

DESCRIPTIVE ANALYTICS AND **VISUALISATION**

**Model Building, Visualising and
Interpreting Predicted Probabilities &
Forecasting Sales**

CONFIDENTIAL

Name: Dhavala Sharma

Contents

INTRODUCTION	3
TASK 1	3
SUMMARY STATISTICS	3
OPEN ON SUNDAYS.....	3
ONLINE SALES CHANNEL	4
TASK 2	4
REGRESSION MODEL TO PREDICT SALES	4
TASK 2.2- ESTIMATING SALES BY REGRESSION ANALYSIS	4
FINAL MODEL	5
INTERPRETATIONS	6
TASK 3	6
STUDYING THE INTERACTION EFFECT.....	6
TASK 4	7
LOGISTIC REGRESSION MODEL TO PREDICT THE LIKELIHOOD OF OPENING ON SUNDAY.....	7
Final Model	7
TASK 5	8
PREDICTING ONLINECHANNEL USING LOGISTIC REGRESSION.....	8
PREDICTED PROBABILITIES (PP)	9
TASK 6	10
TIME-SERIES FORECASTING	10
Forecasted sales:.....	10
Conclusion.....	11
APPENDICES	12

INTRODUCTION

Mitch's superstore is one of the leaders in their sector in the Australian market. The company is doing well for the last two years. There have been regular cash in flows and out flows. There are constant sales. But the company is predicting a downward trend in the next five years. They have contacted BEAUTIFUL-DATA to help them understand their own business performance and components by using analytical tools. They want to understand their staff and their behavior towards new concepts and ideas. Random sample of 150 stores are taken to conduct this study.

TASK 1 SUMMARY STATISTICS

SALES (\$m)

- In the financial year 2016, Mitch's superstores had an average sale of 11.66 million dollars per year. The data derived by using sample of 150 grocery stores of Mitch's.
- Total sales were 1748.9 million dollars. 5.9 million dollars were the minimum sales recorded. Highest sale derived was approximately 23 million dollars.
- 8.1 million dollars is the mode value. It shows right skewed distribution as the mode value is lower than the average sales.
- This can be observed by the median value of 10.95 million dollars which is smaller than the mean value for sales, implying that the distribution of sales is right skewed.
- It can be concluded that distribution of sales is right skewed as our median value which is 10.95 million dollars. Median value is less than average mean sales value.
- 0.7 is the skewness of distribution. Data is moderately skewed because it falls in between 0.5 to 1.
- Range is the difference between the lowest and the highest income. It comes about to be 17.6 million dollars.
- Q3 is the highest 75% of the number of stores. They earned sales revenue of 14 million. On the contrary the lowest 25% also denoted by Q1 made 8.9 million.
- The value of two outliers are, lower outlier is 1 million and the upper outlier is 22 million dollars.

OPEN ON SUNDAYS

- Stores which are open on Sunday and which are not opened on Sunday are coded by 1 & 0.
- Total sample size selected was 150. 96 managers opened their store and 54 store managers were less interested in opening their store on a Sunday.

ONLINE SALES CHANNEL

- A store which is opening their grocery section electronically on an online platform were represented by Online sales channel variable. Grocery store which can operate on a online transaction platform are coded as 1 and left which can not are coded as 0.
- 41 percent of the total stores sell online whereas 59 percent didn't. Total random sample (100%) is 150 stores.

TASK 2

REGRESSION MODEL TO PREDICT SALES

- Dependent variables are affected by other independent variables. In our study, sales is a depended variable. In order to study the effect and causes of their relationship. We plot scatter diagram. The best part is, we can draw each in dependent variables relationship with independent variable. We did the same and came to know that each variable has a linear relationship with the dependent variable i.e. Sales
- In order to understand the strength of relationship between dependent(sales) and independent variables, we make correlation matrix. It can also tell us about the strength between independent variable.
- Findings form the matrix were
 - Time/age of the store & manager, hours of trading, no. of competitors have a negative relationship/effect on sales.
 - There is also a negative relationship between union percent & mng-train with sales respectively.
 - Rest all have a positive relationship/effect with sales
- The reason behind a strong relationship is the value of correlation. Closer value to 1 means a strong relationship between dependent (sales) and independent variables.
- High Multicollinearity of 0.8 is found as wages are vary highly correlated to total staff. wages and number of staffs also highly correlated with 0.92 multicollinearity
- To build a regression model I have to remove one the variables which had high correlation with independent variable, so I have removed wages.
- Basket:2016 and basket:2017 are having a high correlation
- Correlation between basket:2016 and sales is low. Regression model is free from that removed variable.

TASK 2.2- ESTIMATING SALES BY REGRESSION ANALYSIS

Following variable were used to create the regression model in order to predict sales in future.

- Mng-Gender
- Mng-Age
- Mng-Exp
- Basket:2017

- Mng-Train
- Union%
- OnlineChannel
- CarSpaces
- HrsTrading
- Adv.\$'000
- GrossProfit \$m
- No. Staff
- Age (Yrs)
- Competitors

Independent variables were removed which had high multicollinearity in between themselves and correlation with sales was low.

Overall model is significant. F-value comes out to be lower than 0.5. This tells us that at least one variable affects sales at a significance level of 5%. In order to know the significance of individual variables and individual variables relevance, we look at the P-value with 5% level of significance. Value exceeding 0.05 are considered insignificant variables. There is lack of evidence to support that these variables affect the sales. We can say this 5% level of significance. So while doing iterations we should remove the variable with the biggest P-value.

In the first iteration, Union% has the highest P-value = 0.92 > 0.05. Therefore, we have removed this variable.

- Iteration 1: P-value of Union% is highest (0.92 > 0.05), hence removed.
- Iteration 2: P-value of competitors is highest, hence removed (0.64 > 0.05)
- Iteration 3: P-value of onlinechannel is highest, hence removed (0.62 > 0.05)
- Iteration 4: P-value of Mng-train is highest, hence removed (0.54 > 0.05)
- Iteration 5: P-value of carspaces is highest, hence removed (0.52 > 0.05)
- Iteration 6: P-value of mng-gender is highest, hence removed (0.31 > 0.05)
- Iteration 7: P-value of Sundays is highest, hence removed
- Iteration 8: P-value of age is highest, hence removed

And regression analysis was conducted again

Final model was created at the 8th iteration. All variables are significant to predict sales.

FINAL MODEL

$$\text{Sales} = -28.35 + 0.02 * (\text{No. staff}) + 0.25 * (\text{GrossProfit}) + 0.04 * (\text{Adv. \$1000}) + 0.02 * (\text{HrsTrading}) - 0.02 * (\text{Mng-Age}) + 0.06 * (\text{Mng-Exp}) + 0.18 * (\text{Basket:2017})$$

R square (R^2) is equal to ninety five percent sales variation which can be illustrated by independent variables. It shows that model is good to fit. Other variables which are not in the model account for five percent variation. Keeping the same sample size, the independent variables can explain 95 percent of the variability in sales which is *Adjusted R Square*. Variables

which are independent in the model and maintains a linear relationship with sales are *Residual*. In addition, the residual plots indicate that the X-axis disperses well all independent variables.

INTERPRETATIONS

- I. Intercept: There will be reduction in sales by 28350000 dollars. It has no practical significance.
- II. No.staff: There will be addition in sales by 20,000 Dollars if we increase the staff by one, keeping others the same.
- III. Adv.\$1000: There will be addition in sales by 40000 dollars, if we increase advertising by 1000 dollars.
- IV. Mng-Exp: There will be addition in sales by 20000 Dollars, if we increase experience by one year, keeping everything same.
- V. GrossProfit: There will be addition in sales by 250000 Dollars, if we increase it by one unit, keeping others the same.
- VI. Mng-Age: There will be addition in sales by 20000 Dollars, if we increase age by one year, keeping everything same.
- VII. Basket:2017: There will be addition in sales by 180000 dollars, if the cost of basket increases by one dollar for 2017 financial year, keeping everything similar.

TASK 3

STUDYING THE INTERACTION EFFECT

- Independent variable is competitors and dependent is Sales. They both are related to each other. Sunday is moderator. In this study we have studied the relationship between number of competitors in the market, stores which are opening up on Sundays and their respective sales. I have come to the conclusion that there is enough evidence present to say that the model is significant as it can be re assured by seeing the ANOVA table. Our P-value is less than 0.05. thus, we can say this five percent level of significance that the model is significant. We reject H_0 .
- We can also conclude that interaction exist. We can say this five percent of significance level as P-value is less than 0.05 (independent variables including interaction term which are used in the model are significant).
- When there is less competition in the market. Average sales of stores opening on Sundays are higher than those stores which are not opening. On the contrary when there is high competition the situation is vice-versa.
- There is a negative effect of competitors on sales of stores that opens on Sunday. Contrary to that there is a positive effect of competitors on sales of stores that do not opens up on Sunday.

