MGMT 635

DATA MINING AND ANALYSIS FOR MANAGERS

PROJECT REPORT
FALL 2018

Submitted to:

Dr. Stephan P. Kudyba

Submitted by:

Snehal Shukla (sds55)
Akanksha Shirsath (as3447)
Sushmitha Sudarsan (ss3598)
Pushkar Gadgil (pg395)
Sanket Pednekar (sp2363)

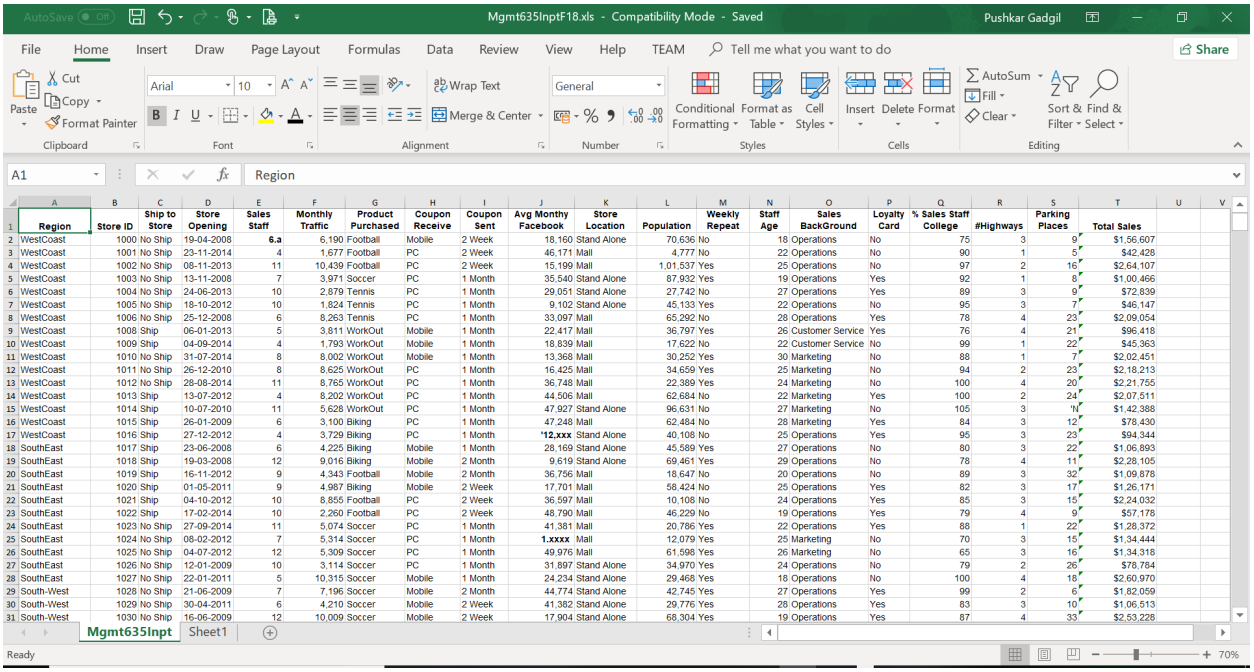
Report Part – 1

1. Business Problem

- We have a dataset which provides the explanatory and target variables to analyze various retail branches a of sporting goods chain around the US.
- From the data available, we are trying to attempt to gain greater insights as to what potentially drives sales for stores. We are also going to transform this file, so it can eventually be mined easily.
- Once we clean and analyse the data, it can be easy to mine the data.

2. Cleaning the dataset

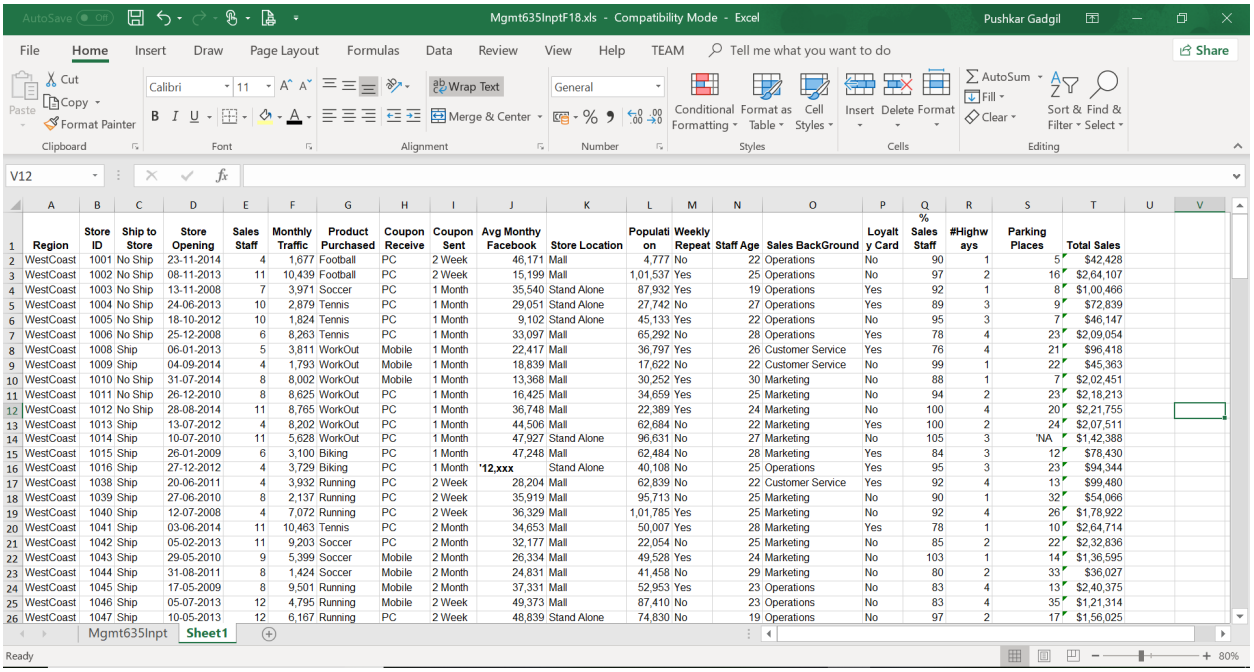
- The screenshot below shows the given original dataset. Refer Figure 1 below.



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
WestCoast	1000	No Ship	19-04-2008	6.a	6,190	Football	Mobile	2 Week	18,160	Stand Alone	70,636	No	18	Operations	No	75	3	9	$1,56,607
WestCoast	1001	No Ship	23-11-2014	4	1,677	Football	PC	2 Week	46,171	Mall	4,777	No	22	Operations	No	90	1	5	$42,428
WestCoast	1002	No Ship	08-11-2013	11	10,439	Football	PC	2 Week	15,199	Mall	1,01,537	Yes	25	Operations	No	97	2	16	$2,64,107
WestCoast	1003	No Ship	13-11-2008	7	3,971	Soccer	PC	1 Month	35,540	Stand Alone	87,932	Yes	19	Operations	Yes	92	1	8	$1,00,466
WestCoast	1004	No Ship	24-06-2013	10	2,879	Tennis	PC	1 Month	29,051	Stand Alone	27,742	No	27	Operations	Yes	89	3	9	$72,839
WestCoast	1005	No Ship	18-10-2012	10	1,824	Tennis	PC	1 Month	9,102	Stand Alone	45,133	Yes	22	Operations	No	95	3	7	$46,147
WestCoast	1006	No Ship	25-12-2008	6	8,263	Tennis	PC	1 Month	33,097	Mall	65,292	No	28	Operations	Yes	78	4	23	$2,09,054
WestCoast	1008	Ship	06-01-2013	5	3,811	WorkOut	Mobile	1 Month	22,417	Mall	36,797	Yes	26	Customer Service	Yes	76	4	21	$96,418
WestCoast	1009	Ship	04-09-2014	4	1,793	WorkOut	Mobile	1 Month	18,839	Mall	17,622	No	22	Customer Service	No	99	1	22	$45,363
WestCoast	1010	No Ship	31-07-2014	8	8,002	WorkOut	Mobile	1 Month	13,368	Mall	30,252	Yes	30	Marketing	No	88	1	7	$2,02,451
WestCoast	1011	No Ship	26-12-2010	8	8,625	WorkOut	PC	1 Month	16,425	Mall	34,659	Yes	25	Marketing	No	94	2	23	$2,18,213
WestCoast	1012	No Ship	28-08-2014	11	8,765	WorkOut	PC	1 Month	36,748	Mall	22,389	Yes	24	Marketing	No	100	4	20	$2,21,755
WestCoast	1013	Ship	13-07-2012	4	8,202	WorkOut	PC	1 Month	44,506	Mall	62,684	No	22	Marketing	Yes	100	2	24	$2,07,511
WestCoast	1014	Ship	10-07-2010	11	5,628	WorkOut	PC	1 Month	47,927	Stand Alone	96,631	No	27	Marketing	No	105	3	N	$1,42,388
WestCoast	1015	Ship	26-01-2009	6	3,100	Biking	PC	1 Month	47,248	Mall	62,484	No	28	Marketing	Yes	84	3	12	$78,430
WestCoast	1016	Ship	27-12-2012	4	3,729	Biking	PC	1 Month	'12,xxx	Stand Alone	40,108	No	25	Operations	Yes	95	3	23	$94,344
SouthEast	1017	Ship	23-06-2008	6	4,225	Biking	Mobile	1 Month	28,169	Stand Alone	45,589	Yes	27	Operations	No	80	3	22	$1,06,893
SouthEast	1018	Ship	19-03-2008	12	9,016	Biking	Mobile	2 Month	9,619	Stand Alone	69,461	Yes	29	Operations	No	78	4	11	$2,28,105
SouthEast	1019	Ship	16-11-2012	9	4,343	Football	Mobile	2 Month	36,756	Mall	18,647	No	20	Operations	No	89	3	32	$1,09,878
SouthEast	1020	Ship	01-05-2011	9	4,987	Biking	Mobile	2 Week	17,701	Mall	58,424	No	25	Operations	Yes	82	3	17	$1,26,171
SouthEast	1021	Ship	04-10-2012	10	8,855	Football	PC	2 Week	36,597	Mall	10,108	No	24	Operations	Yes	85	3	15	$2,24,032
SouthEast	1022	Ship	17-02-2014	10	2,260	Football	PC	2 Week	48,790	Mall	46,229	No	19	Operations	Yes	79	4	9	$57,178
SouthEast	1023	No Ship	27-09-2014	11	5,074	Soccer	PC	1 Month	41,381	Mall	20,786	Yes	22	Operations	Yes	88	1	22	$1,28,372
SouthEast	1024	No Ship	08-02-2012	7	5,314	Soccer	PC	1 Month	1.xxxx	Mall	12,079	Yes	25	Marketing	No	70	3	15	$1,34,444
SouthEast	1025	No Ship	04-07-2012	12	5,309	Soccer	PC	1 Month	49,976	Mall	61,598	Yes	26	Marketing	No	65	3	16	$1,34,318
SouthEast	1026	No Ship	12-01-2009	10	3,114	Soccer	PC	1 Month	31,897	Stand Alone	34,970	Yes	24	Operations	No	79	2	26	$78,784
SouthEast	1027	No Ship	22-01-2011	5	10,315	Soccer	Mobile	1 Month	24,234	Stand Alone	29,468	Yes	18	Operations	No	100	4	18	$2,60,970
South-West	1028	No Ship	21-06-2009	7	7,196	Soccer	Mobile	2 Month	44,774	Stand Alone	42,745	Yes	27	Operations	Yes	99	2	6	$1,82,059
South-West	1029	No Ship	30-04-2011	6	4,210	Soccer	Mobile	2 Week	41,382	Stand Alone	29,776	Yes	28	Operations	Yes	83	3	10	$1,06,513
South-West	1030	No Ship	16-06-2009	12	10,009	Soccer	Mobile	2 Week	17,904	Stand Alone	68,304	Yes	19	Operations	Yes	87	4	33	$2,53,228

