DATA MINING AND ANALYSIS FOR MANAGERS

FINAL REPORT

UNDER THE GUIDANCE OF
Prof. Stephan P. Kudyba

TEAM:
Avinash Rajagopal
Lekshmi Narasimman
Priya Sangale

CONTENT

PART 1

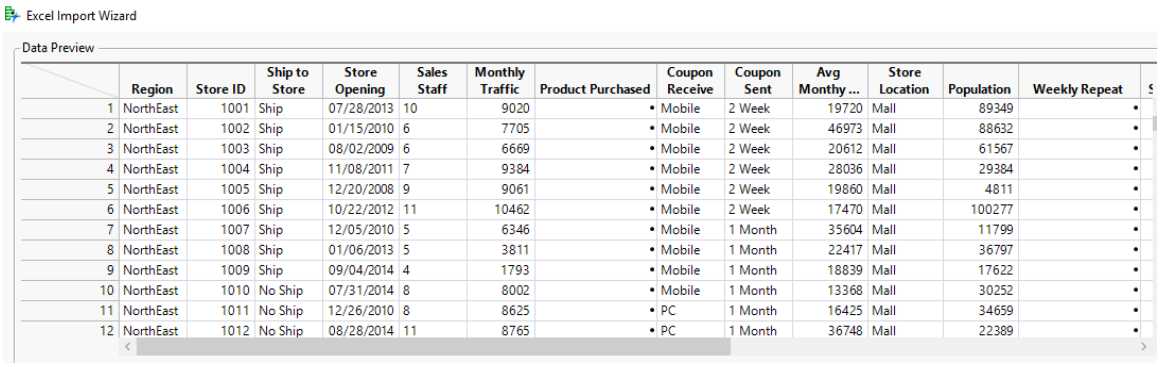
Input Data Analysis

1.1 Objective

The data set given provides the explanatory and target variables to analyze store sales and production. But the dataset has many errors, which makes it hard to analyze the file. We must make required changes (polish it/format it) to make it more useful for the analysis.

1.1 Troubleshoot Dataset

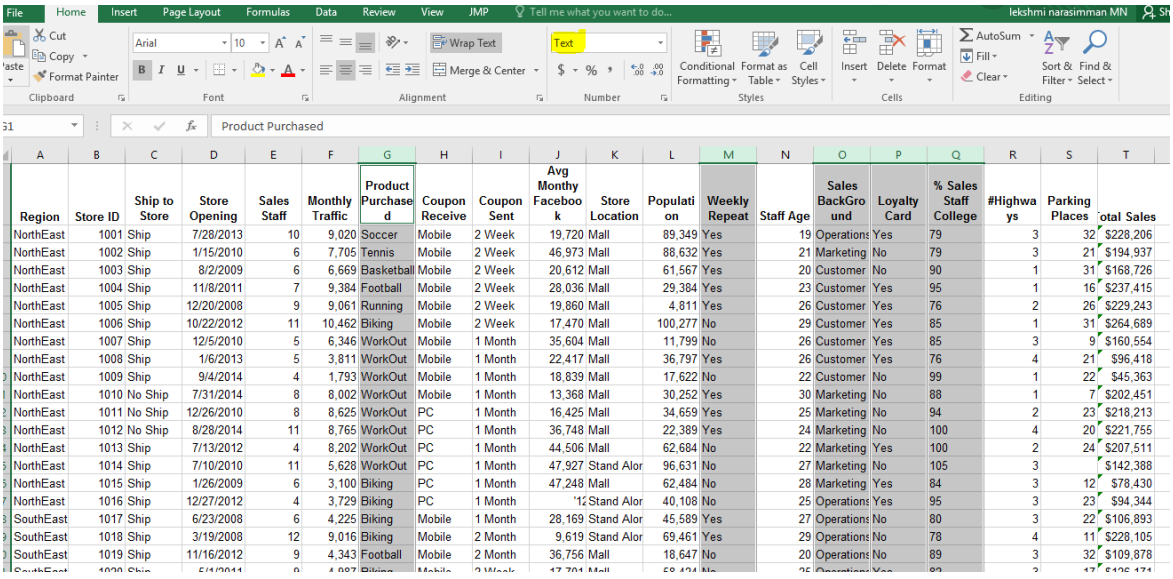
1. When we were trying to import the data in SAS JMP, some of the columns values were not imported, as shown in the screenshot below.
2. The columns are: Product Purchased, Weekly Repeat, Sales Background, Loyalty Card.


Excel Import Wizard

Data Preview

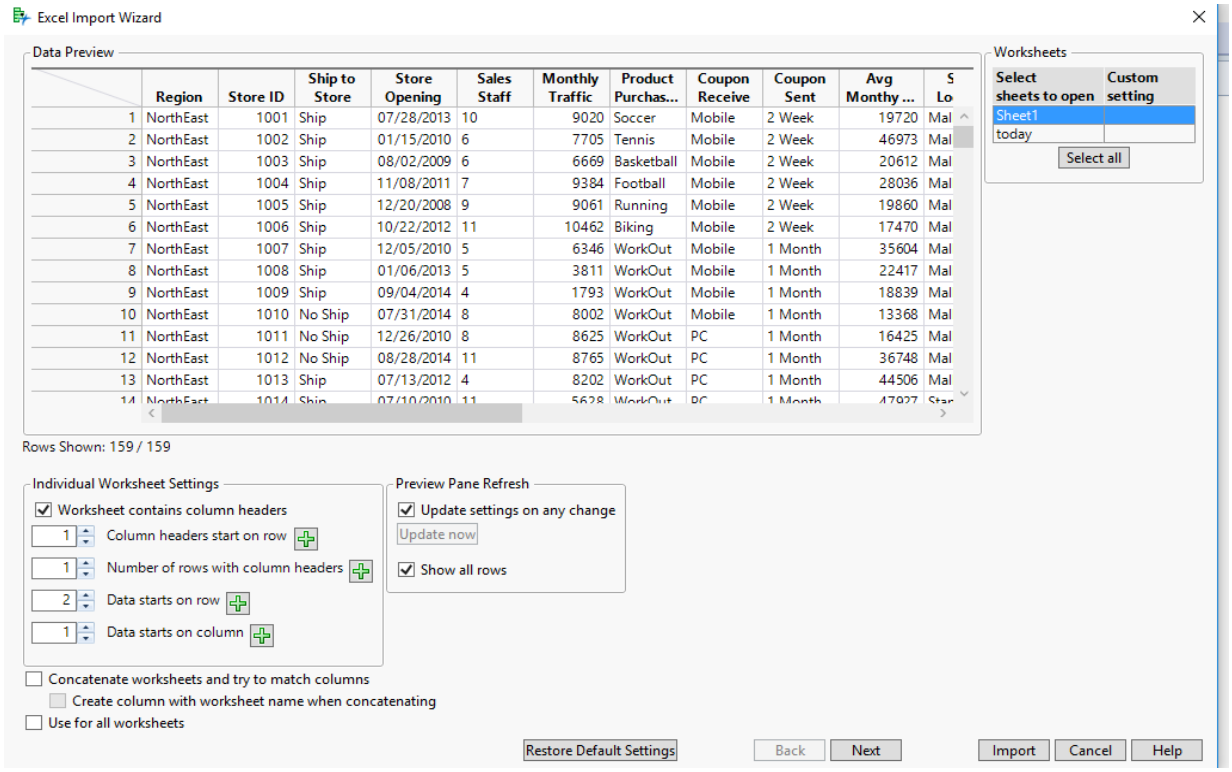
	Region	Store ID	Ship to Store	Store Opening	Sales Staff	Monthly Traffic	Product Purchased	Coupon Receive	Coupon Sent	Avg Monthy ...	Store Location	Population	Weekly Repeat
1	NorthEast	1001	Ship	07/28/2013	10	9020	•	Mobile	2 Week	19720	Mall	89349	•
2	NorthEast	1002	Ship	01/15/2010	6	7705	•	Mobile	2 Week	46973	Mall	88632	•
3	NorthEast	1003	Ship	08/02/2009	6	6669	•	Mobile	2 Week	20612	Mall	61567	•
4	NorthEast	1004	Ship	11/08/2011	7	9384	•	Mobile	2 Week	28036	Mall	29384	•
5	NorthEast	1005	Ship	12/20/2008	9	9061	•	Mobile	2 Week	19860	Mall	4811	•
6	NorthEast	1006	Ship	10/22/2012	11	10462	•	Mobile	2 Week	17470	Mall	100277	•
7	NorthEast	1007	Ship	12/05/2010	5	6346	•	Mobile	1 Month	35604	Mall	11799	•
8	NorthEast	1008	Ship	01/06/2013	5	3811	•	Mobile	1 Month	22417	Mall	36797	•
9	NorthEast	1009	Ship	09/04/2014	4	1793	•	Mobile	1 Month	18839	Mall	17622	•
10	NorthEast	1010	No Ship	07/31/2014	8	8002	•	Mobile	1 Month	13368	Mall	30252	•
11	NorthEast	1011	No Ship	12/26/2010	8	8625	•	PC	1 Month	16425	Mall	34659	•
12	NorthEast	1012	No Ship	08/28/2014	11	8765	•	PC	1 Month	36748	Mall	22389	•