All in all, sales and no. of competitors interact with Sunday. Their relationship is significantly positive with stores that are not open on Sundays. As there are more competitors in the market now, it is advised not to open stores on Sunday.

TASK 4

LOGISTIC REGRESSION MODEL TO PREDICT THE LIKELIHOOD OF OPENING ON SUNDAY

Wages and advertising were excluded from the model as they have high multicollinearity. The model was created to estimate the probability of opening up stores on a Sunday. Excluding wages and advertising, rest all variables were included in the model. Various iteration was involved in order to remove insignificant P-value, in order to reach final model.

Final Model

$$\text{Sunday} = 0.23 + 0.04 * \text{No.Staff} + 0.63 * \text{GrossProfit \$} + 0.35 * \text{competitors} - 1.53 * \text{Onlinechannel} - 0.10 * \text{Mng-Age} + 0.11 * \text{Mng-Exp}.$$

Classification accuracy came out to be 87.5% because 84 stores out of total of 96 stores that were opening up on Sunday were accurately classified where the remaining 12 stores were misclassified by logistic regression. In case of stores that were not opening up on Sunday, which were 54 in total. 26 were accurately and the rest 28 were inaccurately predicted. Therefore, this time the classification accuracy was 48.1%.

In conclusion, 73% superstores were accurately classified. This states that the hit ratio (all in all classification accuracy) is 73%. In order to understand the stores which were misclassified, we need to independent variables.

50% cannot be cut off by us as the findings/observation were unequal. In order to supplement it, logistic regression was added with *standard hit ratio* and *proportional chance criteria* which is also known as *PCC*. Ratios were 67.4% and 54% respectively. From the findings it is clear that logistic regression model is better than random model. Hit ratio is greater than PCC. Hit ratio is also greater than standard hit ratio. Therefore, it can be said that logistic regression is better in classifying observation because of its practicality.

The $LL1 < LL0$, implying that we have improved the final model.

P-value associated with Chi-square is less than 0.05. Therefore, it can be concluded that at 5% significance level, there is sufficient evidence that logistic regression model has some predictive power.

- LL1<LL0
 - final model is better now.
 - P-value chi-square<0.05
 - Model has predictive power.
- Lemeshow's R-square
 - Regression model can explain 12.87 percent change in Sunday
- R-Sq (CS)
 - 15.48% of the variation can be explained through regression model.
- R-Sq (N) (Nagelkarke R-square)
 - Variation of upto 21.22% in Sunday can be described through regression model.
- Probability of a store opening up on a Sunday increases by 82.88% if we increase one person in our staff
- Probability of a store opening up on a Sunday increases by 3.72% if we increase one unit in gross profit
- Where as the probability or likelihood of store opening up on Sunday only increases by 42.12% if there is a increase in competitors by one.
- Onlinechannel opening decreases the likelihood of opening shops on Sunday by 78.26 percent compared to no onlinechannel.
- Probability falls by 9.17% of opening a store on Sunday if we increase the age of manager.
- There is an addition in probability by 12.02% of opening a store on Sunday if managers experience is increased.

ROC curve and the diagonal are near to each other or I should they is not much distance between them. showing the capacity of the model to discriminate between success and failure is due to chance. Success is opening the store on Sunday and failure is not opening a store on Sunday.

TASK 5

PREDICTING ONLINECHANNEL USING LOGISTIC REGRESSION.

$$\text{Onlinechannel} = 0.75 - 0.13*(\text{Mng-Age}) + 0.38*(\text{Mng-Exp}) + 0.97*(\text{Mng-Gender})$$

- 62 stores were there in total who opened an online channel. 43 out of them were correctly predicted/classified where as the rest were misclassified by the regression model.
- 88 stores didn't open an online channel. 72 stores out of the total were rightly predicted and the rest 16 stores were not rightly predicted.
- Regression model can rightly explain the variation up to 76%(hit ratio = 76 %)

- We can not cut off 50% because our findings are unequal. Proportion chance criteria is 51.5% and standard hit ratio is 64.37%.
- Accuracy hit ratio > PCC
- Accuracy hit ratio > standard ratio
- Regression model is better and significant.

- LL1 < LLO
 - Model is enhanced
 - P-value chi-square < 0.05
 - Model has some predicting ability

- R-Sq (L) (Lemeshow's R-square)
 - Variation up to 32.83% can be explained.

- R-Sq (cs) (Cox and R-snell)
 - Variation up to, 35.92% can be explained.

- R-Sq (N) (Nagelkerke R-square)
 - Variation up to, 48.39% can be explained

- Probability decreases by 12.12% of starting a online channel if age is managers is increased.
- Likelihood increases by 45.64% on starting a online platform if the experience of the manager is increased.
- When it comes to gender, if there is male manager then the probability increases by 163.69% when compared with female manager.

In conclusion, model is moderately powered. This is due to the reason because the model is moderately distanced from diagonal.

TASK 5.2

PREDICTED PROBABILITIES (PP)

As the experience improves the likelihood of shops opening an online internet channel also improves for both females and males. But the age is in a complete opposite relationship. As the age increases the chances of opening falls. So, In order to solve this issue, managers should be told to open up a online channel when they are in their thirties. This implies for both male and female managers. For men, there is a declining rate of rise in the predicted likelihood relative to

women. All in all, we have founded that men and women have probability which is less than ninety percent.

TASK 6

TIME-SERIES FORECASTING

The model was created in order to predict sales for the financial year 2019 (Q2, Q3 & Q4) and for the financial year 2020 (Q1) of Mitchies Superstore. The model was designed to predict sales for four quarters. The data which we have has irregularities, seasonal & trend components. So, we need to use time series multiplication to overcome it.

Y	T X C X I X S
T	Trend component
C	Cyclical component
I	Irregular component
S	Seasonal Component.

- 4 moving averages are taken as we have data for four quarters. We have done this to eliminate seasonality component. Roughness of data was eradicated.
- Acquired values were then used to quantify the element of seasonality by ratio. Ratio had irregularities and variation seasonally. To remove that, indices were estimated. This was possible by using the quarter table from ratio.
- In order to normalize, de-seasonalizing was performed.
- After multiplying trend and index, we derive forecasting

Forecasted sales:

2019	2	121.010764
	3	139.531682
	4	162.84601
2020	1	114.444326

- Variation in percentage between forecasted value and observed value is called Absolute percentage.
- Mean Absolute Percent Error = MAPE
 - Is illustrates percentage of error as accuracy.
 - Five percent was the historical mape.

Conclusion

Broadly speaking, the information set in question enables us to predict the sales model, open shops on Sunday and open an online shopping platform.

Space room error is low. Sales could be predicted quite rightly. Independent variables were mostly used.

Quality variable which also effect the probability of opening up a store on Sunday should be added in order to enhance predictive power.

Online channels model had a nicer or better model then Sunday if compared with their predictive powers. We can enhance it more by adding useful variables which effects.