Figure 1

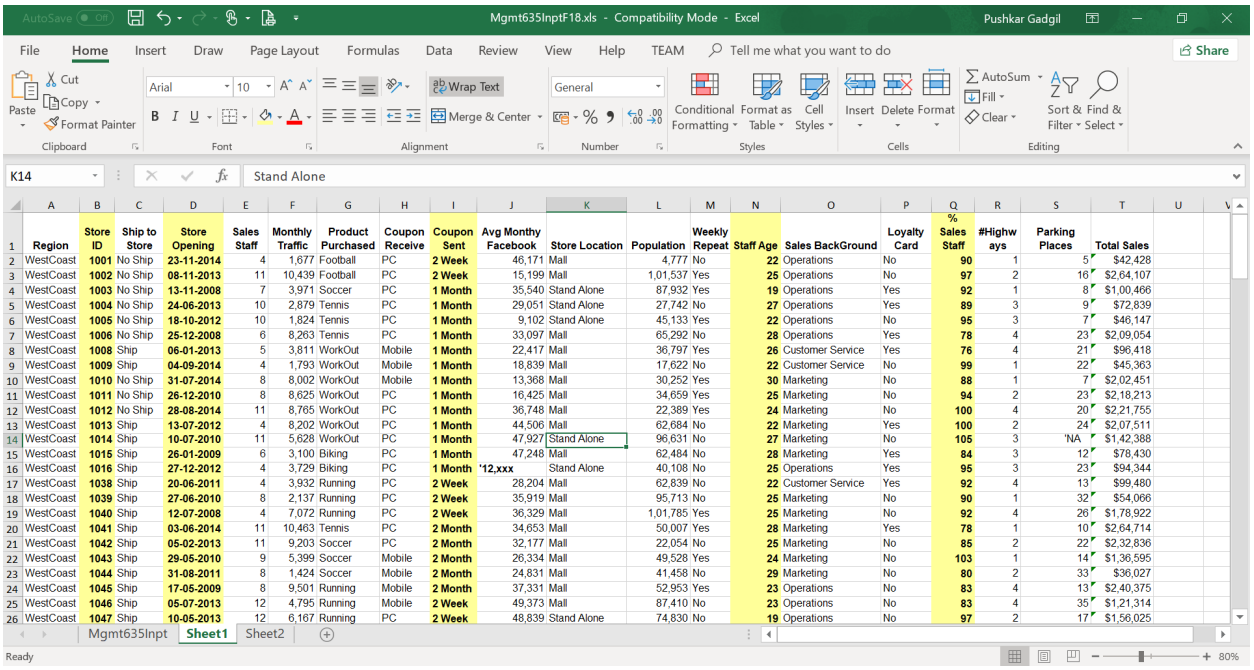
- Now, we clean the given dataset.
- For this, we are first given that we need to focus our analysis on retailers in the West Coast region to better understand what drives their sales.
- Hence, as we are not interested in other rows, we remove them and keep only those which have West Coast in the Region column which the first column.
- We can do this by applying a filter.
- Refer Figure 2 to see the dataset after removing the other rows and keeping only those with Region as West Coast.



	Region	Store ID	Ship to Store	Store Opening	Sales Staff	Monthly Traffic	Product Purchased	Coupon Receive	Coupon Sent	Avg Monthy Facebook	Store Location	Populati on	Weekly Repeat	Staff Age	Sales BackGround	Loyalt y Card	% Sales Staff	#Highw ays	Parking Places	Total Sales
2	WestCoast	1001	No Ship	23-11-2014	4	1,677	Football	PC	2 Week	46,171	Mall	4,777	No	22	Operations	No	90	1	5	$42,428
3	WestCoast	1002	No Ship	08-11-2013	11	10,439	Football	PC	2 Week	15,199	Mall	1,01,537	Yes	25	Operations	No	97	2	16	$2,64,107
4	WestCoast	1003	No Ship	13-11-2008	7	3,971	Soccer	PC	1 Month	35,540	Stand Alone	87,932	Yes	19	Operations	Yes	92	1	8	$1,00,466
5	WestCoast	1004	No Ship	24-06-2013	10	2,879	Tennis	PC	1 Month	29,051	Stand Alone	27,742	No	27	Operations	Yes	89	3	9	$72,839
6	WestCoast	1005	No Ship	18-10-2012	10	1,824	Tennis	PC	1 Month	9,102	Stand Alone	45,133	Yes	22	Operations	No	95	3	7	$46,147
7	WestCoast	1006	No Ship	25-12-2008	6	8,263	Tennis	PC	1 Month	33,097	Mall	65,292	No	28	Operations	Yes	78	4	23	$2,09,054
8	WestCoast	1008	Ship	06-01-2013	5	3,811	WorkOut	Mobile	1 Month	22,417	Mall	36,797	Yes	26	Customer Service	Yes	76	4	21	$96,418
9	WestCoast	1009	Ship	04-09-2014	4	1,793	WorkOut	Mobile	1 Month	18,839	Mall	17,622	No	22	Customer Service	No	99	1	22	$45,363
10	WestCoast	1010	No Ship	31-07-2014	8	8,002	WorkOut	Mobile	1 Month	13,368	Mall	30,252	Yes	30	Marketing	No	88	1	7	$2,02,451
11	WestCoast	1011	No Ship	26-12-2010	8	8,625	WorkOut	PC	1 Month	16,425	Mall	34,659	Yes	25	Marketing	No	94	2	23	$2,18,213
12	WestCoast	1012	No Ship	28-08-2014	11	8,765	WorkOut	PC	1 Month	36,748	Mall	22,389	Yes	24	Marketing	No	100	4	20	$2,21,755
13	WestCoast	1013	Ship	13-07-2012	4	8,202	WorkOut	PC	1 Month	44,506	Mall	62,684	No	22	Marketing	Yes	100	2	24	$2,07,511
14	WestCoast	1014	Ship	10-07-2010	11	5,628	WorkOut	PC	1 Month	47,927	Stand Alone	96,631	No	27	Marketing	No	105	3	'NA	$1,42,388
15	WestCoast	1015	Ship	26-01-2009	6	3,100	Biking	PC	1 Month	47,248	Mall	62,484	No	28	Marketing	Yes	84	3	12	$78,430
16	WestCoast	1016	Ship	27-12-2012	4	3,729	Biking	PC	1 Month	'12,xxx	Stand Alone	40,108	No	25	Operations	Yes	95	3	23	$94,344
17	WestCoast	1038	Ship	20-06-2011	4	3,932	Running	PC	2 Week	28,204	Mall	62,839	No	22	Customer Service	Yes	92	4	13	$99,480
18	WestCoast	1039	Ship	27-06-2010	8	2,137	Running	PC	2 Week	35,919	Mall	95,713	No	25	Marketing	No	90	1	32	$54,066
19	WestCoast	1040	Ship	12-07-2008	4	7,072	Running	PC	2 Week	36,329	Mall	1,01,785	Yes	25	Marketing	No	92	4	26	$1,78,922
20	WestCoast	1041	Ship	03-06-2014	11	10,463	Tennis	PC	2 Month	34,653	Mall	50,007	Yes	28	Marketing	Yes	78	1	10	$2,64,714
21	WestCoast	1042	Ship	05-02-2013	11	9,203	Soccer	PC	2 Month	32,177	Mall	22,054	No	25	Marketing	No	85	2	22	$2,32,836
22	WestCoast	1043	Ship	29-05-2010	9	5,399	Soccer	Mobile	2 Month	26,334	Mall	49,528	Yes	24	Marketing	No	103	1	14	$1,36,595
23	WestCoast	1044	Ship	31-08-2011	8	1,424	Soccer	Mobile	2 Month	24,831	Mall	41,458	No	29	Marketing	No	80	2	33	$36,027
24	WestCoast	1045	Ship	17-05-2009	8	9,501	Running	Mobile	2 Month	37,331	Mall	52,953	Yes	23	Operations	No	83	4	13	$2,40,375
25	WestCoast	1046	Ship	05-07-2013	12	4,795	Running	Mobile	2 Week	49,373	Mall	87,410	No	23	Operations	No	83	4	35	$1,21,314
26	WestCoast	1047	Ship	10-05-2013	12	6,167	Running	PC	2 Week	48,839	Stand Alone	74,830	No	19	Operations	No	97	2	17	$1,56,025

Figure 2

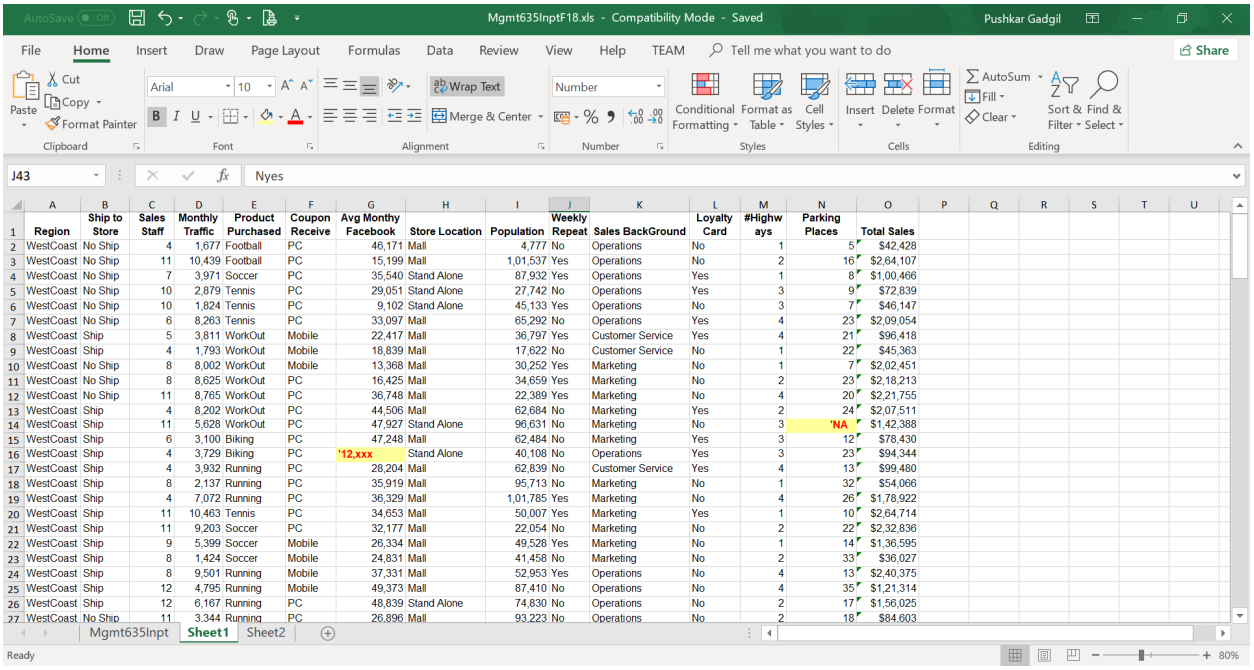
- Now, we find out whether there are any irrelevant columns which are of no use for further analysis.
- Accordingly, in the given dataset, there are some columns which are irrelevant. They are as follows:
  - ➢ Store ID
  - ➢ Store Opening
  - ➢ Coupon Sent
  - ➢ Staff Age
  - ➢ Percentage sales Staff
- We removed above mentioned columns from the dataset. Refer the Figure 3 below:



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	#Highways	Parking Places	Total Sales
WestCoast	1001	No Ship	23-11-2014	4	1,677	Football	PC	2 Week	46,171	Mall	4,777	No	22	Operations	No	90	1	5	$42,428
WestCoast	1002	No Ship	08-11-2013	11	10,439	Football	PC	2 Week	15,199	Mall	1,01,537	Yes	25	Operations	No	97	2	16	$2,64,107
WestCoast	1003	No Ship	13-11-2008	7	3,971	Soccer	PC	1 Month	35,540	Stand Alone	87,932	Yes	19	Operations	Yes	92	1	8	$1,00,466
WestCoast	1004	No Ship	24-06-2013	10	2,879	Tennis	PC	1 Month	29,051	Stand Alone	27,742	No	27	Operations	Yes	89	3	9	$72,839
WestCoast	1005	No Ship	18-10-2012	10	1,824	Tennis	PC	1 Month	9,102	Stand Alone	45,133	Yes	22	Operations	No	95	3	7	$46,147
WestCoast	1006	No Ship	25-12-2008	6	8,263	Tennis	PC	1 Month	33,097	Mall	65,292	No	28	Operations	Yes	78	4	23	$2,09,054
WestCoast	1008	Ship	06-01-2013	5	3,811	WorkOut	Mobile	1 Month	22,417	Mall	36,797	Yes	26	Customer Service	Yes	76	4	21	$96,418
WestCoast	1009	Ship	04-09-2014	4	1,793	WorkOut	Mobile	1 Month	18,839	Mall	17,622	No	22	Customer Service	No	99	1	22	$45,363
WestCoast	1010	No Ship	31-07-2014	8	8,002	WorkOut	Mobile	1 Month	13,368	Mall	30,252	Yes	30	Marketing	No	88	1	7	$2,02,451
WestCoast	1011	No Ship	26-12-2010	8	8,625	WorkOut	PC	1 Month	16,425	Mall	34,659	Yes	25	Marketing	No	94	2	23	$2,18,213
WestCoast	1012	No Ship	28-08-2014	11	8,765	WorkOut	PC	1 Month	36,748	Mall	22,389	Yes	24	Marketing	No	100	4	20	$2,21,755
WestCoast	1013	Ship	13-07-2012	4	8,202	WorkOut	PC	1 Month	44,506	Mall	62,684	No	22	Marketing	Yes	100	2	24	$2,07,511
WestCoast	1014	Ship	10-07-2010	11	5,628	WorkOut	PC	1 Month	47,927	Stand Alone	96,631	No	27	Marketing	No	105	3	'NA	$1,42,388
WestCoast	1015	Ship	26-01-2009	6	3,100	Biking	PC	1 Month	47,248	Mall	62,484	No	28	Marketing	Yes	84	3	12	$78,430
WestCoast	1016	Ship	27-12-2012	4	3,729	Biking	PC	1 Month	'12,xxx	Stand Alone	40,108	No	25	Operations	Yes	95	3	23	$94,344
WestCoast	1038	Ship	20-06-2011	4	3,932	Running	PC	2 Week	28,204	Mall	62,839	No	22	Customer Service	Yes	92	4	13	$99,480
WestCoast	1039	Ship	27-06-2010	8	2,137	Running	PC	2 Week	35,919	Mall	95,713	No	25	Marketing	No	90	1	32	$54,066
WestCoast	1040	Ship	12-07-2008	4	7,072	Running	PC	2 Week	36,329	Mall	1,01,785	Yes	25	Marketing	No	92	4	26	$1,78,922
WestCoast	1041	Ship	03-06-2014	11	10,463	Tennis	PC	2 Month	34,653	Mall	50,007	Yes	28	Marketing	Yes	78	1	10	$2,64,714
WestCoast	1042	Ship	05-02-2013	11	9,203	Soccer	PC	2 Month	32,177	Mall	22,054	No	25	Marketing	No	85	2	22	$2,32,836
WestCoast	1043	Ship	29-05-2010	9	5,399	Soccer	Mobile	2 Month	26,334	Mall	49,528	Yes	24	Marketing	No	103	1	14	$1,36,595
WestCoast	1044	Ship	31-08-2011	8	1,424	Soccer	Mobile	2 Month	24,831	Mall	41,458	No	29	Marketing	No	80	2	33	$36,027
WestCoast	1045	Ship	17-05-2009	8	9,501	Running	Mobile	2 Week	37,331	Mall	52,953	Yes	23	Operations	No	83	4	13	$2,40,375
WestCoast	1046	Ship	05-07-2013	12	4,795	Running	Mobile	2 Week	49,373	Mall	87,410	No	23	Operations	No	83	4	35	$1,21,314
WestCoast	1047	Ship	10-05-2013	12	6,167	Running	PC	2 Week	48,839	Stand Alone	74,830	No	19	Operations	No	97	2	17	$1,56,025

Figure 3

- In the dataset, there are a few outliers which we found out. We can find them in the Screenshots below which are highlighted. Refer Figure 4, Figure 5and Figure 6 below:



Region	Ship to Store	Sales Staff	Monthly Traffic	Product Purchased	Coupon Receive	Avg Monthy Facebook	Store Location	Population	Weekly Repeat	Sales BackGround	Loyalty Card	#Highways	Parking Places	Total Sales
WestCoast	No Ship	4	1,677	Football	PC	46,171	Mall	4,777	No	Operations	No	1	5	$42,428
WestCoast	No Ship	11	10,439	Football	PC	15,199	Mall	1,01,537	Yes	Operations	No	2	16	$2,64,107
WestCoast	No Ship	7	3,971	Soccer	PC	35,540	Stand Alone	87,932	Yes	Operations	Yes	1	8	$1,00,466
WestCoast	No Ship	10	2,879	Tennis	PC	29,051	Stand Alone	27,742	No	Operations	Yes	3	9	$72,839
WestCoast	No Ship	10	1,824	Tennis	PC	9,102	Stand Alone	45,133	Yes	Operations	No	3	7	$46,147
WestCoast	No Ship	6	8,263	Tennis	PC	33,097	Mall	65,292	No	Operations	Yes	4	23	$2,09,054
WestCoast	Ship	5	3,811	WorkOut	Mobile	22,417	Mall	36,797	Yes	Customer Service	Yes	4	21	$96,418
WestCoast	Ship	4	1,793	WorkOut	Mobile	18,839	Mall	17,622	No	Customer Service	No	1	22	$45,363
WestCoast	No Ship	8	8,002	WorkOut	Mobile	13,368	Mall	30,252	Yes	Marketing	No	1	7	$2,02,451
WestCoast	No Ship	8	8,625	WorkOut	PC	16,425	Mall	34,659	Yes	Marketing	No	2	23	$2,18,213
WestCoast	No Ship	11	8,765	WorkOut	PC	36,748	Mall	22,389	Yes	Marketing	No	4	20	$2,21,755
WestCoast	Ship	4	8,202	WorkOut	PC	44,506	Mall	62,684	No	Marketing	Yes	2	24	$2,07,511
WestCoast	Ship	11	5,628	WorkOut	PC	47,927	Stand Alone	96,631	No	Marketing	No	3	'NA	$1,42,388
WestCoast	Ship	6	3,100	Biking	PC	47,248	Mall	62,484	No	Marketing	Yes	3	12	$78,430
WestCoast	Ship	4	3,729	Biking	PC	'12,xxx	Stand Alone	40,108	No	Operations	Yes	3	23	$94,344
WestCoast	Ship	4	3,932	Running	PC	28,204	Mall	62,839	No	Customer Service	Yes	4	13	$99,480
WestCoast	Ship	8	2,137	Running	PC	35,919	Mall	95,713	No	Marketing	No	1	32	$54,066
WestCoast	Ship	4	7,072	Running	PC	36,329	Mall	1,01,785	Yes	Marketing	No	4	26	$1,78,922
WestCoast	Ship	11	10,463	Tennis	PC	34,653	Mall	50,007	Yes	Marketing	Yes	1	10	$2,64,714
WestCoast	Ship	11	9,203	Soccer	PC	32,177	Mall	22,054	No	Marketing	No	2	22	$2,32,836
WestCoast	Ship	9	5,399	Soccer	Mobile	26,334	Mall	49,528	Yes	Marketing	No	1	14	$1,36,595
WestCoast	Ship	8	1,424	Soccer	Mobile	24,831	Mall	41,458	No	Marketing	No	2	33	$36,027
WestCoast	Ship	8	9,501	Running	Mobile	37,331	Mall	52,953	Yes	Operations	No	4	13	$2,40,375
WestCoast	Ship	12	4,795	Running	Mobile	49,373	Mall	87,410	No	Operations	No	4	35	$1,21,314
WestCoast	Ship	12	6,167	Running	PC	48,839	Stand Alone	74,830	No	Operations	No	2	17	$1,56,025
WestCoast	No Ship	11	3,344	Running	PC	26,896	Mall	93,223	No	Operations	No	2	18	$84,603