3. To rectify, we have changed the respective columns format from General to Text in the excel sheet provided.

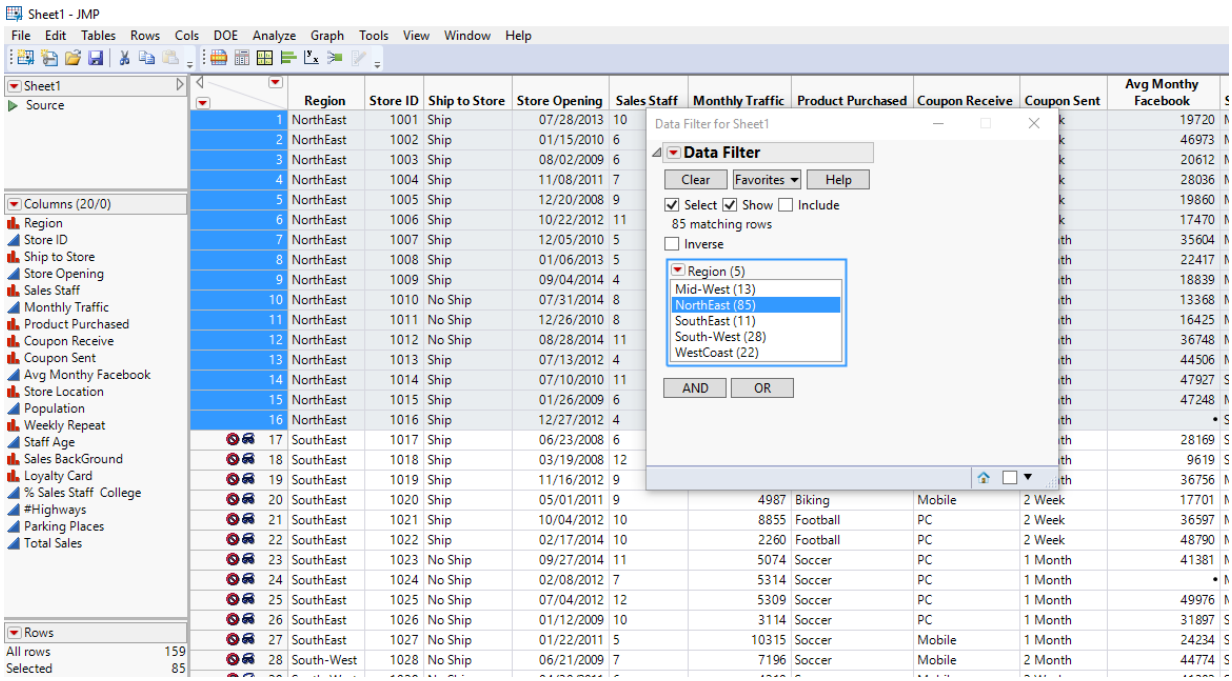


Region	Store ID	Ship to Store	Store Opening	Sales Staff	Monthly Traffic	Product Purchased	Coupon Receive	Coupon Sent	Avg Monthy Facebook	Store Location	Population	Weekly Repeat	Staff Age	Sales BackGround	Loyalty Card	% Sales Staff College	#Highways	Parking Places	Total Sales
NorthEast	1001	Ship	7/28/2013	10	9,020	Soccer	Mobile	2 Week	19,720	Mall	89,349	Yes	19	Operations	Yes	79	3	32	$228,206
NorthEast	1002	Ship	1/15/2010	6	7,705	Tennis	Mobile	2 Week	46,973	Mall	88,632	Yes	21	Marketing	No	79	3	21	$194,937
NorthEast	1003	Ship	8/2/2009	6	6,669	Basketball	Mobile	2 Week	20,612	Mall	61,567	Yes	20	Customer	No	90	1	31	$168,726
NorthEast	1004	Ship	11/8/2011	7	9,384	Football	Mobile	2 Week	28,036	Mall	29,384	Yes	23	Customer	Yes	95	1	16	$237,415
NorthEast	1005	Ship	12/20/2008	9	9,061	Running	Mobile	2 Week	19,860	Mall	4,811	Yes	26	Customer	Yes	76	2	26	$229,243
NorthEast	1006	Ship	10/22/2012	11	10,462	Biking	Mobile	2 Week	17,470	Mall	100,277	No	29	Customer	Yes	85	1	31	$264,689
NorthEast	1007	Ship	12/5/2010	5	6,346	WorkOut	Mobile	1 Month	35,604	Mall	11,799	No	26	Customer	Yes	85	3	9	$160,554
NorthEast	1008	Ship	1/6/2013	5	3,811	WorkOut	Mobile	1 Month	22,417	Mall	36,797	Yes	26	Customer	Yes	76	4	21	$96,418
NorthEast	1009	Ship	9/4/2014	4	1,793	WorkOut	Mobile	1 Month	18,839	Mall	17,622	No	22	Customer	No	99	1	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	1	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	2	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	4	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	2	24	$207,511
NorthEast	1014	Ship	7/10/2010	11	5,628	WorkOut	PC	1 Month	47,927	Stand Alor	96,631	No	27	Marketing	No	105	3		$142,388
NorthEast	1015	Ship	1/26/2009	6	3,100	Biking	PC	1 Month	47,248	Mall	62,484	No	28	Marketing	Yes	84	3	12	$78,430
NorthEast	1016	Ship	12/27/2012	4	3,729	Biking	PC	1 Month	'1	Stand Alor	40,108	No	25	Operations	Yes	95	3	23	$94,344
SouthEast	1017	Ship	6/23/2008	6	4,225	Biking	Mobile	1 Month	28,169	Stand Alor	45,589	Yes	27	Operations	No	80	3	22	$106,893
SouthEast	1018	Ship	3/19/2008	12	9,016	Biking	Mobile	2 Month	9,619	Stand Alor	69,461	Yes	29	Operations	No	78	4	11	$228,105
SouthEast	1019	Ship	11/16/2012	9	4,343	Football	Mobile	2 Month	36,756	Mall	18,647	No	20	Operations	No	89	3	32	$109,878

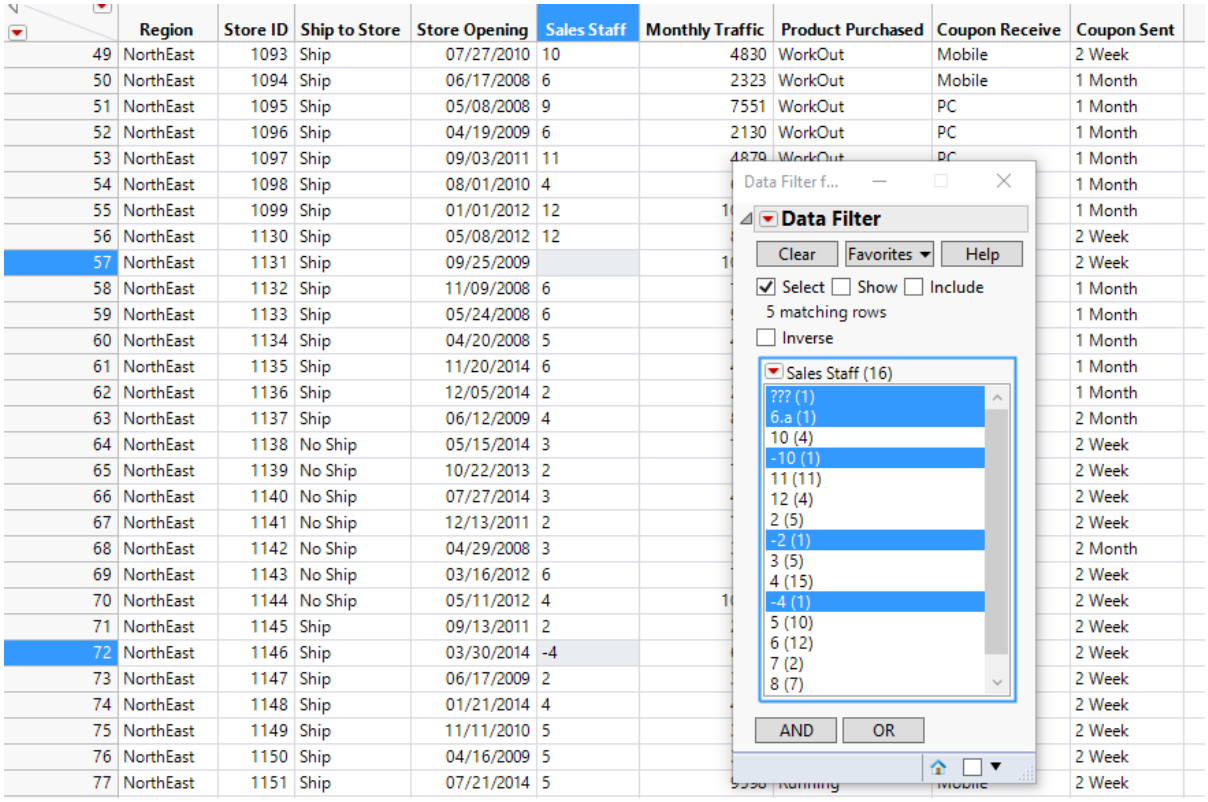
4. After implementing the necessary changes, the data is available for Mining.



