SUPERMARKET SALES AN

FINAL REPORT

Moganaviniith Rathinavel

12/11/22

BDAT 1005 – 22F
by Jonathan Gladstone

INTRODUCTION

The competition in the sales sector is always escalating in the most urban centers. Sales analysis becomes essential for the organisation to manage and expand the firm. One of the sales of a supermarket brand is included in this dataset, containing data of 3 different branches recorded for 3 months. This dataset got my attention because I can learn to analyse and understand the trends of sales by different variables present here like types of products, date and time, rating, and gross income.

This dataset is a population, not a sample. (I have changed the city names for my convenience)

DATA SET DESCRIPTION

This data set consists of 1001 records and 17 columns with a mix of both, qualitative and quantitative data types. Below is the description of each variable present in the data set.

1. **Invoice Id**: It's the unique computer-generated identification number of every printed sales slip.
2. **Branch and City**: 3 different branches of supermarket in 3 different cities categorized as A, B and C.
    A. Toronto
    B. Barrie
    C. Orillia
3. **Customer type**: Type of customer is categorized as member and normal, in which member depicts the customers with membership card and normal depicts the new and those who don't have membership card (dichotomous data).
4. **Gender**: Gender type of customer (dichotomous data).
5. **Product line**: It's a qualitative data of varied item groups such as electronic accessories, fashion accessories, food and beverages, health and beauty, home and lifestyle, sports and travel(nominal).
6. **Unit price**: Price of each product (priced in cad)
7. **Quantity**: Number of products purchased by customer (quantitative data)
8. **Tax**: 5% tax fee for applied for each invoice (continuous data).
9. **Total**: Total price including tax. (Priced in cad)
10. **Date**: Date of generated invoice
11. **Time**: Purchase time
12. **Payment**: Way of payment used by customers from 3 options available – Cash/ Credit / E-wallet
13. **COGS**: Cost of Goods Sold. (Priced in cad)
14. **Gross margin percentage –** percentage of gross margin income
15. **Gross margin income** – (gross revenue – COGS)
16. **Rating**: recorded customer satisfaction rating on their overall shopping experience on a scale of 1 to 10 (interval scale).

RESEARCH QUESTIONS AND ADDITIONAL NEW QUESTIONS OVER TIME

In data exploration part 1, initially the research questions were:

1. Which city leads in sales? On that note, which location's branch should be chosen for expansion and which category of items it should focus on?
    - As we can see, there is data of 3 cities. So, the sales among them may be compared which could answer above question
2. What is the purchasing preference of men and women? Is there any difference in category they prefer more?
    - In this dataset, there are different categories of items which can be preferred by different genders
3. Is there any relationship between COGS and ratings? Using these, shall gross income be predicted?
4. Which is the day, the products are sold maximum? And which hour of the day is busiest?

    Looking at the date and time the products bought, thought of above question

Then new some questions got added which are:

5. Product sales by product line and by city with month slicer so that data statistics can be seen for each month.
6. Same as above for each day and hour.
7. What is the rating distribution across the board?

DESCRIPTION OF DATA ANALYSIS

1. DATASET EXPLORATION PART 1.

In the first part of our data exploration, finding the appropriate dataset and asking research questions was exercised. In the same, it was observed that, measures should be taken to learn to understand and analyse the questions regarding:

- The most selling products
- Analysing preference of men and women's purchasing pattern
- What can be done to attract the non members to member's list
- Relation between COGS and ratings
- Relation between unit price and quantity with gross income

2. DATASET EXPLORATION PART 2.

2nd part of the dataset exploration was all about univariate analysis on at least 8 different variables of dataset. In this part,

- To track total price by gender, first created 2 different columns. Followed by that, formed a central tendency metrics calculating mean, median, mode, Q3, Q1, IQR, upper outlier and lower outlier range. In which female had higher upper outlier range than male.
- Likewise for all variable suitable charts are being made and tables are generated.

During this part, when analysing gross income and branch, 1st question is answered.

Pivot table shows gross income per branch. So, through table, we can see that branch c, Orillia, has highest gross income.

Pivot Table	
Row Labels	Sum of Gross Income
A	5057.1605
B	5057.032
C	5265.1765
Grand Total	15379.369

3. DATASET EXPLORATION PART 3.

This part of the exploration brought some additional research questions and performed hypothesis test such as:

1. Anova test –
    - Analysis of variation – comparing the means of a given variable for multiple groups.
    - There are 2 types of Anova (Analysis of variance test). 1st is 1-way test and 2nd are 2-way test.
    - Here we have used the anova with single factor for our hypothesis.

