes	25	Marketing	No	92	26
3/14	11	10,463	Tennis	PC	2 Month	34,653	Mall	50,007	Yes	28	Marketing	Yes	78	10
5/13	11	9,203	Soccer	PC	2 Month	32,177	Mall	22,054	No	25	Marketing	No	85	22
9/10	9	5,399	Soccer	Mobile	2 Month	26,334	Mall	49,528	Yes	24	Marketing	No	103	14
1/11	8	1,424	Soccer	Mobile	2 Month	24,831	Mall	41,458	No	29	Marketing	No	80	33
7/09	8	9,501	Running	Mobile	2 Month	37,331	Mall	52,953	Yes	23	Operations	No	83	13
5/13	12	4,795	Running	Mobile	2 Week	49,373	Mall	87,410	No	23	Operations	No	83	35
0/13	12	6,167	Running	PC	2 Week	48,839	Stand Alone	74,830	No	19	Operations	No	97	17
6/11	11	3,344	Running	PC	2 Month	26,896	Mall	93,223	No	21	Operations	No	85	18
0/12	5	7,925	Running	PC	2 Month	45,815	Mall	63,404	No	22	Operations	No	94	21
5/14	-10	4,276	Running	PC	2 Month	25,020	Stand Alone	91,057	No	23	Operations	No	104	13
9/09	4	9,942	WorkOut	PC	2 Month	13,885	Mall	51,019	No	23	Operations	No	91	17
3/13	4	8,891	Running	PC	1 Month	11,314	Mall	45,551	No	25	Operations	No	100	34
2/14	7	9,057	Running	PC	1 Month	23,209	Mall	5,441	Yes	22	Marketing	No	80	35
1/13	8	3,442	WorkOut	Mobile	1 Month	41,537	Mall	5,019	No	28	Marketing	No	89	11

Original dataset statistics
Number of rows -163
Number of columns - 19

1.4 b) Screenshot of Cleaned Dataset:



Region	Store ID	Ship to Store	Store Opening	Sales Staff	Monthly Traffic	Product Purchased	Coupon Receive	Coupon Sent	Avg Monthy Facebook	Store Location	Population	Weekly Repeat	Staff Age	Sales BackGround	Loyalty Card	% Sales Staff College	Parking Places	Total Sales
NorthEast	1001	No Ship	11/23/2014	4	1,677	Football	PC	2 Week	46,171	Mall	4,777	No	22	Operations	No	90	5	$42,428
NorthEast	1002	No Ship	11/8/2013	11	10,439	Football	PC	2 Week	15,199	Mall	101,537	Yes	25	Operations	No	97	16	$264,107
NorthEast	1003	No Ship	11/13/2008	7	3,971	Soccer	PC	1 Month	35,540	Stand Alone	87,932	Yes	19	Operations	Yes	92	8	$100,466
NorthEast	1004	No Ship	6/24/2013	10	2,879	Tennis	PC	1 Month	29,051	Stand Alone	27,742	No	27	Operations	Yes	89	9	$72,839
NorthEast	1005	No Ship	10/18/2012	10	1,824	Tennis	PC	1 Month	9,102	Stand Alone	45,133	Yes	22	Operations	No	95	7	$46,147
NorthEast	1006	No Ship	12/25/2008	6	8,263	Tennis	PC	1 Month	33,097	Mall	65,292	No	28	Operations	Yes	78	23	$209,054
NorthEast	1008	Ship	1/6/2013	5	3,811	WorkOut	Mobile	1 Month	22,417	Mall	36,797	Yes	26	Customer Service	Yes	76	21	$96,418
NorthEast	1009	Ship	9/4/2014	4	1,793	WorkOut	Mobile	1 Month	18,839	Mall	17,622	No	22	Customer Service	No	99	22	$45,363
NorthEast	1010	No Ship	7/31/2014	8	8,002	WorkOut	Mobile	1 Month	13,368	Mall	30,252	Yes	30	Marketing	No	88	7	$202,451
NorthEast	1011	No Ship	12/26/2010	8	8,625	WorkOut	PC	1 Month	16,425	Mall	34,659	Yes	25	Marketing	No	94	23	$218,213
NorthEast	1012	No Ship	8/28/2014	11	8,765	WorkOut	PC	1 Month	36,748	Mall	22,389	Yes	24	Marketing	No	100	20	$221,755
NorthEast	1013	Ship	7/13/2012	4	8,202	WorkOut	PC	1 Month	44,506	Mall	62,684	No	22	Marketing	Yes	100	24	$207,511
NorthEast	1015	Ship	1/26/2009	6	3,100	Biking	PC	1 Month	47,248	Mall	62,484	No	28	Marketing	Yes	84	12	$78,430
NorthEast	1038	Ship	6/20/2011	4	3,932	Running	PC	2 Week	28,204	Mall	62,839	No	22	Customer Service	Yes	92	13	$99,480
NorthEast	1039	Ship	6/27/2010	8	2,137	Running	PC	2 Week	35,919	Mall	95,713	No	25	Marketing	No	90	32	$54,066
NorthEast	1040	Ship	7/12/2008	4	7,072	Running	PC	2 Week	36,329	Mall	101,785	Yes	25	Marketing	No	92	26	$178,922
NorthEast	1041	Ship	6/3/2014	11	10,463	Tennis	PC	2 Month	34,653	Mall	50,007	Yes	28	Marketing	Yes	78	10	$264,714
NorthEast	1042	Ship	2/5/2013	11	9,203	Soccer	PC	2 Month	32,177	Mall	22,054	No	25	Marketing	No	85	22	$232,836
NorthEast	1043	Ship	5/29/2010	9	5,399	Soccer	Mobile	2 Month	26,334	Mall	49,528	Yes	24	Marketing	No	103	14	$136,595
NorthEast	1044	Ship	8/31/2011	8	1,424	Soccer	Mobile	2 Month	24,831	Mall	41,458	No	29	Marketing	No	80	33	$36,027
NorthEast	1045	Ship	5/17/2009	8	9,501	Running	Mobile	2 Month	37,331	Mall	52,953	Yes	23	Operations	No	83	13	$240,375
NorthEast	1046	Ship	7/5/2013	12	4,795	Running	Mobile	2 Week	49,373	Mall	87,410	No	23	Operations	No	83	35	$121,314
NorthEast	1047	Ship	5/10/2013	12	6,167	Running	PC	2 Week	48,839	Stand Alone	74,830	No	19	Operations	No	97	17	$156,025
NorthEast	1048	No Ship	10/6/2011	11	3,344	Running	PC	2 Month	26,896	Mall	93,223	No	21	Operations	No	85	18	$84,603
NorthEast	1049	No Ship	4/10/2012	5	7,925	Running	PC	2 Month	45,815	Mall	63,404	No	22	Operations	No	94	21	$200,503
NorthEast	1051	No Ship	9/19/2009	4	9,942	WorkOut	PC	2 Month	13,885	Mall	51,019	No	23	Operations	No	91	17	$251,533
NorthEast	1052	No Ship	3/13/2013	4	8,891	Running	PC	1 Month	11,314	Mall	45,551	No	25	Operations	No	100	34	$224,942
NorthEast	1053	No Ship	10/22/2014	7	9,057	Running	PC	1 Month	23,209	Mall	5,441	Yes	22	Marketing	No	80	35	$229,142
NorthEast	1054	No Ship	6/1/2013	8	3,442	WorkOut	Mobile	1 Month	41,537	Mall	5,019	No	28	Marketing	No	89	11	$87,083
NorthEast	1055	No Ship	12/26/2009	10	5,771	WorkOut	Mobile	2 Month	49,872	Mall	93,730	Yes	30	Operations	No	88	26	$146,006
NorthEast	1056	No Ship	2/25/2014	11	2,502	WorkOut	Mobile	2 Month	29,010	Mall	58,426	No	20	Operations	No	93	5	$63,301
NorthEast	1059	No Ship	11/8/2013	11	10,439	Football	PC	2 Week	15,199	Mall	101,537	Yes	25	Operations	No	97	16	$264,107

