PROJECT REPORT

MGMT 635 – Data Mining and Analysis for Managers

Under the Guidance of

Prof. Stephan Kudyba

ABSTRACT

Performing data mining using data model like Regression and Neural Network and finding solutions to aid business problems.

Team Members

Shashank Parab
Pooja Suvarna
Shubhangi Rakhe
Atish Nayak
Likhit Jain

SPRING 2018

Table of Contents

Part 1

1.1. Objective

The Dataset provides explanatory and target variables to analyze various Call Centers of Insurance companies around the world that deal with their Insurance product and to gain greater insights as to what drives some Centers to cross-sell better than others regarding total sales.

1.2. Identifying Target and Driver Variable

Variable	Type	Description
Region	Not Required	Region of where the call center is located. Because of outsourcing activities, regions have been classified into zones.
Branch ID	Not Required	The identifier of the center.
Center Start Date	Required	The day the center came into functioning
Call Center Type	Not Required	Whether the center has Standard technology supported by simple dashboard system or new system.
System Type	Driver Variable	Classification code of System Type
Agents	Required	The number of call center agents at the center.
Facebook	Driver Variable	Whether the insurance company center actively corresponds to customers via Facebook.
Issue	Driver Variable	This is the focus of the customer issue (why the called the call center) as identified by voice mining conversations or mining notes taken by agents.

Survey Response	Driver Variable	How a caller rated their call experience (1 negative to 10 positive)
Web Support	Driver Variable	Whether the call center has a web self-service option for callers.
Agent Gender	Driver Variable	The gender of a call center agent.
Customer Positive Response	Driver Variable	The number of positive responses by a customer to the call center's cross-sell.
Product Inquiry	Driver Variable	The purpose of the phone call from a customer (e.g. what they called asking about)
# Phone Calls	Driver Variable	Amount of phone calls the center handles over the period of the analysis.
% Agents College Deg.	Driver Variable	The percentage of agents who have a college degree.
Red Flags	Driver Variable	The number of callers who asked for managers to complain to during calls.
Customer Age	Driver Variable	The age of a customer calling in.
Agent Rank	Driver Variable	The quality ranking of an agent at a call center (e.g. experience, tenacity…) Scale of (1 - 5); 1 is weak….5 is excellent.
Sales	Target Variable	Total revenue attributed to cross-sales by agents at a center.

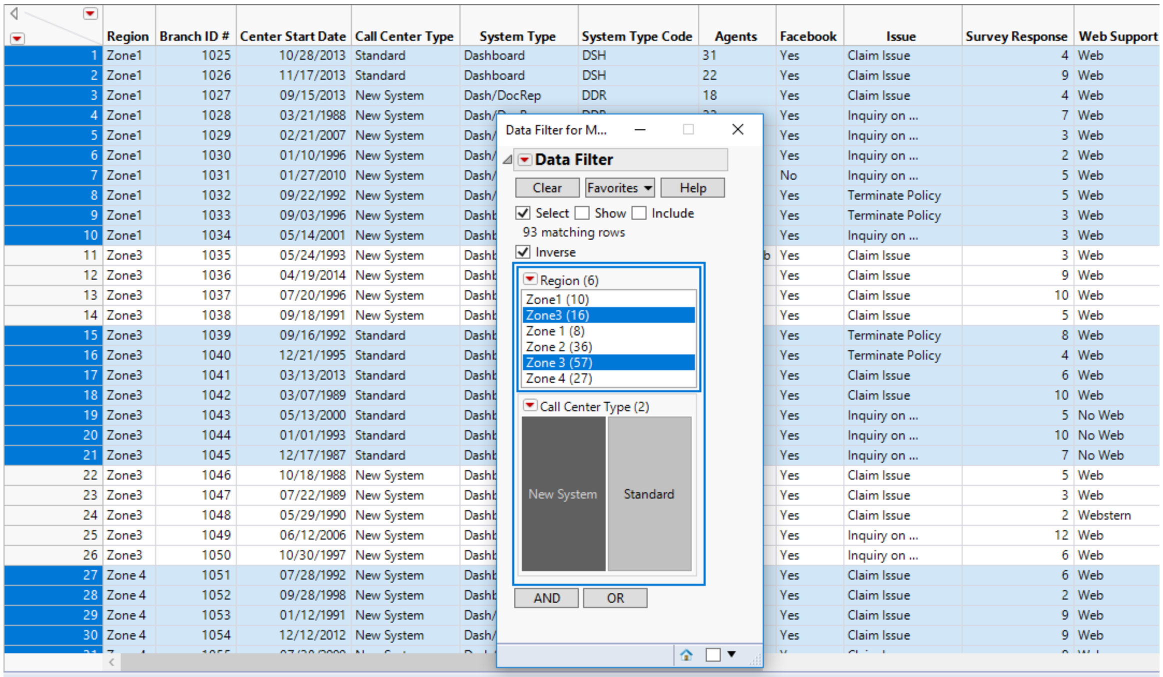
1.3. Troubleshoot Dataset

1.3.1. Filtering Data

As per the dataset criteria, the focus is to analyze the centers in Zone 3 that have New Systems and to find the cause that will help in better performance.

Below we have selected the required data, i.e., Zone 3 and New System.

- The dataset is sorted by region as Zone 3 and Call Center Type as New System using Data Filter in SAS JMP.
- Click on Inverse to select the unwanted data and delete it as per the business problem.



Data Filter for M...

Data Filter

Clear | Favorites | Help

☑ Select ☐ Show ☐ Include

93 matching rows

☑ Inverse

Region (6)

- Zone1 (10)
- Zone3 (16)
- Zone 1 (8)
- Zone 2 (36)
- Zone 3 (57)
- Zone 4 (27)

Call Center Type (2)

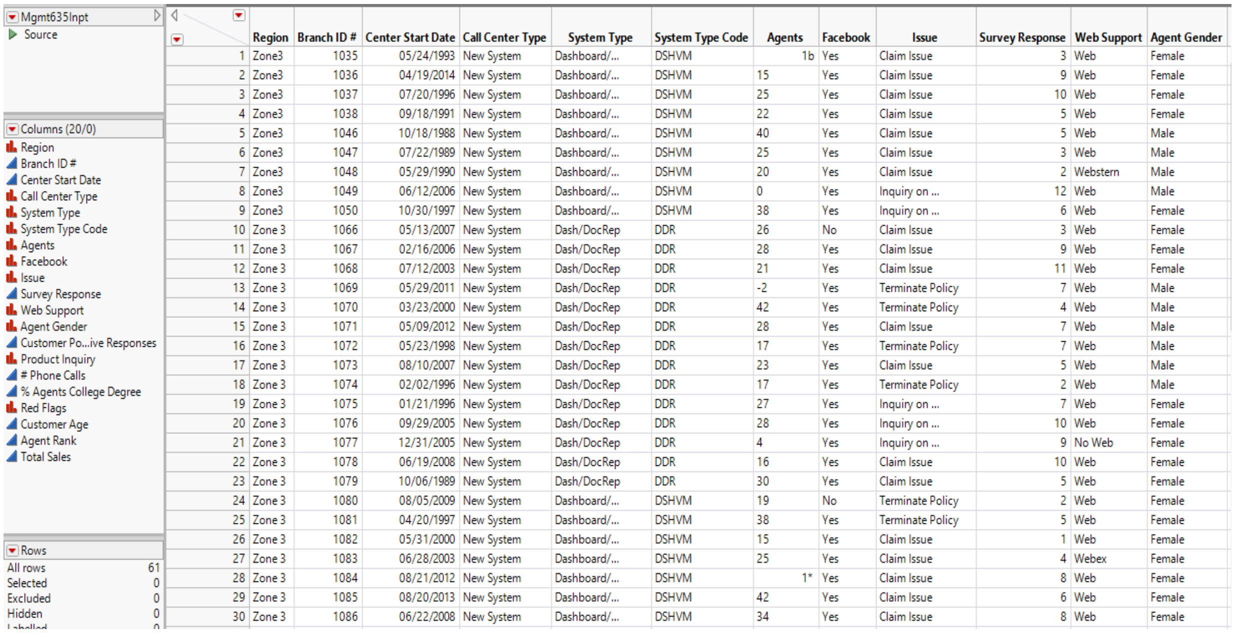
New System | Standard

AND | OR

	Region	Branch ID #	Center Start Date	Call Center Type	System Type	System Type Code	Agents	Facebook	Issue	Survey Response	Web Support
1	Zone1	1025	10/28/2013	Standard	Dashboard	DSH	31	Yes	Claim Issue	4	Web
2	Zone1	1026	11/17/2013	Standard	Dashboard	DSH	22	Yes	Claim Issue	9	Web
3	Zone1	1027	09/15/2013	New System	Dash/DocRep	DDR	18	Yes	Claim Issue	4	Web
4	Zone1	1028	03/21/1988	New System	Dash/			Yes	Inquiry on ...	7	Web
5	Zone1	1029	02/21/2007	New System	Dash/			Yes	Inquiry on ...	3	Web
6	Zone1	1030	01/10/1996	New System	Dash/			Yes	Inquiry on ...	2	Web
7	Zone1	1031	01/27/2010	New System	Dash/			No	Inquiry on ...	5	Web
8	Zone1	1032	09/22/1992	New System	Dash/			Yes	Terminate Policy	5	Web
9	Zone1	1033	09/03/1996	New System	Dashb			Yes	Terminate Policy	3	Web
10	Zone1	1034	05/14/2001	New System	Dashb			Yes	Inquiry on ...	3	Web
11	Zone3	1035	05/24/1993	New System	Dashb		b	Yes	Claim Issue	3	Web
12	Zone3	1036	04/19/2014	New System	Dashb			Yes	Claim Issue	9	Web
13	Zone3	1037	07/20/1996	New System	Dashb			Yes	Claim Issue	10	Web
14	Zone3	1038	09/18/1991	New System	Dashb			Yes	Claim Issue	5	Web
15	Zone3	1039	09/16/1992	Standard	Dashb			Yes	Terminate Policy	8	Web
16	Zone3	1040	12/21/1995	Standard	Dashb			Yes	Terminate Policy	4	Web
17	Zone3	1041	03/13/2013	Standard	Dashb			Yes	Claim Issue	6	Web
18	Zone3	1042	03/07/1989	Standard	Dashb			Yes	Claim Issue	10	Web
19	Zone3	1043	05/13/2000	Standard	Dashb			Yes	Inquiry on ...	5	No Web
20	Zone3	1044	01/01/1993	Standard	Dashb			Yes	Inquiry on ...	10	No Web
21	Zone3	1045	12/17/1987	Standard	Dashb			Yes	Inquiry on ...	7	No Web
22	Zone3	1046	10/18/1988	New System	Dashb			Yes	Claim Issue	5	Web
23	Zone3	1047	07/22/1989	New System	Dashb			Yes	Claim Issue	3	Web
24	Zone3	1048	05/29/1990	New System	Dashb			Yes	Claim Issue	2	Webstern
25	Zone3	1049	06/12/2006	New System	Dashb			Yes	Inquiry on ...	12	Web
26	Zone3	1050	10/30/1997	New System	Dashb			Yes	Inquiry on ...	6	Web
27	Zone 4	1051	07/28/1992	New System	Dashb			Yes	Claim Issue	6	Web
28	Zone 4	1052	09/28/1998	New System	Dashb			Yes	Claim Issue	2	Web
29	Zone 4	1053	01/12/1991	New System	Dash/			Yes	Claim Issue	9	Web
30	Zone 4	1054	12/12/2012	New System	Dash/			Yes	Claim Issue	9	Web

Fig 1.3.1. Select unwanted data using Data Filter

After deletion, the following dataset will be obtained.



Mgmt635Inpt

Source

Columns (20/0)

- Region
- Branch ID #
- Center Start Date
- Call Center Type
- System Type
- System Type Code
- Agents
- Facebook
- Issue
- Survey Response
- Web Support
- Agent Gender
- Customer Po...ive Responses
- Product Inquiry
- # Phone Calls
- % Agents College Degree
- Red Flags
- Customer Age
- Agent Rank
- Total Sales