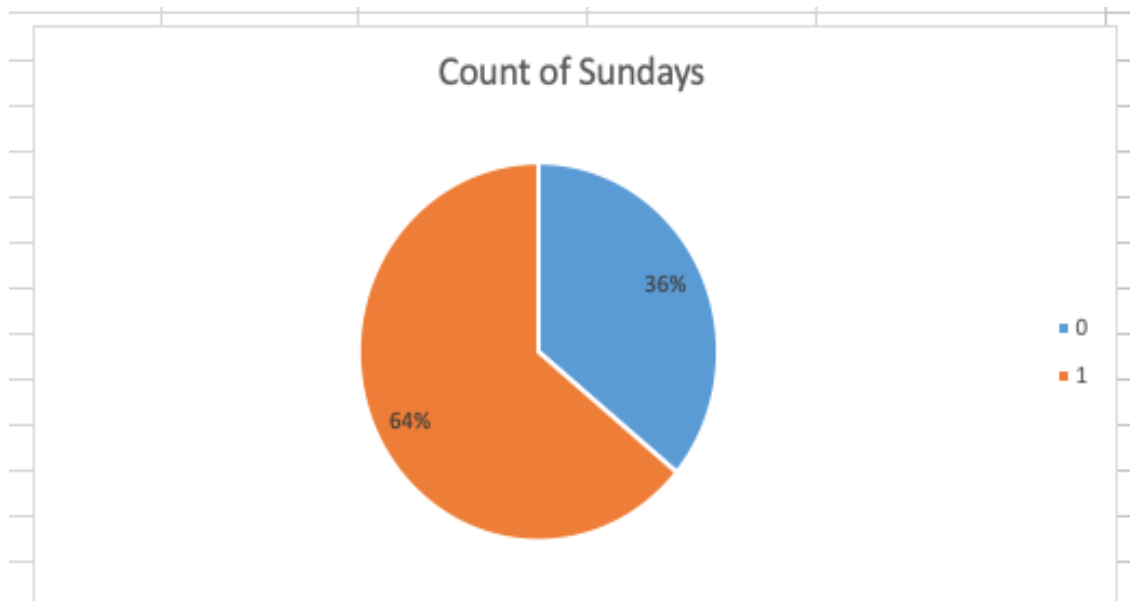
We have limited data set in order to predict sales for future. Probability of opening up a store and opening up a online channel will get effected if we add more useful/relevant variables.

APPENDICES

TASK 1 – SUMMARY STATISTICS

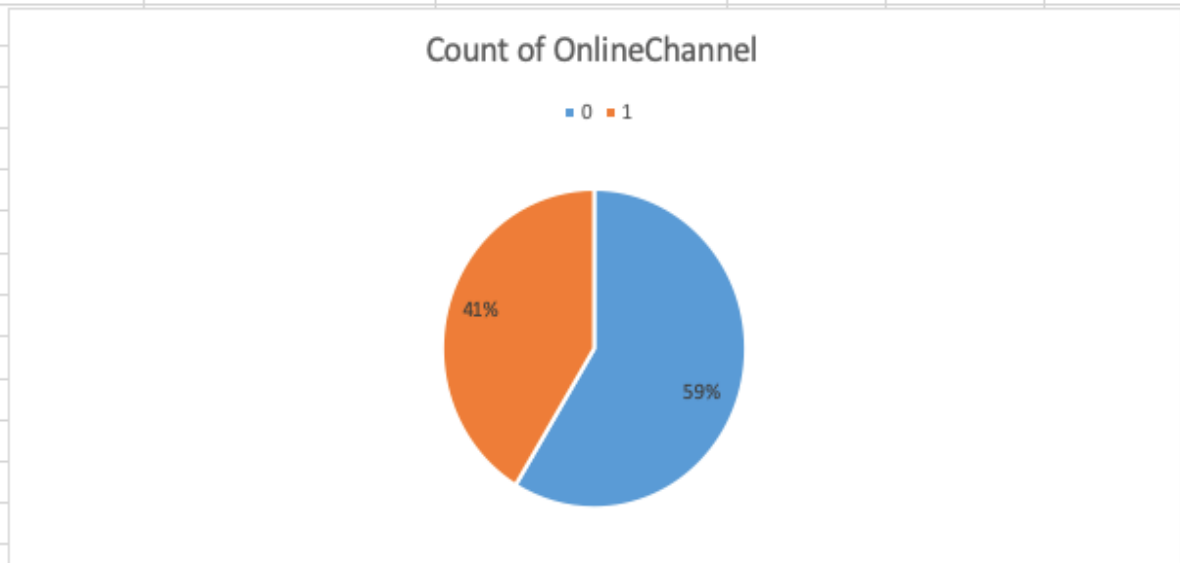
<i>Sales \$m</i>	
Mean	\$11.66
Standard Error	0.29156999
Median	10.95
Mode	8.1
Standard Deviation	3.57098856
Sample Variance	12.7519593
Kurtosis	0.00774116
Skewness	0.74479958
Range	17.6
Minimum	5.9
Maximum	23.5
Sum	1748.9
Count	150
Confidence Level(95.0%)	0.57614615
Quartile 1	8.825
Quartile 3	14
IQR	5.175
outliers (lower limit)	1.0625
Upper limit	21.7625

Row Labels	Count of Sundays
0	54
1	96
Grand Total	150
<i>Sundays</i>	
Mean	0.64
Standard Error	0.039323132
Median	1
Mode	1
Standard Deviat	0.481608045
Sample Variance	0.231946309
Kurtosis	-1.675312111
Skewness	-0.5892422
Range	1
Minimum	0
Maximum	1
Sum	96
Count	150
Confidence Level	0.077703027

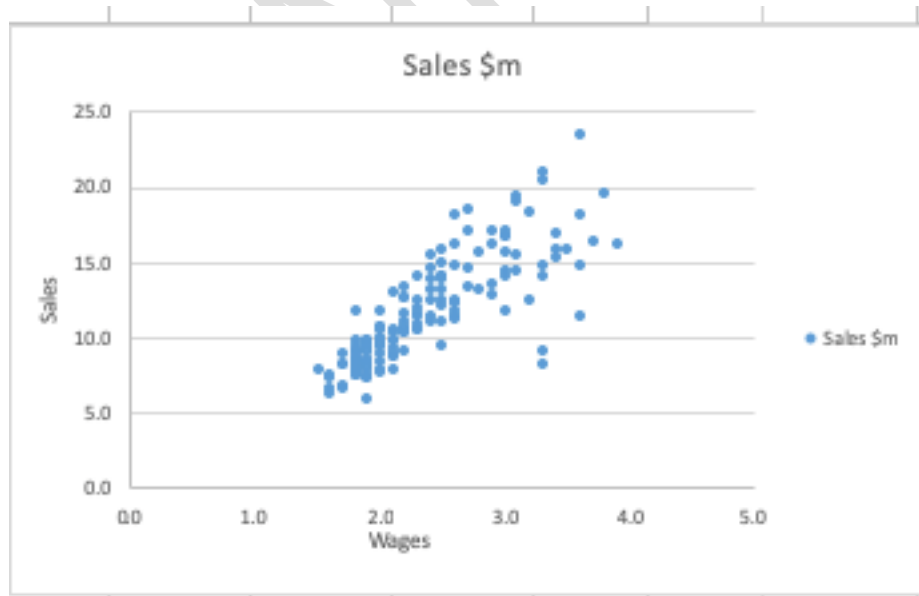


OnlineChannel	
Mean	0.41333333
Standard Error	0.04034157
Median	0
Mode	0
Standard Deviation	0.4940813
Sample Variance	0.24411633
Kurtosis	-1.89907921
Skewness	0.35556026
Range	1
Minimum	0
Maximum	1
Sum	62
Count	150
Confidence Level(95	0.07971547

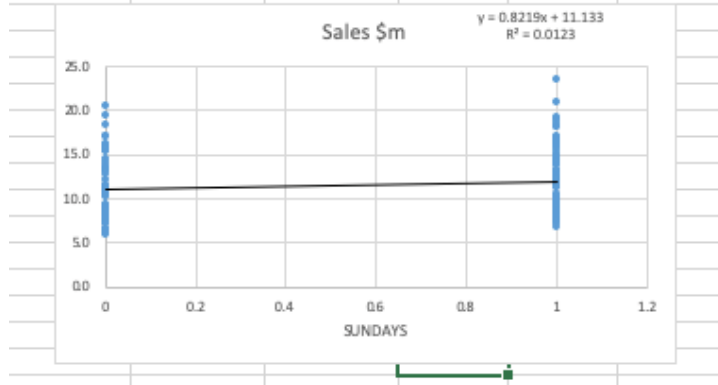
Row Labels	Count of OnlineChannel			
0	88			
1	62			
Grand Total	150			