5. Once data has been loaded, since it is asked to analyze on retailers in the North-East Region, we have filtered the Region only to North East Neglecting Mid-West, South West, South East and West Coast using SAS JMP.
6. We deleted the regions except Northeast.

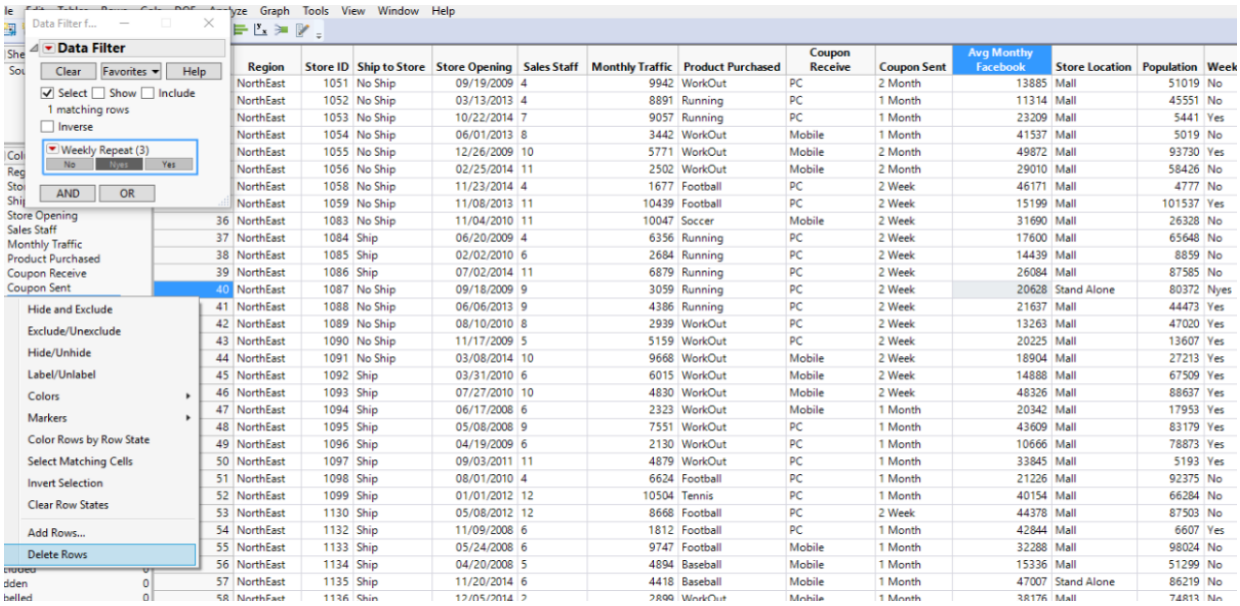


7. Sales Staff is the average amount of store workers that are present daily. It contains negative number and invalid number format. We have removed the insignificant values from Sales Staff Column using Data filter.



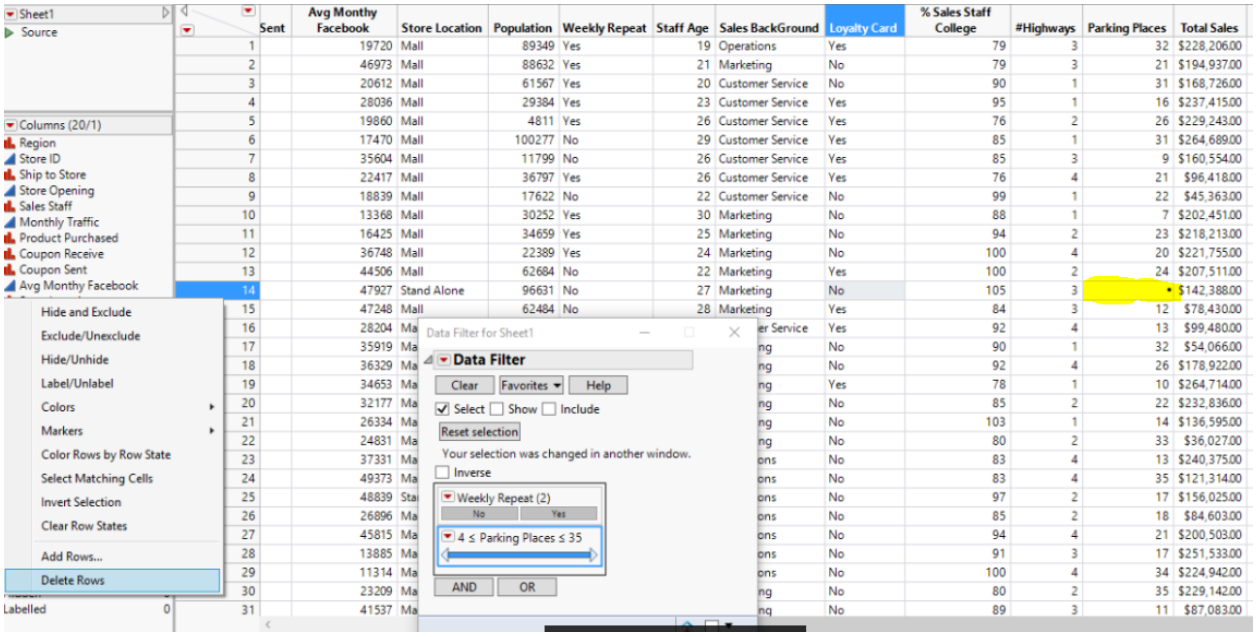
	Region	Store ID	Ship to Store	Store Opening	Sales Staff	Monthly Traffic	Product Purchased	Coupon Receive	Coupon Sent
49	NorthEast	1093	Ship	07/27/2010	10	4830	WorkOut	Mobile	2 Week
50	NorthEast	1094	Ship	06/17/2008	6	2323	WorkOut	Mobile	1 Month
51	NorthEast	1095	Ship	05/08/2008	9	7551	WorkOut	PC	1 Month
52	NorthEast	1096	Ship	04/19/2009	6	2130	WorkOut	PC	1 Month
53	NorthEast	1097	Ship	09/03/2011	11	4879	WorkOut	PC	1 Month
54	NorthEast	1098	Ship	08/01/2010	4				1 Month
55	NorthEast	1099	Ship	01/01/2012	12				1 Month
56	NorthEast	1130	Ship	05/08/2012	12				2 Week
57	NorthEast	1131	Ship	09/25/2009					2 Week
58	NorthEast	1132	Ship	11/09/2008	6				1 Month
59	NorthEast	1133	Ship	05/24/2008	6				1 Month
60	NorthEast	1134	Ship	04/20/2008	5				1 Month
61	NorthEast	1135	Ship	11/20/2014	6				1 Month
62	NorthEast	1136	Ship	12/05/2014	2				1 Month
63	NorthEast	1137	Ship	06/12/2009	4				2 Month
64	NorthEast	1138	No Ship	05/15/2014	3				2 Week
65	NorthEast	1139	No Ship	10/22/2013	2				2 Week
66	NorthEast	1140	No Ship	07/27/2014	3				2 Week
67	NorthEast	1141	No Ship	12/13/2011	2				2 Week
68	NorthEast	1142	No Ship	04/29/2008	3				2 Month
69	NorthEast	1143	No Ship	03/16/2012	6				2 Week
70	NorthEast	1144	No Ship	05/11/2012	4				2 Week
71	NorthEast	1145	Ship	09/13/2011	2				2 Week
72	NorthEast	1146	Ship	03/30/2014	-4				2 Week
73	NorthEast	1147	Ship	06/17/2009	2				2 Week
74	NorthEast	1148	Ship	01/21/2014	4				2 Week
75	NorthEast	1149	Ship	11/11/2010	5				2 Week
76	NorthEast	1150	Ship	04/16/2009	5				2 Week
77	NorthEast	1151	Ship	07/21/2014	5				2 Week

8. In 'Avg Month Facebook' Column, there is one value which is not significant. And thus, removed that respective value using Data Filter.
9. In '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.