Rows

All rows	61
Selected	0
Excluded	0
Hidden	0

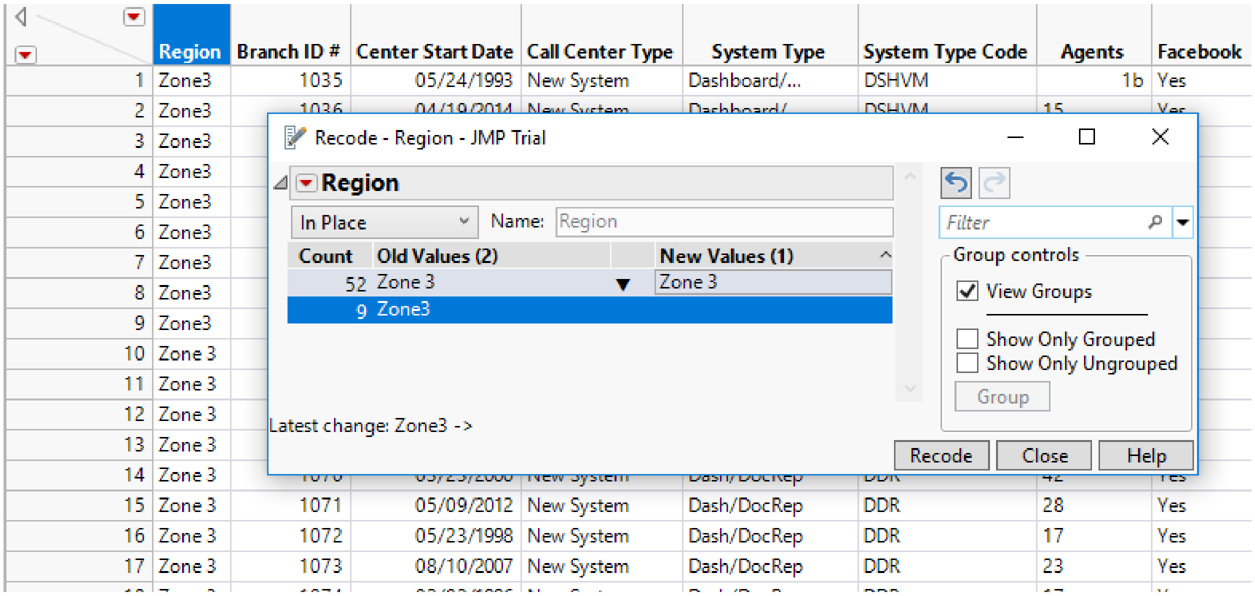
	Region	Branch ID #	Center Start Date	Call Center Type	System Type	System Type Code	Agents	Facebook	Issue	Survey Response	Web Support	Agent Gender
1	Zone3	1035	05/24/1993	New System	Dashboard/...	DSHVM	1b	Yes	Claim Issue	3	Web	Female
2	Zone3	1036	04/19/2014	New System	Dashboard/...	DSHVM	15	Yes	Claim Issue	9	Web	Female
3	Zone3	1037	07/20/1996	New System	Dashboard/...	DSHVM	25	Yes	Claim Issue	10	Web	Female
4	Zone3	1038	09/18/1991	New System	Dashboard/...	DSHVM	22	Yes	Claim Issue	5	Web	Female
5	Zone3	1046	10/18/1988	New System	Dashboard/...	DSHVM	40	Yes	Claim Issue	5	Web	Male
6	Zone3	1047	07/22/1989	New System	Dashboard/...	DSHVM	25	Yes	Claim Issue	3	Web	Male
7	Zone3	1048	05/29/1990	New System	Dashboard/...	DSHVM	20	Yes	Claim Issue	2	Webstern	Male
8	Zone3	1049	06/12/2006	New System	Dashboard/...	DSHVM	0	Yes	Inquiry on ...	12	Web	Male
9	Zone3	1050	10/30/1997	New System	Dashboard/...	DSHVM	38	Yes	Inquiry on ...	6	Web	Female
10	Zone 3	1066	05/13/2007	New System	Dash/DocRep	DDR	26	No	Claim Issue	3	Web	Female
11	Zone 3	1067	02/16/2006	New System	Dash/DocRep	DDR	28	Yes	Claim Issue	9	Web	Female
12	Zone 3	1068	07/12/2003	New System	Dash/DocRep	DDR	21	Yes	Claim Issue	11	Web	Female
13	Zone 3	1069	05/29/2011	New System	Dash/DocRep	DDR	-2	Yes	Terminate Policy	7	Web	Male
14	Zone 3	1070	03/23/2000	New System	Dash/DocRep	DDR	42	Yes	Terminate Policy	4	Web	Male
15	Zone 3	1071	05/09/2012	New System	Dash/DocRep	DDR	28	Yes	Claim Issue	7	Web	Male
16	Zone 3	1072	05/23/1998	New System	Dash/DocRep	DDR	17	Yes	Terminate Policy	7	Web	Male
17	Zone 3	1073	08/10/2007	New System	Dash/DocRep	DDR	23	Yes	Claim Issue	5	Web	Male
18	Zone 3	1074	02/02/1996	New System	Dash/DocRep	DDR	17	Yes	Terminate Policy	2	Web	Male
19	Zone 3	1075	01/21/1996	New System	Dash/DocRep	DDR	27	Yes	Inquiry on ...	7	Web	Female
20	Zone 3	1076	09/29/2005	New System	Dash/DocRep	DDR	28	Yes	Inquiry on ...	10	Web	Female
21	Zone 3	1077	12/31/2005	New System	Dash/DocRep	DDR	4	Yes	Inquiry on ...	9	No Web	Female
22	Zone 3	1078	06/19/2008	New System	Dash/DocRep	DDR	16	Yes	Claim Issue	10	Web	Female
23	Zone 3	1079	10/06/1989	New System	Dash/DocRep	DDR	30	Yes	Claim Issue	5	Web	Female
24	Zone 3	1080	08/05/2009	New System	Dashboard/...	DSHVM	19	No	Terminate Policy	2	Web	Female
25	Zone 3	1081	04/20/1997	New System	Dashboard/...	DSHVM	38	Yes	Terminate Policy	5	Web	Female
26	Zone 3	1082	05/31/2000	New System	Dashboard/...	DSHVM	15	Yes	Claim Issue	1	Web	Female
27	Zone 3	1083	06/28/2003	New System	Dashboard/...	DSHVM	25	Yes	Claim Issue	4	Webex	Female
28	Zone 3	1084	08/21/2012	New System	Dashboard/...	DSHVM	1*	Yes	Claim Issue	8	Web	Female
29	Zone 3	1085	08/20/2013	New System	Dashboard/...	DSHVM	42	Yes	Claim Issue	6	Web	Female
30	Zone 3	1086	06/22/2008	New System	Dashboard/...	DSHVM	34	Yes	Claim Issue	8	Web	Female

Fig 1.3.2. After Data Filter

1.3.2. Format the Data

The Column Region includes entries as ‘Zone3’ and ‘Zone 3’. Therefore, we need to format the data to have consistency in the value.

Here, we recoded the value to ‘Zone 3’.



	Region	Branch ID #	Center Start Date	Call Center Type	System Type	System Type Code	Agents	Facebook
1	Zone3	1035	05/24/1993	New System	Dashboard/...	DSHVM	1b	Yes
2	Zone3	1036	04/19/2014	New System	Dashboard/...	DSHVM	15	Yes
3	Zone3							
4	Zone3							
5	Zone3							
6	Zone3							
7	Zone3							
8	Zone3							
9	Zone3							
10	Zone 3							
11	Zone 3							
12	Zone 3							
13	Zone 3							
14	Zone 3	1070	03/23/2000	New System	Dash/DocRep	DDR	42	Yes
15	Zone 3	1071	05/09/2012	New System	Dash/DocRep	DDR	28	Yes
16	Zone 3	1072	05/23/1998	New System	Dash/DocRep	DDR	17	Yes
17	Zone 3	1073	08/10/2007	New System	Dash/DocRep	DDR	23	Yes

Recode - Region - JMP Trial

Region

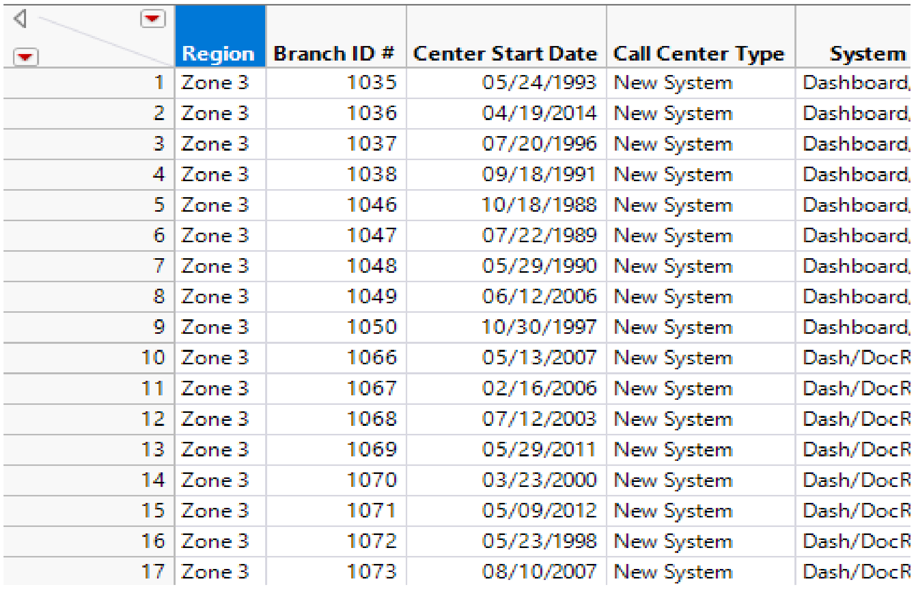
In Place Name: Region

Count	Old Values (2)	New Values (1)
52	Zone 3	Zone 3
9	Zone3	

Latest change: Zone3 ->

Fig 1.3.2.1. Recoding ‘Zone3’ to ‘Zone 3’

After recoding, the column region will have only one value throughout, i.e., Zone 3.

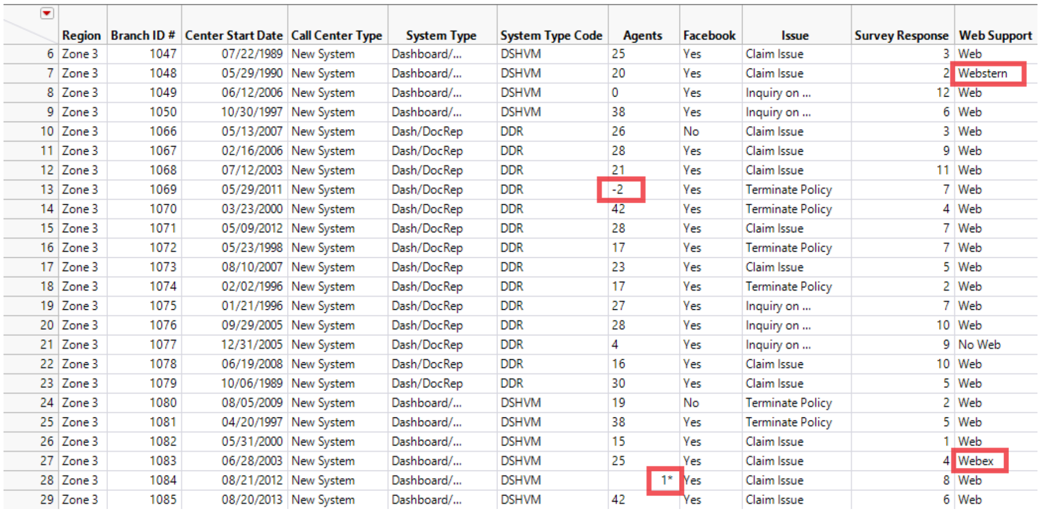


	Region	Branch ID #	Center Start Date	Call Center Type	System
1	Zone 3	1035	05/24/1993	New System	Dashboard/
2	Zone 3	1036	04/19/2014	New System	Dashboard/
3	Zone 3	1037	07/20/1996	New System	Dashboard/
4	Zone 3	1038	09/18/1991	New System	Dashboard/
5	Zone 3	1046	10/18/1988	New System	Dashboard/
6	Zone 3	1047	07/22/1989	New System	Dashboard/
7	Zone 3	1048	05/29/1990	New System	Dashboard/
8	Zone 3	1049	06/12/2006	New System	Dashboard/
9	Zone 3	1050	10/30/1997	New System	Dashboard/
10	Zone 3	1066	05/13/2007	New System	Dash/DocR
11	Zone 3	1067	02/16/2006	New System	Dash/DocR
12	Zone 3	1068	07/12/2003	New System	Dash/DocR
13	Zone 3	1069	05/29/2011	New System	Dash/DocR
14	Zone 3	1070	03/23/2000	New System	Dash/DocR
15	Zone 3	1071	05/09/2012	New System	Dash/DocR
16	Zone 3	1072	05/23/1998	New System	Dash/DocR
17	Zone 3	1073	08/10/2007	New System	Dash/DocR

Fig 1.3.2.2. After Recoding

1.3.3. Clean the Data

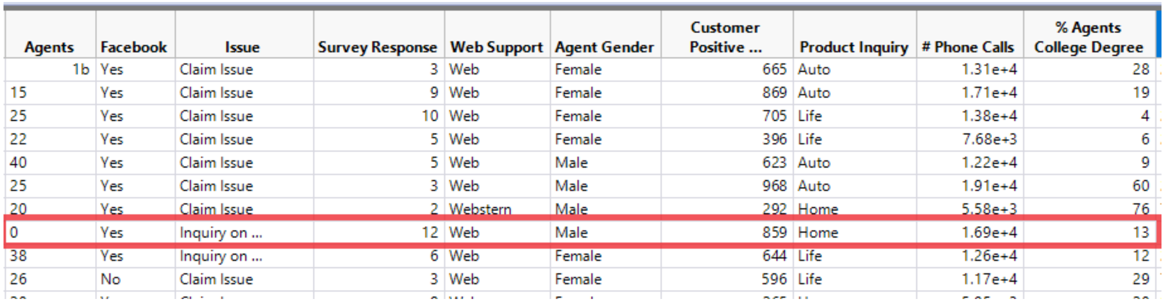
Check for dirty data and delete that row.



	Region	Branch ID #	Center Start Date	Call Center Type	System Type	System Type Code	Agents	Facebook	Issue	Survey Response	Web Support
6	Zone 3	1047	07/22/1989	New System	Dashboard/...	DSHVM	25	Yes	Claim Issue	3	Web
7	Zone 3	1048	05/29/1990	New System	Dashboard/...	DSHVM	20	Yes	Claim Issue	2	Webstern
8	Zone 3	1049	06/12/2006	New System	Dashboard/...	DSHVM	0	Yes	Inquiry on ...	12	Web
9	Zone 3	1050	10/30/1997	New System	Dashboard/...	DSHVM	38	Yes	Inquiry on ...	6	Web
10	Zone 3	1066	05/13/2007	New System	Dash/DocRep	DDR	26	No	Claim Issue	3	Web
11	Zone 3	1067	02/16/2006	New System	Dash/DocRep	DDR	28	Yes	Claim Issue	9	Web
12	Zone 3	1068	07/12/2003	New System	Dash/DocRep	DDR	21	Yes	Claim Issue	11	Web
13	Zone 3	1069	05/29/2011	New System	Dash/DocRep	DDR	-2	Yes	Terminate Policy	7	Web
14	Zone 3	1070	03/23/2000	New System	Dash/DocRep	DDR	42	Yes	Terminate Policy	4	Web
15	Zone 3	1071	05/09/2012	New System	Dash/DocRep	DDR	28	Yes	Claim Issue	7	Web
16	Zone 3	1072	05/23/1998	New System	Dash/DocRep	DDR	17	Yes	Terminate Policy	7	Web
17	Zone 3	1073	08/10/2007	New System	Dash/DocRep	DDR	23	Yes	Claim Issue	5	Web
18	Zone 3	1074	02/02/1996	New System	Dash/DocRep	DDR	17	Yes	Terminate Policy	2	Web
19	Zone 3	1075	01/21/1996	New System	Dash/DocRep	DDR	27	Yes	Inquiry on ...	7	Web
20	Zone 3	1076	09/29/2005	New System	Dash/DocRep	DDR	28	Yes	Inquiry on ...	10	Web
21	Zone 3	1077	12/31/2005	New System	Dash/DocRep	DDR	4	Yes	Inquiry on ...	9	No Web
22	Zone 3	1078	06/19/2008	New System	Dash/DocRep	DDR	16	Yes	Claim Issue	10	Web
23	Zone 3	1079	10/06/1989	New System	Dash/DocRep	DDR	30	Yes	Claim Issue	5	Web
24	Zone 3	1080	08/05/2009	New System	Dashboard/...	DSHVM	19	No	Terminate Policy	2	Web
25	Zone 3	1081	04/20/1997	New System	Dashboard/...	DSHVM	38	Yes	Terminate Policy	5	Web
26	Zone 3	1082	05/31/2000	New System	Dashboard/...	DSHVM	15	Yes	Claim Issue	1	Web
27	Zone 3	1083	06/28/2003	New System	Dashboard/...	DSHVM	25	Yes	Claim Issue	4	Webex
28	Zone 3	1084	08/21/2012	New System	Dashboard/...	DSHVM	1*	Yes	Claim Issue	8	Web
29	Zone 3	1085	08/20/2013	New System	Dashboard/...	DSHVM	42	Yes	Claim Issue	6	Web

Fig 1.3.3.1. Dirty Data in the dataset

The below highlighted row is to be deleted because No. of Agents is 0 and % Agents College Degree is 13.