Statistics after cleaning
Number of rows -79
Number of columns - 19

Part 2- Regression Analysis

2.1 Objective:

You will be given an Excel file with a number of variables to run a regression analysis. You will use the statistical output of the model to help make decisions on where and the tactics you need to pursue to generate the highest revenue for a restaurant. You are to analyze your regression results and devise a simple business plan using whatever information is critical to your strategic initiative. In this case you need to advise your company on the factors/variables that effect the revenue of your proposed restaurant. You are also required to use the results of your model to estimate the expected revenue of planned restaurants to be opened. Finally, you need to analyze the data file with a neural net methodology and compare the results to the regression analysis for expected revenue.

2.2 Regression Analysis

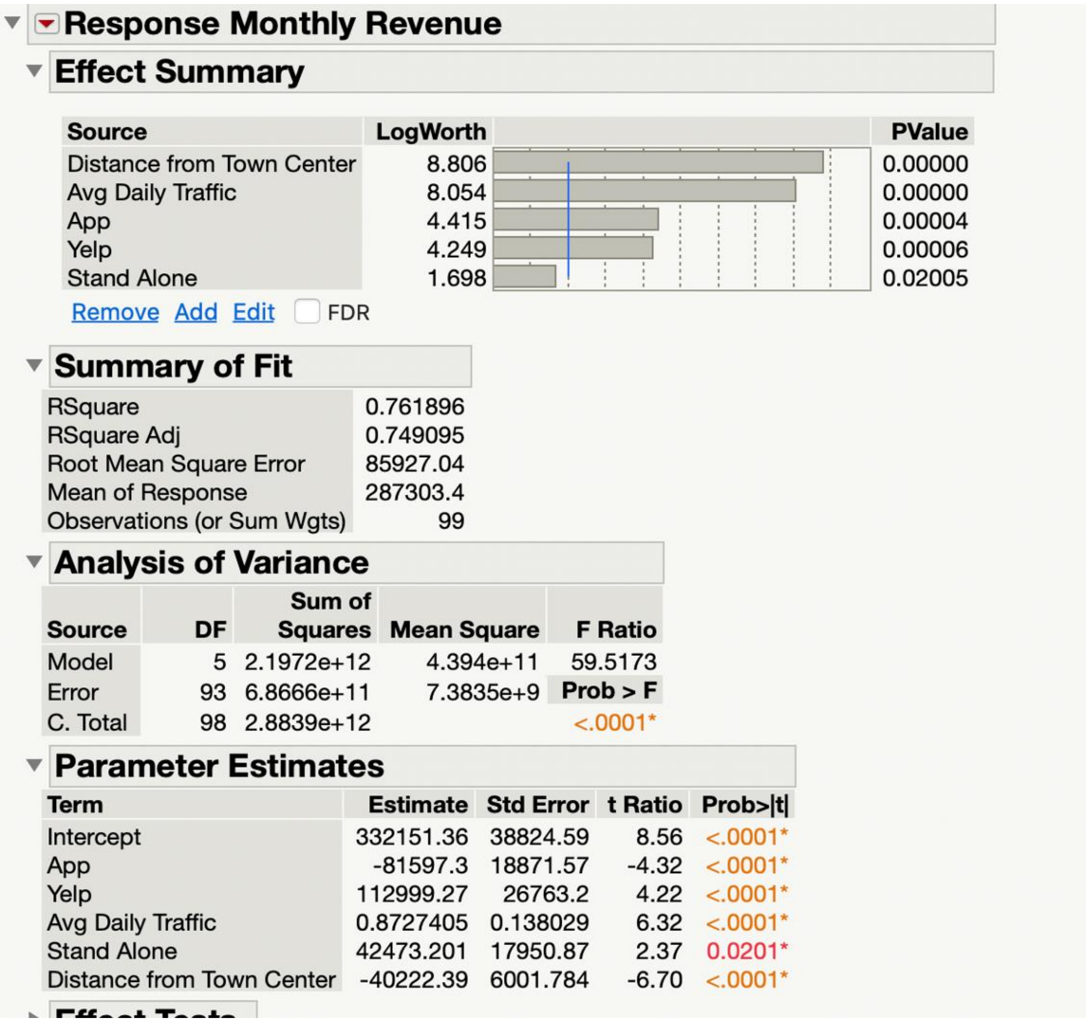
Process Description:

1. We have imported the training data set from the excel into JMP. We build a model using Linear regression including 'Monthly Revenue' as the target variable.
2. When we try to run the model by keeping the target variable and including all the variables as the explanatory variable, we found that the Restaurant column seems to be significant and all other column values are insignificant as all the other columns have P-values greater than 0.05.