	Region	Store ID	Ship to Store	Store Opening	Sales Staff	Monthly Traffic	Product Purchased	Coupon Receive	Coupon Sent	Avg Monthy Facebook	Store Location	Population	Week
	NorthEast	1051	No Ship	09/19/2009	4	9942	WorkOut	PC	2 Month	13885	Mall	51019	No
	NorthEast	1052	No Ship	03/13/2013	4	8891	Running	PC	1 Month	11314	Mall	45551	No
	NorthEast	1053	No Ship	10/22/2014	7	9057	Running	PC	1 Month	23209	Mall	5441	Yes
	NorthEast	1054	No Ship	06/01/2013	8	3442	WorkOut	Mobile	1 Month	41537	Mall	5019	No
	NorthEast	1055	No Ship	12/26/2009	10	5771	WorkOut	Mobile	2 Month	49872	Mall	93730	Yes
	NorthEast	1056	No Ship	02/25/2014	11	2502	WorkOut	Mobile	2 Month	29010	Mall	58426	No
	NorthEast	1058	No Ship	11/23/2014	4	1677	Football	PC	2 Week	46171	Mall	4777	No
	NorthEast	1059	No Ship	11/08/2013	11	10439	Football	PC	2 Week	15199	Mall	101537	Yes
36	NorthEast	1083	No Ship	11/04/2010	11	10047	Soccer	Mobile	2 Week	31690	Mall	26328	No
37	NorthEast	1084	Ship	06/20/2009	4	6356	Running	PC	2 Week	17600	Mall	65648	No
38	NorthEast	1085	Ship	02/02/2010	6	2684	Running	PC	2 Week	14439	Mall	8859	No
39	NorthEast	1086	Ship	07/02/2014	11	6879	Running	PC	2 Week	26084	Mall	87585	No
40	NorthEast	1087	No Ship	09/18/2009	9	3059	Running	PC	2 Week	20628	Stand Alone	80372	Nyes
41	NorthEast	1088	No Ship	06/06/2013	9	4386	Running	PC	2 Week	21637	Mall	44473	Yes
42	NorthEast	1089	No Ship	08/10/2010	8	2939	WorkOut	PC	2 Week	13263	Mall	47020	Yes
43	NorthEast	1090	No Ship	11/17/2009	5	5159	WorkOut	PC	2 Week	20225	Mall	13607	Yes
44	NorthEast	1091	No Ship	03/08/2014	10	9668	WorkOut	Mobile	2 Week	18904	Mall	27213	Yes
45	NorthEast	1092	Ship	03/31/2010	6	6015	WorkOut	Mobile	2 Week	14888	Mall	67509	Yes
46	NorthEast	1093	Ship	07/27/2010	10	4830	WorkOut	Mobile	2 Week	48326	Mall	88637	Yes
47	NorthEast	1094	Ship	06/17/2008	6	2323	WorkOut	Mobile	1 Month	20342	Mall	17953	Yes
48	NorthEast	1095	Ship	05/08/2008	9	7551	WorkOut	PC	1 Month	43609	Mall	83179	Yes
49	NorthEast	1096	Ship	04/19/2009	6	2130	WorkOut	PC	1 Month	10666	Mall	78873	Yes
50	NorthEast	1097	Ship	09/03/2011	11	4879	WorkOut	PC	1 Month	33845	Mall	5193	Yes
51	NorthEast	1098	Ship	08/01/2010	4	6624	Football	PC	1 Month	21226	Mall	92375	No
52	NorthEast	1099	Ship	01/01/2012	12	10504	Tennis	PC	1 Month	40154	Mall	66284	No
53	NorthEast	1130	Ship	05/08/2012	12	8668	Football	PC	2 Week	44378	Mall	87503	No
54	NorthEast	1132	Ship	11/09/2008	6	1812	Football	PC	1 Month	42844	Mall	6607	Yes
55	NorthEast	1133	Ship	05/24/2008	6	9747	Football	Mobile	1 Month	32288	Mall	98024	No
56	NorthEast	1134	Ship	04/20/2008	5	4894	Baseball	Mobile	1 Month	15336	Mall	51299	No
57	NorthEast	1135	Ship	11/20/2014	6	4418	Baseball	Mobile	1 Month	47007	Stand Alone	86219	No

10. In ‘Parking Places’, there is NA in one of the row, we have filtered that respective row.



	Sent	Avg Monthy Facebook	Store Location	Population	Weekly Repeat	Staff Age	Sales BackGround	Loyalty Card	% Sales Staff College	#Highways	Parking Places	Total Sales
1		19720	Mall	89349	Yes	19	Operations	Yes	79	3	32	$228,206.00
2		46973	Mall	88632	Yes	21	Marketing	No	79	3	21	$194,937.00
3		20612	Mall	61567	Yes	20	Customer Service	No	90	1	31	$168,726.00
4		28036	Mall	29384	Yes	23	Customer Service	Yes	95	1	16	$237,415.00
5		19860	Mall	4811	Yes	26	Customer Service	Yes	76	2	26	$229,243.00
6		17470	Mall	100277	No	29	Customer Service	Yes	85	1	31	$264,689.00
7		35604	Mall	11799	No	26	Customer Service	Yes	85	3	9	$160,554.00
8		22417	Mall	36797	Yes	26	Customer Service	Yes	76	4	21	$96,418.00
9		18839	Mall	17622	No	22	Customer Service	No	99	1	22	$45,363.00
10		13368	Mall	30252	Yes	30	Marketing	No	88	1	7	$202,451.00
11		16425	Mall	34659	Yes	25	Marketing	No	94	2	23	$218,213.00
12		36748	Mall	22389	Yes	24	Marketing	No	100	4	20	$221,755.00
13		44506	Mall	62684	No	22	Marketing	Yes	100	2	24	$207,511.00
14		47927	Stand Alone	96631	No	27	Marketing	No	105	3	*	$142,388.00
15		47248	Mall	62484	No	28	Marketing	Yes	84	3	12	$78,430.00
16		28204	Ma				er Service	Yes	92	4	13	$99,480.00
17		35919	Ma				ng	No	90	1	32	$54,066.00
18		36329	Ma				ng	No	92	4	26	$178,922.00
19		34653	Ma				ng	Yes	78	1	10	$264,714.00
20		32177	Ma				ng	No	85	2	22	$232,836.00
21		26334	Ma				ng	No	103	1	14	$136,595.00
22		24831	Ma				ng	No	80	2	33	$36,027.00
23		37331	Ma				ons	No	83	4	13	$240,375.00
24		49373	Ma				ons	No	83	4	35	$121,314.00
25		48839	Sta				ons	No	97	2	17	$156,025.00
26		26896	Ma				ons	No	85	2	18	$84,603.00
27		45815	Ma				ons	No	94	4	21	$200,503.00
28		13885	Ma				ons	No	91	3	17	$251,533.00
29		11314	Ma				ons	No	100	4	34	$224,942.00
30		23209	Ma				ng	No	80	2	35	$229,142.00
31		41537	Ma				ng	No	89	3	11	$87,083.00

11. Once the dataset is filtered and cleaned, we have now 77 rows of data ready to do Mining

1.2 Degrees of freedom

Before Cleaning: Raw Data:
Total Number of Variables: 20
Target Variable: 1
Total rows: 159
Degrees of Freedom: 159- (20-1) = 140

Final Cleansed and Transformed Data:
Total Number of Variables: 20
Target Variable: 1
Total Rows: 77
Degrees of Freedom: 77 – (20-1) = 58