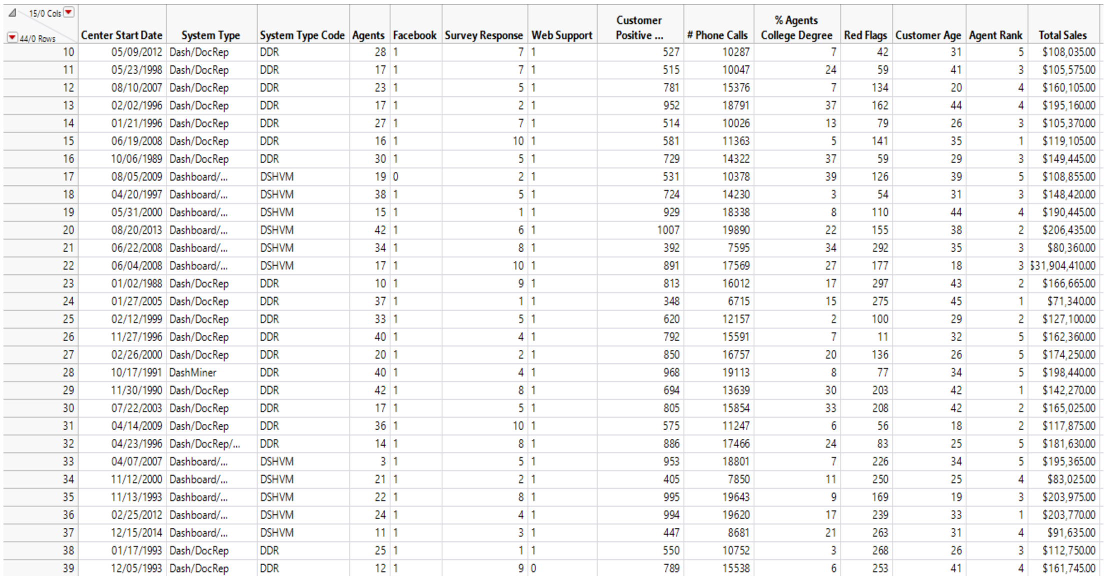


Agents	Facebook	Issue	Survey Response	Web Support	Agent Gender	Customer Positive ...	Product Inquiry	# Phone Calls	% Agents College Degree
1b	Yes	Claim Issue	3	Web	Female	665	Auto	1.31e+4	28
15	Yes	Claim Issue	9	Web	Female	869	Auto	1.71e+4	19
25	Yes	Claim Issue	10	Web	Female	705	Life	1.38e+4	4
22	Yes	Claim Issue	5	Web	Female	396	Life	7.68e+3	6
40	Yes	Claim Issue	5	Web	Male	623	Auto	1.22e+4	9
25	Yes	Claim Issue	3	Web	Male	968	Auto	1.91e+4	60
20	Yes	Claim Issue	2	Webstern	Male	292	Home	5.58e+3	76
0	Yes	Inquiry on ...	12	Web	Male	859	Home	1.69e+4	13
38	Yes	Inquiry on ...	6	Web	Female	644	Life	1.26e+4	12
26	No	Claim Issue	3	Web	Female	596	Life	1.17e+4	29

Fig 1.3.3.2. Invalid Data

Also, delete the irrelevant columns from dataset which have no affect on the Target variable.
The following columns are deleted:

- Branch ID #
- Region
- Call Center Type
- Issue
- Agent Gender
- Product Inquiry
- Customer Age



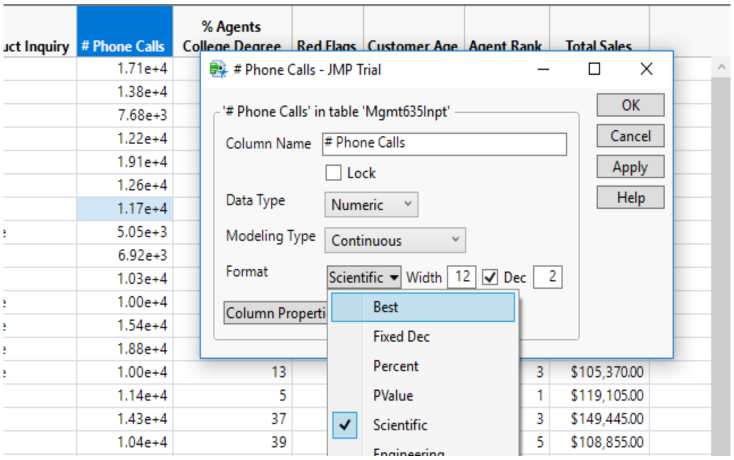
15/0 Cols 44/0 Rows	Center Start Date	System Type	System Type Code	Agents	Facebook	Survey Response	Web Support	Customer Positive ...	# Phone Calls	% Agents College Degree	Red Flags	Customer Age	Agent Rank	Total Sales
10	05/09/2012	Dash/DocRep	DDR	28	1	7	1	527	10287	7	42	31	5	$108,035.00
11	05/23/1998	Dash/DocRep	DDR	17	1	7	1	515	10047	24	59	41	3	$105,575.00
12	08/10/2007	Dash/DocRep	DDR	23	1	5	1	781	15376	7	134	20	4	$160,105.00
13	02/02/1996	Dash/DocRep	DDR	17	1	2	1	952	18791	37	162	44	4	$195,160.00
14	01/21/1996	Dash/DocRep	DDR	27	1	7	1	514	10026	13	79	26	3	$105,370.00
15	06/19/2008	Dash/DocRep	DDR	16	1	10	1	581	11363	5	141	35	1	$119,105.00
16	10/06/1989	Dash/DocRep	DDR	30	1	5	1	729	14322	37	59	29	3	$149,445.00
17	08/05/2009	Dashboard/...	DSHVM	19	0	2	1	531	10378	39	126	39	5	$108,855.00
18	04/20/1997	Dashboard/...	DSHVM	38	1	5	1	724	14230	3	54	31	3	$148,420.00
19	05/31/2000	Dashboard/...	DSHVM	15	1	1	1	929	18338	8	110	44	4	$190,445.00
20	08/20/2013	Dashboard/...	DSHVM	42	1	6	1	1007	19890	22	155	38	2	$206,435.00
21	06/22/2008	Dashboard/...	DSHVM	34	1	8	1	392	7595	34	292	35	3	$80,360.00
22	06/04/2008	Dashboard/...	DSHVM	17	1	10	1	891	17569	27	177	18	3	$31,904,410.00
23	01/02/1988	Dash/DocRep	DDR	10	1	9	1	813	16012	17	297	43	2	$166,665.00
24	01/27/2005	Dash/DocRep	DDR	37	1	1	1	348	6715	15	275	45	1	$71,340.00
25	02/12/1999	Dash/DocRep	DDR	33	1	5	1	620	12157	2	100	29	2	$127,100.00
26	11/27/1996	Dash/DocRep	DDR	40	1	4	1	792	15591	7	11	32	5	$162,360.00
27	02/26/2000	Dash/DocRep	DDR	20	1	2	1	850	16757	20	136	26	5	$174,250.00
28	10/17/1991	DashMiner	DDR	40	1	4	1	968	19113	8	77	34	5	$198,440.00
29	11/30/1990	Dash/DocRep	DDR	42	1	8	1	694	13639	30	203	42	1	$142,270.00
30	07/22/2003	Dash/DocRep	DDR	17	1	5	1	805	15854	33	208	42	2	$165,025.00
31	04/14/2009	Dash/DocRep	DDR	36	1	10	1	575	11247	6	56	18	2	$117,875.00
32	04/23/1996	Dash/DocRep/...	DDR	14	1	8	1	886	17466	24	83	25	5	$181,630.00
33	04/07/2007	Dashboard/...	DSHVM	3	1	5	1	953	18801	7	226	34	5	$195,365.00
34	11/12/2000	Dashboard/...	DSHVM	21	1	2	1	405	7850	11	250	25	4	$83,025.00
35	11/13/1993	Dashboard/...	DSHVM	22	1	8	1	995	19643	9	169	19	3	$203,975.00
36	02/25/2012	Dashboard/...	DSHVM	24	1	4	1	994	19620	17	239	33	1	$203,770.00
37	12/15/2014	Dashboard/...	DSHVM	11	1	3	1	447	8681	21	263	31	4	$91,635.00
38	01/17/1993	Dash/DocRep	DDR	25	1	1	1	550	10752	3	268	26	3	$112,750.00
39	12/05/1993	Dash/DocRep	DDR	12	1	9	0	789	15538	6	253	41	4	$161,745.00

Fig 1.3.3.3. Dataset after cleansing

1.3.4. Transform the Data

The dataset includes certain column with wrong data type. The data types of the following columns were changed:

Agents – from Character to Numeric and Continuous
Facebook – Recode Yes to 1 and No to 0
Web Support – Recode Web to 1 and No Web to 0
Center Start Date – from Character to Date
Phone Calls – Format from Numeric to Best
Red Flags – From Character to Numeric and Continuous



# Phone Calls - JMP Trial

'# Phone Calls' in table 'Mgmt635Inpt'

Column Name: # Phone Calls

☐ Lock

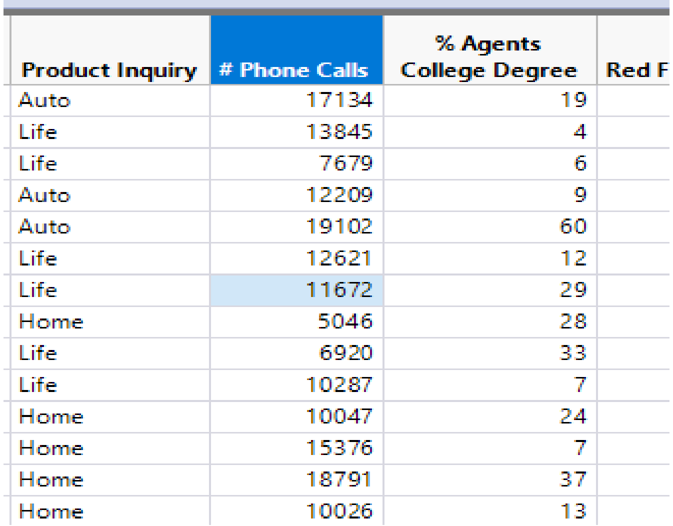
Data Type: Numeric

Modeling Type: Continuous

Format: Scientific, Width 12, ☑ Dec 2

Fig 1.3.4.1. Changing the format of 'Phone Calls' to Best

After transforming,



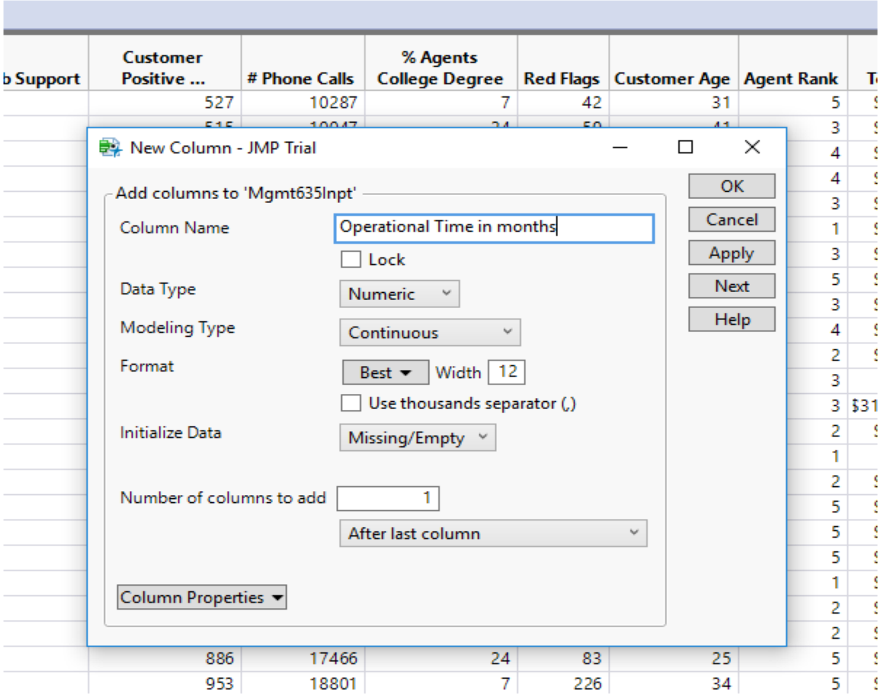
Product Inquiry	# Phone Calls	% Agents College Degree	Red F
Auto	17134	19	
Life	13845	4	
Life	7679	6	
Auto	12209	9	
Auto	19102	60	
Life	12621	12	
Life	11672	29	
Home	5046	28	
Life	6920	33	
Life	10287	7	
Home	10047	24	
Home	15376	7	
Home	18791	37	
Home	10026	13	

Fig 1.3.4.2. After Transforming the data

1.3.5. Addition of Column

As per the business rule, the amount of time a call center has been in operation is to be calculated in months.