2. T-test on branches A, B, C.
3. Odd risk test of rating and gross income.
4. Chi-square test

5. Manova - Multivariate analysis of variation – comparing the means of multiple numerical outcome variables for multiple groups in one or more categorical independent variables.

So, through manova, using invoice id, sum of gross income, sum of cogs and coded variable, question 3. is answerable. Calculations are below:

One-way MANOVA

	stat	*F*	*df1*	*df2*	*p-value*	*part eta-sq*
Pillai Trace	0.0001831	0.0456432	4	1994	0.996075	9.155E-05
Wilk's Lambda	1.0831571	-19.498156	4	1992	#VALUE!	-0.040748
Hotelling Trace	0.0001678	0.0417519	4	1990	0.9966988	8.392E-05
Roy's Lg Root	0.0001866					

SSCP Matrices

T

136959.498	2739190
2739189.95	54783799

H

22.2791772	445.58354
445.583544	8911.6709

E

136937.218	2738744.4
2738744.37	54774887

Group Covariance Matrices

1

149.8947799	2997.8956
2997.895599	59957.912

2

133.5242317	2670.4846
2670.484635	53409.693

3

129.8576311	2597.1526
2597.152623	51943.052

Pooled covariance matrix

137.3492662	2746.9853
2746.985323	54939.706

Correlation matrix

1	1
1	1

Group Covariance Matrices

1

149.8947799	2997.8956
2997.895599	59957.912

2

133.5242317	2670.4846
2670.484635	53409.693

3

129.8576311	2597.1526
2597.152623	51943.052

Pooled covariance matrix

137.3492662	2746.9853
2746.985323	54939.706

Correlation matrix

1	1
1	1

Multiple ANOVA

ANOVA: Single Factor — Sum of Gross Income

DESCRIPTION — Alpha 0.025

Groups	*Count*	*Sum*	*Mean*	*Variance*	*SS*	*Std Err*	*Lower*	*Upper*
1	311	4798.43	15.4290418	149.89478	46467.382	0.6645581	13.937246	16.920837
2	344	5343.17	15.53247093	133.52423	45798.811	0.631879	14.114033	16.950909
3	345	5237.77	15.18193333	129.85763	44671.025	0.6309626	13.765553	16.598314

ANOVA

Sources	*SS*	*df*	*MS*	*F*	*P value*	*Eta-sq*	*RMSSE*	*Omega Sq*
Between Groups	22.279177	2	11.1395886	0.0811041	0.9221038	0.0001627	0.0153682	-0.0018412
Within Groups	136937.22	997	137.3492662					
Total	136959.5	999	137.0965941					

ANOVA: Single Factor — Sum of COGS

DESCRIPTION — Alpha 0.025

Groups	*Count*	*Sum*	*Mean*	*Variance*	*SS*	*Std Err*	*Lower*	*Upper*
1	311	95968.6	308.580836	59957.912	18586953	13.291163	278.74492	338.41675
2	344	106863	310.6494186	53409.693	18319525	12.637581	282.28066	339.01818
3	345	104755	303.6386667	51943.052	17868410	12.619252	275.31105	331.96628

ANOVA

Sources	*SS*	*df*	*MS*	*F*	*P value*	*Eta-sq*	*RMSSE*	*Omega Sq*
Between Groups	8911.6709	2	4455.835441	0.0811041	0.9221038	0.0001627	0.0153682	-0.0018412
Within Groups	54774887	997	54939.70647					
Total	54783799	999	54838.63766					