3. So, we built the model neglecting the Restaurant column and included all the other explanatory variables.
4. The P-value of Standalone column is 0.17332, which is greater than 0.05, we have considered this variable as insignificant and re-build the model.

Future Prediction:

1. Thus, we have trained the dataset and concluded that the variables ‘Area income’ and ‘Restaurant’ are not significant to run the model.
2. We built the model for testing dataset and copied the predicted monthly revenue formula from the dataset and predicted the monthly revenue for testing dataset. The results are shown below in a screenshot of the predicted monthly revenue.
3. We have shown the predicted revenue in a separate column and applied the formula.

2.3 a) Screenshot of regression analysis in JMP:



Response Monthly Revenue

Effect Summary

Source	LogWorth	PValue
Distance from Town Center	8.806	0.00000
Avg Daily Traffic	8.054	0.00000
App	4.415	0.00004
Yelp	4.249	0.00006
Stand Alone	1.698	0.02005

Summary of Fit

RSquare	0.761896
RSquare Adj	0.749095
Root Mean Square Error	85927.04
Mean of Response	287303.4
Observations (or Sum Wgts)	99

Analysis of Variance

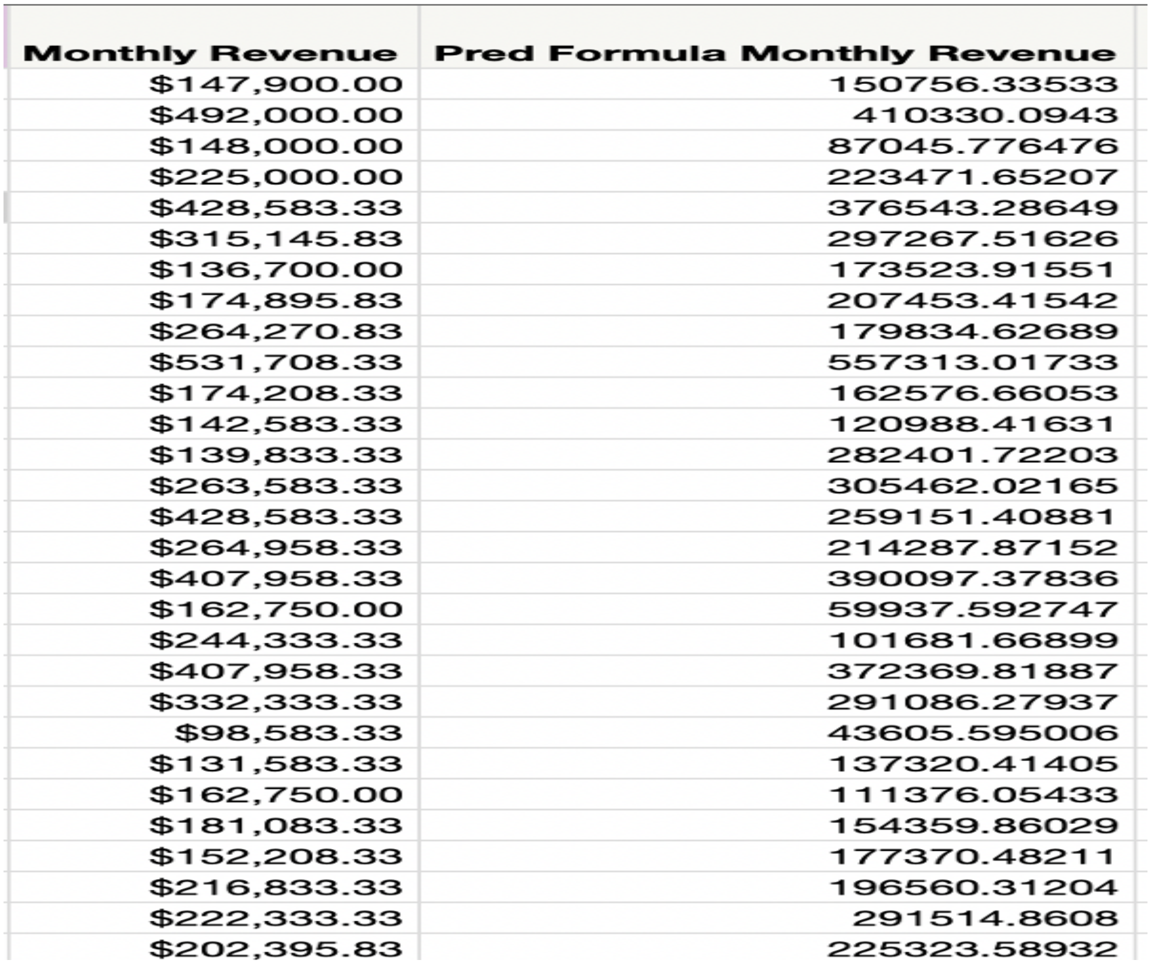
		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	5	2.1972e+12	4.394e+11	59.5173
Error	93	6.8666e+11	7.3835e+9	Prob > F
C. Total	98	2.8839e+12		<.0001*