PART 2

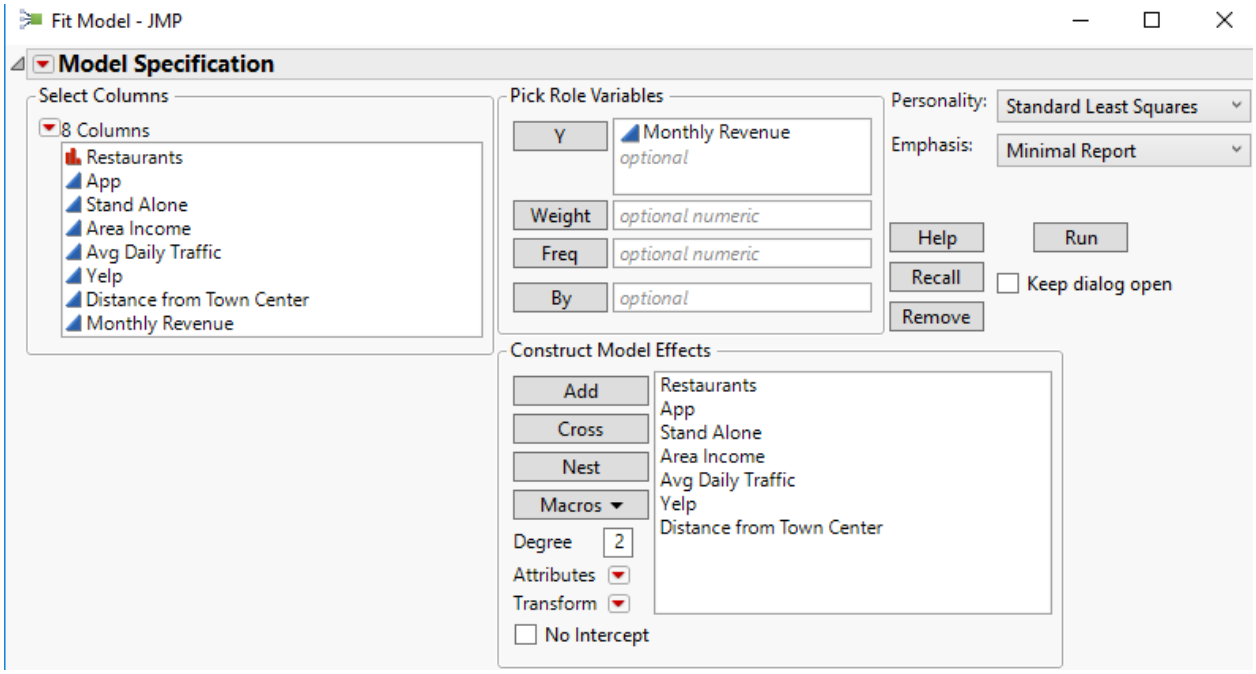
Regression Analysis

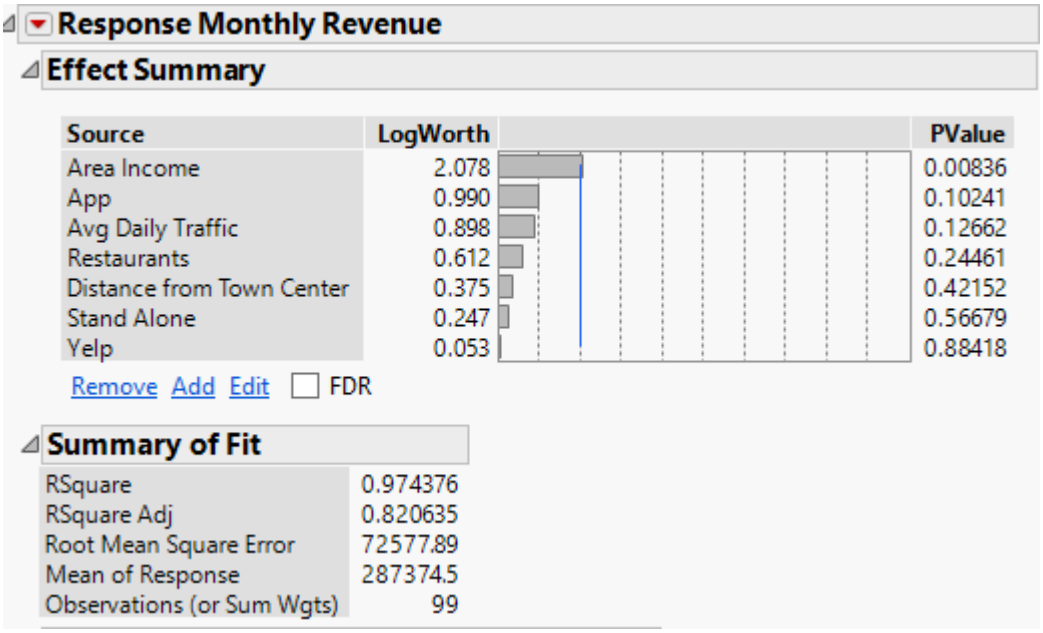
2.1 Objective

For this exercise, we are acting as marketing strategists for restaurant activities, and our objective is to use statistical output of the model to help make decisions on how to adjust the operational activities to best locate and operate a higher-tier restaurant by analyzing regression results. We must devise a simple business plan using whatever information is critical to strategic decision and to use the results of our model to estimate the expected monthly revenue at your proposed restaurant.

2.2 Process Description

1. We have imported the training data set from the excel into JMP. We build a model using Linear regression. We have included Monthly revenue as the target variable
2. When we try to run the model by keeping the target variable and included all the variables as explanatory variable, we found that only the **Restaurant** column seems to be significant and all the other column values are insignificant as all the other columns have P-Values greater than 0.05. Please find below the screenshot.





Response Monthly Revenue

Effect Summary

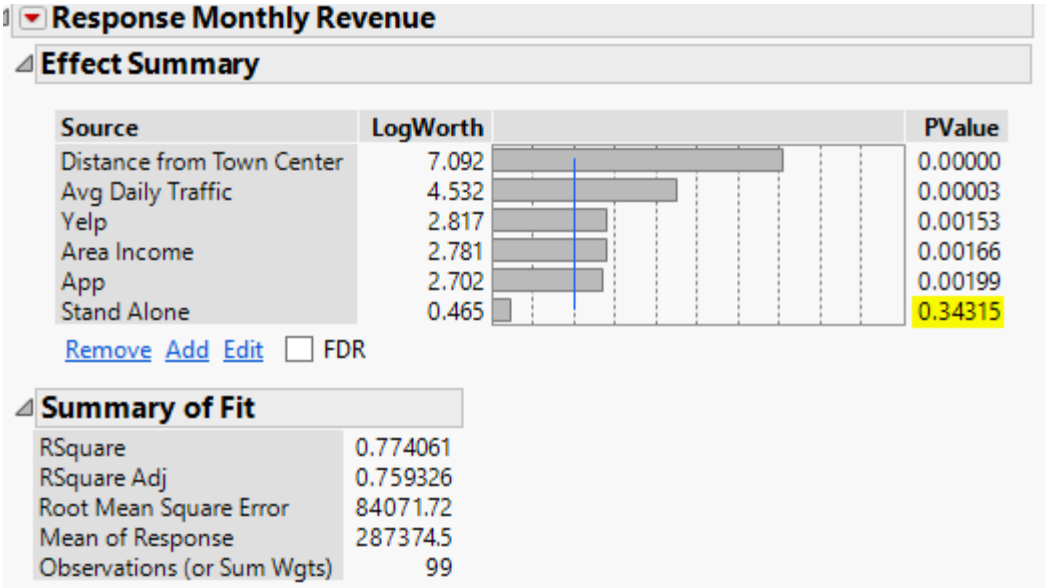
Source	LogWorth	PValue
Area Income	2.078	0.00836
App	0.990	0.10241
Avg Daily Traffic	0.898	0.12662
Restaurants	0.612	0.24461
Distance from Town Center	0.375	0.42152
Stand Alone	0.247	0.56679
Yelp	0.053	0.88418

Remove Add Edit ☐ FDR

Summary of Fit

RSquare	0.974376
RSquare Adj	0.820635
Root Mean Square Error	72577.89
Mean of Response	287374.5
Observations (or Sum Wgts)	99

3. So, we build the model neglecting the Restaurant column, and included all other explanatory variables. As shown in the below the result:



Response Monthly Revenue

Effect Summary

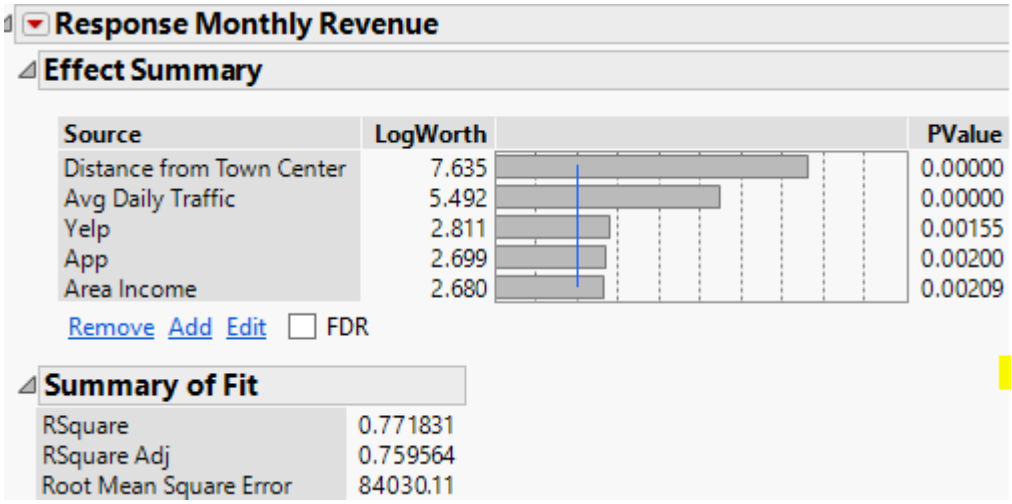
Source	LogWorth	PValue
Distance from Town Center	7.092	0.00000
Avg Daily Traffic	4.532	0.00003
Yelp	2.817	0.00153
Area Income	2.781	0.00166
App	2.702	0.00199
Stand Alone	0.465	0.34315

Remove Add Edit ☐ FDR

Summary of Fit

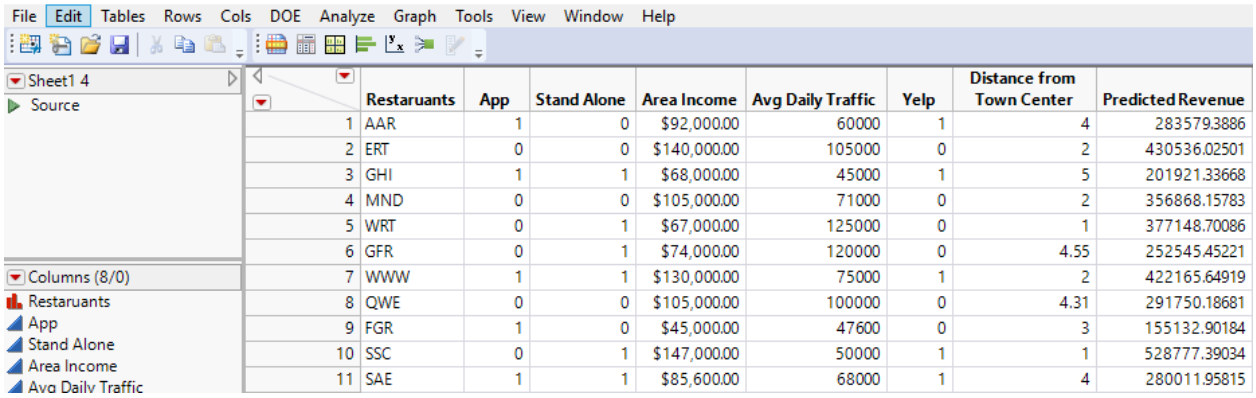
RSquare	0.774061
RSquare Adj	0.759326
Root Mean Square Error	84071.72
Mean of Response	287374.5
Observations (or Sum Wgts)	99

4. The P-Value of Stand Alone column is 0.34315, which is greater than 0.05, we have considered this variable as insignificant and build the model again as shown in the screenshot below.