Figure 4

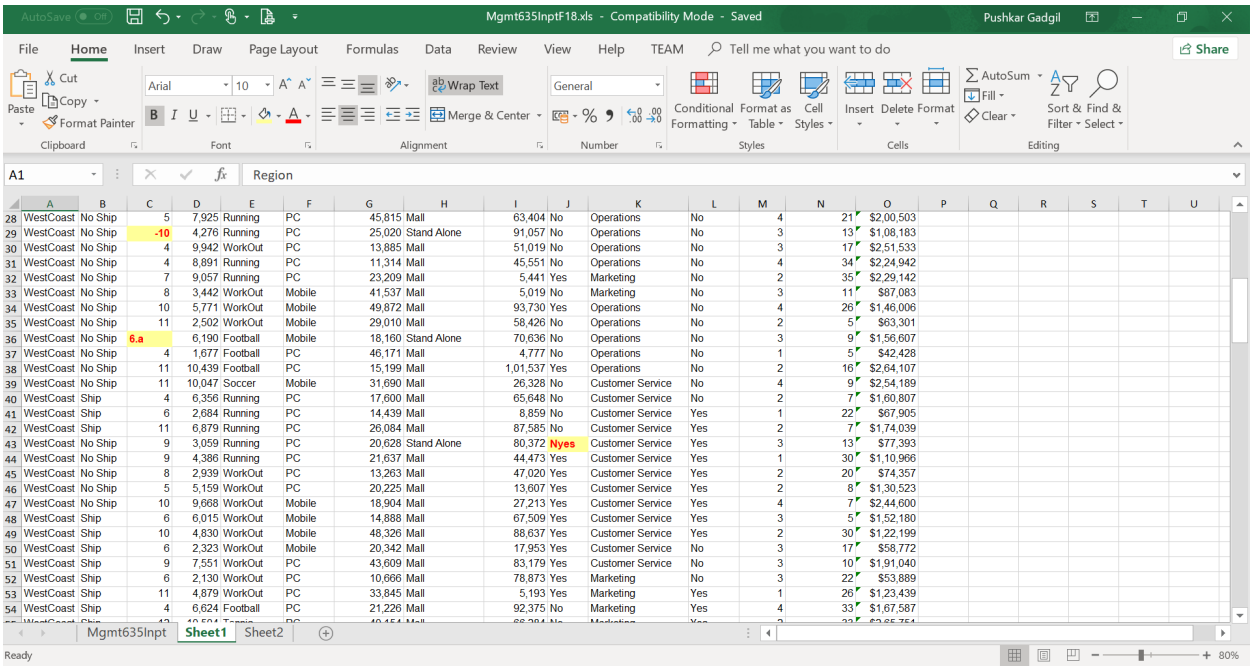


Figure 5

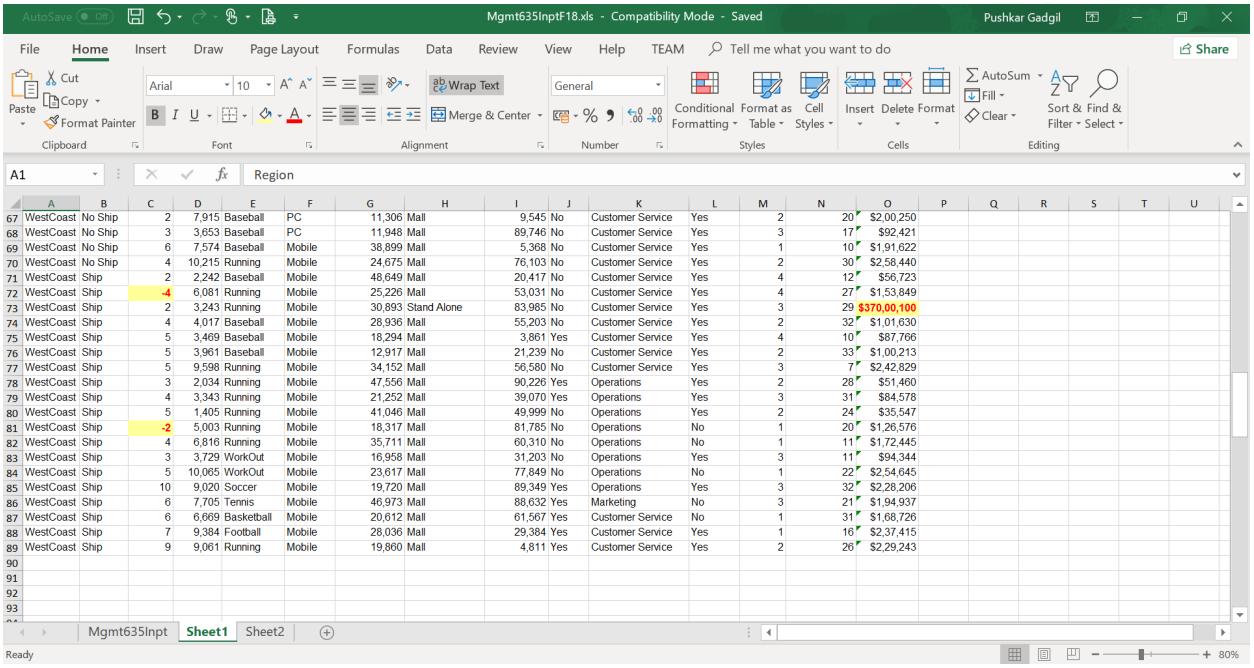
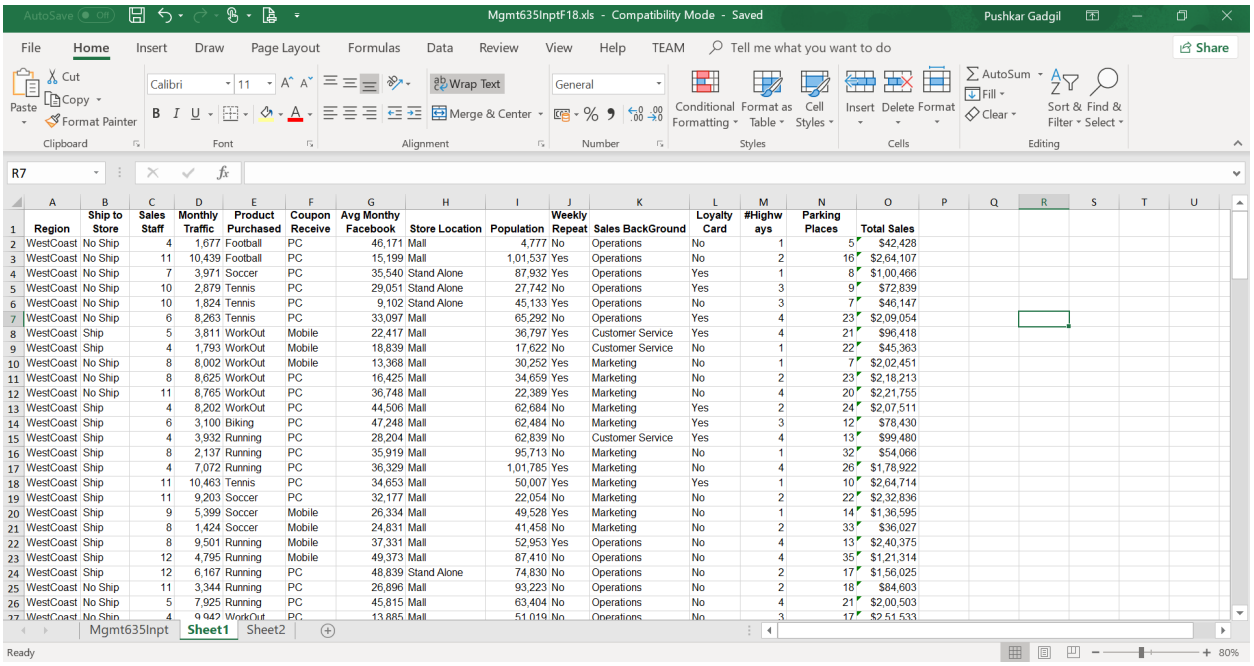


Figure 6

- As these are the outliers, we remove them and hence, we now get the clean data. Refer Figure 7 for the same.



Region	Ship to Store	Sales Staff	Monthly Traffic	Product Purchased	Coupon Receive	Avg Monthy Facebook	Store Location	Population	Weekly Repeat	Sales BackGround	Loyalty Card	#Highways	Parking Places	Total Sales
WestCoast	No Ship	4	1,677	Football	PC	46,171	Mall	4,777	No	Operations	No	1	5	$42,428
WestCoast	No Ship	11	10,439	Football	PC	15,199	Mall	1,01,537	Yes	Operations	No	2	16	$2,64,107
WestCoast	No Ship	7	3,971	Soccer	PC	35,540	Stand Alone	87,932	Yes	Operations	Yes	1	8	$1,00,466
WestCoast	No Ship	10	2,879	Tennis	PC	29,051	Stand Alone	27,742	No	Operations	Yes	3	9	$72,839
WestCoast	No Ship	10	1,824	Tennis	PC	9,102	Stand Alone	45,133	Yes	Operations	No	3	7	$46,147
WestCoast	No Ship	6	8,263	Tennis	PC	33,097	Mall	65,292	No	Operations	Yes	4	23	$2,09,054
WestCoast	Ship	5	3,811	WorkOut	Mobile	22,417	Mall	36,797	Yes	Customer Service	Yes	4	21	$96,418
WestCoast	Ship	4	1,793	WorkOut	Mobile	18,839	Mall	17,622	No	Customer Service	No	1	22	$45,363
WestCoast	No Ship	8	8,002	WorkOut	Mobile	13,368	Mall	30,252	Yes	Marketing	No	1	7	$2,02,451
WestCoast	No Ship	8	8,625	WorkOut	PC	16,425	Mall	34,659	Yes	Marketing	No	2	23	$2,18,213
WestCoast	No Ship	11	8,765	WorkOut	PC	36,748	Mall	22,389	Yes	Marketing	No	4	20	$2,21,755
WestCoast	Ship	4	8,202	WorkOut	PC	44,506	Mall	62,684	No	Marketing	Yes	2	24	$2,07,511
WestCoast	Ship	6	3,100	Biking	PC	47,248	Mall	62,484	No	Marketing	Yes	3	12	$78,430
WestCoast	Ship	4	3,932	Running	PC	28,204	Mall	62,839	No	Customer Service	Yes	4	13	$99,480
WestCoast	Ship	8	2,137	Running	PC	35,919	Mall	95,713	No	Marketing	No	1	32	$54,066
WestCoast	Ship	4	7,072	Running	PC	36,329	Mall	1,01,785	Yes	Marketing	No	4	26	$1,78,922
WestCoast	Ship	11	10,463	Tennis	PC	34,653	Mall	50,007	Yes	Marketing	Yes	1	10	$2,64,714
WestCoast	Ship	11	9,203	Soccer	PC	32,177	Mall	22,054	No	Marketing	No	2	22	$2,32,836
WestCoast	Ship	9	5,399	Soccer	Mobile	26,334	Mall	49,528	Yes	Marketing	No	1	14	$1,36,595
WestCoast	Ship	8	1,424	Soccer	Mobile	24,831	Mall	41,458	No	Marketing	No	2	33	$36,027
WestCoast	Ship	8	9,501	Running	Mobile	37,331	Mall	52,953	Yes	Operations	No	4	13	$2,40,375
WestCoast	Ship	12	4,795	Running	Mobile	49,373	Mall	87,410	No	Operations	No	4	35	$1,21,314
WestCoast	Ship	12	6,167	Running	PC	48,839	Stand Alone	74,830	No	Operations	No	2	17	$1,56,025
WestCoast	No Ship	11	3,344	Running	PC	26,896	Mall	93,223	No	Operations	No	2	18	$84,603
WestCoast	No Ship	5	7,925	Running	PC	45,815	Mall	63,404	No	Operations	No	4	21	$2,00,503
WestCoast	No Ship	4	9,942	WorkOut	PC	13,885	Mall	51,019	No	Operations	No	3	17	$2,51,533

Figure 7

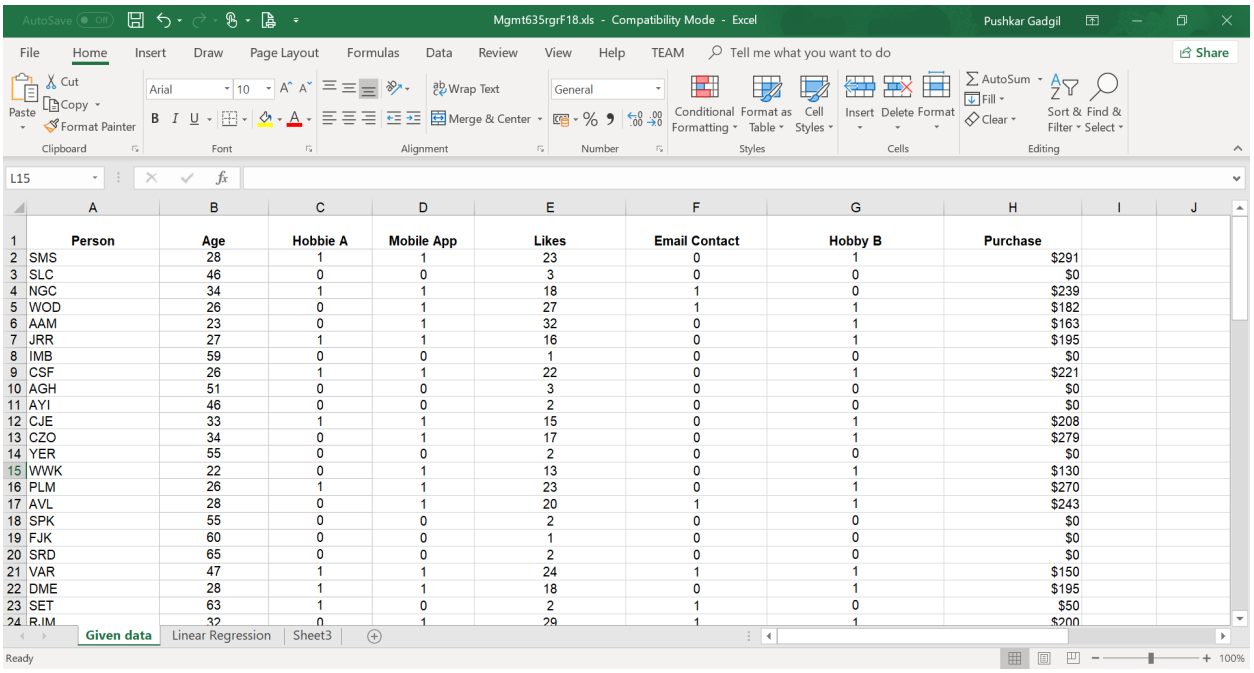
3. Finding Degree of Freedom

- Now we find the Degree of Freedom for both original as well as cleaned data.
  - ➢ For Original Data:
    - Total Number of Records: 163
    - Total Number of Columns: 20
    - Degree of Freedom: 163 – 20 =143
  - ➢ For Clean Data:
    - Total Number of Records: 80
    - Total Number of Columns: 15
    - Degree of Freedom: 80 – 15 =65

Report Part -2

1. Business Problem

- We are given a dataset which tells us about a major retailing organization which wants to better understand the type of customer that reacts more to the coupons (marketing coupons) that have been sent to them.
- The main aim is to run a regression analysis with the data provided to understand the spending patterns of people for the products.
- Also, we are assigned to compare the results obtained using Linear Regression and Neural Networks.
- The Figure 1 shows the Screenshot of the given dataset.