Parameter Estimates

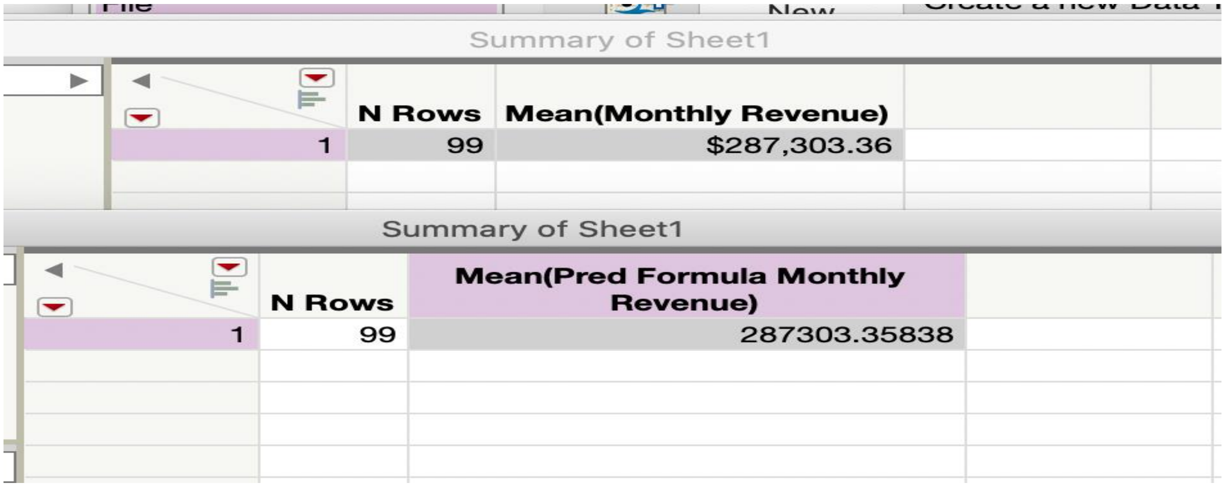
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	332151.36	38824.59	8.56	<.0001*
App	-81597.3	18871.57	-4.32	<.0001*
Yelp	112999.27	26763.2	4.22	<.0001*
Avg Daily Traffic	0.8727405	0.138029	6.32	<.0001*
Stand Alone	42473.201	17950.87	2.37	0.0201*
Distance from Town Center	-40222.39	6001.784	-6.70	<.0001*

Effect Tests

2.3 b) Screenshot of Monthly Revenue and Prediction Formula Monthly Revenue using Regression.



Monthly Revenue	Pred Formula Monthly Revenue
$147,900.00	150756.33533
$492,000.00	410330.0943
$148,000.00	87045.776476
$225,000.00	223471.65207
$428,583.33	376543.28649
$315,145.83	297267.51626
$136,700.00	173523.91551
$174,895.83	207453.41542
$264,270.83	179834.62689
$531,708.33	557313.01733
$174,208.33	162576.66053
$142,583.33	120988.41631
$139,833.33	282401.72203
$263,583.33	305462.02165
$428,583.33	259151.40881
$264,958.33	214287.87152
$407,958.33	390097.37836
$162,750.00	59937.592747
$244,333.33	101681.66899
$407,958.33	372369.81887
$332,333.33	291086.27937
$98,583.33	43605.595006
$131,583.33	137320.41405
$162,750.00	111376.05433
$181,083.33	154359.86029
$152,208.33	177370.48211
$216,833.33	196560.31204
$222,333.33	291514.8608
$202,395.83	225323.58932