2.3 Future Prediction

1. Thus, we have trained the dataset and concluded that the variables stand alone and restaurant are not significant to run the model.
2. We build the model for testing dataset and copied the predicted monthly revenue formula from training dataset and predicted the monthly revenue for testing data set. Please find below the screenshot of the predicted monthly revenue.
3. We have showed the Predicted revenue as a separate column and applied the formula.



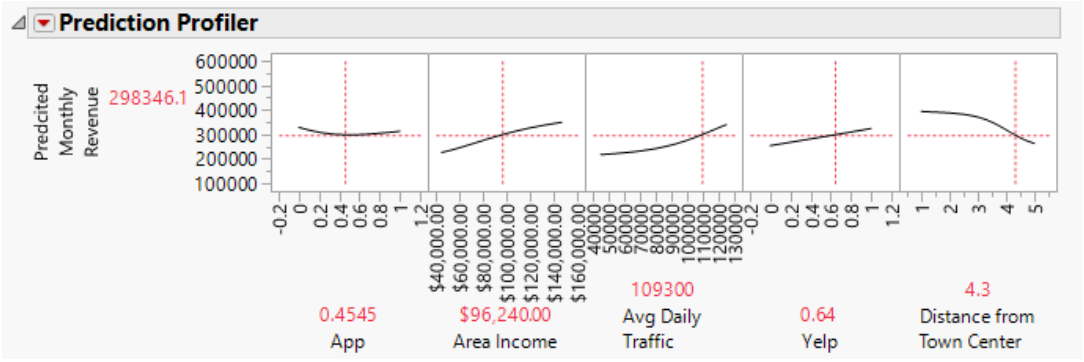
	Restaurants	App	Stand Alone	Area Income	Avg Daily Traffic	Yelp	Distance from Town Center	Predicted Revenue
1	AAR	1	0	$92,000.00	60000	1	4	283579.3886
2	ERT	0	0	$140,000.00	105000	0	2	430536.02501
3	GHI	1	1	$68,000.00	45000	1	5	201921.33668
4	MND	0	0	$105,000.00	71000	0	2	356868.15783
5	WRT	0	1	$67,000.00	125000	0	1	377148.70086
6	GFR	0	1	$74,000.00	120000	0	4.55	252545.45221
7	WWW	1	1	$130,000.00	75000	1	2	422165.64919
8	QWE	0	0	$105,000.00	100000	0	4.31	291750.18681
9	FGR	1	0	$45,000.00	47600	0	3	155132.90184
10	SSC	0	1	$147,000.00	50000	1	1	528777.39034
11	SAE	1	1	$85,600.00	68000	1	4	280011.95815

2.3 Neural Network:

1. We have considered the Monthly Revenue and build the model and please find below the screenshots of the predicted monthly Revenue.

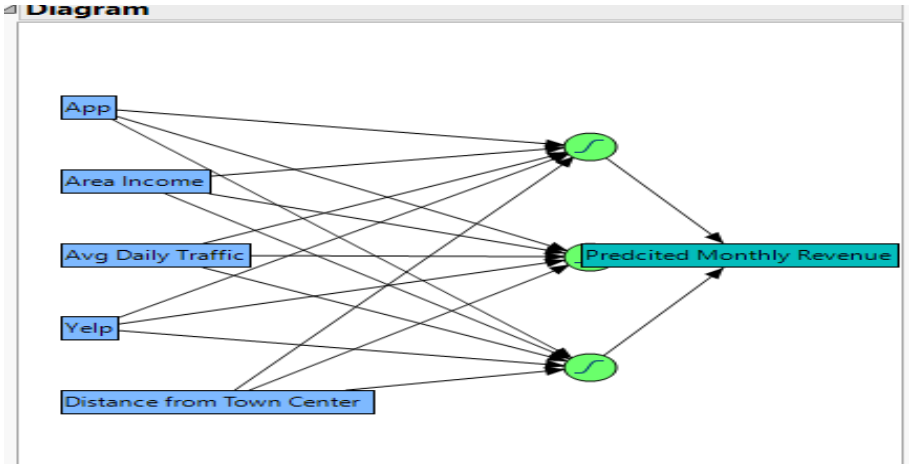
Cols DOE Analyze Graph Tools View Window Help

	Restaruants	App	Stand Alone	Area Income	Avg Daily Traffic	Yelp	Distance from Town Center	Predicted monthly Revenue
1	AAR	1	0	$92,000.00	60000	1	4	232562.46078
2	ERT	0	0	$140,000.00	105000	0	2	487851.65882
3	GHI	1	1	$68,000.00	45000	1	5	208172.62181
4	MND	0	0	$105,000.00	71000	0	2	346405.45976
5	WRT	0	1	$67,000.00	125000	0	1	379764.93589
6	GFR	0	1	$74,000.00	120000	0	4.55	232302.08009
7	WWW	1	1	$130,000.00	75000	1	2	455836.85491
8	QWE	0	0	$105,000.00	100000	0	4.31	262216.3487
9	FGR	1	0	$45,000.00	47600	0	3	111898.81471
10	SSC	0	1	$147,000.00	50000	1	1	564285.75945
11	SAE	1	1	$85,600.00	68000	1	4	228059.91242



2. The above neural net prediction showcase how the variable 'Area Income' when substantially raised using the vertical bar, increases the monthly revenue.
3. 'Yelp' and 'Distance from town' center also tells us how important factor it is in the increase of the monthly revenue.

Neural Network graph builder:



PART 3

Data Analysis on Tobacco use

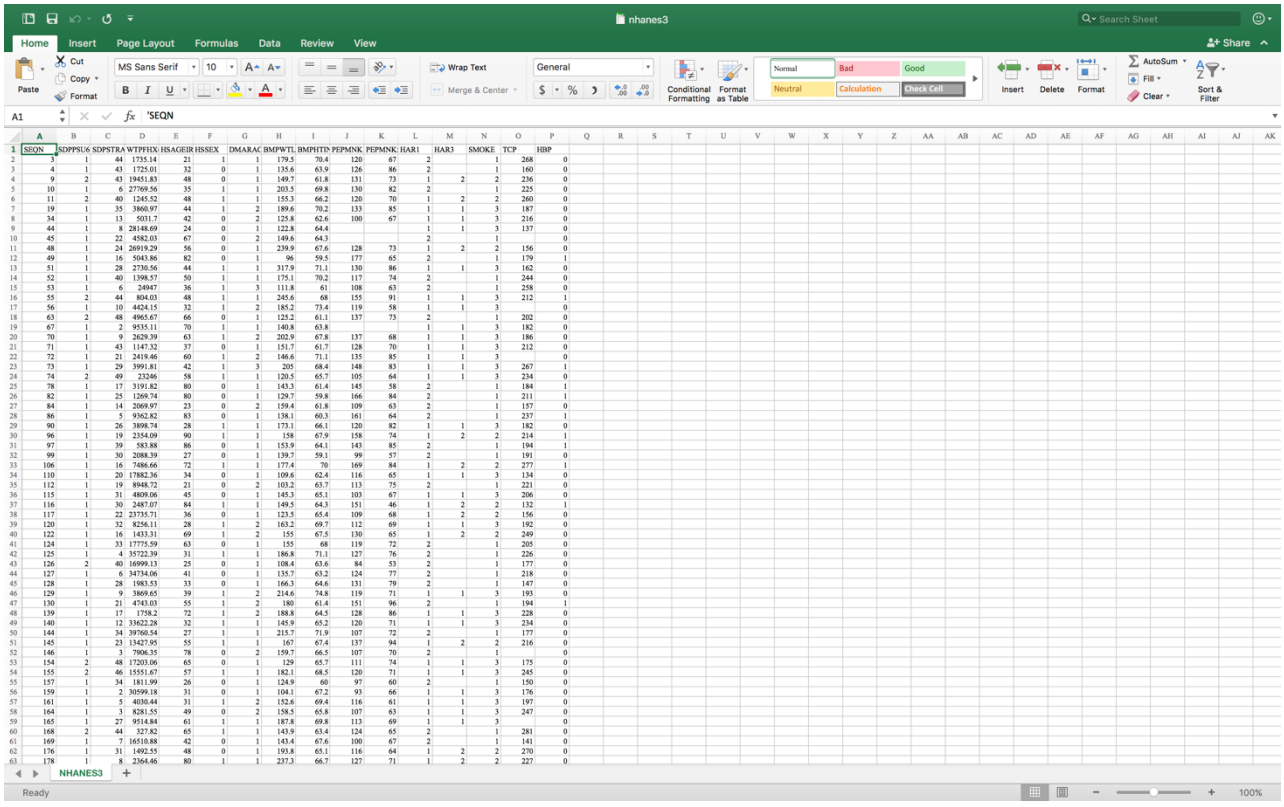
3.1 Data Set:

The dataset used for analysis for this project contains the prevalence and trends of Tobacco use and other characteristics of a person. This data set is originally from National Health and Nutrition Examination Survey (NHANES). The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. This dataset explains the drivers which includes smoking habits that causes high blood pressure.