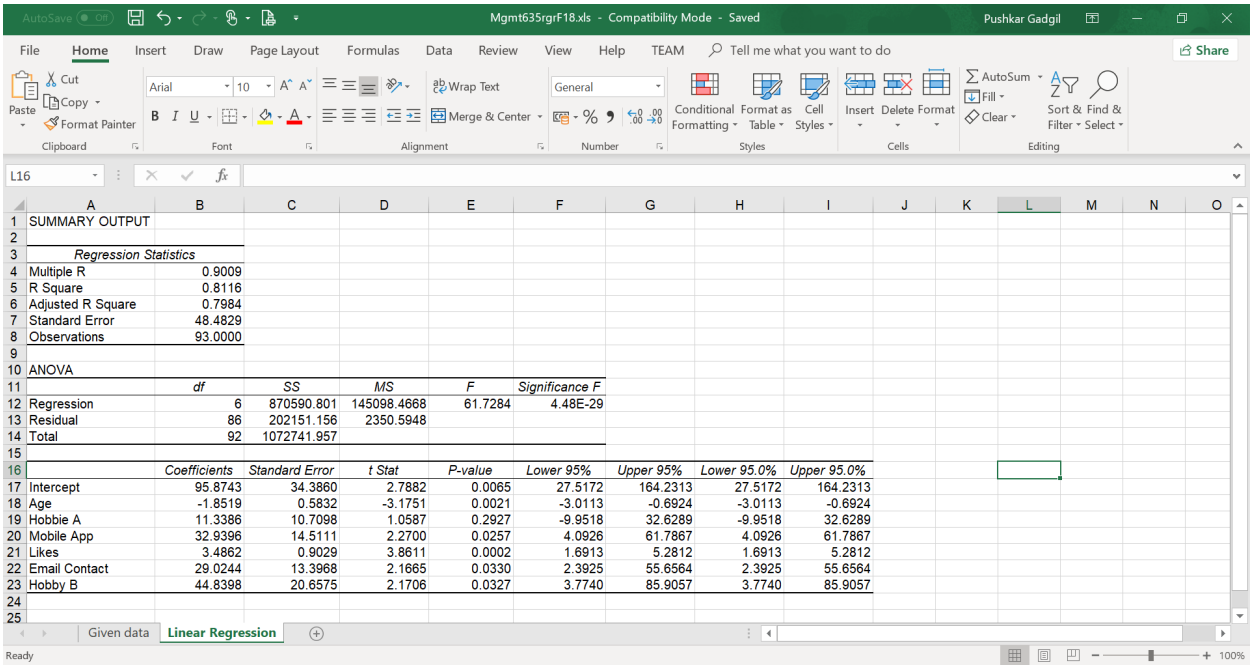


	Person	Age	Hobbie A	Mobile App	Likes	Email Contact	Hobby B	Purchase
2	SMS	28	1	1	23	0	1	$291
3	SLC	46	0	0	3	0	0	$0
4	NGC	34	1	1	18	1	0	$239
5	WOD	26	0	1	27	1	1	$182
6	AAM	23	0	1	32	0	1	$163
7	JRR	27	1	1	16	0	1	$195
8	IMB	59	0	0	1	0	0	$0
9	CSF	26	1	1	22	0	1	$221
10	AGH	51	0	0	3	0	0	$0
11	AYI	46	0	0	2	0	0	$0
12	CJE	33	1	1	15	0	1	$208
13	CZO	34	0	1	17	0	1	$279
14	YER	55	0	0	2	0	0	$0
15	WWK	22	0	1	13	0	1	$130
16	PLM	26	1	1	23	0	1	$270
17	AVL	28	0	1	20	1	1	$243
18	SPK	55	0	0	2	0	0	$0
19	FJK	60	0	0	1	0	0	$0
20	SRD	65	0	0	2	0	0	$0
21	VAR	47	1	1	24	1	1	$150
22	DME	28	1	1	18	0	1	$195
23	SET	63	1	0	2	1	0	$50
24	RJM	32	0	1	29	1	1	$200

Figure 1

2. What is Linear Regression?

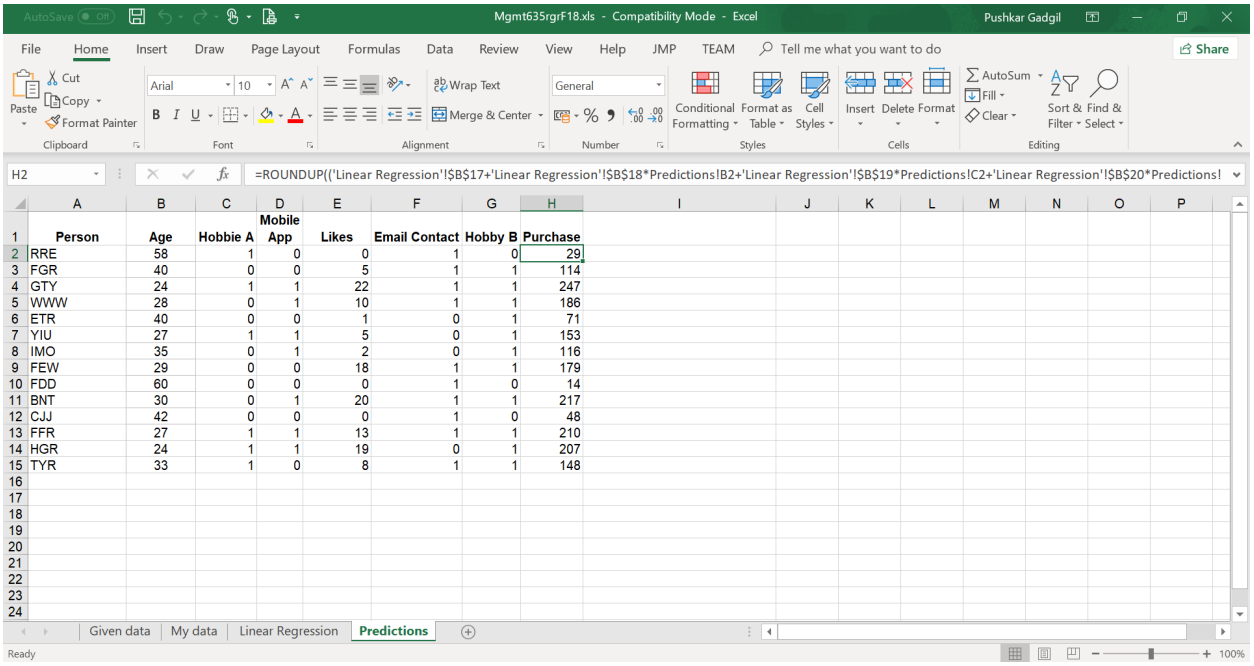
- Linear regression is a basic and commonly used type of predictive analysis.
- These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.
- The overall idea of regression is to examine two things:
  - Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
  - Which variables are particularly significant predictors of the outcome variable, and in what way do they impact the outcome variable indicated by the magnitude and sign of the beta estimates?
- We now perform Linear Regression on the given dataset. Figure 2 is the Screenshot for the Linear Regression Analysis done on the given dataset.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.9009							
5	R Square	0.8116							
6	Adjusted R Square	0.7984							
7	Standard Error	48.4829							
8	Observations	93.0000							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	6	870590.801	145098.4668	61.7284	4.48E-29			
13	Residual	86	202151.156	2350.5948					
14	Total	92	1072741.957						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	95.8743	34.3860	2.7882	0.0065	27.5172	164.2313	27.5172	164.2313
18	Age	-1.8519	0.5832	-3.1751	0.0021	-3.0113	-0.6924	-3.0113	-0.6924
19	Hobbie A	11.3386	10.7098	1.0587	0.2927	-9.9518	32.6289	-9.9518	32.6289
20	Mobile App	32.9396	14.5111	2.2700	0.0257	4.0926	61.7867	4.0926	61.7867
21	Likes	3.4862	0.9029	3.8611	0.0002	1.6913	5.2812	1.6913	5.2812
22	Email Contact	29.0244	13.3968	2.1665	0.0330	2.3925	55.6564	2.3925	55.6564
23	Hobby B	44.8398	20.6575	2.1706	0.0327	3.7740	85.9057	3.7740	85.9057
24									
25									

Figure 2

- After carrying out the Linear Regression using Excel, we use these results for further predictions. This is shown in Figure 3 below.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**Person**	**Age**	**Hobbie A**	**Mobile App**	**Likes**	**Email Contact**	**Hobby B**	**Purchase**								
2	RRE	58	1	0	0	1	0	29								
3	FGR	40	0	0	5	1	1	114								
4	GTY	24	1	1	22	1	1	247								
5	WWW	28	0	1	10	1	1	186								
6	ETR	40	0	0	1	0	1	71								
7	YIU	27	1	1	5	0	1	153								
8	IMO	35	0	1	2	0	1	116								
9	FEW	29	0	0	18	1	1	179								
10	FDD	60	0	0	0	1	0	14								
11	BNT	30	0	1	20	1	1	217								
12	CJJ	42	0	0	0	1	0	48								
13	FFR	27	1	1	13	1	1	210								
14	HGR	24	1	1	19	0	1	207								
15	TYR	33	1	0	8	1	1	148								
16																
17																
18																
19																
20																
21																
22																
23																
24																

Figure 3

- For predicting the purchase value, we used the following formula:

ROUNDUP(('LinearRegression'!$B$17+'LinearRegression'!$B$18*Predictions!B2+'LinearRegression'!$B$19*Predictions!C2+'LinearRegression'!$B$20*Predictions!D2+'LinearRegression'!$B$21*Predictions!E2+'LinearRegression'!$B$22*Predictions!F2+'LinearRegression'!$B$23*Predictions!G2),0)

- This formula directly gives us the Predicted Purchase value in whole numbers.

3. What are Neural Networks?

- Neural networks are notable for being adaptive, which means they modify themselves as they learn from initial training and subsequent runs provide more information about the world.
- The most basic learning model is centred on weighting the input streams, which is how each node weights the importance of input from each of its predecessors. Inputs that contribute to getting right answers are weighted higher.
- This Figure 4 shows the value of R^2 using Neural Network in SAS JMP along with the Neural Network diagram.