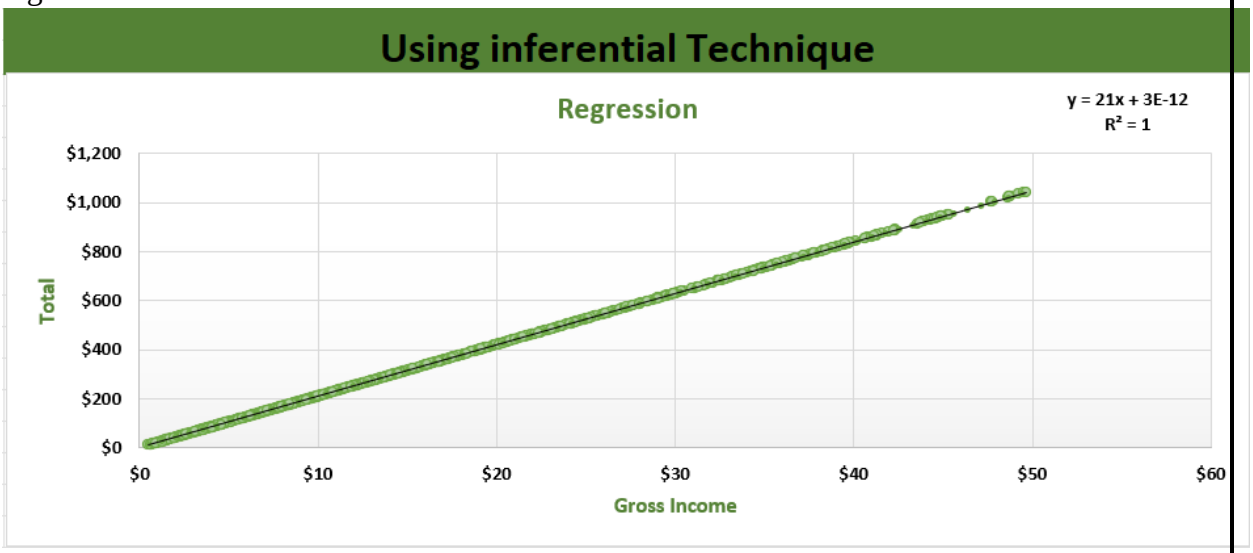
4. DATASET EXPLORATION PART 4.

Added research question.

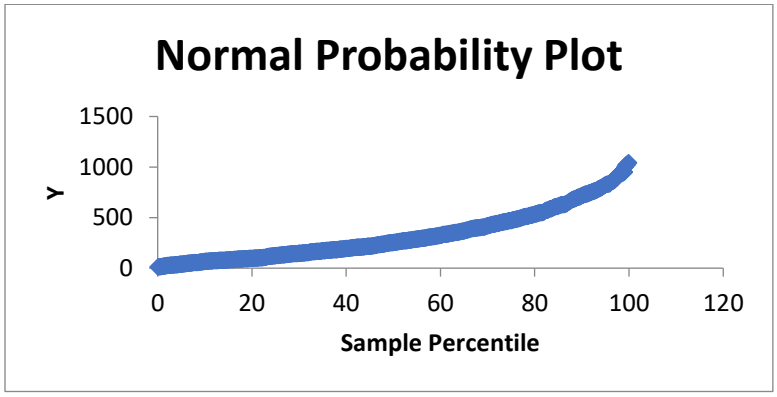
1. **relation between date and gross income. Can future income be predicted?**

In current dataset exploration, inferential techniques have been exercised containing both interpolation and extrapolation analysis.

1. **Regression analysis.**
   Here, gross income is taken as x-axis and total is taken as y-axis to perform regression.



- Probability statistics



- Summary output

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	1
R Square	1
Adjusted R Square	1
Standard Error	2.98422E-13
Observations	1000