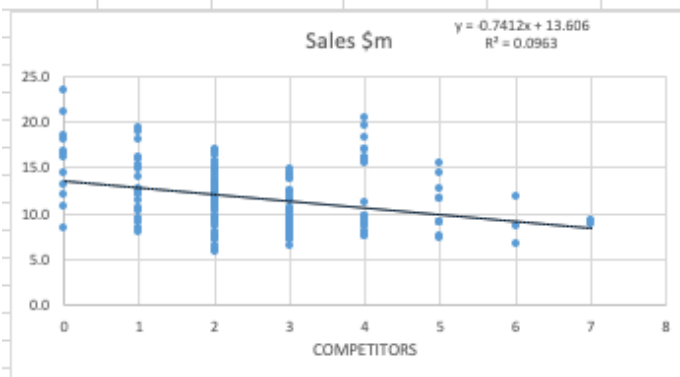
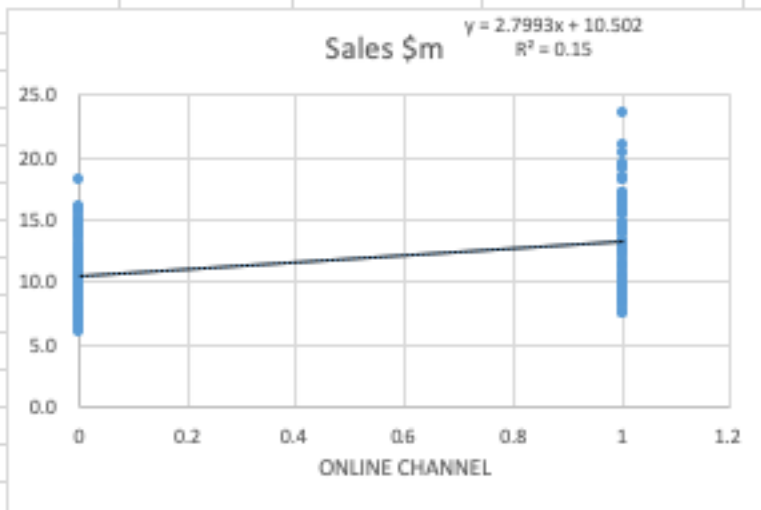
TASK 2 – SCATTER PLOT



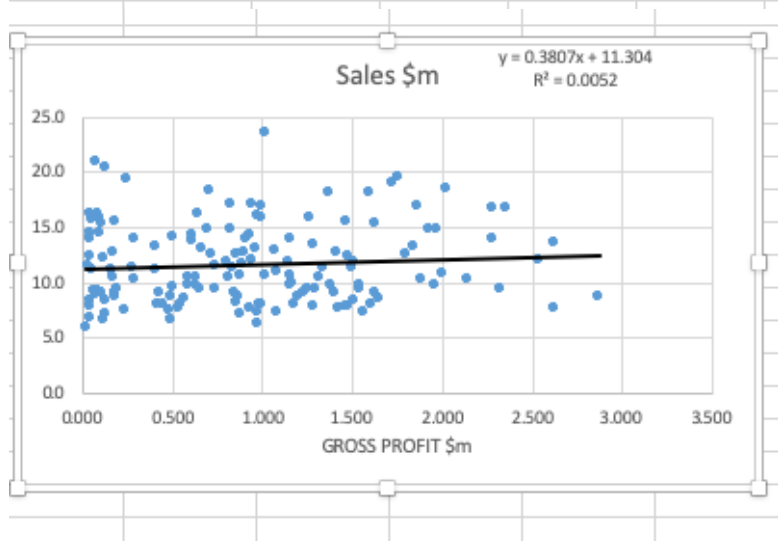
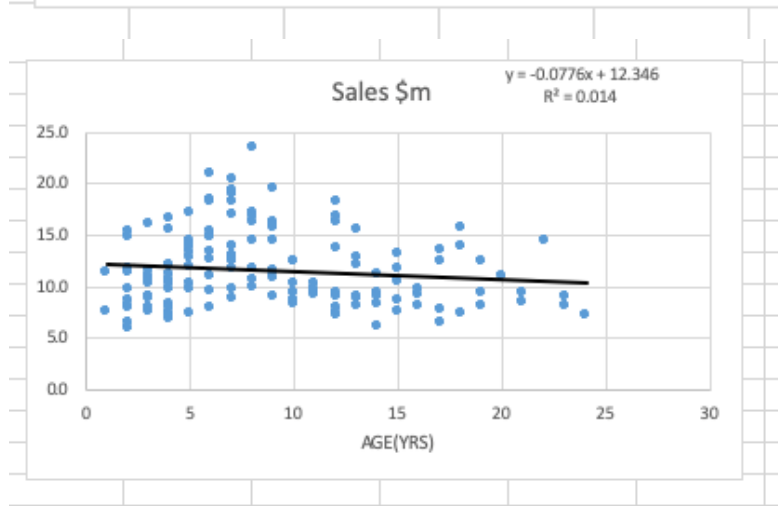
Row Labels	Sum of Sales \$m
0	601.2
1	1147.7
(blank)	
Grand Total	1748.9

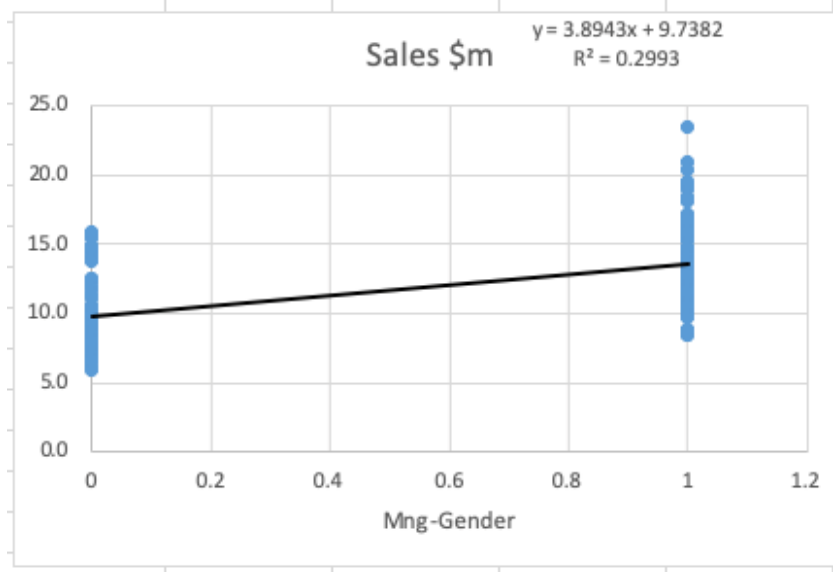
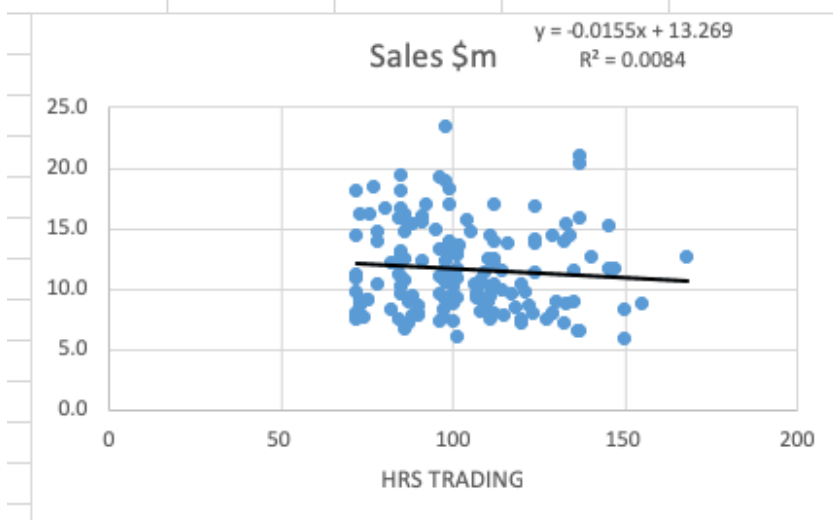
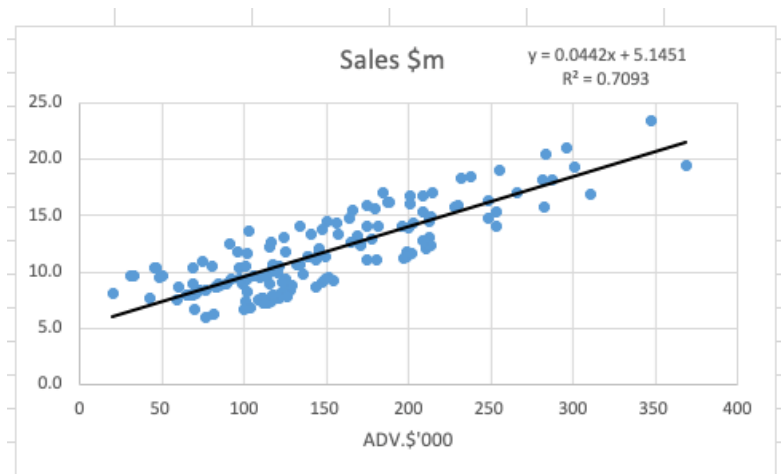