- Create a new Column named ‘Amount of time a call center has been in operation (months)’.
- Set Data type to Numeric and Modeling type to Continuous.



b Support	Customer Positive ...	# Phone Calls	% Agents College Degree	Red Flags	Customer Age	Agent Rank
	527	10287	7	42	31	5
						3
						4
						4
						3
						1
						3
						5
						3
						4
						2
						3
						3
						2
						1
						2
						5
						5
						5
						1
						2
						2
	886	17466	24	83	25	5
	953	18801	7	226	34	5

New Column - JMP Trial

Add columns to 'Mgmt635Inpt'

Column Name: Operational Time in months

□ Lock

Data Type: Numeric

Modeling Type: Continuous

Format: Best Width 12

□ Use thousands separator (,)

Initialize Data: Missing/Empty

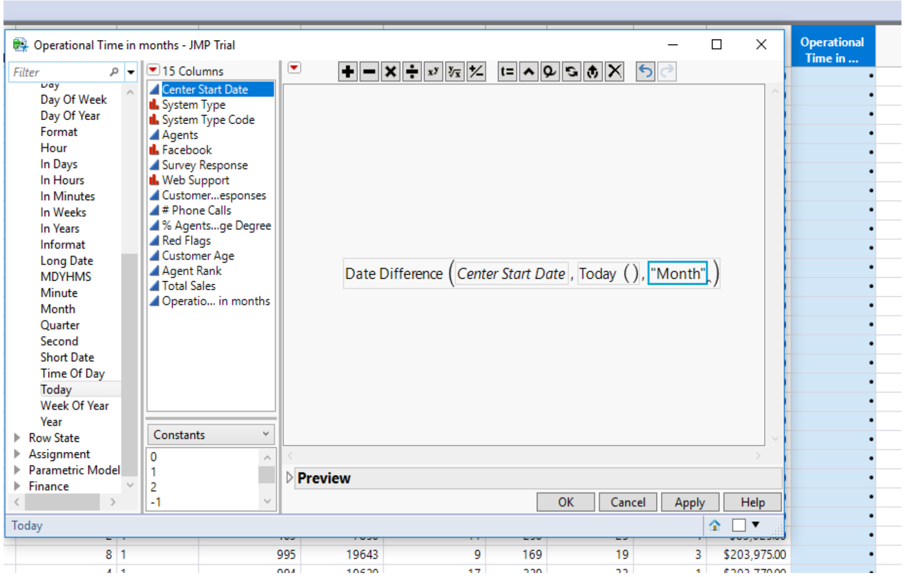
Number of columns to add: 1

After last column

Fig 1.3.5.1. Insert the new column details

Add formula to the column as given below,

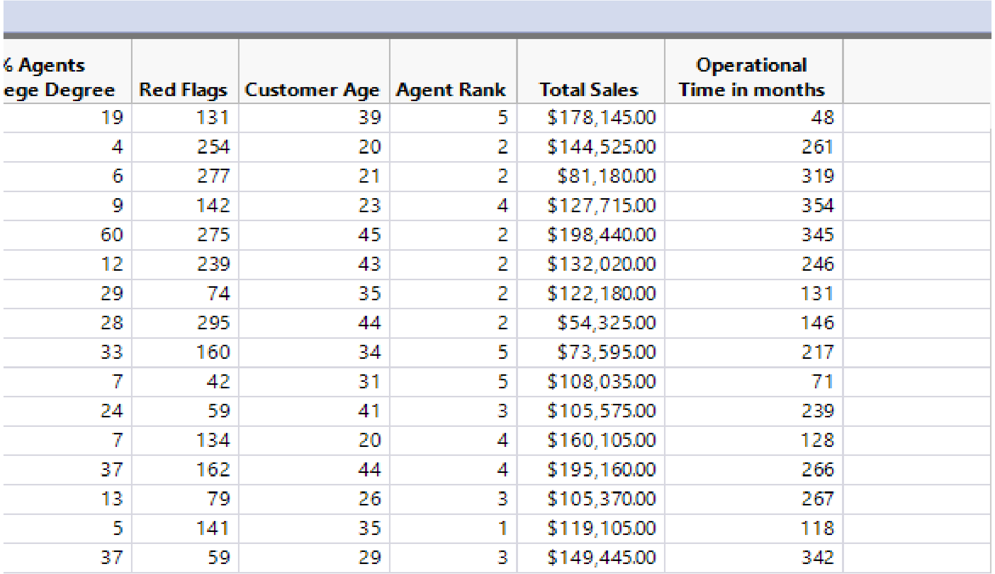
Amount of time a call center has been in operation = Center Start Date – Today



Date Difference (Center Start Date, Today (), "Month")

Fig 1.3.5.2. Add the required formula

Hence, the result.



% Agents ege Degree	Red Flags	Customer Age	Agent Rank	Total Sales	Operational Time in months
19	131	39	5	$178,145.00	48
4	254	20	2	$144,525.00	261
6	277	21	2	$81,180.00	319
9	142	23	4	$127,715.00	354
60	275	45	2	$198,440.00	345
12	239	43	2	$132,020.00	246
29	74	35	2	$122,180.00	131
28	295	44	2	$54,325.00	146
33	160	34	5	$73,595.00	217
7	42	31	5	$108,035.00	71
24	59	41	3	$105,575.00	239
7	134	20	4	$160,105.00	128
37	162	44	4	$195,160.00	266
13	79	26	3	$105,370.00	267
5	141	35	1	$119,105.00	118
37	59	29	3	$149,445.00	342

Fig 1.3.5.3. After Transforming the data

1.4. Degrees of Freedom

1. Before Cleaning Data

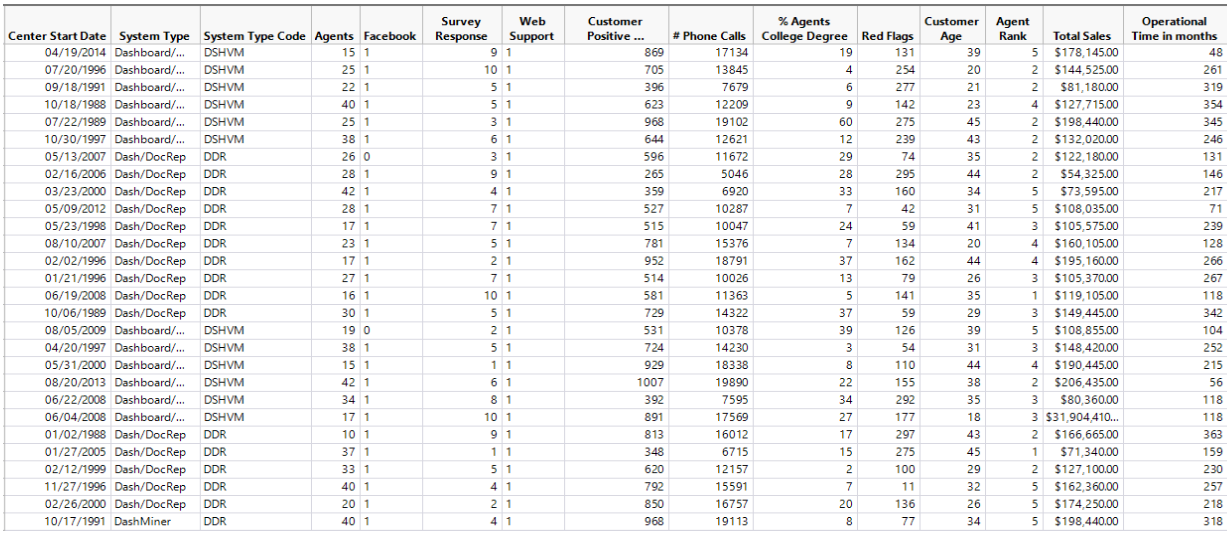
Number of columns- 20
Number of Rows- 154
Target Variable- Total Sales
Number of Target Variable- 1
Number of Explanatory variable- 19
Degree of Freedom= Number of Rows – Number of Explanatory variables
= 154 -19
= 135

2. After Data cleaning

Number of columns - 14
Number of Rows - 44
Target Variable- Total Sales
Number of Target Variable - 1
Number of Explanatory variable - 12
Degree of Freedom= Number of Rows – Number of Explanatory variables
= 44 - 12
= 32

1.5. Final Dataset

Below is the cleaned dataset.



Center Start Date	System Type	System Type Code	Agents	Facebook	Survey Response	Web Support	Customer Positive ...	# Phone Calls	% Agents College Degree	Red Flags	Customer Age	Agent Rank	Total Sales	Operational Time in months
04/19/2014	Dashboard/...	DSHVM	15	1	9	1	869	17134	19	131	39	5	$178,145.00	48
07/20/1996	Dashboard/...	DSHVM	25	1	10	1	705	13845	4	254	20	2	$144,525.00	261
09/18/1991	Dashboard/...	DSHVM	22	1	5	1	396	7679	6	277	21	2	$81,180.00	319
10/18/1988	Dashboard/...	DSHVM	40	1	5	1	623	12209	9	142	23	4	$127,715.00	354
07/22/1989	Dashboard/...	DSHVM	25	1	3	1	968	19102	60	275	45	2	$198,440.00	345
10/30/1997	Dashboard/...	DSHVM	38	1	6	1	644	12621	12	239	43	2	$132,020.00	246
05/13/2007	Dash/DocRep	DDR	26	0	3	1	596	11672	29	74	35	2	$122,180.00	131
02/16/2006	Dash/DocRep	DDR	28	1	9	1	265	5046	28	295	44	2	$54,325.00	146
03/23/2000	Dash/DocRep	DDR	42	1	4	1	359	6920	33	160	34	5	$73,595.00	217
05/09/2012	Dash/DocRep	DDR	28	1	7	1	527	10287	7	42	31	5	$108,035.00	71
05/23/1998	Dash/DocRep	DDR	17	1	7	1	515	10047	24	59	41	3	$105,575.00	239
08/10/2007	Dash/DocRep	DDR	23	1	5	1	781	15376	7	134	20	4	$160,105.00	128
02/02/1996	Dash/DocRep	DDR	17	1	2	1	952	18791	37	162	44	4	$195,160.00	266
01/21/1996	Dash/DocRep	DDR	27	1	7	1	514	10026	13	79	26	3	$105,370.00	267
06/19/2008	Dash/DocRep	DDR	16	1	10	1	581	11363	5	141	35	1	$119,105.00	118
10/06/1989	Dash/DocRep	DDR	30	1	5	1	729	14322	37	59	29	3	$149,445.00	342
08/05/2009	Dashboard/...	DSHVM	19	0	2	1	531	10378	39	126	39	5	$108,855.00	104
04/20/1997	Dashboard/...	DSHVM	38	1	5	1	724	14230	3	54	31	3	$148,420.00	252
05/31/2000	Dashboard/...	DSHVM	15	1	1	1	929	18338	8	110	44	4	$190,445.00	215
08/20/2013	Dashboard/...	DSHVM	42	1	6	1	1007	19890	22	155	38	2	$206,435.00	56
06/22/2008	Dashboard/...	DSHVM	34	1	8	1	392	7595	34	292	35	3	$80,360.00	118
06/04/2008	Dashboard/...	DSHVM	17	1	10	1	891	17569	27	177	18	3	$31,904,410...	118
01/02/1988	Dash/DocRep	DDR	10	1	9	1	813	16012	17	297	43	2	$166,665.00	363
01/27/2005	Dash/DocRep	DDR	37	1	1	1	348	6715	15	275	45	1	$71,340.00	159
02/12/1999	Dash/DocRep	DDR	33	1	5	1	620	12157	2	100	29	2	$127,100.00	230
11/27/1996	Dash/DocRep	DDR	40	1	4	1	792	15591	7	11	32	5	$162,360.00	257
02/26/2000	Dash/DocRep	DDR	20	1	2	1	850	16757	20	136	26	5	$174,250.00	218
10/17/1991	DashMiner	DDR	40	1	4	1	968	19113	8	77	34	5	$198,440.00	318

Fig 1.5.1. The final dataset

Part 2

2.1. Objective

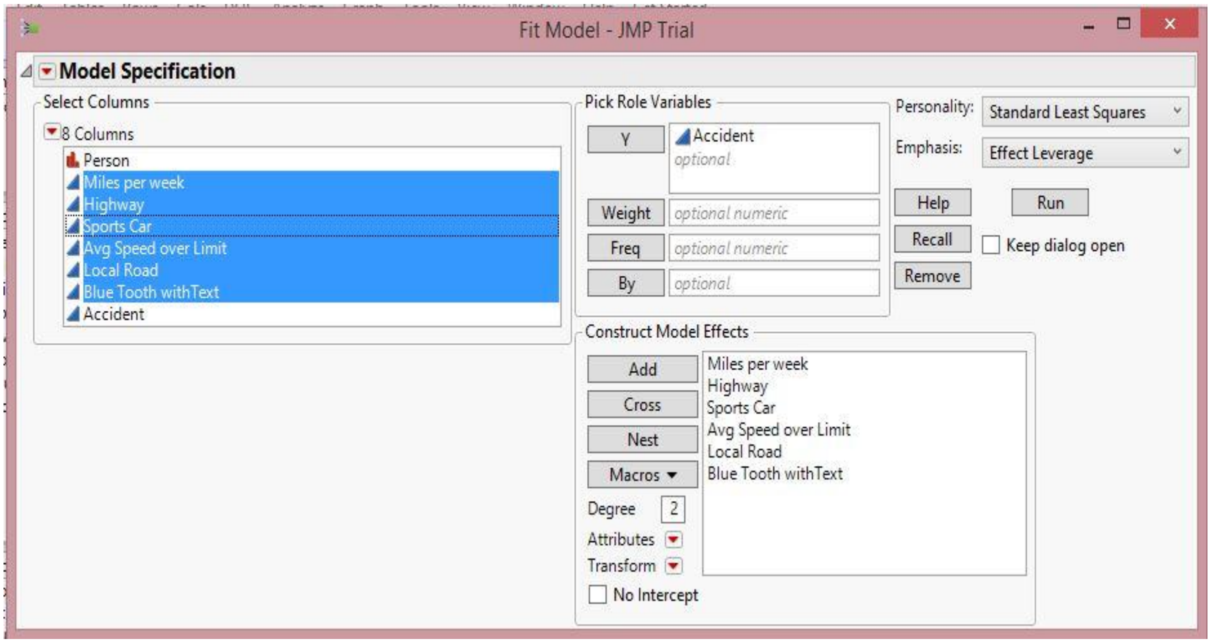
You will use the statistical output of the model to help make decisions on how to estimate risk for automobile insurance customers. You are to analyze your regression results and devise a simple business plan using whatever information is critical to your strategic decision. You also are required to use the results of your model to estimate the risk level for new potential insurance customers.