Summary of Sheet1

	N Rows	Mean(Monthly Revenue)
1	99	$287,303.36

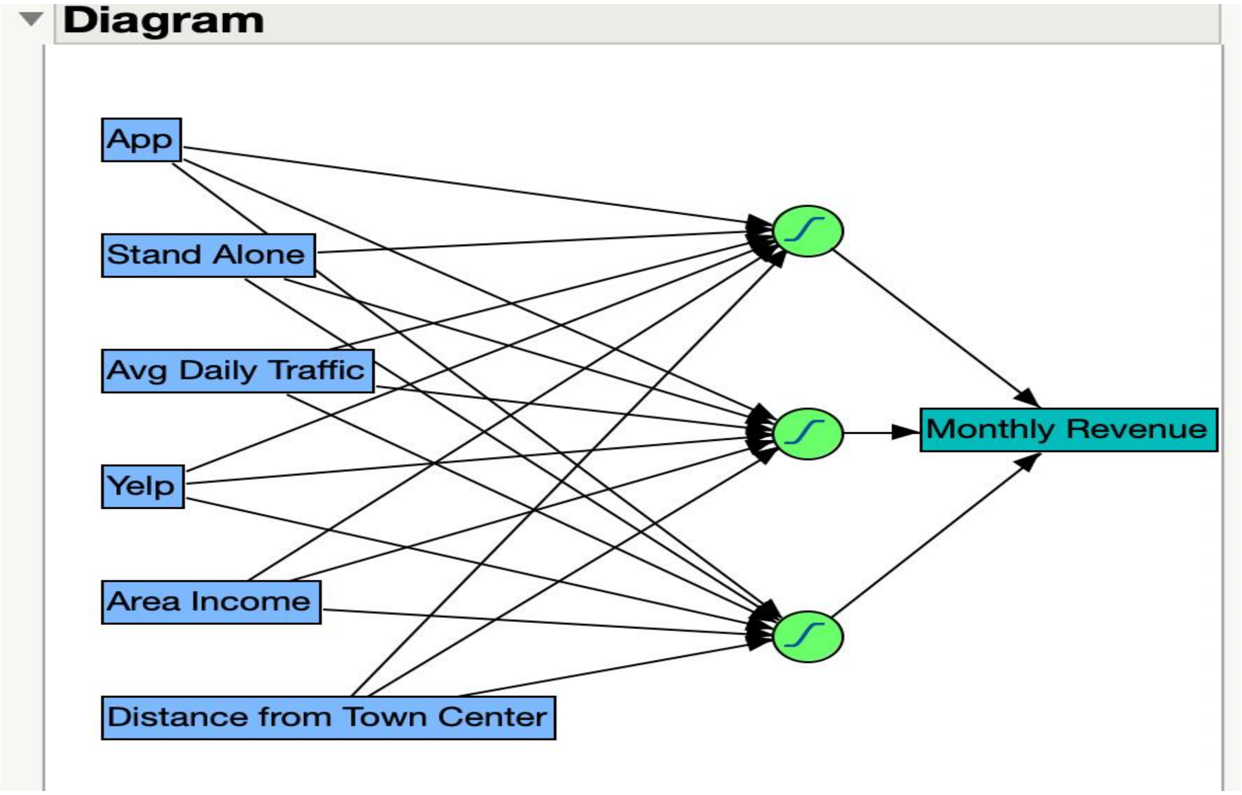
Summary of Sheet1

	N Rows	Mean(Pred Formula Monthly Revenue)
1	99	287303.35838

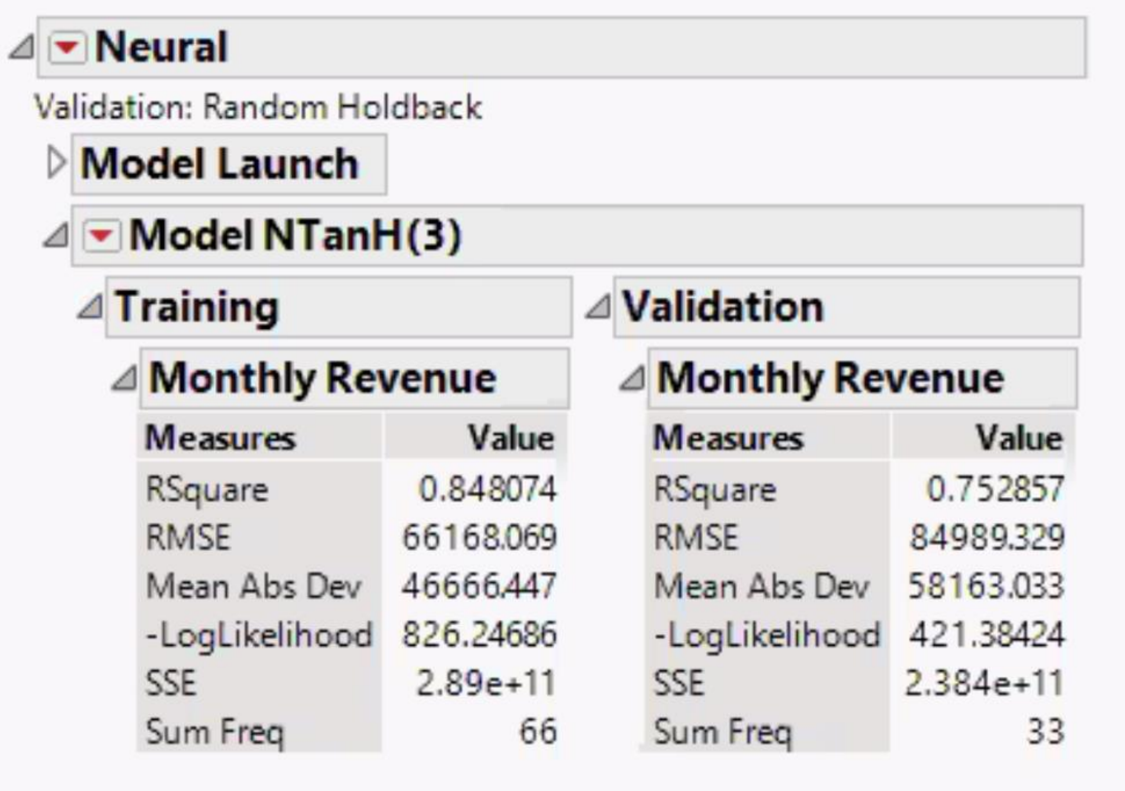
2.4 Neural Network Analysis

Process Description:

1. We have considered the Monthly Revenue and built the model using predictive modelling of neural network as shown in the screenshot below:
2. ‘Yelp’ and ‘Distance from town’ center also tells us how important factor it is in the increase of the monthly revenue.