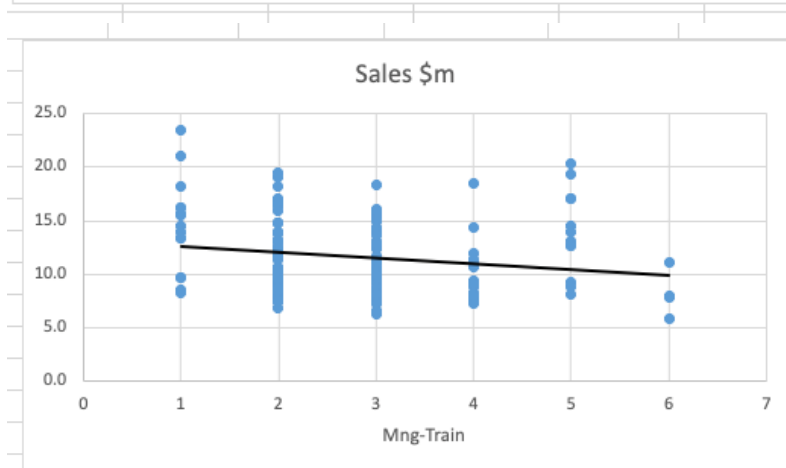
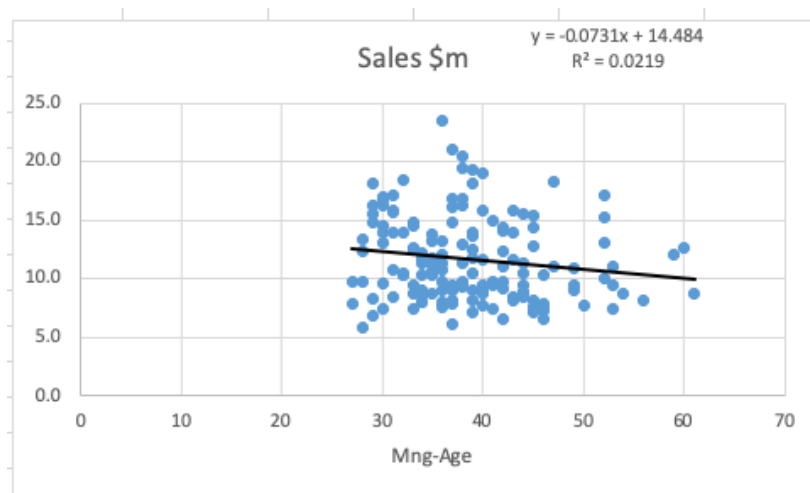
Row Labels	Sum of Sales \$m
0	924.2
1	824.7
Grand Total	1748.9



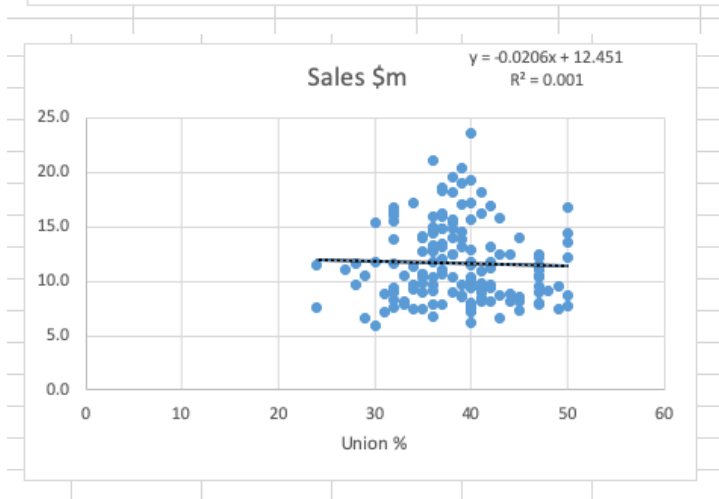
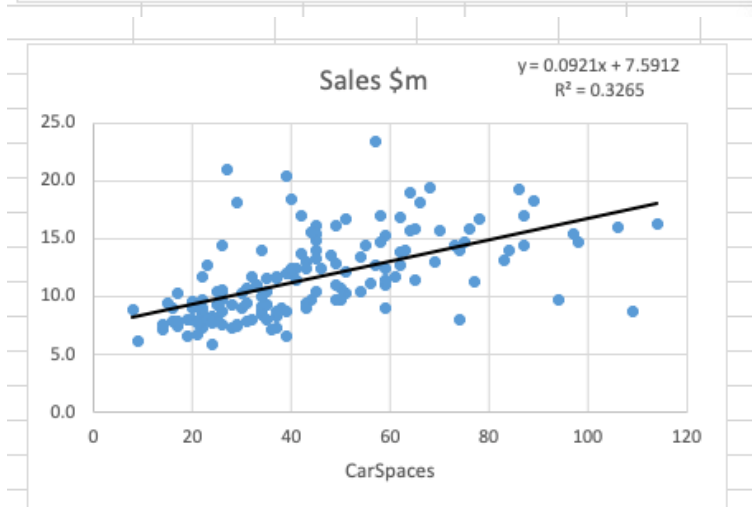
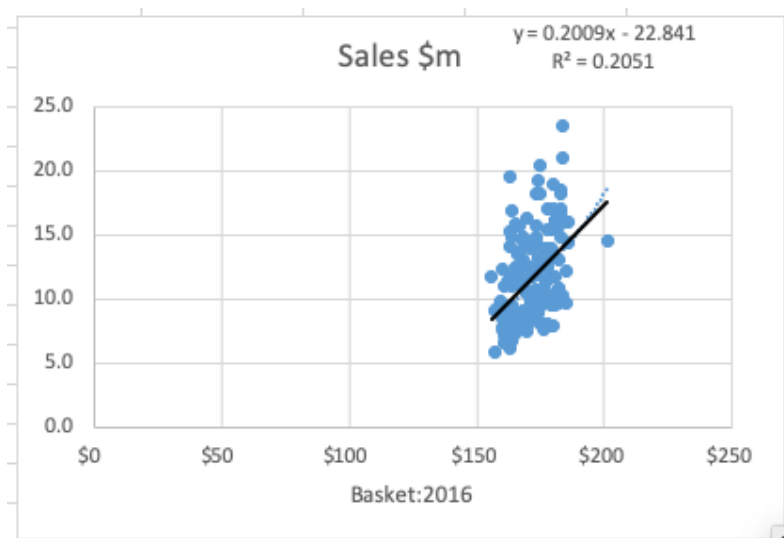
Row Labels	Sum of Sales \$m
0	190.1
1	258.2
2	479.2
3	372.4
4	295.6
5	107.9
6	27.3
7	18.2
Grand Total	1748.9

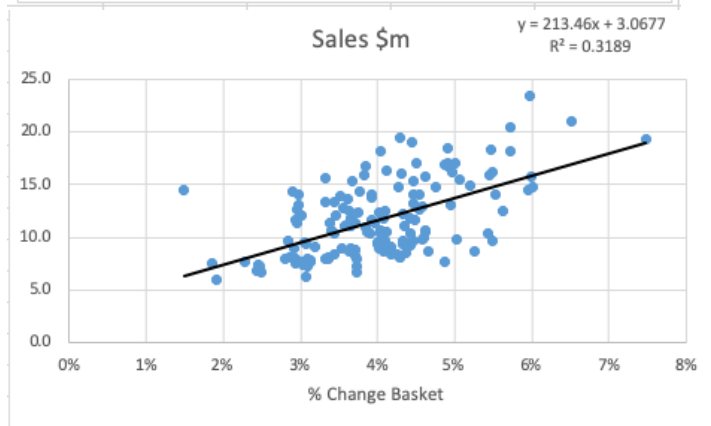
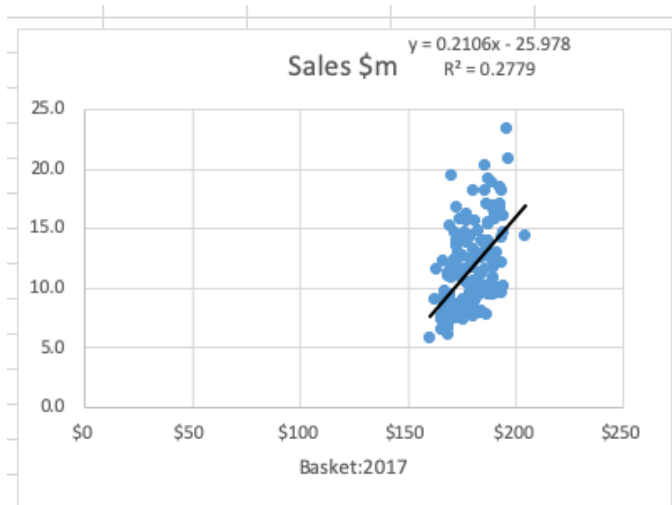