2.2. Process Description

The dataset is imported from Excel to JMP. The Fit Model tool from JMP is used to perform regression analysis.

2.3. Regression Analysis

Following screenshot describes the target variable and the driver variables consider for Regression Analysis:



Fit Model - JMP Trial

Model Specification

Select Columns

8 Columns

- Person
- Miles per week
- Highway
- Sports Car
- Avg Speed over Limit
- Local Road
- Blue Tooth withText
- Accident

Pick Role Variables

Y: Accident

Weight: optional numeric

Freq: optional numeric

By: optional

Personality: Standard Least Squares

Emphasis: Effect Leverage

□ Keep dialog open

Construct Model Effects

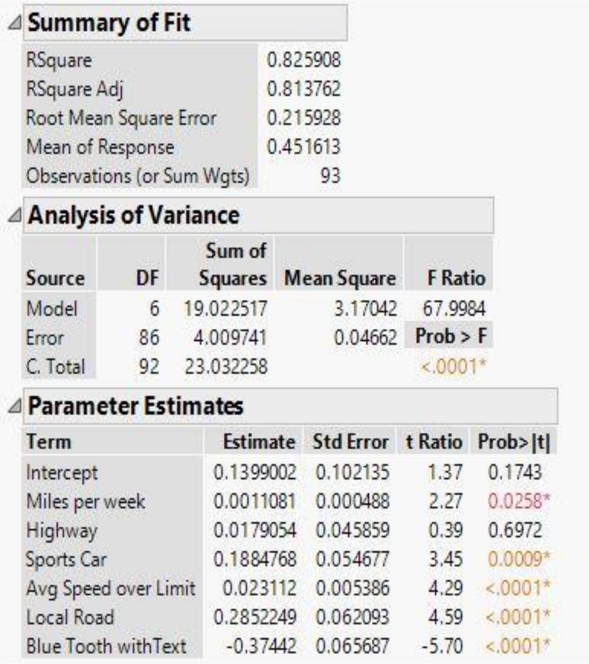
- Miles per week
- Highway
- Sports Car
- Avg Speed over Limit
- Local Road
- Blue Tooth withText

Degree: 2

□ No Intercept

Fig 2.3.1. Target and Driver Variables

Following Screenshot shows the output of the Regression Analysis:



Summary of Fit

RSquare	0.825908
RSquare Adj	0.813762
Root Mean Square Error	0.215928
Mean of Response	0.451613
Observations (or Sum Wgts)	93

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	19.022517	3.17042	67.9984
Error	86	4.009741	0.04662	Prob > F
C. Total	92	23.032258		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.1399002	0.102135	1.37	0.1743
Miles per week	0.0011081	0.000488	2.27	0.0258*
Highway	0.0179054	0.045859	0.39	0.6972
Sports Car	0.1884768	0.054677	3.45	0.0009*
Avg Speed over Limit	0.023112	0.005386	4.29	<.0001*
Local Road	0.2852249	0.062093	4.59	<.0001*
Blue Tooth withText	-0.37442	0.065687	-5.70	<.0001*

Fig 2.3.2. Output after applying Regression Analysis

- The R-Square values measures 82% of the variation in the independent variable Accident. Hence the model created is a perfected model.
- The T-Statistics for the variable Highway is 0.39 which suggest us that it is a result of an accidental outcome.

2.3.1. Future Prediction:

The following image shows the changes in the variable 'Accident' based on the values set for the driver variables Y-axis.

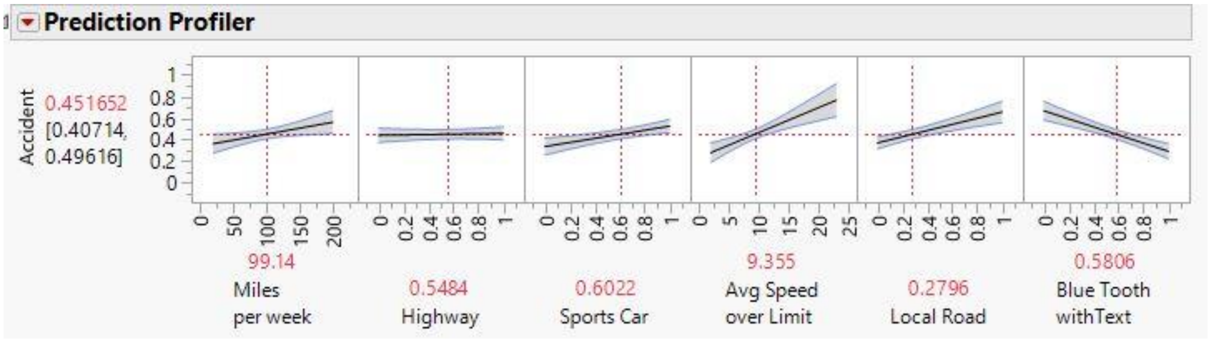


Fig 2.3.1.1. Results generated using Prediction Profiler

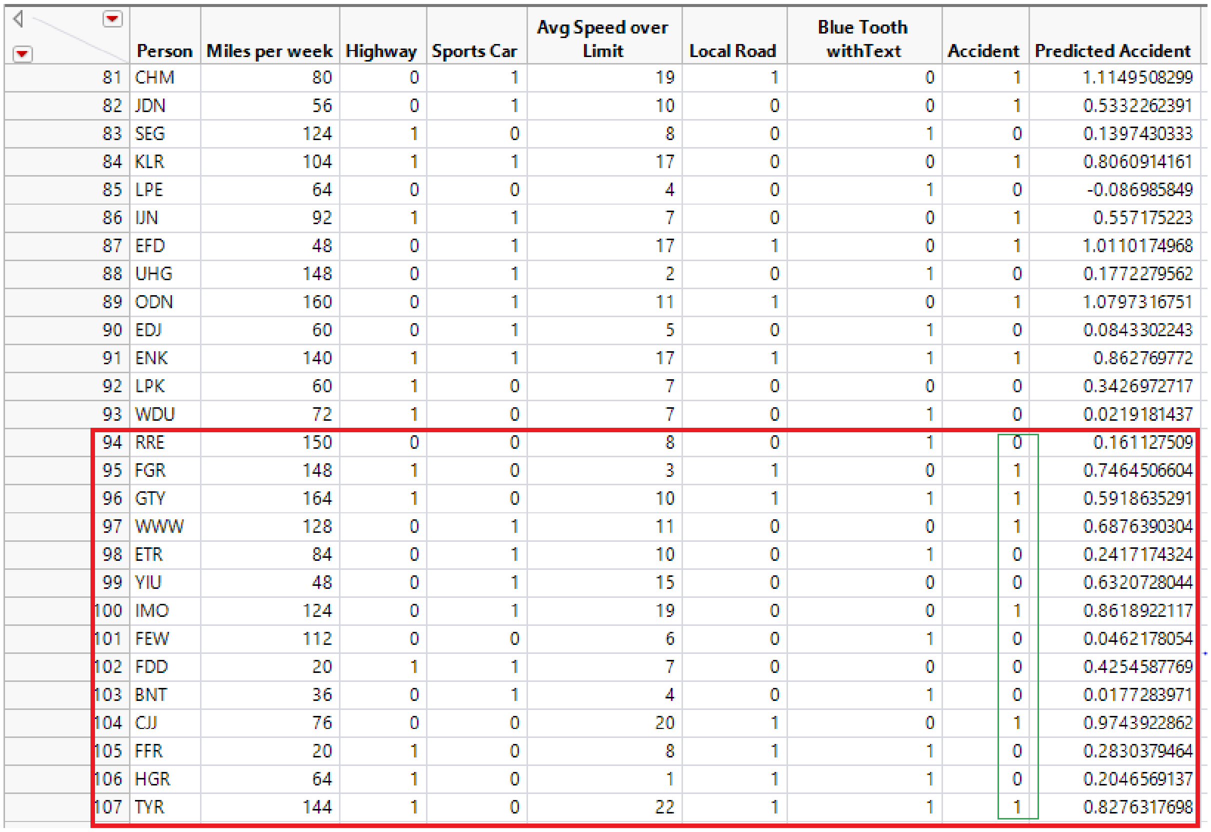
- The dataset with the major independent variables are trained and concluded that the variables Miles per Week, Sports Car, Local Road directly affect the dependent variable Accident. The explanatory variable Average Speed over Limit has the maximum effect on the target variable.
- From the profiler we can conclude that the variable Bluetooth with Text inversely effects the target variable accident no matter the variance.

Elimination of variable Highway:

The variance in the values of the variable ‘highway’ has no significant impact on the variance of target variable ‘accident’. Hence this variable can also be excluded in the analysis.

2.3.2. Output

The screenshot below shows the predicted accident as a separate column and the formula is applied to find out the chance of accident.

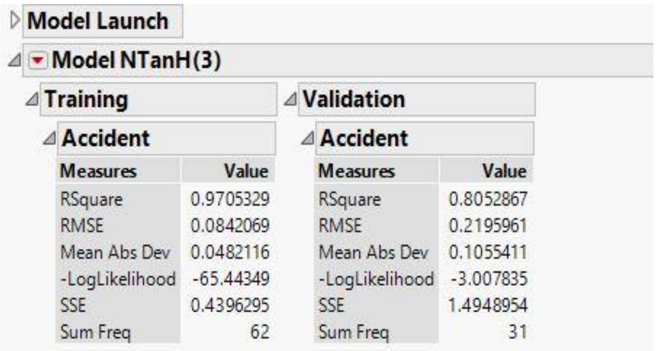


	Person	Miles per week	Highway	Sports Car	Avg Speed over Limit	Local Road	Blue Tooth withText	Accident	Predicted Accident
81	CHM	80	0	1	19	1	0	1	1.1149508299
82	JDN	56	0	1	10	0	0	1	0.5332262391
83	SEG	124	1	0	8	0	1	0	0.1397430333
84	KLR	104	1	1	17	0	0	1	0.8060914161
85	LPE	64	0	0	4	0	1	0	-0.086985849
86	IJN	92	1	1	7	0	0	1	0.557175223
87	EFD	48	0	1	17	1	0	1	1.0110174968
88	UHG	148	0	1	2	0	1	0	0.1772279562
89	ODN	160	0	1	11	1	0	1	1.0797316751
90	EDJ	60	0	1	5	0	1	0	0.0843302243
91	ENK	140	1	1	17	1	1	1	0.862769772
92	LPK	60	1	0	7	0	0	0	0.3426972717
93	WDU	72	1	0	7	0	1	0	0.0219181437
94	RRE	150	0	0	8	0	1	0	0.161127509
95	FGR	148	1	0	3	1	0	1	0.7464506604
96	GTY	164	1	0	10	1	1	1	0.5918635291
97	WWW	128	0	1	11	0	0	1	0.6876390304
98	ETR	84	0	1	10	0	1	0	0.2417174324
99	YIU	48	0	1	15	0	0	0	0.6320728044
100	IMO	124	0	1	19	0	0	1	0.8618922117
101	FEW	112	0	0	6	0	1	0	0.0462178054
102	FDD	20	1	1	7	0	0	0	0.4254587769
103	BNT	36	0	1	4	0	1	0	0.0177283971
104	CJJ	76	0	0	20	1	0	1	0.9743922862
105	FFR	20	1	0	8	1	1	0	0.2830379464
106	HGR	64	1	0	1	1	1	0	0.2046569137
107	TYR	144	1	0	22	1	1	1	0.8276317698

Fig 2.3.2.1. Predicted Accident values after adding observations

2.4. Neural Network Analysis

On analyzing the data file, we get the following results:



Model Launch

Model NTanH(3)

Training

Accident

Measures	Value
RSquare	0.9705329
RMSE	0.0842069
Mean Abs Dev	0.0482116
-LogLikelihood	-65.44349
SSE	0.4396295
Sum Freq	62

Validation

Accident

Measures	Value
RSquare	0.8052867
RMSE	0.2195961
Mean Abs Dev	0.1055411
-LogLikelihood	-3.007835
SSE	1.4948954
Sum Freq	31

Fig 2.4.1. RSquare using Neural Network

The value of RSquare is 97%, from which it can be concluded that the model is perfect model.

2.4.1. Future Prediction:

The predicted model for Accident was built using neural network profiler and can be found below in the attached screenshot.

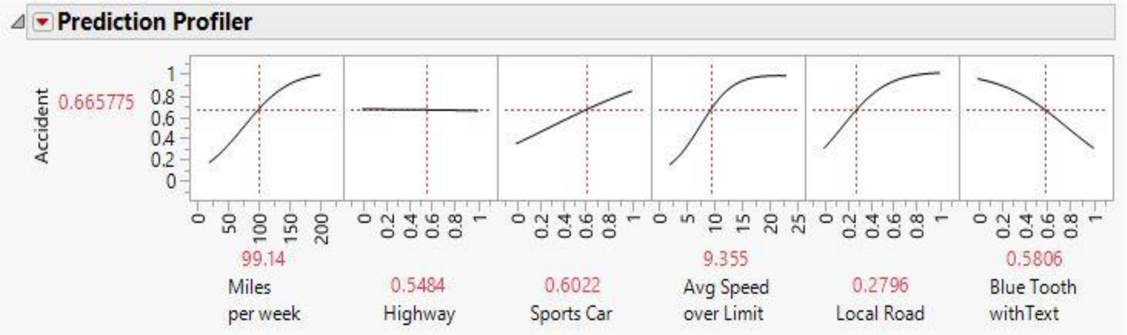


Fig 2.4.1. Graph using Prediction Profiler

- The explanatory variables Miles per Week, Sports Car, Avg Speed over Limit, Local Road and Blue Tooth with Text has a major impact on target variable accident.
- The variable Highway shows no significant impact on the variable accident and could also be not considered for analysis.
- It can also be concluded that with steady increase in Sports Car there is also a steady increase in Accident.