2.3 a) Screenshot of Neural Network analysis in JMP:



Neural

Validation: Random Holdback

Model Launch

Model NTanH(3)

Training

Monthly Revenue

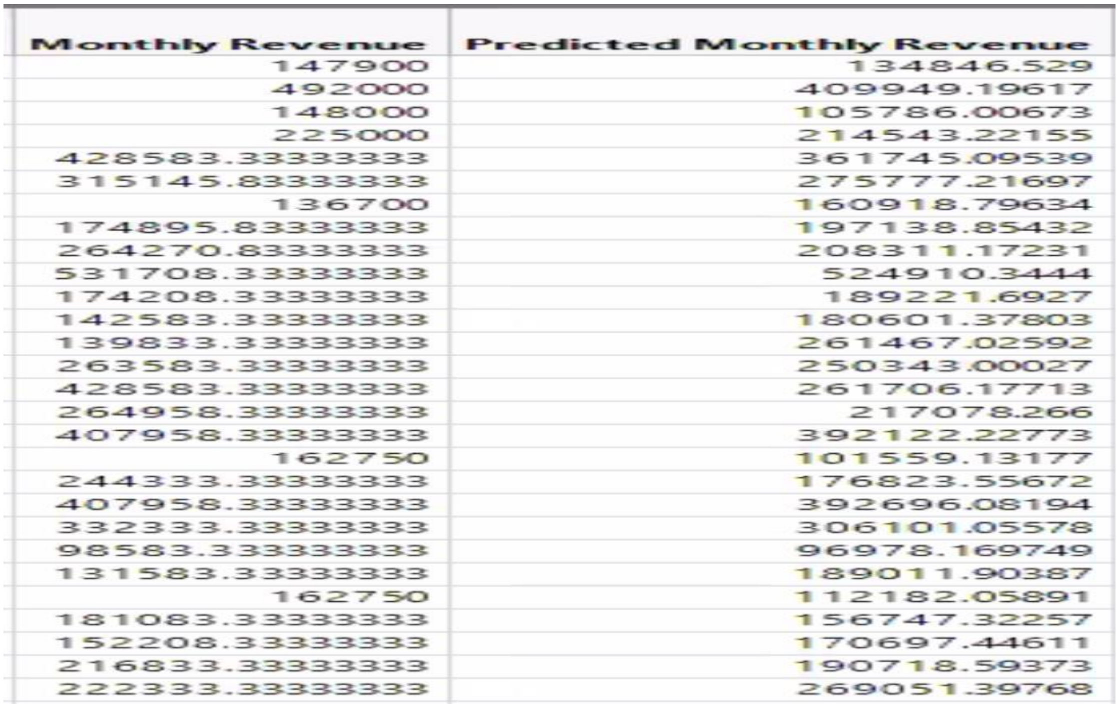
Measures	Value
RSquare	0.848074
RMSE	66168.069
Mean Abs Dev	46666.447
-LogLikelihood	826.24686
SSE	2.89e+11
Sum Freq	66

Validation

Monthly Revenue

Measures	Value
RSquare	0.752857
RMSE	84989.329
Mean Abs Dev	58163.033
-LogLikelihood	421.38424
SSE	2.384e+11
Sum Freq	33