Row Labels	Sum of Sales \$m
1	217.8
2	615.8
3	536.5
4	179
5	167.1
6	32.7
Grand Total	1748.9





TASK 2.1 CORRELATION MATRIX

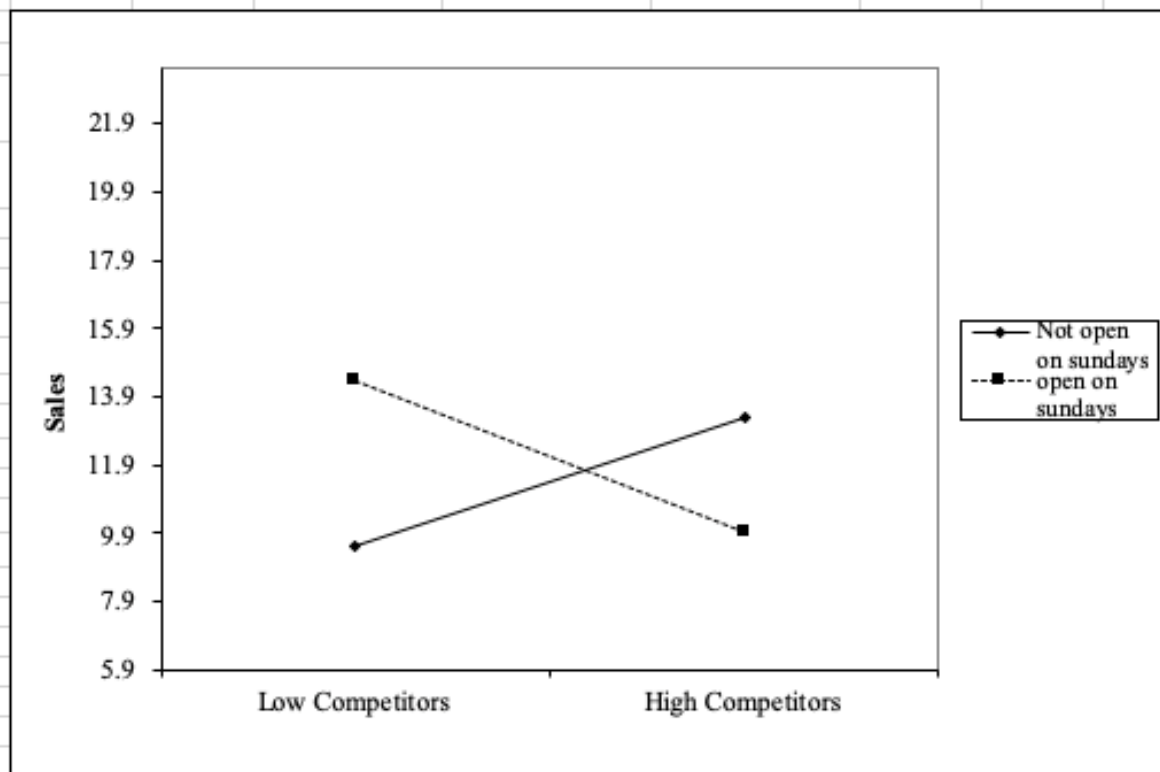
	Wages \$m	No. Staff	Age (Yrs)	GrossProfit \$m	Adv.\$'000	Competitors	HrsTrading	Sundays	Mng-Gender	Mng-Age	Mng-Exp	Mng-Train	Union%	CarSpaces	OnlineChannel	Basket:2016	Basket:2017	%ChangeBasket	Sales \$m
Wages \$m	1																		
No. Staff	0.92	1.00																	
Age (Yrs)	-0.06	-0.03	1.00																
GrossProfit \$m	0.08	0.00	-0.08	1.00															
Adv.\$'000	0.80	0.74	-0.01	0.04	1.00														
Competitors	-0.23	-0.21	0.03	-0.28	-0.20	1.00													
HrsTrading	-0.12	-0.03	-0.07	-0.46	-0.09	0.37	1.00												
Sundays	0.12	0.12	0.01	0.09	0.08	0.12	-0.02	1.00											
Mng-Gender	0.44	0.42	-0.16	0.11	0.46	-0.24	-0.14	-0.04	1.00										
Mng-Age	-0.02	0.05	0.11	0.01	0.02	0.04	0.11	-0.17	-0.13	1.00									
Mng-Exp	0.26	0.23	-0.09	-0.09	0.37	-0.06	0.10	0.05	0.28	0.05	1.00								
Mng-Train	-0.08	-0.11	0.05	-0.06	-0.03	0.13	0.07	-0.11	-0.14	0.47	-0.22	1.00							
Union%	-0.04	-0.11	0.62	0.24	-0.04	-0.12	-0.23	-0.02	-0.01	-0.03	-0.09	-0.02	1.00						
CarSpaces	0.70	0.76	-0.05	0.02	0.54	-0.17	-0.12	0.11	0.37	-0.02	0.18	-0.12	-0.12	1.00					
OnlineChannel	0.25	0.25	-0.02	-0.03	0.32	-0.11	0.01	-0.10	0.34	-0.27	0.50	-0.38	0.04	0.28	1.00				
Basket:2016	0.21	0.14	-0.24	0.12	-0.01	-0.33	-0.25	0.10	0.29	-0.31	0.28	-0.22	0.03	0.16	0.14	1.00			
Basket:2017	0.28	0.21	-0.25	0.10	0.07	-0.34	-0.22	0.11	0.31	-0.30	0.32	-0.24	0.00	0.22	0.19	0.98	1.00		
%ChangeBasket	0.45	0.43	-0.10	-0.05	0.41	-0.15	0.04	0.09	0.26	-0.07	0.37	-0.19	-0.12	0.36	0.30	0.27	0.43	1.00	
Sales \$m	0.81	0.74	-0.12	0.07	0.84	-0.31	-0.09	0.11	0.55	-0.15	0.50	-0.18	-0.03	0.57	0.39	0.45	0.53	0.56	1.00

TASK 2.2 FINAL MODEL (REGRESSION ANALYSIS)

SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.97							
R Square	0.95							
Adjusted R Sq	0.95							
Standard Error	0.83							
Observations	150.00							
<i>ANOVA</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	7.00	1801.68	257.38	371.58	0.00			
Residual	142.00	98.36	0.69					
Total	149.00	1900.04						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	-28.35	1.87	-15.13	0.00	-32.06	-24.65	-32.06	-24.65
No. Staff	0.02	0.01	3.32	0.00	0.01	0.04	0.01	0.04
GrossProfit \$r	0.25	0.11	2.23	0.03	0.03	0.48	0.03	0.48
Adv.\$'000	0.04	0.00	22.67	0.00	0.03	0.04	0.03	0.04
HrsTrading	0.02	0.00	4.20	0.00	0.01	0.02	0.01	0.02
Mng-Age	-0.02	0.01	-2.33	0.02	-0.04	0.00	-0.04	0.00
Mng-Exp	0.06	0.02	2.72	0.01	0.02	0.10	0.02	0.10
Basket:2017	0.18	0.01	18.49	0.00	0.16	0.19	0.16	0.19

TASK 3 – INTERACTION EFFECT

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.60							
R Square	0.36							
Adjusted R Square	0.35							
Standard Error	2.89							
Observations	150.00							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	3.00	682.24	227.41	27.26	0.00			
Residual	146.00	1217.81	8.34					
Total	149.00	1900.04						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	8.15	0.85	9.55	0.00	6.47	9.84	6.47	9.84
Competitors	1.25	0.32	3.93	0.00	0.62	1.87	0.62	1.87
Sundays	7.86	1.04	7.58	0.00	5.81	9.91	5.81	9.91
Comp*Sundays	-2.72	0.37	-7.40	0.00	-3.44	-1.99	-3.44	-1.99