3.2 Objective:

The purpose of this data set mining implementation is to extrapolate the factors triggering High Blood Pressure from smoking and other attributes. The analysis will also explain which variables leads to tobacco use.

3.3 Data sheet snapshot:



3.4 Attributes and Source:

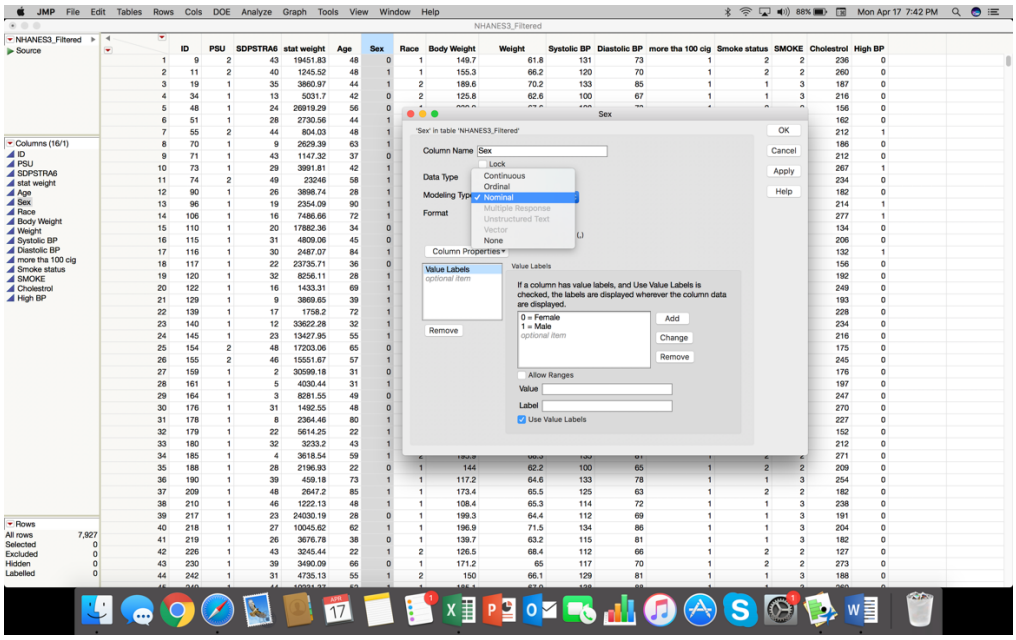
- **ID(SEQN**): Respondent identification number
- **Pseudo PSU(SDPPSU6):** Primary Sampling unit(PSU) is the unit used in NHANES to categorize their samples.
- **Pseudo Stratum(SDPSTRA6):** is the variable used to specify the sample design in NHANES.
- **Statistical Weight(WTPFHX6):** Calculated by NHANES

  NOTE: For NHANES datasets, the use of sampling weights and sample design variables is recommended for all analyses because the sample design is a clustered design and incorporates differential probabilities of selection. Fail to account for the sampling parameters, possibility to obtain biased estimates and overstate significance levels**.**
- **Age(HSAGEIR):** Age in years
- **Sex(HSSEX):** Gender
- **Race(DMARACER):** Black/White/Others
- **Body Weight(BMPWTLBS):** Weight of the body measured in pounds
- **Height(BMPHTIN):** Height of the person in inches
- **Systolic BP(PEPMNK1R):** Average systolic blood pressure describes the higher number, of the two numbers, marked in BP to check the severity of BP.
- **Diastolic BP(PEPMNK5R):** Average Diastolic blood pressure describes the lower number, of the two numbers, marked in BP to check the severity of BP.
- **More than 100 Cig(HAR1):** Has respondent smoked > 100 cigarettes in life
- **Smoke Status(HAR3):** Does respondent smoke cigarettes now?
- **Smoking(SMOKE):** Is the respondent a rigorous smoker or do not smoke now?
- **Cholesterol(TCP):** Serum Cholesterol
- **Blood Pressure(HBP):** High Blood Pressure

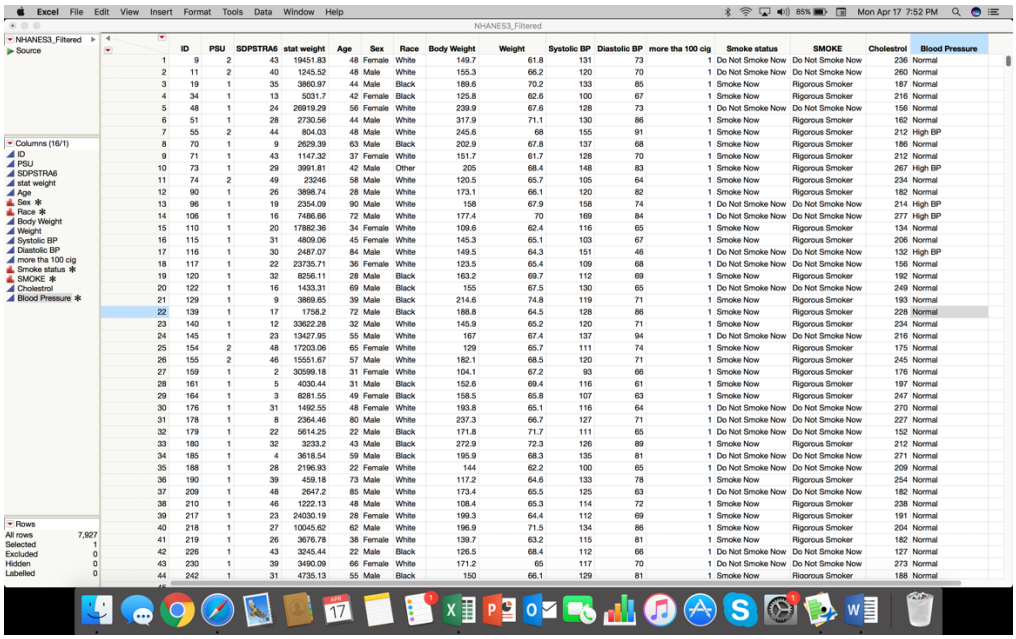
Source: https://www.umass.edu/statdata/statdata/data/ (NHANES-Smoking)

3.5 Data Preparation:

The snapshot shown under the section 3.3 shows the data received from NHANES. The Original data set had 17301 rows. Out of 17301 data rows, 9374 row had few missing values. Out of these, most of the data fields were 'nominal' variables (either 0 or 1). It is difficult to clean the data or transform the nominal data, we then decided to delete the data and we still had 7927 row in hand. We altered the data header for easy understanding and characterized the nominal data in JMP tool.



Snapshot of changing the data type to nominal and setting the data value to character using JMP Software



Snapshot of data after Data Cleaning using JMP Software

3.6 Data Analytical Model:

The Data Analytical technique used for this model is Logistic Regression. The reason to choose this technique was the target variable being the type 'binomial', so Logistic regression will be the best fit. Furthermore, this type of regression will give a better view of the driver variables driving the target with high approximate value.

This model will answer the following questions,
1. We obviously know Smoking leads to Cancer, but does smoking affects Blood Pressure too?
2. With the past data, what qualities does a 'Rigorous Smoker' have?

3.7 Result and Analysis:

PART 1: Does Smoking lead to High Blood Pressure

Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept		4.96129602	0.8453812	34.44	<.0001*	3.30437933	6.6182127
Smoke status[Smoke Now]	Biased	-0.0984645	0.0354107	7.73	0.0054*	-0.1679813	-0.0291517
SMOKE[Do Not Smoke Now]	Zeroed	0	0	.	.	.	.
Cholestrol		-0.0045422	0.0006988	42.25	<.0001*	-0.0059144	-0.0031739
PSU[1]		0.04073064	0.0315805	1.66	0.1971	-0.0211577	0.10265236
SDPSTRA6		-0.0021974	0.0024994	0.77	0.3793	-0.0070907	0.00270858
stat weight		7.2267e-6	3.7141e-6	3.79	0.0517	1.01528e-7	0.00001466
Age		-0.0775498	0.0025227	944.98	<.0001*	-0.0825499	-0.0726593
Sex[Female]		0.07734722	0.0456788	2.87	0.0904	-0.0121101	0.16697795
Race[White]		0.21486246	0.0789197	7.41	0.0065*	0.05726286	0.36710235
Race[Black]		-0.2969533	0.0864079	11.81	0.0006*	-0.4690514	-0.1299513
Body Weight		-0.0069151	0.0009405	54.06	<.0001*	-0.0087543	-0.0050663
Height		0.041178	0.0129174	10.16	0.0014*	0.0158603	0.0664957

Confidence limits are likelihood-based.
For log odds of Normal/High BP

Estimate is the value of every variable which denotes the level of impact on the target variable