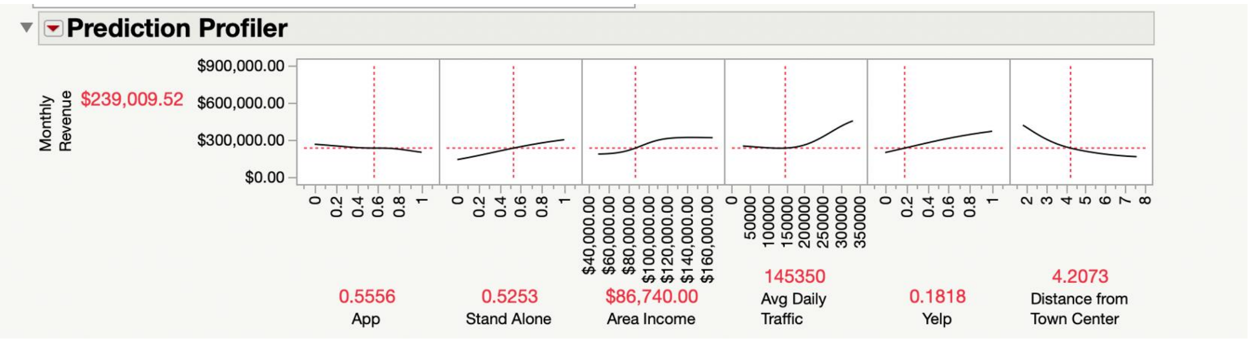
2.3 b) Screenshot of Monthly Revenue and Prediction Formula Monthly Revenue using Neural Network Analysis



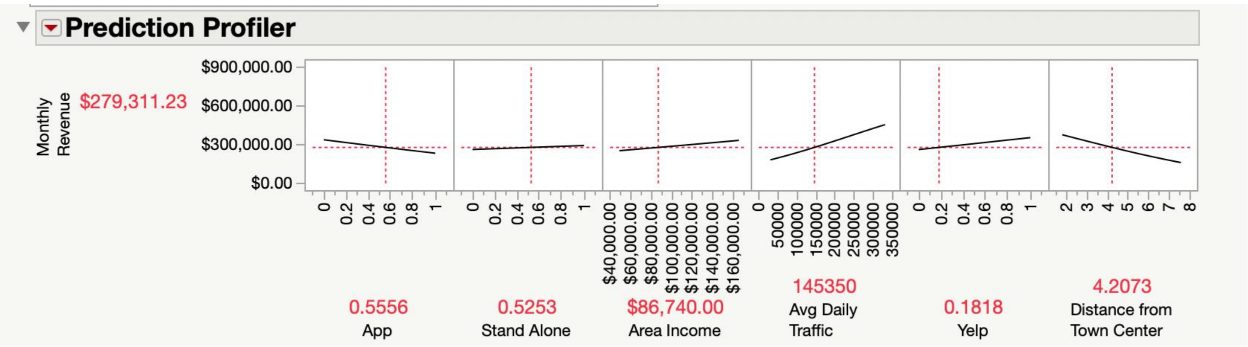
Monthly Revenue	Predicted Monthly Revenue
147900	134846.529
492000	409949.19617
148000	105786.00673
225000	214543.22155
428583.33333333	361745.09539
315145.83333333	275777.21697
136700	160918.79634
174895.83333333	197138.85432
264270.83333333	208311.17231
531708.33333333	524910.3444
174208.33333333	189221.6927
142583.33333333	180601.37803
139833.33333333	261467.02592
263583.33333333	250343.00027
428583.33333333	261706.17713
264958.33333333	217078.266
407958.33333333	392122.22773
162750	101559.13177
244333.33333333	176823.55672
407958.33333333	392696.08194
332333.33333333	306101.05578
98583.333333333	96978.169749
131583.33333333	189011.90387
162750	112182.05891
181083.33333333	156747.32257
152208.33333333	170697.44611
216833.33333333	190718.59373
222333.33333333	269051.39768