TASK 4 - CORRELATION MATRIX

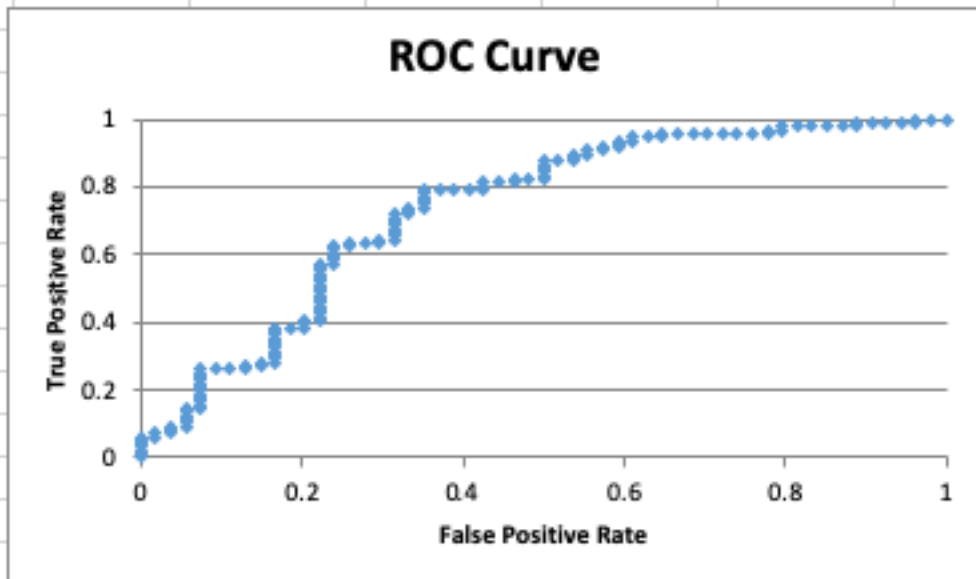
	Wages \$m	No. Staff	Age (Yrs)	GrossProfit \$m	Adv.S'000	Competitors	HrsTrading	OnlineChannel	Mng-Gender	Mng-Age	Mng-Exp	Mng-Train	Union%	CarSpaces	Basket:2016	Basket:2017	%ChangeBaske	Sales \$m	Sundays
Wages \$m	1.00																		
No. Staff	0.92	1.00																	
Age (Yrs)	-0.06	-0.03	1.00																
GrossProfit \$m	0.08	0.00	-0.08	1.00															
Adv.S'000	0.80	0.74	-0.01	0.04	1.00														
Competitors	-0.23	-0.21	0.03	-0.28	-0.20	1.00													
HrsTrading	-0.12	-0.03	-0.07	-0.46	-0.09	0.37	1.00												
OnlineChannel	0.25	0.25	-0.02	-0.03	0.32	-0.11	0.01	1.00											
Mng-Gender	0.44	0.42	-0.16	0.11	0.46	-0.24	-0.14	0.34	1.00										
Mng-Age	-0.02	0.05	0.11	0.01	0.02	0.04	0.11	-0.27	-0.13	1.00									
Mng-Exp	0.26	0.23	-0.09	-0.09	0.37	-0.06	0.10	0.50	0.28	0.05	1.00								
Mng-Train	-0.08	-0.11	0.05	-0.06	-0.03	0.13	0.07	-0.38	-0.14	0.47	-0.22	1.00							
Union%	-0.04	-0.11	0.62	0.24	-0.04	-0.12	-0.23	0.04	-0.01	-0.03	-0.09	-0.02	1.00						
CarSpaces	0.70	0.76	-0.05	0.02	0.54	-0.17	-0.12	0.28	0.37	-0.02	0.18	-0.12	-0.12	1.00					
Basket:2016	0.21	0.14	-0.24	0.12	-0.01	-0.33	-0.25	0.14	0.29	-0.31	0.28	-0.22	0.03	0.16	1.00				
Basket:2017	0.28	0.21	-0.25	0.10	0.07	-0.34	-0.22	0.19	0.31	-0.30	0.32	-0.24	0.00	0.22	0.98	1.00			
%ChangeBasket	0.45	0.43	-0.10	-0.05	0.41	-0.15	0.04	0.30	0.26	-0.07	0.37	-0.19	-0.12	0.36	0.27	0.43	1.00		
Sales \$m	0.81	0.74	-0.12	0.07	0.84	-0.31	-0.09	0.39	0.55	-0.15	0.50	-0.18	-0.03	0.57	0.45	0.53	0.56	1.00	
Sundays	0.12	0.12	0.01	0.09	0.08	0.12	-0.02	-0.10	-0.04	-0.17	0.05	-0.11	-0.02	0.11	0.10	0.11	0.09	0.11	1.00

TASK 4.1 – FINAL MODEL (LOGISTIC REGRESSION TO PREDICT SUNDAY)

	coeff b	s.e.	Wald	p-value	exp(b)	lower	upper	change in odds
Intercept	0.23	1.42	0.03	0.87	1.26			
No. Staff	0.04	0.01	6.09	0.01	1.04	1.01	1.07	3.71787348
GrossProfit \$m	0.63	0.32	3.99	0.05	1.88	1.01	3.50	88.283441
Competitors	0.35	0.14	5.92	0.01	1.42	1.07	1.89	42.1182834
OnlineChannel	-1.53	0.50	9.35	0.00	0.22	0.08	0.58	-78.259876
Mng-Age	-0.10	0.03	11.21	0.00	0.91	0.86	0.96	-9.173366
Mng-Exp	0.11	0.06	4.21	0.04	1.12	1.01	1.25	12.0230506

Classification Table			
	Suc-Obs	Fail-Obs	
Suc-Pred	84	28	112
Fail-Pred	12	26	38
	96	54	150
Accuracy	0.875	0.48148148	0.73333333
Cutoff	0.5		
PCC	0.64		
MCC	0.5392		
Standard	0.674		

LL0	-98.012729
LL1	-85.397896
Chi-Sq	25.2296671
df	6
p-value	0.00030958
alpha	0.05
sig	yes
R-Sq (L)	0.12870607
R-Sq (CS)	0.15481334
R-Sq (N)	0.21226935
Hosmer	158.066584
df	148
p-value	0.27063552
alpha	0.05
sig	no

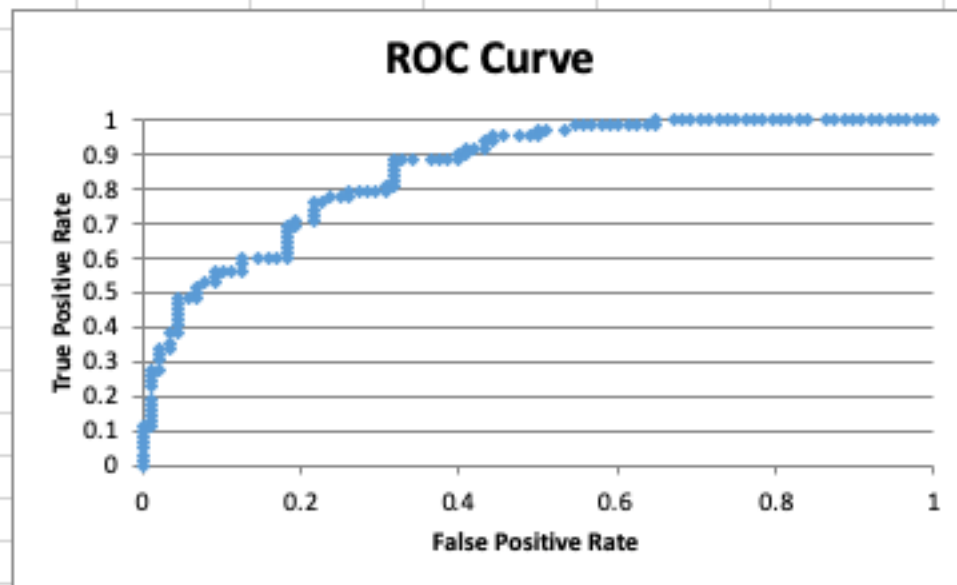


TASK 5.1- FINAL MODEL (LOSISTIC REGRESSION TO PREDICT ONLINECHANNEL)

	<i>coeff b</i>	<i>s.e.</i>	<i>Wald</i>	<i>p-value</i>	<i>exp(b)</i>	<i>lower</i>	<i>upper</i>	<i>% change in odds</i>
Intercept	0.42	1.30	0.10	0.75	1.52			
Mng-Age	-0.13	0.04	13.35	0.00	0.88	0.82	0.94	-12.1200134
Mng-Exp	0.38	0.07	25.95	0.00	1.46	1.26	1.68	45.64571404
Mng-Gender	0.97	0.43	5.13	0.02	2.64	1.14	6.10	163.6962773