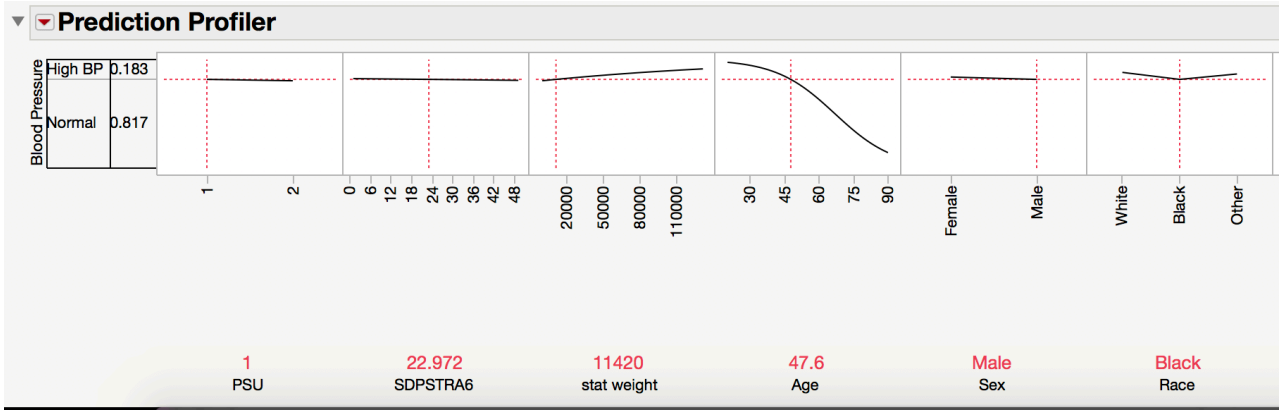
RSquare (U)	0.2365
AICc	6320.27
BIC	6410.94
Observations (or Sum Wgts)	7927

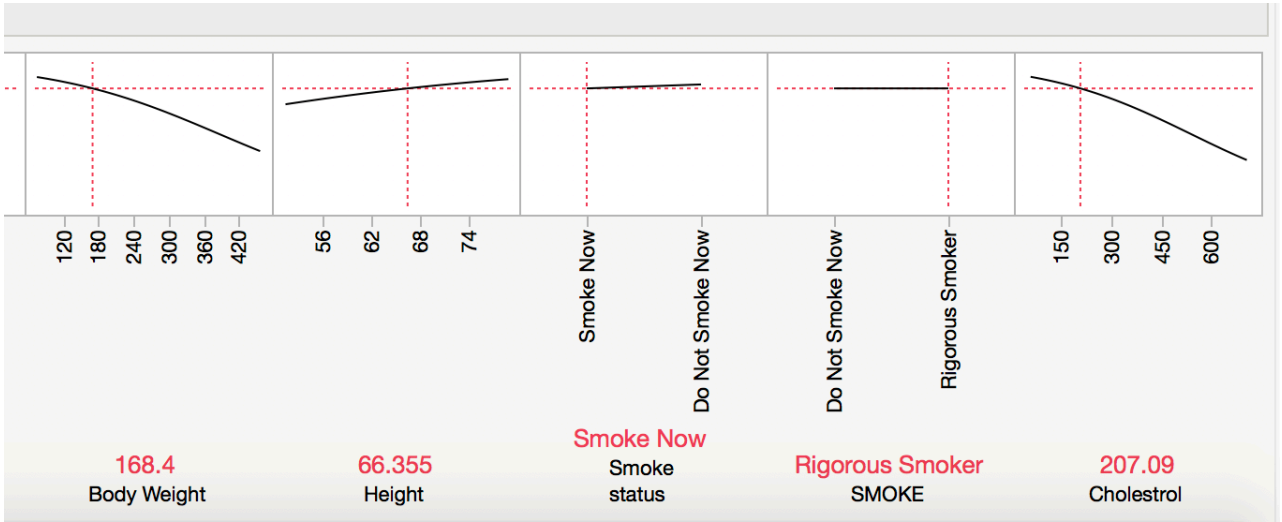
RSquare is 23%

Effect Summary

Source	LogWorth	PValue
Age	283.002	0.00000
Body Weight	12.395	0.00000
Cholestrol	10.089	0.00000
Race	8.225	0.00000
Height	2.851	0.00141
Smoke status	2.272	0.00534
stat weight	1.330	0.04675
Sex	1.045	0.09017
PSU	0.705	0.19709
SDPSTRA6	0.421	0.37961
SMOKE	.	.

The PValue should be less than 0.05, to consider the variable impacting the target variable. Here 'Age', 'Body Weight', 'Cholesterol', 'Race', 'Height' are the variables driving the target Variable.



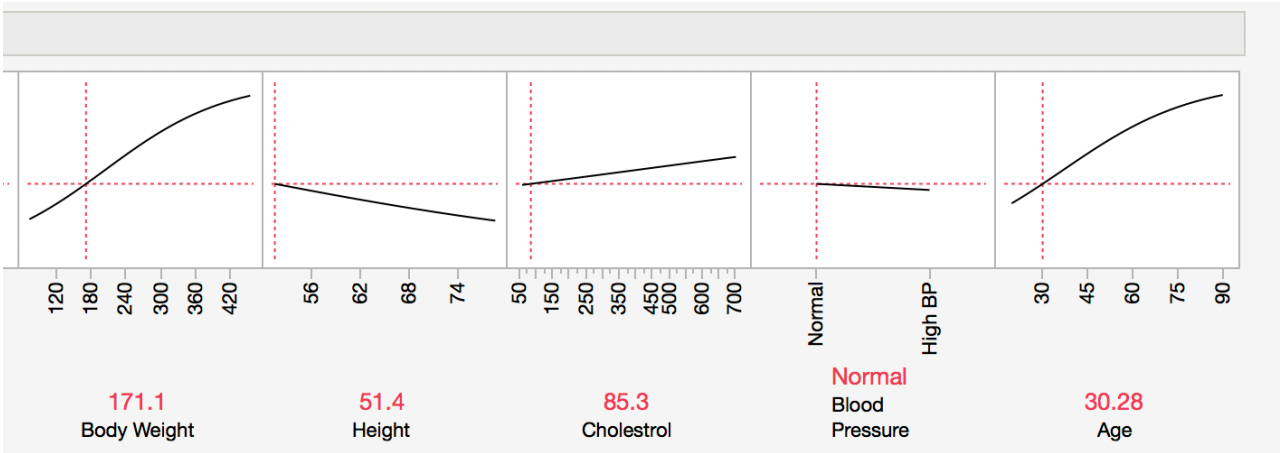


The above graphs explain the behavior of each variable on the target variable 'Blood Pressure'. For Example: As the age increase, the tendency to have a normal Blood Pressure decreases. The Two variable 'Smoke Status' and 'SMOKE' does not impact the Blood Pressure at all. The graph shows that these values are constant and does not affect the Blood Pressure, Moreover, PValue of these variables are exceeds 0.05. **Thus, Smoking does not lead to High Blood Pressure.**

PART 2: With Past Data Analysis, Variables leading to Rigorous Smoker

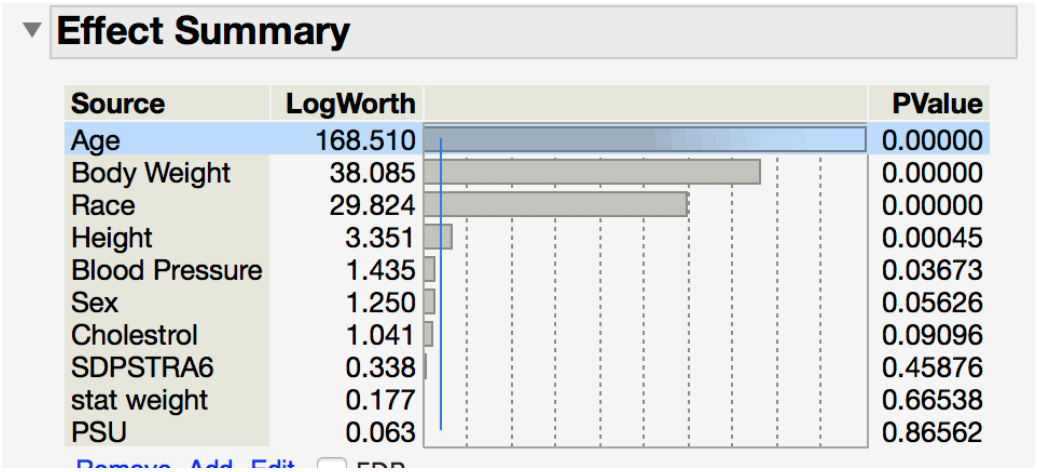
RSquare (U)	0.2365
AICc	6320.27
BIC	6410.94
Observations (or Sum Wgts)	7927

RSquare Value is 23.6%



The curves the Prediction Profiler indicates that the Variables 'Age', 'Race', 'Body Weight' and 'Height' are the variables driving the Variable 'SMOKE'. From the above Profiler analysis, even

though the ‘blood Pressure’ is set to ‘Normal’, the value of ‘Rigorous Smoker’ is high. The values of the Blood Pressure are almost constant and hence, not affecting a person to be a rigorous smoker.



From the above result, ‘Age’, ‘Body Weight’, ‘Race’ and ‘Height’ and ‘Blood Pressure’ has PValue less than 0.05. ‘Blood Pressure’ gives a very less impact, than other variables, to the target variable. Hence, we can say from the past data that the most Rigorous smoker is impacted only by Age, Height, Weight

3.8 Benefits of Model:

With the given data analysis by NHANES, we can say that Blood Pressure is not influenced much by smoking. And, the Data Analysis say that the person’s age, height, weight, Cholesterol and race are the factors majorly affecting the Blood Pressure as well as the tendency to be a rigorous smoker.