2.5 Comparison of Regression analysis and Neural Network Analysis

Regression Analysis



Neural Network Analysis



2.6 Conclusion:

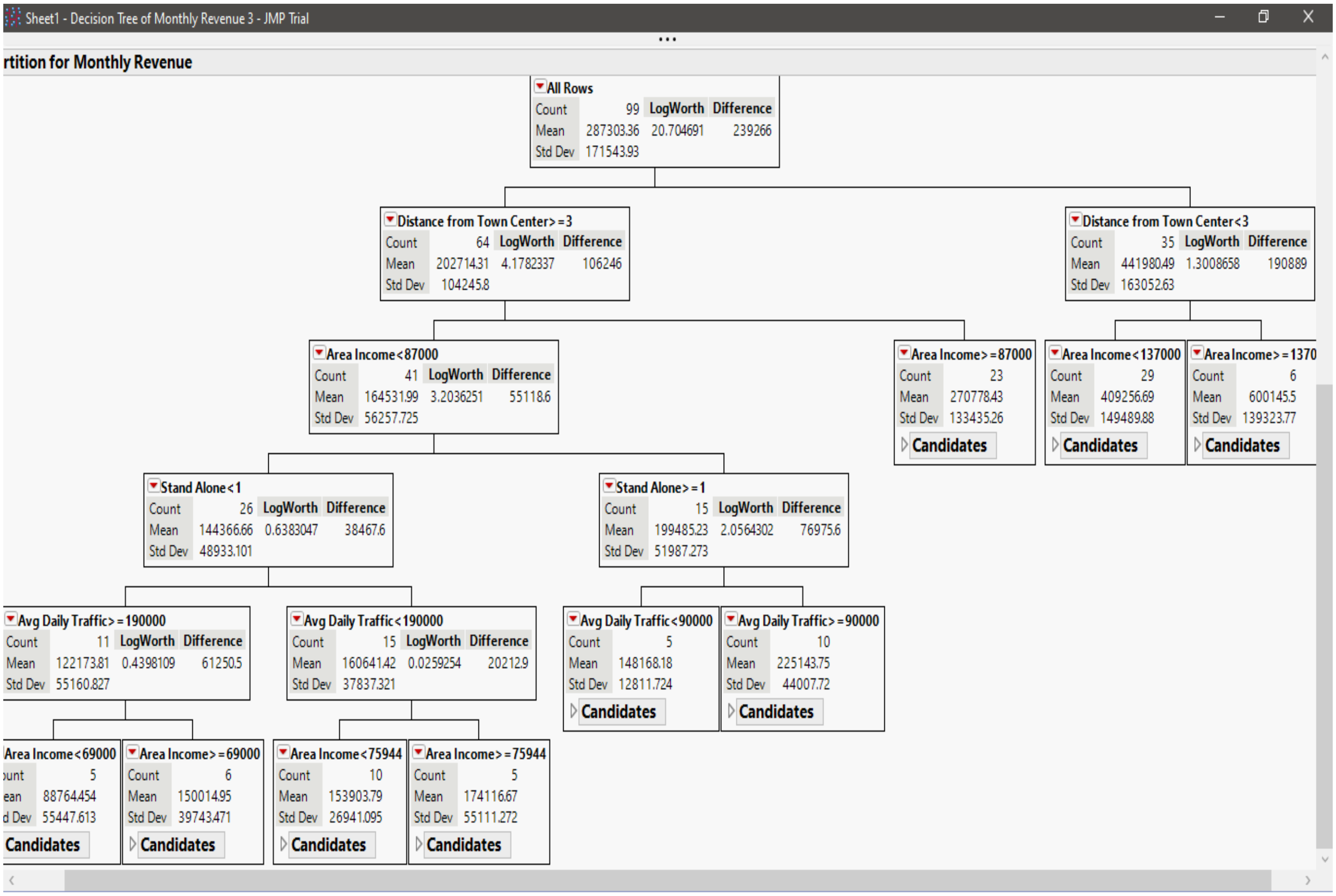
From the above screenshots, we see that the predictions from the neural networks analysis keeps changing according to the weights. The Predicted monthly revenue is similar in both regression analysis and neural network analysis.

Part-3 Segmentation Analysis

3.1 Objective:

Your organization has run into restrictions in the Morristown NJ area. Commercial real estate prices within a 3-mile radius of the town are extremely expensive. It has also been determined that the average income of areas outside the 3-mile perimeter drops to below $86,000. Therefore, the restaurant needs to adjust its prices and cuisine accordingly. Using a Segmentation/decision tree (use partition in SAS) methodology, provide two more descriptions/variables of the type and place of restaurant that you would open to attain the highest average expected revenue. You are to use the file in Part 2 and conduct a segmentation analysis to guide your response.

3.2 Screenshot of decision tree


Sheet1 - Decision Tree of Monthly Revenue 3 - JMP Trial

rtition for Monthly Revenue

All Rows

Count	99	LogWorth	Difference
Mean	287303.36	20.704691	239266
Std Dev	171543.93		

Distance from Town Center>=3

Count	64	LogWorth	Difference
Mean	202714.31	4.1782337	106246
Std Dev	104245.8		

Distance from Town Center<3

Count	35	LogWorth	Difference
Mean	441980.49	1.3008658	190889
Std Dev	163052.63		

Area Income<87000

Count	41	LogWorth	Difference
Mean	164531.99	3.2036251	55118.6
Std Dev	56257.725		

Area Income>=87000

Count	23
Mean	270778.43
Std Dev	133435.26

Area Income<137000

Count	29
Mean	409256.69
Std Dev	149489.88

Area Income>=1370

Count	6
Mean	600145.5
Std Dev	139323.77

Stand Alone<1

Count	26	LogWorth	Difference
Mean	144366.66	0.6383047	38467.6
Std Dev	48933.101		

Stand Alone>=1

Count	15	LogWorth	Difference
Mean	199485.23	2.0564302	76975.6
Std Dev	51987.273		

Avg Daily Traffic>=190000

Count	11	LogWorth	Difference
Mean	122173.81	0.4398109	61250.5
Std Dev	55160.827		

Avg Daily Traffic<190000

Count	15	LogWorth	Difference
Mean	160641.42	0.0259254	20212.9
Std Dev	37837.321		

Avg Daily Traffic<90000

Count	5
Mean	148168.18
Std Dev	12811.724

Avg Daily Traffic>=90000

Count	10
Mean	225143.75
Std Dev	44007.72

Area Income<69000

ount	5
ean	88764.454
d Dev	55447.613

Area Income>=69000

Count	6
Mean	150014.95
Std Dev	39743.471

Area Income<75944

Count	10
Mean	153903.79
Std Dev	26941.095

Area Income>=75944

Count	5
Mean	174116.67
Std Dev	55111.272

3.3 Conclusion:

The two variables that are:
1. Type of restaurant, for example what kind of food it serves.
2.Pricing of restaurant

The above two additional variables can play a role, they can help in knowing the food choices of people residing in the areas and also the type of restaurant they go to i.e Food prices.
The restaurant that can be open where standalone <1 , Avg Daily Traffic >=190000 & <190000 , area income >=69000 & area income >=75944 and distance from town center >5.27.