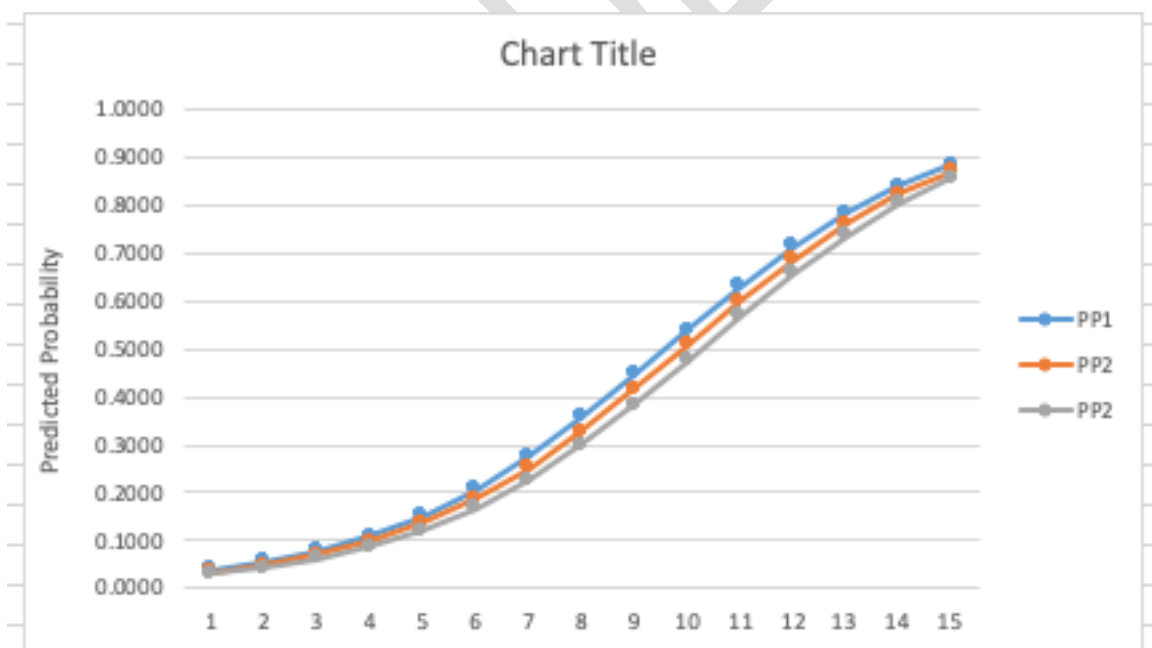
LL0	-101.70732
LL1	-68.323436
Chi-Sq	66.7677723
df	3
p-value	2.0997E-14
alpha	0.05
sig	yes
R-Sq (L)	0.32823483
R-Sq (CS)	0.35925165
R-Sq (N)	0.4839477
Hosmer	120.76885
df	135
p-value	0.80443757
alpha	0.05
sig	no

Classification Table			
	Suc-Obs	Fail-Obs	
Suc-Pred	43	16	59
Fail-Pred	19	72	91
	62	88	150
Accuracy	0.693548387	0.81818182	0.76666667
Cutoff	0.5		
MCC	0.413333333		
PCC	0.515022222		
Standard	0.643777778		

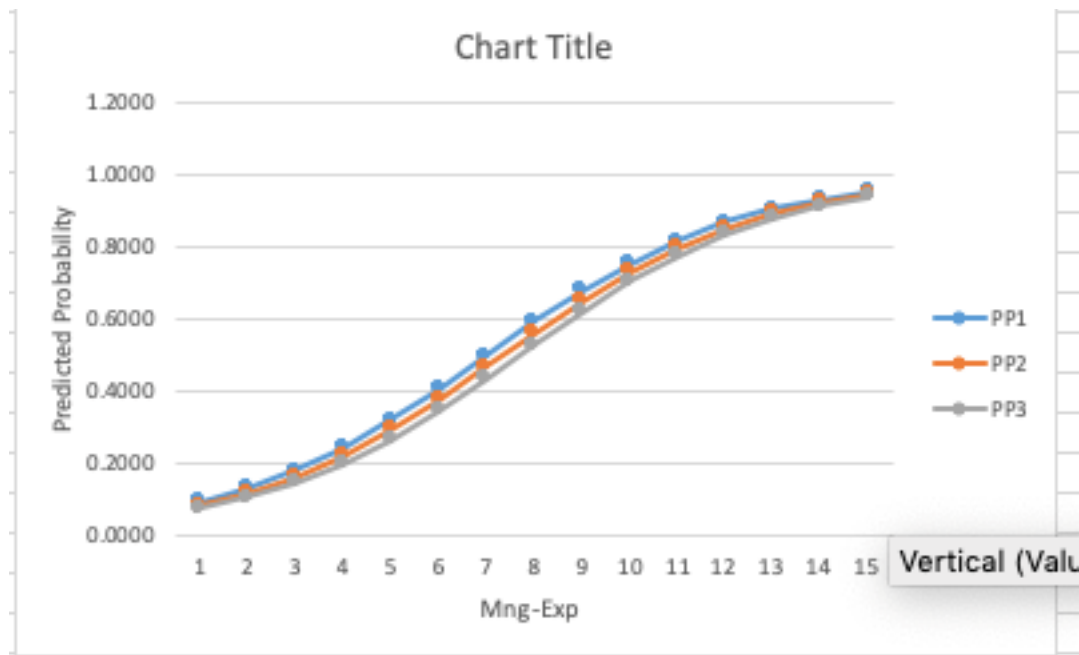


TASK 5.2- PREDICTED PROBABILITIES

FEMALES



MALES



TASK 6- TME-SERIES FORECASTING

Time Period	Year	Quarter	Sales (\$ million)	4 MA	4 Centered MA	Ratio (Obs/MA)	Index	Deseasonalised	Trend	Forecast	Observed	APE		
1	2015	2	60.9				0.94331835	64.601733	54.634		60.9	0.15429513	51.5372546	
2		3	64.8				1.0500183	61.70368623	59.237		64.8	0.03997632	62.1999343	
3		4	72.7	64.0	64.7	1.123194562	1.18443799	61.38776431	63.84		72.7	0.03994665	75.6145211	
4	2016	1	57.4	65.5	67.0	0.856615385	0.8054298	71.29112909	68.443		57.4	0.03995068	55.126032	
5		2	67.1	68.6	71.2	0.942995169	0.94331835	71.13187207	73.046		67.1	0.02690957	68.905632	
6		3	77.0	73.8	74.7	1.030548399	1.0500183	73.33205494	77.649		77.0	0.05886846	81.5328713	
7		4	93.5	75.7	77.9	1.200564972	1.18443799	78.94039283	82.252		93.5	0.04195073	97.4223933	
8	2017	1	65.1	80.1	82.6	0.788807462	0.8054298	80.85124218	86.855		65.1	0.07425684	69.9556055	
9		2	84.7	85.0	88.3	0.95890411	0.94331835	89.78941229	91.458		84.7	0.01858335	86.2740094	
10		3	96.8	91.6	93.5	1.035751067	1.0500183	92.18886907	96.061		96.8	0.04200215	100.865808	
11		4	119.9	95.3	97.5	1.229554428	1.18443799	101.2294449	100.664		119.9	0.00558578	119.230266	
12	2018	1	79.8	99.7	103.5	0.77086656	0.8054298	99.01545706	105.267		79.8	0.06313704	84.785179	
13		2	102.5	107.2	110.5	0.928055763	0.94331835	108.6801718	109.87		102.5	0.01094798	103.642387	
14		3	126.5	113.8	116.7	1.083755448	1.0500183	120.4740903	114.473		126.5	0.04981229	120.198745	
15		4	146.3	119.7			1.18443799	123.518497	119.076		146.3	0.03596625	141.038138	MAPE
16	2019	1	103.4				0.8054298	128.3786616	123.679	103.4	103.4	0.03660781	99.6147525	5%
17		2					0.94331835		128.282	121.010764				
18		3					1.0500183		132.885	139.531682				
19		4					1.18443799		137.488	162.84601				
20	2020	1					0.8054298		142.091	114.444326				

Quarter	1	2	3	4	
2015				1.123	
2016	0.857	0.943	1.031	1.201	
2017	0.789	0.959	1.036	1.230	
2018	0.771	0.928	1.084		SUM
Average	0.805	0.943	1.050	1.184	3.983
Indices	0.80882597	0.94729594	1.05444581	1.18943228	4