2.4.2. Neural Network Diagram

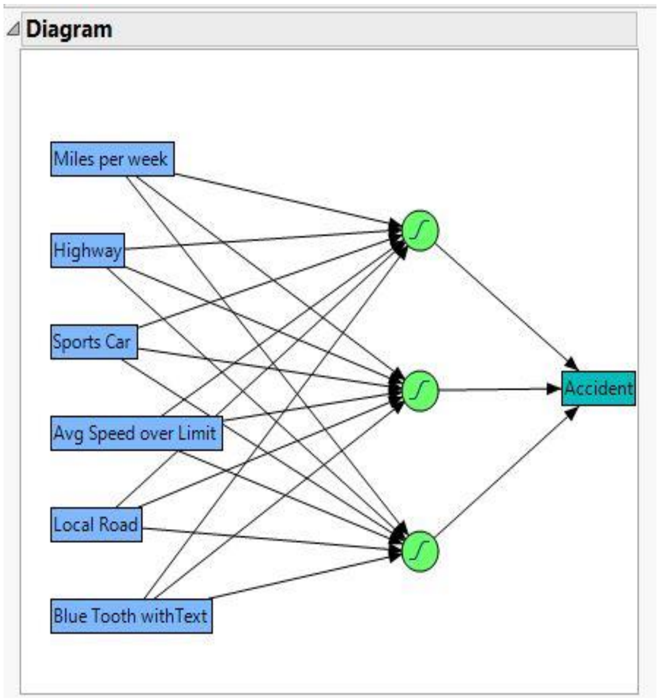
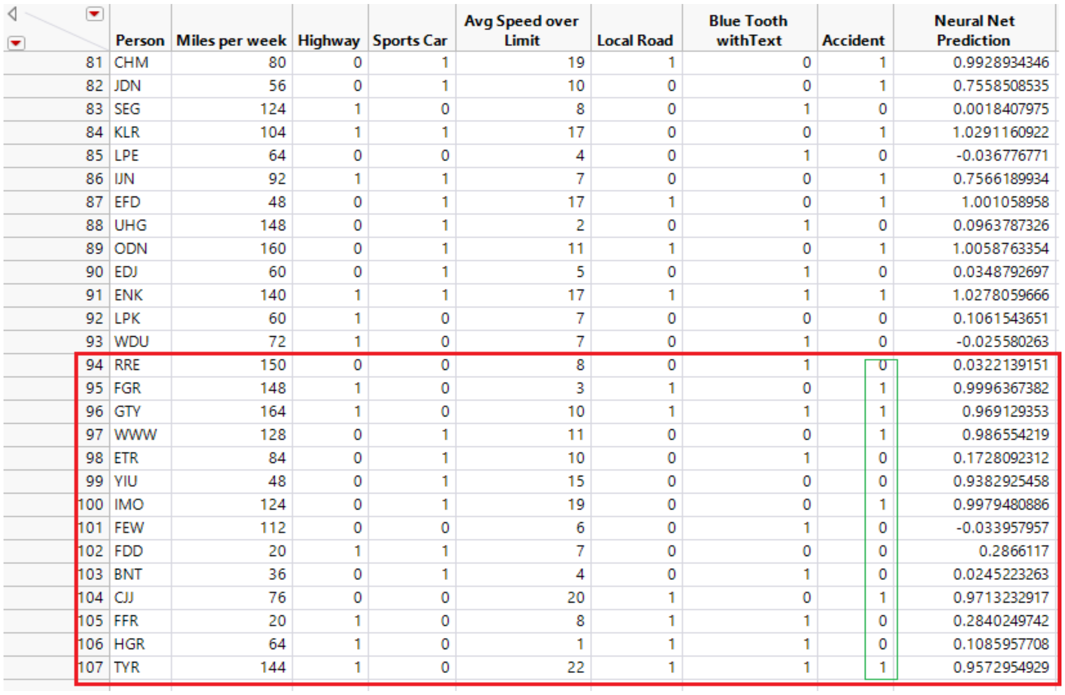


Fig 2.4.1. Diagram using Neural Network

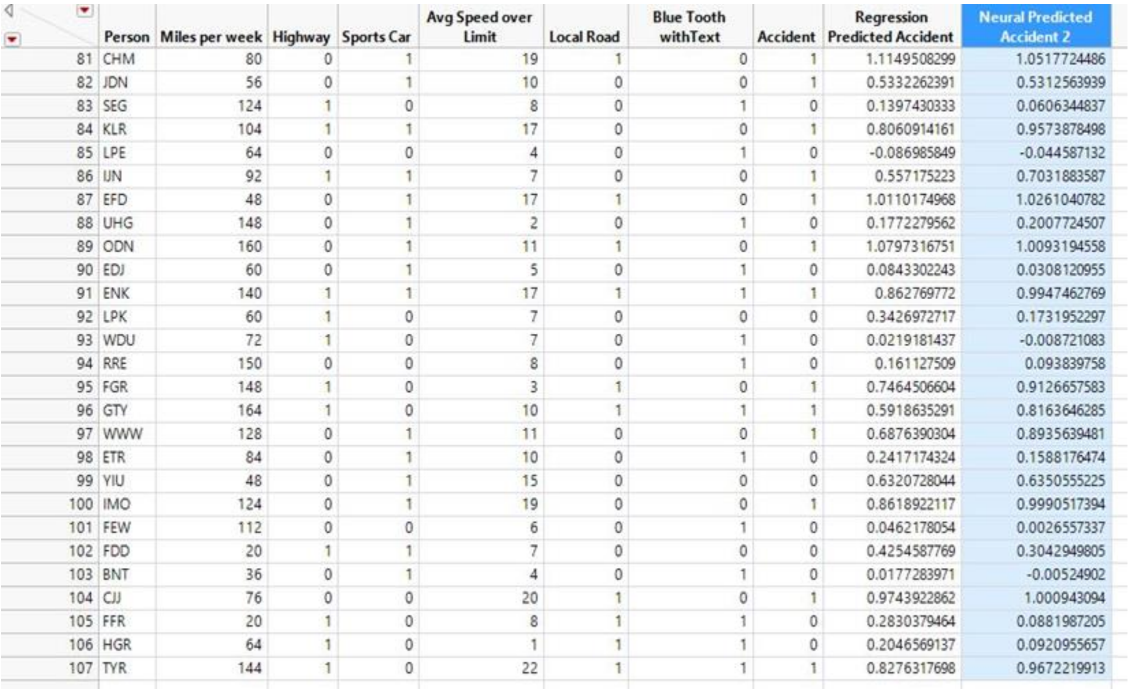
2.4.3 Predicted value using Neural Network Analysis



	Person	Miles per week	Highway	Sports Car	Avg Speed over Limit	Local Road	Blue Tooth withText	Accident	Neural Net Prediction
81	CHM	80	0	1	19	1	0	1	0.9928934346
82	JDN	56	0	1	10	0	0	1	0.7558508535
83	SEG	124	1	0	8	0	1	0	0.0018407975
84	KLR	104	1	1	17	0	0	1	1.0291160922
85	LPE	64	0	0	4	0	1	0	-0.036776771
86	IJN	92	1	1	7	0	0	1	0.7566189934
87	EFD	48	0	1	17	1	0	1	1.001058958
88	UHG	148	0	1	2	0	1	0	0.0963787326
89	ODN	160	0	1	11	1	0	1	1.0058763354
90	EDJ	60	0	1	5	0	1	0	0.0348792697
91	ENK	140	1	1	17	1	1	1	1.0278059666
92	LPK	60	1	0	7	0	0	0	0.1061543651
93	WDU	72	1	0	7	0	1	0	-0.025580263
94	RRE	150	0	0	8	0	1	0	0.0322139151
95	FGR	148	1	0	3	1	0	1	0.9996367382
96	GTY	164	1	0	10	1	1	1	0.969129353
97	WWW	128	0	1	11	0	0	1	0.986554219
98	ETR	84	0	1	10	0	1	0	0.1728092312
99	YIU	48	0	1	15	0	0	0	0.9382925458
100	IMO	124	0	1	19	0	0	1	0.9979480886
101	FEW	112	0	0	6	0	1	0	-0.033957957
102	FDD	20	1	1	7	0	0	0	0.2866117
103	BNT	36	0	1	4	0	1	0	0.0245223263
104	CJJ	76	0	0	20	1	0	1	0.9713232917
105	FFR	20	1	0	8	1	1	0	0.2840249742
106	HGR	64	1	0	1	1	1	0	0.1085957708
107	TYR	144	1	0	22	1	1	1	0.9572954929

Fig 2.4.3.1. Predicted Values after adding observations

2.5. Comparison between Regression and Neural Network Analysis



	Person	Miles per week	Highway	Sports Car	Avg Speed over Limit	Local Road	Blue Tooth withText	Accident	Regression Predicted Accident	Neural Predicted Accident 2
81	CHM	80	0	1	19	1	0	1	1.1149508299	1.0517724486
82	JDN	56	0	1	10	0	0	1	0.5332262391	0.5312563939
83	SEG	124	1	0	8	0	1	0	0.1397430333	0.0606344837
84	KLR	104	1	1	17	0	0	1	0.8060914161	0.9573878498
85	LPE	64	0	0	4	0	1	0	-0.086985849	-0.044587132
86	IJN	92	1	1	7	0	0	1	0.557175223	0.7031883587
87	EFD	48	0	1	17	1	0	1	1.0110174968	1.0261040782
88	UHG	148	0	1	2	0	1	0	0.1772279562	0.2007724507
89	ODN	160	0	1	11	1	0	1	1.0797316751	1.0093194558
90	EDJ	60	0	1	5	0	1	0	0.0843302243	0.0308120955
91	ENK	140	1	1	17	1	1	1	0.862769772	0.9947462769
92	LPK	60	1	0	7	0	0	0	0.3426972717	0.1731952297
93	WDU	72	1	0	7	0	1	0	0.0219181437	-0.008721083
94	RRE	150	0	0	8	0	1	0	0.161127509	0.093839758
95	FGR	148	1	0	3	1	0	1	0.7464506604	0.9126657583
96	GTY	164	1	0	10	1	1	1	0.5918635291	0.8163646285
97	WWW	128	0	1	11	0	0	1	0.6876390304	0.8935639481
98	ETR	84	0	1	10	0	1	0	0.2417174324	0.1588176474
99	YIU	48	0	1	15	0	0	0	0.6320728044	0.6350555225
100	IMO	124	0	1	19	0	0	1	0.8618922117	0.9990517394
101	FEW	112	0	0	6	0	1	0	0.0462178054	0.0026557337
102	FDD	20	1	1	7	0	0	0	0.4254587769	0.3042949805
103	BNT	36	0	1	4	0	1	0	0.0177283971	-0.00524902
104	CJJ	76	0	0	20	1	0	1	0.9743922862	1.000943094
105	FFR	20	1	0	8	1	1	0	0.2830379464	0.0881987205
106	HGR	64	1	0	1	1	1	0	0.2046569137	0.0920955657
107	TYR	144	1	0	22	1	1	1	0.8276317698	0.9672219913

Fig 2.5.1. Comparing the predicted value obtained from Regression and Neural Network respectively

PART 3

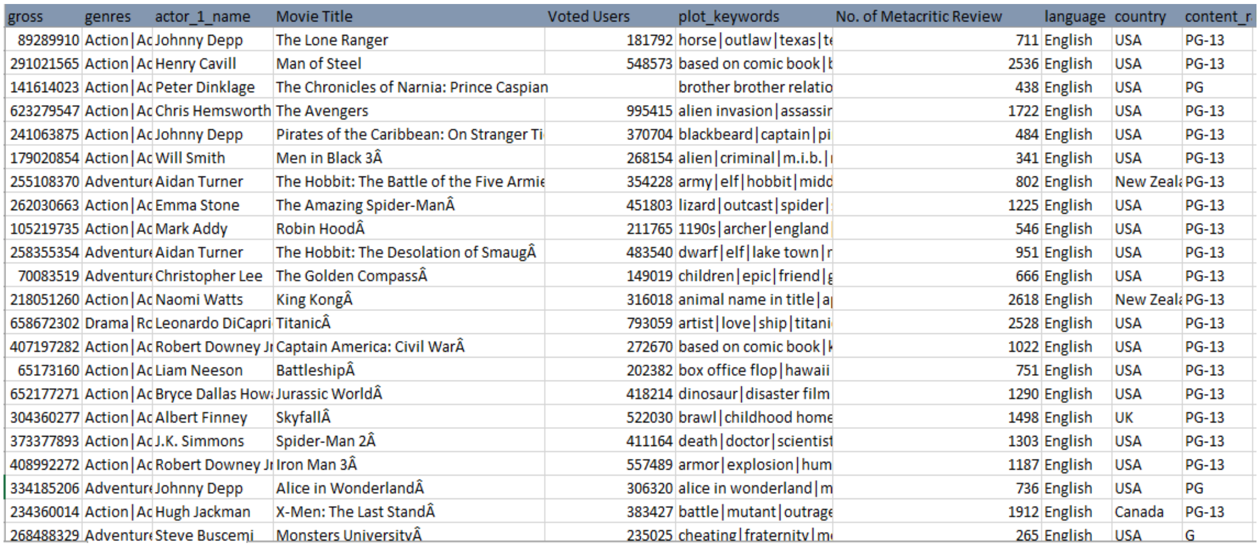
3.1. About the Dataset

The dataset includes the list of directors and the gross earnings from their movies collected from year 1920 to 2016 to figure out how director name, genre and other factors affects the gross of the movie. This data set contains 20 variables and 5044 observations.

3.2. Objective

Our objective is to come up with a business solution which will help increase the gross of a movie. The analysis would be performed after cleaning the dataset. There are 9 variables which can be used to analyze the patterns and trends on the cleaned data and determine how different variable affects the target variable.