Figure 4

Summary of Fit

RSquare	0.811557
RSquare Adj	0.798409
Root Mean Square Error	48.48293
Mean of Response	101.3118
Observations (or Sum Wgts)	93

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	870590.8	145098	61.7284
Error	86	202151.2	2351	Prob > F
C. Total	92	1072742.0		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	95.874254	34.38597	2.79	0.0065*
Age	-1.851875	0.583245	-3.18	0.0021*
Hobbie A	11.33859	10.70979	1.06	0.2927
Mobile App	32.939631	14.51106	2.27	0.0257*
Likes	3.4862349	0.902906	3.86	0.0002*
Email Contact	29.024427	13.39678	2.17	0.0330*
Hobby B	44.839804	20.65755	2.17	0.0327*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Age	1	1	23697.274	10.0814	0.0021*
Hobbie A	1	1	2634.719	1.1209	0.2927
Mobile App	1	1	12112.024	5.1527	0.0257*
Likes	1	1	35043.408	14.9083	0.0002*
Email Contact	1	1	11033.263	4.6938	0.0330*
Hobby B	1	1	11075.100	4.7116	0.0327*

Here, as we can see that P values for Hobbie A, Hobby B and Email Contact is significantly high, so we will try removing Hobbie A(highest P value) as the decision variables and re-run the regression using Fit Model, we get below output-

Summary of Fit

RSquare	0.809101
RSquare Adj	0.798129
Root Mean Square Error	48.5166
Mean of Response	101.3118
Observations (or Sum Wgts)	93

Analysis of Variance

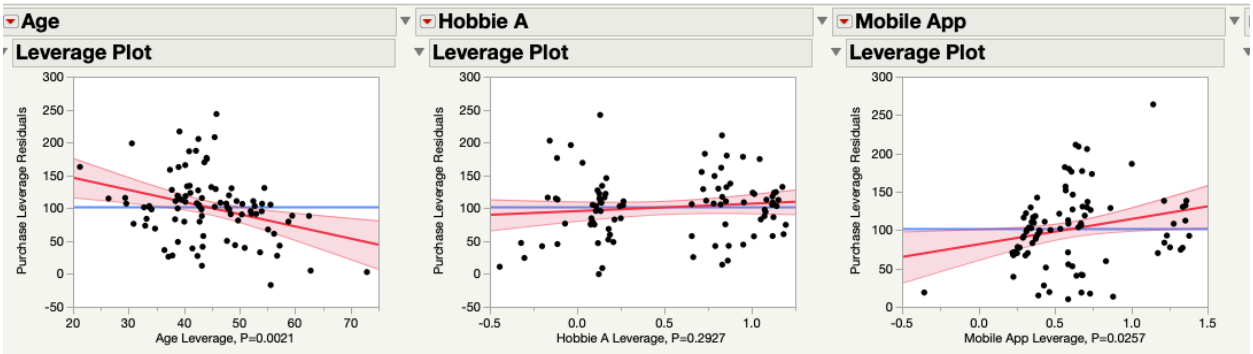
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	867956.1	173591	73.7474
Error	87	204785.9	2354	Prob > F
C. Total	92	1072742.0		<.0001*

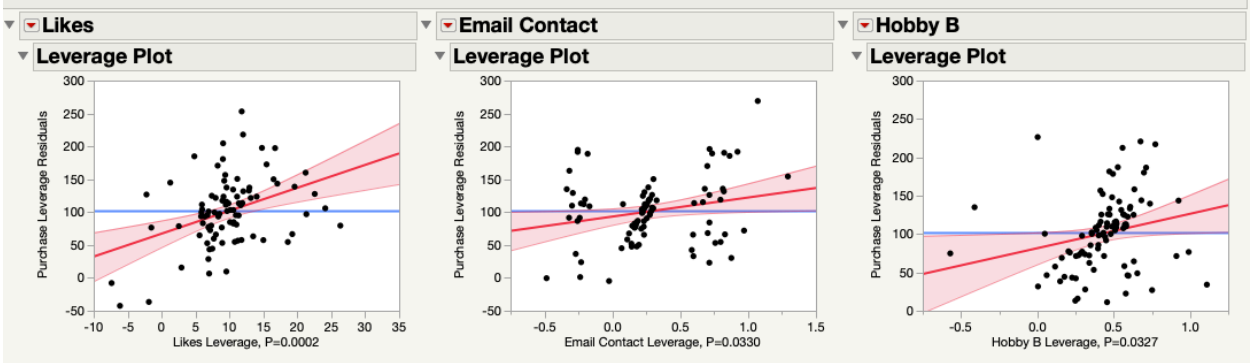
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	101.23577	34.03463	2.97	0.0038*
Age	-1.874646	0.583253	-3.21	0.0018*
Mobile App	31.729142	14.476	2.19	0.0311*
Likes	3.6869921	0.883382	4.17	<.0001*
Email Contact	28.427067	13.39419	2.12	0.0367*
Hobby B	45.547376	20.66107	2.20	0.0301*

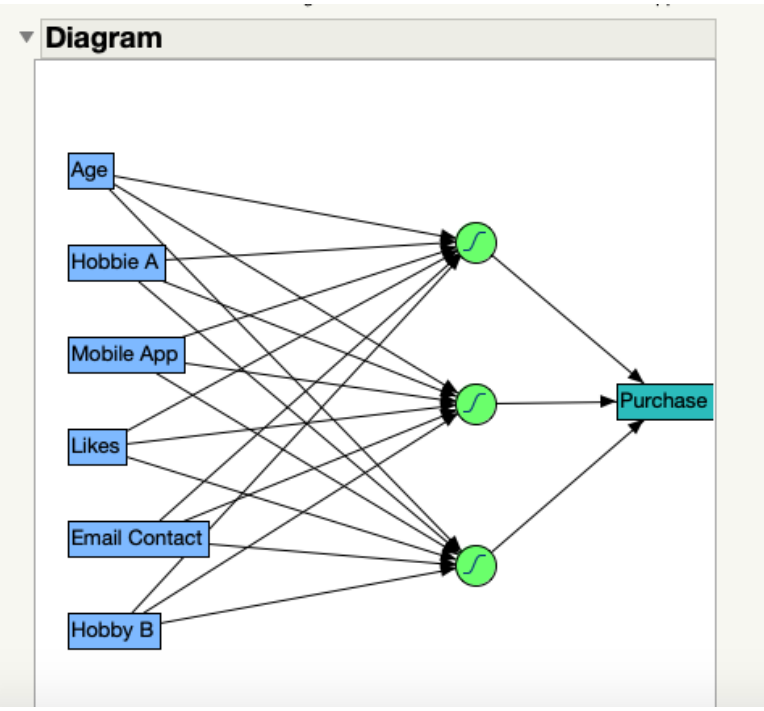
We can see that it has impacted the R2 value if we remove one of the decision variable. And this is same for Hobbie B. Also we can't remove Email contact, because that's necessary for our prediction analysis. So we will keep all the variables as present in Figure 4.

The other diagram and graphs for this Fit model is as below-





Below diagram shows the connectivity of the decision variable using 3 hidden nodes.



Below is the output for Neural Net prediction analysis-

If we select the Purchase as a target Variable(Y) and Rest other variables as decision variables (X)-

Neural

Validation: Random Holdback

Model Launch

Model NTanH(3)

Training

Purchase

Measures	Value
RSquare	0.8537185
RMSE	40.536961
Mean Abs Dev	28.134892
-LogLikelihood	317.51147
SSE	101881.2
Sum Freq	62

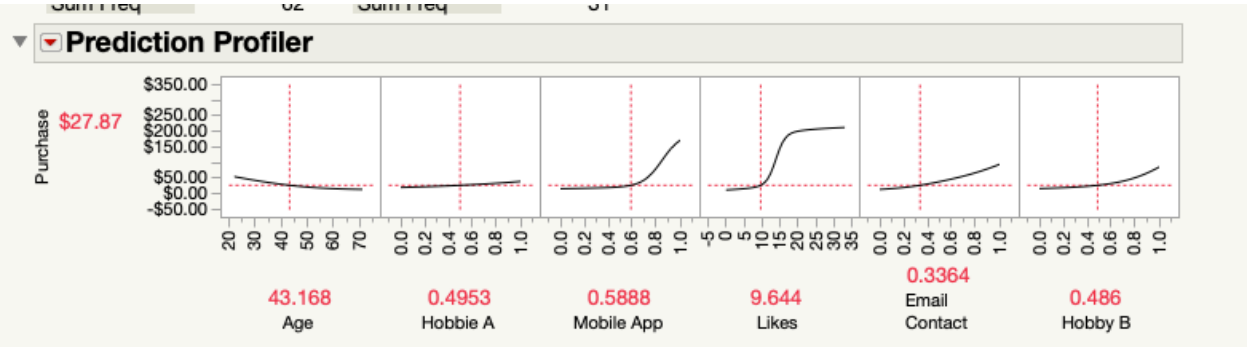
Validation

Purchase

Measures	Value
RSquare	0.9034372
RMSE	34.198947
Mean Abs Dev	23.674624
-LogLikelihood	153.48514
SSE	36256.608
Sum Freq	31

We can see that the Training R^2 value is 0.85 and while validating it is almost 90% which is quite a good model and predicts the values quite perfectly.

Below are the profiler for **Prediction Analysis-**



We can see that –

1. Age, Hobbie A has not that significant relation in deciding the Purchase variable as we can see the almost flat curve.
2. Mobile App has direct relationship with the decision variable (Purchase) after a certain point and thus, higher (1) the Mobile App value, higher the purchase value.

3. Likes has also a direct relationship with the Purchase but after a point it becomes stagnate in terms of affecting Purchase. This implies that the higher number of likes by a customer upto a point has great impact on Purchase.
4. Email Contact and Hobby B has somewhat direct relationship with Purchase but it is also not that significant. This implies that Whether a customer has been sent the marketing messages over email has nominal impact on Purchase but it is also dependant on Likes on Facebook and Mobile App presence on customer's device.

Below is the predicted values column beside the purchase column for our historical data and we can see that there are some difference but mostly it is quite near.

	Person	Age	Hobbie A	Mobile App	Likes	Email Contact	Hobby B	Purchase	Predicted Purchase
1	SMS	28.00	1.00	1.00	22.67	0.00	1.00	$291.00	212.16110542
2	SLC	46.00	0.00	0.00	2.53	0.00	0.00	$0.00	19.51980239
3	NGC	34.00	1.00	1.00	17.50	1.00	0.00	$239.00	167.22226503
4	WOD	26.00	0.00	1.00	27.33	1.00	1.00	$182.00	249.81978791
5	AAM	23.00	0.00	1.00	32.00	0.00	1.00	$163.00	242.62008208
6	JRR	27.00	1.00	1.00	15.73	0.00	1.00	$195.00	189.84175185
7	IMB	59.00	0.00	0.00	0.67	0.00	0.00	$0.00	-11.06221002
8	CSF	26.00	1.00	1.00	22.00	0.00	1.00	$221.00	213.54069868
9	AGH	51.00	0.00	0.00	3.00	0.00	0.00	$0.00	11.887337394
10	AYI	46.00	0.00	0.00	2.00	0.00	0.00	$0.00	17.660477122
11	CJE	33.00	1.00	1.00	15.00	0.00	1.00	$208.00	176.17393008
12	CZO	34.00	0.00	1.00	17.07	0.00	1.00	$279.00	170.18835044
13	YER	55.00	0.00	0.00	2.00	0.00	0.00	$0.00	0.9936028317
14	WWK	22.00	0.00	1.00	13.20	0.00	1.00	$130.00	178.9307413
15	PLM	26.00	1.00	1.00	23.46	0.00	1.00	$270.00	218.62479121
16	AVL	28.00	0.00	1.00	20.00	1.00	1.00	$243.00	220.55031563
17	SPK	55.00	0.00	0.00	2.00	0.00	0.00	$0.00	0.9936028317
18	FJK	60.00	0.00	0.00	1.00	0.00	0.00	$0.00	-11.75200665
19	SRD	65.00	0.00	0.00	2.00	0.00	0.00	$0.00	-17.52514638
20	VAR	47.00	1.00	1.00	24.42	1.00	1.00	$150.00	212.10081964
21	DME	28.00	1.00	1.00	17.53	0.00	1.00	$195.00	194.26509971
22	SET	63.00	1.00	0.00	2.00	1.00	0.00	$50.00	26.541620286
23	RJM	32.00	0.00	1.00	29.08	1.00	1.00	$200.00	244.80944942
24	MJW	53.00	1.00	0.00	0.00	0.00	0.00	$25.00	9.063473057
25	SJP	64.00	0.00	1.00	1.00	0.00	0.00	$0.00	13.780125087
26	PT	26.00	1.00	1.00	29.17	1.00	1.00	$190.00	267.54980865
27	CHH	25.00	1.00	1.00	19.20	0.00	1.00	$250.00	205.63111594
28	MCE	60.00	1.00	0.00	2.00	0.00	0.00	$0.00	3.0728183649
29	LJP	53.00	0.00	0.00	0.67	0.00	0.00	$25.00	0.0490395034
30	SMM	55.00	1.00	0.00	0.00	0.00	0.00	$0.00	5.3597232147
31	SLV	63.00	0.00	1.00	0.00	0.00	0.00	$24.00	12.145765131
32	NGT	51.00	0.00	0.00	0.00	0.00	0.00	$0.00	1.4286327604
33	WWD	57.00	1.00	1.00	14.00	0.00	0.00	$100.00	83.402893088
34	ARR	47.00	0.00	1.00	3.00	0.00	0.00	$23.00	52.234468503
35	ART	42.00	0.00	0.00	1.00	0.00	0.00	$0.00	21.581741929