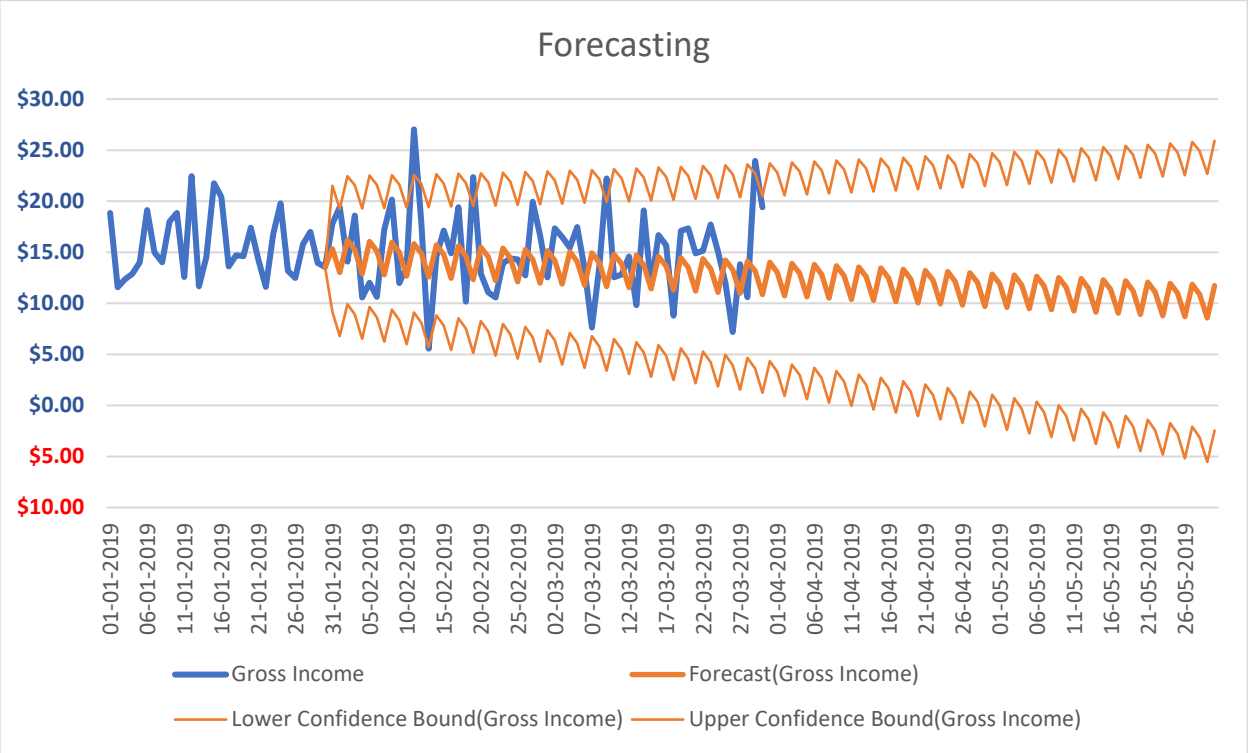
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	60399138.42	60399138.42	6.78219E+32	0
Residual	998	8.88774E-23	8.90555E-26		
Total	999	60399138.42			

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***	***Lower 95.0%***	***Upper 95.0%***
Intercept	5.68434E-14	1.55837E-14	3.647620302	0.000278369	2.62628E-14	8.7424E-14	2.62628E-14	8.7424E-14
X Variable 1	21	8.0637E-16	2.60426E+16	0	21	21	21	21

RESIDUAL OUTPUT		
Observation	***Predicted Y***	***Residuals***
1	457.443	-6.25278E-13
2	399.756	-5.68434E-13
3	470.673	-6.25278E-13
4	388.29	-4.54747E-13
5	132.762	-2.27374E-13
6	132.027	-2.27374E-13
7	621.243	-7.95808E-13
8	113.568	-1.98952E-13
9	779.31	-9.09495E-13
10	184.086	-2.55795E-13
11	177.408	-2.84217E-13
12	888.615	-1.13687E-12
13	44.5935	-1.06581E-13
14	209.622	-2.84217E-13
15	359.205	-5.11591E-13
16	383.7645	-5.11591E-13
17	138.663	-1.98952E-13
18	262.458	-3.41061E-13
19	266.028	-3.41061E-13
20	281.169	-3.97904E-13
21	367.5525	-5.11591E-13
22	217.6335	-3.12639E-13
23	44.352	-1.13687E-13
24	352.2225	-4.54747E-13
25	79.674	-1.42109E-13

PROBABILITY OUTPUT	
Percentile	***Y***
0.05	10.6785
0.15	12.6945
0.25	13.167
0.35	13.419
0.45	14.679
0.55	16.107
0.65	16.2015
0.75	16.275
0.85	17.094
0.95	18.6375
1.05	19.194
1.15	19.2465
1.25	20.1075
1.35	20.685
1.45	22.386
1.55	22.659
1.65	23.499
1.75	23.751
1.85	24.108
1.95	25.263
2.05	26.25
2.15	26.5545
2.25	26.7225
2.35	26.733
2.45	26.796