3.3. Original unclean Dataset



gross	genres	actor_1_name	Movie Title	Voted Users	plot_keywords	No. of Metacritic Review	language	country	content_r
89289910	Action|Ac	Johnny Depp	The Lone Ranger	181792	horse|outlaw|texas|te	711	English	USA	PG-13
291021565	Action|Ac	Henry Cavill	Man of Steel	548573	based on comic book|b	2536	English	USA	PG-13
141614023	Action|Ac	Peter Dinklage	The Chronicles of Narnia: Prince Caspian		brother brother relatio	438	English	USA	PG
623279547	Action|Ac	Chris Hemsworth	The Avengers	995415	alien invasion|assassir	1722	English	USA	PG-13
241063875	Action|Ac	Johnny Depp	Pirates of the Caribbean: On Stranger Ti	370704	blackbeard|captain|pi	484	English	USA	PG-13
179020854	Action|Ac	Will Smith	Men in Black 3Â	268154	alien|criminal|m.i.b.|	341	English	USA	PG-13
255108370	Adventur	Aidan Turner	The Hobbit: The Battle of the Five Armie	354228	army|elf|hobbit|midd	802	English	New Zeal	PG-13
262030663	Action|Ac	Emma Stone	The Amazing Spider-ManÂ	451803	lizard|outcast|spider|	1225	English	USA	PG-13
105219735	Action|Ac	Mark Addy	Robin HoodÂ	211765	1190s|archer|england	546	English	USA	PG-13
258355354	Adventur	Aidan Turner	The Hobbit: The Desolation of SmaugÂ	483540	dwarf|elf|lake town|r	951	English	USA	PG-13
70083519	Adventur	Christopher Lee	The Golden CompassÂ	149019	children|epic|friend|g	666	English	USA	PG-13
218051260	Action|Ac	Naomi Watts	King KongÂ	316018	animal name in title|a	2618	English	New Zeal	PG-13
658672302	Drama|Rc	Leonardo DiCapri	TitanicÂ	793059	artist|love|ship|titani	2528	English	USA	PG-13
407197282	Action|Ac	Robert Downey Jr	Captain America: Civil WarÂ	272670	based on comic book|k	1022	English	USA	PG-13
65173160	Action|Ac	Liam Neeson	BattleshipÂ	202382	box office flop|hawaii	751	English	USA	PG-13
652177271	Action|Ac	Bryce Dallas How	Jurassic WorldÂ	418214	dinosaur|disaster film	1290	English	USA	PG-13
304360277	Action|Ac	Albert Finney	SkyfallÂ	522030	brawl|childhood home	1498	English	UK	PG-13
373377893	Action|Ac	J.K. Simmons	Spider-Man 2Â	411164	death|doctor|scientist	1303	English	USA	PG-13
408992272	Action|Ac	Robert Downey Jr	Iron Man 3Â	557489	armor|explosion|hum	1187	English	USA	PG-13
334185206	Adventur	Johnny Depp	Alice in WonderlandÂ	306320	alice in wonderland|m	736	English	USA	PG
234360014	Action|Ac	Hugh Jackman	X-Men: The Last StandÂ	383427	battle|mutant|outrage	1912	English	Canada	PG-13
268488329	Adventur	Steve Buscemi	Monsters UniversityÂ	235025	cheating|fraternity|m	265	English	USA	G

Fig 3.3.1. Original Dataset

3.4. Identifying Target and Driver Variables

Variable	Type	Description
Director Name	Driver	Director of the movie
Duration	Driver	The duration of the movie
Gross	Target Variable	The gross earning of the movies after release

Genre	Driver	The type of the movie based on the plot
Movie Title	Not Required	The name of the movie
Voted User	Driver	The number of users voted for the movie
No. of Metacritic Review	Driver	The number of review given by the Metacritic.
Year	Driver	Year the movie was released
IMDB Score	Driver	The rating given by the users in IMDB website scaling from 1 to 10
Facebook Likes	Driver	The number of likes the movie received in the Facebook
Actor Name	Not Required	The name of the lead actor in the movie
Face number in Poster	Not Required	No. of actor's faces in the poster of the movie
Plot keywords	Not Required	The storyline of the movie
IMDB link	Not Required	The link to the movie details in IMDB's website
Language	Not Required	The language in which the movie is made
Country	Not Required	The country to which the movie belongs
Budget	Not Required	The budget of the movie
Director Facebook like	Not Required	The number of likes the director received on Facebook for the movie
Actor Facebook like	Not Required	The number of likes the actor received for the movie on Facebook.
Cast Facebook like	Driver	The number of likes the entire cast received for the movie on Facebook.

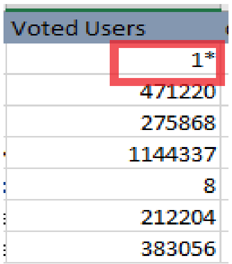
3.5. Troubleshoot the data

3.5.1. Clean the data

The dataset is browsed and checked for dirty data, if any. Rows containing such data is deleted from the dataset to avoid errors in the analysis results.

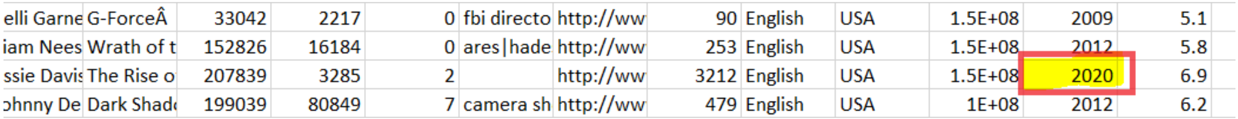
The following rows has been deleted since it had dirty data,

- Voted user has value 1*



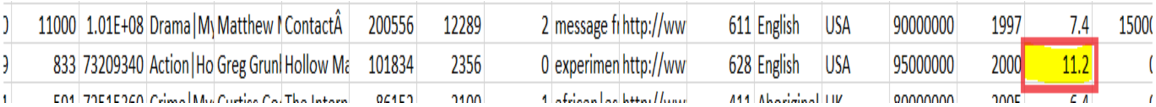
Voted Users
1*
471220
275868
1144337
8
212204
383056

- Year mentioned is 2020 which is future



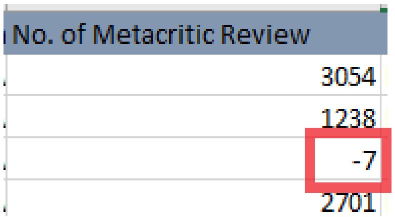
elli Garne	G-ForceÂ	33042	2217	0	fbi directo	http://ww	90	English	USA	1.5E+08	2009	5.1
iam Nees	Wrath of t	152826	16184	0	ares|hade	http://ww	253	English	USA	1.5E+08	2012	5.8
ssie Davis	The Rise o	207839	3285	2		http://ww	3212	English	USA	1.5E+08	2020	6.9
ɔhnny De	Dark Shad	199039	80849	7	camera sh	http://ww	479	English	USA	1E+08	2012	6.2

- IMDB Score is 11.2 which is greater than 10



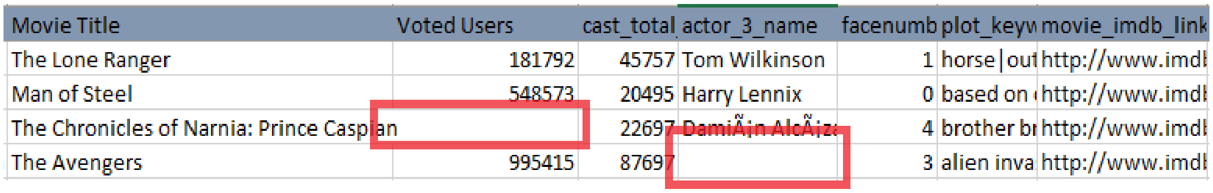
11000	1.01E+08	Drama|My	Matthew I	ContactÂ	200556	12289	2	message fi	http://ww	611	English	USA	90000000	1997	7.4	1500(
833	73209340	Action|Ho	Greg Grunl	Hollow Ma	101834	2356	0	experimen	http://ww	628	English	USA	95000000	2000	11.2	(

- No. of Metacritic Review is -7 which is again irrelevant



No. of Metacritic Review
3054
1238
-7
2701

- Remove the rows which has empty data in any column.



Movie Title	Voted Users	cast_total	actor_3_name	facenumb	plot_keyw	movie_imdb_link
The Lone Ranger	181792	45757	Tom Wilkinson	1	horse|out	http://www.imdl
Man of Steel	548573	20495	Harry Lennix	0	based on	http://www.imdl
The Chronicles of Narnia: Prince Caspian		22697	DamiÃ¡n AlcÃ¡z	4	brother bı	http://www.imdl
The Avengers	995415	87697		3	alien inva	http://www.imdl

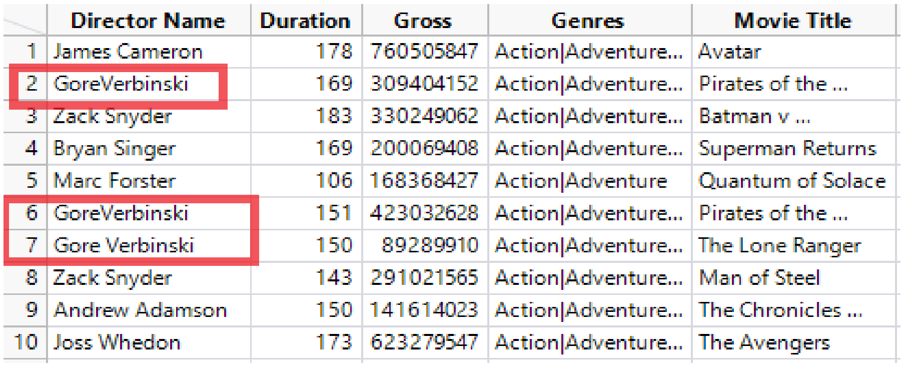
Also, delete the irrelevant columns which has no effect on the target variable.

The following columns are deleted:

- Actor name
- Movie Title
- Face number in poster
- Plot key words
- IMDB link
- Language
- Country
- Budget
- Director FB likes
- Actor FB likes

3.5.2. Format the data

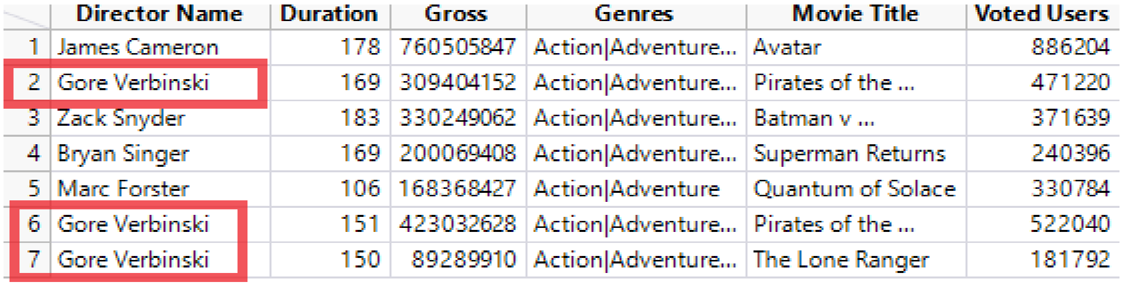
In the dataset the column 'Director name' includes entries with no space between the first name and the last name. Therefore, we need to format the data to have consistency in the value.



	Director Name	Duration	Gross	Genres	Movie Title
1	James Cameron	178	760505847	Action|Adventure...	Avatar
2	GoreVerbinski	169	309404152	Action|Adventure...	Pirates of the ...
3	Zack Snyder	183	330249062	Action|Adventure...	Batman v ...
4	Bryan Singer	169	200069408	Action|Adventure...	Superman Returns
5	Marc Forster	106	168368427	Action|Adventure	Quantum of Solace
6	GoreVerbinski	151	423032628	Action|Adventure...	Pirates of the ...
7	Gore Verbinski	150	89289910	Action|Adventure...	The Lone Ranger
8	Zack Snyder	143	291021565	Action|Adventure...	Man of Steel
9	Andrew Adamson	150	141614023	Action|Adventure...	The Chronicles ...
10	Joss Whedon	173	623279547	Action|Adventure...	The Avengers

Fig 3.5.2.1. The Director Name should be changed from 'GoreVerbinski' to 'Gore Verbinski'

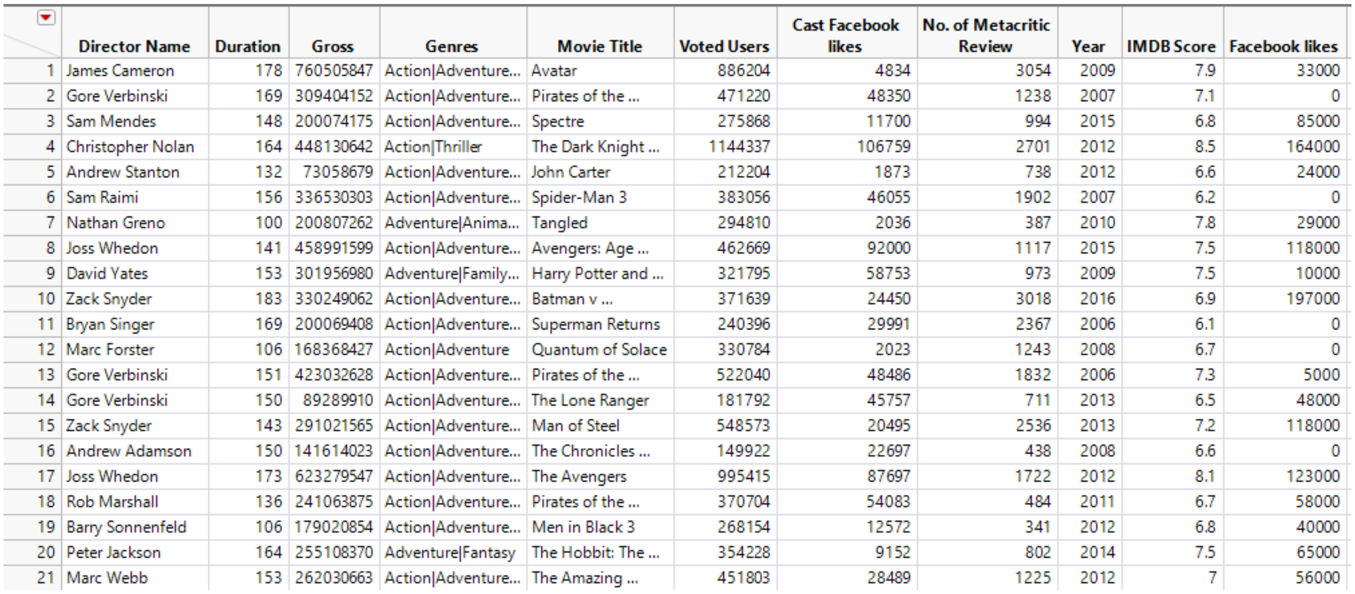
Here, we recoded such value as follows.