And below is the prediction for new Value we are trying to find values of-

88	UHG	42.00	0.00	1.00	2.00	0.00	0.00	$20.00	58.007608231
89	ODN	24.00	1.00	1.00	22.33	1.00	1.00	$250.00	247.4309535
90	EDJ	61.00	0.00	1.00	1.00	0.00	0.00	$24.00	19.335749851
91	ENK	29.00	1.00	1.00	13.33	1.00	1.00	$145.00	206.79546499
92	LPK	32.00	0.00	1.00	10.00	0.00	1.00	$250.00	149.25604048
93	WDU	38.00	1.00	0.00	2.00	0.00	0.00	$24.00	43.81406663
94	RRE	58.00	1.00	0.00	0.00	1.00	0.00	•	28.828525136
95	FGR	40.00	0.00	0.00	5.00	1.00	1.00	•	113.09466198
96	GTY	24.00	1.00	1.00	22.00	1.00	1.00	•	246.2688752
97	WWW	28.00	0.00	1.00	10.00	1.00	1.00	•	185.68796685
98	ETR	40.00	0.00	0.00	1.00	0.00	1.00	•	70.125295781
99	YIU	27.00	1.00	1.00	5.00	0.00	1.00	•	152.42283083
100	IMO	35.00	0.00	1.00	2.00	0.00	1.00	•	115.81053669
101	FEW	29.00	0.00	0.00	18.00	1.00	1.00	•	178.78633952
102	FDD	60.00	0.00	0.00	0.00	1.00	0.00	•	13.786185155
103	BNT	30.00	0.00	1.00	20.00	1.00	1.00	•	216.84656578
104	CJJ	42.00	0.00	0.00	0.00	1.00	0.00	•	47.119933736
105	FFR	27.00	1.00	1.00	13.00	1.00	1.00	•	209.33713654
106	HGR	24.00	1.00	1.00	19.00	0.00	1.00	•	206.78574389
107	TYR	33.00	1.00	0.00	8.00	1.00	1.00	•	147.8550812

4. Comparison value for Regression & Neural Networks

Below is the comparison of Regression & Neural Net predicted values for the below data and we can see that both methods are giving us almost similar values-

Person	Age	Hobbie A	App	Likes	Email Contact	Hobby B	Purchase- Regression	Purchase - Neural Net
RRE	58	1	0	0	1	0	29	28.82852514
FGR	40	0	0	5	1	1	113	113.094662
GTY	24	1	1	22	1	1	246	246.2688752
WWW	28	0	1	10	1	1	186	185.6879669
ETR	40	0	0	1	0	1	70	70.12529578
YIU	27	1	1	5	0	1	152	152.4228308
IMO	35	0	1	2	0	1	116	115.8105367
FEW	29	0	0	18	1	1	179	178.7863395
FDD	60	0	0	0	1	0	14	13.78618516
BNT	30	0	1	20	1	1	217	216.8465658
CJJ	42	0	0	0	1	0	47	47.11993374
FFR	27	1	1	13	1	1	209	209.3371365
HGR	24	1	1	19	0	1	207	206.7857439
TYR	33	1	0	8	1	1	148	147.8550812

Part- 3

1. Business Problem

- Here we have the data set for Breast Cancer presence in a set of population of Wisconsin region which was collected while doing the diagnostic of these patients.
- According to the data available, we are trying to mine this data to predict the possibility of the Breast Cancer presence.
- Once we do the regression analysis on this data, we are creating a model which will predict whether the patient's cancer is benign or malignant (harmless at the moment or malicious).
- This prediction model is based on the characteristics of the cancer cell nuclei present in the patient's body such as Radius, Texture, Perimeter, Area, Smoothness, Compactness, etc.

2. Variables of the data sets

- Target Variable Diagnosis
  Possible values (Value Range): 1 – Malignant, 0 – benign

- Decision variables:

Sr. No	Variable Name	Description
1	Radius Mean	Mean of distances from center to points on the perimeter
2	Texture Mean	Standard deviation of gray-scale values
3	Perimeter Mean	Mean size of the core tumor
4	Area Mean	Mean size of the area of the cancer cells
5	Smoothness Mean	Mean of local variation in radius lengths
6	Compactness mean	Mean of perimeter^2 /area -1.0
7	Concavity Mean	Mean of severity of concave portions of the contour
8	Concave Points Mean	Mean for number of concave portions of the contour
9	Symmetry Mean	Mean for Symmetry of the tumor
10	Fractal dimension Mean	Mean for coastline approximation -1

3. Data Source

- We have obtained this healthcare related data from Kaggle which shows the historical data for set of population having patients from Wisconsin region
- https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/home
- Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. n the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

- This database is also available through the UW CS ftp server: ftp ftp.cs.wisc.edu cd math-prog/cpo-dataset/machine-learn/WDBC/
- Also can be found on UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

4. Analysis on Data

- **Transforming data:** First we transformed below column to have quantitative variable instead of categorical variable as we need Target Variable to be quantitative for prediction.

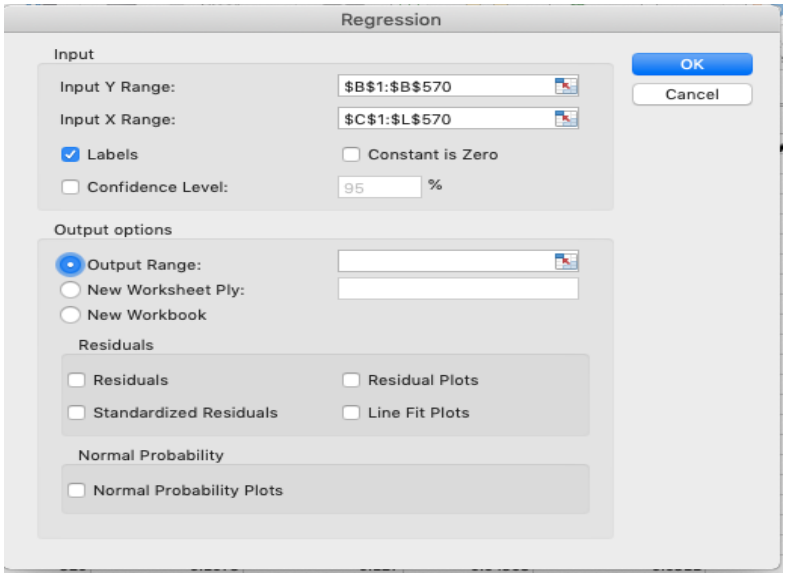
Column 1- Diagnostic

Previous values:		New Values:	
M=	Malignant	1 =	Malignant
B =	Benign	0 =	Benign

A	B	C	D	E	F	G	H	I	J	K	L
id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
842302	1	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
842517	1	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
84300903	1	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999
84348301	1	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744
84358402	1	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883
843786	1	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613
844359	1	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742
84458202	1	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451
844981	1	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389
84501001	1	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243
845636	1	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697
84610002	1	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082
846226	1	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078
846381	1	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338
84667401	1	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682
84799002	1	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077
848406	1	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922
84862001	1	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356
849014	1	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395
8510426	0	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766
8510653	0	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811
8510824	0	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905
8511133	1	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032
851509	1	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278
852552	1	16.65	21.38	110	904.6	0.1121	0.1457	0.1525	0.0917	0.1995	0.0633
852631	1	17.14	16.4	116	912.7	0.1186	0.2276	0.2229	0.1401	0.304	0.07413
852763	1	14.58	21.53	97.41	644.8	0.1054	0.1868	0.1425	0.08783	0.2252	0.06924
852781	1	18.61	20.25	122.1	1094	0.0944	0.1066	0.149	0.07731	0.1697	0.05699
852973	1	15.3	25.27	102.4	732.4	0.1082	0.1697	0.1683	0.08751	0.1926	0.0654
853201	1	17.57	15.05	115	955.1	0.09847	0.1157	0.09875	0.07953	0.1739	0.06149

- **Linear Regression:**

Step 1: We first performed the linear regression on the set of data as shown below –



Step 2: Below is the **output of the regression-**

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.826295401							
R Square	0.68276409							
Adjusted R Square	0.677078858							
Standard Error	0.274991948							
Observations	569							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	10	90.81602346	9.081602346	120.0943365	4.40E-132			
Residual	558	42.19627882	0.075620571					
Total	568	133.0123023						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-2.052084249	0.416653914	-4.92515294	1.11E-06	-2.870486052	-1.233682445	-2.870486052	-1.233682445
radius_mean	0.490012293	0.131245854	3.733544941	0.000208178	0.232215979	0.747808608	0.232215979	0.747808608
texture_mean	0.021973203	0.002922766	7.517948376	2.23E-13	0.016232235	0.027714171	0.016232235	0.027714171
perimeter_mean	-0.054974678	0.021001833	-2.617613329	0.009094986	-0.096226992	-0.013722364	-0.096226992	-0.013722364
area_mean	-0.000954771	0.000245981	-3.881484756	0.000116233	-0.001437933	-0.00047161	-0.001437933	-0.00047161
smoothness_mean	1.94086211	1.410797058	1.37572027	0.169460202	-0.83025996	4.711984179	-0.83025996	4.711984179
compactness_mean	0.097260808	1.039078697	0.093602927	0.925458183	-1.94372297	2.138244586	-1.94372297	2.138244586
concavity_mean	0.809767523	0.495398551	1.634577901	0.102701553	-0.163306421	1.782841467	-0.163306421	1.782841467
concave points_mean	6.431011461	1.385580999	4.641382544	4.32E-06	3.709419391	9.152603532	3.709419391	9.152603532
symmetry_mean	1.011900043	0.561293282	1.802800917	0.071958894	-0.090605943	2.114406029	-0.090605943	2.114406029
fractal_dimension_mean	-0.119292419	4.157825795	-0.028691058	0.977121247	-8.28619549	8.047610652	-8.28619549	8.047610652

Output Interpretation:

The R^2 measures how well the model explains the variation in the target variable.
It Ranges from (0 to 1) where the better the model the closer to (1) the R^2 registers.
The R^2 of (.68) reflects that the multivariate equations accurately models the diagnostic output from breast cancer. In other words, almost 68% of the variation in Output is explained by the explanatory variables.