2. Forecasting



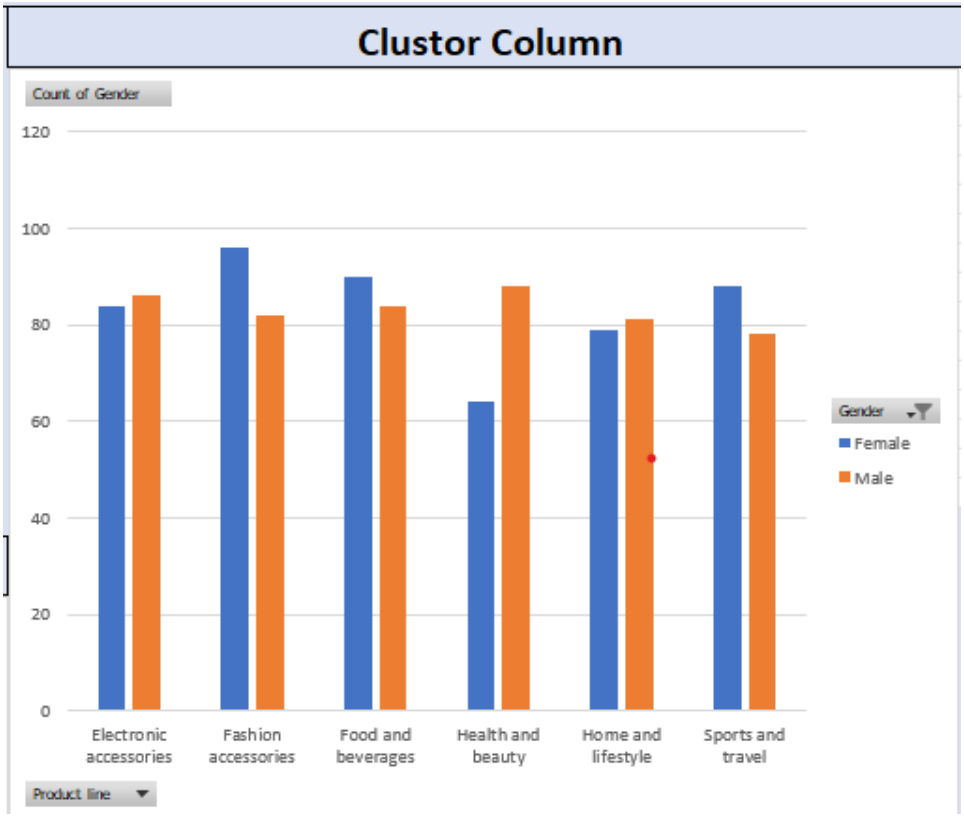
Forecasting is performed on date and gross income. Start date is took as 30/1/2019 and end date is given as 30/5/2019.

Through this result, company can estimate the future gross income, and they can plan budgets or plan demand accordingly. And this answers the updated question in this dataset exploration.

Results

- As we can see, there is data of 3 cities, when analysing gross income and branch, 1st question is answered. That is, branch C, Orillia has the highest gross income, so if the organisation is planning for expansion, Orillia can be considered as the best pick.

- Frequency of sales of diff product line is shown based on gender. In which ratio of both male and female is almost equal in purchase of electronic accessories and home & lifestyle products. Women leads in purchase rate of fashion and accessories, while men lead contributing more sales in health and beauty products.
- So, second question has been answered.



- Relation between cogs and rating by performing chi-square and OR
  1. OR

	OR			
Count of Condition Check2	Column Labels			
Row Labels	Yes	No	(blank)	Grand Total
Yes	127	198		325
No	284	391		675
(blank)				
Grand Total	**411**	**589**		**1000**
(a*d)/(b*c)	0.883073695		**1.1324**	Inverted

	CONDITION CHECK 2>15	CONDITION CHECK <15
Rating in {5:7}	a	b
rating NOT on {5:7)	c	d

2. Chi-square

Expected values			
Row Labels	**Yes**	**No**	**Grand Total**
Yes	133.575	191.425	325
No	277.425	397.575	675
Grand Total	**411**	**589**	**1000**

	Chi-square		Grand Total
Row Labels	**Yes**	**No**	
Yes	0.324	0.226	
No	0.156	0.109	
Grand Total			**0.479**

p-value for Chi-square	**0**

- Through manova, using invoice id, sum of gross income, sum of cogs and coded variable, question 3. is answerable, whose output is already given above in dataset exploration part 3's explanation part.
- By dataset exploration 4, its clearly seen that gross income can be calculated as both regression and forecasting is carried out.

Conclusion

To conclude, all the dataset exploration parts have been successfully accomplished by performing step by step analysis.

References

Historical record of sales data in 3 different supermarkets. 2019. Supermarket sales | Kaggle

Create a forecast in Excel for Windows - Microsoft Support