	Director Name	Duration	Gross	Genres	Movie Title	Voted Users
1	James Cameron	178	760505847	Action|Adventure...	Avatar	886204
2	Gore Verbinski	169	309404152	Action|Adventure...	Pirates of the ...	471220
3	Zack Snyder	183	330249062	Action|Adventure...	Batman v ...	371639
4	Bryan Singer	169	200069408	Action|Adventure...	Superman Returns	240396
5	Marc Forster	106	168368427	Action|Adventure	Quantum of Solace	330784
6	Gore Verbinski	151	423032628	Action|Adventure...	Pirates of the ...	522040
7	Gore Verbinski	150	89289910	Action|Adventure...	The Lone Ranger	181792

Fig 3.5.2.2. After recoding the data

Hence, the cleaned data is as follows:



	Director Name	Duration	Gross	Genres	Movie Title	Voted Users	Cast Facebook likes	No. of Metacritic Review	Year	IMDB Score	Facebook likes
1	James Cameron	178	760505847	Action|Adventure...	Avatar	886204	4834	3054	2009	7.9	33000
2	Gore Verbinski	169	309404152	Action|Adventure...	Pirates of the ...	471220	48350	1238	2007	7.1	0
3	Sam Mendes	148	200074175	Action|Adventure...	Spectre	275868	11700	994	2015	6.8	85000
4	Christopher Nolan	164	448130642	Action|Thriller	The Dark Knight ...	1144337	106759	2701	2012	8.5	164000
5	Andrew Stanton	132	73058679	Action|Adventure...	John Carter	212204	1873	738	2012	6.6	24000
6	Sam Raimi	156	336530303	Action|Adventure...	Spider-Man 3	383056	46055	1902	2007	6.2	0
7	Nathan Greno	100	200807262	Adventure|Anima...	Tangled	294810	2036	387	2010	7.8	29000
8	Joss Whedon	141	458991599	Action|Adventure...	Avengers: Age ...	462669	92000	1117	2015	7.5	118000
9	David Yates	153	301956980	Adventure|Family...	Harry Potter and ...	321795	58753	973	2009	7.5	10000
10	Zack Snyder	183	330249062	Action|Adventure...	Batman v ...	371639	24450	3018	2016	6.9	197000
11	Bryan Singer	169	200069408	Action|Adventure...	Superman Returns	240396	29991	2367	2006	6.1	0
12	Marc Forster	106	168368427	Action|Adventure	Quantum of Solace	330784	2023	1243	2008	6.7	0
13	Gore Verbinski	151	423032628	Action|Adventure...	Pirates of the ...	522040	48486	1832	2006	7.3	5000
14	Gore Verbinski	150	89289910	Action|Adventure...	The Lone Ranger	181792	45757	711	2013	6.5	48000
15	Zack Snyder	143	291021565	Action|Adventure...	Man of Steel	548573	20495	2536	2013	7.2	118000
16	Andrew Adamson	150	141614023	Action|Adventure...	The Chronicles ...	149922	22697	438	2008	6.6	0
17	Joss Whedon	173	623279547	Action|Adventure...	The Avengers	995415	87697	1722	2012	8.1	123000
18	Rob Marshall	136	241063875	Action|Adventure...	Pirates of the ...	370704	54083	484	2011	6.7	58000
19	Barry Sonnenfeld	106	179020854	Action|Adventure...	Men in Black 3	268154	12572	341	2012	6.8	40000
20	Peter Jackson	164	255108370	Adventure|Fantasy	The Hobbit: The ...	354228	9152	802	2014	7.5	65000
21	Marc Webb	153	262030663	Action|Adventure...	The Amazing ...	451803	28489	1225	2012	7	56000

Fig 3.5.2.3. The Dataset after cleansing that can be used for analysis

3.6. Degrees of Freedom

1. Before Cleaning Data

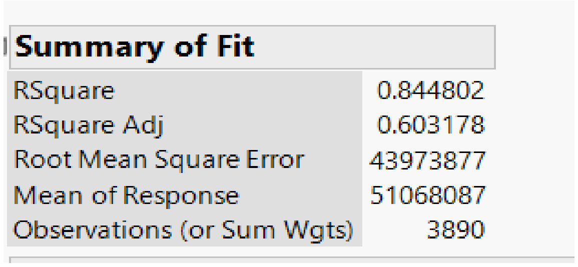
   Number of columns - 20
   Number of Rows - 5044
   Target Variable - Gross
   Number of Target Variable - 1
   Number of Explanatory variable - 19
   Degree of Freedom= Number of Rows – Number of Explanatory variables
   = 5044 -19
   = 5025

2. After Data cleaning

   Number of columns – 11
   Number of Rows - 3890
   Target Variable - Gross
   Number of Target Variable - 1
   Number of Explanatory variable - 9
   Degree of Freedom= Number of Rows – Number of Explanatory variables
   = 3890 - 9
   = 3881

3.7. Data Analysis

3.7.1. Analysis using Regression

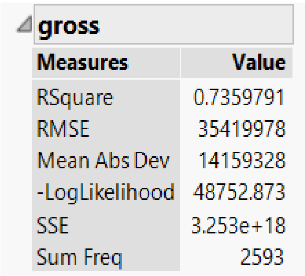


Summary of Fit	
RSquare	0.844802
RSquare Adj	0.603178
Root Mean Square Error	43973877
Mean of Response	51068087
Observations (or Sum Wgts)	3890

Fig 3.7.1.1. RSquare obtained from Regression Analysis

Here, the R^2 is 84% which shows that the model is nearly perfect.

3.7.2. Analysis using Neural Nets



gross

Measures	Value
RSquare	0.7359791
RMSE	35419978
Mean Abs Dev	14159328
-LogLikelihood	48752.873
SSE	3.253e+18
Sum Freq	2593

Fig 3.7.2.1. RSquare obtained from Neural Network

Here, the R^2 is 73% which is a decent value for a model.

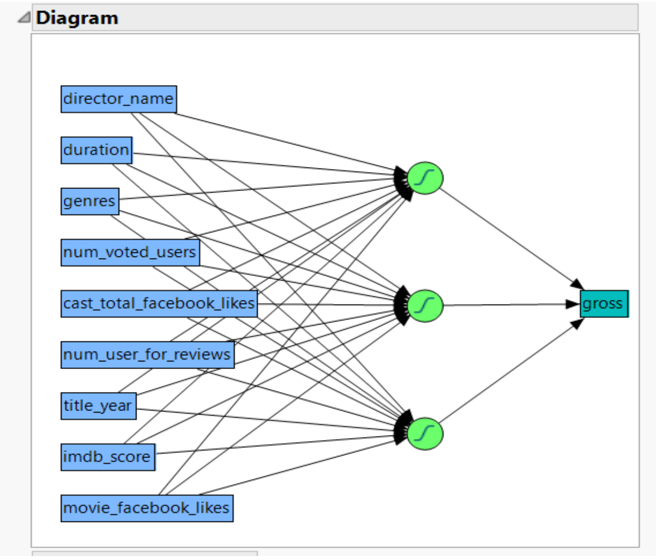


Fig 3.7.2.2. Diagram using Neural Network

Prediction Profiler:

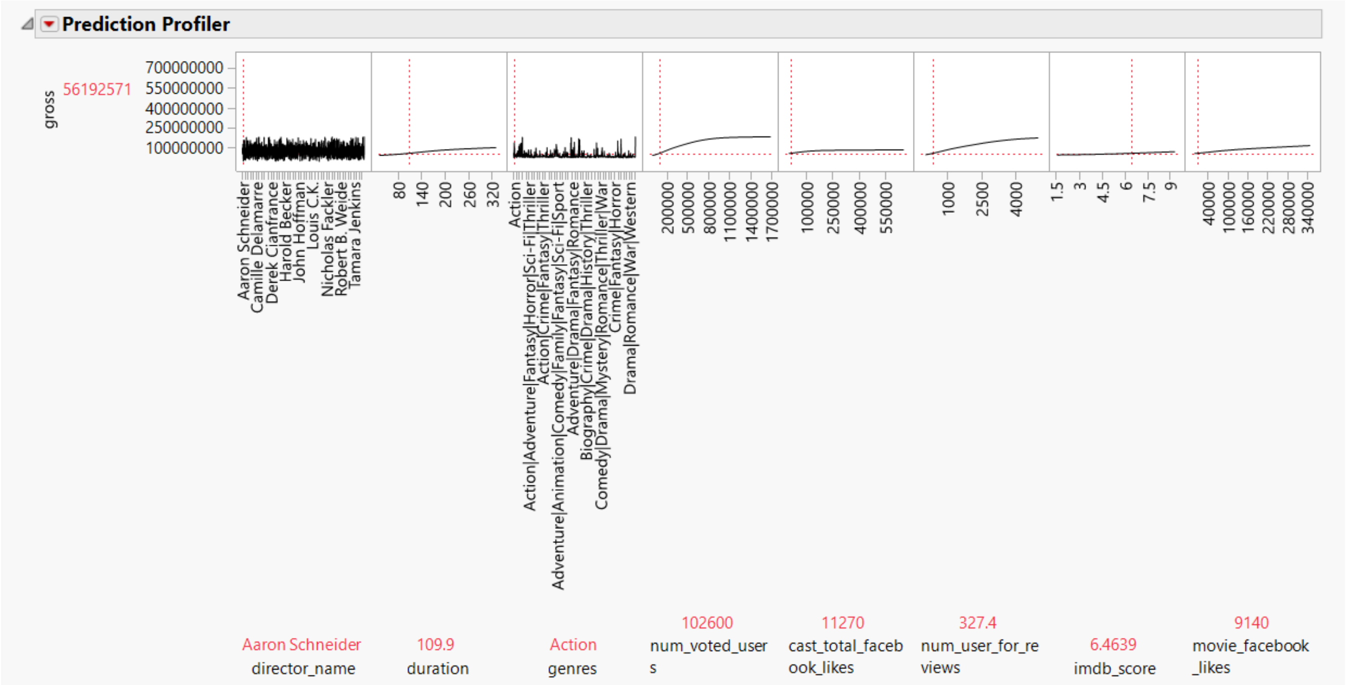


Fig 3.7.2.3. Graph obtained using Prediction Profiler

The profiler provides several highly interactive cross-sectional views of any response surface. This will help find good factor settings and produce desirable outcomes. Therefore, if the target is to achieve $50million+ gross, the values on the X-axis can be changed till the value of gross on the Y-axis does not reach the desired value.

Here, in the above image, it can be stated that the movie will gross $55million+ if the values on the Y-axis are maintained as obtained above.

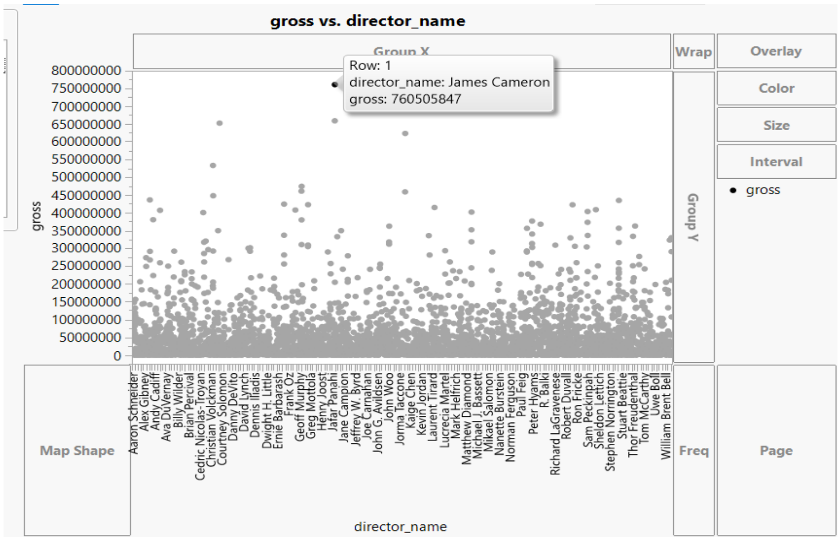
3.7. Analysis of the Output

The analysis shows that

- The 6 factors viz., Director Name, Duration, Genres, Voted Users, No. of Metacritic Review, Facebook Likes, highly impact the target variable ‘Gross’.
- Other variables can affect the analysis, but their impact is almost negligible. These include IMDB Score, Year, Cast Facebook likes with T ratio between -2 and 2. Therefore, it can be eliminated from the analysis.
- Dirty data was cleaned as it was deceiving the output of the analysis.

3.8. Managerial Suggestions after analyzing results

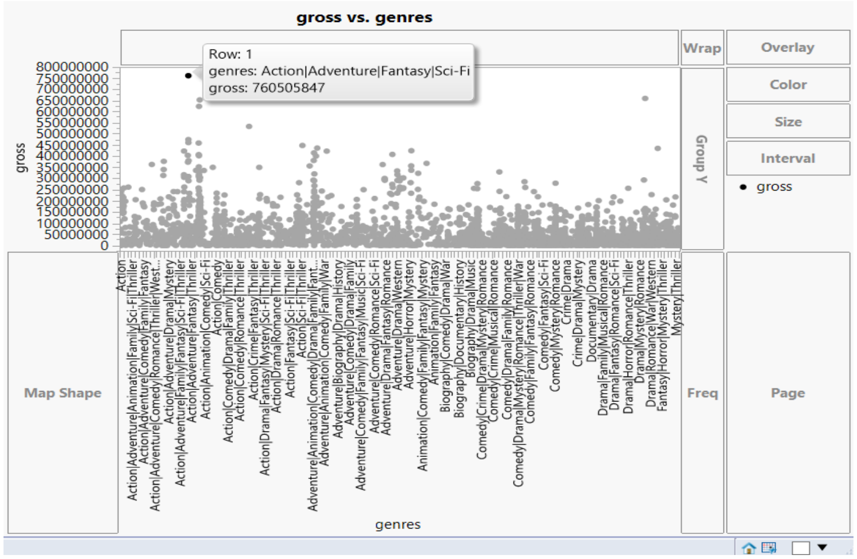
- The movies directed by James Cameron gross more profit than any other movies.



Row: 1
director_name: James Cameron
gross: 760505847

Fig 3.8.1. Gross Vs Director Name showing James Cameron as the top most director with the highest gross

- Genres like Action, Adventure, Fantasy and Sci-fi are trending more.



Row: 1
genres: Action|Adventure|Fantasy|Sci-Fi
gross: 760505847

Fig 3.8.2. Gross Vs Genres showing Action/Adventure/Fantasy/Sci-fi with the highest gross

- Also, it doesn't depend on the Facebook likes, IMDB score or the year in which the movie was released.
- The focus should be on the genre, director, duration and the votes received by the audiences to achieve the desirable gross profit.

3.9. Conclusion

- The impact of the driver variable on the target variable was analyzed successfully using JMP tool.
- We were able to find out the factors that can be used to achieve our objective of increasing the gross of a movie.
- Future Scope would be to mine the data further to dive deep into the specific correlation these variables might have.

References

Dataset has been picked from Kaggle.com