Also, as we can see that the P values for Smoothness, compactness, concavity, symmetry & fractal dimension is quite low, we can ignore those and see the value of R^2 is not getting affected that much. Thus, finalizing only rest of the 5 decision variables, we have the output as below-

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.823176253							
R Square	0.677619144							
Adjusted R Square	0.674756081							
Standard Error	0.275979186							
Observations	569							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	5	90.13168242	18.02633648	236.6763231	8.0721E-136			
Residual	563	42.88061986	0.076164511					
Total	568	133.0123023						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.538645149	0.167486221	-9.1866969	7.72328E-19	-1.867619328	-1.209670969	-1.867619328	-1.209670969
radius_mean	0.373725826	0.081521283	4.58439578	5.61355E-06	0.21360282	0.533848832	0.21360282	0.533848832
texture_mean	0.021678822	0.002866078	7.563934063	1.60175E-13	0.016049311	0.027308333	0.016049311	0.027308333
perimeter_mean	-0.039122136	0.012542838	-3.11908167	0.001906943	-0.06375861	-0.014485663	-0.06375861	-0.014485663
area_mean	-0.000974	0.000209341	-4.652685287	4.08855E-06	-0.001385186	-0.000562814	-0.001385186	-0.000562814
concave points_mean	9.182603809	0.90043362	10.19797973	1.59547E-22	7.413984222	10.9512234	7.413984222	10.9512234

Step 3: According to the Regression output, below is the Regression Equation which is calculated as below and can be used for predicting the values-

= -1.538645149 + (AV34 * 0.373725826) + (AW34 * 0.021678822) + (AX34 * -0.039122136) + (AY34 * -0.000974) + (AZ34 * 9.182603809)

radius_mean	texture_mean	perimeter_mean	area_mean	concave points_mean	diagnosis
17.99	10.38	122.8	1001	0.1471	0.981297353
20.57	17.77	132.9	1326	0.07017	0.687615194
19.69	21.25	130	1203	0.1279	1.19754668
11.42	20.38	77.58	386.1	0.1052	0.725971386
20.29	14.34	135.1	1297	0.1043	0.764193172
13.54	14.36	87.46	566.3	0.04781	0.298732493
13.08	15.71	85.63	520	0.0311	0.119333421
9.504	12.44	60.34	273.9	0.02076	-0.153847784
13	14.3	87.45	566.9	0.05	0.115536541
20.29	14.34	135.1	1297	0.12	0.908360051

Here, Last two rows are the values which we have predicted on the basis of the values of the decision variables.

As we can see in the screenshot above-
The first prediction depicts that the breast cancer is Benign or harmless in the current situation.

The second prediction depicts that the breast cancer is Malignant or Malicious in the current situation.

Below is the output of the **Neural Net fit model** –

diagnosis Predicted

Summary of Fit

RSquare	0.677619
RSquare Adj	0.674756
Root Mean Square Error	0.275979
Mean of Response	0.372583
Observations (or Sum Wgts)	569

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	90.13168	18.0263	236.6763
Error	563	42.88062	0.0762	Prob > F
C. Total	568	133.01230		<.0001*

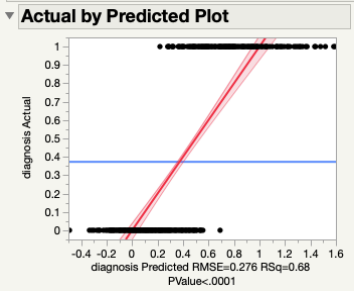
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.538645	0.167486	-9.19	<.0001*
radius_mean	0.3737258	0.081521	4.58	<.0001*
texture_mean	0.0216788	0.002866	7.56	<.0001*
perimeter_mean	-0.039122	0.012543	-3.12	0.0019*
area_mean	-0.000974	0.000209	-4.65	<.0001*
concave points_mean	9.1826038	0.900434	10.20	<.0001*

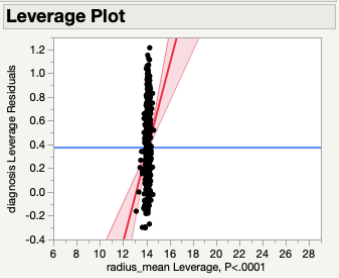
Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
radius_mean	1	1	1.6007255	21.0167	<.0001*
texture_mean	1	1	4.3576077	57.2131	<.0001*
perimeter_mean	1	1	0.7409794	9.7287	0.0019*
area_mean	1	1	1.6487698	21.6475	<.0001*
concave points_mean	1	1	7.9210171	103.9988	<.0001*

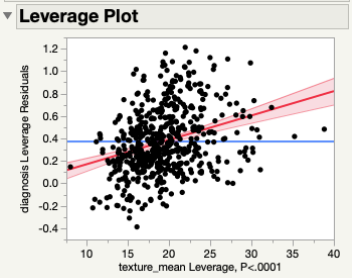
Actual by Predicted Plot



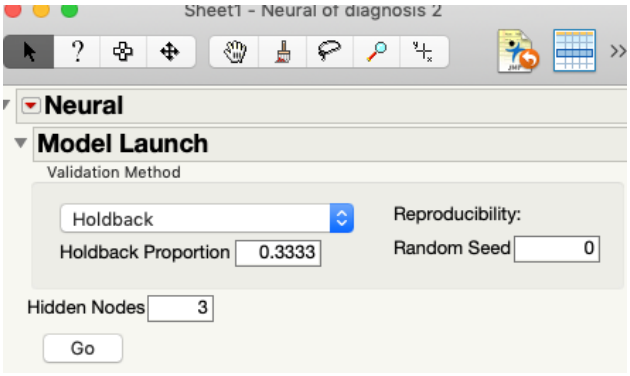
Leverage Plot

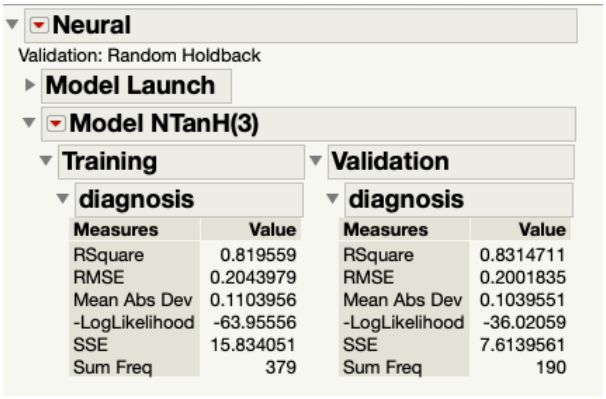


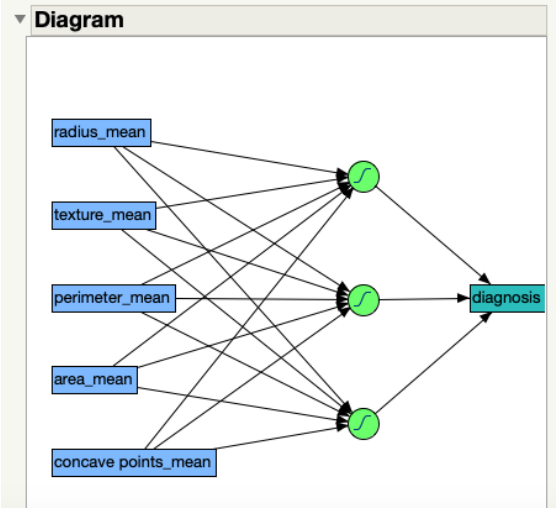
Leverage Plot

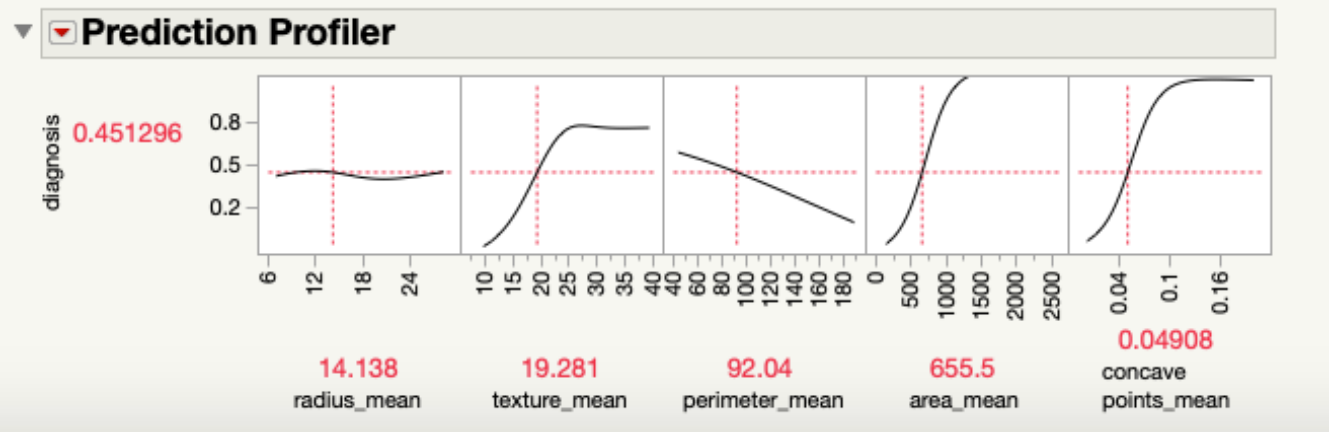


Neural Net Prediction Model:









Sheet1

	rea_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	Predicted diagnosis
1	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	0.9659019439
2	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.7161169726
3	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	1.1807409791
4	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.7854693401
5	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7546513346
6	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.5134316575
7	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.6978369207
8	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5340393158
9	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.737829179
10	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.6936391544
11	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.4549463323
12	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5179545547
13	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.8679295943
14	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.5714567589
15	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.6052431906
16	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.7986086471
17	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4717944805
18	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.8582827729
19	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.8770693196
20	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.292442586
21	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1474179404
22	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	-0.162419022
23	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.741925985
24	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.9494034224
25	904.6	0.1121	0.1457	0.1525	0.0917	0.1995	0.0633	0.8047844265
26	912.7	0.1186	0.2276	0.2229	0.1401	0.304	0.07413	1.0911836532
27	644.8	0.1054	0.1868	0.1425	0.08783	0.2252	0.06924	0.717240954
28	1094	0.0944	0.1066	0.149	0.07731	0.1697	0.05699	0.7314170105
29	732.4	0.1082	0.1697	0.1683	0.08751	0.1926	0.0654	0.7843434408
30	955.1	0.09847	0.1157	0.09875	0.07953	0.1739	0.06149	0.6165651047
31	1088	0.1064	0.1887	0.2319	0.1244	0.2183	0.06197	1.1551298027
32	440.6	0.1109	0.1516	0.1218	0.05182	0.2301	0.07799	0.3411169041
33	899.3	0.1197	0.1496	0.2417	0.1203	0.2248	0.06382	1.1911775927
34	1162	0.09401	0.1719	0.1657	0.07593	0.1853	0.06261	0.8330785937

20.92	25.09	143	1347	0.1099	0....	0....	0.1474	0.2...	0....	1.2...	1....	1.2706929104
21.56	22.39	142	1479	0.111	0....	0....	0.1389	0.1...	0....	1.2...	1....	1.28384661
20.13	28.25	131.2	1261	0.0978	0....	0....	0.09791	0.1...	0....	1.1...	1....	1.1349127424
16.6	28.08	108.3	858.1	0.08455	0....	0....	0.05302	0.159	0....	0.6...	0....	0.6880896606
20.6	29.33	140.1	1265	0.1178	0....	0....	0.152	0.2...	0....	1.4...	1....	1.4785810331
7.76	24.54	47.92	181	0.05263	0....	0	0	0.1...	0....	-0....	-0...	-0.15756125
14.127	19.29	91.97	359	•	•	•	0.04892	•	•	•	0....	0.6606481144
20.29	14.34	135.1	1297	•	•	•	0.14	•	•	•	0....	1.0920119199

Here, Last two rows are the values which we have predicted on the basis of the values of the decision variables.

As we can see in the screenshot above-
The first prediction depicts that the breast cancer is Benign or harmless in the current situation.

The second prediction depicts that the breast cancer is Malignant or Malicious in the current situation.

5. Benefits of the model

- This model helps in predicting the status or current effect of the cancer cells present in the patient's body.
- If the input values of the decision variables such as- Area, Smoothness, radius, texture, etc details of the cancer cells nuclei is present for diagnostic to a particular physician, it becomes helpful to have this predictive model which is based on the historical data.
- Although the model doesn't guarantee a defined decision, but it gives the fair idea of in which direction the treatment cycle should be run.
- As we know, Cancer treatments are very critical and requires timely attention towards the treatment and diagnosis. If there are such predictive modelling present for specific treatments, it's helps the physician to know whether there are chances of diagnosing the cancer in a particular patient based on the medical data-set.